Messy Understandings

Mia Ridge @mia_out

Speaking in Code, November 2013

University of Virginia Library Scholars' Lab

Image by Flickr member rosipaw under CC-BY-NC-SA

Fundamental tension between tools and cultural heritage data: trying to fit a square peg into a round hole. **Do you craft the tools to the data or the data to the tool?**

So what do you do with square pegs and round holes? You can chop off the interesting edges to fit something into a round hole, you can reduce the size of the entire peg so it'll slip through, or you can make a new bespoke hole that'll fit your peg. But then how do we make the choices we've made obvious to people who encounter the data we've squeezed through various holes? It's particularly important if people are using these collections in scholarly work to make the flattenings, exclusions that shape a dataset visible.

The choices you make will depend on your resources and skills, the audience for and the purpose of the final product… Will look at some examples of visualisations for exploring collections where I had to tidy the mess to make them work, and an example of designing software to cope with the messy reality it was trying to reflect.

I want to set the scene with my own experiences with cultural heritage data, but am curious to hear about your own experiences with messy data in your respective fields, and the solutions you've explored for dealing with it and conveying your decisions.

Image credits: http://www.flickr.com/photos/rosipaw/4643095630/ rosipaw

The 'about me' bit. I've spent over a decade working in cultural heritage - mostly at museums, but also with library and archive collections. Arts then software engineering then HCI/UX at uni. Much of the time I was working as an analyst and database or web programmer - mostly backend stuff; sometimes hooking directly into collections management systems and sometimes working on the user experience on the frontend. I also found myself working as a translator, explaining why the internal systems for registering collections, managing events, publishing content on the websites, make some things easy and other things hard. (Over time, I realised that working at the intersection of museums and technology required 'double domain expertise' and that it's a combination of skills worth recognising, and started talking about 'museum technologists' as a label. Others in UK academia have come up with other terms - strategic developer, research software engineers - to describe people who 'not only develop the software, they also understand the research that it makes possible', this event, etc More at http://openobjects.blogspot.com/search/label/museum%20technologists). NB: will often use 'GLAM' as a shortcut - galleries, libraries, museums, archives.

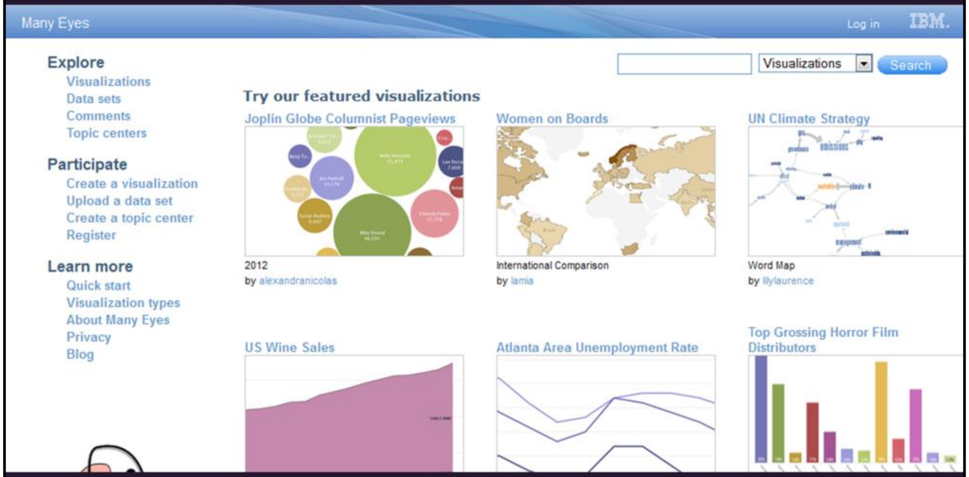
I'll start by trying to convey some of my own tactic knowledge about GLAM collections and my methods for dealing with their messiness. The infovis examples I'll look at are based on my work with Science Museum (London) data and work others did after we released the data and a short residency I did at the Cooper-Hewitt design museum in New York - shaping the data to the tools - and to think about shaping tools to the data I'll look at some on recording systems in archaeology.

#musetech

Museum background possibly also important means I've been part of a big, friendly, international community; they share ideas through blogs, social media (#musetech, #drinkingaboutmuseums), meetups. Another important aspect of the museum technologist experience is that the big international (North American-based) 'Museums and the Web' conference requires that all accepted presentations write a 5000 word paper, which is published on their website a few months beforehand. This means you can get more out of the conference, or follow it from a distance, but it also means that technologists have to learn to write formally about their work - often non-academic technologists first experience of writing a 5000 word paper, and as such a bit of a shock to the system but ultimately worthwhile.

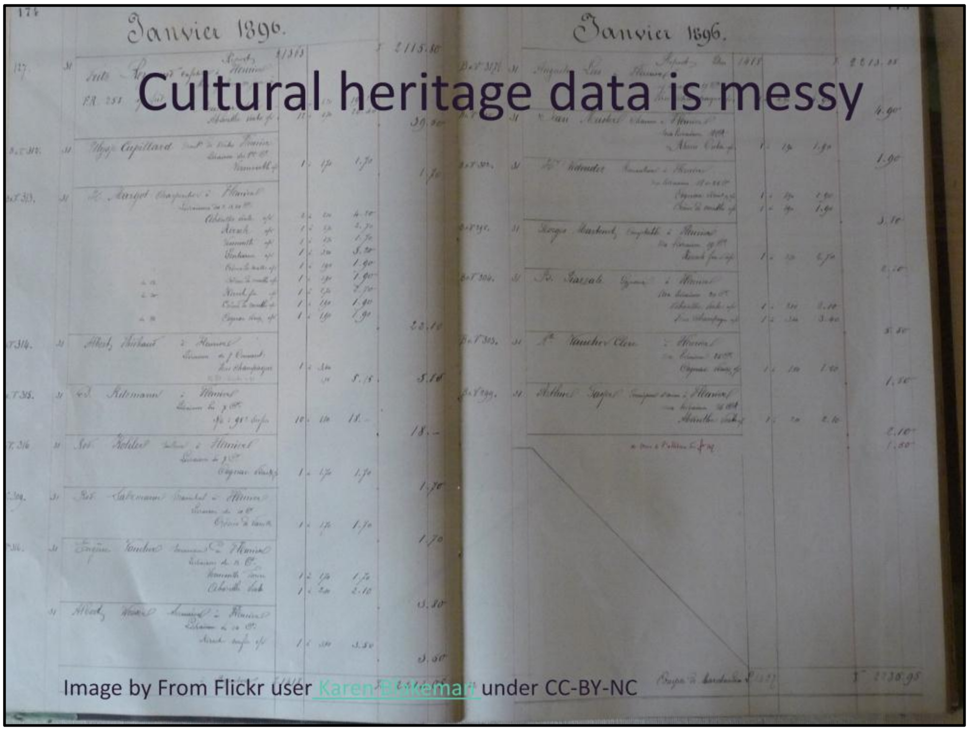
Museum background also means I'm interested in public engagement with history and culture and tend to assume that any research project has a public face. I've been interested in 'open cultural data', publishing museum, library, archive records to help researchers and make content more discoverable by other users. Got into data visualisation through museum work, provide ways in to content, make something when open cultural data provided. But run into big issues trying to use general visualisation tools on cultural heritage data, watched it happen to other people at hack days etc… It makes it hard to explore playfully; have to commit to data cleaning to suit tool or writing custom code with eg javascript libraries.

