Annex 4a: Metadata and metadata catalogues – full report

Condensed summary in Final consolidated WS2 report above

Introduction to the metadata work in VALO2

The task *Metadata and metadata catalogues in relation to EHDS* was part of WS2 in VALO2 project. The aim of the task was to continue the work done in the first VALO project on collaboration around health data catalogues and implementation of EHDS2 metadata requirements in the Nordics (and possibly also Baltics).

The objectives of the metadata work of VALO2 were to develop:

- common understanding of the HealthDCAT-AP to support its implementation in the national catalogues;
- common understanding on Nordic health data landscape (EHDS Article 51), including how to present the data as datasets for secondary use, and
- preliminary understanding on the requirements efficient use of AI sets for metadata.

The metadata work in the VALO project builds on already longer history of Nordic collaboration around health data catalogues in secondary use. The VALO projects were preceded by two Nordic Commons projects funded by the Nordic Council of Ministers, both of which had a task force on metadata issues. There is also ongoing informal collaboration bridging over from the Nordic Commons to the VALO projects within the framework of the Nordic Health Metadata Network.

The metadata task was advanced through Competence Forum Metadata breakout sessions, engaging with the informal Nordic Health Metadata Network involving monthly meetings, and organisation of some additional meetings to discuss specific tasks.

Previous work and key outcomes

The metadata work during the first VALO project (VALO1) focused mainly on forming a shared understanding of the HealthDCAT-AP metadata specification draft and EHDS2 metadata requirements. The task also actively sought to influence the development of HealthDCAT-AP through the TEHDAS2 HealthDCAT-AP development work, which was underway at the same time. In addition to specific EHDS-related work, a common Nordic variable-level metadata specification was further developed up to version 0.8 and it was also submitted as input from the VALO1 project to the HealthDCAT-AP development work being carried out within the TEHDAS2 project.

The identified key challenges of HealthDCAT-AP draft from the Nordic perspective included:

- lack of shared understanding on the information model,
- lack of hierarchy in the information model,
- ambiguity about how datasets should be defined,
- some technical complexity issues, and
- terminology inconsistencies.

For upcoming work, it was proposed to:

- continue building a shared Nordic understanding of HealthDCAT-AP,
- seek the most harmonised way possible to implement the HealthDCAT-AP in the national health data catalogues.

The proposed solutions included:

- creating possible enrichments of HealthDCAT-AP in the national catalogues in a harmonised way, including common Nordic variable metadata specification,
- implementing data quality and utility labelling in a harmonised way, and
- aiming for harmonised dataset definitions.

For more information, see the previous VALO report on metadata catalogues[1].

The metadata work in the VALO2 project built on where the first VALO project ended. All identified objectives were still relevant, as well as most of the identified HealthDCAT-AP related challenges, even though the HealthDCAT-AP draft had evolved since the first VALO. The VALO1 work was based on the HealthData@EU Pilot project's and TEHDAS2 versions of HealthDCAT-AP draft; at the time of writing, HealthDCAT-AP Release 7 published by EU's DG SANTE is in use.

Outcomes of the metadata work

Common understanding of HealthDCAT-AP

The primary objective of the metadata task was to enhance shared understanding of the HealthDCAT-AP specification, with the overall goal to achieve the highest possible level of harmonisation in its implementation in the Nordics.

This objective was promoted mainly through:

- mappings between already existing or emerging Nordic national health data catalogues and the mandatory dataset-level properties of HealthDCAT-AP (using HealthDCAT-AP Release 7 and Implementing Act on Article 77[2] draft's mandatory properties);
- pilot descriptions using the mandatory HealthDCAT-AP properties;
- discussions on which non-mandatory HealthDCAT-AP properties the countries would want to encourage their data holders to describe;
- testing the HealthDCAT-AP Release 5 and the Dataset Description Assistant provided in the HealthData@EU Central Platform and sending shared feedback with questions to DG SANTE;
- collection of information on the status of HealthDCAT-AP understanding and implementation in the countries;
- sharing experiences with the EU project Cancer Watch Joint Action metadata work;
- by showcasing the development made in the catalogues.

In addition, shared feedback on the EHDS Article 77 Implementing Act on dataset descriptions was compiled and submitted on behalf of the VALO2 project in the implementing act public consultation (https://ec.europa.eu/info/law/better-regulation/have-

[1] VALO Metadata catalogues – final report. 2025. https://www.sitra.fi/en/publication/valo-metadata-catalogues-final-report/
[2] EHDS Article 77: *Dataset description and dataset catalogue*

your-say/initiatives/15673-European-Health-Data-Space-dataset-descriptions/F33400889_en).

Summary of findings

The HealthDCAT-AP specification has evolved in a positive direction to some extent, and understanding of it has also increased in the participating countries during the VALO2 project. However, the specification is still considered mature enough only in certain respects to support planning for its national implementation, and the understanding of HealthDCAT-AP across VALO2 participant organisations is perceived as uneven. Most of the challenges identified during the VALO1 project (introduced under section 4.2) remain valid and still require further clarification, despite the continued evolution of the HealthDCAT-AP specification and the important revisions that have been made.

The identified key needs to support the national implementation of HealthDCAT-AP include:

- **Clear guidance with examples on the information model**: What is a dataset? What is a catalogue in the information model? And how should dataset series or distributions be used?
- **Clear guidance on datasets in different data categories:** The key concept dataset should be defined more precisely, and concrete examples should be given of what a single dataset could be for different EHDS Article 51[3] data categories. There are currently very different views on the matter in the community. Inconsistent implementation can jeopardise the ability to gain a harmonious understanding of the available data in different countries.
- **Development of property definitions and vocabularies, taking into account different data categories**: HealthDCAT-AP is still in the development stage, and in particular specific mandatory properties' definitions and vocabularies should be developed to enable both the correct understanding and also the possibilities to comprehensively and with sufficient accuracy describe datasets from all different concerned data categories.
- **Detailed instructions taking into account different data categories**: More detailed EHDS Art. 51 data category specific guidelines are needed for dataset descriptions in order to make them harmonious and ensure the findability of similar data from different countries.
- **Possibility to publish pilot descriptions**: In addition that the HealthDCAT-AP pilot descriptions can be made using the Commission's Dataset Description Assistant, a possibility to also publish the descriptions in the HealthData@EU Central Platform's EU Dataset catalogue more easily is needed. Currently, the descriptions can only be published via national infrastructure, and not all countries have started setting it up yet. Easily accessible pilot descriptions would support the development of common understanding and allow feedback from different stakeholder groups. Discussions and feedback are needed both at the EU and Nordic level and within individual countries, considering the perspectives of various data holders, communities and different types of data categories, as well as different data user groups and HDAB personnel who process data permit applications – making the descriptions visible and accessible is a prerequisite for progress.

[3] Article 51: *Minimum categories of electronic health data for secondary use*

- **Concrete training, examples and exercises in the EU level:** Also, more concrete examples and exercises in the EU level are needed and clarification on how different user needs (e.g. statistic, research, or combined datasets) should be supported.
- **Additional resources**: The need for additional resources and especially technical expertise support was identified as key factors in facilitating implementation in the countries – currently more in the HDAB side, but also in the health data holder side. In general, manual work on metadata is time consuming but automation requires resources and technical investment.
- **Possible Commission support regarding technical solutions**: Also, the possibility to use the Dataset Description Assistant[4] provided by the European Commission as part of national technical metadata solutions in the future was identified as a possible implementation path.

The mandatory dataset level properties' mapping exercise (see "Annex 4b The HealthDCAT-AP mapping tables to the national health data catalogues and pilot descriptions") demonstrated that countries primarily aiming to implement the specification into an existing catalogue, Finland and Norway, will need to add several new properties and also adapt some existing ones in order to be HealthDCAT-AP compatible. In general, it seems that building a new catalogue based on HealthDCAT-AP might be easier than adapting an old catalogue to it.

Countries also share the view that the mandatory HealthDCAT-AP properties are not enough to describe Nordic health data consistently and support cross-border discovery. In addition to the HealthDCAT-AP properties, all countries also intend to implement other properties – either by retaining existing ones or by adding others deemed necessary. It is seen important to continue cooperation and strive to implement additional properties as harmoniously as possible, possibly with the aim of "Nordic HealthDCAT-AP extension". It should include both other than mandatory HealthDCAT-AP properties the countries consider necessary to be described, and additional properties outside the HealthDCAT-AP specification that are needed to be able to describe Nordic health data in a meaningful way.

Doing the pilot descriptions (see Annex 4b "The HealthDCAT-AP mapping tables to the national health data catalogues and pilot descriptions") and discussion with the VALO Legal group demonstrated that the HealthDCAT-AP specification has evolved in a better direction, but it still needs further development. To highlight a few examples among others, the following development needs were identified during the work:

- **Development of vocabularies**: To highlight a few examples, the current version of certain vocabularies (e.g. *Coding systems*), currently have more suitable options for Cancer registry descriptions than biobank data descriptions. The vocabulary in *Has personal data* lacks many key concepts for describing personal data related to health data.
- **Development of instructions**: For example, instructions on how numerical data in properties like *Number of records* and *Number of unique individuals* will be presented for continuously updating datasets, is needed.
- **Addition of new properties**: Also, some missing key properties were identified in the specification, like for example:

[4] https://acceptance.data.health.europa.eu/healthdata-central-platform/datasetDescriptionAssistant?locale=en

- **The unit of observation**: The mandatory property *Number of unique individuals* would need to be accompanied by the observation unit to make more sense for the user.
- **IP rights and trade secrets**: In discussion with the VALO Legal group it was noted that a dedicated property for IP rights would be needed. Data holders must communicate to HDABs if the data includes IP rights or trade secrets, and it would be necessary to have a structured and harmonised way to describe it in metadata.

Differences between countries were observed in the interpretation of properties, as the guidance is not yet sufficient. More pilots with different datasets are needed to be able to highlight more specific needs.

The implementation status of HealthDCAT-AP is briefly outlined in the table below. More detailed information on the countries' national health data catalogues can be found in the Final metadata catalogues report from the earlier VALO project[5].

Table III. Status of HealthDCAT-AP implementation in the Nordic national health data catalogues.

Country	National health data catalogue exists	Status	Key challenges in HealthDCAT-AP implementation
Denmark	No	• Minimum Viable Product (MVP) for a national health data catalogue based on HealthDCAT-AP Release 6 has been developed in EU funded HDAB Denmark project. • The catalogue was soft-launched in May 2026, with the dataset descriptions visible only internally. It currently contains descriptions on key population-based health registries in Danish language.	
Finland	Yes	• Goal to implement HealthDCAT-AP in the existing national catalogue Data resources catalogue. • The catalogue has been preliminarily mapped to HealthDCAT-AP in the dataset and variable description level. Quite many HealthDCAT-AP properties	• Understanding of the HealthDCAT-AP information model. • Adapting the old catalogue's more hierarchical information model (data resource and dataset descriptions vs. dataset

[5] https://www.sitra.fi/en/publication/valo-metadata-catalogues-final-report/

		are missing and need to be implemented. • Plan for the HealthDCAT-AP implementation will be drawn up during the autumn 2027. Catalogue usability and search functions were developed in the EU-funded direct grant FinHITS project during the year 2025.	descriptions) to the new one. • An additional resource with technical expertise would be needed to assist with planning. • There is funding for the planning, but not yet for the implementation.
Iceland	No	• Applied for a direct grant to start planning a national HealthDCAT-AP based catalogue. • Descriptions of key health registers including variable description in Excel files are already published at webpage. Most of the registries are owned by Directorate of Health.	• Understanding of the HealthDCAT-AP information model. • Resources.
Norway	Yes	• The national metadata standard mapped to HealthDCAT-AP. • The national metadata catalogue is currently being developed within MUNIN, a metadata management system that supports both the national metadata specification and HealthDCAT-AP. The platform enables efficient metadata governance and facilitates compliance with emerging European interoperability requirements. • To support EHDS compliance, a HealthDCAT-AP validator has been developed, providing feedback on the quality and completeness of metadata imported from health data registries and helping data providers identify and address compliance gaps. • HealthDCAT-AP will be implemented in the national health data catalogue, helsedata.no, which serves as Norway's public portal for health data metadata	Norway has established the foundations for HealthDCAT-AP implementation through initiatives such as MUNIN, the Dataset Assistant, and the national metadata catalogue. However, the key challenges are: • **Limited adoption across the health sector**: HealthDCAT-AP is still a relatively new standard, and implementation among health data providers remains limited. • **Metadata harmonisation**: Health registries and other data holders use different metadata models, terminologies, and code lists, making it difficult to achieve interoperability and consistency. • **Metadata quality and completeness**: Existing metadata are often incomplete, inconsistent,

		managed in MUNIN. In addition, new functionalities, including the Data Source Explorer (Kildeutforsker - Source explorer - Munin) and Variable Explorer-Variable Explorer - Munin, are currently under development and testing, with launch planned for October. • The primary motivation for developing these new functionalities has been to improve metadata discoverability by enhancing search capabilities, introducing more advanced filtering mechanisms, and enriching metadata records with HealthDCAT-AP properties.	or outdated, reducing their value for data discovery and reuse. • **Unclear implementation requirements**: Some uncertainty remains regarding the required level of metadata granularity and the interpretation of certain HealthDCAT-AP requirements. • **Awareness and competence among data holders**: Many organisations need additional guidance and support to understand their responsibilities under EHDS and to implement HealthDCAT-AP correctly. • **Resource constraints**: Creating, maintaining, and continuously improving metadata requires time, expertise, and dedicated resources. • **Variable-level metadata**: Researchers increasingly expect detailed information at variable level, but such metadata are often not standardised or easily available. • **Long-term governance and maintenance**: Sustainable metadata management processes must be established to ensure that metadata remain accurate, up to date, and compliant over time.
Sweden	Yes	• Sweden has a national metadata catalogue for registry data. Rather than altering said catalogue to meet EHDS criteria, a new catalogue will be assembled.	• Awareness, resources and capabilities among data holders to provide metadata for the catalogue.

		• A prototype catalogue (POC), which is partially mapped to HealthDCAT-AP has been developed as part of EU-funded SENASH direct grant project. The POC is not built for launch but has gathered valuable information for creation of a production catalogue. • Sweden's HDAB will now build on all lessons learned in the direct grant-funded POC catalogue when assembling a national health data catalogue.	• The time restraints for development: HealthDCAT-AP is not yet stable, and time is ticking to implement it.

Recommendations

- **Making rationale visible and fostering communication:** The next steps should be to move to concrete piloting by publishing pilot descriptions using datasets covering a wide range of different data categories and thereby enabling making the rationale visible to different user groups and asking for feedback. The findings and possible suggestions should be presented and brought for further discussions in the EU level in a consolidated manner, for example in the Community of Practice (CoP) subgroup 2 on dataset catalogue (SG2), to demonstrate open questions in a practical way and to foster the formulation of common guidelines.
- **Explore the need to continue forming a "Nordic HealthDCAT-AP extension":** To ensure that Nordic health metadata is as harmonised as possible, a shared metadata specification on additional metadata properties is needed. It should include both other than mandatory HealthDCAT-AP properties the countries consider necessary to be described, and additional properties outside the HealthDCAT-AP specification that are needed to be able to describe Nordic health data in a meaningful way. The group has started the work towards this goal by starting to discuss which other than mandatory HealthDCAT-AP properties would be needed to be filled in in the descriptions and earlier during the previous projects developing the draft of common Nordic variable metadata specification (see as annex in the VALO1 metadata final report[6]). Confirmation from all countries as to whether this goal is still supported is needed.

Common understanding of the Nordic health data landscape

EHDS expands the field of data categories under secondary use in relation to the current data categories familiar to many in the Metadata group. Many countries already have a national catalogue or sector specific catalogues of health data, but not all data categories (minimum categories of electronic health data for secondary use) of EHDS Art. 51 are included in them. Also, many in the Metadata group are not familiar with what these data categories are in practice, and which health data holders are found in the countries that fall

[6] https://www.sitra.fi/en/publication/valo-metadata-catalogues-final-report/

within the scope of the legislation. Those countries that are already engaging with the future EHDS2 health data holders, like Norway, have reported that the data holders which have not yet described their data in existing catalogues, are often unfamiliar with the EHDS Regulation and may lack understanding of their role and responsibilities as health data holders under the Regulation. Additionally, the suitability of HealthDCAT-AP to describe data from all data categories has not been tested either.

Work on building an understanding of the health data landscape and of different data categories according to EHDS Art. 51 was started with a light touch. The metadata group conducted a practical exercise to identify examples of health data holders and datasets belonging to various data categories from each country (see Annex 4c "Nordic health data landscape analysis tables"). The work also included considering in what kind of datasets it would be desirable to offer these data in the catalogues for users. A more detailed reasoning on why to present data in such datasets was gathered and discussed with examples from a couple of different data categories. Also, a joint workshop with the VALO2 Legal group was organised to discuss open questions. The Legal group had discussed preparing a self-assessment template for potential data holders, to help them assess whether they are health data holders according to EHDS, based on work done in Sweden. A shared tool of this kind would be warmly welcomed by the Metadata group.

Disclaimer: This mapping is illustrative only and does not constitute a determination that any listed organisation is, or is not, a health data holder under Regulation (EU) 2025/327. Each organisation remains responsible for assessing its status and compliance obligations for each dataset and processing context, including whether it acts as a controller or joint controller for personal electronic health data under Articles 2(1)(a) and 2(2)(t), whether it controls the availability of non-personal electronic health data, and the controllership rules in Article 74. Natural persons and microenterprises are exempt from the Chapter IV health data holder obligations under Article 50(1), unless a Member State extends those obligations pursuant to Article 50(2). Final national legislation, implementing acts, contractual arrangements and actual processing roles may affect the assessment.

Summary of the findings

To summarise the findings and remaining open questions, it can be stated that:

- **Clarification of the health data holder definition and a self-assessment tool needed:** The health data holder definition in the EHDS Regulation is perceived as too vague. Furthermore, many Nordic countries have not yet initiated a comprehensive discussion or established a legal interpretation regarding which data holders fall within the scope of the definition. A more specific definition within the Nordics could be useful and a Nordic self-assessment framework for data holders is needed.
- **No clear understanding yet on the Nordic EHDS data landscape:** Based on a brief exercise to identify data resources belonging to different Art. 51 categories, the group found it somewhat challenging to identify examples for many of the less familiar data categories. For example, there is no clear overview yet on the research datasets generated by hospitals and universities and many collaborative research datasets, or data from apps and wearables, or health data produced by private companies.
- **Clarification needed for Art. 51 definitions:** Many of the definitions of the minimum data categories set out in Art. 51 of the EHDS Regulation require further

clarification. In some cases, it is currently unclear which data are actually covered by these categories. For example, it remains unclear which data held by healthcare providers are included in the EHR data category and which are not.

- **More guidance on metadata implications needed:** There is currently no general guideline on how the HealthDCAT-AP health category property should be used. In particular, it is unclear whether it should reflect the original primary data source (e.g., EHR data in case of national registry data) or only the current data category of the dataset. This issue was discussed with the VALO2 Legal group, which concluded that it is a practical dilemma, rather than a legal matter.

The exercise to present the current views on how to divide national cancer registries into datasets demonstrated varied views among Nordic countries. It also demonstrated that the contents of the cancer registries are not similar, or the data are organised differently under different registries in the countries. For example, the National cancer registry in Norway includes quality registries, which in some other countries might be as a separate registry. The different views on how to present data as datasets are guided by, among other things:

- views on what the goal and core ambition of the catalogue should be (a starting point to provide a general overview – "a high level directory", or a place to get detailed description on specific data),
- already implemented ways of describing data in the catalogues,
- views on how feasible it is to describe the data (e.g. the description of a complex dataset as one single dataset most probably might compromise the accuracy of description),
- views on how feasible it would be to report on the quality of data from the defined entities (data quality and utility label),
- the known data holders' different views on how detailed metadata they would like to share,
- the level of funding and resources for the metadata work, and
- views on already known user needs.

A good example of the complexity of the subject – and of the specificity regarding data categories and intended use – is that the Cancer Watch JA shared with the group their conclusion on the most optimal dataset division of national cancer registries. Their conclusion was that frozen versions of the entire registry should be described as datasets.

It is clear that the VALO2 project was only able to scratch the surface of the EHDS health data landscape analysis in the Nordics, as well as other relevant metadata issues, and it is important to continue the work.

Recommendations

- **Publishing health data holder self-assessment tool**: A shared Nordic tool should be developed and published to help both the data holders and HDABs in identifying health data holders.
- **Data holder and dataset mapping:** Continuing the mapping of data holders and datasets and sharing the methodology and results among the countries.
- **Forming more detailed descriptions on data categories:** Once the category descriptions are clarified at the EU level, publishing the descriptions including examples from Nordic datasets in connection with the health data holder self-assessment tool could facilitate the identification of data holders.

- **Publishing HealthDCAT-AP metadata and utilising user feedback:** Continue experimenting with data category selection in the published metadata descriptions and collecting user feedback on the consistent discoverability of specific data categories in the catalogues

Preliminary understanding of needs AI places on metadata

Expectations towards AI have grown greatly in recent years. The potential applications of AI are expanding rapidly, and it has been recognised that it is important to identify and learn about the connections between metadata and effective utilisation of AI applications in connection to health data secondary use.

The topic was approached from two different angles: By examining the AI Act's requirements regarding data quality when training high-risk AI models, and by learning from practical findings regarding the experiences of a Norwegian project exploring metadata-driven AI for variable discovery.https://folkehelse.sharepoint.com/:p:/r/sites/sp000002719/Shared Documents/AI in metadata_20260820_VALO2.pptx?d=wa51c2c2fdfee41aeb39765d09d6f682b&csf=1&web=1&e=uTF45f

The AI Act Art. 10 states that the data used to train high-risk AI models must meet certain quality criteria, including criteria related to data governance processes. AI training, validation and testing datasets must be subject to data governance and management practices appropriate for the intended purpose of the high-risk AI system. Those practices concern in particular relevant design choices, data collection processes and the original purpose of data collection, relevant data preparation processing operations, formulation of assumptions, assessment of the availability, quantity and suitability of the datasets that are needed, examination in view of possible biases, appropriate measures to detect possible biases, and identification of relevant data gaps. Training, validation and testing datasets shall be relevant, sufficiently representative, and to the best extent possible, free of errors and complete in view of the intended purpose. (https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng)

A preliminary analysis was made on how well the potential users could assess and gain an understanding on the aspects mentioned in the quality criteria by descriptions following the current EHDS metadata and data quality and utility label documentation requirements - HealthDCAT-AP Release 7 metadata specification (EHDS Art. 77) and the data quality and utility label metrics developed by the QUANTUM project[7,8] (EHDS Art. 78). When discussing the outputs of the QUANTUM project, it is worth noting that they will serve as the basis for the implementing act on data quality and utility label, but there is no guarantee that the final EU requirements will align with them in the end.

The objective was promoted by:

- Seeking to identify criteria among those listed in AI Act Article 10 that could potentially be met through metadata descriptions or data quality reporting.

[7] Deliverable 1.1 Specification of the data sets' quality and utility label. 2025

https://zenodo.org/records/14937423

[8] https://quantumproject.eu/

- Conducting a preliminary mapping of the equivalent HealthDCAT-AP metadata properties and the QUANTUM data quality and utility labelling dimensions and metrics and analysing the correspondences.

