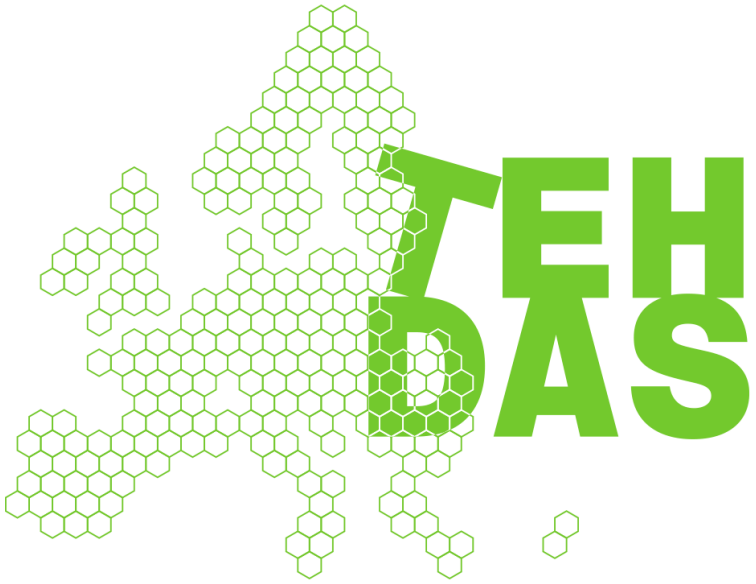


Towards
European
Health
Data
Space

Milestone M7.1

Technical and operational analysis report of existing data sharing and/or secondary use initiatives in health and biomedical sciences

30 September 2021

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

0 Document info

0.1 Authors

Author	Partner
Juan González-García	IACS – Aragon Institute of Health Sciences, Spain
Helena Lodenius	CSC - IT Center for Science, Finland
Cátia Pinto	SPMS - Shared Services of the Ministry of Health, EPE, Portugal
Rémi Quilliet	FR-HDH – Health Data Hub, France
Philipp Schardax	ATNA – Ministry of Health, Austria
Carlos Tellería	IACS - Aragon Institute of Health Sciences, Spain

0.2 Keywords

Keywords	TEHDAS, Joint Action, Health Data, Health Data Space, Data Space, HP-JA-2020-1, Data Hub, Infrastructure

Accepted in Project Steering Group 28 September 2021.

Disclaimer

Copyright Notice

Contents

Executive summary

Task 7.1 in work package 7 (WP7) of TEHDAS aimed to survey several data sharing initiatives on their current practices, addressing the requirements for the deployment of data sharing and secondary use of health and biomedical data. The information gathered via the survey will help further to identify how the technical infrastructure of the European Health Data Space (EHDS) should be developed to support the current and future needs on sharing and accessing health data across the EU.

The resulting technical and analysis report presents an overview of current technical solutions in selected data sharing initiatives and other relevant entities. The analysis report will serve to aid in the design of the services (deliverable D7.1) and in the proposal of architecture and infrastructure options (deliverable D7.2).

The results of the data hub survey conducted show a rather heterogeneous landscape of data, technologies and processes among responding data sharing initiatives. At the same time, some common patterns are observable:

- Electronic Health Records, Prescriptions, Laboratory Results and Surveys represent the most common data sources.

- R and Python are the data analysis tools most commonly provided by the data sharing initiatives to researchers. Roughly half of the responding data sharing initiatives offer search portals for data, which allow searches based on metadata, variables, number records, or statistical information.

- External access to the data is restricted, and logging of user actions is commonly performed. The majority uses local or centralised user authentication (e.g., LDAP), which in turn are mostly password-based – sometimes complemented by IP addresses or other additional factors. The trend among data sharing initiatives is towards allowing data access only in a secured computing environment hosted by the initiative.

- Data and privacy protection are ensured by different sets of security measures. Among these techniques are anonymisations, pseudonymisation, and data aggregation.

- Data is most commonly ingested in flat file format – such as .csv - followed by XML, RDF, JSON, FASTA (genomic data) and DICOM (medical imaging). Mechanisms of quality control include format validation, detection of wrong measurement units, temporal data consistency, and management of outliers and missing data.

- Typical international terminology standards for the attributes included in the data sets are ICD-10 (diagnoses and procedures), SNOMED-CT (clinical terminology) and LOINC (laboratory data). Less frequently, data follows terminologies such as WHO-ATC (drug substances), and other domain specific standards (ICDO3, HemOnc, RxNorm, Orpha, etc.). Only two of the surveyed data sharing initiatives facilitate interoperability by employing HL7-FHIR.

1. Background

WP7 will detail the technical options to provide an effective secondary use of health data through the EHDS. In this way, these solutions represent the technical interoperability elements of the EHDS, according to the European Interoperability Framework. The infrastructure technologies proposed for EHDS should be based on global standards where appropriate as this ensures maximum interoperability both within Europe and worldwide. The Global Alliance for Genomics and Health (GA4GH), for example, sets the technical standards and frames the policy for the responsible sharing of human genomic data within a human rights framework. Many pilots are using this framework as a way to innovate in the use of sensitive data while enhancing privacy and security. Progress has also been made in healthcare with improving data interoperability. A number of clinical data standards and common data models have been developed to provide a normalised representation of EHR data, including HL7 FHIR and OMOP Common Data Model.

In this report we document the technical requirements, functionalities and standards used by different European data sharing initiatives. Each initiative can be seen as a 'potential approach' to serve as a practical example to provide information and inspiration on what can be done.

The EHDS requires a secure and interoperable infrastructure with well-defined services to facilitate the consistent and secure secondary use of health data. The key concept for the envisaged EHDS is a federated peer-to-peer system, that interconnect nodes of different typologies (data permit authorities, European level health data providers, research infrastructures, data altruism bodies etc.). The EHDS nodes will operate regularly in an independent manner, providing services to their users, but as part of the federation they will be connected. The interconnection will be implemented through a series of specific EHDS services deployed on to provide federation-wide features, such as: announcement, authentication, data discovery, data retrieval, data analytics, etc.