Visualising complex data



Another way of looking at the tension between crafting the tools to the data or the data to the tool. Two options for visualising complex data - **find a visualisation type that can harbour the data in a meaningful way or reduce the data in a meaningful way**. E.g. go from individual values to distribution of values; or introduce interaction: overview, zoom and filter, details on demand - useful to keep in mind.

NB: I tend to use a lot of generally-available tools for teaching as they're more accessible to people who want to keep learning after a workshop, and because they can help you work out where to invest more time in custom code. IBM ManyEyes, Google Fusion Tables rather than D3. People are often surprised to learn that they can make up new types of visualisation.



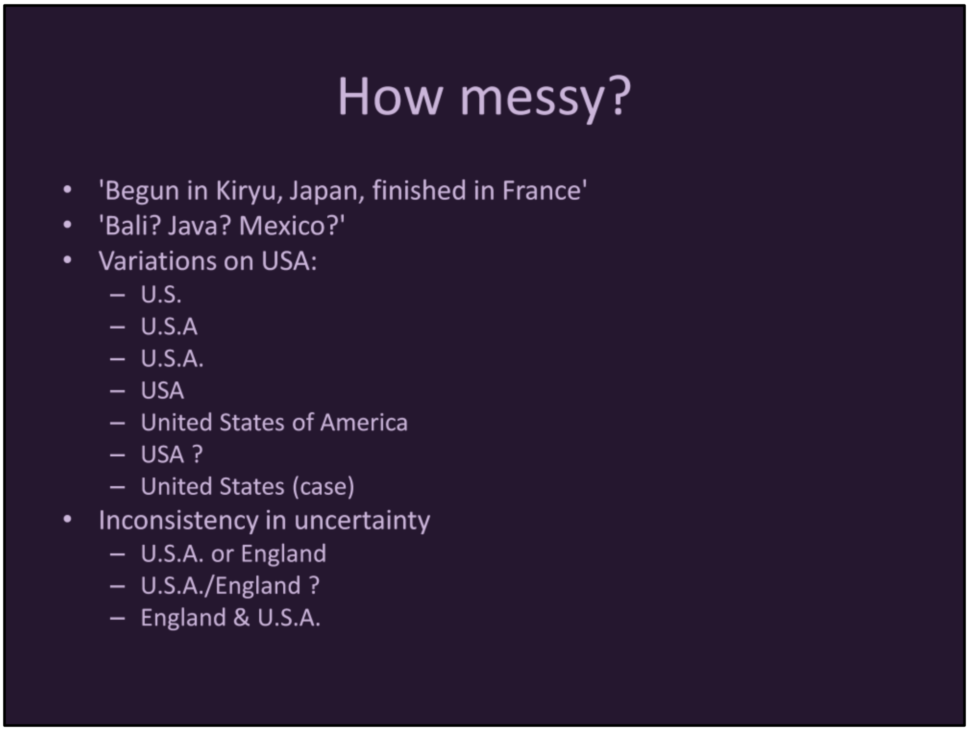
Cultural heritage data is messy

Image by From Flickr user Karen Blakeman under CC-BY-NC

Museum collections - whether art, history or specialist - are often accidents of history, the result of the personalities, trends and politics that shaped an institution over its history. Many older collections are based on private collections or include private bequests, so they're shaped by idiosyncratic impulses as well as the vagaries of collecting over time. Collections can be extremely lumpy and oddly shaped conglomerates.

Collections data can be even more so, because the randomness of the collecting process itself is multiplied by the variability of documentation practices and standards over the past decades or centuries. Very different to born-as-data scientific material or literature that's created to be read, the existence of cultural data is secondary to other processes eg managing loans, conservation, exhibition logistics, catalogues and interpretation. The decisions GLAM staff have made about links, relationships, media, etc also affect what's possible and what's easily discoverable in various systems. 'Data' is also variously to describe any combination of item- or collection-level metadata, transcribed texts, descriptions of objects, or images of objects - when I'm looking at visualisations I'm often looking at the metadata as it's easiest to get hold of, but it's not a very satisfactory form of 'distant reading'.

http://www.flickr.com/photos/rbainfo/6581426941/ Flickr User Karen Blakeman



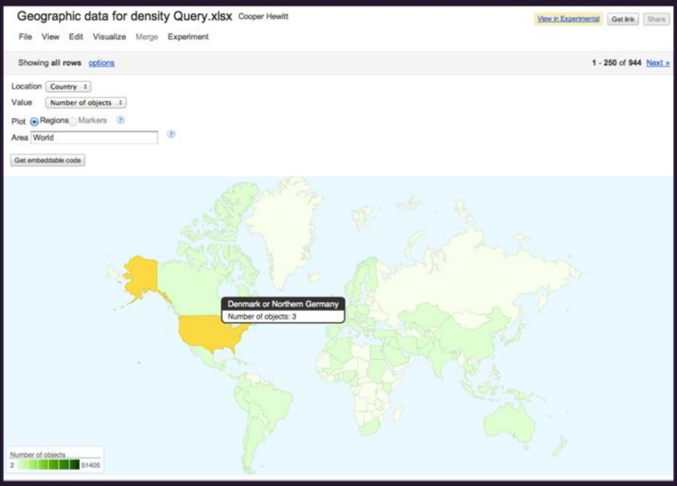
How messy?

- 'Begun in Kiryu, Japan, finished in France'
- 'Bali? Java? Mexico?'
- Variations on USA:
  - U.S.
  - U.S.A
  - U.S.A.
  - USA
  - United States of America
  - USA ?
  - United States (case)
- Inconsistency in uncertainty
  - U.S.A. or England
  - U.S.A./England ?
  - England & U.S.A.

More from the Cooper Hewitt collection. I spent 3/5 of my week at the Cooper Hewitt just trying to get the data clean enough to vaguely represent the collection. The first problem is that computers think U.S., U. S. , U.S.A., U. S. A. , United States, United States of America are six different strings, though a good geocoder might place them all at the same coordinates. But those aren't the worst issues - with enough resources, you could resolve those inconsistencies in the source database. It's harder when you're dealing with uncertainty - it might not be possible to resolve uncertain provenance even with research. You will often want to preserve necessarily complex values - 'place made' usually ends up being a qualified field - eg place designed, place prototyped, place manufactured - but you'd rarely design a database around exceptions like 'begun in Japan, finished in France'. More common museum issues - what year is 'early 18th century'? What do you do with '1836 (probably)' - introducing certainty fields can be tricky if that information has to be assumed for existing records. Date ranges (in turn derived from artistic periods, historic movements) are common, and can't easily be reduced to a more precise meaningful date in order to be rendered on a conventional timeline. Individual fields can contain gotchas - e.g. measurements in different or unstated units - as a result of changing practices over time.

