Curating archaeological research data in practice, from the field to the archive

Zack Batist

November 4, 2024

- Hi! I'm Zack!
- I appreciate you all being here at this early hour.
- I remember being in a very similar Digital Curation class at University of Toronto's iSchool, and it was one of my favourites.
- Ilja invited me to speak about my dissertation work, which involves some Digital Curation-adjacent components, specifically relating to the curation of research data
- Even more specifically, I look at the practical application of data work, and how the infrastructures that scaffold data-sharing mesh with established practices.
- Ilja mentioned that you already had someone speak about open data-sharing in scientific research contexts – and while there is a lot to love about open data, there are some significant challenges that many advocates just aren't thinking about.
- So what I'm going to present is a bit critical of some of the assumptions that the open data movement takes for granted, with a goal of improving the infrastructures that scaffold open data, rather than tearing it all down.

1 About me

1.1 Archaeologist

- But first, maybe I can share a bit more about me.
- Much of my work is based in archaeology.
- I've excavated in Greece and Cyprus, but my main focus was applying network analysis and other computational methods using published data.

- Specifically, I integrated obsidian sourcing data from dozens of published sources to trace large-scale regional interaction in the prehistoric near east.
- I also managed archaeological databases in rough fieldwork environments.
- This involved establishing workflows and protocols for data collection, processing and analysis.
- I also maintain open-archaeo, a list of open source archaeological software and resources, which has grown into a valuable community resource and dataset.
- And on top of all this, I help administer and moderate the archaeo-social mastodon instance and social media collective.

1.2 Scholar of Scientific Practice

- These experiences contributed to my role as a scholar of scientific practice.
- My specific focus is on scientific data management places, and framing data-sharing as collaborative experience.
- I'm especially interested in articulating:
  - how open science reconfigures (or attempts to reconfigure) collaborative experiences;
  - the bureaucratic apparatus of open science;
  - the emergence of a new "informational" perspective among researchers, who are now made to think about their work from a curatorial lens; and
  - resistances against the power relations that emerge from these aforementioned developments.
- I think archaeology is a really great lens to examine these things because of how explicitly data appears to be constructed and transformed.
- Archaeology has a long history of introspection, and the centrality of recording systems and interfaces really drives awareness of the tension between situated experiences and objective representations.

2 Meta-Science Work

- I do this through various vectors of empirical research.
- For instance, in a recent paper, I supplemented open-archaeo with data down from the GitHub API to examine how the open source model of software development actually impacts collaborative coding experiences.

- My colleague and I found that archaeologists still code rather independently, or in groups that correspond with existing working groups.
- As such, we found that open source is not really changing how archaeologists collaborate on software development, and that research software engineering in fact firmly embedded in established academic power structures.
- I'm also currently doing a bibliometric analysis archaeological data papers, specifically trying to elucidate different kinds of ways in which data are cited in situations of re-use, but this is still very much a work in progress.

3 Key questions and concerns

- This builds on my dissertation work, which examined how archaeologists contribute to communal data streams in local and relatively private research environments.
- I wanted to show how data work is embedded throughout the archaeological process, and trace the technical and social structures that scaffold the flow of information, which has implications for how data are shared more broadly, beyond the scope of the projects from which they originated.
- I take to heart my supervisor's ideas on curation, which he conceives as occurring throughout the research process, in all decisions and activities that produce or make use of data.
- In other words, scientific activities, which take in and push out information, effectively re-interpret meanings produced prior to their application while simultaneously re-presenting new meanings in anticipation of future use cases.
- So while data has descriptive and evidentiary value, in that they serve as stable representations of an object of phenomenon that can be used to infer deeper meaning about them, I focused on data's capacity to translate meaning across time, space and social circumstance.

4 Notions of Data

- We commonly think about data as formally-structured, consistent, discrete, corpuscular, and scalable.
- They are thought to be objective representations of reality, exhibiting an aura of authority and truthfulness, and as asocial, a-contextual and lacking material basis.
- And they are thought to be akin to building blocks that can be easily recombined to produce synthetic forms of knowledge.

