THE CHALLENGES OF DIGITAL HUMANITIES: COMMON REQUIREMENTS FOR HUMANITIES RESEARCHERS

JAMES O’SULLIVAN
UNIVERSITY OF SHEFFIELD

@JAMESCOSULLIVAN
JOSULLIVAN.ORG
DIGITAL HUMANITIES INSTITUTE
@DHISHEF

• ~20-25 live projects, all externally funded
• External partners in higher education and other public and commercial sectors
• “DH as service”
• Director: Michael Pidd

• Digital Humanities Research Associate
• “Build capacity”
• Computer Science / Literary Studies
• Electronic Literature / Computer-assisted Criticism
OVERVIEW

- Defining DH
- Requirements & Challenges
- Case Studies
WHAT IS A DH PROJECT?

It can be big…

Linguistic DNA
Modelling concepts and semantic change in English 1500–1800

Computational linguistic techniques and data visualisation to identify lexical patterns in 250,00 printed books (approx. 30 million pages).

but it doesn’t have to be…
THE DH LIFECYCLE

Acquisition & Processing
- Scope
- Preparation
- Management

Analysis
- Adding value
- Visualisation
- Interpretation

Dissemination
- Representation
- Sharing
ACQUISITION & PROCESSING

• Scope
  – What is your research question?
  – How can computation help you answer that question?
  – What data do you need? How can you get it?
  – “Big data” is relative to current infrastructure and processing conventions
  – Total archive of documents held by The National Archives dating up to the 1970s is less than the Home Office’s annual deposits
ACQUISITION & PROCESSING

How you prepare and structure your data can have significant repercussions for the questions you ask and the results you receive…
ACQUISITION & PROCESSING

Gathering your data

• How reliable is your source?
  – Do you need to clean the content?
    • OCR vs “born-digital”
    • Outsourced, in-house, or crowdsourced transcription?
    • How clean is clean?
  – Which edition are you getting?

• What labour will be required?
  – Correcting errors
  – Spelling conventions
  – Removing boilerplate
  – Critical commentary etc

• Copyright/licensing issues

British Library Nineteenth Century Newspapers
Keyword search for “pidd” gives 2,730 results…

higher forms have access to a good peculiar attention is paid to the Exercises taking and Classical Composition.

Bid me rise—
And cheer in endless smart—
Bid me a weary wanderer be—
But never bid me part from Thee

in the morning till eight in Cheapside, where any con Moore’s interest... will he

Regularly cited as the most used online resource by Jisc, but it has an extremely high error rate; little more than a substitution for microfilm
And all of that is dependent on…

Datasets in the Humanities are usually:
1. Small (discrete sources created by individuals)
2. Broad (many different types of sources have to be assembled)
3. Complex (because humans are not spreadsheets)
Management

Data management is all about *stewardship*! It is an ongoing commitment in any DH project.

“An effort to deliberately care for information resources that enable and inform research saves time and effort both during a project and after it has ended or evolved along new paths of inquiry.”

“A dedicated effort to care for the physical or even digital content of relatively traditional studies aided by our computers seems like a reasonable undertaking, but it will not happen on its own – we have to commit to it!”

-- Sarah Pickle (Assessment Librarian at the Claremont Colleges)
ACQUISITION & PROCESSING

• Practical Considerations
  – Ensuring your work has the best chance at being useful to you and others
  – Down the line, you know where and what your materials are; where they came from, how you produced them in the first place, and how to work with them.
  – Create a plain text README!

• Technical Considerations
  – What is your data’s current and long-term storage needs?
  – How are you structuring your dataset? Naming conventions etc.
  – What backup policies are in place?
  – What formats are you using? Plain text, XML, TIFF, jpg?
ANALYSIS

• The application of sophisticated computer-assisted methods to Humanities research
• What is the added value?
• DH methods aren’t better, they’re just different; they lend new types of evidence
• DH methods allow us to read in different ways, lending quantitative evidence to qualitative arguments
• Capacity: machines are stupid, but they can process information a lot quicker than we can
“We cannot say precisely what literature is, but we can recognise the literary when we see it, and by extension, the various characteristics of that which makes it so. If we view literary texts as systems, however formal or informal, these systems will be comprised of a series of elements, which, while detectable yet insignificant to a machine, can be useful to a human in the construction of meaning. Digital Literary Studies seeks to equip literary and cultural scholars with the instruments necessary to isolate specific literary elements and to use these to conduct some experiment or calculation in an effort to provide additional insight.”
**ANALYSIS**

**Visualisation & Interpretation**

- How are you going to represent your results?
- Macro-level information is most intuitive when visual.
- Visualisation is not just about dissemination, it’s about interpretation.
- Detecting trends and anomalies.
- Misinterpretation is always possible.
  - Discoveries vs mistakes.
  - **Researcher** determines significance.
• Will your analysis be limited by your data?
• Irish Literary Network
• (This is on top of typical issues like canonicity etc...)
What constitutes a finding?
ANALYSIS

• How do we present our findings to Humanities scholars?
  – Sean G. Weidman, “The Limits of Distinctive Words: Re-evaluating the Gender Marker Debate”
    Forthcoming in *Digital Scholarship in the Humanities* (2017)

• Should we be conducting experiments that aren’t reproducible?

• How do you overcome the scholarly issues posed by black box tools?

• Is collaboration the answer?
  O’Sullivan, James, Diane Jakacki, and Mary Galvin. “Programming in the Digital Humanities.”

• Understanding is essential, coding is not
DISSEMINATION

Representation & Sharing

• **If you can** share your dataset, **you should**
  – Reproducibility!

• Datasets are most effectively shared with others when accompanied by documentation concerning how it was obtained and “massaged”

• Where do you publish your research?
  – Many editors / reviewers are still very suspicious of DH methods
  – If we just continue to publish in DH journals, are we simply building a silo?
  – Who is our audience?
THANK YOU!

j.c.osullivan@sheffield.ac.uk
@jamescosullivan
josullivan.org