When cleaning data like this, you're making constant decisions about what to support and what's an unsupported edge case.

Messy data fails



Geographic data for density Query.xlsx Cooper Hewitt

Showing all rows

1 - 250 of 944

Location: Country

Value: Number of objects

Plot ● Regions ○ Markers

Area: World

Denmark or Northern Germany

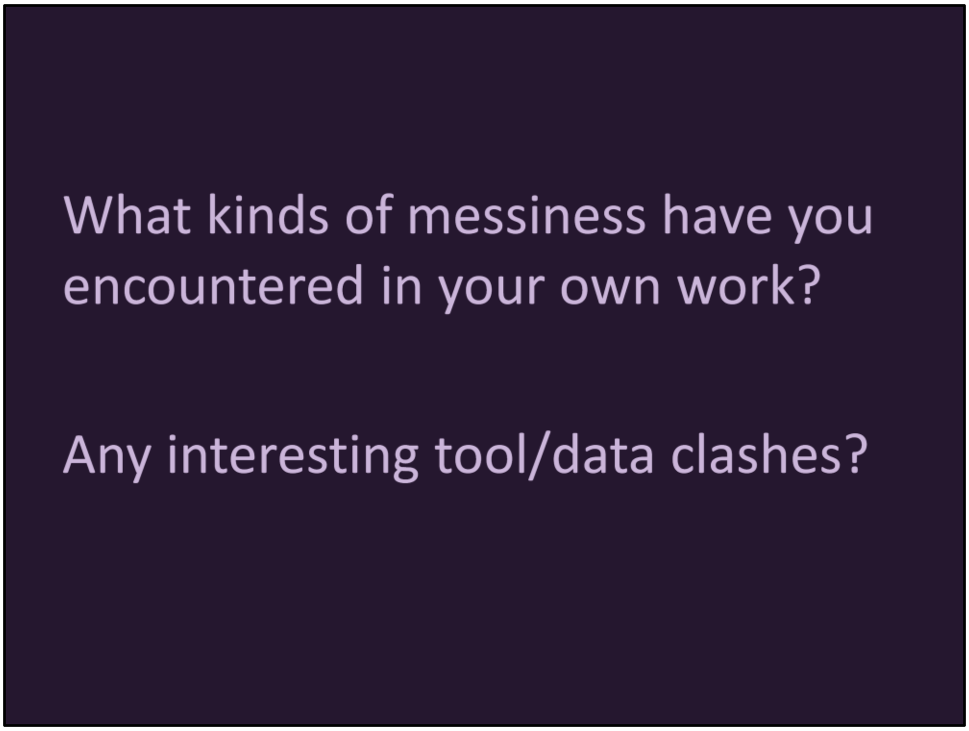
Number of objects: 3

Number of objects

2 51405

Most visualisation tools don't cope with messy or fuzzy data
This is what happens when tools encounter messy data when they expect something neat. Who knows what Google Fusion Tables thought was going on here, but it's effectively hidden the hundreds of thousands of records that should be shown in the USA.

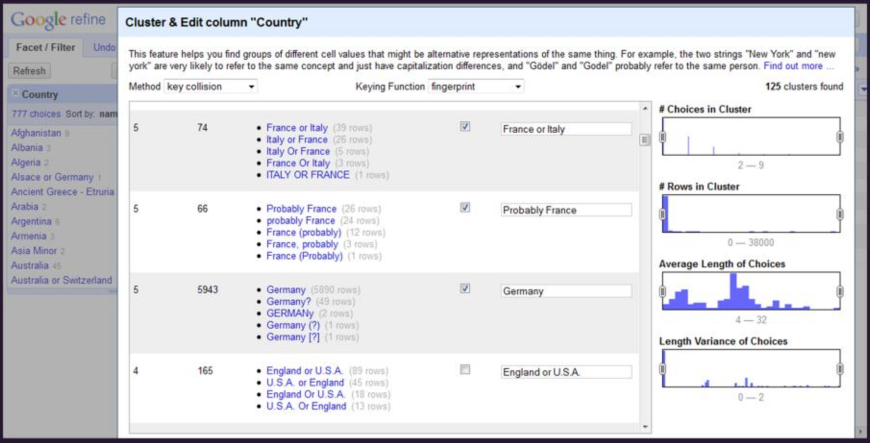
Commercial tools often assume complete, born-digital datasets – no missing fields, consistent data entry over time.



What kinds of messiness have you encountered in your own work?

Any interesting tool/data clashes?

Data cleaning is a PITA



Cluster & Edit column "Country"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

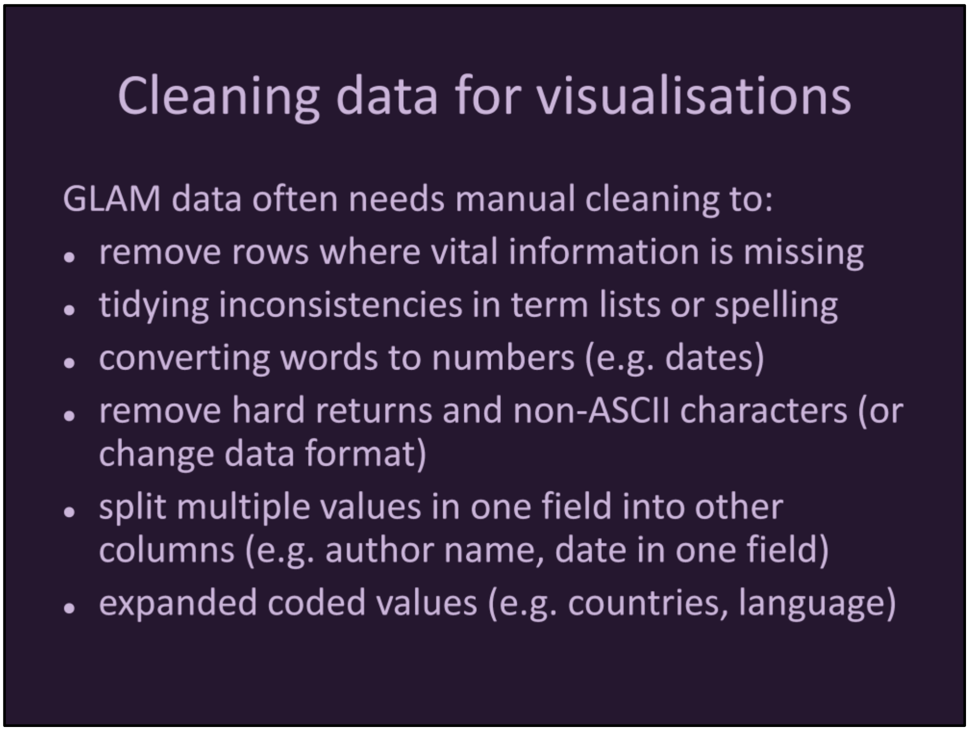
Method key collision   Keying Function fingerprint   125 clusters found

5	74	• France or Italy (39 rows) • Italy or France (26 rows) • Italy Or France (5 rows) • France Or Italy (3 rows) • ITALY OR FRANCE (1 rows)	☑	France or Italy
5	66	• Probably France (26 rows) • probably France (24 rows) • France (probably) (12 rows) • France, probably (3 rows) • France (Probably) (1 rows)	☑	Probably France
5	5943	• Germany (5890 rows) • Germany? (49 rows) • GERMANy (2 rows) • Germany (?) (1 rows) • Germany [?] (1 rows)	☑	Germany
4	165	• England or U.S.A. (89 rows) • U.S.A. or England (45 rows) • England Or U.S.A. (18 rows) • U.S.A. Or England (13 rows)	□	England or U.S.A.

I had a great big list of questions I wanted to explore in my week at the Cooper Hewitt. I thought I'd iterate through stages of cleaning the data, trying it in different visualisations, then going back to clean up more precisely as necessary. Ended up spending 3 of the 5 days wrestling with data cleaning, limitations of tools. Not big data per se, but big enough - difficult to load 270,000 rows in Excel. Overall I spent about a day of my time dealing with the sheer size of the dataset: it's tricky to load 60 meg worth of 270,000 rows into tools that are limited by the number of rows (Excel), rows/columns (Google Docs) or size of file (Google Refine, ManyEyes), and any search-and-replace cleaning takes a long time.

This screenshot shows examples of inconsistent data from Cooper-Hewitt being cleaned in Refine. If term lists have been used, the data won't be quite as messy as this, but often standards have emerged over time, and might have been different between different departments.

Fields also contained random extra things like internal notes about potential duplicates, unexpected extra information - notes on what type of location, etc.



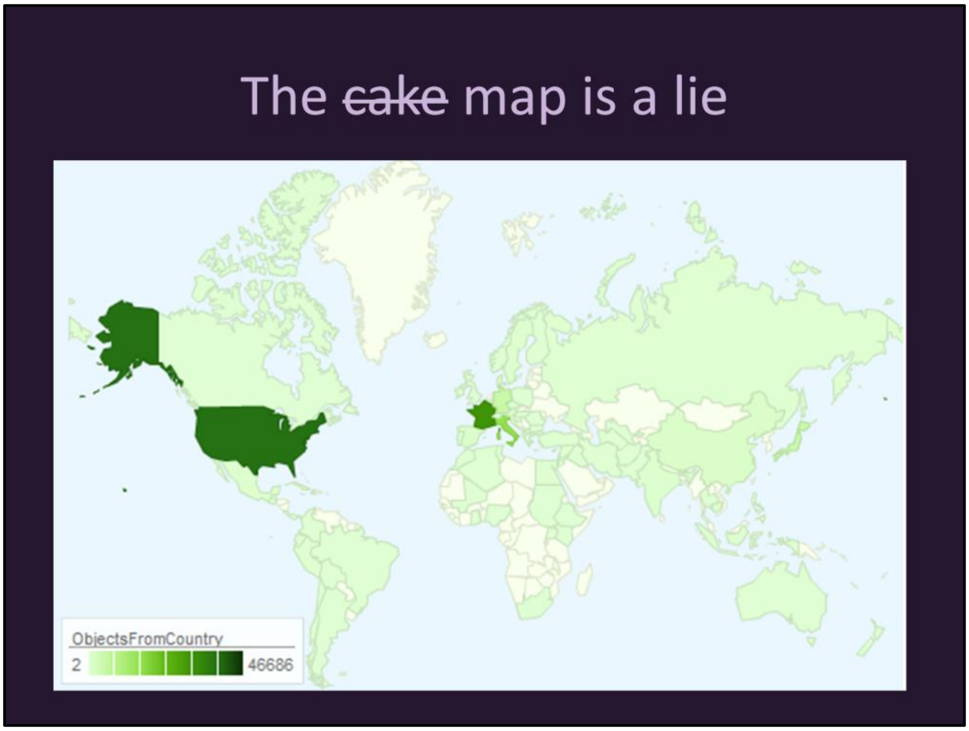
Cleaning data for visualisations

GLAM data often needs manual cleaning to:

- remove rows where vital information is missing
- tidying inconsistencies in term lists or spelling
- converting words to numbers (e.g. dates)
- remove hard returns and non-ASCII characters (or change data format)
- split multiple values in one field into other columns (e.g. author name, date in one field)
- expanded coded values (e.g. countries, language)

The necessity to leave out data that isn't clean enough is one reason visualisations should be taken with a pinch of salt... Documenting decisions and tracking versions of files as you go so you can explain the provenance and representativeness of your data is useful, but if you're trying to present a few simple interactive visualisations, you need to think carefully about how to explain what data's missing without losing people in detail.

Tools like Refine are great, but almost too powerful... One issue is that museums tend to use question marks to record when a value is uncertain, but Refine strips out all punctuation, so you have to be careful about preserving the distinction between certain and uncertain records (if that's what you want).



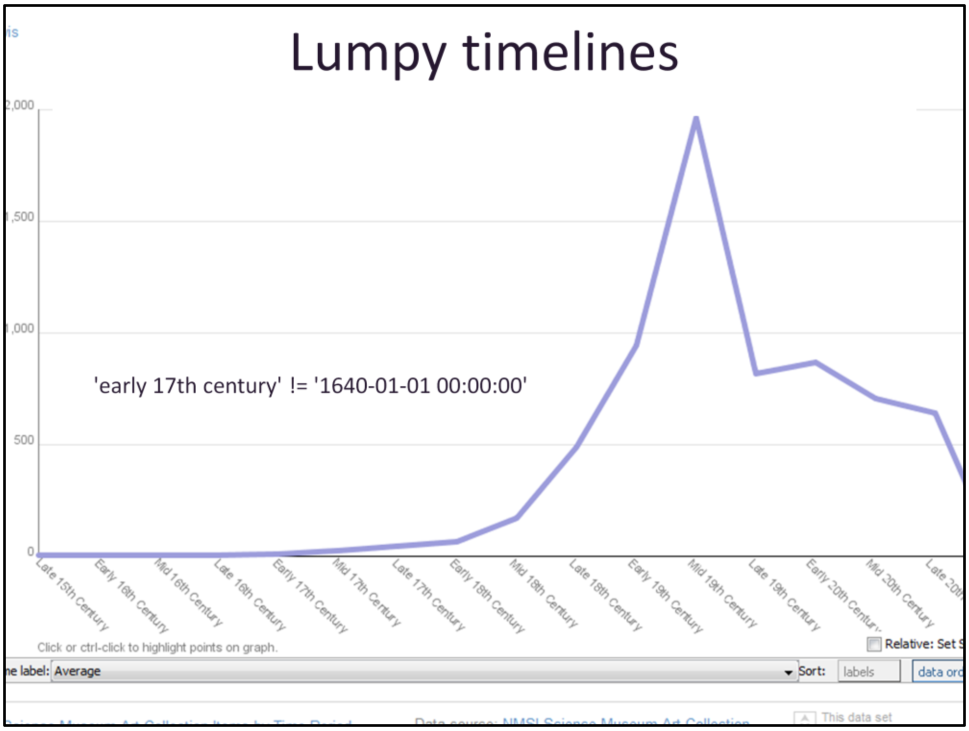
The ~~cake~~ map is a lie

Where are (some) objects from? At Cooper Hewitt I made a map which shows which countries have been collected from most intensively. I had to remove out any rows that had values that didn't exactly match the name of just one country, etc, so it doesn't represent the entire collection. But you can get a sense of the shape of the collection – for example, there's a strong focus on the US and Western Europe objects.

