Summary page:
TEHDAS2 public consultation on Draft guideline for data user navigating the catalogue

Preliminary remark: Before you start reviewing the user guide, please open the page (https://acceptance.data.health.europa.eu/healthdata-central-platform?locale=en) and try to follow the steps and instructions as if you were a non-technical first-time user.

This consultation has 7 pages and 32 questions. The first and the second pages are common to all TEHDAS2 public consultations and cover demography of the responder and overall quality of the document. Pages 3, 4, 5, 6 and 7 consist of questions specific to this document.

Demography

1. Country *

Netherlands

2. Type of responder *

Other
Research Infrastructure

3. Are you responding on behalf of several organisations? *

If yes: On behalf of how many organisations?

No

4. Sector *

Other
Research Infrastructure

5. Organisation size *

Small to medium enterprise (10–249 employees)

6. Professional role / function

Program Manager

Quality

7. Is the document easy to understand? *

2

8. How well does the document address the key issues related to its subject matter? *

4

9. How feasible do you find the guidelines or technical specifications to implement, as outlined in the document? *

2

10. Generic feedback

Do you have any suggestions for improving the document? Are there any additional topics or areas that should be covered? Max. 5000 characters.

Confusing order: Please start with the overview of the central platform menu items, explaining which are part of the catalogue, and which are not. After that go into depth of the catalogue functions. If some of the menu items are in future meant for HDAB staff only, please also describe that, to avoid confusion about the menu.

Paragraph 6.3 does not make the clear the distinction between data access application and data request. It also seems that selecting a dataset is mandatory for both, while in EHDS it's not mandatory for Data Request.

Page 27 "compare datasets side by side". Can screenshots and explanation be added?

The quick guide definitely adds value. Preparing the user for complexity of the request is something quite different from guiding the user through the catalogue.

Stakeholders indicate to us that it's important to make a clear distinction in the catalogue between datasets from registries and research datasets versus dynamic data like EHR. As registries and research datasets have been prepared already for research and the request process could run smoothly for them. For dynamic data like EHRs there will always be a need for contact first between data user and data holder before requesting.

Also, to make it realistically possible for data holders to comply: Start easy, e.g. high over metadata, only the most important variables, lighter burden for the data holders. Gradually improve complexity of dataset descriptions, increase granularity and include more variables over time. Describing everything you have in detail is not feasible right at the start. It would help data holders a great deal, if the implementing act for (EHR etc) systems would require them to put a list of possible variables per license model in their technical documentation.

More detailed comments:

Population characteristics and size: Filters related to population characteristics, such as Age range or Population size, help you understand the scale and scope of a dataset. For example, population size can give an early indication of how many records or individuals are included, which can be useful when assessing whether a dataset is suitable for statistical analysis or large-scale studies

Please note that population size is something different than the number of records!

Public: The dataset is openly accessible and can be used without restrictions. Data in this category can typically be accessed directly, without the need for a formal access request. This applies mainly to datasets that do not contain personal or sensitive information. Examples are High-Value Datasets (HVD) according to the EU Open Data Directive.

The document states that a public dataset “can be used without restrictions”. But later licenses and terms for use are discussed. Also for direct downloadable data restrictions can apply, so the first statement is not entirely correct.

Generic remarks about the documents:

We observe that the connections and alignment between the different guidelines are not sufficiently clear. Greater consistency and cross-referencing between the guidelines would improve coherence, reduce the risk of conflicting interpretations, and facilitate more effective implementation. For example, a reference to the DAAMS guidelines to which a request from this one would connect.

We observe that the guidelines have largely been developed and described as separate documents by different authors (which is understandable), but therefore there is limited visibility of how they relate to one another within the broader process.

Greater horizontal alignment across the guidelines is needed to support a clear and coherent data journey. This would help stakeholders understand how the different components fit together, ensure consistency across the framework, and facilitate practical implementation throughout the entire data lifecycle.

The implementability of the guidelines is a concern. Implementation should start with a minimum viable end-to-end data journey and evolve through an iterative learning and growing process. Rather than regulating all details from the outset, requirements should be introduced in phases, allowing experience and evaluation to inform further development and requirements.

At the same time, sufficient harmonisation is needed to prevent divergent national interpretations and implementation approaches. Therefore, each phase should include clear, directly applicable requirements that ensure consistent implementation across Member States growing into a more robust and mature legislation.

Another concern we would like to address is that the TEHDAS guidelines timelines do not always match the comitology implementing acts timelines, which make it hard to assess the effect of the

input on both trajectories and how input on the guidelines will have effect on the implementing acts. Timeline alignment is needed.

