



Summary page:

TEHDAS2 public consultation on Draft guideline for data enrichment

This consultation has 5 pages and 25 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 and 5 consist of questions specific to this document.

Demography

1. Country *

Netherlands

2. Type of responder *

Other

Research Infrastructure

3. Are you responding 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? *

2

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.

The document is very repetitive. Especially the starting sections explain multiple times the difference between data enrichment and data linkage. There is a risk that a reader loses track of what is being discussed. The structure and readability of the guideline should be improved.

Due to the non-binding nature of the guideline, it is hard to grasp the concrete recommendations from it. For example, Table 1 states numerous guiding questions, while a reader would have the expectation from a guideline that it provides answers to these questions.

The non-binding nature of the guideline also creates a risk of misalignment of enrichment procedures between member states. This might be especially worrying for multi-country applications.

The examples should be improved and should all be in one place in the document. The differences between examples and use-cases are not clear and this makes it confusing.

Especially regarding standardization/harmonisation: better describe the difference between data holder responsibility and what data user can do. Providing a dataset with clear, unambiguous semantics is the responsibility of a data holder. This should not be up to the data user. If the data user is involved here, the outcomes should be verified by the data holder.

More in general, 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.

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 enrichment

Section A: Background & Experience

11. What perspective describes best your affiliation?

The perspective of an HDAB

12. Does your organisation currently perform data enrichment on health datasets in secure environments?

No

If yes, please describe the types of enrichment performed. Are there established internal procedures or national guidance for documenting and sharing enrichment activities?

Please provide examples or references if available, max. 5000 characters.

No answers

Section B: Key Concepts & Implementation Approaches

13. What approach do you consider most appropriate for handling data enrichment outputs?

Please choose one or more

Sharing enrichment methods, documentation, and code only (without transferring enriched datasets)

Direct communication between data user and data holder (without HDAB intermediation)

Other approach (please specify below)

We believe a data user and data holder(s) should decide together if the enriched dataset is a version of the original dataset or if the outcome is a new dataset. It should be made clear who is responsible for making this dataset findable again (with a link to the original dataset).

14. In your experience or context, what are the main barriers to sharing enrichment outputs?

Please provide feedback, max. 5000 characters.

- Sharing enrichment outputs, particularly in a form reusable by others, requires more effort than enrichment performed solely for a project's own purposes. Documentation must be more thorough, methods more generalisable, and outputs more structured. This additional work is not always budgeted in research projects. This is also acknowledged in the guidelines' section 5,1 reporting that it 'is an additional administrative burden'.
- Preparing enrichment outputs for broader sharing requires data stewardship skills (standardization, documentation, version control) that are not always available.
- Insufficient motivation ('what's in it for me?'). Despite the open science rationale there are no explicit incentives for data users to share enrichment outputs.
- Perhaps intellectual property or responsibility uncertainty. Is an enriched dataset just a version of an existing dataset or a new dataset?

Section C: Roles, Responsibilities & Governance

15. Are the roles and responsibilities of data users, HDABs, and data holders in relation to data enrichment clear and appropriate?

Disagree

If you disagree or strongly disagree, please explain what clarifications or adjustments would be needed.

Max. 5000 characters.

The role of the HDAB is not entirely clear. We believe the HDAB should be monitoring the process; communication between data holder and data user should go via established HDAB communication channels, but HDAB does not have to interfere actively. We believe that the HDAB might not have sufficient knowledge to evaluate the enrichment documentation or code.

Responsibilities of data holders are unclear in case a data user has performed data enrichment with datasets from multiple sources. Who will have to make the updated datasets available? How to decide which data holder should perform this?

16. Is the proposed feedback loop (data user → HDAB → data holder) realistic and workable in your context?

Not realistic/workable

If not realistic or only somewhat realistic, please describe the barriers.

Max. 5000 characters.

The feedback loop places an additional burden on data holders. Data holders are first expected to make their datasets findable, described, and available for secondary use (quite an investment in itself). The feedback loop then asks them to evaluate incoming enrichment utilities. We are not sure if this is workable in practice (especially for smaller data holders). The criteria based on which a dataholder (or HDAB) should evaluate the enrichment (relevance) are unclear but should be elaborated on. When is an enrichment valuable to share with a broader community? It would help to have some examples or best practices for enrichment evaluation.

And it can only be the data holder who will be able to validate the enrichment, but this also places another large burden on the dataholder. It would help to have some examples or best practices for enrichment validation.

A fee can be asked for the effort of making the data available in the SPE, but additional work to evaluate the enrichment, perform it on the dataset based on the received documentation and code and making it findable again in the catalogue are not accounted for.

At the same time, new collaborations may emerge like this. If the enrichment plans are known already when the data user gets access to the data, it would be useful if the data holder could already indicate at that stage that they are interested in the enriched data.

For some enrichment procedures that require a lot of (manual) work, we are doubting that the enrichment will at all be reproducible solely based on the documentation by the data user. It may require many hours of work by the data holder to reproduce the enrichment or might not even be feasible at all (e.g. in case of manual annotation of imaging data).

We doubt the HDAB will have sufficient knowledge (and capacity) to do this evaluation and therefore suggest that the HDAB is not actively involved in evaluating the enrichment.

Section D: Legal & Technical Feasibility

17. Do you see legal constraints (e.g., GDPR, data permits, intellectual property, trade secrets) that would limit the sharing or reuse of enrichment outputs?

Significant legal constraints

If you selected "Significant constraints" or "Major barriers," please elaborate on which legal issues are most problematic.

Max. 5000 characters.

- Data permits define the scope of authorised processing within the SPE. Enrichment activities may not be explicitly anticipated at the time of application, for example if the data user does not have sufficient knowledge to describe which types of data enrichment will be necessary.
- GDPR and re-identification risk, see guideline. Also related: data use agreements
- Intellectual property of the data user that has performed the enrichment (particularly for industry data users?). Can a fee be charged by the data user?

18. Do you consider the proposed implementation model (prioritising documentation/code sharing over dataset transfer) feasible in your context?

Not applicable to my context

If not feasible or only somewhat feasible, please explain why and what would be needed to make it work.

Max. 5000 characters.

No answers

19. How feasible is it for data users to document enrichment activities as suggested in Sections 5.1–5.3 (e.g., methodological descriptions, code sharing, reporting to HDABs)?

Please provide feedback, max. 5000 characters.

The proposed approach might not be applicable to all use-cases. We are missing a process description for multi-dataset applications. Who is responsible for making the new or updated dataset findable in the catalogue?

Section E: Clarity & Completeness by Chapter

20. Is chapter 4 clear and easy to understand?

Please rate from 1 (not clear nor easy to understand) to 4 (very clear and easy to understand).

No answers

21. If there are specific topics missing in chapter 4, please elaborate on these below

Please provide feedback, max. 5000 characters.

The chapter is very repetitive. Structure and readability should be improved heavily.

For example, section 4.1. defines data preparation, enrichment, linkage while they have been discussed earlier.

The difference between data correction and data standardisation is unclear. Furthermore, 4.1.1 and 4.1.5 both discuss data correction. The difference between the two sections is not entirely clear.

22. Is chapter 5 clear and easy to understand?

Please rate from 1 (not clear nor easy to understand) to 4 (very clear and easy to understand)

2

23. If there are specific topics missing in chapter 5, please elaborate on these below

Please provide feedback, max. 5000 characters.

It is unclear how data enrichment and the feedback loop should be handled in case a data-user has performed data enrichment on data that was linked from different sources. To which dataholder does the data user report back the data enrichment code and documentation? And do the dataholders have to communicate about their evaluation of the value of the enrichment with each other (and the HDAB)?

The description of how an updated dataset should be added/described in the catalogue by the data holder is not correct and not according to HealthDCAT-AP. 'isVersionOf' is currently not a property in HealthDCAT-AP. Furthermore, the proposed example is incorrect. The 'isVersionOf' property should point to another dataset (catalogue entry), and should not be the version number of that other dataset.

24. Is chapter 6 clear and easy to understand?

Please rate from 1 (not clear nor easy to understand) to 4 (very clear and easy to understand).

2

25. If there are specific topics missing in chapter 6, please elaborate on these below

Please provide feedback, max. 5000 characters.

- It aggregates a range of loosely related topics (legal, operational, semantic, organisational) without a clear organising logic. Considerations relevant to implementation are also scattered across Chapters 4 and 5, making it difficult to understand what Chapter 6 adds distinctly. A clearer delineation of scope at the outset would help
- Section 6.4 ('Organisational aspects') is narrower than its title suggests. The section focuses almost exclusively on real-world data (RWD) and the practical reasons why enriched datasets should not be returned to data holders. While these are valid points, they do not constitute a comprehensive treatment of organisational aspects. The section title should either be narrowed to reflect its actual content (e.g., 'Why dataset return is generally not feasible') or broadened to include governance structures, capacity requirements, and roles more generally.
- Section 6.5 ('Practical examples') lacks a clear relationship to Section 4.2.3 and 8.6, which also provides illustrative examples of enrichment operations. It is unclear what distinguishes the three sections. The distinction should be made more explicit or perhaps a 'practical examples' chapter can be added in which all practical examples are mentioned together (because practical examples do help for understanding)
- Section 6.3 ('Semantic interoperability') reads more as background information than as a practical consideration. It describes what semantic interoperability is and why it matters in general terms. But what is the consideration? Also, if semantic interoperability is indeed foundational for the EHDS then a critical question arises that the document does not adequately address: to what extent should semantic responsibilities rest with the data user? What is the role of a data holder and HDAB regarding harmonization, standardization, and semantic interoperability (before the data are in a SPE environment) vs what is the role of the data user?