Responsible Data Collection

Lucia Michielin, Digital Skills Training Manager

Jessica Witte, Digital Research Analyst

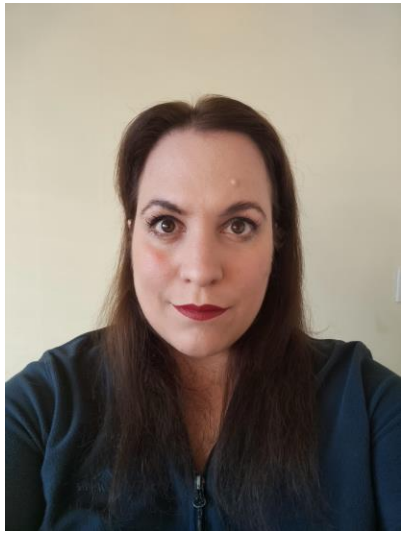
Introductions

Lucia Michielin

Digital Skills Training Manager

PhD in Computational Archaeology

Specialised in: Webscraping, Text and Data Analysis, Data Visualisation, GIS, 3D reconstructions, Photogrammetry



Jessica Witte

CDCS Digital Research Analyst

PhD in Literary Studies

Specialised in: text analysis, natural language processing, web scraping, generative AI

Plan for Today

Data Collection for Research

- Important part of project design
- Methodologies for acquisition and analysis vary by research question and data type
- Often a challenging component of a digital research project—methodologically, ethically, and even legally

Image from Unsplash

Discussion

Have you ever faced (potential) ethical or legal challenges in acquiring data?

If so, what happened?

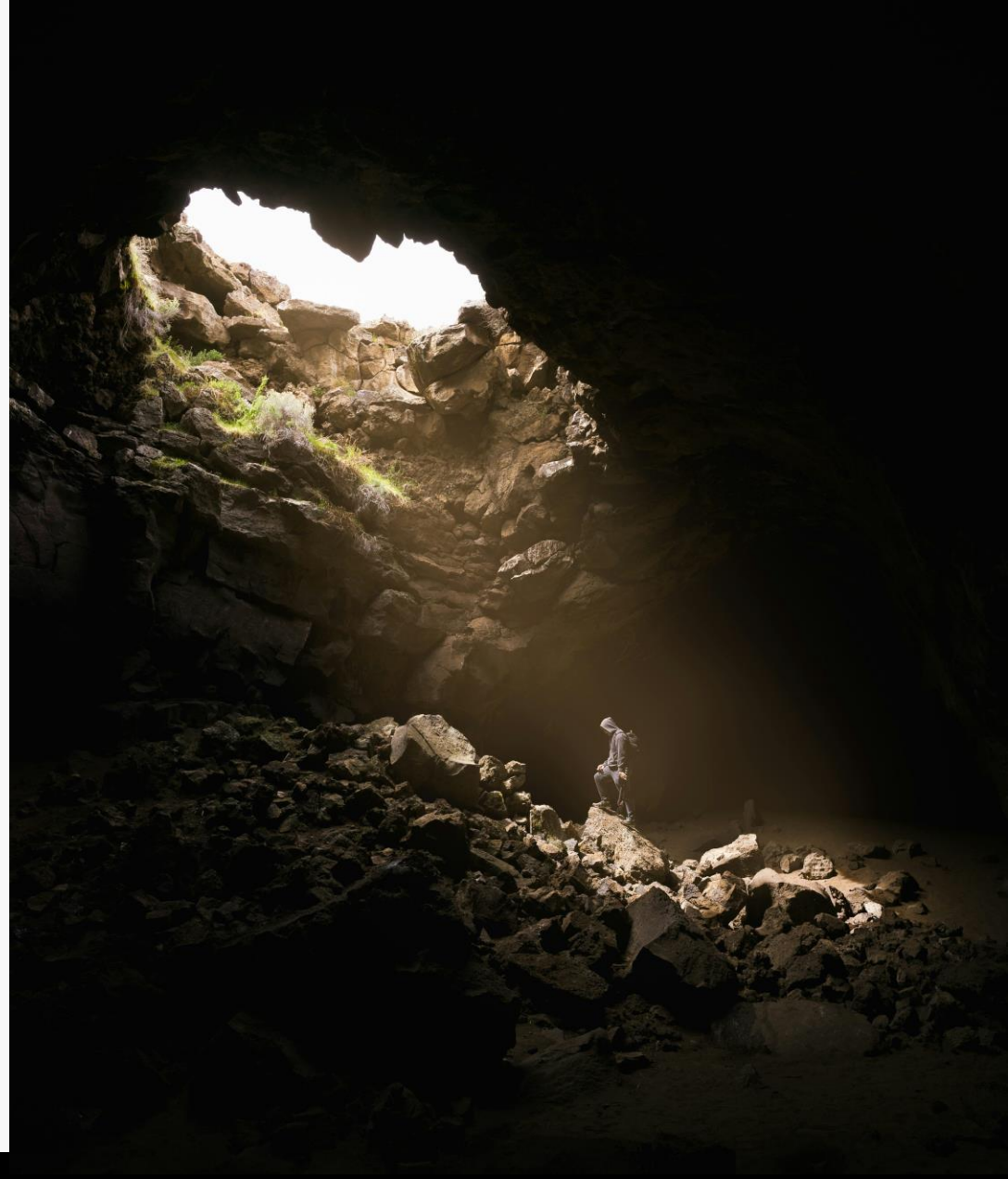


Image from Unsplash

Digital Humanities Data Sources

- Books
- Newspapers
- Magazines
- Websites
- Transcriptions of audio
- Social media

NB! Always read the licensing/copyright information and terms of use
[Text and data mining for non commercial research exception]

Image from Unsplash

Finding Data

Libraries – British Library Datasets, NLS Data Foundry (data.nls.uk)

Project Gutenberg (Gutenberg.org)

Hathi Trust Digital Library (hathitrust.org)