The following questions are specific for TEHDAS2 Draft Guideline for data user navigating the catalogue

Framework and Infrastructure Understanding

Context: Section 3 of the User Guide and the Quick Guide describe the federated architecture of the EHDS, explaining how national catalogues interact with the central HealthData@EU platform, the overall user journey and provides screenshots of the platform.

16. On a scale of 1 to 4, how clearly does the User Guide explain the functional distinction between a National Dataset Catalogue and the Central EU Dataset Catalogue?

2

17. Is the explanation regarding the fact that the Central Platform does not store underlying health datasufficiently prominent to manage user expectations?

Yes

18. On a scale of 1 to 4, how helpful is the presentation of the user journey for understanding the user guide and overall orientation?

4

19. On a scale of 1 to 4, how well do the screenshots help you to follow the steps described in the user guide?

4

20. Please provide any suggestions for improving the description of the EHDS infrastructure for non-technical users.

Max. 5000 characters.

P.11 Login confused me: “You will need ... what to expect”. Next screenshot shows basket instead of login.

“Finally ... dataset descriptions.1” Please use dynamic HealthDCAT-AP link the implementing act refers to.

About page 7 the paragraph “Each member ... national level”: Is search and request via HealthData@EU the only way? Please write it here explicitly if the user also has other options according to EHDS (e.g. request via national HDAB catalogue & request service). If not described, the manual implies that this is the data user's only option.

Search Functionality and Precision

Context: Section 5 details the mechanisms for locating data, ranging from basic keyword searches to advanced SPARQL queries for machine-readable metadata.

21. How would you rate the clarity of the instructions for performing a basic keyword search in Section 5.1.1?

4

22. For advanced users, to what extent do the SPARQL query examples (e.g., Figure 5) assist in understanding how to perform property-based retrieval?

4

Interpreting Dataset Descriptions (Metadata)

Context: Section 6 and the Quick Guide explain how to assess "Dataset Cards," including content, provenance, and standards like HealthDCAT-AP.

23. How specific is the guidance in helping you anticipate the procedural burden of a future data application?

2

Basket Functionality and Workflow Transition

Context: Section 7 defines the "Basket" as a preparatory tool for shortlisting datasets prior to formal authentication.

24. Does the guide clearly communicate that adding an item to the "Basket" does not constitute a formal request for data access?

Yes

25. On a scale of 1 to 4, how well does the manual describe the transition from the "pre-authentication workflow" (discovery) to the "post-authentication workflow" (application)?

3

Quick Guide Utility and Visual Aids

Context: The Quick Guide provides a two-page visual overview of the platform's key elements using numbered UI components.

26. To what extent does the Quick Guide serve as an effective "standalone" document for a first-time user's initial orientation?

4

27. Is the visual mapping (the use of numbered icons in Figures 1–3) accurately linked to the corresponding textual descriptions in the User Guide?

Yes

Use Case Scenarios (Annex 10.4)

Context: The Annex provides tailored scenarios for Healthcare Professionals, Policy Makers, Researchers, and Industry users.

28. How well does your specific professional profile (e.g., Academic Researcher vs. Industry) align with the search behaviours described in the Use Case Scenarios?

3

29. Which aspect of the Use Case Scenarios did you find most valuable for your understanding? (e.g., filter selection, citing datasets, or identifying standards).

Max. 5000 characters.

The combination of the scenarios showing different options is valuable in itself. I do wonder if SPARQL will really only be used by programmers or that scientists with SPSS background would also be comfortable writing queries. Especially with the help of a good query editor that helps you to click to assemble the query.

30. Please provide any suggestions for improving the Use Case Scenarios.

Max. 5000 characters.

For the survey: It's 10.3, not 10.4. And is it an option to connect with an API instead of manually entering SPARQL queries? If so, a use case would be helpful. And are citizens allowed to request data? Then a use case could be helpful like: Citizen E would like to conduct investigations about a rare disease in their family, as professional research is not being funded. (...).

Conventions and Terminology

Context: The guide aims for a non-technical tone suitable for CEFR B2-C1 language levels.

31. Do you find the terminology used (e.g., 'metadata', 'provenance', 'distribution') to be consistently applied and sufficiently explained?

Please rate from 1 (Not consistent at all) to 4 (Very consistent).

2

32. Are there any sections where the academic or technical language acts as a barrier to understanding for non-experts?

Max. 5000 characters.

4.1 Nature and function of the catalogues can be a bit confusing. Dataset vs Dataset record is extra confusing because a dataset is something that can contain dataset records, but the catalogue contains metadata records that refer to datasets. So, a dataset is an overarching table, but it's also the subject that a dataset record refers to. Then the text talks about a metadata catalogue, where at all other places it is more correctly called the dataset catalogue. I do see the importance to explain that the catalogue does not contain the actual data itself. But still, it's a confusing text.