# Open Data in Practice

Zachary Batist

McGill University
Epidemiology, Biostatistics and Occupational Health

October 28, 2025

How do researchers *actually* engage with open science practices, principles and tooling?

vs.

How advocates *claim* science works, and data's role in their vision of scientific knowledge production

Speaker notes

So today I’m gonna talk about open science, and more specifically about open data.

In particular, I’ll be addressing some assumptions that many advocates of open science make about science and data’s role in the production of scientific knowledge.

And I’ll highlight how digital curation can contribute to resolving some challenges resulting from these false assumptions.

**Open Science:** a movement to make scientific research more accessible to scientists and society in general.

---

**Open Data:** the application of open science principles to data

Speaker notes

Before going forward, it may be helpful to define what I'm referring to.

So for clarity, open science is a movement to make scientific research more accessible to scientists and society in general.

It's supported by administrative policies and technical workflows meant to encourage greater access to research outcomes and greater transparency concerning research practices.

Open data is the application of open science principles to data, and largely involves establishing systems to document, host, preserve and disseminate research data.

- **About me**
- What are data?
- How open science tends to imagine data
- Examples
- Take-Aways
- Questions?

# Primary Interests

- Collaborative research practices
- How digital tools and infrastructures reconfigure knowledge production
- Specifically:
  - Infrastructures that support data sharing
  - Open source communities of practice
  - The cultural and epistemic implications of research infrastructures and policy

- How does open science (esp. open data) reconfigure — *or attempt to reconfigure* — collaborative experiences?
- What values are associated with open data, and how do they intersect with scientists actual needs, desires and values?

Currently:

- Technical and administrative systems that scaffold the **Covid Immunity Task Force Databank**

Previously (and ongoing, to a lesser extent):

- Data management within archaeological projects
- Archaeological data sharing, integration and reuse

Also:

- Collaborative software development practices
- Small-scale and community-driven data sharing initiatives

- About me
- **What are data?**
- How open science tends to imagine data
- Examples
- Take-Aways
- Questions?

Speaker notes

- So when I talk about data, we should be clear about what we mean.
- This is true as researchers, as curators, and as scholars of scientific practice.
- So before I go further, I'll briefly discuss the notion of data and disentangle its multiple meanings.

- It turns out, that when researchers say the word "data", they mean a few different things.
- So I've identified three notions of data that are often conveyed:

1. Descriptive accounts of observed phenomenon
2.
3.

---

Cats in my neighbourhood:

| **Name** | **Colour** | **Sex** | **Feral** |
|---|---|---|---|
| Ellie | Mix | F | Y |
| Bob | Brown | M | N |
| Charlie | Brown | M | N |
| Jasper | Black | M | N |
| ???? | Tortie | ? | Y |
| Spencer | White | M | N |

Speaker notes

- The first is data as descriptive accounts of observed phenomenon;
- So, consistent and precise records about objects or phenomena, used to document the state of a thing through a series of relevant properties.
- For example: a table describing cats in the neighbourhood.

- The data is going to follow a data model that delimits what's relevant, and it will always paint an incomplete picture.
- For instance, this does not capture that Spencer always wears a bowtie collar, or that the unnamed feral tortie loves playing in the autumn leaves.

1. Descriptive accounts of observed phenomenon
2. Evidence that forms the basis of a claim
3.

---

Data were collected to document Ellie's dietary preferences, by comparing the quantity of food consumed of various flavours.

Based on analysis of this data, we found that Ellie has a clear preference for chicken pate relative to all other options presented to her.

Speaker notes

- The second notion of data presents them as evidence that forms the basis for a claim;
- For example: Based on analysis of data pertaining to Ellie’s eating habits, we found that she prefers chicken pate over beef.

1. Descriptive accounts of observed phenomenon
2. Evidence that forms the basis of a claim
3. Means of communicating observations from one set of circumstances to another

---

Dearest Colleagues,

I’m enclosing the data from my neighbourhood observations.

My observations were done between 2PM-6PM, during my Sunday strolls.

It includes cats I saw in ground-floor apartment windows, and cats I encountered on the street.

Note that I saw no outdoor cats on rainy days, and I skipped my walk last week.

Feel free to make use of this very important information as you see fit.

Speaker notes

- And finally, the third notion of data presents them as means of communicating observations from one set of circumstances to another.
- It tends to acknowledge the material circumstances under which data were collected, and therefore presents the data as outcomes of decisions and actions, rather than reflections of stable knowledge.
- This can be imagined as someone documenting their data collection protocols, the rationales behind those plans, and also the ways in which their actions deviated from their planned courses of action.

# What are data?

1. Descriptive accounts of observed phenomenon
2. Evidence that forms the basis of a claim
3. Means of communicating observations from one set of circumstances to another

Speaker notes

- To be clear, these are not mutually exclusive definitions.
- Data is and can be all of these things at the same time, and people even refer all these aspects of data in the same breath.
- By pulling these apart, I’m trying to articulate what people value about data in different contexts.

# What are data?

1. Descriptive accounts of observed phenomenon
2. Evidence that forms the basis of a claim
3. Means of communicating observations from one set of circumstances to another

---

Implies:

- Data are fixed and stable
- Data are generated prior to and separate from analysis
- Aura of authority, stability, and truthfulness.

Speaker notes

- For example:
- These first two notions of data are really suited for each other.
- They each present data as fixed and stable observations, which are generated prior to and separate from analysis.
- Additionally, they imbue data with an aura of authority, stability, and truthfulness.
- This is extremely useful in a computational analytic environment, where a program requires users to ingest a clean and rectangular spreadsheet to perform some form of statistical analysis.
- The assumption that data are stable is necessary for the program to function as intended, and the sense that data are truthful is necessary to justify the relevance of the findings (garbage in, garbage out!).

# What are data?

1. Descriptive accounts of observed phenomenon
2. Evidence that forms the basis of a claim
3. Means of communicating observations from one set of circumstances to another

---

Implies:

- Data are created under specific circumstances
- Data are created for targeted purposes
- Data are informed by partial background knowledge

Speaker notes

- At the same time, conceiving of data as means of communicating observations between specific circumstances adds a layer of subjectivity.
- It acknowledges that data are not universal, that they’re created for targeted purposes, and that they’re limited by partial knowledge and material factors.

- This is especially important in a curatorial context, since it presents data as a product of human creativity.
- And like all things made my people, they carry traces of the social and material contexts of their creation.

- As digital curators, a significant part of our job is ensuring that the information contained in these records can be understood by those who seek to read them.
- And so our role is that of a mediator, connecting the perspectives of data creators and designated communities who will be accessing the data using the knowledge that each party already has.

Speaker notes

- In other words, those who spend a lot of time curating research data must continually reconcile the past decisions and actions that contributed to a dataset's state while simultaneously anticipating and supporting potential use cases by designated user communities.
- This tension is fascinating to me, and it's what drives my critical research on the assumptions and values that undergird infrastructures and policies pertaining to data sharing, integration and reuse.

- About me
- What are data?
- **How open science tends to imagine data**
- Examples
- Take-Aways
- Questions?

Speaker notes

- So now I’m going to talk a bit about open science, open data, and the infrastructures that support it.

# The Open Data Imaginary

- Non-material and non-political
- Reflections of natural reality
- Infinitely re-configurable

Speaker notes

- Specifically, I’m gonna address how the open data movement imagines data and data’s role in science — and how these figurations miss the mark.
- And while what follows may come across as more vibe-y and un-systematic, there is some truth to what I’m saying, and perhaps we can discuss these things in greater depth afterward.

# The Open Data Imaginary

- Non-material and non-political
- Reflections of natural reality
- Infinitely re-configurable

---

## Implies:

- Spreadsheets, databases and articles are considered value-neutral representations
- Belief that anyone can create and access data
  - BUT: Requires scientific resources, internet access, computational resources, requisite expertise to make sense of data

Speaker notes

- So first off, there is a strong tendency in the open science imaginary to conceive of data as non-material.
- That is, the accumulation and assembly of data is thought to contribute to a species-level understanding of the world, which is not held by any one individual but is stored as a form of common knowledge.
- While there is some acknowledgement that information is stored in physical media like books, scientific reports and internet-connected archives, these documents are often perceived of as value-neutral and purely representational in nature.
- However, not everyone can create data, and standards for validating the legitimacy of data are rooted in material constraints.
- For instance, if I work for a small or under-funded university, I may be unable to afford equipment that enables me to generate data at the same scale or calibre as my peers at Harvard or Oxford.
- Moreover, you need internet access, computational resources and knowledge of how to make sense of data in order for data to be useful, and if you’re not thinking about non-English speakers or people without internet access, you’re not actually making things universally accessible.

# The Open Data Imaginary

- Non-material and non-political
- Reflections of natural reality
- Infinitely re-configurable

---

Implies:

- Belief that data are reflections of reality, rather than *outcomes of decisions and actions*
- Data's legitimacy considered to derive from their objectivity

Speaker notes

- Next is an assumption that data are truthful reflections of natural reality.
- We already touched on this briefly, but it's no coincidence (in my opinion) that most of the people who are really into open data are data scientists and researchers who primarily do secondary data analysis (analysis of published data).
- People who work in those positions tend to be much more comfortable abstracting away the circumstances under which data were created.
    - In fact, they need to obscure those things in order to render their work legitimate.
- However, data are in fact products of human decisions and actions, and stripping away any sense of authorship provides a false sense that data are accountable only to nature.