This also demonstrates the impact of the different tools – I'm sure the Cooper-Hewitt has more than 43 objects from the countries (England, Scotland, Wales and Northern Ireland) that make up the UK but Google's map has only picked up references to 'United Kingdom', effectively masking the geo-political complexities of the region and hiding tens of thousands of records. It might also be hiding records from under-represented regions like Africa - hard to know whether records under previous names for countries might have shown up on a historically-aware map or geocoder.

So we have two problems - in the first map I showed, data has mistakenly been elided for us; in the second, the map is only representing a subset of the overall collection.
http://bit.ly/Ls572u or
https://www.google.com/fusiontables/embedviz?viz=GVIZ&t=MAP&gco_region=world&gco_dataMode=regions&containerId=gviz_canvas&q=select+gvizcountry%28col0%29%2C+col1%2C+col0+from+19Wuxyb12xrM1vn828Xi3XPhb4nlCqZnb_Hu188k&qrs=+where+gvizcountry%28col0%29+%3E%3D+&qre=+and+gvizcountry%28col0%29+%3

C%3D+&qe=+limit+134&width=500&height=300



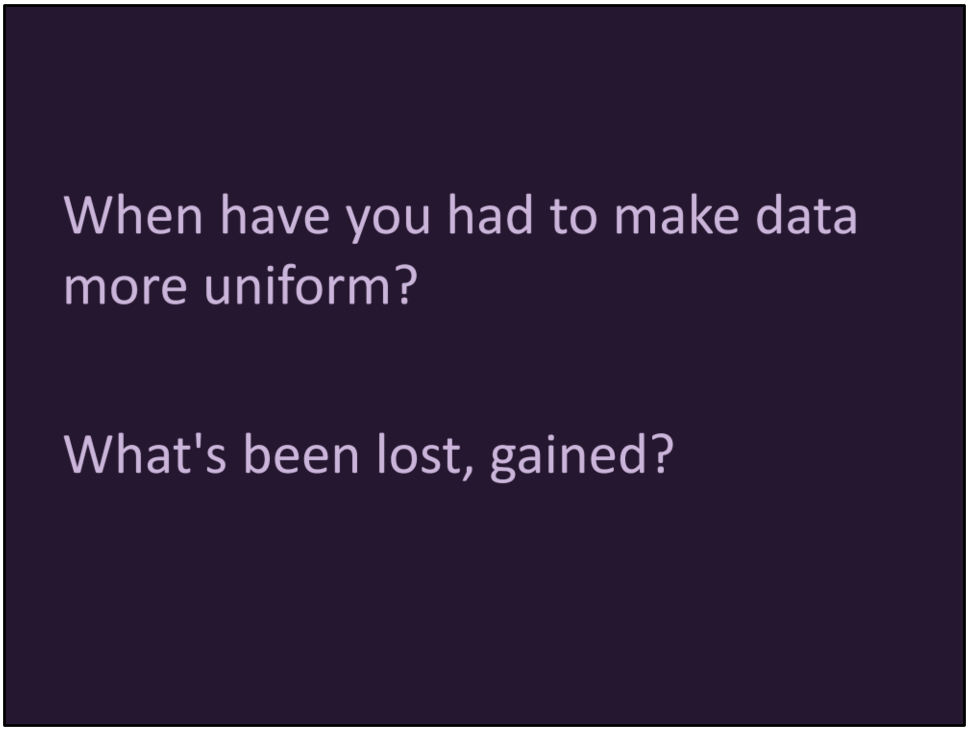
Lumpy timelines

'early 17th century' != '1640-01-01 00:00:00'

A third problem is the temptation to clean data up so it can be displayed on generic tools, introducing false precision and odd lumps in datasets. Date ranges are common in museum data - objects might be linked to a Period which has date ranges, or might be assigned 'earliest' and 'latest' possible dates when nothing more precise can be known, but apart from tools like Neatline, very few timelines deal with fuzzy date or date ranges.

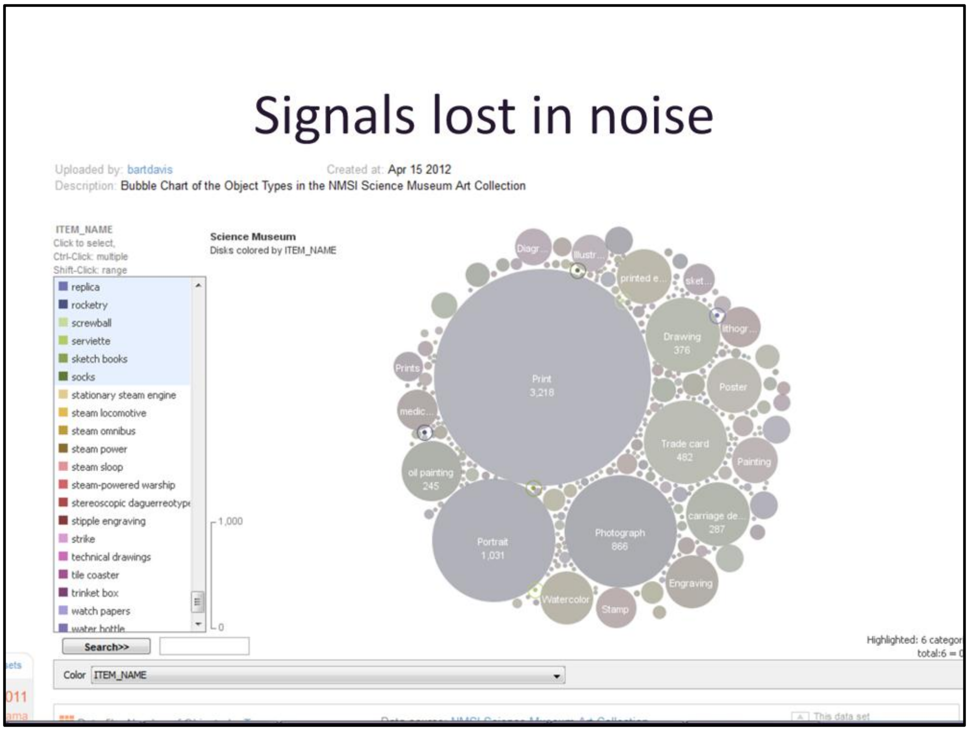
If you're working with a generic timeline generator that wouldn't know what to do with 'early 17th century' as an input value, flattening it to 1640 or another representative year creates a false level of precision that has the appearance of accuracy. Fudging it means it'll display, but at what cost?