In addition to this, the Norwegian team shared their findings of a project on exploring metadata-driven AI for variable discovery, application support and case processing in the Norwegian Institute of Public Health (FHI) called AI-Powered Metadata Discovery for Health Data Applications and Case Processing, and issues related specifically to metadata quality were discussed. The project was initiated in August 2024 through a pilot with Microsoft Copilot and evolved during 2025-2026 into a purpose-built AI agent architecture and MCP (Model Context Protocol)-based service for structured metadata search and variable discovery. The solution aims to help researchers, applicants and case handlers identify relevant datasets and variables from FHI's health registry metadata, with the objective of reducing the time from data request to data delivery while simultaneously identifying areas where metadata quality can be improved.

Summary of findings

The role of metadata in ensuring the effective and reliable use of AI for the secondary use of health data appears to be absolutely central. This very preliminary analysis identified metadata – both metadata describing data quality and aiding in assessing the suitability and quality of the data – as well as the quality and coverage of the metadata itself, as key factors.

Regarding the requirements introduced in the AI Act, it appears that:

- Some of the quality-related information needed to assess whether a dataset meets the criteria set out in Art. 10 could potentially be provided through HealthDCAT-AP-compliant dataset descriptions and data quality and utility labelling, including data quality reporting based on the requirements proposed by the QUANTUM project.
- At the very least, the metadata and quality label could provide the AI developers with a starting point for assessing whether the data are worth further consideration and analysis, especially if detailed and up-to-date variable-level data quality reporting is available. For example:
    - Higher level information on the data collection processes, the origin of data and the original purpose of data collection can be provided by the mandatory HealthDCAT-AP property *Provenance* and non-mandatory *Purpose*. However, the documentation in the HealthDCAT-AP metadata is in free text fields and likely not in a very detailed level, and more detailed information might be needed for the high-risk AI model development. In QUANTUM's data quality and utility label in turn, the data holder needs to indicate whether dataset source and processes and operations are documented or not.
    - For assessment of the availability, quantity and suitability of the datasets that are needed, it seems that accurate and detailed metadata based on HealthDCAT-AP and the related QUANTUM data quality label's metrics facilitates this assessment. However, in the case of health data for example from national registries or EHRs, reporting on the availability of specific diagnoses in the data could support this assessment.
    - For the examination of potential biases, it might be that deeper analysis is needed. Information on all results of such analyses should be made visible in the metadata for subsequent users to see.

Therefore, it seems that the requirements in the AI Act go deeper than what can be fulfilled by the current HealthDCAT-AP or QUANTUM requirements, and more in-depth information and analysis of data quality are needed to be able to state whether the data used meets the quality criteria. However, well documented metadata and data quality and utility label and reporting provide the necessary starting point for the AI model developers to start digging deeper to evaluate the data quality.

The Norwegian project demonstrated that the effectiveness of AI-assisted discovery depends heavily on the completeness, consistency and standardisation of metadata.

The project found out in practice that metadata quality is a key factor in ensuring that generative AI can reliably suggest which variables could be used in a study from among the datasets described in the catalogue, based on the study description. The metadata needs to be well structured, standardised, all necessary information completed and regularly updated. A common problem is that the metadata has not been described with sufficient comprehensiveness or updated regularly. Harmonising variable descriptions that cover the similar content is also very important: if the descriptions are not created consistently, AI may not necessarily recognise the similarity and be able to suggest the needed variables.

