Inclusive Data Research Skills Hackathon for Arts and Humanities (DAReS)

Data Skills Session

DAReS Hackathon, 26 Jan (10:00-16:00)

Facilitators: Gauti Sigthorsson (U. of Roehampton), Carla Fernandez (UAL), Harry Solomons (UAL, LCC)

Session Summary

This session supports the overall aim of today's event, to co-create inclusive data skills assets and materials for arts and humanities researchers with Wikimedia UK and the DAReS project team.

The theme of this session is "What works for Arts & Humanities practitioners when working with data?" Therefore, we aim to explore arts and humanities-specific approaches to data, tools/methods and research outputs. The data skills section of our wikibook is here.

Participants are welcome to move between groupings as they wish.

Facilitators will work with the following themes as starting points:

Grouping 1: What counts as data?

(Facilitator: Harry Solomons)

This session will work on broader questions on how data and arts and humanities disciplines interact, particularly around the question: what data is useful and relevant in arts and humanities disciplines?

- What counts as "data" in arts/hums?
    - The distinction between qualitative and quantitative data
    - Quantitative: most readily apparent in social sciences, measurable
        - Case studies:
        - Pedestrianisation work
        - UK deprivation visualisation: combination of old techniques with new data
        - Untitled (Portrait of Ross in L.A.) by Felix Gonzalez-Torres: applying a form of quantitative data analysis into an artistic piece (would this fit better in storytelling session?)
        - Work of Nathalie Miebach: art with data as a resource
    - Qualitative: descriptive, language-based, very useful for humanities research (prompt: how is it useful in arts?)
        - Case studies:

        - Living with Practical Realities by Stephen Willats: artistic representation of what is, at its core, research data
        - Is this easier to understand as arts/humanities professionals? Why/why not?
    - Follow-up: how can you move between the two?
    - Extension resources:
        - Frayling: research in art and design
- How are primary and secondary data sources used in arts/hums?
    - What data can be collected primarily?
    - Collect case studies
    - How is secondary data used?
    - Collect case studies
- What sources of data are available to arts/hums professionals?

Grouping 2: Data Visualisation and Storytelling

(Facilitator: Carla Fernandez)

Focusing on data visualisation as both a research method and a practice, this session focuses on the route to exploration, discovery and the storytelling of information.

- **Explore**

  We are all visualisers: Drawing your breath. 5 min exercise

  - What is data?
  - Pieces of information that convey a message about the subject of our research
  - Who can visualise information?
  - What tools are available for data visualization?
  - How can researchers in arts and humanities utilise data visualisation as a tool, and for what purposes?
- **Discover**

  Data relationships: How to find the stories in data

  - What information can be documented in data?
  - What insights can you extract from data?
  - Which chart should be selected?

- **Explain**

  Storytelling with data: What I ate today. 15 min exercise.

  How can researchers in arts and humanities utilise data visualization to articulate and narrate their findings?

  - Data Visualisation process
  - Narrative in Data Storytelling

Grouping 3: Data practices

(Facilitator: Gauti Sigthorsson)

Connecting projects/topics to specific tools and methods. What implications do choices of tools and methods have for inclusion/exclusion?

- What works for artists and humanities researchers when working with data?
- Methods, techniques, tools: What are the tools of choice for you as A&H researcher/practitioner?
    - Miro
    - Zotero (collection) + Obsidian (themes coding + visualisation)
    - https://kinopio.club

- What data and digital research skills are most relevant for arts and humanities practitioners/researchers?

Drafting / sketching section

Define [Object of study] and the [Data problem]

What are the ethical implications of the methodological choice of tools for handling data?

- Institutional ethics
- GDPR
-

Organise tools into the following:

Gather	Access	Analyse	Present
Interviews	Facilitate access of marginalised groups to reach outputs, the researchers	Thematic Content analysis	Powerpoint
Focus groups		Knowledge graphs?	Keynote
Story telling		Lives/stories from different backgrounds	Thick detailed descriptions of phenomenon
Personal observations			

What tools can support the Gathering of data?

1. Language translators
2. Voice recorders
3. Cameras
4. pens/writing books

What tools can support the Access of data?

1. Hard tools like legal policy frameworks, regulations,
2. Soft tools like convention/ traditions
3. Information sharing
4. Training/empowerment of user community

What tools can support the Analysis of data?

1. Computer Assisted Qualitative Data Analysis Software,
2. Software like Python,
3. Nvivo?
4.

How to have a less binary relation with a tool? Ex.: Data analysis done through GAI or researcher's insights

Dataset Assessments

Data Skill Sets and Practices for Humanity and Arts could be explored through the DIKW framework [https://en.wikipedia.org/wiki/DIKW_pyramid]

The journey of Dataset exploration for insight into Research Questions goes through

Data Collection

Humanities and Arts data collection methods are varied and often include structured , tabular datasets as well as unstructured datasets from interviews, surveys , books , photos, images, blogs or social media. The growing trends towards digitisation has expanded the potential datasets to other modalities like Audio and Video. The sensory devices and IoT devices continue to add to the existing types of datasets.

