Humanities Data and their Research Use

Open Science Infrastructures for Big Cultural Data
International Advanced Masterclass
Plovdiv, Bulgaria ❖ 13—15 December 2018

Daniel Paul O'Donnell
University of Lethbridge

DOIs:
10.5281/zenodo.2246389 (latest).
10.5281/zenodo.2246390 (this version).

About this paper

- Going to be speaking of how data are used in the humanities
- Background is *small* data:
  - 9-line Anglo-Saxon poem (http://caedmon.seenet.org/);
  - 5 object digital library (http://visionarycross.org). [hacked and being rebuilt]
- But data that are treated as data
  - FAIR (Findable, Accessible, Interoperable, Reusable)
  - Open
  - Focus on long-term preservation
- And data as they are used by Humanists in the Humanities
  - Very traditionally trained Germanic Philologist and Medievalist

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

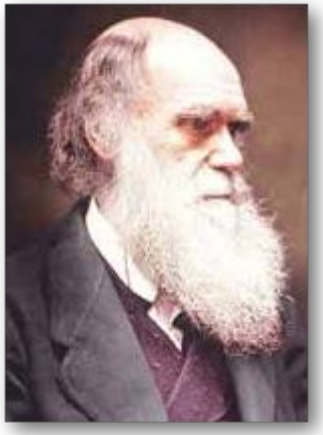
- Our definitions are highly contingent
  - “Primary source” in one context, can be the “secondary source” in another (and vice versa)
  - Or simultaneously “Primary” and “Secondary” (e.g. a critical edition)
- Also hard to constrain

  “[a]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?

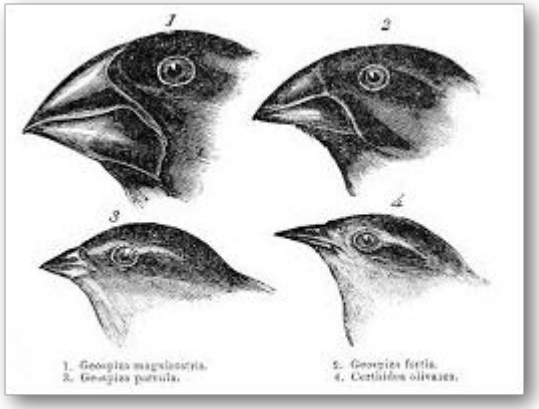
How does data work in other fields?

- 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 traditional humanities, “data” can be both “data” and “capta”, but most often ~“data”
- Interest is specific and often provisional; depend on understanding of purpose, context, identity, and form that are also open to analysis and modification
- We might base our work on ~“capta” (e.g. editions?), but also work from interpretation and without clear intermediate stage

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 to form large data sets that 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 method questions 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?
- Analogy to (and valorisation of) sciences can make this all quite challenging and disturbing
- How does it interact with our (largely intuitively understood) “humanistic method”?

Fish 2012: Minding your P’s and B’s

- 2012 New York Times “Opinionator” column
- Argues “against” Digital Humanities by attempting to demonstrate something it “can’t” (or doesn’t) do: provide close reading of *Areopagitica* (Milton)
- Fish argues that Milton understands censorship in Protestant England as a kind of *de facto* counter-reformation
- A repeat of the previous mistakes made by the Catholic and Episcopal churches

Fish 2012: Minding your P's and B's

Halfway through the Areopagitica *(1644), his celebration of freedom of publication, John Milton observes that the Presbyterian ministers who once complained of being censored by Episcopalian Bishops have now become censors themselves. Indeed, he declares, when it comes to exercising a "tyranny over learning," there is no difference between the two: "Bishops and Presbyters are the same to us both name and thing." That is, not only are they acting similarly, their names are suspiciously alike.*

Fish 2012: Minding your P’s and B’s

- This is also reflected in the sound pattern of the piece.

  In the sentences that follow the declaration of equivalence, “b’s” and “p’s” proliferate in a veritable orgy of alliteration and consonance. Here is a partial list of the words that pile up in a brief space: prelaty, pastor, parish, Archbishop, books, pluralists, bachelor, parishioner, private, protestations, chop, Episcopacy, palace, metropolitan, penance, pusillanimous, breast, politic, presses, open, birthright, privilege, Parliament, abrogated, bud, liberty, printing, Prelatical, people.