- However, based on my experience working with archaeological data, I recognize that these are rather aspirational or idealistic notions, and do not really correspond with how people actually work with data.
- If we think about data as the products of human intervention, then – as with all things of anthropogenic origin, we must think of data as being informed by values and material circumstances.
- This has been recognized previously in a few studies published over the past decade, where archaeologists were surveyed about the challenges they experienced while trying to reuse data.
- They consistently responded by stating that they needed more context about the circumstances of the datasets' creation, including information about the character of the people who created the data.
- In other words, they effectively sought out information that could only really be obtained through a deeper collaborative relationship, or by being more closely involved in the datasets' creation.
- Moreover, when archaeologists download spreadsheets from a web server, they must also do a lot of additional interpretive work to make sense of the data they contain, and this involves parsing context and subtext not contained in the data itself.
- All of this got me thinking about data-sharing at a distance as an extension of existing collaborative ties, or as **one form** of collaboration that has been devised in the context of the global span of scholarship and the modern telecommunication technologies that connect disparate research endeavours.
- So, in thinking about data as records that enable people to translate meaning from one research context to another, we are now dealing with what it means to establish mutual understanding.
- And this in turn means that it is necessary to examine science as a humanistic enterprise, driven by coordinated action and scaffolded by both social and technological systems.

5 Methods

- In order to do this, I rely on a variety of data sources and analytical methods.
- I already mentioned some of my more recent work analyzing data pulled from APIs maintained my coding and publishing platforms.
- But my dissertation work, which is what I'll be focusing on today, was much more in-depth.

- That essentially involved observing and interviewing archaeologists while they worked, and examining the documents they produced.
- I put cameras on archaeologists' heads and in the corners of the trench or lab environment while they excavated or sorted through artefacts, and interviewed them about their work practices.
- I also conducted numerous retrospective interviews to ascertain bigger picture values and priorities.
- I was therefore able to compare what archaeologists said they were doing with how they were actually behaving, and trace connections between past and future activities recorded across different times and places.
- I did this at three cases from 2016-2019; one case was longitudinal over 3-4 years, and one was explicitly focused on a data archive.

6 Data Collection

- I wanted to look at how data are collected and transformed into more stable and transmissible media, so I'll talk about these aspects, in sequence.

6.1 Recording Sheets

- The most acute and visible mode of information work within archaeological projects consists of acts of recording: filling in recording sheets and writing notes in a field journal.
- So one priority was to articulate recording practices and how these practices were situated within the broader apparatus of archaeological knowledge production.
- Specifically, I found it valuable to compare the use of recording sheets and field journals, which afford different kinds of behaviours and communicative outcomes.
- In my analysis of the context recorded sheet used at one of my cases, I identified five main sections.

- First, a section is dedicated to storing indexical information that identifies the locus, or excavation unit, that the sheet pertains to.
- This takes the form of a unique identifiers for the excavation unit, trench, survey unit, or feature, as well as the dates when these things were being uncovered.

- This section also identifies the people who were responsible for identifying, articulating, and recording the locus and its attributes.

- Second, throughout the document, and clustered into subsections corresponding with different kinds of materials, the recording sheet prompts users to provide structured information according to a controlled vocabulary, as documented in the fieldwork manual or on the recording sheets themselves.
- Fields prompt users to record the things they found, the depth of the trench in various locations, the properties of the soil, and the equipment they used for excavation.
- The third section prompts users to describe the excavation unit in their own words.
- This allows them to highlight relationships among entities within the locus and among loci.
- The fourth section prompts users to relate the record with other media pertaining to the same archaeological entity (such as photographs or illustrations) and documents corresponding to other related archaeological features.
- Finally, the fifth section contains a blank grid and relationship chart where users could draw and identify the locus, surrounding loci, and any significant aspects in a visual or schematic form.
- The fieldworkers I spoke with perceived recording sheets as formal documents that are meant to contain official or authoritative accounts of each locus and its material properties.
- Consequently, new fieldworkers often felt a need to ask questions about how they should fill out these forms to meet the expectations of the project.
- In some ways, filling the recording sheets seemed to represent a somewhat bureaucratic obligation.
- While recording sheets were considered official records, they were sometimes also viewed as cumbersome obstacles that distract from ongoing work or that fail to capture what was really occurring in the trench.
- For instance, Theo, a seasoned fieldworker, was somewhat dismissive of the recording sheets and resentful of the demands that they impose.
- He believed that context sheets force him to write his observations in unnatural ways, forcing naturally fuzzy information into strict and arbitrary forms.

- For Theo, recording sheets are tools that warp reality into *abstractions of reality*.