Recommendations

- **Aim for detailed variable metadata:** Detailed and structured variable-level metadata, along with clear information on classifications and changes to them, is important. The variable-level metadata in the HealthDCAT-AP draft is not yet this detailed and the attention so far in the development has been more in the dataset level metadata; it is important for the Nordic countries to develop their own solution and share their lessons learned.
- **Support for data holders**: To ensure sufficient metadata quality, Nordic countries are encouraged to develop metadata description support for data holders that is as resource efficient as possible while remaining comprehensive. Particular attention should be paid to supporting the description of datasets that are most frequently utilised for secondary use and those otherwise assessed as key datasets in detailed variable level. Especially if the goal is metadata harmonisation, a metadata expert familiar with the metadata of various organisations on a broader scale can offer assistance, as opposed to a deep-level data expert who is skilled at describing details of specific data precisely. It is advisable to share best practices and lessons learned among the Nordic countries, as well as any support approaches specific to particular data categories and use cases. It is important for countries to fund metadata work sufficiently to create the conditions for providing support.
- **Harmonised implementation of EHDS data quality and utility label and data quality reporting:** It is worth investing in the consistent implementation of the EHDS data quality and utility label across the Nordic countries. Furthermore, it is important to share experiences and strive to find common methods for reporting data quality.
- **Collecting quality related user feedback:** It is recommended to explore enabling a direct customer feedback channel on the metadata and data quality from data users to data holders and HDABs, to enable the continuous development of metadata especially with regards to quality considerations. In this context, it is also important to share experiences and learn from the experiences of other Nordic countries. The possibility of publishing user feedback on data quality and utility alongside the dataset description should be explored, and potential European developments – such

as the possible further development of the fit-for-purpose strategy developed in the QUANTUM project – should be monitored.

The importance of Nordic cooperation

Nordic cooperation around metadata for secondary use of health data has been considered important among the national actors. Now, under the EHDS Regulation and the significantly increased requirements and expectations regarding metadata and data quality and utility reporting, such cooperation has become even more important.

The following are some of the identified achievements and future possibilities that would not be possible to achieve without collaboration:

- HealthDCAT-AP, including data quality and utility label, is not easy to understand and interpret. It would not be wise use of resources for the countries to think about implementation solutions solely on their own. The metadata work presented here demonstrates concrete examples of the value of Nordic collaboration, in this case the joint interpretation and implementation of HealthDCAT-AP. In an increasingly harmonised way
- Many of the Nordic countries have extensive practical experience with health data catalogues. It is important to leverage this experience to have a constructive influence on the solutions being developed at the EU level. Working together on such efforts is more effective than countries acting individually. This influencing work has been carried out during VALO project, among other things, through discussions on the implementing acts on dataset descriptions and data quality and utility labelling, the exchange of national views, and the submission of joint feedback on the implementing act on dataset descriptions and the previous version of HealthDCAT-AP.
- Many countries have limited resources for metadata work, and under the EHDS the requirements for metadata catalogues increase significantly. National health data catalogues will need to contain new information elements, include data quality and utility labels, and describe a wider range of data categories. In addition, new types of data holders will need to be involved in the work related to metadata. It is valuable to discuss open questions and implementation plans together.
- The first Nordic Commons project established a vision of a common Nordic metadata catalogue. Consistent and harmonised implementation across countries would enhance the feasibility of realising this vision in the future, provided it continues to be considered beneficial also in the EHDS era.

The VALO project and this report account for several examples of initiated Nordic harmonisation such as developing of a common variable-level metadata specification draft version, aiming for creating possible enrichments of HealthDCAT-AP in the national catalogues in a harmonised way and implementing data quality and utility labelling in a harmonised way, as well as aiming for harmonised dataset definitions.

Metadata implications for the User Journey

As described in this chapter, metadata – created and used in a standardised manner - is a key prerequisite for making datasets searchable and findable in the User Journey (the Data discovery step, see chapters 3 and 7). Accordingly, there are several steps prior to that (during Pre-discovery) where metadata can make a significant change: The health data

holders must first understand their role as a health data holders in EHDS and be able to identify what data subject to EHDS Regulation they hold. Health data holders must provide adequate metadata for their datasets, which in turn must be clearly identified, scrutinised and up to date. The metadata must be of sufficiently high quality. Data holder must also provide the data quality and utility label for all publicly funded datasets.

In addition, the catalogue search features must be adequate, and guidance – as well as support for using the catalogues, if needed – must be available. These matters were not addressed in detail in this report, but sharing good practices related to them within the Nordic countries is important.