One solution is to record separate 'display dates' - a public-facing summary date for individual objects - but this requires curatorial knowledge. Another that I haven't explored because it seems as almost as misleading as other solutions is to add random noise to spread data out across years within a date range.



When have you had to make data more uniform?

What's been lost, gained?



Signals lost in noise

Uploaded by: bartdavis Created at: Apr 15 2012
Description: Bubble Chart of the Object Types in the NMSI Science Museum Art Collection

ITEM_NAME
Click to select,
Ctrl-Click: multiple
Shift-Click: range

- replica
- rocketry
- screwball
- serviette
- sketch books
- socks
- stationary steam engine
- steam locomotive
- steam omnibus
- steam power
- steam sloop
- steam-powered warship
- stereoscopic daguerreotype
- stipple engraving
- strike
- technical drawings
- tile coaster
- trinket box
- watch papers
- water bottle

Search>>

Color: ITEM_NAME

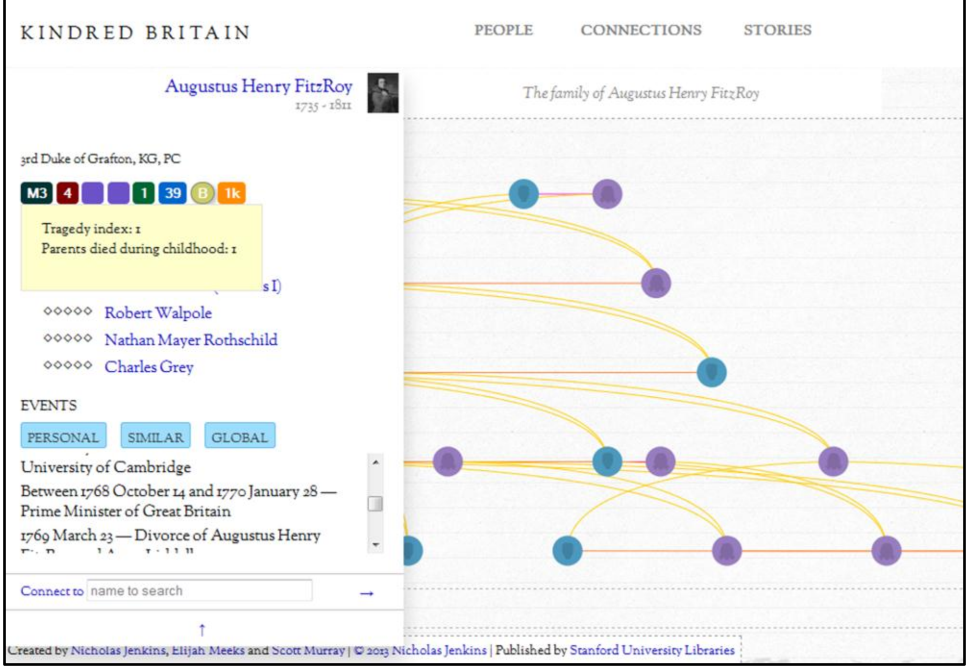
Highlighted: 6 categor
total:6 = 0

Science Museum
Disks colored by ITEM_NAME

http://www-958.ibm.com/software/analytics/manyeyes/visualizations/object-types-in-the-nmsi-science-m

Replicas, rocketry, screwball, serviette, sketch book, socks… none of these have enough comparative volume to show up in most visualisations, but surely they have interesting facets to reveal…
Messiness between individual records within a dataset (whether from different departments collecting different types of objects record different information, or objects spanning a couple of thousand years have different information recorded about them) translates into noise at the aggregate level. We need better methods for smoothing variation in the distant view and revealing it in the zoomed in, detail view. But this means deciding what's important, which differences can be squished without being misleading. Attempts to provide shortcuts can be problematic as they assume certain modes of use.

Not on the label?



Augustus Henry FitzRoy
1735 - 1811

The family of Augustus Henry FitzRoy

3rd Duke of Grafton, KG, PC

Tragedy index: 1
Parents died during childhood: 1

s I)

◇◇◇◇◇ Robert Walpole
◇◇◇◇◇ Nathan Mayer Rothschild
◇◇◇◇◇ Charles Grey

EVENTS

University of Cambridge

Between 1768 October 14 and 1770 January 28 — Prime Minister of Great Britain

1769 March 23 — Divorce of Augustus Henry

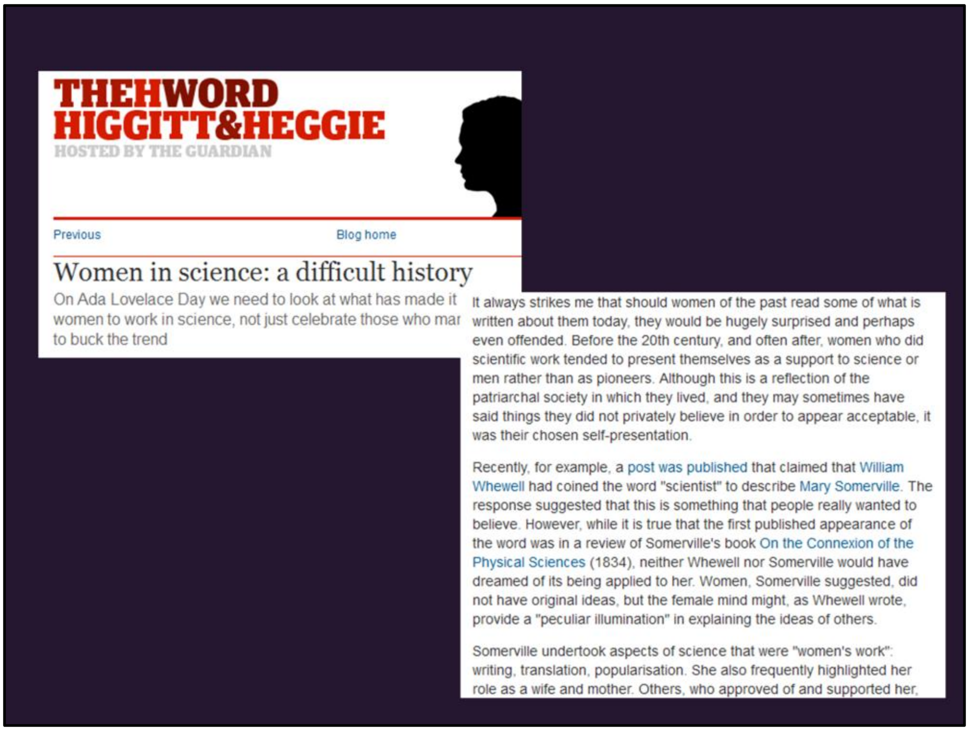
Created by Nicholas Jenkins, Elijah Meeks and Scott Murray | © 2013 Nicholas Jenkins | Published by Stanford University Libraries

http://kindred.stanford.edu/#/kin/none/full/none/I9190// Not to pick on this project, but when it was launched, something on the page really stood out to me. A tragedy index? Isn't that ahistoric?