6.2 Field Journals

- Archaeologists often compared recording sheets with field journals, which are the other primary way archaeologists record their observations in fieldwork settings.
- According to Theo, field journals record a "stream of consciousness" and provide a more genuine account of what occurred in the field.
- They enable a reader "to understand what the excavator was thinking ... whilst they were excavating".
- In other words, they serve as mnemonic devices that preserve memories of the reasoning behind decisions that excavators made, but which they may forget during the flurry of activities that they must perform or that may fade from institutional or collective memory as fieldworkers move on to other projects or otherwise become inaccessible.
- While Theo claimed that "in the journal you can just write the fuck you want" there are professional expectations that guide what information supervisors should record in field journals and how they should structure that information.
- As with recording sheets, the field journals I examined comprised a few distinct sections.
- First, they contain indexical information that identifies the endeavour (i.e. the trench or specialty) to which the journal pertains, as well as information about who was responsible for leading or carrying out the work.
- This typically occurs on the cover or first page.
- Then, the journal entries themselves follow.
- These are typically recorded on a day-by-day basis rather than ordered by unit or locus.
- Each entry may contain its own indexical information, such as the date, a list of people involved in the work, unique identifiers of contexts being worked on, etc.
- Entries also typically mention the conditions or circumstances under which work is occurring, such as the weather, remarks about the crew's general attitude and morale, or any disruptions that may have occurred that day.
- They may also list the goals set out for each day of work, relating entries to each other and leading to the formation of quasi-narratives about work progress.

- The main content of journal entries consists of a log of decisions that the supervisor made and instructions to and carried out by assistants.
- They also include fleeting interpretations of phenomena being uncovered, revealing why and how certain decisions were made during the work process.
- Journal entries also commonly use colloquial language and refer to entities they recover in a very casual way.
- For instance, the journal entry depicted in this image refers to areas of the trench as the "sand pit of doom" and "bouldery hell".
- The field journals I examined are crafty, multi-media documents.
- They often contain sketches or schematic visualizations of the trench, of the landscape, of relevant features, or of mental models scattered throughout the notebook.
- Sketches are without scale, and entities are labelled only when the illustrator deems it necessary at the time of drawing.
- They also sometimes contain hand-drawn tables recording regularly formatted data, such as running lists of photographs taken, contexts opened, special finds and their spatial coordinates, or samples taken.
- Because these tables are typically recorded at the end of the notebook and are filled in as new pertinent info comes across their radar, they tend either to run out of space or to reserve too many extra pages.
- Sometimes, tabs are added to the edges of those pages using a piece of paper reinforced with scotch tape to make them easier to access. Notebooks sometimes have pages ripped out or have pages informally added with tape, glue, or a stapler.
- The journal entries switched between atomic and descriptive characterizations of specific elements within the trench and more speculative associations that draw the trench within a broader understanding of the site as a whole.
- They exhibit greater flexibility than more formal records in that they often refer to a variety of related entities or observations on the basis of the judgement and experience of the writer.
- In this way, field journals are discursive media that describe and discuss particular aspects of the project from the situated perspectives of their authors and contextualize and define an object's significance on the basis of particular experiences with it.

- Trench supervisors sometimes elaborated on these rough interpretations during site tours, which, at least at one of my cases, were regularly scheduled events whereby the whole team went around the site to learn what was going on in each trench.
- When the team arrived at a trench, its supervisor described its principal features, typically in a fashion that recalls the work and decisions involved in its exploration.
- Usually, the project director or analysts supplemented this account by making interjections or rebuttals, helping to situate the trench in relation to broader project-wide narratives.
- Site tours were informal and were never recorded, but they conveyed a great deal of information to listeners.
- Tours used imprecise language and referred to things whose meanings may not have been well understood outside the project team.
- For instance, members of Case A often referred to the "red shit," which signifies a layer of red clay that appears throughout the site and which nearly all excavators have had to struggle with.
- Project directors also liked to give these tours to visiting scholars, notable guests, and new project participants so that they could get a better understanding of what was going on in the site, rather than being limited to what was published in a paper or report.

6.3 Comparison

- This echoes other mentioned statements made by my informants regarding the value of personal and informal modes of communication when trying to relate the character of a site to those who are less familiar with it.
- In each case, there was a general consensus among fieldworkers that journals captured much more information than recording sheets, though of a different kind.
- This is especially interesting in light of the fact that the information contained in journals occupy are rarely transcribed as more formal records, and occupy an entirely separate data stream as that which gets processed into a project's relational database.