Websites – Internet Archive (archive.org)'s Wayback Machine

UK Web archive(webarchive.org.uk)

Newspaper archives (Universities often subscribe to them)

Social media data – More difficult now but still options

Image from Unsplash

Methods for Obtaining Data

Optical character recognition (OCR) and **handwritten text recognition** (HTR)

- Digitises documents in a **machine-readable** format that can be searched, edited, and analysed computationally
- Code and code-free options available

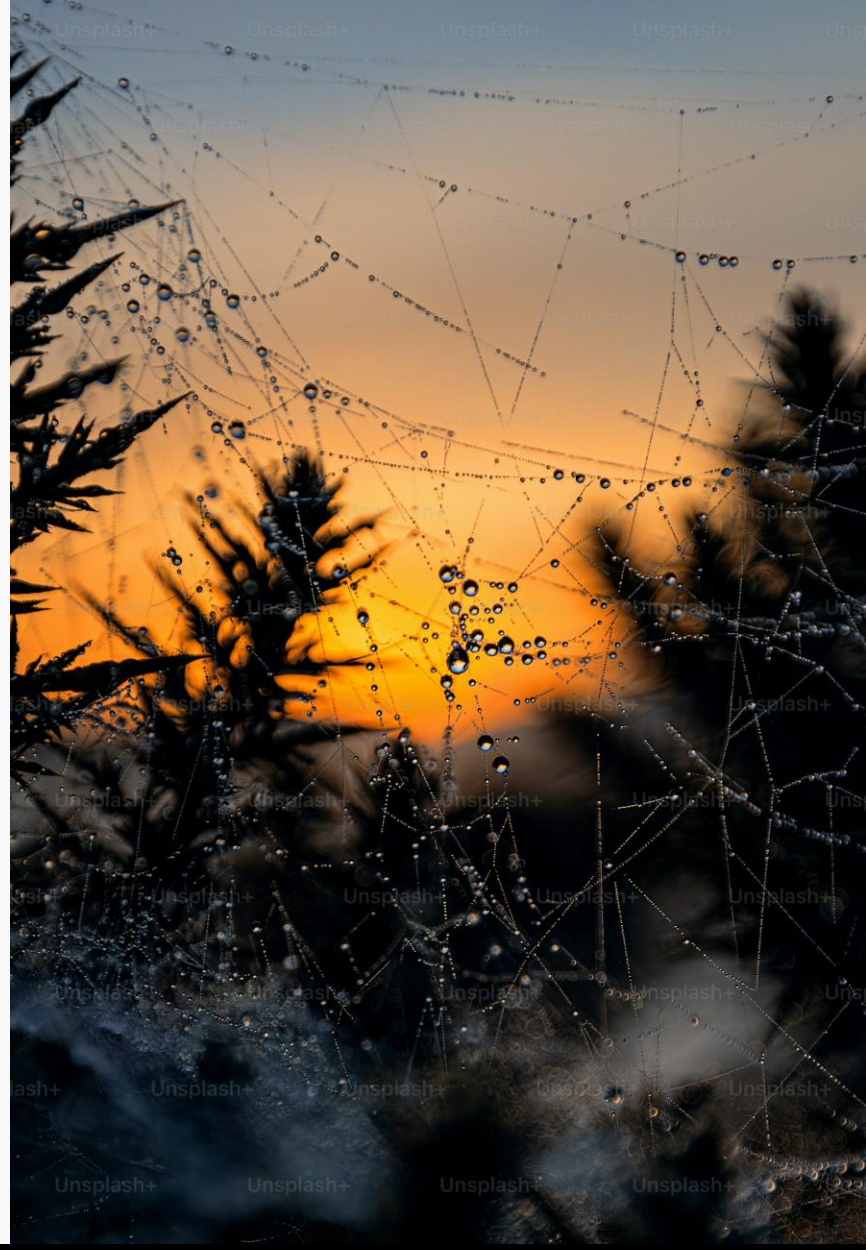
Downloading digital resources

- Individual file download
- Bulk downloading (e.g. using an API)

Web Scraping or web crawling

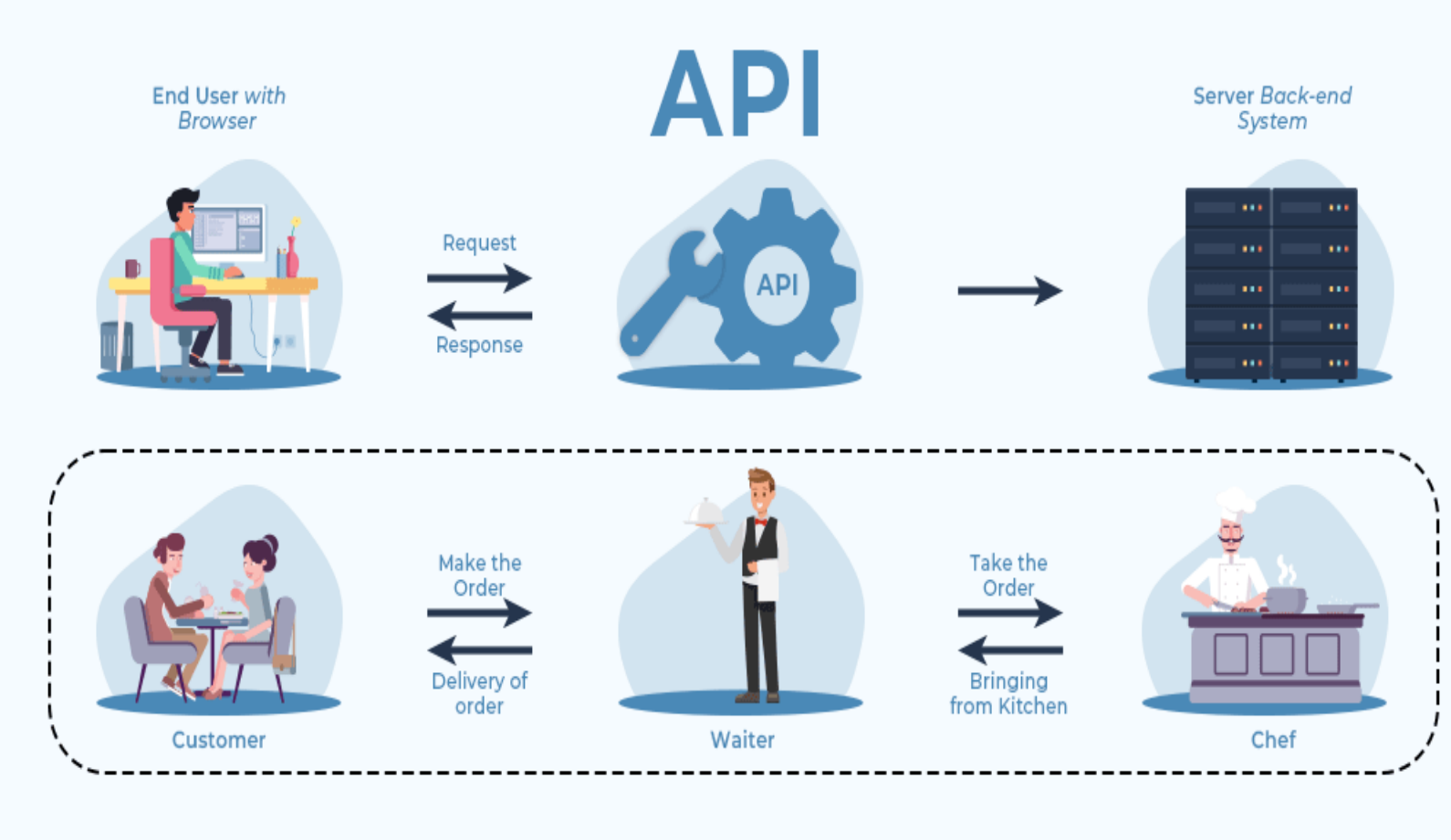
- Crawling static websites (e.g. forums or news sites)
- Techniques such as browser automation for scraping dynamic websites (e.g. content-focused platforms like social media)

Image from Unsplash



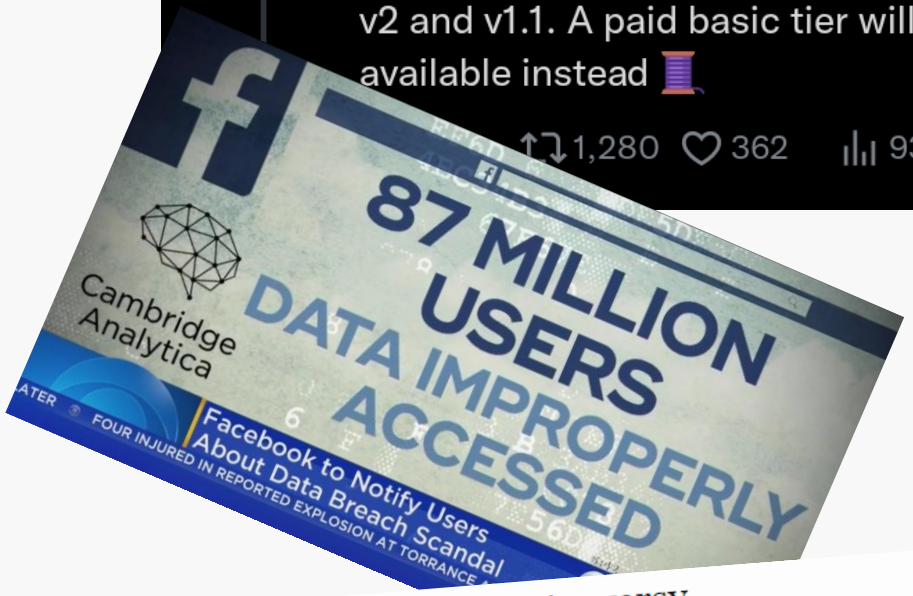
API

- Application Programming Interface
- Software that connects your computer to another computer or server for data transfer
- Facilitates batch data collection that is (relatively) user-friendly



Ethical and Legal issues: The ‘Post-API ERA’

- Extracting social media data using platforms’ APIs has become standard research practice
- Recently, some platforms have paywalled or closed their APIs
- A consensus has yet to be reached on how to ethically acquire data from these platforms
  - Paying for data
  - Finding a new source of data
  - Collecting data through web scraping




2023 Reddit API controversy

Article Talk

From Wikipedia, the free encyclopedia

In April 2023, the discussion and news aggregation website Reddit announced its intentions to charge for its application programming interface (API), a feature which had been free since 2008, causing a dispute. The move forced multiple third-party applications to shut down and threatened accessibility applications and moderation tools.

Responsible Data Management

Ethics

- Personal and private data
- Importance of research on topics like misinformation
- Green digital research
- Transparency and open research

Legal

- Copyright
- Platform governance
- ToS/ToC

Where is the balance?
Who should decide?

CDCS Mashup image

Terms of Service (ToS)

- Also called ‘Terms of Use’
- The **fine print** users agree to when accessing a particular platform or site
- Most people accept the ToS without reading them
- **Violating ToS can lead to a range of consequences**
  - rate limiting
  - account deletion
  - possible legal action
  - for researchers, a breach of ethics

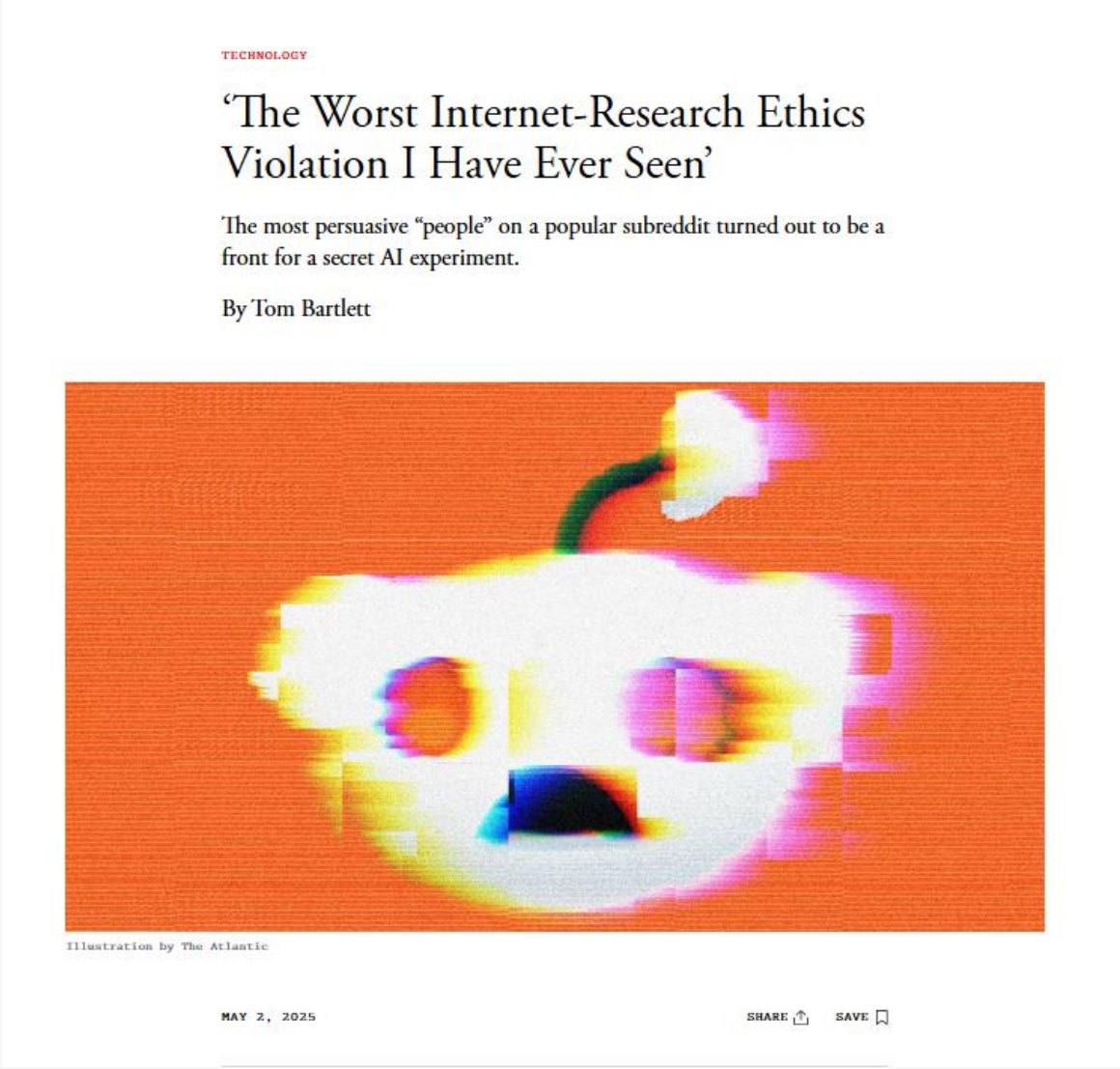
The Importance of Responsible Data Collection

Case Studies:

1. The University of Zurich Reddit Study (2025)
2. The Aarhus University OK Cupid Study (2018)

University of Zurich Reddit Study

- Examined whether AI-generated comments could change people's opinions in the Reddit community r/changemyview
- AI-generated responses were tailored to users based on their profiles
- After the experiment, the researchers debriefed the community
- Reddit users found the experiment "'violating,' 'shameful,' 'infuriating,' and 'very disturbing.'"



TECHNOLOGY

'The Worst Internet-Research Ethics Violation I Have Ever Seen'

The most persuasive "people" on a popular subreddit turned out to be a front for a secret AI experiment.