More broadly, what's the effect when you preconfigure certain 'attributes' for analysis? After all, 'tragicity' might be less important than latent lefthandedness, birth order or access to education but you can't as easily search or browse on those. It's great to provide one way in, but what explorations does it prioritise and which might a database optimised for that preclude?

Or if this term comes out of the researchers' personal interests and the data has been collected through a particular lens, how should the site render that context visible?

'Tragedy Badge: A total of events in a person's life defined as tragic and derived from the database: 1 point for dying young or to violence, 1 point for each child that dies before the age of 13, 1 point for each sibling that dies before the age of 13, 1 point for each parent that died during childhood, 1 point for each spouse that's outlived by more than 20 years, and 1 point for mental illness. See the conceptual story on Tragedy.' http://kindred.stanford.edu/notes.html



THEHWORD
HIGGITT&HEGGIE
HOSTED BY THE GUARDIAN

Previous

Blog home

Women in science: a difficult history

On Ada Lovelace Day we need to look at what has made it women to work in science, not just celebrate those who mar to buck the trend

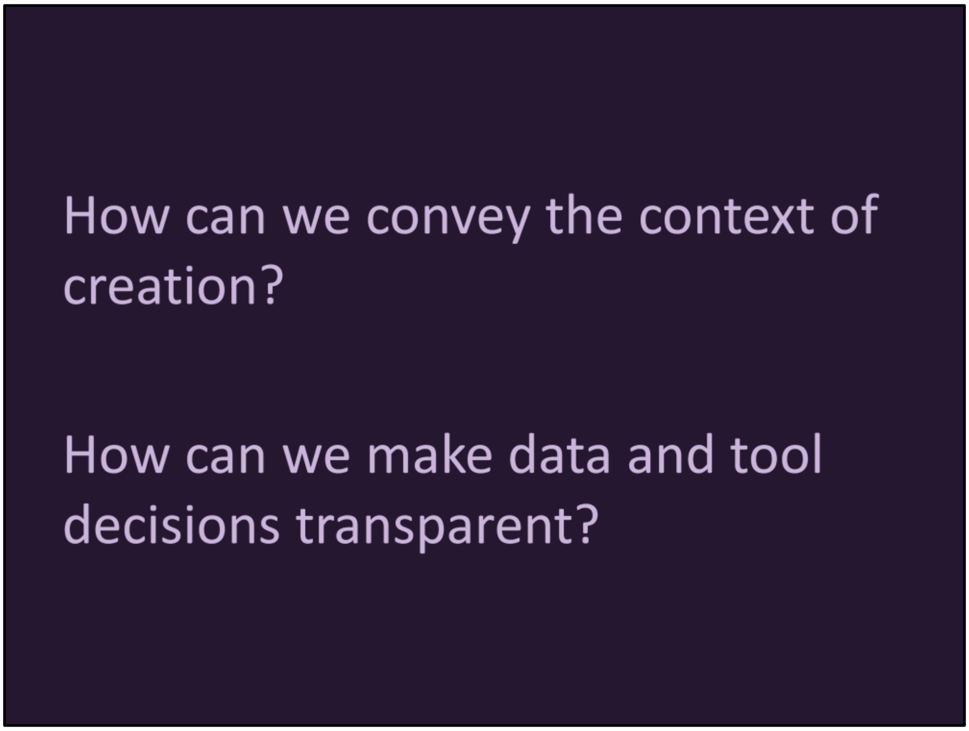
It always strikes me that should women of the past read some of what is written about them today, they would be hugely surprised and perhaps even offended. Before the 20th century, and often after, women who did scientific work tended to present themselves as a support to science or men rather than as pioneers. Although this is a reflection of the patriarchal society in which they lived, and they may sometimes have said things they did not privately believe in order to appear acceptable, it was their chosen self-presentation.

Recently, for example, a post was published that claimed that William Whewell had coined the word "scientist" to describe Mary Somerville. The response suggested that this is something that people really wanted to believe. However, while it is true that the first published appearance of the word was in a review of Somerville's book On the Connexion of the Physical Sciences (1834), neither Whewell nor Somerville would have dreamed of its being applied to her. Women, Somerville suggested, did not have original ideas, but the female mind might, as Whewell wrote, provide a "peculiar illumination" in explaining the ideas of others.

Somerville undertook aspects of science that were "women's work": writing, translation, popularisation. She also frequently highlighted her role as a wife and mother. Others, who approved of and supported her,

This is where the humanist side of being a DHer comes in… I'm interested in women in intellectual history, but get tangled up in questions about naming - I'd love to reclaim the term 'bluestocking' but in gathering biographies under that label, I'm retrospectively applying a label that most of those women wouldn't have applied to themselves, and wouldn't recognise. This issue isn't unique to heritage data but as DH eyes 'big data' it's worth keeping in mind.

http://www.theguardian.com/science/the-h-word/2013/oct/15/women-science-history-ada-lovelace-day



How can we convey the context of creation?

How can we make data and tool decisions transparent?

Image by Flickr user open_knowledge under CC-BY-NC

Moving from shaping the peg to the hole, to shaping the hole to the peg.

Catalhoyuk is a neolithic site in Turkey, considered the world's first city. It's a long-term research dig, with specialists from different institutions based in different labs and teams. I spent a few years designing and building their database systems, including some summers on site. My first task was to normalise, consolidate and centralise the various existing research databases held by different specialists.

While interviewing people about their recording requirements, I realised that there were many variations on similar term lists for materials, sizes, etc, and core values for different types of objects, and that the boundaries between some object types were very uncertain. It could be hard to tell whether something was a clay ball, an incomplete or broken figurine, a sherd of pottery, a stamp seal, building materials or just a blob of clay (nothing like this beautiful figurine) - each of which had a separate database with term lists that had diverged from the other databases. Ambiguous objects had to be entered in one specialist database or another to be recorded at all, but were at risk of being mislabelled, also artificially increasing the count of figurines or sherds found at the site, and perhaps increasing the workload of some specialists who had to record each item. The specialism-lead system was artificially labelling uncertain objects as being specific types then imposing database-specific constraints on what could be recorded about them. As a result of assigning artefacts to particular specialists to record, they were subtly but implicitly labelled by the specialism and specialist application within which they were recorded. The prompts provided by database fields possibly also affected what people expected to see in those blobs of clay. (e.g. figurines mixed up materials (type of clay, inclusions, etc) with fields for representational aspects). The lack of prompts for other artefact types may also have lead to useful attributes not being recorded. This was definitely a case of the tools affecting the data.