Data Organisation

For multimodal datasets , organisation of the datasets by the object(s) of research questions is an important step before any evaluation or analytics could be performed in a meaningful way. This step includes looking at the different sources of data -

primary, secondary , different types of datasets - quantitative , qualitative as well as different modalities of datasets - text, audio, image, video

The key concern of this step is to be able to organise the varied data sources , types and modalities into a coherent form that would enable it for further evaluation and analytics.

An example of such organisation is given here through the assessment of online ecommerce behavioural changes since the onset of pandemic. The research datasets for this example could potentially be a combination of the publically available datasets with consumer in store and online purchases before and after the pandemic , social blog posts with consumer sentiments, newscast about the shifting habits, interviews with Retail business representatives, geo spatial movement of people and vehicles.

Organising such a vast collection of data needs to be done using merging techniques that would help to group the subjects into meaningful categories for further evaluation. Whilst the current methodologies mainly focus on organising each datasets individually, it could be argued no such independent evaluation of datasets would ever be deemed complete until they are combined for the research purpose.

Data Evaluation

Independent of modalities - structured or unstructured, sources - public, private, qualitative, quantitative methods, datasets very often have gaps of issues with individual values of a given attribute or certain data types could even be restricted under GDPR and other Privacy laws. In such scenario two types of evaluation of the combined datasets become important consideration

Ethics and Provenance

Understanding and authenticating sources for all datasets is a very important step. This includes checking the usage of datasets under licence - Creative Commons, Data Licensing, Data Sharing Agreement. Each data field would need to be evaluated for Personal Protected Information and Sensitive Information to assess if the data used for the research has been reviewed through all possible guardrails. Where conflicts are found, this step would include corrective measures like data aggregation or synthetic data as privacy protection techniques

Missing, Incorrect data or Outliers

Decisions for Missing , Incorrect datasets as well as outliers within a dataset should be taken based on the subject and context of the research. Under certain circumstances , the missing or incorrect data elements could simply be removed from the dataset without impacting the research whilst in other instances a prudent approach would be to replace the missing, incomplete data elements with the average of the sample set of observations, records.

This step could sometimes also lead back to the data collection methods and result in the need for additional data collection or a different method for collecting data if the

entire dataset is found to be biased or not allowed to be used for the purpose of research

Histograms, Bar charts are helpful tools to quickly scan the datasets for outliers or skewness for an assessment of underlying issues with bias in datasets or inaccuracies in the datasets.

Data Transformation

Conversion between modalities (e.g., text to audio)

Conversion of analogue archival materials into digital formats (Optical Character Recognition), card catalogues to digital databases.

In preparation for running various data analytics on research datasets it would be sometimes useful to augment one modality with another for e.g. in the field of art - it would be very useful to combine the textual data about the work of an artist with the paintings, images by the artist with other documentaries or archived records. Such an augmented dataset would be very helpful to study the socio political landscape during the period of a given painting or image. Knowledge Graphs could potentially be used as a tool at this stage to combine the modalities to create relationship between various modalities for a given subject

Data Analytics

Asking questions and answering them using data, to derive insight(s).

An important part of the data analytics step is to differentiate numerical insights from data vs context assessment. Whilst numerical analysis based on statistical methods could provide insights about percentage of population, rate of growth or decline , context assessment techniques help to understand the subject better this includes - topic analysis, summary of long paragraphs or passages, sentiment analysis

Generative AI models or classical NLP techniques have been found to be good at doing such contextual analysis.

Access

It is necessary to give less privileged groups/communities voices through new ways to access existing data/content platforms.
The Wikimedia/wiki book or a university department could create a new system of accreditations to allow those groups to access the many block/paywall websites.

Making connections

Knowledge graphs (webs) map connections between data points, connected nodes in a network.

An important concept for making connection is related to representation and learning from datasets. Humans often struggle to make sense of huge datasets but also varied datasets. We might have tabular data about population and news clippings about socio economic sentiments and time series data about inflation, but combining them

together to deduce wisdom is often very challenging for humans. Machine are good at processing huge volumes of data , having better memory at defined connections.

Connections made by humans (using logic, analogy, similarity).

Connections made with machine learning (statistics).

Data Protection

Governance

What is the current legislation around the type of data and region that it is being collected?

Does your organisation have to register it's data usage?

Where is this data being stored?

Compliance

Do you have data compliance officer?

Who has access to this data?

What are you data collection methods?

Will this data be anonymised?

How long will this data be stored for?

How will this data be disposed of after collection?

Can people opt out or remove themselves or their data from your database?

#Recycled for re0use actually

- What data and digital research skills are most relevant for arts and humanities practitioners/researchers?

  The existing data/research supplied by those knowledgeable in their fields/themes can allow new practitioners to develop those subjects further, adding the many missing voices. Also, translating the many existing works from other continents, namely from the Global South, could add much value to make the current Arts and humanities texts more factual.

- What works for artists and humanities researchers when working with data?
  - ➢ Open access

- Trustworthy sources
- Sources
- Diversity

Data Storage

Images from session: