Deliverable 6.3

Recommendations on a Data Quality Framework for the European Health Data Space for secondary use

26 September 2023

This project has been co-funded by the European Union's 3rd Health Programme (2014-2020) under Grant Agreement no 101035467.

1 Document info

1.1 Authors

Author	Partner
Enrique Bernal-Delgado	Institute for Health Sciences in Aragon, Spain
Francisco Estupiñán-Romero	Institute for Health Sciences in Aragon, Spain
Natalia Martínez-Lizaga	Institute for Health Sciences in Aragon, Spain
Persephone Doupi	Finnish Institute for Health and Welfare (THL), Finland
Mari Mäkinen	Finnish Institute for Health and Welfare (THL), Finland
Malou Mulkholm	Central Denmark Region, Denmark
Annelise Nyvold Lundbye	Central Denmark Region, Denmark
Juan Gonzalez-García	Institute for Health Sciences in Aragon, Spain
Carlos Telleria-Orriols	Institute for Health Sciences in Aragon, Spain

1.2 Contributors

Contributor	Institution
Hans Aage Huru	Norway Helsenett (Norway)
Roxana Arzideh	Federal Institute for Drugs and Medical Devices (Germany)
Antal Bodi	EMK (Hungary)
Petronille Bogaert	Sciensano (Belgium)
Kristina Bränd Persson	NBHW (Sweden)
Lorien Brenda	French Data Hub (France)
Fidelia Cascini	Unicatt (Italy)
Pablo Chaves de Luis	MoH (Spain)
Shona Cosgrove	Sciensano (Belgium)
Sarah Craig	HRB (Ireland)
Pascal Derycke	Sciensano (Belgium)
Sérgio Dinis	Shared Services of the Ministry of Health (Portugal)
Lara Dirian	French Data Hub (France)

Jean-Charles Dufour	Aix-Marseille Université (France)
Kerstin Engelhardt	eHelse (Norway)
Barbara Foley	Health Information and Quality Authority (Ireland)
Luis Frauca	MoH (Spain)
Roch Giorgi	Aix-Marseille University (France)
Zdenek Gütter	Ministry of Health (Czech Republic)
Tala Haddad	INSERM (France)
Magnus Häll	Statistics Sweden
Anne Heidi Skogholt	Directorate of e-health (Norway)
Johanna HOLM	The National Board of Health and Welfare (Sweden)
Josef Horak	Masaryk University (Czech Republic)
Henrik Jensen	Sundhedsdata (Denmark)
Beatrice Kluge	Gematik (Germany)
Toth Kornel	SU-EMK (OKFŐ)
Truls Korsgaard	Directorate of e-health (Norway)
Vanessa Lima	Shared Services of the Ministry of Health (Portugal)
Vanessa Mendes	Shared Services of the Ministry of Health (Portugal)
Catia Pinto	SPMS (Portugal)
Lenka Radova	Masaryk University (Czech Republic)
Gregoire Rey	Inserm (France)
Anton Rivall Andersen	Sundhedsdata (Denmark)
Miia Ryhanen-Tompuri	THL (Finland)
Nina Sahkertz Kristiansen	AUH (Denmark)
Philipp Schardax	Ministry of Health ATNA (Austria)
Nienke Schutte	Sciensano (Belgium)
Hana Svozilova	Masaryk University (Czech Republic)
Marije van Melle	Nictiz (Netherlands)
Jan Vorisek	FNUSA-ICRC (Czech Republic)
Lukas Wrosch	Gematik (Germany)
Dickjan Zijda	NICTIZ (Netherlands)

Alexander Zlotnik	MoH (Spain)

1.3 Keywords

Keywords	Data Quality Framework, TEHDAS, Joint Action, Health Data, Health Data Space, Data Space, HP-JA-2020-1

Accepted in Project Steering Group on 27 June 2023. The European Commission gives final approval to all joint action's deliverables.

Disclaimer

Copyright Notice

Contents

1 Document info 1
1 Executive summary 5
2 Glossary 6
3 Acronyms 12
4 TEHDAS Data Quality Framework 15
4.1 Definition of data quality and utility 15
4.2 Pragmatic approach: Data life cycle in HealthData@EU 16
5 Key elements in the TEHDAS DQF 17
5.1 Data quality management and quality assurance procedures 17
5.2 Semantic interoperability 18
5.3 Datasets publication and cataloguing 19
5.3.1 Cross-border registries 22
5.3.2 Quality and utility label meta-data specifications 23
5.4 Minimisation and purpose limitation 24
5.5 Return of research outputs 25
6 Governance in the TEHDAS DQF 26
7 Guidance for the implementation of the TEHDAS DQF in HealthData@EU 30
Annex 1 - Definitions in article 55 33
Annex 2 - Data Quality and Utility Labelling: survey results 35
2.1 Methodology 35
2.2 Results 35
Annex 3 - Minimum data requirements in cross border registries 48
3.1 Methodology 48
3.2 Cross-border registries reviewed 48
Annex 4 - Results of the survey voting the final recommendations 49
4.1 Methodology 49
4.2 Results 49

1 Executive summary

This report describes the TEHDAS JA Data Quality Framework (DQF) as an approach to data quality and utility for the secondary use of health data in the context of HealthData@EU.

In TEHDAS DQF, Data quality refers to how data fits data users' needs. This fit-for-purpose definition implies an approach to data quality that includes both, elements of technical quality and utility. From TEHDAS perspective, quality and utility dimensions are relevance, accuracy and reliability, coherence, coverage, completeness and timeliness. This fitting-for-purpose approach in TEHDAS DQF implies focusing both on datasets and data holders. In this approach data holders' maturity in data quality management becomes paramount.

TEHDAS DQF identifies several activities and services along the Data life cycle that HealthData@EU actors should provide with a view to ensure data quality and utility. Some of those activities are placed at the data preparation phase when data holders process the data for reuse; among those activities, the use of data management and data quality assurance procedures, the semantic mapping of the datasets using international standards, the linkage of datasets and the application of privacy enhancement technologies, the publication of meta-data referring to their datasets and, eventually the enrichment of their datasets and procedures after the reuse of the datasets is terminated. Some other activities affecting data quality and utility are put in place when the data user interacts with the HealthData@EU; those activities include the cataloguing of meta-data, activities procuring data minimisation and purpose limitation, data processing in secure processing environments (SPE), activities aimed at returning the research outputs pursuing data and procedures for the enrichment of data at the data holder level and activities enabling data users' feedback on fitness-for-purpose.

TEHDAS DQF proposes using both the legal enforcement (and subsequent implementing and delegated acts) and the use of guidance and recommendation as governance mechanisms at the disposal of the HealthData@EU actors. Among those to be legally bound: the publication and cataloguing of datasets, the labelling of the datasets according to quality and utility, the implementation of privacy enhancement technologies and procedures, and the supervision of measures pursuing data quality and utility. Among those to be implemented under a mechanism of guidance and recommendation: implementation and supervision of the data quality assurance maturity models, implementation of semantic standards, the implementation of mechanisms for privacy enhancement, the implementation of mechanisms for datasets enrichment, or the implementation of mechanisms for the return of other digital objects to enrich the data quality procedures.

Finally, TEHDAS DQF provides 13 recommendations affecting the implementation of data quality and utility in the HealthData@EU. The recommendations were voted on and reached a high level of agreement.

1 Glossary

Term	Definition (source)
General	
Data holder maturity	In the context of TEHDAS, it refers to the maturity of the data quality management procedures (i.e., data governance) at data holder level. It is included as a condition for data quality. In the Capability Maturity Model (CMM), maturity can be viewed as a set of structured levels that describe how well the behaviours, practices and processes of an organisation - data holder - can reliably and sustainably produce required outcomes. A maturity model can be used as a benchmark for comparison and as an aid to understanding and comparative assessment of different organisations.
Data quality	In the context of TEHDAS, data quality refers to how data fits data users' needs. These needs refer to the secondary use of health data for health research, policy making and regulation.
Data utility	Data utility is a dimension of data quality that relies both on a priori conditions and post-hoc conditions of the data centred on the data user. Commonly, *a priori* conditions of the data configuring utility can be categorised within the fit-for-use approach in which main data quality dimensions can be assessed to inform about the operational status of the data to be used (or reused). Post-hoc conditions of the data to inform utility are mostly based in fulfilling the expectations of a potential user that are specific to a certain purpose - following a fit-for-purpose approach. Utility can be measured using metrics of utilisation (i.e., in how many studies a data set has been exploited), metrics of interest (i.e., how many users has queried the data), or value scores (i.e., how the data user value the data provided) (see https://www.hdruk.ac.uk/helping-with-health-data/data-utility-evaluation/)
Dataset	A collection of data published or curated by a single agent, and available for access or download in one or more representations (distributions, data services, or media/format). A dataset can be a subset formed on the basis of a wider data collection (e.g., a data set formed based on a national patient registry).
Fit-for-purpose	The degree to which a dataset is suitable for a particular application or purpose, encompassing factors such as quality, credibility, scale, interoperability, accessibility, cost, format, timeliness, and so on. Data fits the purposes of the user - post hoc judgement. May implies the need for collecting users' feedback, or users' returns (i.e., incentives for data enrichment).
Fit-for-use	The suitability of data for the intended use, that is, the degree to which the data meets the needs of a user for their use. Data preparation main aim will be fit-for-use.

Primary use of data	'Primary use of electronic health data' means the processing of personal electronic health data for the provision of health services to assess, maintain or restore the state of health of the natural person to whom that data relates, including the prescription, dispensation and provision of medicinal products and medical devices, as well as for relevant social security, administrative or reimbursement services. (https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52022PC0197)
Secondary use of data	'Secondary use of electronic health data' means the processing of electronic health data for purposes set out in Chapter IV of this Regulation. The data used may include personal electronic health data initially collected in the context of primary use, but also electronic health data collected for the purpose of the secondary use. (https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52022PC0197)
Actors (Roles)	
Data holder	'Data holder' means any natural or legal person, which is an entity or a body in the health or care sector, or performing research in relation to these sectors, as well as Union institutions, bodies, offices and agencies who has the right or obligation, in accordance with this Regulation, applicable Union law or national legislation implementing Union law, or in the case of non-personal data, through control of the technical design of a product and related services, the ability to make available, including to register, provide, restrict access or exchange certain data. (https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52022PC0197)
Data types	As specified in Chapter IV, article 33 on Minimum categories of electronic data for secondary use in the EHDS regulatory proposal.
Data user	'Data user' means a natural or legal person who has lawful access to personal or non-personal electronic health data for secondary use. (https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52022PC0197)
Data life cycle	
Data access	Processing by a data user of data that has been provided by a data holder, in accordance with specific technical, legal, or organisational requirements, without necessarily implying the transmission or downloading of such data (as in the TEHDAS Glossary in WP7)
Data collection	Refers to data collection to secondary use and entails all the procedures to make this data available before standardisation and harmonisation as in WP7
Data preparation	In the data life cycle refers to any operation or set of operations which is performed on data, whether or not by automated means, included in the three steps of collection, standardisation and publication aimed to make the data available, interoperable and findable for reuse (*as defined in TEHDAS*

	WP7)
Data publication	Refers to the activities aimed to make the data findable, interoperable and accessible including metadata documentation and cataloguing in public repositories that can be queried following a standard syntaxis enabling precise searching. Data publication can also be referred to the activities leading to the cataloguing and indexing of data in a public and interoperable data repository as stated in TEHDAS WP7
Data standardisation	Refers to the critical process of applying standards to the data to make it syntactically, semantically, and technically interoperable. Data may be stored in different formats using different database systems and information models; and despite the growing use of standard terminologies in healthcare, the same concept (e.g., blood glucose) may be represented in a variety of ways from one setting to the next. Data standardisation imply the processing of data aimed at achieving compliance with a certain standard (as in TEHDAS WP7).
Data use	Refers to the analysis of the data (i.e., processing or querying) for the purposes stated in the EHDS regulatory proposal of research, regulation and informing policies. More concretely, the data use phase starts when the data access/use has been granted to the data user. In this phase, the data user finally performs the data analyses he or she needs to as part of their work. The data use phase finishes when the data user has finished its research project or have found the evidence to support new or existing policies or regulations. The finalisation of the data analysis phase may be also subject on contractual arrangements, for example, limiting the amount of time a data user has access to the data (as in TEHDAS WP7).
Discoverability	Discoverability is the degree to which a data set or source can be found in a search, a file, a database, or other information systems. Discoverability is related to data publication, metadata documentation, and harmonisation. It is different from accessibility and usability, other qualities that affect the usefulness of a piece of information. In the data discovery phase, the data user looks for the data needed to perform their work (answer a research question and/or make decisions regarding new or existing policies or regulations). Once the search is performed, he or she decides on the feasibility of carrying on their study according to the data found, possibly with the advice of data experts.
Finalisation	The finalisation phase is the last phase in the Users' Journey. It starts when the research question is answered, or the evidence required to support a legislative proposal or regulation has been found. In this phase the data user needs to ensure a proper disclosure of its findings to the rest of the EHDS2 users, following the FAIR principles for results data. The findings should also be notified to data controllers to finally inform data subjects. The results archival and validation services include all the software related to enable the effective storage and cataloguing of the projects results, including the related metadata and other for its potential re-use in further projects, including the safeguards to validate the dissemination levels authorised by the involved parties (data controllers of the original data, data permit authorities). Data and metadata, and any other supplemental material (analysis scripts, manuals,

	others), should be included to guarantee the reproducibility of the analyses by other data users, following the FAIR principles; thus, reproducible results should be interoperable, findable and accessible. Results cataloguing is expected to facilitate the further re-use in connection to the data search services of the data discovery phase of the Users' Journey. The finalisation of the data analysis phase may be also subject on contractual arrangements, for example, limiting the amount of time a data user has access to the data. Also defined as 'termination' in TEHDAS WP7.
Return of research outputs	Formerly 'devolution'. Refers to sharing or making openly available datasets or other digital objects produced as research outputs and cataloguing them as available resources supporting further research in a virtuous cycle. In the context of HealthData@EU this could also entail promoting the use of these outputs to enrich existing datasets and data quality assurance procedures at the data holder level.
Users' journey	Second group of activities and services that starts when the researcher is interested in searching for existing data. This phase includes: Collection, Standardisation, Publication, Discovery, Access, Use, and Finalisation (*see figure below*).
Data management	
Data curation	Data curation is the process of creating, organising and maintaining data sets so they can be accessed and used by people looking for information.
Data staging	A staging process imports information as streams, transforms it someway to produce integrated, cleaned data, and stages it for loading into permanent or long-term storage (i.e., operational data stores). A staging area, or landing zone, in a data architecture, is an intermediate storage area used for data processing during the extract, transform and load (ETL) process at capturing the data.
Data profiling	Refers to the process of examining, analysing, reviewing and summarising data sets to gain insight into the quality of data.
Data remediation	Data remediation is the process of cleansing, organising and migrating data so that it's properly protected and best serves its intended purpose.
Data servicing	Data operationalisation and servicing refers to the development and implementation of data services. Data services are self-contained units of software functions that give data characteristics it might not originally have to make data more available, resilient, and comprehensible, which makes data more useful to users and programs. Data service functions turn inputs into relevant outputs to the operational process that require information support.
Interoperability	
Data portability	As in CAMSS (link).

European Interoperability Framework (EIF)	The European interoperability framework is a commonly agreed approach to delivering European public services in an interoperable manner. It defines basic interoperability guidelines through common principles, models and recommendations.
Interoperability	Following the European Interoperability Framework, interoperability refers to a) full compliance with the legal and ethical provisions in each constituent node, b) an organisation that supports knowledge exchange and software transference across nodes, c) a compatible technological environment that supports the communication between nodes and allows the deployment of the computational tasks, and d) the existence of common data models that enables semantic standardisation across data sources. In a distributed research infrastructure, interoperability is a key feature for its governance and achievements.
Openness	The level of openness of a specification/standard is decisive for reusing software components implementing that specification. This also applies when such components introduce new European public services.
Organisational interoperability	This refers to the way in which public administrations align their business processes, responsibilities and expectations to achieve commonly agreed and mutually beneficial goals. In practice, organisational interoperability means documenting and integrating or aligning business processes and relevant information exchanged. Organisational interoperability also aims to meet the user community's requirements by making services available, easily identifiable, accessible and user-focused.
Reusability	Reuse means that public administrations confronted with a specific problem seek to benefit from the work of others by looking at what is available, assessing its usefulness or relevance to the problem at hand, and, where appropriate, adopting solutions that have proven their value elsewhere. This requires the public administration to be open to sharing its interoperability solutions, concepts, frameworks, specifications, tools and components with others.
Semantic interoperability	Semantic interoperability ensures that the meaning of exchanged data and information is preserved and understood throughout exchanges between parties; in other words, ‘what is sent is what is understood’. In the EIF, semantic interoperability refers to the meaning of data elements and the relationship between them. It includes developing vocabularies and schemata to describe data exchanges and ensures that all communicating parties understand data elements in the same way.
Syntactic interoperability*	The syntactic aspect describes the exact format of the information to be exchanged in terms of formats, conceptual and logical models, and

	organisation of the information (i.e., variable structure, units, type of data, transformation and validation rules, etc.)
Technical interoperability	Technical interoperability covers the applications and infrastructures linking systems and services. Aspects of technical interoperability include interface specifications, interconnection services, data integration services, data presentation and exchange, and secure communication protocols.
Technological neutrality	As in CAMSS, (link) decision on technologies supporting data reuse should focus on functional needs minimising technological dependencies, to avoid imposing specific technical implementations or products on their constituents and to be able to adapt to the rapidly evolving technological environment.
Transparency	Transparency in the EIF context refers to Enabling visibility inside the administrative environment of a public administration, ensuring the availability of interfaces with internal information systems and securing the right to the protection of personal data.

2 Acronyms

Acronym	Definition
AF	Atrial Fibrillation
API	Application Programming Interface
BBMRI	Biobanking and Biomolecular resources Research Infrastructure
CAMSS	Common Assessment Method for Standards and Specifications
CMD	Common Data Model
CMM	Capability Maturity Model
DCAT-AP	Data CATalog vocabulary (DCAT)- Application Profile for data portals in Europe (AP)
DPA	Data Protection Authority
DPV	Data Privacy Vocabulary
DQF	Data Quality Framework
DQV	Data Quality Vocabulary
DUOS	Data Usage Ontology Systems
ECDC	European Centre for Disease Prevention and Control
ECHO	European Collaboration for Healthcare Optimisation
ECIS	European Cancer Information System
ECRIN	European Clinical Research Infrastructure Network
EHDS	European Health Data Space
EHR	Electronic Health Record

EMA	European Medicines Agency
ENCR	The European Network of Cancer Registries
EORP-ESC	European Observational Registry Programme (EORP) - European Society of Cardiology (ESC)
EOSC	European Open Science Cloud
EPIRARE	Platform for Health and Genetics
EUCERD	European Union Committee of Experts on Rare Diseases
EURD	European Platform on Rare Disease Registration
EuroHEART	Data standards for heart failure: the European Unified Registries for Heart Care Evaluation and Randomised Trials
EUROSTAT	European Statistical Office
FAIR	Findable, Accessible, Interoperable, and Reusable
FOAF	Friend Of A Friend is a machine-readable ontology describing persons, their activities and their relations to other people and objects.
GDPR	General Data Protection Regulation
HDAB	Health Data Access Bodies
HDRUK	Health Data Research-UK
HealthData@EU	European Health Data Space for secondary use (EHDS2)
HF	Heart Failure
ID	IDentifier
LOST	Legal, Organisational, Semantic/Syntactic, and Technical Interoperability
MyHealth@EU	Electronic cross-border health services in the EU (EHDS1)
NSTEMI	Non-ST- segment Elevation Myocardial Infarction

ODPRL	Open Digital Rights Language
OHDSI	Observational Health Data Sciences and Informatics
OMOP-CDM	Observational Medical Outcomes Partnership- Common Data Model
OSSE project	Open-Source Registry System for Rare Diseases in the EU
PARENT	PAtient REgistries iNiTiative
PET	Privacy Enhancement Technologies
PHIRI	Population Health Information Research Infrastructure
PROV-O	PROVenance Ontology
RCTs	Randomised Controlled Trials
RDF	Resource Description Framework
REC	Research Ethics Committee
SKOS	Simple Knowledge Organisation System
SPE	Secured Processing Environment
TEHDAS	Towards a European Health Data Space Joint Action
URI	Uniform Resource Identifier
URL	Uniform Resource Locators
WP	Work Package

3 TEHDAS Data Quality Framework

The TEHDAS data quality framework (DQF) aims at setting up the basis for a trustworthy and reliable secondary use of data and providing guidance on its implementation. The TEHDAS DQF builds on the principles of a) trust across data institutions and between data institutions and users; b) transparency in the processing of the data, from the collection to the publication of metadata; and c) continuous improvement, benchmarking and promotion.

The TEHDAS DQF develops upon those elements relevant to secondary use of data in the context of the HealthData@EU; where the DQF is relevant and what for, who are the actors in the eventual deployment of the DQF and how the DQF should be implemented.

We have to highlight the importance of standardisation in the process of the primary collection of data (e.g., the extensive use and appropriate use of controlled vocabularies and standards and the quality of coding), as these will be strong determinants of data quality when that data is made available for secondary purposes, as well as highly determining the amount of processing efforts that data holders will face to prepare the data according to required levels of quality. However, the TEHDAS data quality framework will not provide specific recommendations for standardisation in the primary collection of data.

3.1 Definition of data quality and utility

In the context of TEHDAS, data quality refers to how the extent the data fits the data users' needs. These needs refer to the secondary use of health data for health research, policy making and regulation.

This fit-for-purpose definition implies an approach to data quality that includes both elements of technical quality and utility. Relative to this observation, quality and utility dimensions in TEHDAS approach are [1]:

Dimension	**Definition**
Relevance	How well data meets users' needs.
Accuracy and Reliability	How closely data reflects what it was designed to measure and whether this is consistent over time.
Coherence	How consistent is data across data sources and data holders and can be combined and compared.
Coverage	The degree of representativeness in the population to which the dataset refers to, its exposures and events.
Completeness	Level of missingness at variable level.

Timeliness	How up-to-date the information is collected and delivered.

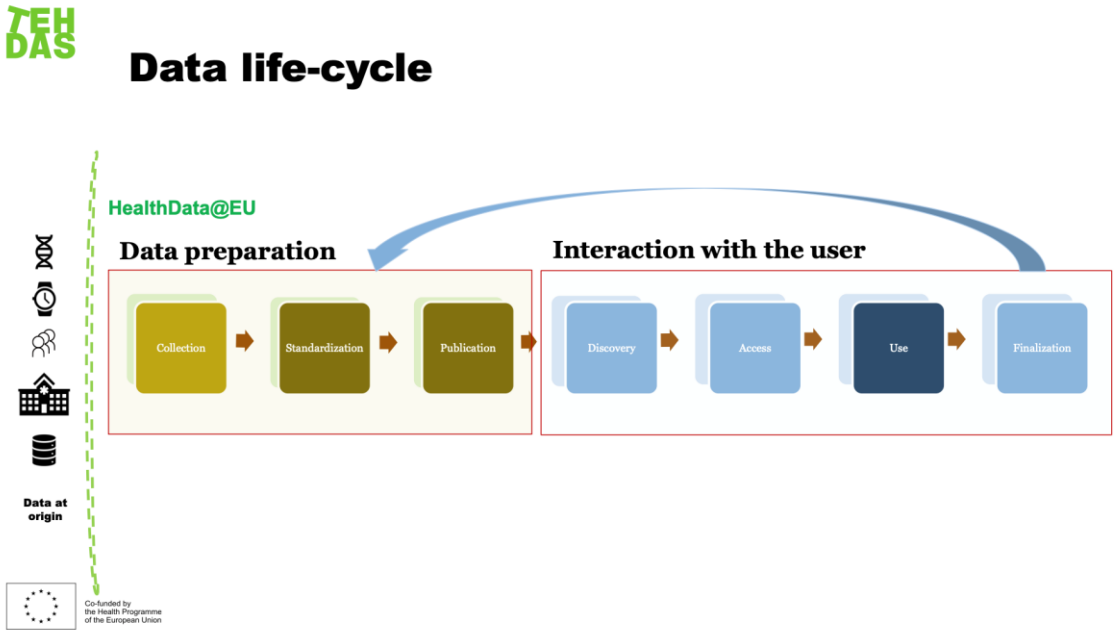
This fitting-for-purpose approach also implies a two-fold perspective. The first one focuses on data quality and utility at dataset level; and the second one, focuses on the maturity of the data quality management procedures at data holder level.

3.2 Pragmatic approach: Data life cycle in HealthData@EU

The TEHDAS data life cycle and user's journey seek to describe the process that the different actors interacting within the HealthData@EU should follow once data collected for primary purposes is made available for secondary uses.

The data life cycle distinguishes between two overarching phases - data preparation and interaction with the end user as in figure 1. The former entails the retrieval of data or collection of metadata from the primary sources [i.e., data collection for primary purposes represented at the left of the green dotted border in figure 1], their preparation for secondary use making them interoperable, and the publication of preparation procedures, data sources and data collections in a way that is easily findable. The latter describes the stages comprising the users' journey, the interaction of the end user with the institutions that may grant access to data; so, once data collections of interest are discovered, how to ask for access permissions, how to access and use the actual data, and how to finalise the use of data including return of intermediate outputs and enriched dataset to the data preparation institutions.

Figure 1: Data life cycle for secondary use

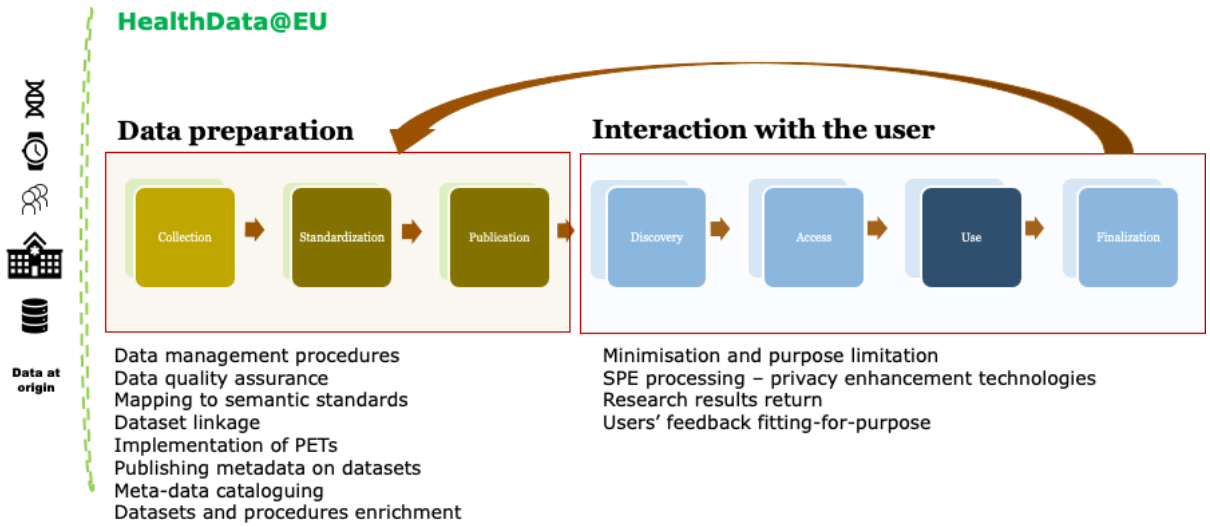


4 Key elements in the TEHDAS DQF

The data life cycle has helped to identify what services HealthData@EU actors should provide at each stage (see in TEHDAS WP7 deliverable on minimum technical services). In relation to the data quality, figure 2 provides a notion of the different quality elements to be taken into consideration at each stage of the journey.

Thus, in the preparation phase, quality would highly rely on: 1) establishing clear data requirements (*for example, the minimum data to make available, the preparation for data source linkage, the harmonisation of data sources and data collection to be semantically interoperable*); 2) a programmatically interoperable cataloguing of the sources (*implementing meta-data standards that allow a description of the data sources, their provenance, and preparation procedures*). Once the users have been granted access, the quality will depend on the impact of linkage between data sources, de-identification and minimisation procedures before data is made effectively available for use.

Figure 2: Services within the data life cycle that may have an impact on data quality



4.1 Data quality management and quality assurance procedures

The rich landscape of data holders across Europe includes statistical organisations, public health agencies, healthcare providers, health insurances, research infrastructures, EU agencies, etc. hosting a variety of data (electronic health records, patient or population registries, biological data, claims and administrative data, statistical data, research data, etc.). According to the regulation on the European Health Data Space, article 33, data holders shall make their dataset available for secondary purposes. Thus, data holders are the very starting point of the TEHDAS data life cycle, where data preparation for further use is needed.

In the context of data quality, data holders will have to make a big effort in the preparation of data and to implement data quality management and quality assurance procedures. In the case of research data, data from registries or statistical data, where data collection is based on strict protocols and data are thoroughly curated and maintained, data quality management is implemented by design, though discipline of data providers and collection methods still play an important role. However, in the case of routine data, where data is purely collected for care purposes (as in electronic health data in EHR), data holders will have to implement data quality management and quality assurance procedures, especially when datasets are updated regularly, and when there is a need for datasets linkage.

Data quality management and quality assurance are integral pieces within the data governance of an organisation. A data quality assurance framework should be transversal to all data management processes including monitoring, detection and resolution of incidences, and data enrichment procedures. Data quality management should be applied throughout the full data life cycle focusing on a) data collection, curation, storage and staging, b) data integration with relevant sources and systems, c) data description and metadata management (i.e. use of meta-data standards), c) data quality assessment, data profiling and remediation, d) data modelling, transformation, operationalisation and servicing.

Depending on the level of maturity of the data holders, these procedures are more or less automated. The Capability Maturity Model (CMM) provides 5 levels of maturity according to the actual capacity of the data holders to implement continuous data quality improvements. The five levels are: 1) Initial: there is an implementation of *ad hoc (non-systematic)* data quality checks poorly documented; 2) Repeatable: Data Quality Management procedures documented sufficiently such that repeating the same steps may be attempted; 3) Defined: Data Quality Management procedures sufficiently defined and implemented as a standard process; 4) Managed: The Data Quality Management process includes quantitative metrics of quality; and, 5) Optimised: Data Quality Management implies a deliberate process of optimisation and continuous improvement.

In addition to this more general CMM approach, in the specific domain of secondary use of health data there are some best practice in maturity worth highlighting; thus; OHDSI QA/QC methodology, BBMRI QA/QC methodology, ECRIN QA/QC certification model, FAIRplus Cookbook, HealthyCloud approach to FAIR Maturity,, Health Data Research UK in the case of research purposes; Health Information and Quality Authority in Ireland in the case of policy decision-making; and the EMA maturity model in the case of regulatory purposes.

4.2 Semantic interoperability

Once data quality management and quality assurance procedures ensure a certain level of quality, data holders may need to implement a semantic and syntactic interoperability layer across datasets. Syntactic as how data is structurally persisted within a dataset (i.e., literal name, standard abbreviation and encoding) and semantic as consistency in data meaning across datasets (i.e., clear description, operational definition rules, mapping across controlled vocabularies and standards).

There are two potential frameworks to data harmonisation and standardisation. One approach relies on the transformation of all datasets held by a data holder against a specific standard, ideally although not necessarily, an international widely adopted standard; another approach, relies on the specific preparation of the dataset to be delivered according to a specific data schema that contains the required harmonisation rules, controlled vocabularies and standards. In the first case, there is an important effort at the beginning that could lead to a more efficient data preparation in the future; in the second case, each transformation entails limiting the transformation of data to those specific entities and variables of interest. A good example of the first one is Darwin EU® and the OHDSI network approach; known examples for the second approach are EUROSTAT, ECDC, ECHO or PHIRI federated approach or those cross-national disease registries, as for example EURD, EPIRARE, EORP-ESC, EuroHEART, ECIS, ENCR (see annex 3), among others.

Some lessons learnt from the assessment of interoperability standards recently released in Deliverable 6.2 (see recommendations 5 to 8) would recommend adopting the second framework, at least in the short to medium term. Among the reasons of this second approach to semantic and syntactic interoperability: a) none of the semantic standards can cover all the data types of interest in the HealthData@EU (according to article 33); b) there is no single ontology or controlled vocabulary that covers all the semantic field of medical concepts, which impedes a univocal transformation from any coding system to that standard; c) there are substantial gaps in semantic mapping for other determinants of health (i.e., social, cultural and economic determinants, environmental determinants, genetic determinants); and, d) semantic and syntactic interoperability at variable level is not sufficient to get datasets harmonised and standardised; thus, the preparation of data for secondary use should not be limited to the mapping of concepts. It also requires the development of data models providing a logical harmonised schema integrating different health data sources among data holders and over time. Therefore, using standard entities and mapping their relationships in the model: cohorts, individuals, place of residence, place of treatment, contacts with the system, treatments, and time.

4.3 Datasets publication and cataloguing

Within the TEHDAS DQF, datasets cataloguing and publication lies more on the utility of the datasets than on the technical quality side. Cataloguing and publication implies making datasets programmatically findable, with the information that is relevant to users' purposes. Underlying to both requirements, it is the use of meta-data standards as those assessed in the deliverable 6.2.

When it comes to what information is relevant to the users' purposes, a balance is needed to reduce the level of false positive and false negative retrievals. However, different users' communities have opted for different standards. Some lessons learnt in the assessment of meta-data standards in deliverable 6.2 were: a) the standards developed in the context of Open Data and Public Administration are those better equipped for data discoverability and no so well equipped to respond to the needs of the research communities; b) on the contrary, those standards developed in the context of research communities ranked top in the user-centric approach highlighting how relevant for the specific research communities would be maintaining those standards; c) none of the metadata standards are part of a national framework or national strategy in the context of health data or have been widely adopted as a standard for discoverability.

Out of these facts, it seems sensible the combined use of general and domain-specific meta-data standards; at a first stage, focusing on gathering high-level knowledge on the data sets available that is agnostic to the domain or the type of data (as in DCAT); at a second stage, by providing further details on the quality and utility of the dataset for various secondary use purposes; and, at a third stage, where the focus is the actual content of the data source (i.e., variable level) as in Beacon. In article 55 in the Regulation on the EHDS legislative proposal, high-level knowledge on a dataset has been formalised as information on the data source, the nature of the data, the type of data, the key characteristics of the dataset and the conditions to make data available. Similarly, the EHDS proposes the creation of a data quality and utility label to provide the characteristics and the potential usefulness of datasets and to support to data holders in identifying and addressing areas of improvement.

The DCAT specification

After the application of the Common Assessment Method for Standards and Specifications (CAMSS), the European guide for assessing and selecting standards and specifications for an eGovernment, DCAT was found the best equipped generalist meta-data standard in all the domains of analysis; thus, Openness, Transparency, Reusability, Technological neutrality and portability, User centricity, Inclusion and Accessibility, Security and Privacy, Multilingualism, Administrative simplification, Preservation of Information, Assessment of effectiveness and efficiency, Governance and Legal, Organisation, Semantic and Technological interoperability.

DCAT is an RDF vocabulary representing data catalogues, and builds upon six interconnected classes whose properties enable a publisher to describe datasets and data services in a catalogue using a standard model and vocabulary that facilitates the consumption and aggregation of metadata from multiple catalogues:

- dcat:Catalog
    - dcat:CatalogRecord
        - dcat:Resource
            - **dcat:Dataset**
                - dcat:Distribution
            - dcat:DataService
            - *dcat:Catalog (nested)*

For the purposes of HealthData@EU, the class *dcat:Dataset* may provide the opportunity for a description of the elements enforced in article 55.

Generally, *dcat:Dataset* represents a collection of data, published or curated by a single agent (*foaf:Agent*) that could be an Organisation (*foaf:Organisation*) or less frequently a person (*foaf:Person*). The Dataset is considered a *dcat:Resource* included as a *dcat:CatalogRecord* primarily concerning some information on the registration of the resource, such as when a new resource has been added and who has added the resource into the *dcat:Catalog*.

DCAT is a general standard that incorporates terms from pre-existing standard vocabularies where stable terms with appropriate meanings could be found (i.e., Dublin Core, SKOS, FOAF, etc.) offering also the possibility of extending the standard through the specification of DCAT profiles. Profiles are

defined by expanding the classes and properties of DCAT to cover more detailed information and by specifying in technical detail how to fulfil the expected values for each class and property, an example being the information requirement for a record in a catalogue, describing the registration of a single data resource (being a data set or a data service) is described by the properties title, description, listing date, update/modification date, primary topic, and conforms to. For each of this properties a DCAT profile could specify a format, a closed list of categories, or the inclusion of standardised keywords in the description, the standards used to complete the dates, a concept list for the specification of the primary topic (i.e., by using the property *skos:Concept* or by establishing a relationship with a relevant standard ontology of topics), etc. This comprehensive capability would enable the configuration of a DCAT application profile (DCAT-AP) specific for the domain of health, and to HealthData@EU providing support to the collection, publication and discovery of health data sources for secondary use.

There is a question, though, on how well the DCAT meta-data standard, in particular the current DCAT specification, covers the article 55 information requirements regarding the minimum set of elements to describe a dataset within HealthData@EU. An operational definition of nature of the dataset, the source and the scope of the dataset, the main characteristics and the conditions to make the dataset available (Annex 1). According to these definitions:

Nature of the data can be addressed within DCAT by using some properties from the *Class: Catalogued Resource* applied to the *Class:Dataset*, and further specifying a standard vocabularies to fill in the properties: *Property: access rights*, *conforms to, contact point, resource creator, licence, rights, and has policy*, providing the information required to appropriately targeting the data application process or the data request within the governance of the HealthData@EU. There is also the possibility of using standardised definitions of some properties of the *Class: Distribution*, such as *licence*, *access rights*, *has policy*, *access URL*, *access service*, and *conforms to*, also applied to the *Class:Dataset.* In addition, the technical specification of the DCAT-AP for HealthData@EU can make use of the Open Digital Rights Language (ODPRL), the Data Privacy Vocabulary (DPV), the Data Usage Ontology (DUO) and the data usage ontology system (DUOS) to standardise the expected values of the commented properties. In addition, the use of a standardise specification of the *Class: Catalogued Resource* is the best option for metadata catalogues maintain a supported by HDABs with the responsibility of managing the health data application processes, therefore sharing these 'nature of data' properties across all health data sets under their stewardship.

Source of the data can be addressed within DCAT by using a standardised definition of the *Class: Organisation/Person* and the *Class: Role* to provide information on the properties regarding *resource creator*, *resource publisher*, and *qualified attribution* configuring the source of the data as the data holder responsible for their collection, management and usage. Further information on the source of the dataset can be provided by specifying the provenance using classes, properties and restrictions within the provenance ontology (PROV-O).

Scope of the data can be addressed within DCAT by defining the standardised vocabulary for scope based on the list of scopes of health data to be made accessible by HealthData@EU as provided in article 33. This standardised vocabulary can be defined using the DCAT classes *Class: Concept Scheme* and *Class: Concept,* within a HealthData@EU DCAT-AP.

4.3.1 Cross-border registries

A specific category of data types are cross-border data collections (for example, cross-border registries). In Annex 3 detail is provided on a piece of work carried out to figure out what have been data sharing commonalities in those cross-border initiatives, and to derive what could be the minimum information requirements at variable level as stated in article 58 in the EHDS proposed regulation. In short, these are the main information requirements:

- All information is at the individual level
- Patient Identification (i.e., national healthcare patient ID, personal identification, full name and address, patients' pseudonym, EU global unique identifier, etc.), including personal information (i.e., full name, address, date of birth, insurance status and identification, contact information, etc.)
- Patients' socio demographic information (i.e., age, sex, gender, socioeconomic level, income, education level, occupational status, race, etc.)
- Information on patients' status (i.e., death, disability, etc.)
- Information on patients' inclusion in EU registry (i.e., rare diseases, etc.)
- Information on patients' compliance with inclusion criteria for an EU registry (specific for each registry)
- Information on medical history - e.g., allergies, illnesses, diagnoses
- Information on patients' informed consent to be included in the registry
- Information on patients' participation in clinical trials (i.e., URI of the clinical trial in which the patient is included)
- Information on patients' possible availability for other research projects
- Information on availability of genomic and genetic data
- Information on availability of biological samples
- Information on availability of medical images
- Information on usage of medical devices

Less common but also shared by some EU registries

- Information on *treatments*
- Information on *adverse effects of medical treatments* (or adverse events)
- Information on *quality of life*

In the eventual case of mobilising data for the purpose of generating a cross-border registry-like data collection, a DCAT profile specification should consider the need to standardise the reporting of these features, following the guidance for describing a dataset based on article 55 information requirements.

Main characteristics of a dataset can be addressed within DCAT by defining standard concept schemes for main properties of the DCAT classes Class: Catalogued Resource (in particular, description, title, release date, update/modification date, language, theme/category, type/genre, resource relation, qualified relation, keyword/tag, qualified attribution, and is referenced by), Class: Record (in particular, description, listing date, update/modification date, primary topic, and conforms to), Class: Dataset (in particular, dataset distribution, frequency, spatial/geographical coverage,

spatial resolution, temporal coverage, and temporal resolution), Class: Distribution (in particular, title, description, release date, update/modification date, spatial resolution, temporal resolution, conforms to, media type, format, compression format, and packaging format). In addition, Class: Distribution can be expanded to attend to the information requirements defined by article 56 proposing a data quality and utility label by building a detailed technical specification for quality assessment based on the classes and properties in the Data Quality Vocabulary (DQV). Finally, this data quality and utility label can be extended to include information on the maturity level of data quality of the data holders (i.e., Class: Organisation/Person) via classes, properties and restrictions of the provenance ontology (PROV-O).

4.3.2 Quality and utility label meta-data specifications

For this latter, WP6 partners were consulted with the aim of eliciting their views with regard to the degree of agreement on the relevance of those categories and dimensions referred in article 56. Some of those characteristics were deemed highly relevant in the survey and participants strongly agreed on this level of relevance; some other categories and dimensions lacked that level of relevance, and discrepancy across participants was the rule. (See results in Annex 2) A summary of the results is provided hereinafter:

"Data documentation" was rated as **highly relevant** (average vote 8.7) in qualifying the quality and utility of a dataset, with a **high agreement** between respondents (values ranged from 8 to 9). Within this category, although scored 7, high disagreement was found in providing a data profile of the dataset at variable level as part of the documentation (meta-data at variable level, number of observations, range of values per variable, visual distribution of values, visual quality assessment, etc); including the end users' assessment as part of the documentation was found of low relevance (scored 6.2) and the level of discrepancy was high.

"Technical quality" was rated as **highly relevant** (average 8.3) when qualifying the quality of a dataset; however, the level of **agreement was lower** than in the previous category (values ranged 7 to 9). All the dimensions of quality were found highly relevant (relevance, accuracy and reliability, completeness, coherence, and timeliness) although the latter scored lower (7) and the survey found more discrepancy

"Coverage" was rated as **relevant** (average 7.5) in qualifying the utility of a dataset, although the level of **agreement was moderate** (most of the values ranged from 7 to 9). Among its dimensions, the representativeness of the population and time-span covered were deemed relevant with a fairly high degree of agreement while, the variety of data types and data sources showed high discrepancy.

"Access and provision" were rated as **moderately relevant** (average 7) in qualifying the utility of a dataset, although the level of **agreement was fairly low** (values evenly ranged from 5 to 9). Major discrepancies in considering access and provision relevant in the labelling of a dataset were found in the time-lag until datasets are made available by the data holder after data collection (measures preparation phase), time-lag between data access application and delivery, and time-lag between

return and enrichment.

"Value and interest" were rated as hardly **relevant** (average 6.5) in qualifying the utility of a dataset, but the level of **agreement was low** (values sparse across the scale). This small agreement is not so evident in some dimensions- allowing over-time update, inclusion of audit and continuous improvement mechanism, maps to interoperability standards and datasets map to a data model that is standard.

Conditions to make data available can be addressed within DCAT by using a standardised definition of the Property: has policy within Class: Catalogued Resource, and of the properties has policy, access URL, and access service of the Class: Distribution. In addition, conditions to make the data available can be provided by the HDABs as data services linked to the dataset distribution via a standard technical specification of the Class: Data Service, providing HealthData@EU standardise endpoints (URL and endpoint descriptions) to complete the dataset request or the data access application process leading to serving the data. These data services can be defined following the preferred technical specifications of the European Commision using available technologies such as the eDelivery tools and services, in particular the eDelivery SMP profiles (i.e., eDelivery Service Metadata Publisher and Access Points).

4.4 Minimisation and purpose limitation

In pursuing data minimisation and purpose limitation, health data access bodies will tend to make accessible only those data that are needed in relation to the purpose of research. Depending on how strict health data access bodies implement minimisation and purpose limitation policies, data may become useless to respond some relevant research questions (for example, delivering data that is anonymous may impede the analysis of changing exposures – as in the case of the implementation of COVID 19 vaccination programs).

The current state of play shows us that in a majority of EU countries, access to sensitive health data is granted by the authorising body after the evaluation of a research protocol including a detailed data management plan. Usually the authorising body (i.e., Health Data Access Body) is subject to the previous approval of a Research Ethics Committee (REC) and/or a Data Protection Authority (DPA). Once authorisation is provided, researchers sign contractual arrangements with the access body in particular when commercial entities are involved or commercial interests are at stake. In addition to a data access agreement, a principal researcher can sign a self-declaration committing not to re-identify individuals based on combining shared data with other public or non-public data sources, sub-processing agreements or confidentiality statements. In turn, access bodies should ensure that access is only provided to requested electronic health data relevant for the purpose of processing indicated in the data access application by the data user and in line with the data permit granted (art.44 (1) of the EHDS legislative proposal).

Provided this general procedure for minimisation and purpose limitation, GDPR and article 44 (3) of the Regulation on the EHDS legislative proposal (under consultation) provide legal backing for those occasions where data users need to access and use personal data. In those occasions, the

application of de-identification techniques reduces the risks for privacy while keeping data useful for research. According to article 45 (2) the data access requestor will have to justify the need to access pseudonymised data and describe the safeguards in place.

Beyond anonymity, there are many privacy enhancement technologies (PET) that can ensure privacy. K-anonymity may work in many situations, for example, ecological studies; Outputs aggregation and meta-analysis of aggregated data in federated approaches as in PHIRI play the same role; the use of synthetic datasets when these datasets are a faithful representation of the population where the original data come from can provide the basis for more advance research queries. However, there are research questions that may require continuous data update particularly when multiple sources provide data, and rapid cycle analysis is required - for example, in cohort studies aiming at the discovery of adverse events associated with the uptake of a new drug or observational studies aiming to discover real-life beneficial effects of an intervention on the population. In those cases, pseudonymisation techniques are preferable. In this case, there is a need for more minimisation efforts, in particular the use of a Secured Processing Environment (SPE) [Goldacre review]. This has become a requirement in the forthcoming EHDS Regulation in a way that only non-personal electronic health data could be transferred out or extracted from such a secure processing environment (article 50.2).

In order to preserve the value of data for those research queries where continuous data refreshment from multiple data sources is required and synthetic data do not fully mirror the original datasets, an alternative may be a multifaceted approach based on a data application that includes a detailed research protocol and a thorough data management plan (as in ARGOS) and, in addition, at the analytical phase, implementing a federated secure multiparty processing or the use of homomorphic encryption under the jurisdiction of an accredited SPE (see TEHDAS Deliverable 7.2).

4.5 Return of research outputs

Research outputs are typically scientific and policy reports, but in the context of the secondary use of data, research outputs of interest will be digital objects that are the results of the research process - data schemas, common data models, synthetic datasets, data quality checks, analytical workflows, partial outputs, enriched datasets, etc.

Return of digital outputs is conditioned to the quality of the data upon which research outputs have been built. Said that, it is important to highlight the relevance of research outputs return within the TEHDAS data-life cycle in coherence with Article 37 in the current Regulation on the EHDS legislative proposal.

An obligation of data users should be preparing those data sources in a way that are reproducible and interoperable, not just for the broader research community, but specifically to improve data holders' procedures, enrich their datasets and add new tools to the SPEs. To make the most of their research, data users should be advising on how to best devolve those digital objects in a FAIR way ideally using a programmatic approach (e.g., workflows publication in GitHub, CDM publication in Zenodo, API development for programmatic harvesting, etc.) allowing HDAB and data holders to implement the procedures to include and these outputs straightforwardly.

6 Governance in the TEHDAS DQF

In the table 1 key elements referred to in the previous section, are allocated to HealthData@EU actors. Importantly, main tasks with impact on data quality and utility are also provided and linked to the main two governance mechanisms for implementation 1) the legal enforcement (and subsequent implementation and delegated acts); and 2) guidance and recommendation.

The governance of data quality at **data holders level** includes managing: 1) the implementation of measures to achieve highest possible level of maturity in data quality management and quality assurance; 2) the implementation of layer(s) of interoperability, ideally international well-recognised semantic standards and data models; 3) implement an international meta-data standard, ideally DCAT and publish all datasets according to the meta-data specification; 4) implement the HealthData@EU data quality and utility label in a way that is publishable as part of the meta-data); 5) if a number of datasets are available with the possibility of linkage, pre-process the datasets to get them linked (1 to 1, 1 to N, or N to N), and when linkage is at individual level, pseudonymise the ID allowing updates; and 6) According to the level of risk, privacy enhancement technologies (PETs) should be used to minimise risks on privacy.

The governance of data quality at **health data access bodies (HDAB)** level includes managing: 1) the interoperable publication of meta-data according to the common HealthData@EU specification; according to Deliverable 7.2, the preferred option will be implementing a pushing mechanism that programmatically updates the National Catalogue and then the EU catalogue; 2) the supervision of the labelling mechanism, procuring external auditing and certification, when needed (for example, the initial levels of quality and utility can be based on self-assessments, but the rest of level would require external assessment of certified actors); 3) providing guidance and supervising the implementation of the data holders maturity model including assessment mechanisms as in the previous paragraph; 4) fostering dialogue on the governance of semantic interoperability taking inspiration from the works developed in MyHealth@EU and, 5) the supervision of the enrichment of datasets with annotations or new attributes may require that both the datasets provided access to and the final dataset share the same pseudonymised ID. HDAB has to implement a mechanism for the implementation and persistence of pseudonyms. Likewise, HDAB should provide data users with guidelines for a proper procedure for datasets enrichment and other digital outputs publication.

The governance of data quality at **SPE level** requires managing: 1) the implementation of analytical PETs to further minimise privacy risks. Deliverable 7.2 provides further elaboration on the use of PETs, specifically homomorphic encryption and secure multiparty computation; and 2) implementing procedures for the persistence of digital object and management of tools versioning.

At users' level, governance relies more on the actual uptake of the HDAB requirements of return after the access to data is granted. One of those requirements may be the assessment and report of the quality and utility of the datasets granted access.

Although it is out of the scope of the TEHDAS DQF, the collection of data for primary purposes determines the quality and utility of datasets when they are made available for secondary use. This

is particularly the case of data collected in the context of health care. Governance efforts have to be made to increase the data quality at the point of care.

Table 1: Actors, actions and preferred governance mechanism

Actor	Key element	Implementation	GM	Reference
Data Holder	Data quality management and quality assurance	Implement Capability Maturity Model	R	Examples in HDRUK, ECRIN, BBMRI
	Semantic and syntactic interoperability	Implement layer(s) of interoperability	R|L	Deliverable 6.2
	Describing the datasets	Interoperable publication	L	HEALTHDATA@EU article 41, article 55, Deliverable 6.2
	Quality and utility labelling (Q&U label)	Implement self-assessment and publish	L	HEALTHDATA@EU article 41, article 56
	Data linkage	Using single pseudonymous ID	R	Deliverables 6.1 and 7.2
	Anonymisation | pseudonymisation	Pre-processing PETs after linkage	L	HEALTHDATA@EU article 44. Deliverable 7.2
	Datasets enrichment	Requires pseudonymisation	R	
HDAB	Datasets cataloguing	Interoperable publication	L	HEALTHDATA@EU article 37 q, article 55, Deliv 6.2
	Labelling supervision	Self-assessment, Audit, Certification	L	HEALTHDATA@EU article 37 j
	Supervision of data holders' maturity	Guidance and assessment	R	
	Submission of annotated/enriched	Requires pseudonymisation	L	HEALTHDATA@EU article 37p

	datasets			
SPE	Privacy enhancement technologies for analysis	Analytical PETs	R	Deliverable 7.2
	Management of digital objects	Software versioning persistence	R	Deliverable 7.2
EC	EU datasets cataloguing	Interoperable publication	L	HEALTHDATA@EU article 57, Deliverable 6.2
	Semantic interoperability	Fostering dialogue on governance	R	Taking inspiration from MyHealth@EU
USERS	Return of enriched/annotated datasets	Report on provenance	R	
	Return of digital objects	FAIR by design using Open Science	R	Deliverable 6.1
	Users' experience	Including experience in the DQ&U label	R	

GM: main governance mechanism
R: recommendation
L: legally enforced

7 Guidance for the implementation of the TEHDAS DQF in HealthData@EU

This report has described the building blocks of the TEHDAS DQF - the rationale behind, what is relevant to data quality and utility, who should be in charge, and which are the main implementation tasks to govern with which governance tool.

Hereinafter, we provide guidance for the implementation of the TEHDAS Data Quality and Utility Framework within HealthData@EU. These recommendations collate previous deliverables in WP6 as well as documents produced in WP4, WP5 and WP7. The recommendations have been voted on, reaching a strong level of support among WP6 participant institutions (see method and results in Annex 4).

Recommendation 1: A HealthData@EU DQF should include not just the technical quality of data but also the utility of datasets, with a view of fostering a fit-for-purpose approach.

Recommendation 2. A HealthData@EU DQF should include as main data quality features relevance, accuracy and reliability, and coherence; likewise, as main utility features coverage, completeness, and timeliness.

Recommendation 3. A HealthData@EU DQF should also take into account the data holders' perspective by implementing actions towards improving their maturity in data collection, curation, storage and staging.

Recommendation 4. A HealthData@EU DQF should be applied along the whole Data life cycle with particular emphasis on data preparation at data holder level, at the dataset publication and discovery phase, when preprocessing the data before delivery, and when enriching the datasets, procedures and tools once research outputs are provided.

Recommendation 5. There is a need for a dedicated plan aimed at the implementation of a data holders maturity model to improve their data quality management and quality assurance procedures, and to reduce gaps across HealthData@EU data holders. All the data holders should be evaluated according to the levels of maturity established in such a model and an agreed notion of their maturity should be included as part of the meta-data of their datasets when made available. Health Data Access Bodies would specify the type of assessment procedure required in the evaluation of maturity; in this respect, data quality management experiences recommend a data holders self-assessment methodology for the initial phase of maturity and external audit and certification for the rest of the levels of maturity. Finally, the implementation of the maturity model should be progressive, and foster incentives for continuous improvement and level promotion.

Recommendation 6. In HealthData@EU, there is a need for data holders to implement a layer of semantic interoperability using widely adopted standards (see recommendations 5 to 8 in deliverable 6.2). As a preferred framework, in the short run, data holders should follow an incremental approach to progressively map their regular controlled vocabularies to

international general and domain-specific ontologies. The European Commision should support continuous dialogue on this governance mechanism, taking as an inspiration how the initiative fostering OMOP-CDM has addressed openness, transparency, technological neutrality, data portability, and cooperation among public institutions.

Recommendation 7. In HealthData@EU, data holders are expected to publish information on their datasets (article 41 of the current version of the Regulation on the EHDS legislative proposal) and health data access bodies to catalogue them all (article 55 of the current version of the Regulation on the EHDS legislative proposal). It is recommendable to combine the use of generic meta-data standards and domain-specific meta-data standards in a two-step approach to discoverability; at a first stage, users should know about the source, scope of the datasets, nature of the data, main characteristics and features of distribution; at a second stage, to allow further knowledge on the datasets to allow federated querying (e.g., providing data profiles). As enforced by law in the article 55, an implementing act should provide the technical specifications for this specific development.

Recommendation 8. In HealthData@EU, data holders are expected to publish a notion on the quality and utility of their datasets that are obliged to make available (articles 41 and 55 of the current version of the Regulation on the EHDS legislative proposal). Although there is a general agreement on the main categories that the label should contain, there are some discrepancies in the operational definitions of some dimensions. An implementation act for the implementation of a quality and utility label should stem from a formal consensual exercise for an operational definition of quality and utility that is instrumental to the development of the label, including the technical specifications for its implementation. One of the specifications should include the procedure for the publication of the label as part of the meta-data describing the dataset. Health Data Access Bodies will have to specify the type of assessment procedure required in the evaluation of quality and utility. Data quality management experiences recommend a data holders self-assessment methodology for the initial phase of maturity and external audit and certification for the rest of the levels of maturity, including upgrade. Finally, Horizon Europe has planned a CSA meant the development of a data quality and utility label for the HealthData@EU. We recommend the consortium for this CSA to take into account and build on the findings of this report.

Recommendation 9. In HealthData@EU, Health Data Access Bodies are expected to publish and maintain a metadata catalogue of all the datasets made public by the data holders under their purview. Those catalogues should be standardised by defining a Health DCAT profile specification. In addition, the publication of the information on the datasets required in articles 55, 56, and 58 in the Regulation on the EHDS proposal should be systematic.

Recommendation 10. Data holders should implement data management procedures to allow datasets linkage and linkage IDs persistence. In the case of sensitive data, those individual IDs should be pseudonymised and persisted across datasets and overtime. Likewise, data holders should implement procedures before dataset delivery to allow the enrichment of the dataset out of the research outputs (article 37(p) of the current version of the Regulation on the EHDS legislative proposal).

Recommendation 11. The application of privacy enhancement technologies in the pre-analytical processing at SPE level, should not put at stake the utility of the data wherever there is a need for the use of non-anonymised data. *De facto*, the use of a permit application with a research protocol and a data management plan compliant with the minimisation principle, and the use of pseudonymised data within an SPE have been found effective in reducing data privacy risks while maintaining the value of the data.

Recommendation 12. In the context of HealthData@EU, data users should be incentivised to provide feedback on the quality and utility of the datasets delivered to them. To make this possible, health data access bodies should enable a feedback procedure. The development of article 55 in the EHDS Regulation should include the technical specifications for the implementation and governance of a feedback procedure.

Recommendation 13. When providing access, data users have to be advised on the need of the return of the research outputs in a way that datasets can be enriched and digital objects (e.g., data models, annotations, algorithms) can be reused. Research outputs should then be reproducible and interoperable. Health Data Access Bodies have to implement a specific procedure, as part of the application process, SPEs have to implement a specific procedure for the acceptance and eventual inclusion of digital objects, and Data holders have to implement a specific procedure for the inclusion of enriched datasets.

Annex 1 - Definitions in article 55

The source of the dataset is defined both by the data holder (i.e., creator as in data steward or publisher as in the organisation making the data available) of each dataset **and further specifying the data provenance** (i.e., in terms of the origin of the health data *as in article 33, section 3 - EHR, human genetic, genomic and proteomic data, person-generated data, health data registries, medical registries, clinical trials, administrative health information systems, etc.*)

The scope of the dataset is defined as in article 33 describing the minimum categories of electronic health data for which a data holder is obliged to make available for secondary use: a) EHRs; b) data impacting health, including social, environmental, behavioural, determinants of health; c) relevant pathogen genomic data, impacting on human health; d) health-related administrative data, including claims and reimbursement data; e) human genetic, genomic and proteomic data; f) person-generated electronic health data, including medical devices, wellness applications or other digital health applications; g) identification data related to health professionals involved in the treatment of a natural person; h) population-wide health data registries (public health registries); i) electronic health data from medical registries for specific diseases; j) electronic health data from clinical trials; k) electronic health data from medical devices and from registries for medicinal products and medical devices; l) research cohorts, questionnaires and surveys related to health; m) electronic health data from biobanks and dedicated databases; n) electronic data related to insurance status, professional status, education, lifestyle, wellness and behaviour data relevant to health; and o) electronic health data containing various improvements such as correction, annotation, enrichment received by the data holder following a processing based on a data permit.

Nature of the dataset is defined as per the level of sensitivity in GDPR. The General Data Protection Regulation (GDPR) includes the requirement for the data holders to classify their data depending on their level of sensitivity, as high, medium and low level - considering both privacy issues in terms of personal vs non-personal data, requirement of consent from a natural person for their use, and subjected to intellectual property rights and trade secrets. Each sensitivity level is rated according to the potential impact that data may have for an individual if confidentiality and privacy were breached.

This translates into different levels of restriction that have to be reflected as part of the data holders' data security and privacy policy. Thus,

- High sensitivity: reserved for data that may produce a major impact in the life of an individual, such as personal data (i.e., some health data, such as certain medical diagnoses, genetic data, etc.). In this case, a data breach would likely cause harm to both the individual and the organisation hosting the data, so it should be processed and maintained within strict cybersecurity controls. This data should also have strict authorisation controls, auditing procedures to detect access requests, as well as encryption mechanisms applied to data storage and transfer.

- Medium sensitivity: characterising data that would not likely harm individuals, but it still is considered sensitive information that may describe operational details (i.e., medical appointments, surgical history). These files could be deemed medium sensitive.
- Low sensitivity: Data intended for public consumption or open publication (i.e., health statistics, information on healthcare resources, etc.) could be considered low sensitivity and would not need any strict control.

Main characteristics of a dataset are defined as those providing relevant insight about the quality of the available data to assess their utility for a certain purpose within the scope of the secondary use of health data (i.e., research, regulation and policy information) as in article 56 on data quality and utility label. In this regard article 56 introduces some elements as mandatory characteristics to be informed by the data holders for the data they should make available. Those characteristics are expected to be reported at dataset level, and can be broadly classified within the data documentation as part of the metadata or referenced by it as part of a quality and utility label including: a) support documentation, such as data model and data dictionary, including information on the used standards, b) technical quality measurements for several data quality dimensions, such as completeness, uniqueness, accuracy, validity, timeliness and consistency; c) information on the provenance and the maturity level of the data quality management processes of the institution stewarding or producing the data, including review and audit processes, and biases examination; d) information on coverage, including population representativeness and follow-up time for the population covered; e) information on possible data enrichment, such as the possibility to enhance a dataset by linking or merging with other datasets; and f) information on data access and provision, such as data latency (i.e., from collection to availability for secondary use), and time from data access application or data request to actual access; and

Conditions for data availability are defined as in article 56(3)(e) information on data access and provision, such as licence, data access application process, available data services, data latency (i.e., from collection to availability for secondary use) and time from data access application or data request to actual access.

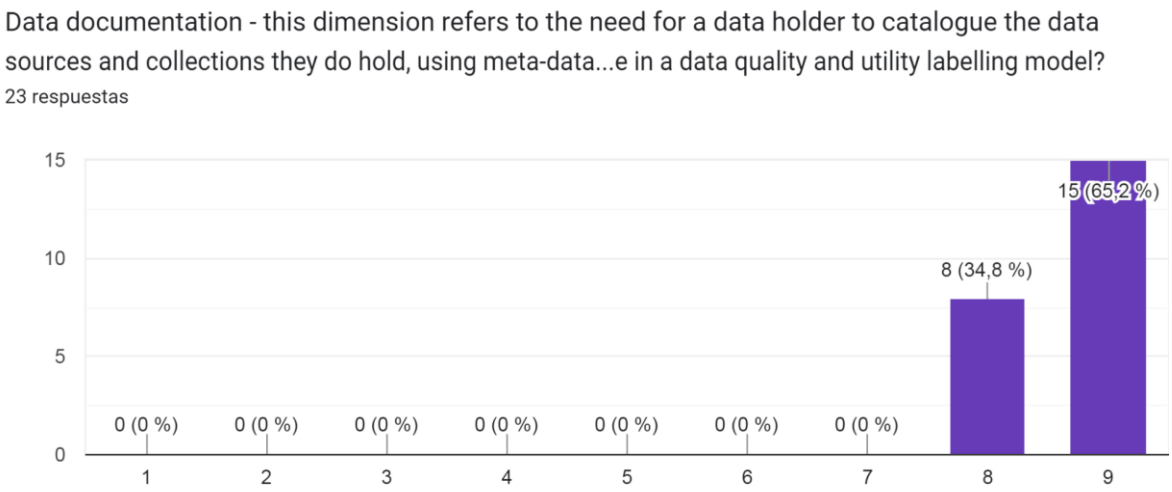
Annex 2 - Data Quality and Utility Labelling: survey results

2.1 Methodology

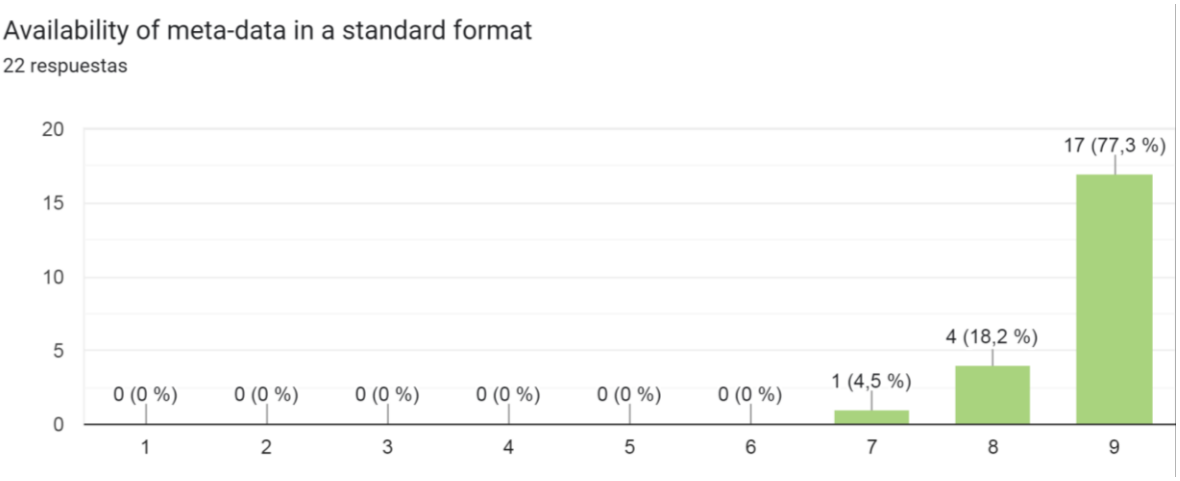
An exploratory two-round survey was conducted to define the level of agreement of WP6 participants with the categories of quality and utility, and specific dimensions, as shown in article 56, of the European Health Data Space regulation. A Likert scale (1 to 9) allows to measure level of importance and level of agreement on a particular category and, within each category, on each specific domain. Votes between 7 and 9 suggest high importance of the category and dimensions when labelling a dataset; voting 4 to 6 would suggest neutrality, while voting 1 to 3 would suggest lack of relevance. When it comes to the agreement, when 75% of the votes are within one of these 3 ranges of values, we may consider there is an agreement on the level of importance of a particular category or dimension.

2.2 Results

Category 1 figure

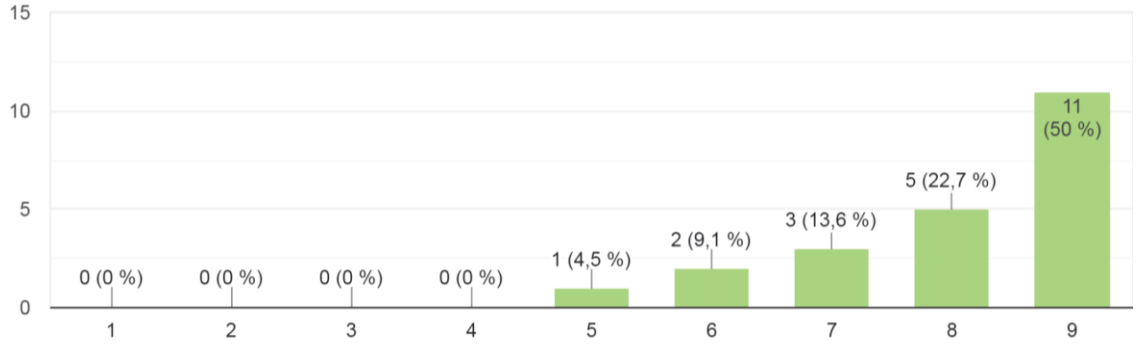


Dimensions in category 1 figures



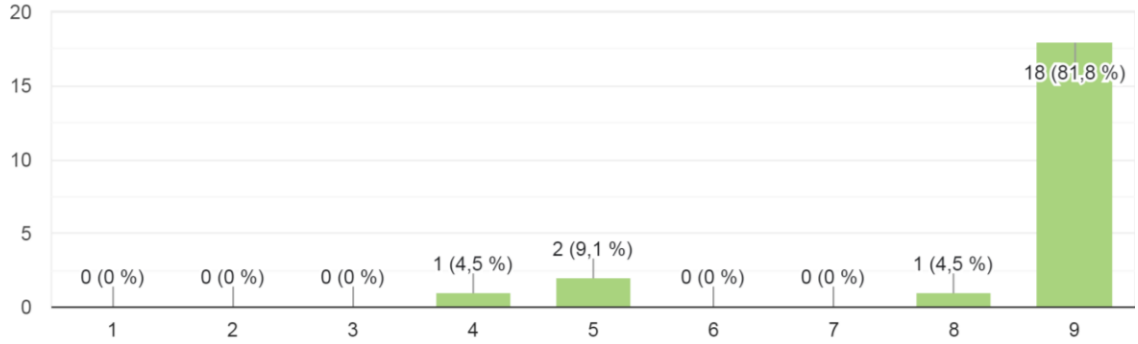
Availability of a clear and documented data model

22 respuestas



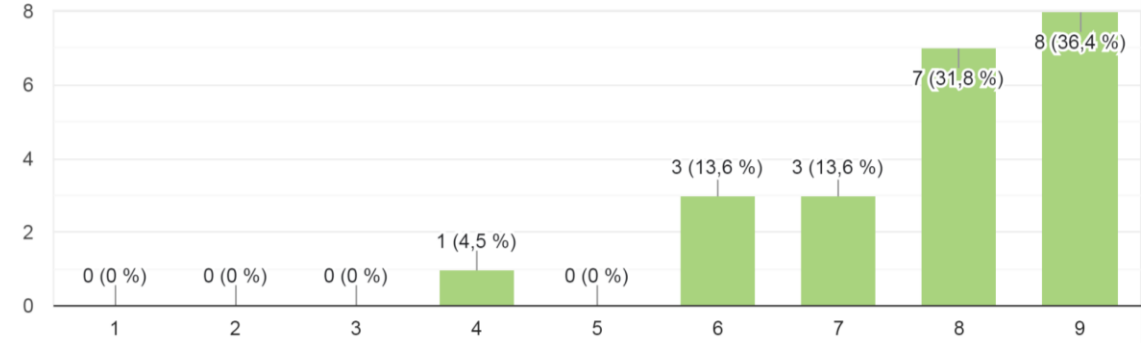
Documented data dictionaries and terminologies

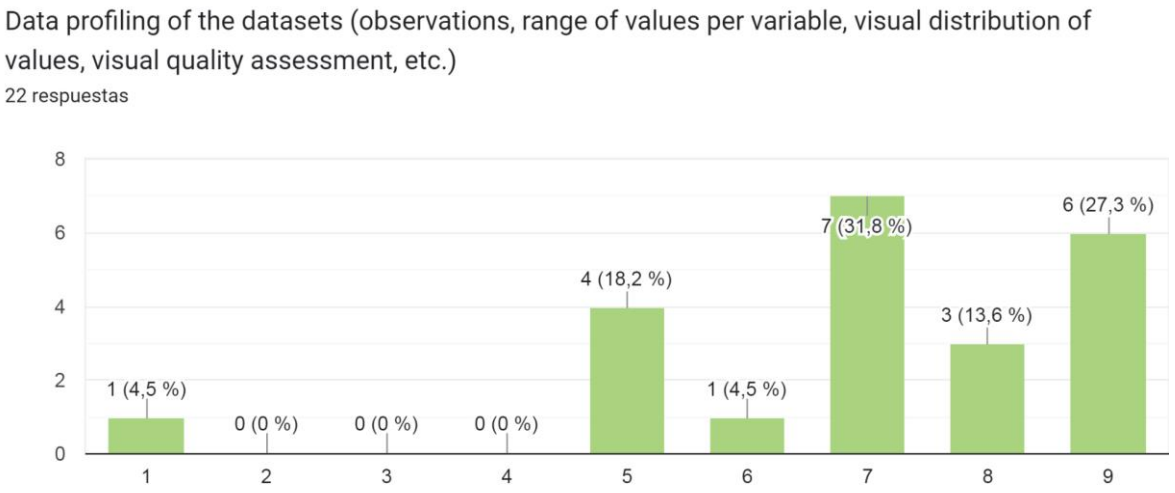
22 respuestas

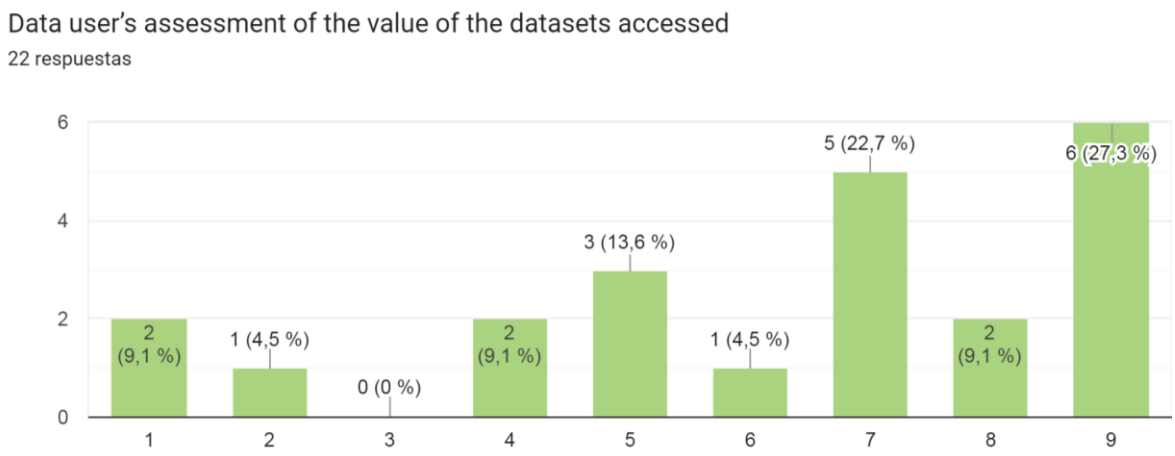


Clear description of source and history of the dataset preparation, (ie transparent data provenance pipeline)

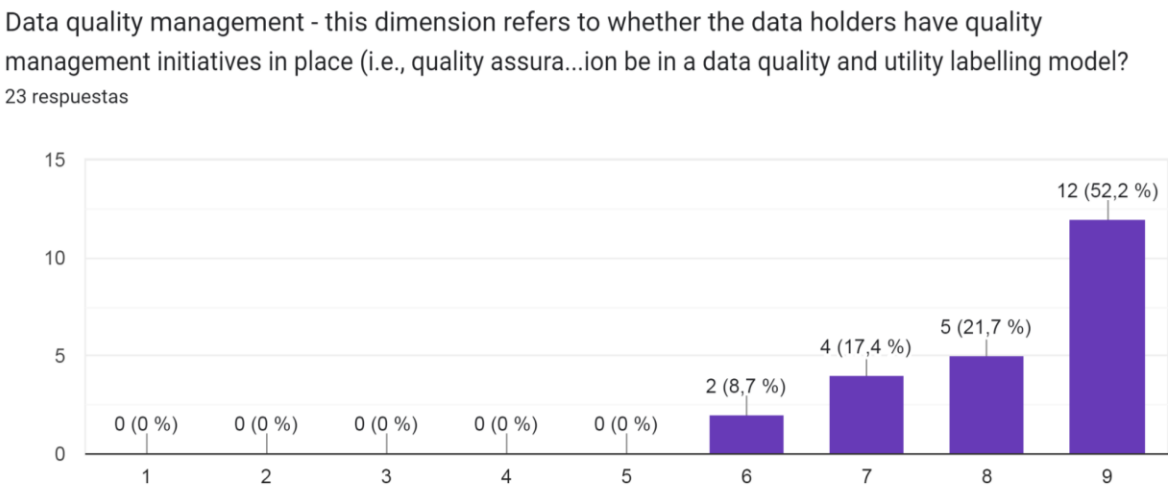
22 respuestas







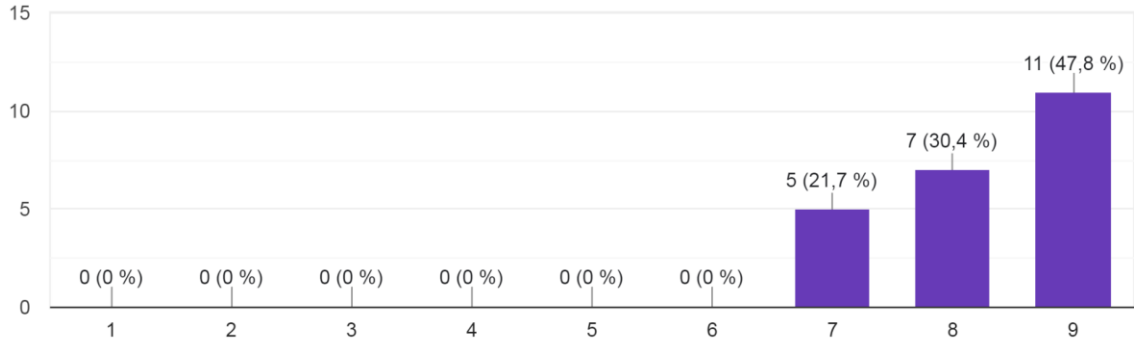
Category 2 figure



Category 3 figure

Technical quality - this dimension refers to whether the data holder carries out formal assessments on technical aspects of data quality at dataset level...ion be in a data quality and utility labelling model?

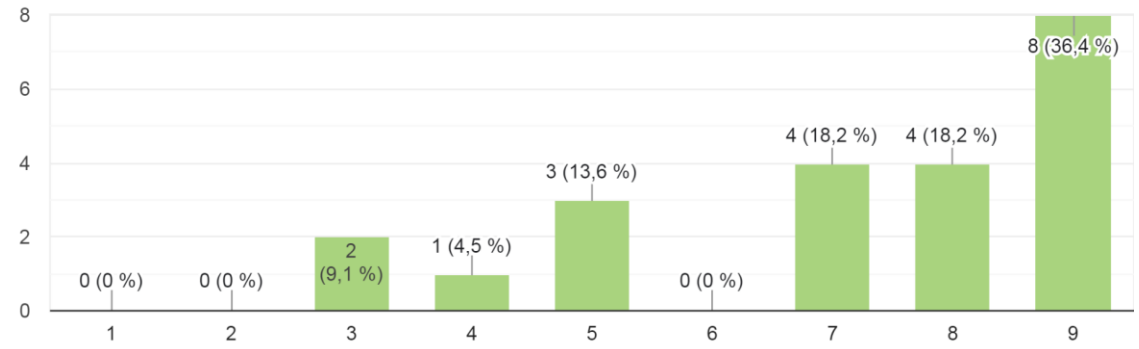
23 respuestas



Dimensions in category 3 figures

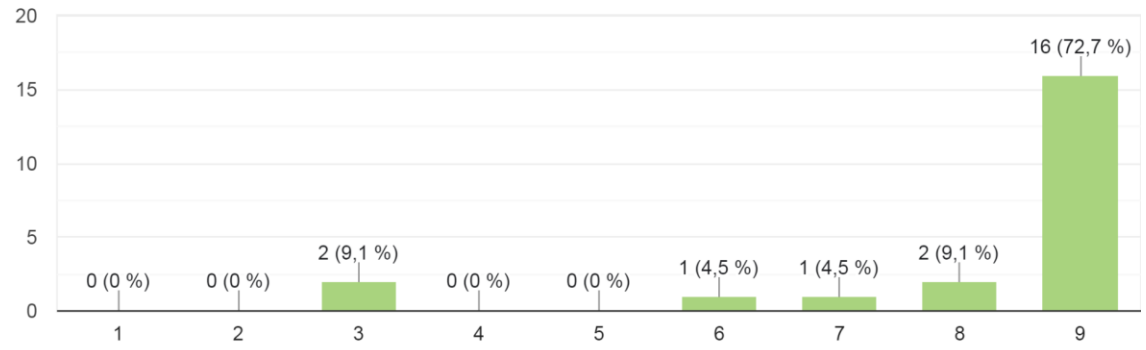
Relevance: Does data meet the users' needs?

22 respuestas



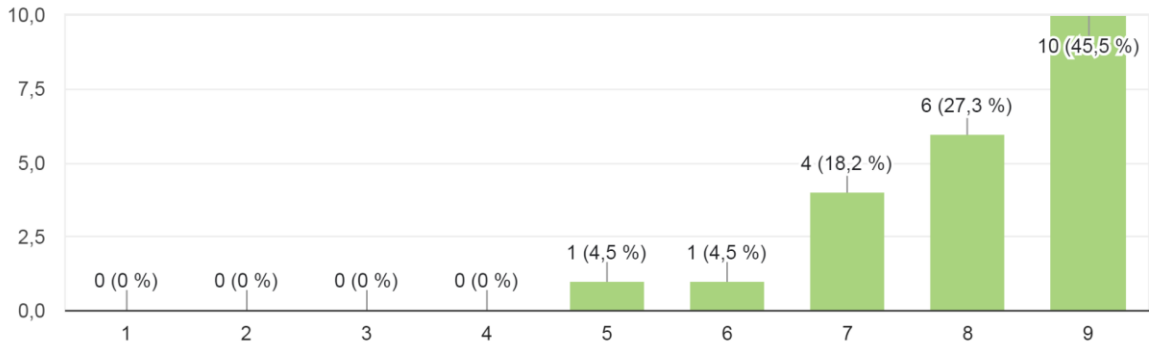
Accuracy and reliability: data reflects accurately and reliably the reality it aims to depict?

22 respuestas



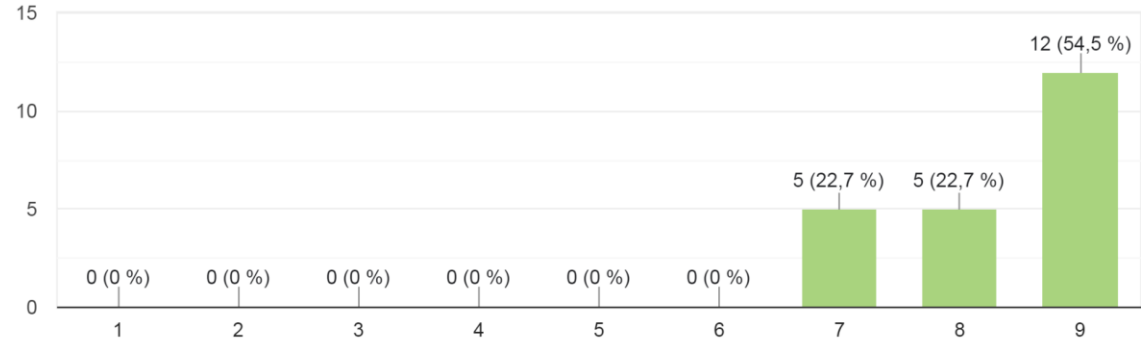
Completeness: at variable level, what is the percentage of missing values in key variables?

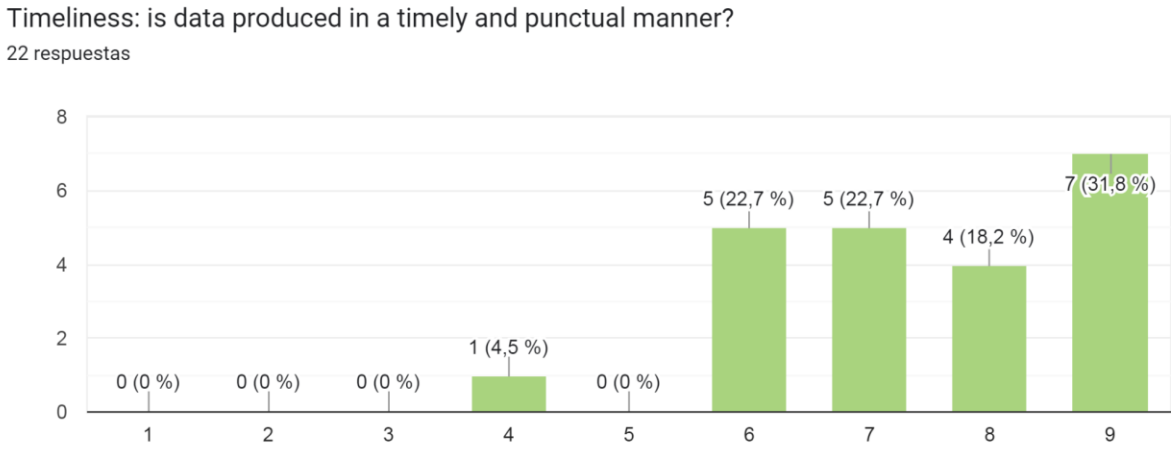
22 respuestas



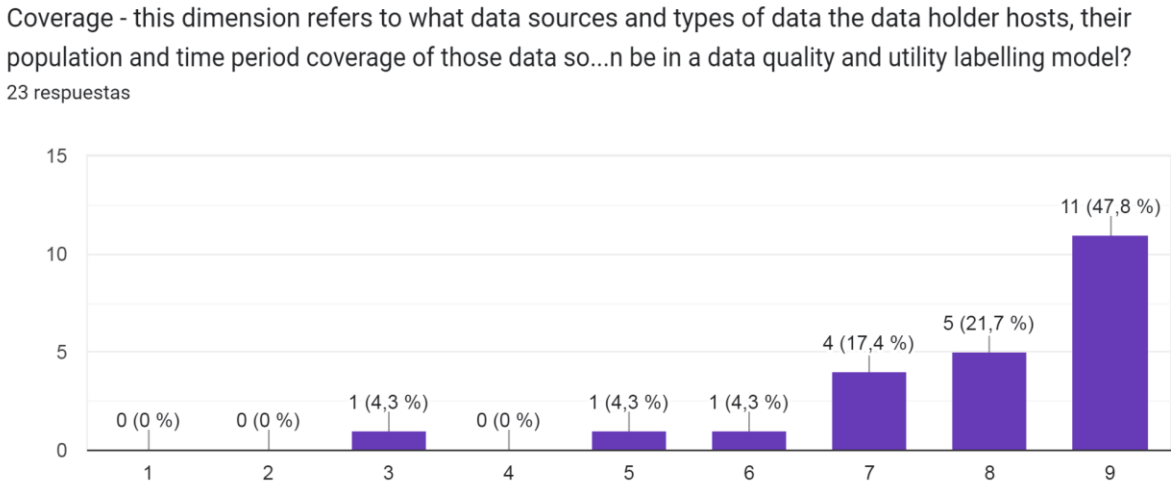
Coherence: data is consistent internally, across data sources, over time and it is comparable between regions and countries

22 respuestas





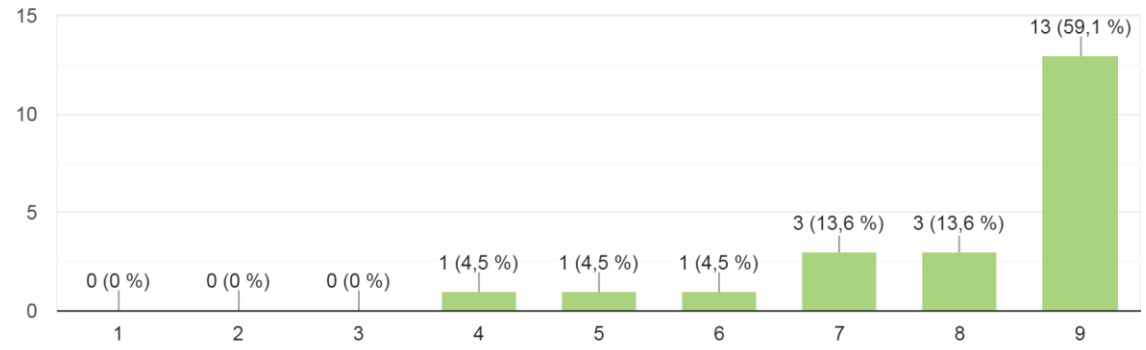
Category 4 figure



Dimensions in category 4 figures

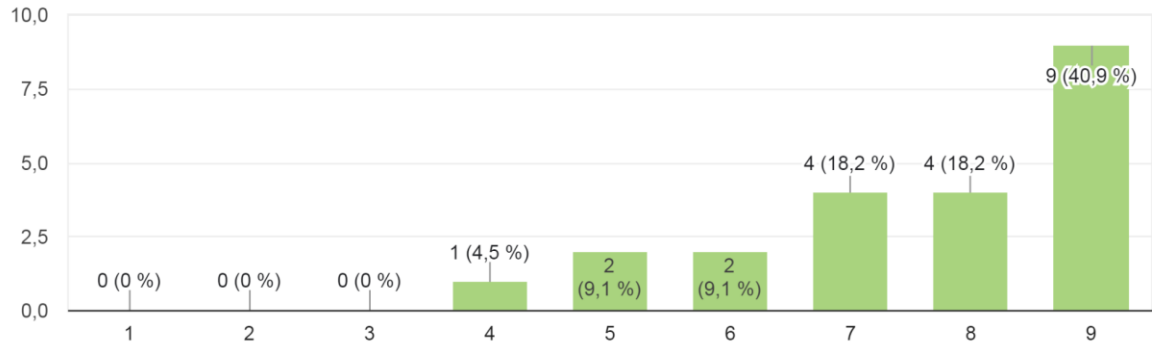
Population included in the dataset (eg, all population, sample)

22 respuestas



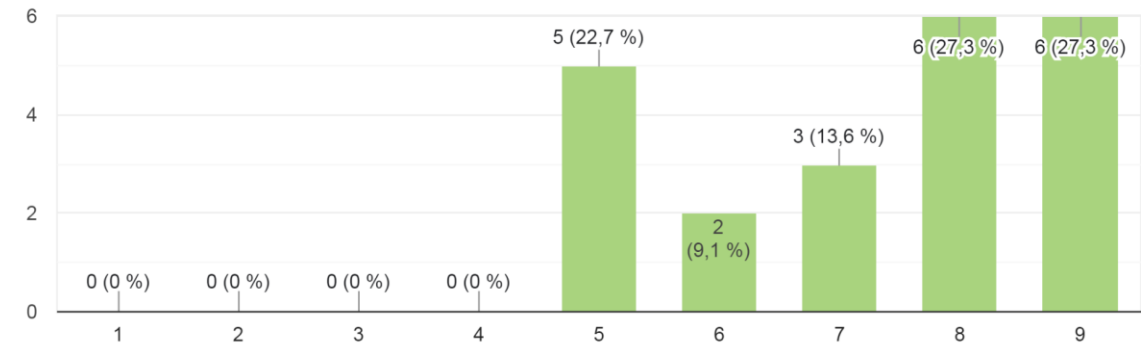
Time-span covered in the data set

22 respuestas

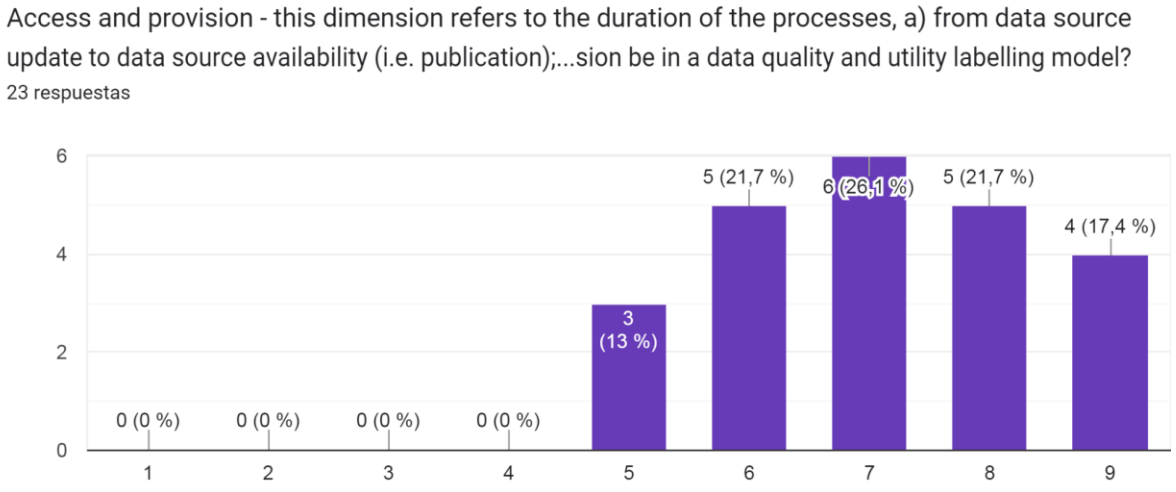


Variety of data sources and data types included

22 respuestas



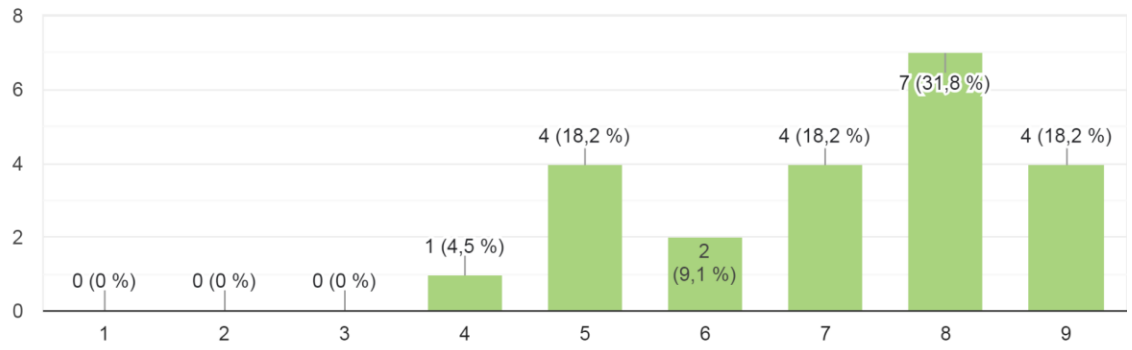
Category 5 figure



Dimensions in category 5 figures

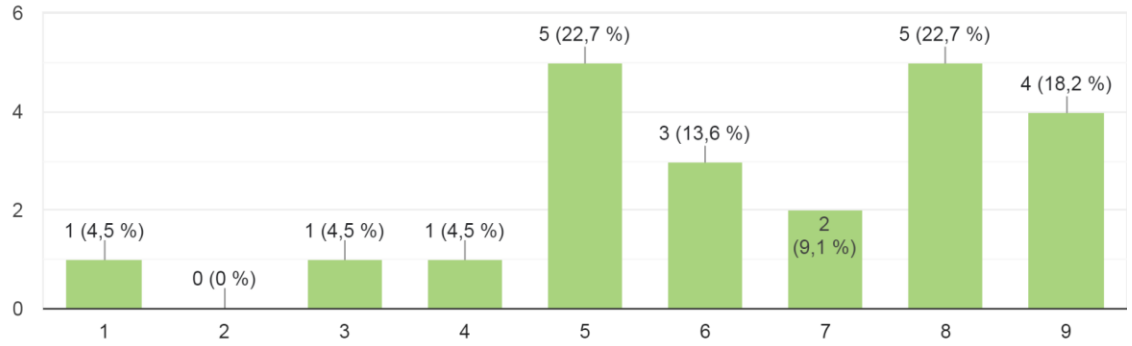
Time-lag between access application and delivery

22 respuestas



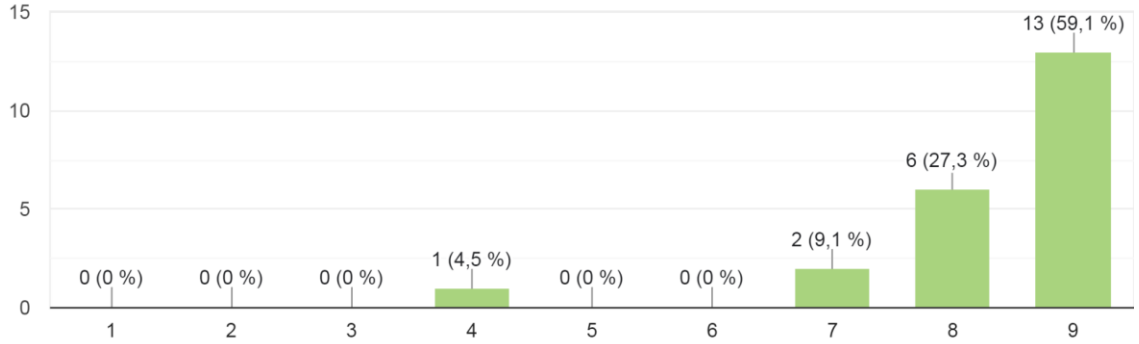
Time-lag until datasets are improved after receiving feedback from the users

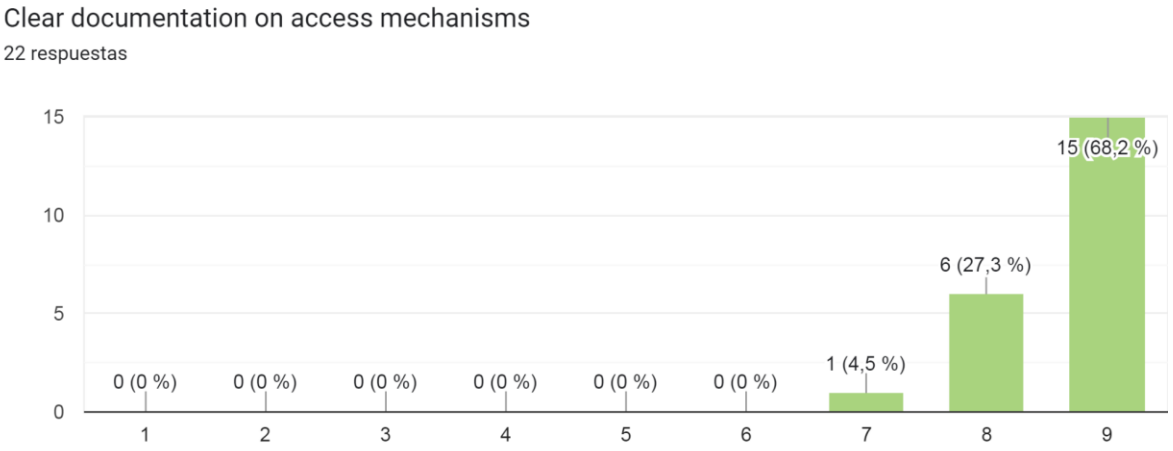
22 respuestas

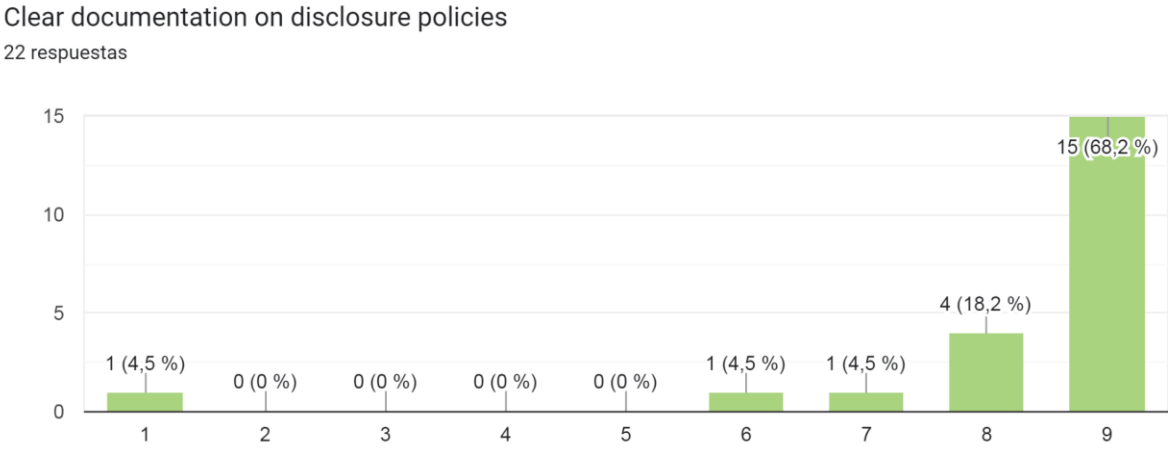


Clear documentation on allowable uses

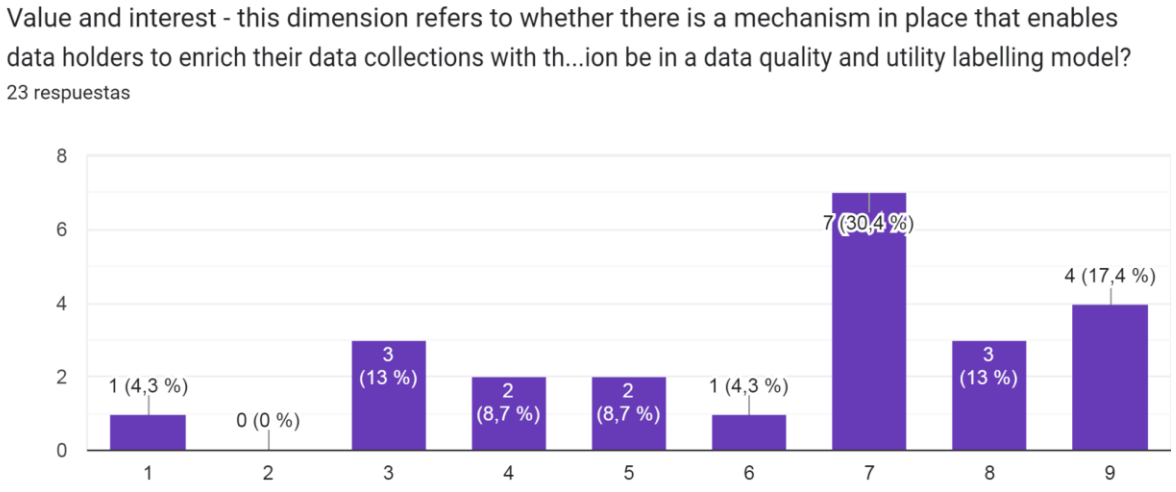
22 respuestas



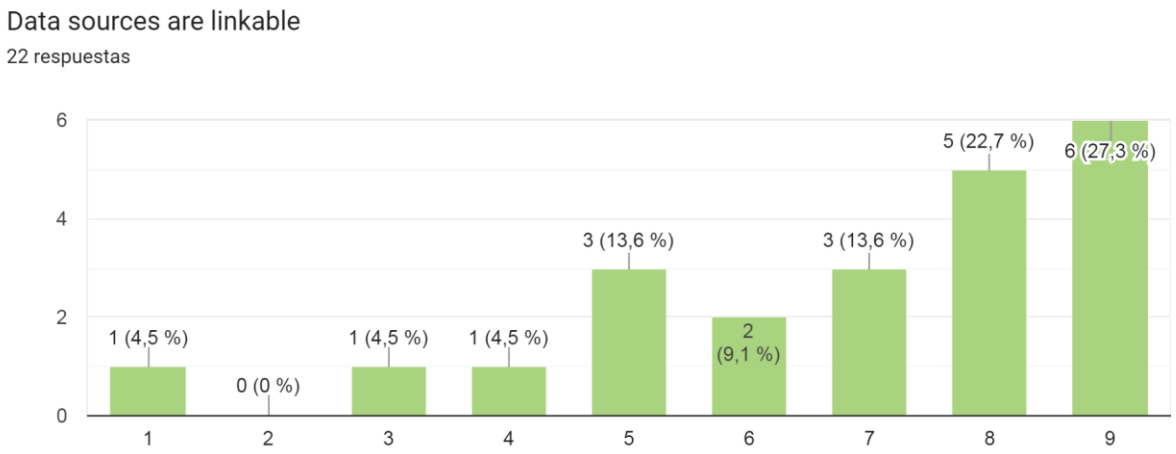




Category 6 figure

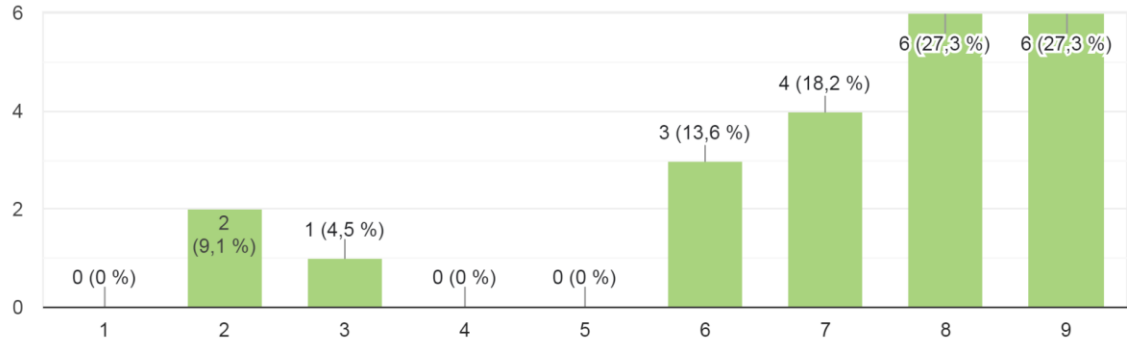


Dimensions in category 6 figures



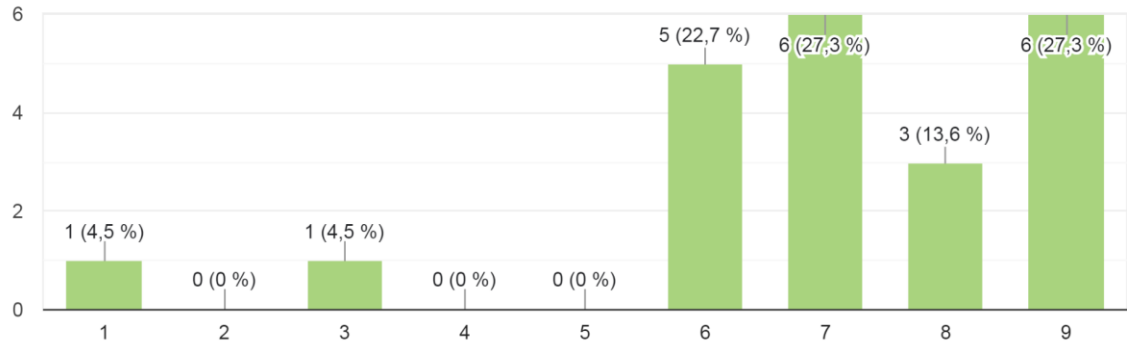
Data sources allow over-time follow up of the units of analysis (individuals, households, geo-areas, etc)

22 respuestas



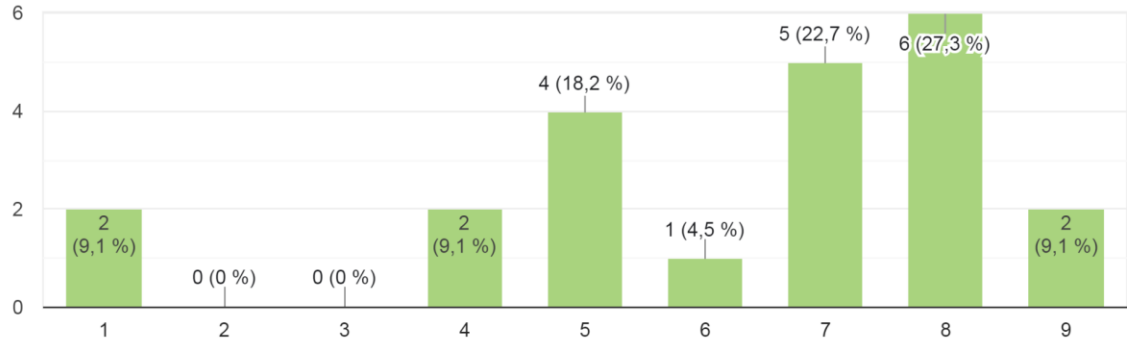
Data quality management includes audit and continuous improvement mechanisms

22 respuestas



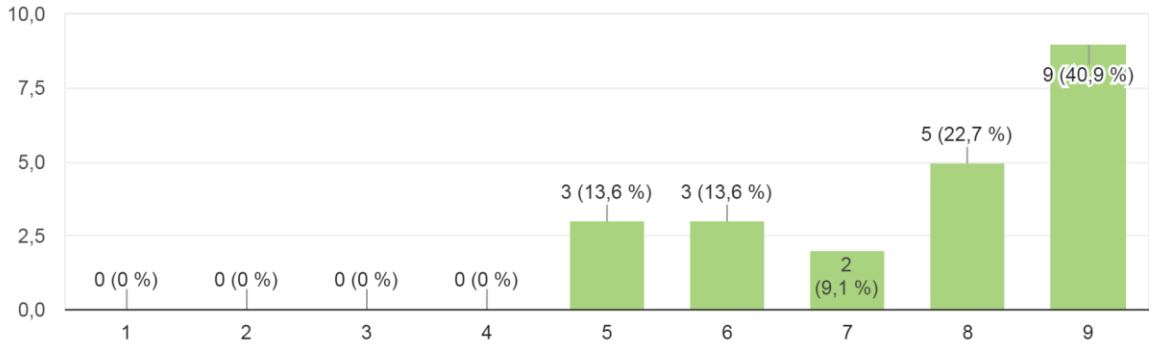
There is a mechanism to enrich data sources with outputs derived from the projects built on them (annotations, data models,

22 respuestas



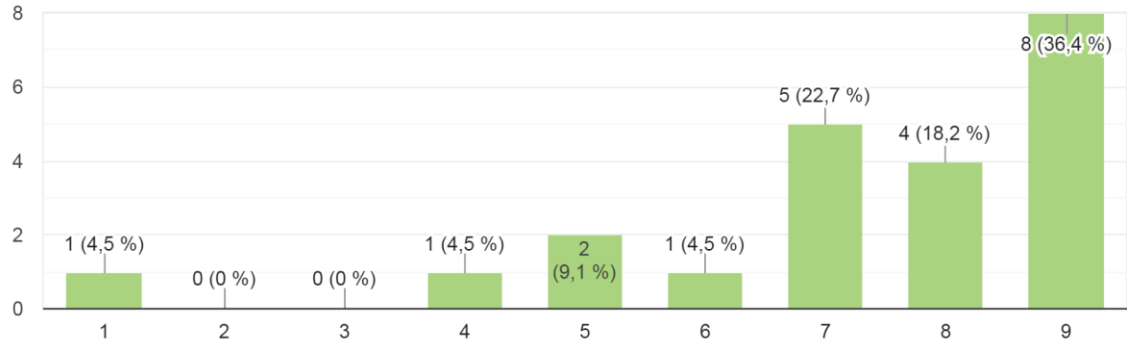
Data sources have been mapped to international interoperability standards

22 respuestas



Data holders have implemented a standard data model

22 respuestas



Annex 3 - Minimum data requirements in cross border registries

3.1 Methodology

The term 'cross-border datasets' in article 58 in the EHDS regulatory proposal was interpreted as EU registries. A scoping review on EU-wide cross-border registries was performed specifically looking for those EU registries that a) were cross-border in nature (i.e., centralising health data from multiple EU countries and potentially others), b) with international coverage and scope, c) set up for research, d) with openly published documentation on their data model structure and health information collected. The scoping review aimed to identify key examples of cross-border registries currently operating in the EU and map their commonalities in terms of scope and health data collected to build a common minimum health information requirement for a cross-border registry within HealthData@EU.

Below is a list of the cross-border registries reviewed with referenced links to their available documentation.

*MyHealth@EU patient summary data model standard was used as main reference of the minimum health data requirement for cross-border healthcare from which elicit additional data requirements for potential cross-border registries within HealthData@EU.

3.2 Cross-border registries reviewed

- MyHealth@EU (patient summary documentation)*
- Joint Action cross-border PAtient REgistries iNiTiative (PARENT) methodological guidelines
- Rare diseases
    - European platform of Registries on Rare Diseases (EURD) (documentation)
    - The European Platform for Rare Disease Registries (EPIRARE)(article)
    - RD-Connect project
    - European Union Committee of Experts on Rare Diseases (EUCERD)(documentation, article)
    - Open-Source Registry System for Rare Diseases in the EU (OSSE project)
- Cardiovascular diseases
    - European Observational Research Programme (EORP) by the European Society of Cardiology (ESC) Registries
        - Atrial fibrillation III (AF III) Registry
        - NSTEMI Registry (article)
        - Other cardiovascular registries within EORP-ESC
    - EuroHEART data standards
- Cancer
    - European Cancer Information System (ECIS)
    - European Network of Cancer Registries (ENCR) recommendations
- European Medicines Agency (EMA) guidelines on registry-based studies
- Health Surveys
    - Health Interview Survey (HIS)
    - Health Examination Survey (HES)

Annex 4 - Results of the survey voting the final recommendations

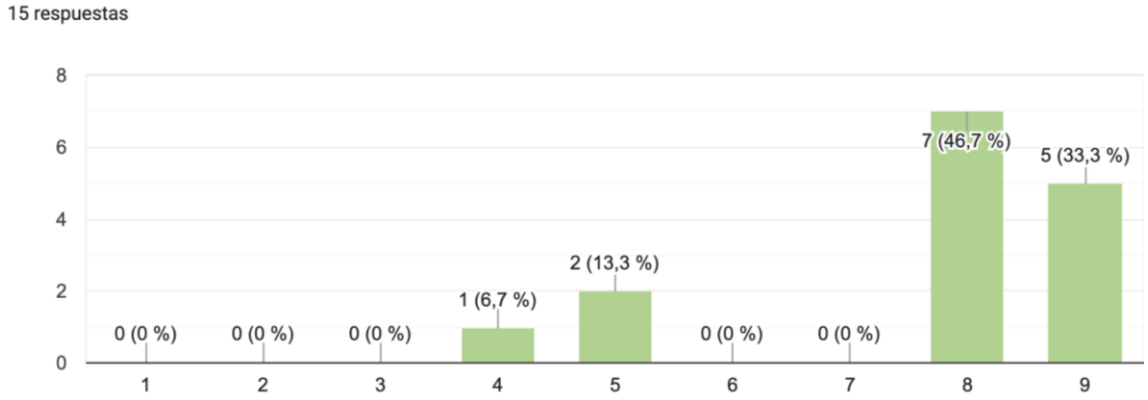
4.1 Methodology

Participants were asked to assess the 13 recommendations that compose the TEHDAS Data Quality Framework. Specifically, you have to respond to this question **"how relevant is it for you to provide the European Commission and the Member States with this recommendation?"**

For each recommendation - votes with a value of 7 and more were interpreted as strong support; votes between 4 and 6 were interpreted as neutral; votes equalling 3 or less were interpreted as no support. In addition, the concentration of votes within each range of votes was assessed in terms of agreement - at least 10 (out of 15) votes within the range were interpreted as high agreement on favouring, being neutral or opposing the recommendation.

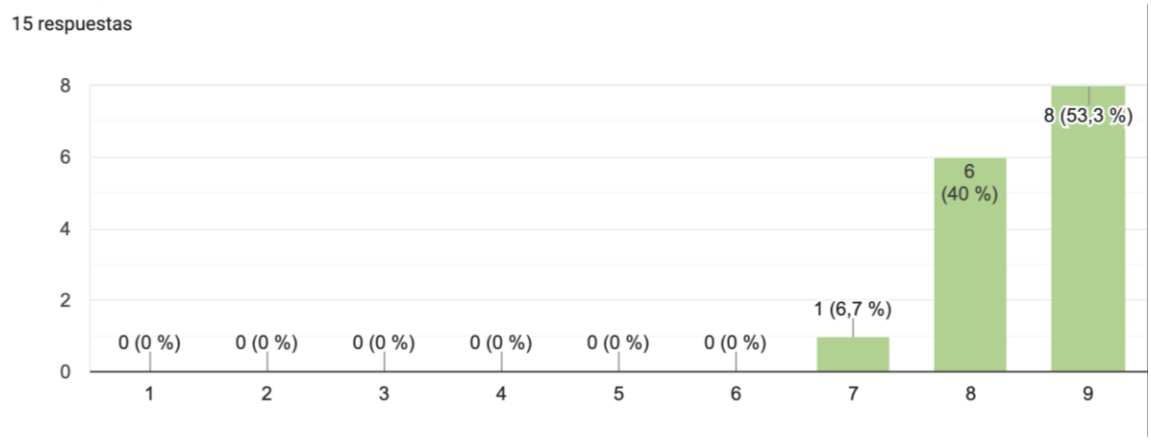
4.2 Results

Recommendation 1. A HealthData@EU DQF should include not just the technical quality of data but also the utility of datasets, with a view of fostering a fit-for-purpose approach.



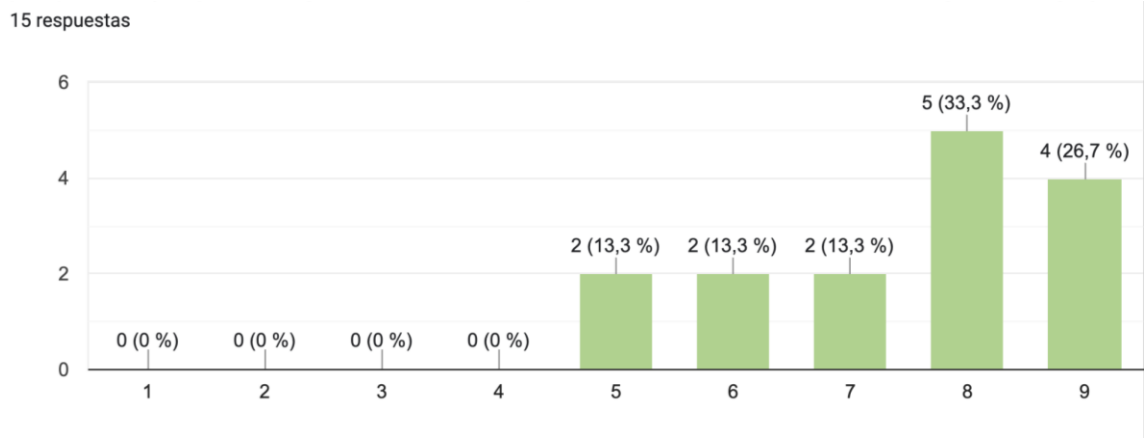
***12 out of the 15 responses were above 7**, denoting high agreement in the support of this recommendation. Among the comments, the implementation of this recommendation requires a consensual definition of utility.*

Recommendation 2. A HealthData@EU DQF should include as main data quality features relevance, accuracy and reliability, and coherence; likewise, as main utility features coverage, completeness, and timeliness.



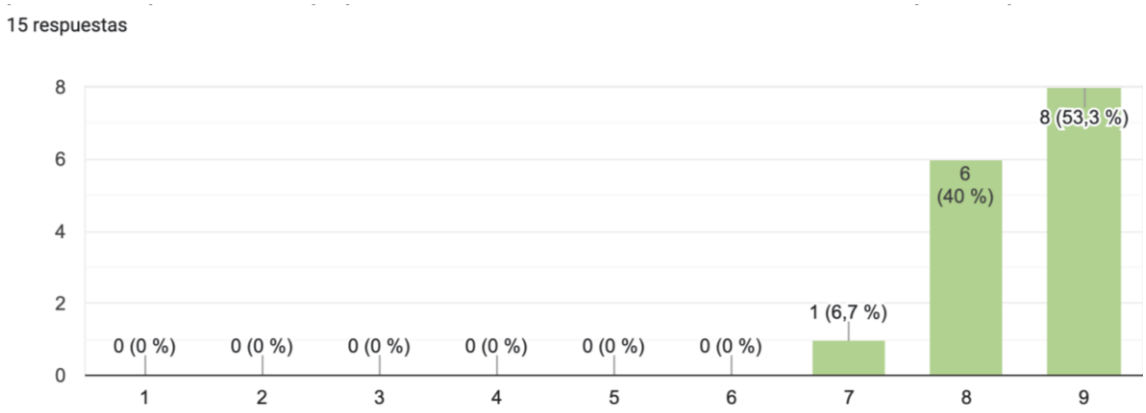
15 out of 15 votes were equal to more than 7, *denoting a high agreement on recommending the adoption of the TEHDAS definition of quality and utility.*

Recommendation 3. A HealthData@EU DQF should also take into account the data holders' perspective by implementing actions towards improving their maturity in data collection, curation, storage and staging.



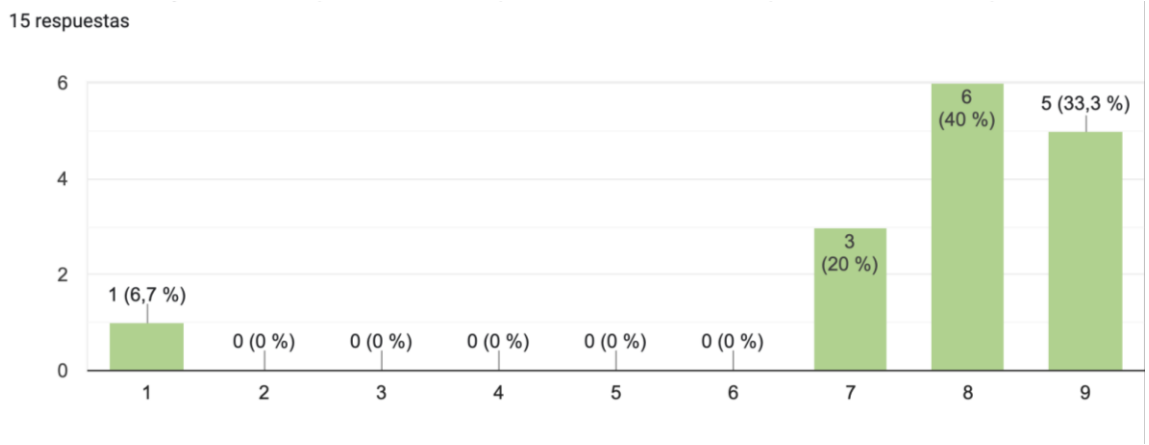
This recommendation was highly supported (11 out of 15 votes) *although the level of agreement was smaller, ranging votes between 5 and 9, maybe denoting uncertainties on the implementation.*

Recommendation 4. A HealthData@EU DQF should be applied along the whole Data life cycle with particular emphasis on data preparation at data holder level, at the dataset publication and discovery phase, when preprocessing the data before delivery, and when enriching the datasets, procedures and tools once research outputs are provided.



All votes casted are strongly supporting this recommendation. *No specific comments nuancing the recommendation were referred*

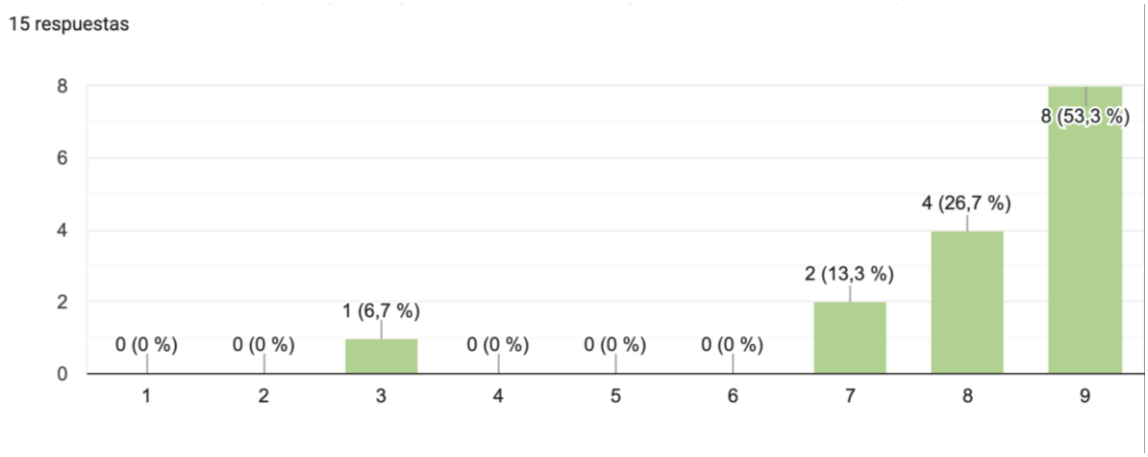
Recommendation 5. There is a need for a dedicated plan aiming the implementation of a data holders maturity model to improve their data quality management and quality assurance procedures, and reduce gaps across HealthData@EU data holders. All the data holders should be evaluated according to the levels of maturity established in such a model and an agreed notion of their maturity should be included as part of the meta-data of their datasets when made available. Health Data Access Bodies would specify the type of assessment procedure required in the evaluation of maturity; in this respect, data quality management experiences recommend a data holders self-assessment methodology for the initial phase of maturity and external audit and certification for the rest of the levels of maturity. Finally, the implementation of the maturity model should be progressive, and foster incentives for continuous improvement and level promotion.



14 out of 15 votes strongly supported this recommendation*, with two additional reflections on the need for a definition of maturity that is consensual and the need for the assessment of costs that inevitably data holders will incur.*

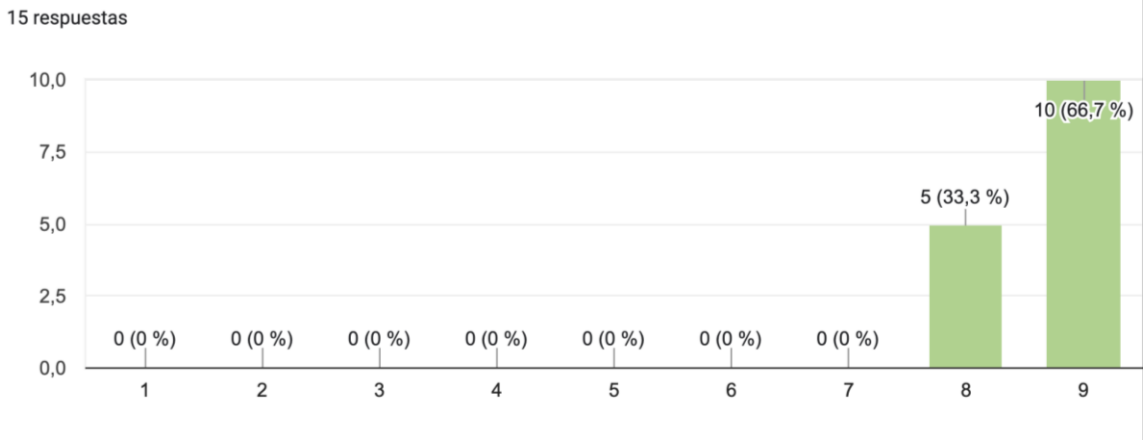
Recommendation 6. In HealthData@EU, there is a need for data holders to implement a layer of semantic interoperability using widely adopted standards (see recommendations 5 to 8 in deliverable 6.2). As a preferred framework, in the short run, data holders should follow

an incremental approach to progressively map their regular controlled vocabularies to international general and domain-specific ontologies. The European Commision should support continuous dialogue on this governance mechanism, taking as an inspiration how the initiative fostering OMOP-CDM has addressed openness, transparency, technological neutrality, data portability, and cooperation among public institutions.



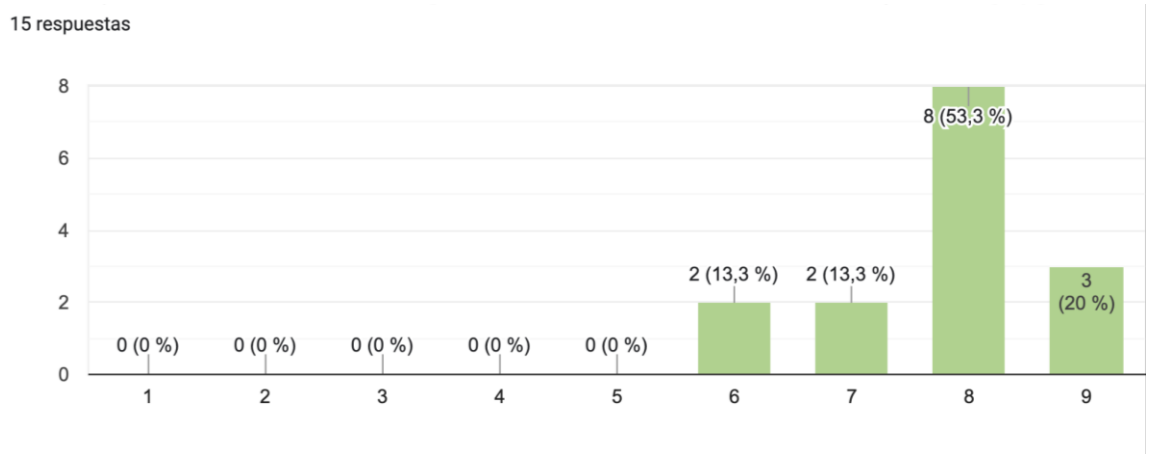
14 out 15 participants showed strong agreement with this recommendation. *No specific comments to highlight.*

Recommendation 7. In HealthData@EU, data holders are expected to publish information on their datasets (article 41 of the current version of the Regulation on the EHDS legislative proposal) and health data access bodies to catalogue them all (article 55 of the current version of the Regulation on the EHDS legislative proposal). It is recommendable to combine the use of generic meta-data standards and domain-specific meta-data standards in a two-step approach to discoverability; at a first stage, users should know about the source, scope of the datasets, nature of the data, main characteristics and features of distribution; at a second stage, to allow further knowledge on the datasets to allow federated querying (e.g., providing data profiles). As enforced by law in the aforementioned article 55, an implementing act should provide the technical specifications for this specific development.



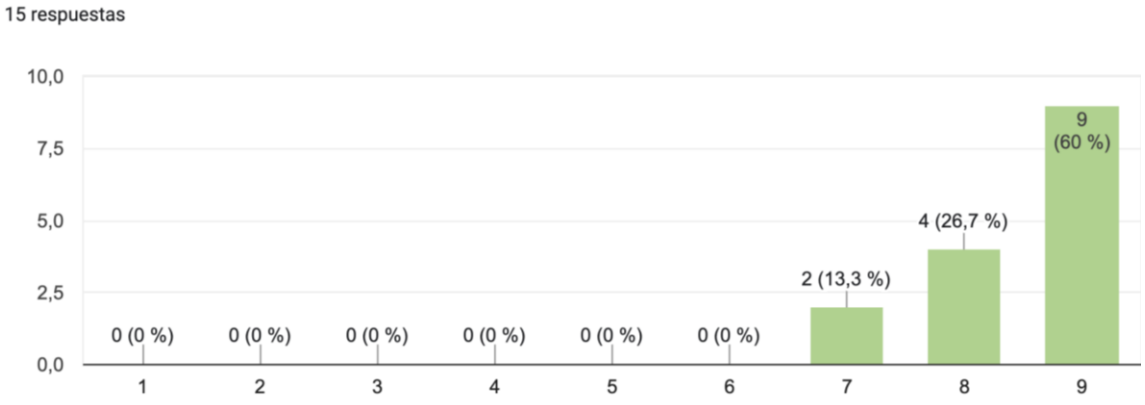
15 out of 15 participants showed strong agreement with this recommendation. *A participant recommended clarity on the definition of a dataset as a critical element to properly fit with the definitions in publication standards.*

Recommendation 8. In HealthData@EU, data holders are expected to publish a notion on the quality and utility of their datasets that are obliged to make available (articles 41 and 55 of the current version of the Regulation on the EHDS legislative proposal). Although there is a general agreement on the main categories that the label should contain, there are some discrepancies in the operational definitions of some dimensions. An implementation act for the implementation of a quality and utility label should stem from a formal consensual exercise for an operational definition of quality and utility that is instrumental to the development of the label, including the technical specifications for its implementation. One of the specifications should include the procedure for the publication of the label as part of the meta-data describing the dataset. Health Data Access Bodies will have to specify the type of assessment procedure required in the evaluation of quality and utility. Data quality management experiences recommend a data holders self-assessment methodology for the initial phase of maturity and external audit and certification for the rest of the levels of maturity, including upgrade.



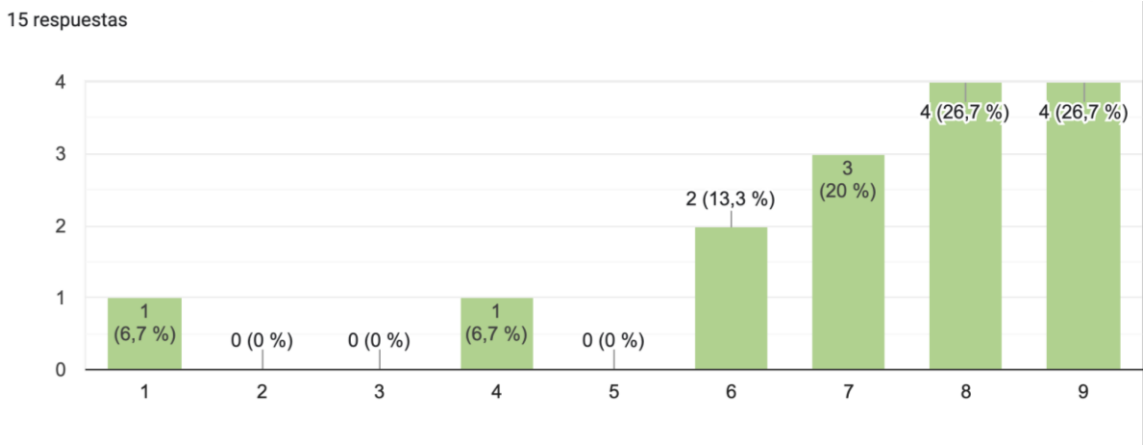
13 out of 15 participants strongly supported this recommendation. *Some notion of the difficulty of implementation is discussed in one comment making advisable the development of a joint roadmap, where guidance on when (if so) to trust in self-assessment and when in external auditing should be provided.*

Recommendation 9. In HealthData@EU, Health Data Access Bodies are expected to publish and maintain a metadata catalogue of all the datasets made public by the data holders under their purview. Those catalogues should be standardised by defining a Health DCAT profile specification. In addition, the publication of the information on the datasets required in articles 55, 56, and 58 in the Regulation on the EHDS proposal should be systematic.



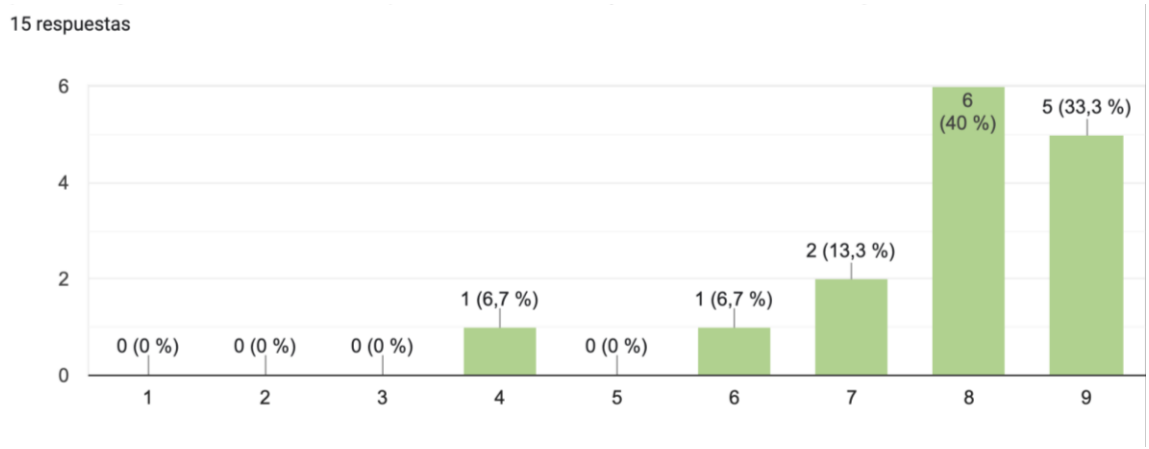
15 out of 15 participants showed strong support for this recommendation. *A comment suggests the importance of timely updates in the catalogues.*

Recommendation 10. Data holders should implement data management procedures to allow datasets linkage and IDs persistence. In the case of sensitive data, those individual IDs should be pseudonymised and persisted across datasets and overtime. Likewise, data holders should implement procedures before dataset delivery to allow the enrichment of the dataset out of the research outputs (article 37(p) of the current version of the Regulation on the EHDS legislative proposal).



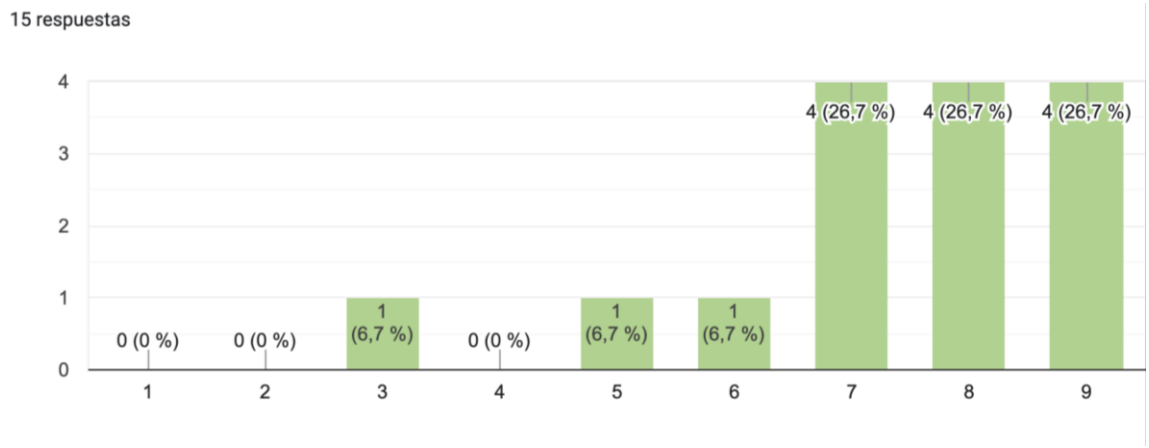
11 out of 15 participants showed strong support for this recommendation. However, a wider range of votes denotes a lower level of agreement. *Comments warn on the difficulties of implementation as well as on the advantages of pseudonyms in research. Further discussion has to be taken maybe not in relation to the impact in quality but the actual policies to implement across HealthData@EU.*

Recommendation 11. The application of privacy enhancement technologies in the pre-analytical processing at SPE level, should not put at stake the utility of the data wherever there is a need for the use of non-anonymised data. *De facto*, the use of a permit application with a research protocol and a data management plan compliant with the minimisation principle, and the use of pseudonymised data within an SPE have been found effective in reducing data privacy risks while maintaining the value of the data.



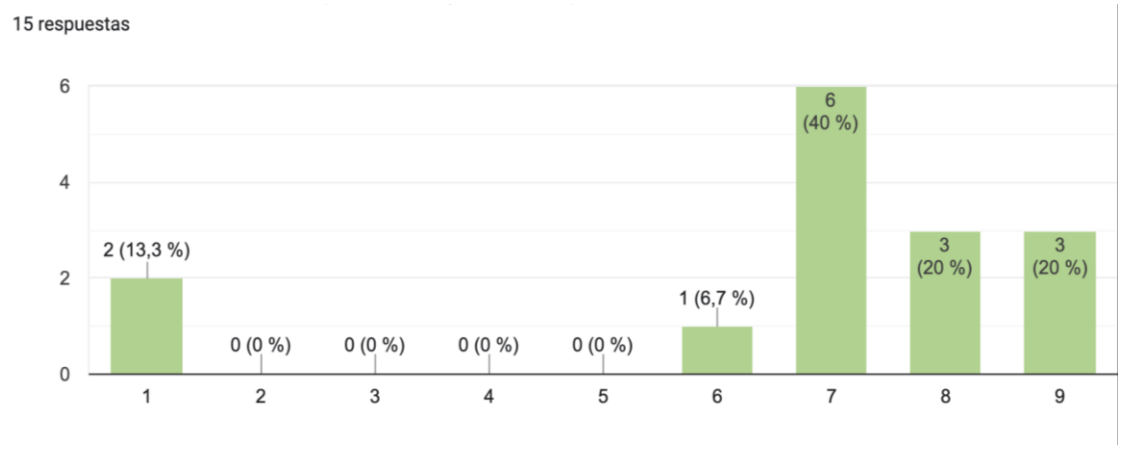
13 out of 15 participants strongly support this recommendation. *One comment raises doubts on whether this recommendation goes straight to the core of data quality.; likewise, a comment on recommendation 10.*

Recommendation 12. In the context of HealthData@EU, data users should be incentivised to provide feedback on the quality and utility of the datasets delivered to them. To make this possible, health data access bodies should enable a feedback procedure. The development of article 55 in the EHDS Regulation should include the technical specifications for the implementation and governance of a feedback procedure.



12 out 15 participants showed strong support for the recommendation*. Albeit a certain level of disagreement (votes casted ranged from 3 to 9) all the comments were very supportive.*

Recommendation 13. When providing access, data users have to be advised on the need of the return of the research outputs in a way that datasets can be enriched and digital objects (e.g., data models, annotations, algorithms) can be reused. Research outputs should then be reproducible and interoperable. Health Data Access Bodies have to implement a specific procedure, as part of the application process, SPEs have to implement a specific procedure for the acceptance and eventual inclusion of digital objects, and Data holders have to implement a specific procedure for the inclusion of enriched datasets.



12 out of 15 participants showed strong support for this recommendation. *However, most of the votes went to 7 denoting certain doubts on the actual definition and also, on difficulties about implementation.*