To define services and architecture options for the EHDS, establishing a common understanding of the technical solutions of existing data sharing initiatives is crucial. The objective of Task 7.1 is to inquire and gather existing knowledge on public and/or private initiatives on data sharing and the secondary use of health and biomedical data, focusing on the requirements for their deployment. This report will help identify the most promising approaches, and serve to aid in the proposal of architecture and infrastructure options (deliverable D7.2), as well as the design of the services (deliverable D7.1).

2. Methodology

2.1. Survey

The findings presented in this report are based on analysis of the survey responses collected from the selected data sharing initiatives.

In order to identify the appropriate data sharing initiatives to be analysed, the criteria for the data sharing initiatives were defined. The selection criteria were to ensure that a wide typology of the data sharing initiatives would be covered (national data hubs or permit authorities, data sharing repositories from health services, European Reference Networks, and other regional, national, or supra-national institutions or initiatives of special interest, etc.). Furthermore, the organisations should represent different data domains, including omics repositories, biobanks,

clinical data repositories etc. No more than 15 data sharing initiatives were expected to be analysed.

To meet the most important criteria, i.e., diversity, every partner in WP7 was requested to suggest four data sharing initiatives that are relevant to their countries. Each suggested data sharing initiative should represent different categories (public / private / national / regional / European), as well as different data domains. In addition to the suggestions by the WP7 partners, a few data sharing initiatives were also selected from the TEHDAS stakeholders list.

The survey was distributed among various data sharing initiatives and other relevant entities across Europe. The survey, which was completed online, consisted of four main sections. The first section was a general section, containing questions about scientific scope and data collection methods. The following three sections addressed legal and governance aspects, as well as technical and operational information. The focus was on the technical information. All sections included a mix of multiple-choice questions and open-ended questions.

The survey created (Appendix 1) was sent to the selected organisations for answering on 4th May 2021. The deadline for submission was 25th May 2021. The survey invitations were sent via the survey tool Webropol, through which the survey was also created.

In total 14 organisations responded to the online survey. Some responses were full of detail, others were more concise.

3. Observations on current practice

3.1. Scope and scientific quality

Fourteen data sharing initiatives from six different European countries (Austria, Belgium, Finland, Germany, Spain, Switzerland) responded to the survey. Four of the contributors are European projects or networks. The majority of these organisations provide data for different research purposes on different conditions or topics, such as COVID-19, prostate cancer, haematological malignancies and adverse effects of medicines in order to accelerate research, improve patient outcomes, increase public health knowledge and to enable patient empowerment. A couple of the analysed data sharing initiatives process data accumulated from various data sources into statistics and bring them to the benefit of citizens, decision-makers, researchers and other information users. The statistics describe for example the population's state of health, prevalence of diseases, and causes of death. One of the contributors provides generic services for data storing and sharing, where the main user groups are research organisations, research infrastructures and research projects, but the services are available for others as well.

Table 1. The following 11 initiatives took part in the survey. Three organisations wished to remain anonymous.

Data sharing initiative	Description	Website
BIFAP	BIFAP is a computerised database of primary care medical records for pharmacoepidemiological studies.	http://www.bifap.org

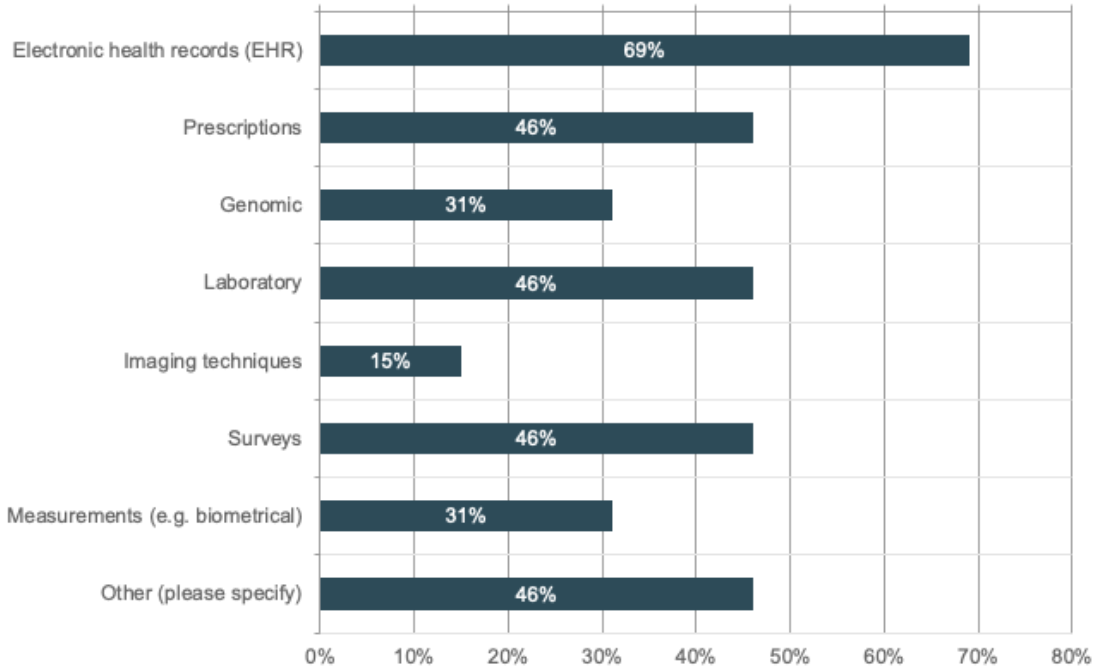
BIGAN	BIGAN is the platform for secondary use of health data of the health system of the Aragón Region (Spain).	https://tinyurl.com/biganplatform
EUDAT	The EUDAT Collaborative Data Infrastructure (or EUDAT CDI) is one of the largest infrastructures of integrated data services and resources supporting research in Europe.	https://eudat.eu/
European Association of Urology	The EAU represents the leading authority within Europe on urological practice, research and education. Through PIONEER they have developed both a centralised and a federated big data platform.	https://uroweb.org/
Findata	Findata is a Finnish Health and Social Data Permit Authority, a one-stop shop for the secondary use of health and social data. Its task is to issue data permits for the secondary use of health and social data in a centralised manner and to process data requests when data is combined from more than one register.	https://findata.fi/en/
Finnish Social Science Data Archive	Finnish Social Science Data Archive provides access to a wide range of digital research data for reuse.	https://www.fsd.tuni.fi/en/
Genesis-Online	Database of the Federal Statistical Office of Germany.	https://www.destatis.de/EN/Home/_node.html
Healthdata.be	The healthdata.be platform, developed by Sciensano, offers new perspectives on e-Health by simplifying the registration and storage of	https://healthdata.sciensano.be/en/home

	health data sent by various healthcare providers.	
Kanta	Data entered in the Finnish social welfare and healthcare service and in pharmacies is recorded in the Kanta Services, after which it can be processed for secondary use with the Kanta data platform.	https://www.kanta.fi/en/citizens
SIB	SIB leads and coordinates the field of bioinformatics in Switzerland and provides the life science community with a state-of-the-art bioinformatics infrastructure, including resources, expertise and services.	https://www.sib.swiss/
Statistics Finland	Statistics Finland processes data accumulated from various data sources into statistics and brings them to the benefit of citizens, decision-makers, researchers and other information users.	https://www.stat.fi/org/index_en.html

Some of the data sharing initiatives provide platforms centralising registries about health and healthcare, but there are also federated models, whereby the data stays at the respective sites and the analyses are executed at the local data sources or in other secure computing environments. The organisations collect many types of data from a wide variety of sources as depicted in Figure 1. The majority of the collected data are electronic health data records (EHR). Other types of data specified included for example interviews, epidemiology data and social security benefits.

The size of the data in the resources vary significantly between the data sharing initiatives - from mere megabytes to petabytes.

Figure 1. Data sources and collection methods. Other data types included interviews, epidemiology data and social security benefits.



3.2. Operational information

The survey requested operational information on data sharing initiatives' users. Users were defined as individuals that submit, access and control access to data. The purpose of this question was to identify the number of sustained users of the technical solutions in place to support data processing. Data sharing initiatives able to support large numbers of sustained users can be valuable to identify best practices that could inform the development of the EHDS. However, the diversity of the data sharing initiatives made it difficult to have a comparable approach to answers in this question, as the volume of data and functionalities associated with data sharing are very different among respondents (Table 2).

Table 2. The analysed data sharing initiatives (12/14) show great diversity regarding operational information on sustained users.

No. of sustained users who submit or store data	No. of sustained users who access data	No. of sustained users who control access to data (Data Access Committee)
821	2824	Not applicable
Not applicable	Currently 710 researchers who have a valid license. Many of them participate in more than one research project.	Our staff and admins, around 25 persons
Automated. Public health service information systems. Two persons develop, deploy and manage loading and storage system.	Two persons access, manage and extract data to researchers	Access to data approval is a business process involving several people: Ethical and scientific committee, corporate CEO (must sign approval), and data hub Platform Security Officer.
Order of 5	Order of 15-20	Project Leaders
1	50	10
10 (one per regional health system involved)	Around 20 active projects expected in 2021. This means around 30 active and current users (staff plus external investigators).	Seven members of the Scientific Commitee (initial authorisation to data access) and ten staff members participating in the process of delivering data to investigators.
Not applicable	Not applicable	Not applicable
Approx. 50	Varying number of anonymous users, approx. 500 concurrent users at any given time including search engine bots.	

Hundreds of controllers	Hundreds of researchers	Data permit authority
Hundreds		
All hospitals (n = 150), all laboratories (n = 250), all general practitioners (n=15,000)	Approx. 300 researchers	HD SteerCo = 20 members; Information Security Committee = 7 members
Approx. 10-15 users	Approx. 100 users	Defining the access to data is an automated process initiated by the study lead

3.3. Governance and regulatory framework for data sharing

Governance models of data sharing initiatives are a cornerstone for personal data protection, meaningful data reuse and transparency of data related procedures, including data quality. The technical and operational features are also strongly related to the organisational and regulatory landscape of data sharing initiatives.

This survey was not intended to characterise the governance and regulatory framework of data sharing initiatives in detail, as this matter is out of the scope of WP 7 (and will be analysed by other WPs of TEHDAS JA). Therefore, the requested information was intended to provide background for interpretation of the technical and operational analysis of survey results. As such, this section lays out a general description of the surveyed data sharing initiatives' governance framework, as a reference point to the subsequent technical analyses.

Governance models reported by the surveyed data sharing initiatives varied widely. The organisation supporting the data sharing initiative, the purpose of the data collection/sharing, and the magnitude of data processed by the initiative determine different types of governance models that support data processing for secondary use. Some data sharing initiatives are supported by a public entity, the ministry of health or a public institute, that have a legal mandate for data processing. Other initiatives are settled based on contractual agreements done by networks of different data controllers and processors.

The purpose of data collection varies from research-oriented (the vast majority of data sharing initiatives have this purpose, with or without other purposes) to health policy and planning as well as public health related.

As such, the surveyed data sharing initiatives have different legal regulatory frameworks for operation. Some are data controllers, others are data processors or a combination of both, with different legal support mechanisms and contractual arrangements for data sharing procedures and data-related policies.

All data sharing initiatives stated having data related policies in place. Some have general statements of data protection requirements, others have detailed written requirements for data access and data protection, including informed consent from data subjects. All, except one, have publicly available information on data protection and/or data policies. However, the content and detail of data related policies vary widely, besides the statements related to "GDPR compliance". Some initiatives declare restricting the access to data to specific purposes, not allowing data to be used when "commercial interests" are in place.

Consent management schemes are not mandatory, but may be required for some data processing initiatives and, in this case, included in the governance model /data policies. This survey did not collect information to analyse whether or not a consent management policy should be in place, so this issue is not addressed in this report, only a description of findings is provided. Among the surveyed data sharing initiatives, consent management schemes are implemented in 42% (5/12), but they are not managed centrally in most data sharing initiatives that have consent management schemes. Data sharing initiatives that are mainly "data repositories"/networks of data controllers commonly derive the responsibility of informed consent in their data suppliers. In these cases, consents usually exist, as datasets are collected in clinical trials or size limited cohort studies. Those initiatives that consist in "populational cohorts" do not usually have a consent management scheme, supporting its

legal basis for data processing in other provisions/regulations or compliance with requirements. One initiative declares that they are working on a general consent scheme, not yet implemented.

3.4. Technical information

The main part of the body of the survey is about technical characteristics of the architecture and infrastructure deployed by the data sharing initiatives to support their purposes. In the next subsections, we review the technical aspects of the initiatives following the logical data flow and user journey, from the primary data source to the final data processing and results, passing through pre-process, storage, discovery and delivery of the data. The service process, which was defined during the Advisory Group workshops in Task 7.2, is depicted in Figure 2.

Figure 2. 1High level service process for secondary use of data

3.4.1. General architecture model

Among the different data sharing initiatives analysed, we find a great diversity of architectures and configuration of solutions, mainly due to two aspects: the nature of the project or infrastructure analysed, and the level of technological maturity of the same.

In this sense, we will find differences in several of the aspects analysed (access protocols, standards used, authentication mechanisms or short and long-term storage systems) depending, for example, on whether we are talking about single databases or registries, systems of continuous integration of multi provider health data, namely health *data hubs*, or on platforms designed for more specific purposes.

In the first case, we are talking about platforms that aggregate and store data obtained from different data providers that have collected them for specific purposes (often research projects), and that are persisted for reuse in future projects. This allows the establishment of perfectly defined conditions, requirements and transfer agreements and access to data. It facilitates the ingestion and cataloguing of data, as well as the procedures for accessing them.

On the other hand, in the case of continuous data concentrators, the providers are usually hospitals, groups of hospitals or regional or national health services. In these cases, data capture and normalization is often more complex, and informed consent from patients is rarely available. This aspect makes governance more complicated, and requires additional security measures, which are not always fully developed.

From the perspective of the maturity of the models, we also find important differences. Some initiatives have been working for enough time, and have enough effort invested in configuring platforms that reliably and securely cover all the services necessary to collect, persist and give access to data or to secure analysis environments. By contrast, there are platforms that, for different reasons, have not reached the same level of maturity. These data sharing initiatives only partially facilitate researchers' access to adequate and quality data to carry out their research. Some of them make it easy to capture and search for information, but data selection

and extraction are done manually, and data must be downloaded for use by researchers. In other cases, classic BI tools are provided (OLAP analysis, Dashboards and reporting).

Similarly, we found that, in general, the most advanced initiatives provide cloud, distributed and / or federated analysis. Most of those that allow some type of analysis to be carried out, are able to do it only in a centralized manner, and several initiatives require data extraction by the researcher for local analysis, separate from the original repository.

In the following sections we will go into detail on these technical aspects that define the architecture and services of the different solutions analysed.

3.4.2. Data reception

With regard to data capture and collection, the procedures used largely depend on the type of platform. In the case of databases, whose function is to add collections of data collected for a specific purpose, and which are made available to others through a common repository, the capture is usually carried out directly by the data provider, and the process requires compliance with a series of requirements, such as metadata of the embedded datasets, or a Data Sharing Agreement. The procedure and the nature of the upload generally lead to uploading systems based on secure Web portals, where it is easy to implement the interactive functionality required for the on-time incorporation of a dataset.

On the contrary, in those data sharing initiatives that deal with continuous information (either in real time, or with a certain daily, weekly or monthly periodicity), the loading procedures tend to be less interactive and more automated. Different technologies and protocols specific to ETL procedures are frequently used. No metadata is created on each upload, because periodic uploads are recurring, and there is usually no variation in the meta-information. In any case, it is not uncommon in this data capture to find data capture protocols based on flat files (csv, txt) or specific to analytical tools (SAS / SPSS / Stata), transferred through FTP / SFTP protocols. In one case we found that the capture, transfer and storage of data from EHR systems is done using FHIR standards.

In this sense, when data is ingested by loading data files, the most frequent format used is plain text (csv or similar). Other structured text formats being used only in specific cases are for example XML, RDF or JSON, as well as FASTA, BAM or CRAM for genomic data, DICOM for medical imaging, or other proprietary formats from statistical analysis tools.

Database access protocols, such as ODBC or JDBC, represent the most common approaches to access data sources directly. In a couple of initiatives, it is specified that access to the data or its persistence is done in accordance with the OMOP data model, but always through access to relational databases, and in one of the cases FHIR is indicated, which implies not only defined access protocols, but also the format of the messaging used to access and load data.

When asked for community-recognised vocabularies and standards used for metadata and data to facilitate interoperability, the most frequently indicated standards are ICD-10 for diagnoses and procedures (in some cases also ICD-9), SNOMED-CT as clinical terminology and LOINC (now part of SNOMED) for laboratory data. Less frequently, WHO-ATC for active substances in drugs, ICDO3 and HemOnc in oncology, RxNorm for radiology, Orpha

(genetics in rare diseases) and ICPC-2 are also mentioned. As at least one of the analysed initiatives deals with social science data, ELSST thesaurus is also used.

The HL7-FHIR standard is mentioned by two data sharing initiatives as being employed to facilitate interoperability. HL7-FHIR has custom coding capabilities, and leverages upon other known standards (SNOMED, LOINC). Along SNOMED-CT, FHIR is powerful and largely used for interoperability reasons and for being useful for concept mapping and metadata specifications, so can also be used in metadata catalogues. In relation to this, JSON-LD is mentioned by one initiative as its metadata access framework.

One question in the survey referred to Clinical Document Architecture level used. Apparently, none of analysed data sharing initiatives uses CDA as an interoperability standard, maybe because CDA is applicable for the primary use of health data (delivery of care) interoperability contexts, and is not common for secondary use repositories. Hence, this question has been considered as not applicable by the responders.

Another important issue reviewed in the survey, and relevant in the process of data ingest and transfer, is whether data is anonymised or pseudonymised, and how this is performed. Similar to other aspects analysed in this survey, the anonymisation requirements and the specific techniques used to achieve it depends on the nature of the data sharing initiative. We will find different solutions in a data collections repository, where several data providers upload given datasets for reuse, than in a pure data hub with continuous ingestion and cross linkage of patients' data prepared to provide data subsets to demanding researchers. From the answers received to this question in the survey, we can draw some conclusions:

- In a data collections repository, pseudonymisation or anonymisation is, in general, a responsibility of the original data provider, and no data-linkage is possible between datasets. In some case a trusted third party is in charge of the pseudonymisation procedure, which potentially permits the linkage.
- In a federated data infrastructure, as researchers have not direct access to data, anonymisation/pseudonymisation is not a requirement at this level, though it can be done at node level depending on the use of data at that level.
- In a centralized infrastructure, where data is gathered into one single node from several providers (e.g., several hospitals loading data to a central repository), anonymisation/pseudonymisation is a requirement, as well as the possibility of cross data-linkage. In these cases, any kind of symmetric encryption is used, based on a single patient ID. This encryption can be done using double key or a trusted third-party encryption. This way, linkage and reidentification are possible, but controlled and limited, requiring the participation of several agents. In these cases, first encryption is done near or in the data source, so the data flow is yet encrypted, and second encryption is done in the repository.
- Besides ID encryption, other techniques are used to minimize the possibility of patient reidentification and the access to unneeded information. Among them are k-anonymisation and l-diversity, noise addition, data permutation, data generalisation.

Regarding data quality control, the maturity level is quite low in general, as quality control is mostly done manually, with basic checking controls. In some initiatives we can find some kind of automation in the checking process, either self-made developments or using third party tools, like Achilles or DataQualityDashboard from the OHDSI community.

Among checking processes, the most common controls are data format validation, detection of wrong measurement units, temporal data consistency, detection of missing mandatory data as well as management of outliers. These insights will also represent important input for TEHDAS' work package 6, aiming at establishing common quality controls.

One initiative describes an ETL pattern where data is evaluated during ingestion, and separated between valid and quarantine data according to a calculated quality score. Quarantine data can be reprocessed and sent to valid data after review and/or correction, or definitely discarded.

The most of the initiatives analysed declare to provide some kind of persistent unique identifiers for patients. In some cases, these identifiers are project specific, but they are usually based on an official ID (URN, Social Security Number, National Health Service ID etc), so patient data linkage is in principle possible at regional or national level.

3.4.3. Data storage

Technologies and infrastructures used in data storage for data persist is a very important aspect to describe how data is actually managed in data sharing initiatives. According to responses given to this question, most of the data sharing initiatives report the use of relational databases, either proprietary or open source. Some of them complement this solution with other technologies like NoSQL databases or even file servers.

Almost all data sharing initiatives examined use on-premise owned infrastructures to store their data, and their capacity is limited, although big enough for their purposes, ranging from a few GB to several TB. Three of the initiatives use cloud solutions or large-scale infrastructures provided by general purpose technological providers (Amazon, Azure or Universities). In these cases, hundreds of TB or *de facto* no limit capacity are reported.

Regarding differences between active data and long-term preservation, we can find multiple possibilities, depending on the nature of the data sharing entity. There are initiatives that only manage active (project linked) data, and this data is deleted once the project is finished. Others, on the contrary, only store long-term data, and active data is extracted and managed outside their infrastructure when needed, and others offer both active and long-term storage, with integrated tools for extracting and analysing data.

3.4.4. Data discovery

To ease the task of discovering which data is available in the repository of the data sharing initiatives, almost all of them offer some kind of catalogue with metadata describing datasets, individual variables, number of records, timestamp (for continuous cohort datasets), and any other statistic information. About one half of these platforms offer this service through a search portal with the capacity to launch queries with some level of complexity, including concept browsers, topic or keyword search, actual data scanners or info about publications based on archived data.

3.4.5. Data accessibility and security

This section covers how the data sharing initiatives identify and authenticate the researcher and how they manage different researchers' permission to access the data (authorisation). When a researcher has identified a dataset of interest, they are required to present a data

access application to a data access committee (DAC). When permission is granted to access the data, the next task is to technically enable the access. This might be implemented in various ways, depending on for example the technical capabilities of the service.

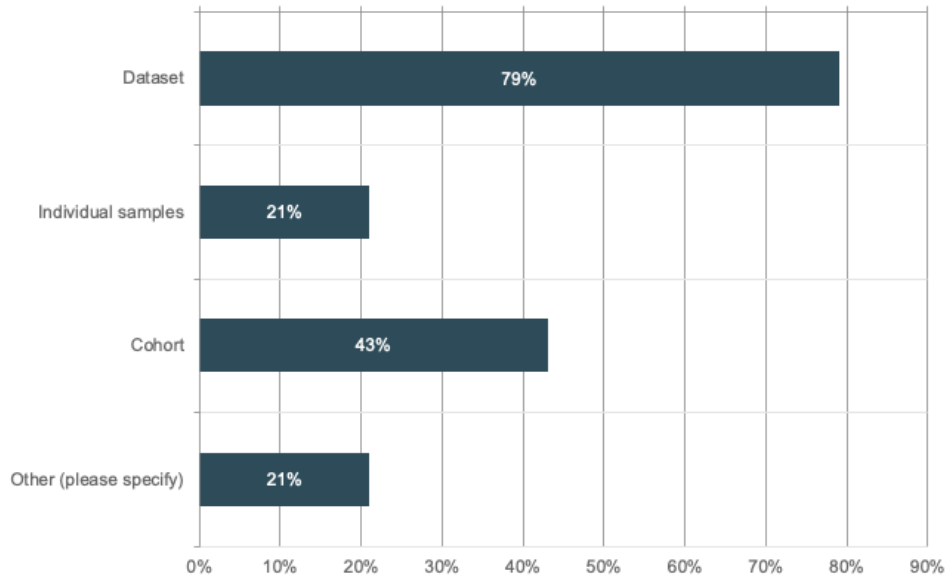
In this survey, the data application process itself was not studied, but rather the data sharing initiatives' current practices for researcher authentication and authorisation. Identification, authentication and authorisation (IAA) are essential concepts of identity and access management, as well as good security design. To manage access to any kind of data the user needs to be correctly identified. Regulating user access has traditionally involved authentication methods for verifying a user's identity, for example user demonstrates the ownership of an institutional email address. In this survey, we studied how many different user identification processes are in use.

Some of the analysed data sharing initiatives use federated authentication, but the majority have local or centralised authentication systems (for example LDAP). Local authentication systems used for accessing data and registration procedures are mostly made using password-based authentication. Some organisations also require a pre-approved national IP address, as well as a multi-factor authentication step before accessing the data to enhance security.

Using a federated authentication, users are able to access data using a single user account and password. Affiliated users can employ the user IDs assigned to them by their home universities to access and use numerous services. In one of the analysed data sharing initiatives, those that do not have credentials can apply for a username and password. The usernames are primarily issued to students and researchers from foreign universities as well as to staff members of those research organisations that do not belong to the federation. Registration rights are not granted for persons whose study or research is not attached to a university or a research institute.

The survey also aimed at understanding the granularity of access rights, that is the data objects to which an applicant can apply and a data access committee grant access. The survey reveals that most data sharing initiatives grant access to either complete datasets or specific cohorts (Figure 3).

Figure 3.2 Most data sharing initiatives grant access to either complete datasets or specific cohorts.



Researchers are often conducting research as an employee of the university, institution or other organisation they are affiliated with. If data access rights are granted to the applicant as an employee of the organisation, that would imply that access must be revoked if the researcher departs from their home organisation. The survey studies how the data sharing initiatives learn of the departures to revoke the access. Most of the data sharing initiatives grant access rights to the applicant personally, but it is also required for the applicant to be an employee of an organisation or member of a university or research group. In most cases, the data sharing initiatives impose an obligation to the applicant to inform the data sharing initiative on their departure so their access rights can be closed. Often this is a contractual issue, and any modification of the affiliation of the Principal Investigator of the research study should be communicated and approved. The access rights can then be revoked via the data sharing initiatives' IT-systems. The access rights are often also granted only for a predefined amount of time, for example the estimated duration of a research project. Furthermore, if users change their email address, they need to verify the new address. However, as long as the predominant way to access the data is to download them to the applicants' own environments, closing the access rights is not a strong control. As access to the datasets in a secure computing environment becomes more common, the ability to revoke the access on departure becomes more important.

The survey also studied how the data sharing initiatives store information on granted access rights. Every organisation is storing users' access rights in a different form, mostly, in databases as described in Table 3.

Table 3. 3The analysed data sharing initiatives (13/14) store users' access rights in a different form.

How data access permissions are stored
Internal systems
Licenses stored in their document control platform. Technical access rights are stored in AD and accessible by admins
Access rights granted by hosting partner. The hosting partner is holding the access rights information on their internal servers
Corporate LDAP server
Central database
Data platform documentation system
Intranet. A new IT tool is about to be implemented to facilitate traceability and archiving of all documents related to a specific research project
To be decided
Centralised user access system
Relational database
Documented in the granted data permit and technically stored in service providers systems
Central Identity and Access Management database (using CyberArk)
Access Control Lists

With regard to access control enforcement, the data sharing initiatives have different ways to ensure that the data is shielded from unauthorised use. The client environments are segregated, and access per user environment is granted only to those who are mentioned in the data permit. Processing, utilisation and transfer rights are restricted, and there is no direct access to the database. Furthermore, purpose specific pseudonyms, anonymisation services and/or anonymous credentials are used. Some data sharing initiatives also require NDAs from each researcher, as well as the approval of their terms of use. The license states the legal

consequences of violations. Company policies are also in place to minimise the exposure to hackers.

It is interesting to note that one of the data sharing initiatives inquired operates a federated data platform, where research questions must be submitted to a centralised committee who assesses the merits of the application. If approved, a research team is established consisting of research project members and, where applicable, external data providers. The research team presents their question to all data providers who can then choose to opt in or out with their data for that specific question. In this way, authorisation for use of the data remains within the data provider.

3.4.6. Data delivery and processing

When the data permit authority has approved the data access application and signed the data access agreement, the applicant is able to use their access rights. Traditionally this is done by downloading the data from the data sharing initiative's data download service to the applicant's own environment, for example using a secure FTP with encrypted files or authenticated web portals where authorized is available. Alternatively, the data is delivered to the applicant in a secure computing environment provided by the data sharing initiative. Most of the analysed data sharing initiatives have moved or are moving towards a direction of allowing the data to be accessed in a secure computing environment. The applicants can use the basic analytical tools provided there, but also request more tools given that their licensing and security aspects do not contradict the data sharing initiatives' policies. Usually, the maturity of the data sharing initiative sets the data delivery mechanism, the data download being the less mature approach, and the secure computing environment the most advanced.

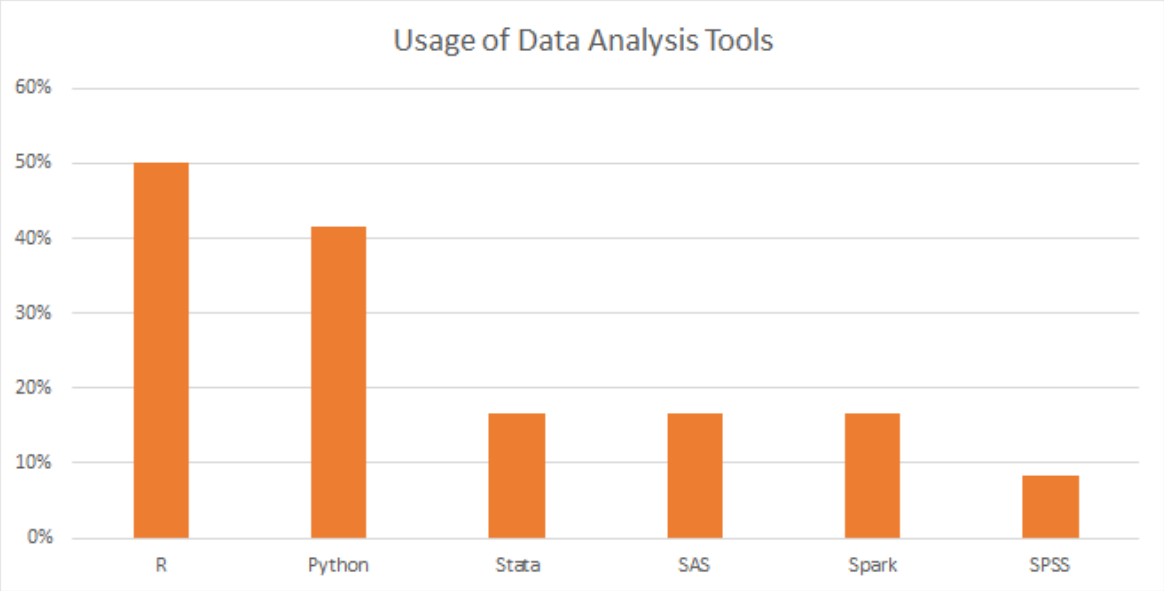
For the vast majority of responding data sharing initiatives, data processing associated with the data delivery is done in batches, either on-demand for specific research projects or in intervals for continuous data reporting (daily, weekly, or monthly). For the respondents running data warehouses, this task is performed daily through the existing ETL processes.

3.4.7. Data analysis

The landscape of tools available for data (pre-)processing and analysis is heterogeneous. While some data sharing initiatives do not provide data processing services at all, and some only on demand, others support a wide range of applications or even provide powerful virtual machines for remote access.

The most commonly supported data analysis languages are R and Python, including their corresponding statistical packages – of which Python is mostly run in Jupyter (Lab or Hub) environments. The following diagram (Figure 4) compares the popularity of stated solutions.

Figure 4. Popularity of different data analysis tools among the analysed data sharing initiatives.



Further, it is worth mentioning that not all the responding data sharing initiatives stated their analysis tools, but some only their hardware / virtual machines provided. Hence, one can assume that the percentages for the data analysis tools displayed are higher in reality, but most likely following the same distribution between them (i.e., R and Python being used by far most commonly).

This picture further is very much in line with broader surveys on the usage of statistical environments, which also state R and Python to be the most commonly used [3] – of which Python recently experienced a large increase in popularity, due to its nature of being also a general-purpose programming language.

Both Python and R also provide basic visualization capabilities. In addition, standardized reports and interactive dashboards are generated regularly at some of the data sharing initiatives surveyed: for instance, powered by Microsoft Azure or RShiny, or by custom developed dashboard solutions.

Data processing hardware varies between use of CPUs, GPUs and scalable cloud computing infrastructure, of which CPUs and GPUs represent the most common option.

Data exports are handled in various ways. Some of the data sharing initiatives provide .csv downloads either directly via websites, or via SFTP connections. Others only provide data exports upon request by individual solutions. In some cases, data sharing initiatives do not provide raw data – also due to data protection concerns, but only summaries of the data to outsiders. In the case of the federated network, highlighted in the previous section, data providers participating in the analyses receive the R package to be executed and return aggregated results to the research team. In this instance the initial data sharing initiative contacted by the research team never has access to the data available in other data sharing sites participating in the overall analysis. As in the data access stage, the data providers can choose to opt in or out for the specific question.

All data sharing initiatives responding to the survey perform either pseudonymisation or anonymisation on the data provided for analysis. Where necessary, pseudonymisation is performed at data ingestion. Other data sharing initiatives either already process anonymised data only, or anonymise data upon export.

At 67% of the data sharing initiatives, researchers can import tools for analysis, such as libraries, to the analysis platform. The imports range from smaller code snippets to libraries and packages, and even additional data. However, the process to import such additional resources is heterogenous. Several of the responding data sharing initiatives perform individual designated security checks upon the imports.

Logging of user actions also varies greatly between data sharing initiatives. Especially logins and exports of data are closely logged. At most initiatives, a broad range of user interactions with the system is clearly logged in relational data bases. At some of them, sophisticated search and auditing services provide further insights on the logging, and one integrated automated alert system in case of suspicious behaviour.

References

1. Stewart D, Simmons M. The Business Playground: Where Creativity and Commerce Collide. Berkeley: New Riders Press; 2010.

2. Patrias K. Citing medicine: the NLM style guide for authors, editors, and publishers [Internet]. 2nd ed. Wendling DL, technical editor. Bethesda (MD): National Library of Medicine (US); 2007 - [updated 2015 Oct 2; cited Year Month Day]. Available from: http://www.nlm.nih.gov/citingmedicine

3. Ozgur, C., Alam, P., & Booth, D. R, Python, Excel, SPSS, SAS, and MINITAB in Research.

Annex 1

TEHDAS SURVEY FOR DATA SHARING INITIATIVES

ORGANISATIONAL INFORMATION

1) State the general purpose of the data collection/generation. If possible, describe shortly a few “iconic” cases solved by your organisation

2) Data sources and collection methods

   i. Electronic Health Records (EHR)

   ii. Prescriptions

   iii. Genomic

   iv. Laboratory

   v. Imaging techniques

   vi. Surveys

   vii. Measurements (e.g. biometrical)

   viii. Other (please specify)