Artefact-led recording

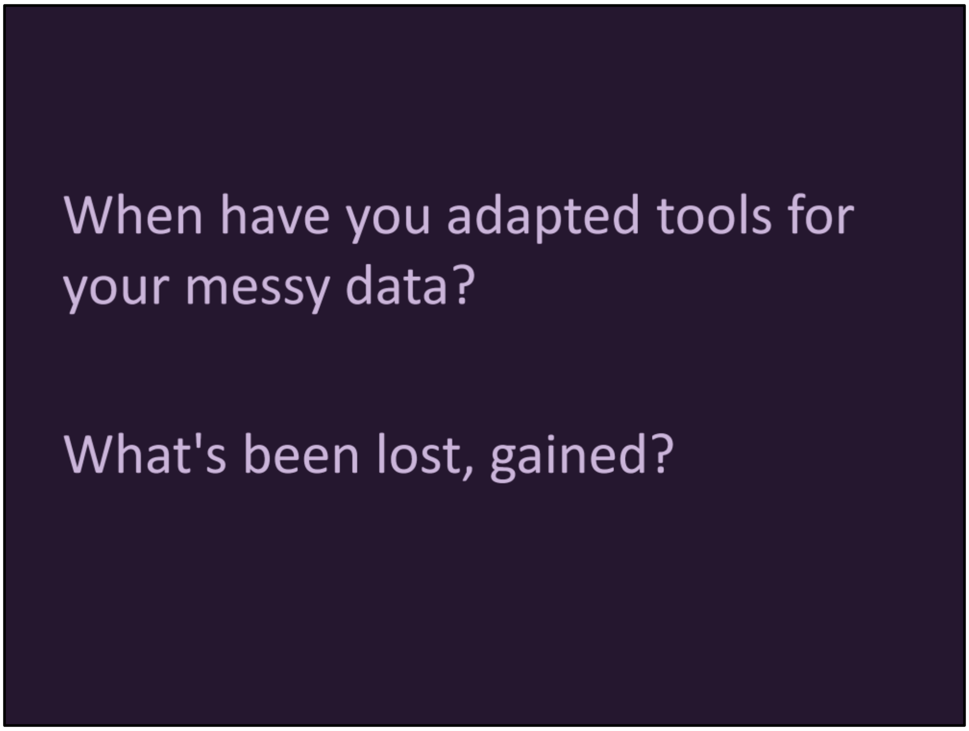
(Image looks like a figurine maybe but is recorded as a stamp seal. Bear? Maybe.)
The cross-specialisms view I took to negotiate where shared term lists could be used (designed to improve cross-database searching) meant I ended up with an overview of all the recording systems and ended up introducing a system of artefact-lead recording, based around a model of core and extension fields. The idea was that artefacts can be recorded as core fields first (recording the existence of the artefact, linking it to the finds and excavation databases), moving from technical or material-specific recording to interpretative or specialist recording if and when it is supported by the characteristics of the artefact itself, rather than recording artefacts through the filter of specialist interpretation that look to fit it into a preconceived model. The recording and analysis moves from the objective to the subjective, the general to specialist, from technical to interpretative data.

The implementation of shared lists of values as part of the recording of interpretative aspects of an artefact allows the direct comparison and analysis of artefacts of the same type and different materials. For example, the characteristics of stone, bone, shell, glass and clay beads could be analysed and compared, regardless of the material from which they were constructed.

Recording all artefacts in core tables enables incomplete, miscellaneous and indeterminate artefacts to be recorded without affecting the quality of the entire dataset, rather than requiring incomplete recording in specialist tables. Importantly, it avoids forcing an interpretation on an artefact at an early stage for convenience during data entry. For example, this model saves recording an indeterminate blob as a 'figurine' or 'clay ball' when it may not have enough surviving characteristics to determine it as either.

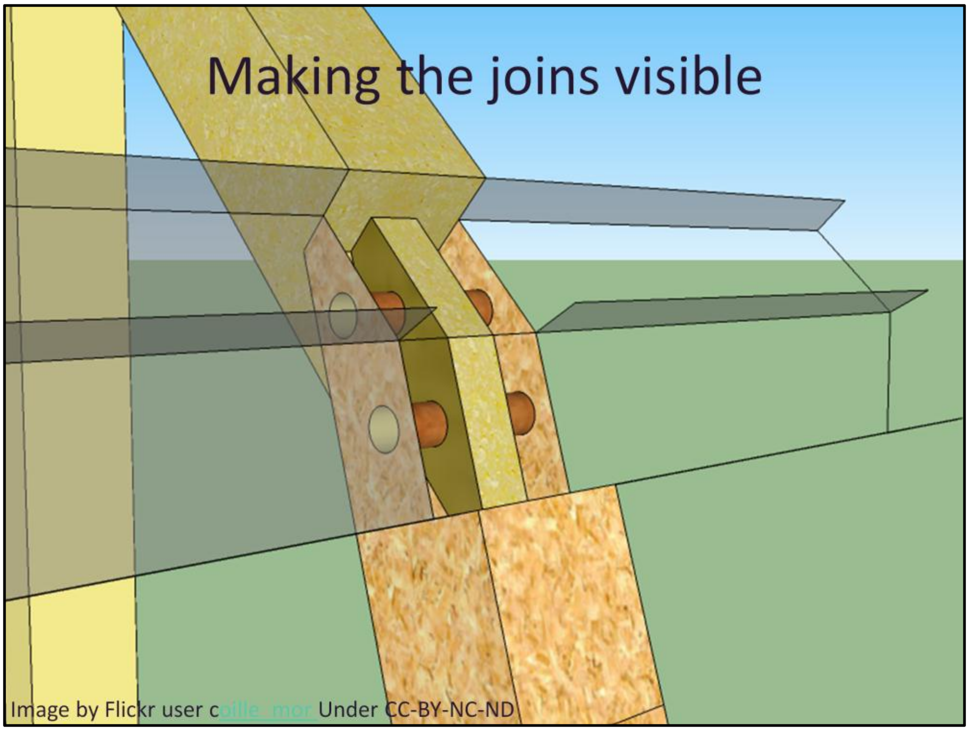
On the other hand, much more complicated to program so created a lot of complexity for the project, needs commitment to listening to the finds. More at http://www.catalhoyuk.com/archive_reports/2005/ar05_37.html

Image credits me or http://www.flickr.com/photos/catalhoyuk/



When have you adapted tools for your messy data?

What's been lost, gained?



Making the joins visible

Image by Flickr user coillie_mor Under CC-BY-NC-ND

A bit of a SQL pun there…

You used to be able to see the joins, the marks of construction in the objects around us, but that's no longer the case. How to carry contingency, data loss, conflation etc in a dataset? When visualisations might appear in print, in contexts far from original source?

Discussion: Share your own stories. How have you worked to incorporate ambiguity or contradictory evidence in humanities computing projects? When have you decided to elide it, why, and with what impact on the scholarly arguments your tool enabled?

Image credit http://www.flickr.com/photos/akottenstette60account/6064112314/