# The Open Data Imaginary

- Non-material and non-political
- Reflections of natural reality
- Infinitely re-configurable

---

Implies:

- Diverse data can click together to form new knowledge
- More data $\Longrightarrow$ more possible configurations
- Ignores the extreme challenges involved in making heterogeneous datasets compatible

Speaker notes

- And third, the technocratic system imagined by many open science advocates envisions a global web of information, whereby new forms of knowledge emerge through novel integrations.
- It’s often imagined that if bits of data click together and are consistent with a pre-existing understanding of the world (which already clicks together), the new knowledge is deemed legitimate.
- However, as anyone who has actually tried to wrangle dozens of disparate datasets will tell you, this actually takes a lot of work and produces extremely unsatisfying outcomes!

**Science is intrinsically material and positional.**

**Open science fails when it does not account for these things.**

Speaker notes

- So open science is predicated upon some very significant false assumptions.
- And the mismatch between these assumptions and reality leads to extremely underwhelming outcomes and failures to meet initial expectations.

- About me
- What are data?
- How open science tends to imagine data
- **Examples**
- Take-Aways
- Questions?

Speaker notes

- So now, if there’s still time, I’ll go through one or two examples to showcase some of these mis-alignments in action.

# Database of Obsidian Sourcing Studies (DObsiSS)

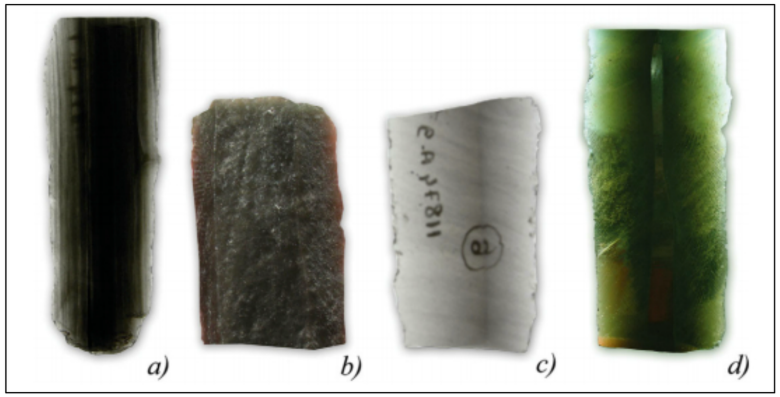
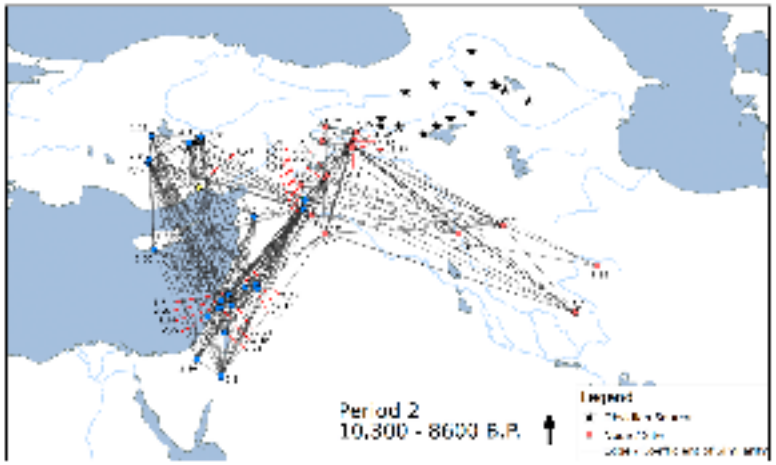
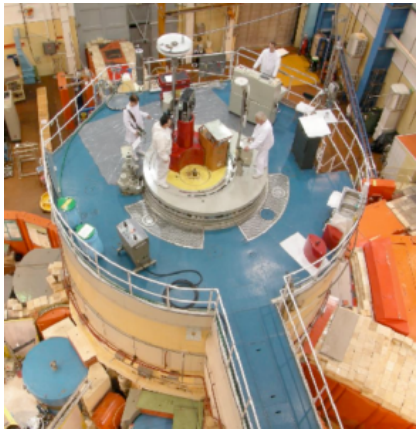