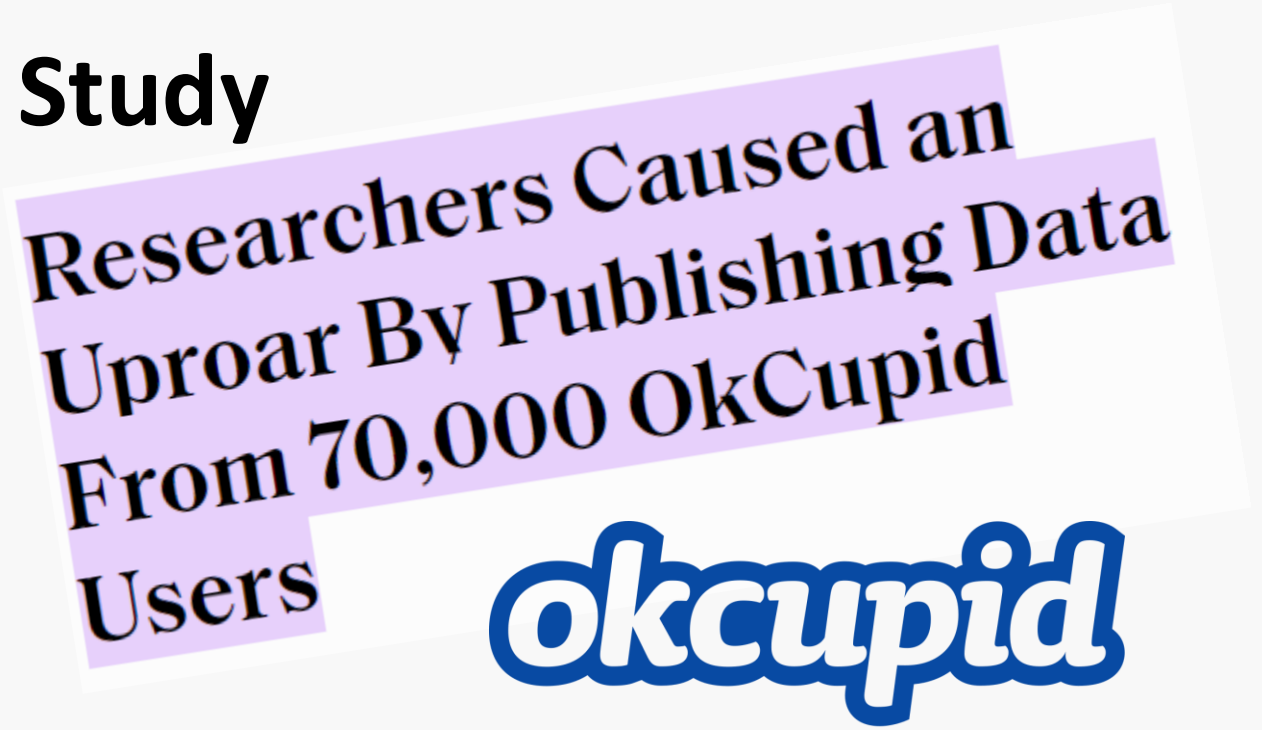
By Tom Bartlett

Illustration by The Atlantic

MAY 2, 2025 SHARE SAVE

Aarhus University OKCupid Study

- In 2016, researchers analysing the dating site OkCupid published their dataset on the Open Science Framework
  - The data included users' personal information
  - Researchers argued the data was already in the public domain
- Open research is not always ethical (and vice-versa)



"No. Data is already public."

—The study's lead researcher, when asked whether the data was anonymised

Open Science & Fair Data

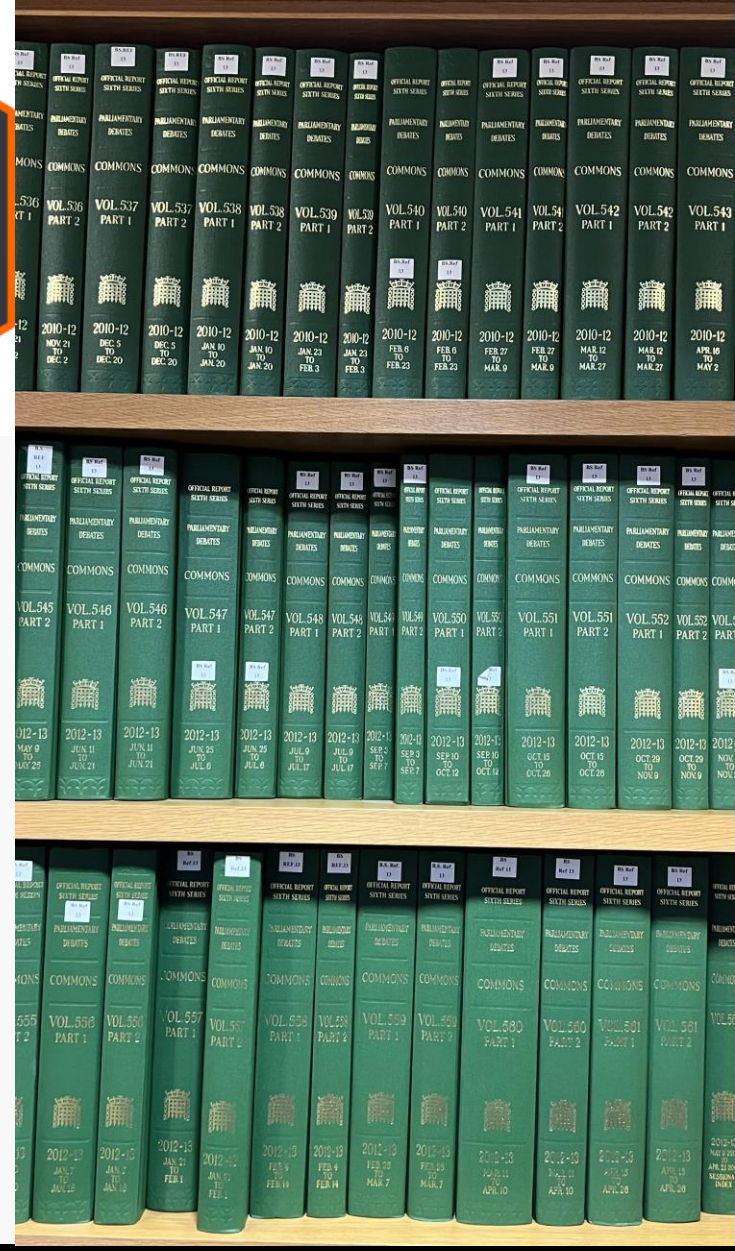
- Transparency, reproducibility and sharing
- Open-source data, software, resources, hardware, publications
- Engaging the public through crowdsourcing, citizen science, crowdfunding, and collaboration
- Inclusion of historically marginalised groups

Data should be:

