Summary of Dataset Contextualisation workshops (June-July 2025)

Background

Humanities and cultural heritage datasets typically have a complicated context. Understanding this context is key to correct, responsible use of the data. The Huygens Institute is working on contextualisation descriptions of datasets, in a project together with the City Archives of Amsterdam and Sound & Vision, and in consultation with other experts investigating this topic, such as the Europeana Datasheets for Digital Cultural Heritage working group and SSHOC-NL. Dataset providers need a format to help them document and share context, while dataset users need this context to help them use datasets. It is essential to include these groups in concept development, so we held a series of workshops to consult both dataset providers and dataset users.

This document summarises the results of dataset contextualisation workshops held in June/July 2025 at the DHBenelux conference, the Clariah Summer School, Sound & Vision and the Huygens Institute. In total, 30-40 participants took part, representing a wide range of academic institutes, cultural heritage organisations, museums, archives and libraries.

In the workshops, we discussed the underlying principles of contextualisation descriptions, illustrated by examples of real datasets documented with the data-envelopes format.

The workshops explored the following questions:

- Where do users search for datasets and what information do they want?
- Which information is needed in contextualisation descriptions?
- Where can providers find the information for contextualisation descriptions?
- Which vocabularies are required?
- Which export formats are required?
- Which systems are used for datasets and context?
- How to incorporate contextualisation descriptions in workflows and systems?
- General feedback on contextualisation descriptions

In the following sections, we summarise the responses to each question, with the most important notes and insights in bold. Finally, we will discuss our conclusions.

Where do users search for datasets and what information do they want?

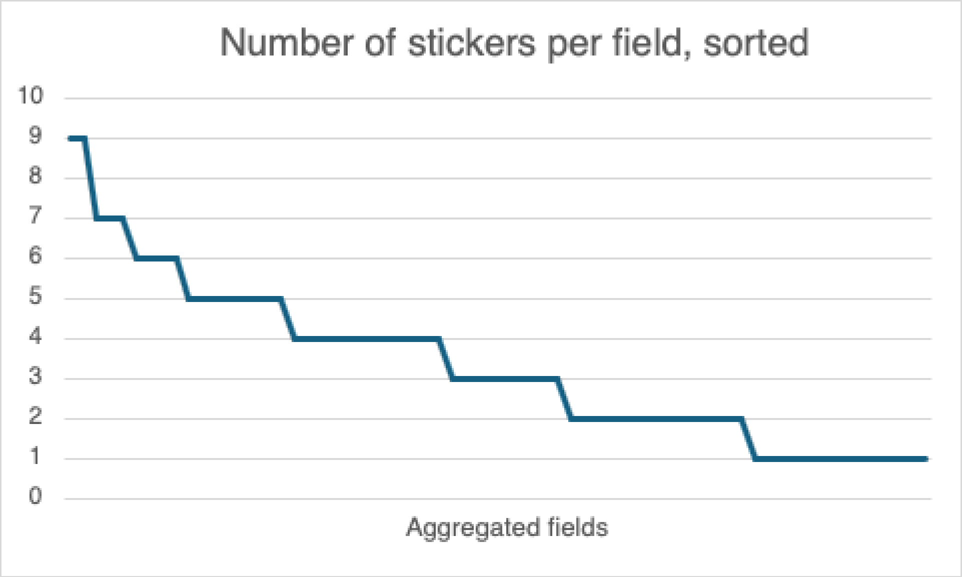
This question was discussed in the Clariah Summer School and Sound & Vision workshop. A summary of the discussion is:

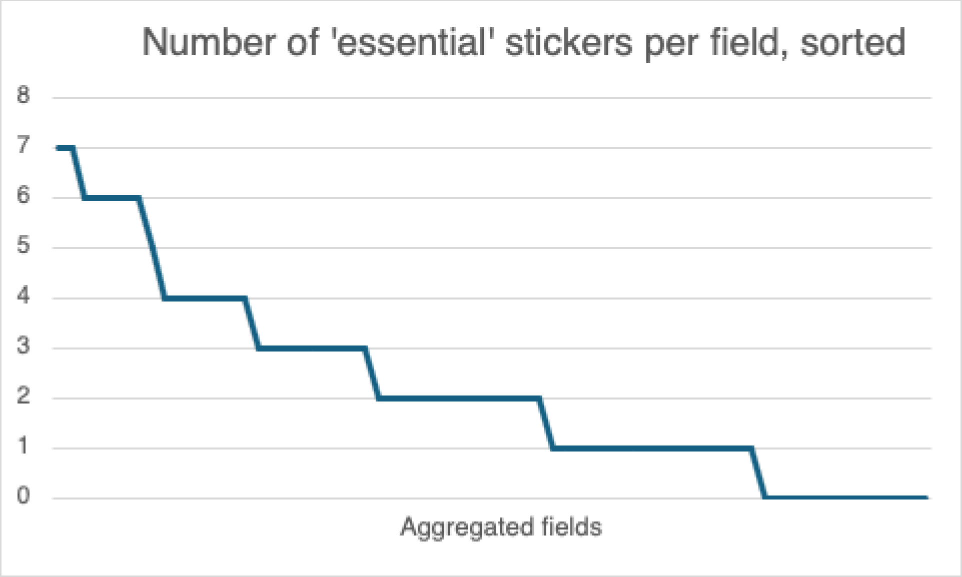
- Users search using generic tools (e.g. Google), dataset repositories, specific archives/platforms, literature, human experts and other
- Search for datasets is conducted by looking for persons/place/time, language, keywords and terms from vocabularies
- Key challenges are rights/access/ethics, technical and quality issues, dataset structure.
- **Discussion of search morphs from searching for datasets to searching for individual items quite seamlessly – users don't draw a strict line between the two**
- Users want to know: what is in the dataset, how was it produced, information about quality/biases/gaps, licenses and uses. Also, individual item metadata

Which information is needed in contextualisation descriptions?

This question was discussed in all four workshops. The fields contained in the data-envelope format were aggregated to produce a simplified version with 66 fields. Participants used stickers to indicate which fields they found 'essential', 'important' and 'nice-to-have'. They wrote post-its describing information they required that was not provided in the format.

- All 66 aggregated fields received at least one sticker
- The fields 'Biases' and 'Errors' received the most stickers, 9 each, of which 7 were 'essential' for 'Biases', and 6 were 'essential' for Errors.
- The fields 'Biases' and 'Access' received the most 'essential' stickers, 7 each.
- There was a gradual reduction in the perceived importance of the fields, as can be seen in the graphs below. There was no obvious cut-off point between important and less important fields





It is hard to compare dataset providers to dataset users as the workshops contained participants with a mix of roles, and some people have both roles. Comparing the Sound & Vision workshop (more focus on users) with the Huygens workshop (more focus on providers), we see that approximately the same number of fields were regarded as essential (32 and 34 respectively). However, only 13 fields were regarded as essential by participants in both workshops.

Only 7 fields were regarded as essential by participants in all four workshops.

There were many suggestions for additional fields or additional detail in an existing field. These were very diverse, ranging from data schemas to geographical distributions to information on consent. Recurring themes were fields related to machine readability, more precise definitions and standardisation.

In summary, all information currently suggested for contextualisation descriptions is essential to at least one participant, and a significant amount of additional information is desired. This, together with the relatively small overlap between groups as to which fields are regarded as essential, **may suggest that contextualisation descriptions should be made up of a large pool of possible fields, from which dataset providers can select those relevant to their dataset, and dataset users can view those relevant to their use case.**

Where can providers find the information for contextualisation descriptions?

This question was explored in the DHBenelux conference and Huygens workshops, where many dataset providers were present. Using the aggregated 66 fields of the data-envelopes, participants used stickers to indicate how they would fill in the required information.

- Participants did not sticker all the fields
- Of the 42 fields stickered:
  - 20 participants would fill in the information themselves
  - 14 would ask a colleague
  - 6 would like to retrieve the information automatically from the dataset
  - 2 did not know how to find this information

As not all fields were stickered, more work is needed to investigate this question. Possibly this is a question participants can best answer while creating a contextualisation description themselves. **It is in any case apparent that filling in such descriptions mostly requires knowledge about that data from more than one person.**

Which vocabularies are required?

This question was explicitly asked in the DHBenelux conference and Huygens workshops, where many dataset providers were present. Participants in the Sound & Vision workshop answered it spontaneously.