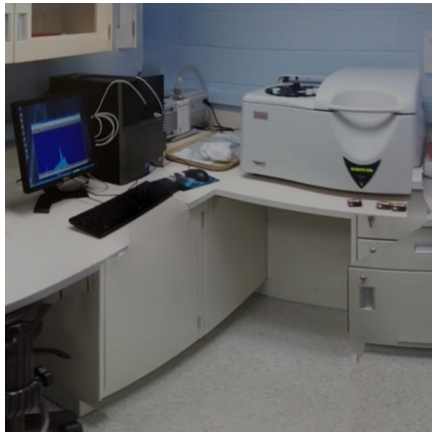

**Figure 3.1 – Obsidian blades made from material derived from various sources: a-b) Nenezi Dağ; c) Göllü Dağ East; and d) Bingöl A / Nemrut Dağ (from Carter et al. 2008).**



Speaker notes

[DObsiSS story]





| *Labora-toire* | *Acıgöl-ouest* | *Acıgöl-est* | *Nenezi Dağı* | *Göllü Dağı-ouest* | *Göllü Dağı-est* | *Bingöl calco-alcalin* | *Bingöl per-alcalin* | *Nemrut Dağı* | *Zarnaki, Meydan Dağı* | *Source incon-nue* | *Références* |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Cambridge | 4f | 1e-f | 1e-f | 1e-f | 2b | 1g | 4c | 4c | 3a | 3c | Renfrew *et al.* 1966, 1968, 1976 |
| Michigan | | 1e-f loc1-3 | 1e-f | 1e-f | 2b loc.6 | 1g | 4c | 4c Nemrut B | 3a | | Wright 1969 |
| Bradford | | B5 | | | B1 | B2 | G2 | G1 | B4 | | McDaniels *et al.* 1980 |
| Jerusalem | KRUD | HTMS A/B/C | NNZD | | GLD A/B/C | | | NMRD1 NMRD2 | ZNKT | | Yellin & Perlman 1981 |
| NIST | Koru | Hotmis I-III | | | Göllü | groupe D | | Nemrut I-IV | groupe E | | Blackman 1984 |
| Strasbourg | Acıgöl | | | | Kömür cü | Bingöl B | Bingöl A | | | | Cauvin *et al.* 1986, 1991 |
| Orléans | | gr.7 | gr.5 | | gr.3 | gr.2, gr.4 | gr.1b | gr.1a | | gr.6 | Gratuze *et al.* 1993 |

*Fig. 2. Équivalences entre les attributions données par les différents laboratoires (voir Chataigner, ce vol.).*

Different geological sources identified by different labs (as of 1998, ~27 years ago)