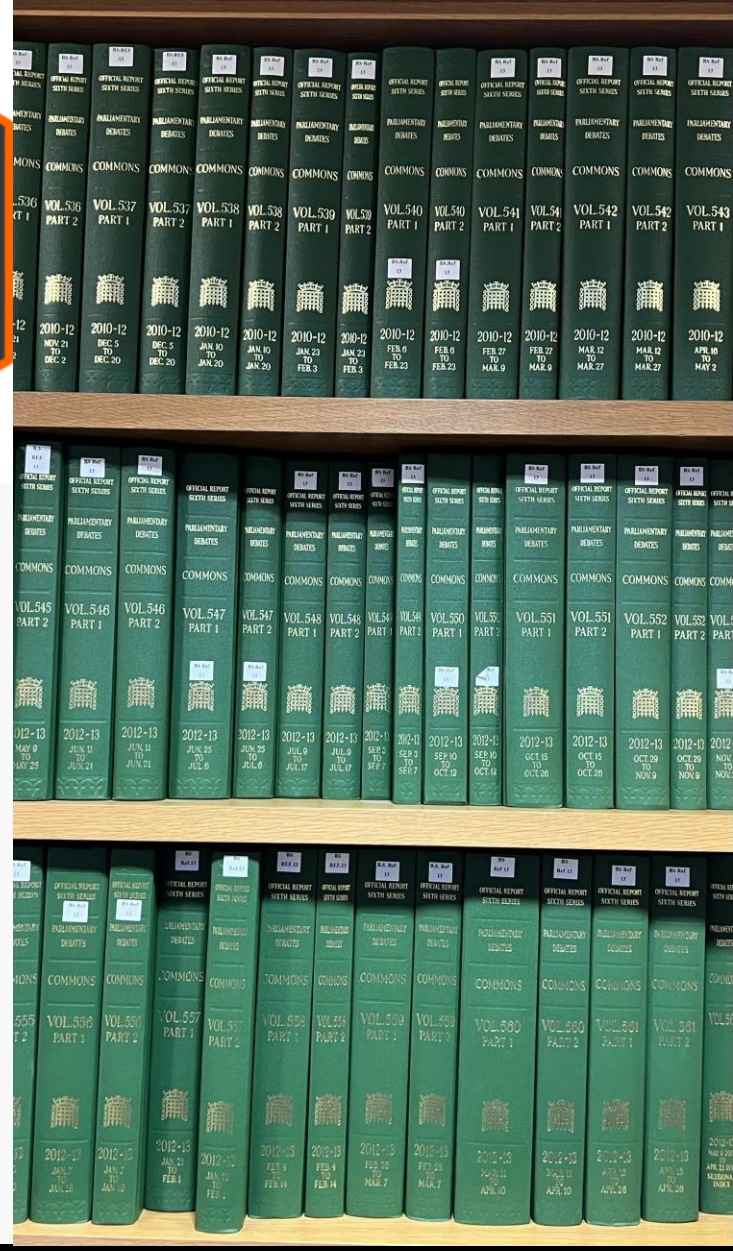
- **F**indable
- **A**ccessible
- **I**nteroperable, or compatible with other data, systems, and technologies
- **R**eusable for other purposes

The Hansard Dataset

- **Hansard** is a database of **official transcripts** of the debates and proceedings in UK parliament
- It provides a **verbatim** (word-for-word) record of what was said by members during sessions.
- Published **daily** or **periodically**, depending on the parliament
- Named after **Luke Hansard**, a printer for the British Parliament in the early 19th century
- Used for **public accountability**, research, legal references, and historical records



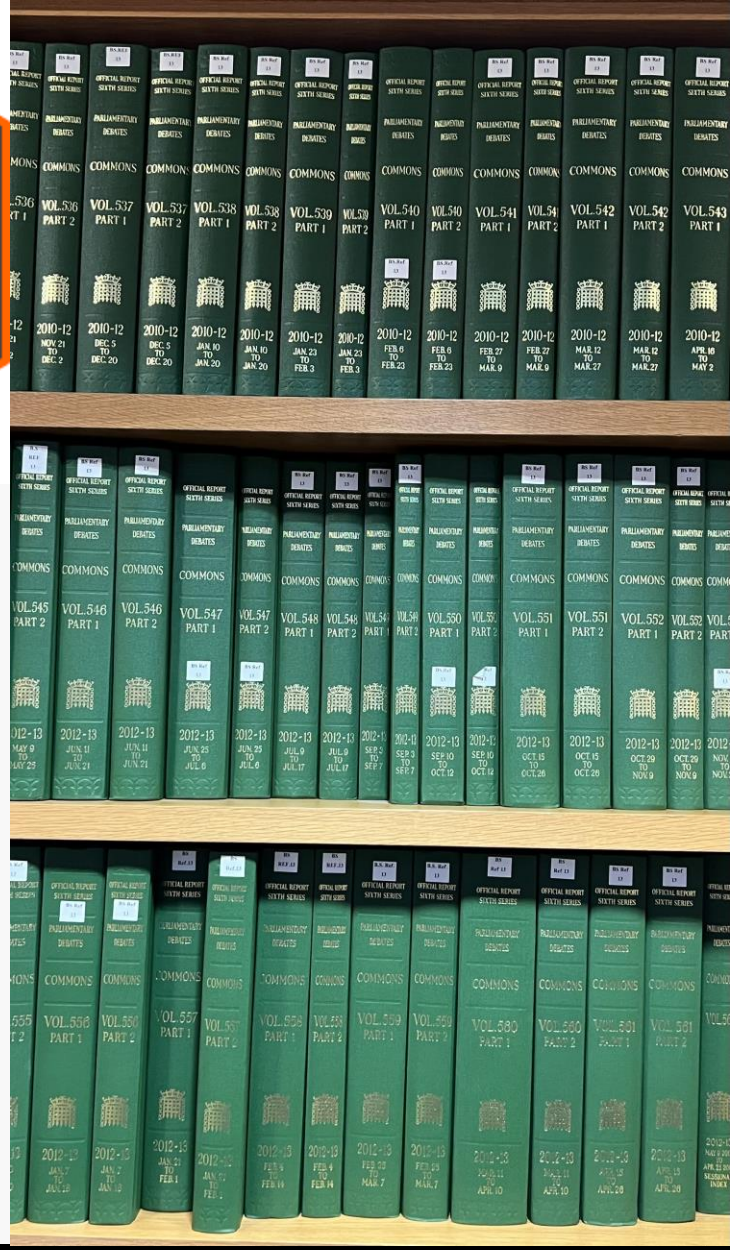
The Hansard Dataset for Discourse Analysis



- You cannot go more open and public of this but…
- Analyse evolution political discourse on immigration and asylum
- Need a large corpus to work on
- Too long to collate it manually

Surely it will be all nice and straightforward, isn't it?

Ethical Approach to Web Scraping?

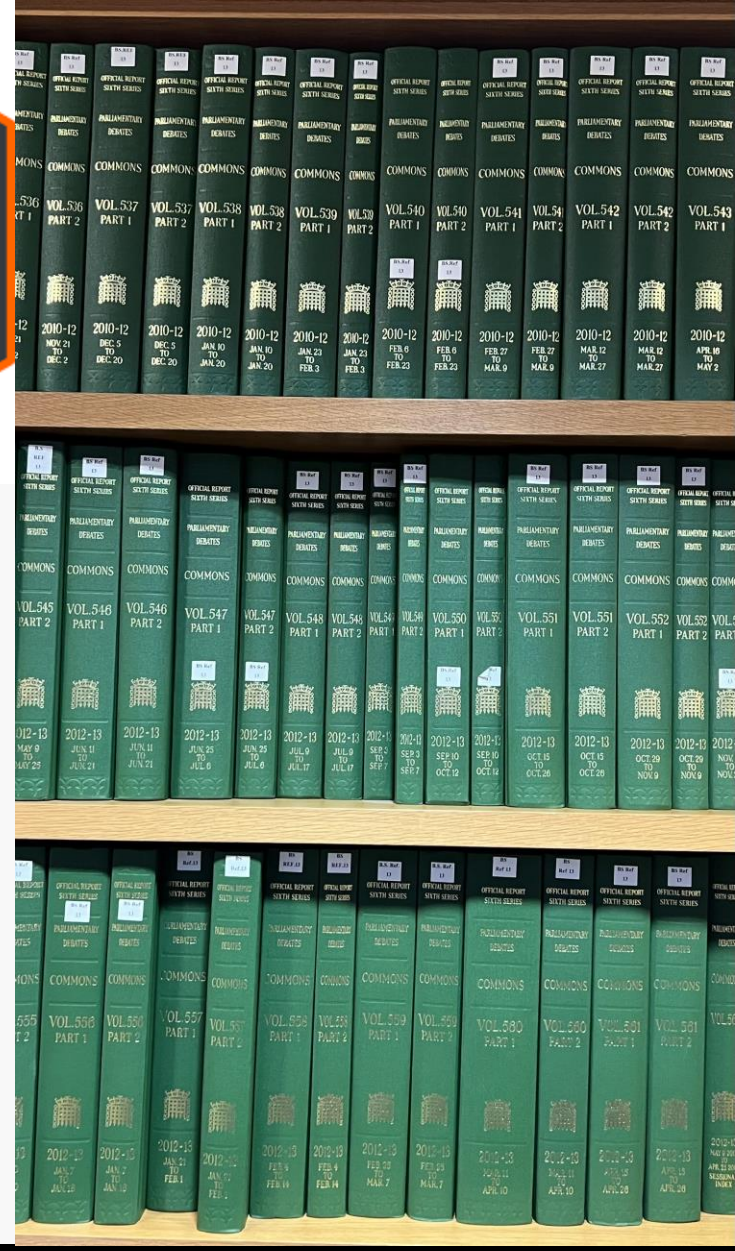


- Tried standard web scraping, but encountered Cloudflare blocking
- Empty Robots.txt page
- API limitations
- R package not working on Hansard data
- Contacted the data provider (Hansard) - No answers

Semi-automatic workflow using API+ Open Refine

More Questions than Answers...

- Is it Legal?
- Is it Ethical?
- How can we assure that Open Data/FAIR data are really so?



More questions than answers. . .

- Where and when do digital humanists encounter challenges in collecting, analysing, and storing/sharing their data?
- Where frameworks for best practices conflict, how should digital humanists make data management decisions?
- We consider the risks of research projects, but what are the risks of *not* doing research on certain topics?
- As a DH community, can/should we come to a consensus?

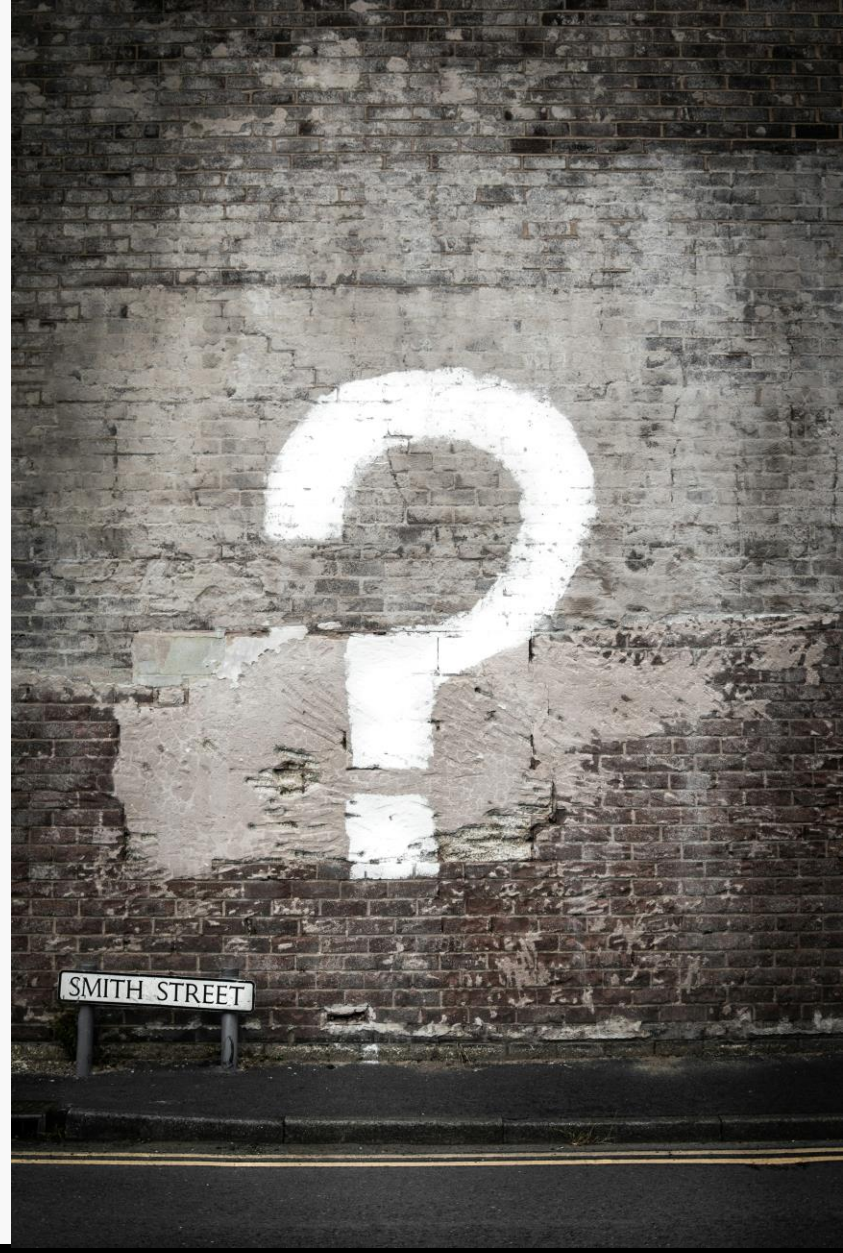


Image from Unsplash

Responsible Data Practices

- **Project planning**
  - Locate the data
  - Determine whether it can be collected (and how)
  - Determine whether the data contains sensitive or personal information
  - Review applicable legal frameworks—e.g. GDPR, ToS

- **Data Acquisition**
  - Consider the carbon costs of various data collection and storage methods
  - Where relevant and feasible, anonymise the data
  - Ensure data is stored in a secure location

Image from Unsplash

Responsible Data Practices

- **Data Preservation**
  - Determine whether the data should be shared and if any restrictions should apply
  - Consider costs of long-term storage—financial and environmental

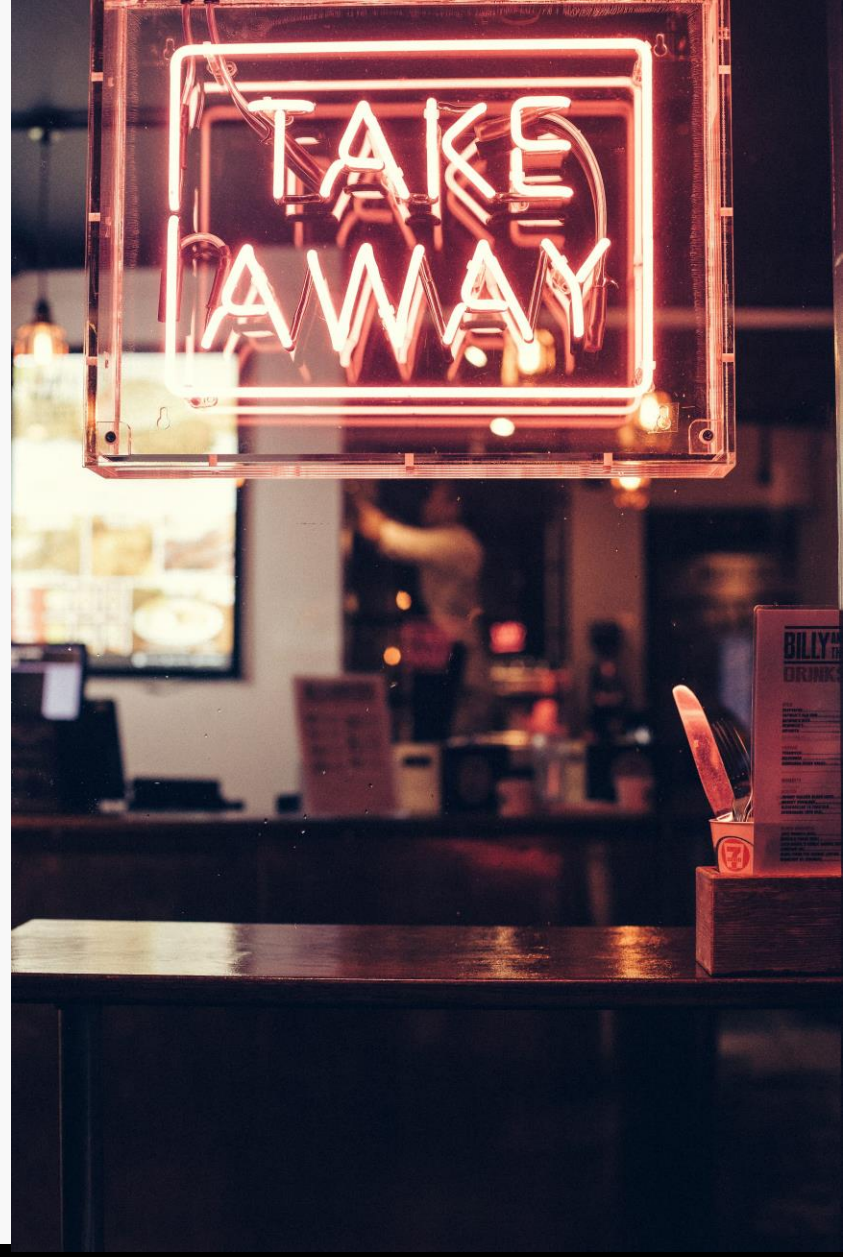
Throughout the project, consult relevant ethics guidelines—institutional, professional organisations, funding bodies.

Image from Unsplash

Takeaways

- Data management challenges can arise throughout the project life cycle
- Frameworks for open science, such as FAIR data practices, can conflict with GDPR and ethics compliance
- Legal and ethical challenges are often conflated, but addressing them requires different solutions
- Knowing your data is essential to working with it responsibly

Image from Unsplash



Questions?


1729
1870
·I·WAS·A·STRANGER·
·AND·YE·TOOK·ME·IN·
PATET
OMNIBUS
·I·WAS·SICK·AND·
·YE·VISITED·ME·