The following vocabularies or types of vocabularies were suggested:

- Dataset features
  - Languages
  - Topics

    - Data subjects
    - Geographical locations (e.g. Geonames)
    - Genres
    - Temporal
    - Dates (e.g. CIDOC)
- Domain information
    - ELSST
    - ISIL
    - Institutional vocabularies
    - Termennetwerk
- Other
    - ORCID
    - Licenses (e.g. ODRL)

Which export formats are required?

Formats for exporting contextualisation descriptions were discussed in the Sound & Vision and Huygens workshops. The following formats were mentioned:

- Formatted text
    - Markdown
    - PDF
- Linked data
    - RDF
- Tabular
    - TSV
    - CSV
- Interchange
    - JSON

One comment was that any format is fine **as long as it is in common use and machine readable.**

Which systems are used for datasets and context?

This was discussed during the Huygens workshop, which focused on dataset providers.

The following systems are currently used by workshop participants for storing/sharing datasets:

- Data dump
- Dataverse
- Dataset register (such as the NDE register)
- APIs

- CMS (e.g. MAIS-Flexis, Memorix Nexus)
- Yoda
- OSF
- Data stations
- Figshare
- S3 data storage
- Own RDF-based system
- OAI-PMH
- SPARQL endpoint

The following ways of communicating context are currently used:

- Manuals
- README files
- Introduction to inventories
- Articles
- Github
- Datasheets
- Elements in machine-readable metadata (e.g. Description)
- Human contact point

How to incorporate contextualisation descriptions in workflows and systems?

This was discussed during the Huygens workshop, which focused on dataset providers. A summary of the discussion is:

- Participants see possibilities for combining contextualisation descriptions with their workflow, data model and systems
- Preference for in own system
- Opportunities for contextualisation descriptions for improving standardisation and shareability
- Problems with required workload. Potential solutions: search for balance, lighter version, involve more people in filling in, use of AI
- Problems with organisation/domain dependency of the information
- Many questions still remain to be answered

General feedback on contextualisation descriptions

The concept of contextualisation descriptions was positively received and sparked many discussions. The following points were identified that require further attention:

- Unclear fields/structure
  - More explanation needed (manual and instructions)
  - Clearer, unambiguous field names
  - Meaning of fields that are empty in the final dataset contextualisation description
  - Reconsider grouping and repetition in fields and levels
- More information needed in some fields, primarily in levels 2 and 3
- Presentation
  - Format (e.g. clickable, static)
  - Order in which information is shown (may be task dependent)
  - Collapsing/expanding levels
  - Show empty fields or not
- Languages
  - Language of the form itself: Multiple languages of the form itself are needed to accommodate diversity of users. For example, quite often the volunteers creating descriptions may prefer a certain language that they are more comfortable with for reading the fields descriptions.
  - Language of the information collected and shared: Different user communities may need different languages of the contextual information to fill in or work with the information
- Contributors
  - Tricky to fill in
  - Privacy issues
- Who fills it in?
  - How to involve the community?
  - Importance of 'collections as data' perspective for the person filling the description in
  - How to maintain different versions
  - How to incorporate points of view of different groups
  - Collaboration
- When should a dataset have a contextualisation description?
  - Not all datasets?
  - What if no access?
  - Depends on the value of the dataset for CH
  - Are there users who need the extended metadata?

    - Cultural differences between universities and archives
        - Priority/urgency for creating descriptions
        - ‘Ownership’ of datasets (individual researcher or the archive)
        - Who creates the descriptions (individual researcher or archive employee)
- Relationship to other documentation/rules
    - README
    - NDE Dataset register
    - Do contextualisation descriptions adhere to the EU AI act?

Conclusions

The positive reactions and engagement in the workshops confirmed the value of contextualisation descriptions for this community. The participants were united in the need for this information, but at the same time very diverse in the specifics of which information and how it should be presented. This diversity, combined with the existing diversity of systems, workflows and datasets, suggests a need for flexibility in the implementation of contextualisation descriptions. At the same time, there was an emphasis on standardisation and machine-readability. Combining flexibility with standardisation is a challenge for future development.

Further work is needed to improve contextualisation descriptions and to answer open questions about their practical implementation.

The feedback gathered will be shared with the institutes and organisations involved in investigating contextualisation descriptions, to guide further development of such descriptions, tools for their creation and viewing, and their incorporation into workflows and systems.