Became methodological/theoretical battleground

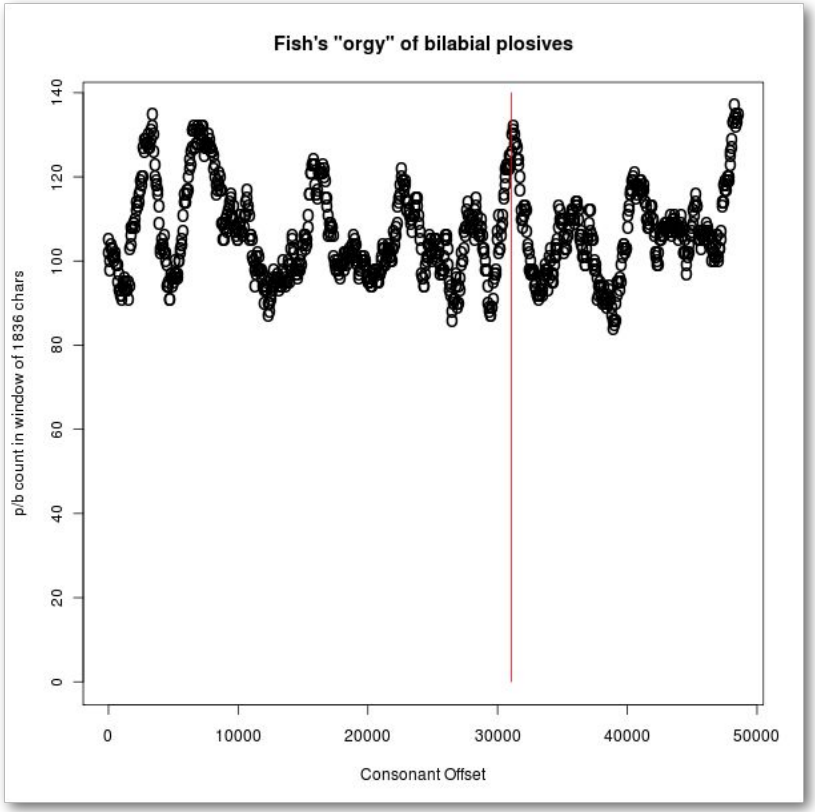
- Fish's piece was intended to contrast against "DH method":

  I began with a substantive interpretive proposition... and, within the guiding light... of that proposition I noticed a pattern that could, I thought, be correlated with it. I then elaborated the correlation. The direction of my inferences is critical: first the interpretive hypothesis and then the formal pattern, which attains the status of noticeability only because an interpretation already in place is picking it out….

  The direction is the reverse in the digital humanities: first you run the numbers, and then you see if they prompt an interpretive hypothesis…. You don't know what you're looking for or why you're looking for it. How then do you proceed? The answer is, proceed randomly or on a whim, and see what turns up. You might wonder, for example, what place or location names appear in American literary texts published in 1851, and you devise a program that will tell you. You will then have data.

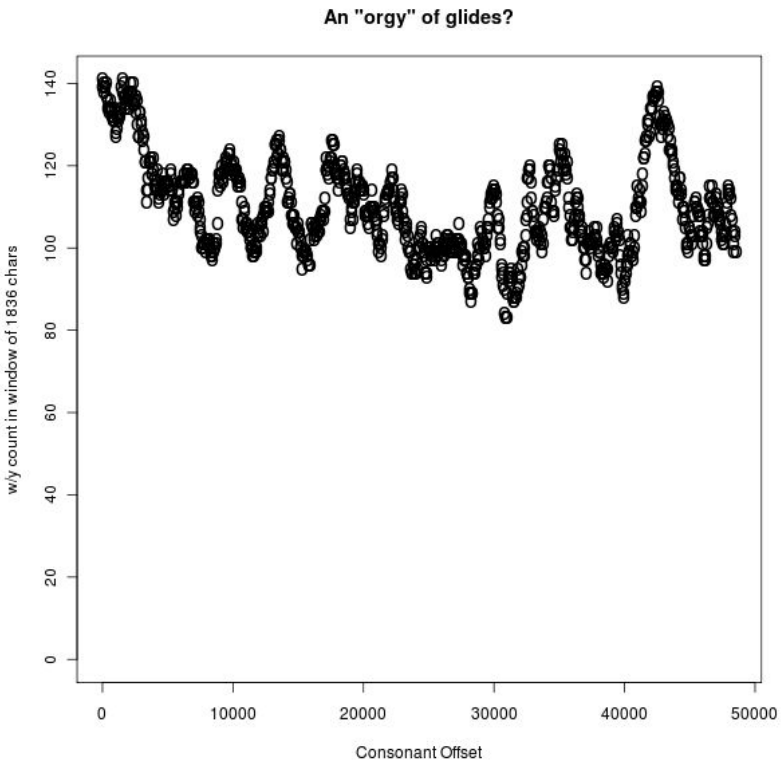
Liberman 2012: Falsifying Fish

- Most disturbing of those who took up Fish's challenge was Mark Liberman
- Did what a scientist might do: attempt to falsify his conclusions with additional data:
  - First looked at the distribution of Ps and Bs in the *Areopagitica*



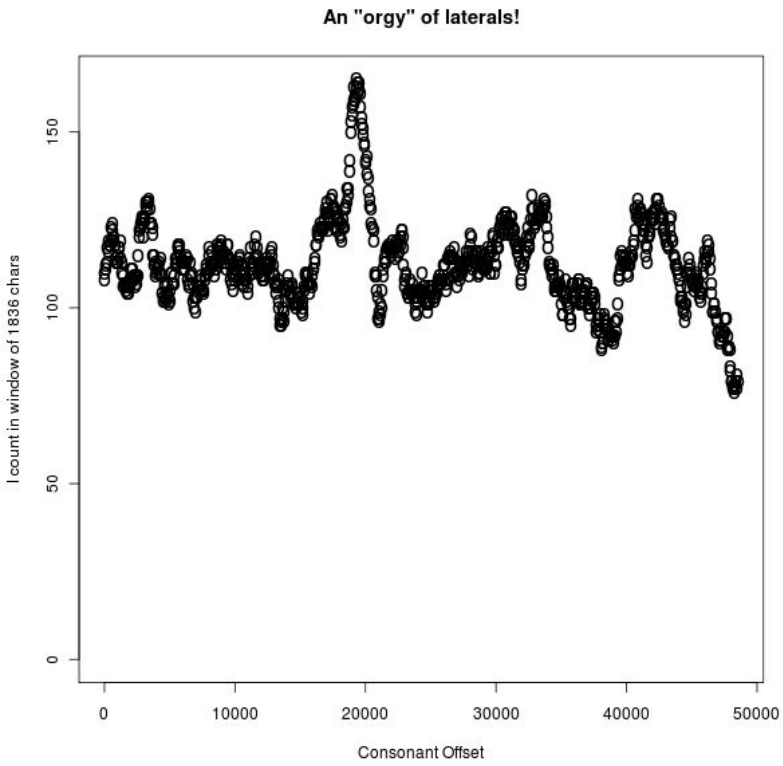
Liberman 2012: Falsifying Fish

- Most disturbing of those who took up Fish's challenge was Mark Liberman
- Did what a scientist might do: attempt to falsify his conclusions with additional data:
  - First looked at the distribution of Ps and Bs in the *Areopagitica*
  - Then looked at the distribution of w, y, and l



Liberman 2012: Falsifying Fish

- Most disturbing of those who took up Fish's challenge was Mark Liberman
- Did what a scientist might do: attempt to falsify his conclusions with additional data:
  - First looked at the distribution of Ps and Bs in the *Areopagitica*
  - Then looked at the distribution of w, y, and l



Liberman 2012: Falsifying Fish

- Concluded that this falsified his argument

  Prof. Fish begins with an "insight" about the alleged dance of p's and b's surrounding Milton's assertion that ""“Bishops and Presbyters are the same to us both name and thing". Despite the paradoxically semi-quantitative nature of his idea, he presents it as an example (though clearly not a very interesting one) of the kind of literary analysis to which "digital humanities" methods are not relevant, the kind of "criticism that insists on the distinction between the true and the false, between what is relevant and what is noise, between what is serious and what is mere play". But it seems to me that a trivial application of statistical methods, humanistic or not, suggests that his idea is probably "false", "noise", and "mere play". Have I missed something?

An important exchange

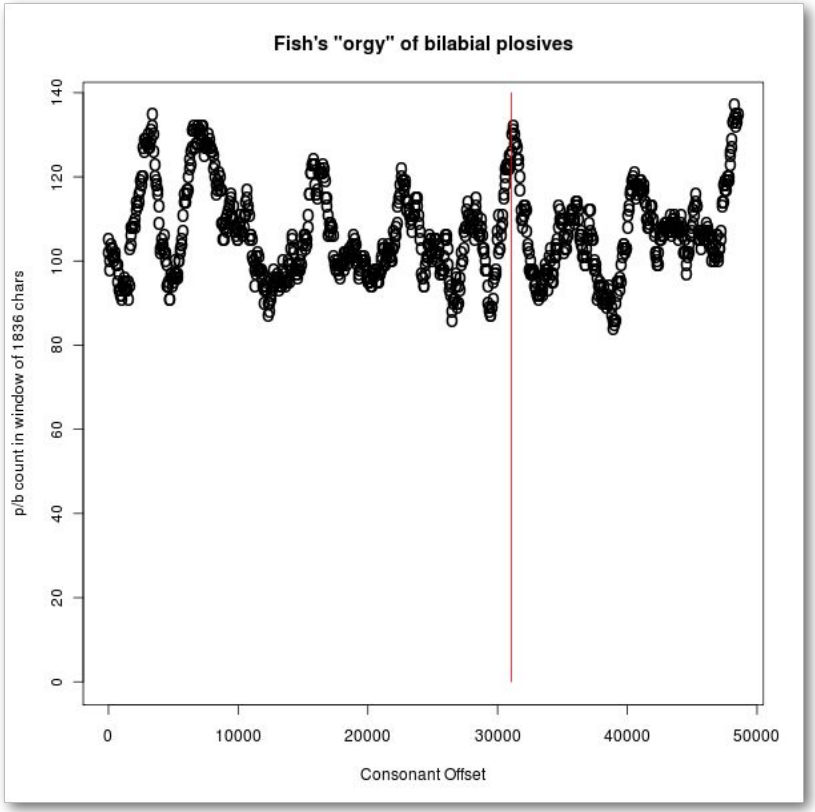
- In a domain in which the fundamental questions of method have not received much attention by practitioners, these posts are both about fundamental questions of evidence, discovery, and argumentation.
- If Fish is right about the degree to which DH requires us to do the opposite of what he is doing, then it represents a fundamental break with at least 125 years of previous work.
- But if Liberman is right, then DH isn't just a fundamental *break* with previous ways of doing things, it is a fundamental *threat*
    - Introduces a new test that had not been used before: falsification.
    - But is showing up more frequently (cf. Matt Jockers vs Ian Watt and the *Rise of the Novel*).

An important exchange

- Fortunately, Liberman isn't actually right in his exchange with Fish--i.e. he doesn't falsify him
- Although Fish is sloppy in his terminology, he's not actually making a hypothesis-driven analysis of data he's collected
- Rather, he's providing an inductive, thesis-driven *reading* of a historical text:
  - Not a claim Milton did this on purpose;
  - Not a claim you can't read the text any other way nor that no other consonants (or vowels or anything else) is important;
  - Just an argument that at this place, Ps and Bs interact in a way that can be read as supporting Milton's argument.
- And having established this thesis, he went out and found evidence for it

An important exchange

- And on these terms Liberman shows that Fish is being reasonable:
  - Shows that Ps and Bs peak where Fish says they do (one of the main peaks)
  - As Fish says, the two sounds are similar
  - That other sounds have other distributions isn't important to argument
- Even improves it because it shows other places to look!



So what have I been doing here?

- Point of this talk has been to disparage neither the (data-driven) Digital Humanities nor the (sometimes more impressionistic) traditional Humanities
- Rather it has been to point out
  - Some fundamental differences between data as we understand them in the Humanities
  - Methodological implications (and origins) of those differences
- Data-driven DH (big or small) is going to open new vistas for work in our domain
- But we have to remain vigilant and sensitive to what it is we ultimately do with these things once we have both “data” and “capta”
- In a field that is not methodologically precise, this is going to be a core challenge

Thank you

DOIs:
10.5281/zenodo.2246389 (latest).
10.5281/zenodo.2246390 (this version).