3) State the expected size of the data in the resource

LEGAL AND GOVERNANCE

4) In the scope of the EU GDPR, what is your organisation’s role in relation to personal data?

   i. Data controller

   ii. Joint controller

   iii. Data processor

   iv. None of the above (please specify)

5) Describe how your organisation meets the technical requirements of the data protection principles (GDPR)

6) Describe or provide links to relevant documentation regarding Data Policy, License Model and Terms of Use

TECHNICAL INFORMATION

7) Describe the platform's architecture model (general workflow from data ingestion to analysis)

DATA RECEPTION

8) Describe how the data is collected (data retrieval (e.g. FTP, API), parsing, transforming, loading)

9) Specify the types and formats of data collected

   i. Plain text
   ii. FASTA
   iii. XML
   iv. RDF
   v. Dublin Core
   vi. Tsv
   vii. JSON
   viii. DICOM
   ix. Parquet
   x. Other (please specify)

10) Which community-recognised vocabularies, standards or methodologies are used for metadata and data to facilitate interoperability (e.g. HL7 FHIR, SNOMED CT, LOINC, ICD-10)?

11) If the resource contains EHR data, state which Clinical Document Architecture (CDA) level the records follow

   i. Level 1
   ii. Level 2
   iii. Level 3
   iv. Not applicable

12) How is the data anonymised/pseudonymised?

13) What data quality control procedures are applied?

14) Does the resource provide persistent and unique identifiers?

    i. No

    ii. Yes (please specify)

15) Does the resource have a consent management scheme in place?

    i. No

    ii. Yes (please elaborate)

DATA STORAGE & INTERFACES

16) Describe the technologies used for data storage (e.g. relational database, NoSQL)

17) How do you handle active data (data stored during analysis) vs. archived data (long term preservation)?

18) How much storage capacity is in use?

19) Describe the services through which data is shared (e.g. website, APIs, FTP)

DATA ACCESS MECHANISM SUPPORT

20) On a general level, describe how the access control mechanism has been implemented (authentication and authorisation)

21) What is the data object to which access rights are granted?

    i. Dataset

    ii. Individual samples

    iii. Cohort

    iv. Other (please specify)

22) How do you electronically identify and authenticate an applicant (e.g. email address, username, two-factor authentication, eIDAS)?

23) Are access rights granted to the applicant personally or as an employee of an organisation?

24) If the applicant ceases their affiliation with the organisation is their access revoked? If access is revoked, how do you enforce it?

25) Where and how is the information on granted access rights stored?

26) How is the data shielded from unauthorised use (Access Control Enforcement)?

27) How can the applicant use their access rights (e.g. the applicant can download the data to their own environment/the applicant can access the data in a dedicated IaaS cloud environment where the data are available as a filemount/the applicant can access the data only using the tools provided in...)?

DATA PROCESSING

28) How is the data processed for analysis (e.g. batch with regular updates or real-time data processing)?

29) Describe the data processing services offered by your platform (used technologies (IDEs), software, languages, libraries etc. If there are any in-house developed services and software, please detail their purpose)

30) Outline the computing capabilities tied with the infrastructure (CPU, GPU)

31) Describe the data analysis and visualisation capabilities and methods (if any available)

32) How is data export handled?

33) How do you make sure that the exported data is pseudonymised/anonymised?

34) Are researchers allowed to import resources (e.g., code, data, algorithms)?

i. No

ii. Yes (please specify)

35) Describe the logging and auditing of user actions

DATA DISCOVERY

36) Describe the components, tools and services that allow users to discover the data (e.g. search engine, browsing catalogue, viewing metadata)

OPERATIONAL INFORMATION

37) State the number of sustained users

i. Who submit or store data

ii. Who access data

iii. Who control access to data (Data Access Committee)

38) State the number of objects stored

   i. Samples / Individuals

   ii. Studies / Datasets

39) State the number of downloads (including FTP downloads and programmatic access)

Annex 2

GLOSSARY

BAM: Binary Alignment Map

CDA: Clinical Document Architecture

CPUs: Central Processing Unit

CRAM: Compressed columnar file format for storing biological sequences aligned to a reference sequence

DAC: Data Access Committee

DICOM: Digital Imaging and Communications in Medicine

EHDS: European Health Data Space

EHR systems: Electronic Health Records systems

ELSST: The European Language Social Science Thesaurus

ETL: Extract – Transform – Load: the typical steps in which data is ingested into a data warehouse

FASTA: Text-based format for representing either nucleotide sequences or amino acid (protein) sequences

FHIR: Fast Healthcare Interoperability Resources - a standard developed and maintained by HL7

FTP / SFTP: File Transfer Protocol / Secure File Transfer Protocol

GDPR: General Data Protection Regulation

GPUs: Graphics Processing Unit

HemOnc: Hematology/Oncology

IAA: Identification, authentication and authorisation

ICD-10: International Classification of Diseases and Related Health Problems - 10th Revision

ICDO3: International Classification of Diseases for Oncology 3rd Revision

ICPC-2: The International Classification of Primary Care

JSON: JavaScript Object Notation, the most common way in which (smaller pieces) are exchanged in networks and by APIs, based on key-value-pairs

JSON-LD: JavaScript Object Notation for Linked Data

Jupyter (Lab or Hub): Popular environments to execute Python or R scripts

LDAP: Lightweight Directory Access Protocol

LOINC: Logical Observation Identifiers Names and Codes

ODBC / JDBC: Open Database Connectivity / Java Database Connectivity

OLAP analysis, Dashboards and reporting: Online Analytical Processing, typically based on data warehouses and data cubes

OMOP: Observational Medical Outcomes Partnership

RDF: Resource Description Format

RxNorm: Prescription for Electronic Drug Information Exchange

SAS / SPSS / Stata: some commercially available applications / environments for data analysis

SNOMED-CT: Systematized Nomenclature of Medicine Clinical Terms

XML: Extended Markup Language, a very versatile way to structure files or documents