7 Digital Transformation

- As archaeological data are collected, it is necessary to render them in ways that are more amenable to systematic analysis.
- This is typically achieved by inputting and organizing data using digital systems such as relational databases, file systems, and digital archives.

7.1 Databases

- Databases served to centralize data, relate the outputs generated by complementary streams of investigation, and ensure that the data are structurally consistent.
- The databases used by the projects I examined were custom-built and used conventions specific to the project.
- Practical decisions about the database were often made "on the fly" or were derived through trial and error.
- Database managers often learned their skills on the job and assembled code that was previously published on various blogs, tutorials, and online forums.
- At the same time, database managers often struggled to reconcile the information presented by these disparate sources, and the products they eventually cobbled together did not always perform optimally.
- So the databases are in one sense representations of a project's priorities, and in another sense manifest the meandering journey of the individual who put it all together.

7.2 Messy Records

- I observed that formal data contained within databases, which are characterized by being clean and tidy and are arranged so that they are more conducive to complex retrieval queries and patterned analysis, often originated as relatively messy analog records that are more amenable to fieldwork conditions.
- Through data entry and data-cleaning processes, the values written down on paper recording sheets were copied to homologous and homogenous digital tables.
- However, fieldwork documentation was performed in ways that were responsive to that specific work environment and did not actively account for those transformations that would occur down the line.
- For instance, fieldworkers used imperfect spelling and grammar, used shorthand representations, deviated from controlled vocabularies, and crossed out and re-wrote text.
- According the database managers I spoke to, they were responsible for correcting somewhat trivial errors, like different spelling of words referring to the same thing.
- This aspect of their job involved transcribing written records into formats optimized for computer-assisted data retrieval.

- This sometimes involved significant editing and ommission of information contained on the handwritten records.
- Words or values that have been crossed out or revised were not copied over;
- different handwriting or penmanship, which implies different authors or circumstances under which the records were made, were disregarded;
- and drafted versions of recording sheets, which were entirely re-written, sometimes never made their way to the database manager at all.
- Additionally, some elements that are difficult to represent as distinct database records, like sketches or mind-maps were excluded from the database altogether.
- Acts of transcription therefore involved a significant amount of transformation, including information loss.

7.3 Anxiety

- And this often produced a sense of anxiety, since this work failed to meet the initial expectation of a smooth and frictionless workflow that database managers seemed to expect going in.
- For instance, Jamie – who was one of the database managers for the primary case I've been discussing today – recognized that fieldwork should be modified to support analysis by producing cleaner and tidier records.
- She specifically advocated for implementing and enforcing standards on fieldwork activities to ensure that the data were more amenable to analytical purposes down the line.
- Similar, Paul, who maintained an archaeological data archive, considered it his job to help projects conceive if their research as data – by which he means as concrete and formally-modelled records.
- These statements imply a perceived disconnect between the rough and improvised experience of fieldwork and a conception of what constitutes "proper" research and "proper" data.

7.4 Digital Archives

- Projects sometimes hire archaeological data services to help maintain their data, with an eye toward curating, preserving, and publishing the data after the project is complete.
- By paying digital archives to curate their data, archaeological projects effectively delegate responsibility to sanitize, document, preserve, and distribute their data to dedicated experts who are committed to these tasks.

- Project leaders stated that depositing data in a digital archive also satisfied projects' commitments to funding agencies, who often mandate that funded projects plan for proper and long-term care of their research materials, which includes ensuring that all data are publicly accessible.
- Interestingly, while digital archives' role in data reuse is often touted as their primary function and benefit, the project directors I spoke with generally considered this a secondary concern.
- Altogether, my informants stated that digital curation services enable them to move forward with new projects without having to worry about the state of their prior work.
- So from one perspective, archives served as a kind of final resting place, and from another perspective they were places where old data could be given a new lease on life.
- But as I said earlier, this notion of archives as loci for simply uploading and downloading spreadsheets is a bit of a myth, and there will always be a need for discursive engagement to enable practical re-use.
- People who re-use data know, on an intuitive level, that there is more to the data than what is documented in the supplementary materials, and they make efforts to circumvent these systems to get at the information that will actually support their research.
- It's really ironic, then, that the formal and transactional protocols meant to streamline mutual comprehension of a dataset reveal their own inadequacy for achieving their stated purpose, while also revealing the strengths of the system that they are meant to replace, e.g. socially-mediated forms of collaboration.
- But archivists' focus on the technical processes of cleaning and documenting data shields them from grappling with this tension.

8 Take-Aways