

Data Provenance Maturity Assessment

Please select exactly one answer per question. Select the option that best describes your current situation today, not the situation you are working toward.

Dimension 4 — Data Provenance

Definition*: Clear description of the lineage of the dataset, providing transparent and comprehensive documentation of the data pipeline used to produce it, covering where the data came from, what was done to it, and what that did to its quality.

1. What do you document about where your dataset comes from?*

- | | |
|---|--------|
| ☐ Nothing is documented. | 1 pt. |
| ☐ We document the source of the dataset. | 2 pts. |
| ☐ We document the source plus the transformations, rules, and exclusions applied. | 3 pts. |
| ☐ We list all original data items and all transformations, rules, and exclusions, and document what effect they had. | 4 pts. |
| ☐ We can show earlier versions, including the raw source data, and review the effect of each step. | 5 pts. |

You can support this answer with: your provenance or lineage document for one representative dataset.

2. How complete is that documentation?*

- | | |
|---|--------|
| ☐ Nothing is written down. | 1 pt. |
| ☐ The source system is named. | 2 pts. |
| ☐ Source, plus the main transformation rules. | 3 pts. |
| ☐ Source, all transformations, rules, and exclusions, with record counts before and after. | 4 pts. |
| ☐ As above, maintained automatically by the pipeline itself, so it can never be out of date. | 5 pts. |

You can support this answer with: the lineage document, or the pipeline metadata it is generated from.

3. Can you retrieve and compare earlier versions of the dataset?*

- | | |
|---|--------|
| ☐ No version control. There is only "the current file." | 1 pt. |
| ☐ Old versions are kept somewhere, informally. | 2 pts. |
| ☐ Dataset versions are numbered and dated. | 3 pts. |
| ☐ Versions are numbered and dated, and the differences between them are documented. | 4 pts. |
| ☐ Any earlier version, including the raw source, can be retrieved and compared step by step. | 5 pts. |

You can support this answer with: your version register, repository, or archive of previous releases.

4. Do you document what transformations and exclusions do to the quality of the dataset?*

- | | |
|---|--------|
| ☐ Not at all. | 1 pt. |
| ☐ We know exclusions happen, but not how large they are. | 2 pts. |
| ☐ We record which records were excluded and why. | 3 pts. |
| ☐ We record how many records each rule affected. | 4 pts. |
| ☐ We report the effect on data quality, such as completeness, plausibility, and representativeness, for each step. | 5 pts. |

You can support this answer with: your exclusion log, record counts per pipeline step, data quality report.

5. Can the people who use your data see this provenance information?*

- | | |
|--|--------|
| ☐ It is not shared outside the team. | 1 pt. |
| ☐ Available on request, if somebody asks the right person. | 2 pts. |
| ☐ A provenance document is supplied together with the dataset. | 3 pts. |
| ☐ Provenance is part of the standard machine-readable metadata delivered with the dataset. | 4 pts. |
| ☐ As above, and published in a dataset catalog so users can read it before requesting data. | 5 pts. |

You can support this answer with: your dataset catalog entry or the metadata file shipped with a release.

YOUR NOTES FOR THIS DIMENSION

What sits behind these answers: the practices and tools you rely on, recent improvements, and the gaps that explain the lower scores. This is the material for the supporting narrative on the Overall Score sheet.

DIMENSION SCORE

Add up the points of all 5 answers and divide by 5. Keep one decimal.

YOUR MATURITY LEVEL

Round the dimension score down to the nearest whole number.

Level 1 Initial Starting out · 1 Point	Level 2 Emerging Written down · 2 Points	Level 3 Defined In practice · 3 Points	Level 4 Managed Measured · 4 Points	Level 5 Optimized Improving · 5 Points
---	---	---	--	---

The five maturity levels in this dimension

Your five answers average to one level. Each row describes what that level typically looks like across all five questions, so you may be ahead of it on some points and behind on others.

Level 1 Initial	Nothing documented; no versions; no view of what transformations do; nothing shared.
Level 2 Emerging	The source system is named; old versions kept informally; exclusions known but not quantified.
Level 3 Defined	Source, transformations, and exclusions documented and versioned; supplied with the dataset.
Level 4 Managed	All items, rules, and record counts documented; version differences tracked; machine-readable.
Level 5 Optimized	Lineage maintained by the pipeline itself; any version retrievable; quality effects published.

Your next step: find the row for the level you reached, then read the row directly below it. That row is what you need to achieve to move up one level in this dimension.

Note: Every question in this dimension uses the same 1–5 scale, and the same real-world situation is worth the same number of points in every question, so both the lowest and the highest level are reachable. Transfer your maturity level for this dimension to the *Maturity Overall Score Calculation* sheet, which brings all ten dimensions together.

References*

Question 1 follows the maturity levels of this dimension in the QUANTUM data holders' maturity model (Deliverable 1.2 v3.5); the wording is EHDS4ALL's own. Underlying source given there for this dimension: Adapted HDRUK. Questions 2 to 5 are EHDS4ALL, developed from the WP3 workshop of August 2024 and the expert interviews of October 2024. Question 4 assesses data quality, identified as a necessary part of the model in the EHDS4ALL workshop of 19 August 2024. The five level statements below combine all five questions, because the dimension score is the average of all five answers.

Jones, M., & Bernal-Delgado, E. (2025). **Deliverable 1.2 — Specification for the assessment of data holders' maturity** (Version 3.5.0). QUANTUM project. Zenodo. doi: 10.5281/zenodo.14944767

Gordon, B., Barrett, J., Fennessy, C., et al. (2021). Development of a data utility framework to support effective health data curation. *BMJ Health & Care Informatics*, 28(1), e100303.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018.

EHDS4ALL (2024). Protocol, Workshop Kick-off WP3, Cologne, 19 August 2024. Work Package 3, dimension validation and maturity levels.

EHDS4ALL (2024). Expert interviews with health data holders and maturity model developers, October 2024. Conducted by inav and the University of Cologne; participants anonymized.