## Tell Aray 1

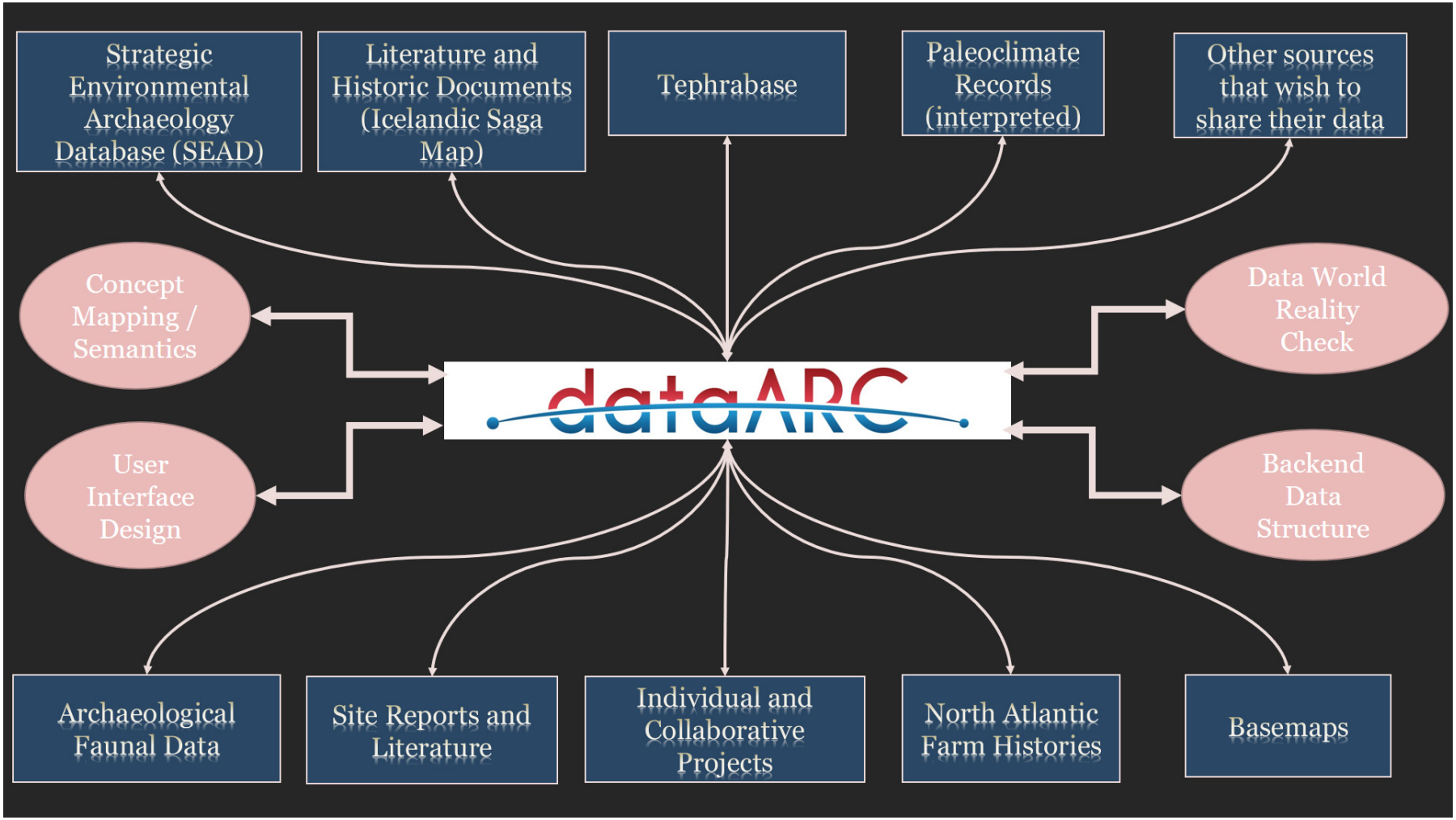
A total of 96 obsidian artifacts have been found during two seasons of excavations at Tell Aray 1. The collection consists of 42 blades, 48 flakes, four cores, and two retouched tools. All of analysed samples were recovered from the El-Rouj 3 (roughly contemporary with the Halaf) context except for one from El-Rouj 2c. A relatively high proportion of flakes and the presence of bladelet cores imply some bladelet production existed, while blade production was not likely to have been carried out consistently on the site because all flakes, some of which have natural surfaces, are small and thin. The typological features of the blades and bladelets do not show noticeable differences from those of Tell el-Kerkh 2 and Tell Aray 2. However, the relative frequencies of their color, ranging from greenish, blackish, transparent grayish and, to a lesser extent, brownish in transmitted light, show a remarkable difference from those two sites.

This difference of color variations was confirmed by the chemical analyses, by which a total of 26 pieces were analyzed. Remarkably, the results show that the relative frequencies of each source at Tell Aray 1 are completely different from those of Tell el-Kerkh 2 and Tell Aray 2. All but two pieces, which were identified as KMR obsidian, were classified into either Group A, C or D (Table 39). This sudden change of the variation and relative frequencies of obsidian sources between Tell Aray 2 and Tell Aray 1 is highly significant, suggesting that Tell Aray 1 had a closer relationship to the eastern regions than Tell el-Kerkh 2 and Tell Aray 2. Although it is not certain whether the distribution from Cappadocia was completely abandoned or not because of the rather small quantities of analyzed samples, there must have been a change in the obsidian distribution pattern. After this break between Tell Aray 2 and Tell Aray 1, no remarkable change of the source variation can be seen in the sequence of Tell Aray 1 itself except for the first appearance of Groups C2 and D obsidian in layer 10 (Table 40).

As for typology, blades and bladelets consist of obsidian from various sources: Group A1, A3, C1, C2, D and KMR. On the other hand, flakes and cores consist of only Group C1 and C2 (Table 41). Two bladelet cores analyzed by ED-XRF were both made of Group C1, that is Bingöl calco-alcalin, obsidian.

Data presented as prose, and mixed with analysis of non-chemical characteristics (2003)

# dataARC




Graphic overview of data sources embedded in dataARC (https://data-arc.org)

Speaker notes

[dataARC story]

- About me
- What are data?
- How open science tends to imagine data
- Examples
- **Take-Aways**
- Questions?

# Take-Aways

- Open data as currently practiced encourages taking data at face-value
- Most researchers intuitively understand data to be more complex than that
- More valuable insights may be obtained by
  - Smaller-scale initiatives
  - Directed by specific objectives
  - With specific designated communities in mind

- About me
- What are data?
- How open science tends to imagine data
- Examples
- Take-Aways
- **Questions?**