# 1+MG – Data Protection by Design and Default

14 February 2023

Regina Becker

# Mandate of the 1+MG WGs

- 1+MG will be established **by governments** of member countries
- WGs were created to **enable this implementation**
- We need to **plan on behalf of the government**, not just the nodes or the central coordination
- We need to provide a **concept** to governments that
  - Ensures **legal and ethical compliance**
  - Offers **sufficient protection** of data and data subjects
  - Covers the **entire data life cycle** from bringing data into the 1+MG to the use to achieve better health
  - Is **feasible**, also financially – operations must scale
  - **Integrates** with the **EHDS** and is as much as possible **reusable** in the EHDS: between EHDS and 1+MG investment and maintenance, the **choice of governments** will be clear

# The concept of DPbDD

- Design your systems **first with the idea** that you need to ensure **security and DP principles**.
- Then see how you **open up the system** for cases where protection is not or less needed.
- The **default setting** should always be: **closed / not permitted**
- **Access / permission** should always be an **active step**
- **NEVER** design the other way around
  (i.e. never introduce considerations on data protection as an afterthought)

# “Data protection” rather than “privacy”

- NOT “Privacy by Design and Default”
- Data Protection is **more than privacy** of data subjects
  DPbDD measures are not only about **security**
- Data Protection is about **rights and freedom** of data subjects with regard to their data
- DPbDD includes compliance with **ALL data protection principles**
- Best source to understand requirements:
  **EDPB Guidelines 04/2019**
  https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201904_dataprotection_by_design_and_by_default_v2.0_en.pdf

# DPbDD measures

- DPbDD ensure compliance with DP principles
  - DP principles are listed in Art. 5 GDPR
  - DP principles are further specified in Articles 6-49
- DPbDD measures translate into
  - Technical measures
  - Organisational measures
  - Workflows
  - Documentation
- DPbDD translates into
  - The data governance and contractual framework
  - The information management
  - The IT infrastructure

# The data protection principles

- Lawfulness
- Fairness
- Transparency
- Purpose Limitation
- Data Minimisation
- Accuracy
- Storage limitation
- Integrity and confidentiality
- Accountability

All these principles need to be **complied with** for the data processing

# Lawfulness of processing

- Lawfulness means
  - is an appropriate legal basis / legal coverage of the processing in place; includes GDPR Art. 6 legal basis, Art. 9.2 exemption to process health and genetic data; suitable instrument to allow international transfer where applicable; any other law with respect to the processing from national or EU level
- Lawfulness involves
  - Documentation of existing legal basis and legitimations
  - ELSI metadata on all requirements and restrictions
  - Being able to build into the the technical safeguards that only lawfully accessible data are processed for a purpose

# Fairness of processing

- Fairness means
  - Processing should not be **unexpected** or **detrimental** for the data subject
  - Data subjects should be able to exercise their rights
- Fairness involves
  - Balancing of interests between data subjects and society
  - Technical and organisational / procedural measures need to be in place to allow the exercise of rights

# Transparency of processing

- Transparency means
  - Data subjects must be aware of the processing of their data
  - Data subjects are informed if a purpose or other relevant information about the processing of their data changes
  - The public is sufficiently informed about the processing including safeguards
- Transparency involves
  - Capturing and providing information on data use both publicly and individually
  - Being able to regularly update the data subjects
  - Provide information in native languages of the data subjects

# Purpose limitation

- Purpose limitation means
  - Ensure that data are only processed for the foreseen, lawful and communicated purposes
- Purpose limitation involves
  - Documentation and awareness of permitted purposes
  - Procedures to update purposes
  - Annotation / [machine readable] ELSI metadata that inform about the permitted use
  - Settings that non-permitted use is automatically not possible → Default is ‘closed’

# Data Minimisation

- Data minimisation means
  - Only data relevant to achieve the purpose must be processed
- Data minimisation involves
  - Limiting data collection to data necessary to achieve the intended purpose
  - Limiting access to data types necessary for the purpose
  - Limiting access to data subjects necessary for the purpose
  - Minimise number of dataset copies to what is necessary

# Accuracy
## Ensure the data is correct

- Accuracy means
  - Data should be accurate and up-to-date for the purposes for which they are envisaged
  - No “absolute” accuracy is needed
- Accuracy involves
  - Possibility to allow **rectification** of data (as under Fairness)
  - **Evaluation / verification** of automated and manual processing steps (e.g. avoid transfer of data use conditions by hand)

# Storage limitation

- Storage limitation means
  - Personal data must not be kept longer than permitted
  - Duration is justified by the necessary of the data for the purposes
- Storage limitation involves
  - Definition of the “life time” of each dataset versions in dependency on the purpose
  - Annotation with such life time
  - Technical and organisational safeguards to ensure that data are not kept longer
  - Technologies to **effectively** delete data (**sanitisation**)

# Integrity and confidentiality

- Integrity and confidentiality means
  - Protection against unauthorised or unlawful processing
  - Protection against accidental loss, destruction or damage
- Integrity and confidentiality involves
  - **Security measures** in transfer, rest, during and following access (i.e. transmission of aggregated data)
  - Back-ups and possibility for disaster-recovery

# Accountability

- Accountability means
  - To demonstrate compliance
  - The ability to prove that the above aspects have been sufficiently implemented
- Accountability involves
  - **Documentation** of processing and of measures taken (both technical and organisational) including changes over time
  - Risk analyses of the chosen measures (also documented): demonstrate effectiveness and appropriateness
  - Data protection impact assessment
  - Trust measures such as audits and certification

# The Master Plan

- Draw **data life cycle** and corresponding **data flows**
- **Map** the DPbDD requirements on each step
- Who is responsible?
  - What needs to be agreed jointly for all?
  - Where are local flexibilities?
  - What are mandatory joint tools / standards?
  - What is done / provided at the central level?
  - What on the national level?
- Are there tools that fulfil requirements?
- Where are **gaps**?
- Can they be **overcome**?
- Is there a **temporary fix**? Is that fix still sufficient?
  [Risk assessment! Impact assessment!]

# 1+MG data governance

- 1+MG is (currently) planned to be a distributed data infrastructure
- Distribution means location of data, not legal distribution
- One harmonised data governance
- Planned setup
  - Data holder / provider legally transfer data for purpose of making them available according to agreed governance to 1+MG:
    Data holders are controllers of this legal transfer and the associated processing of curation and characterisation
  - 1+MG legal entity will be legally responsible for data disclosure to users but build on feedback from national level on their respective datasets
  - Users are sole controller for the processing of their data
- Infrastructure must appear harmonised towards the outside
- Where special use requirements must be considered, these have to be machine-readable

# Data inclusion

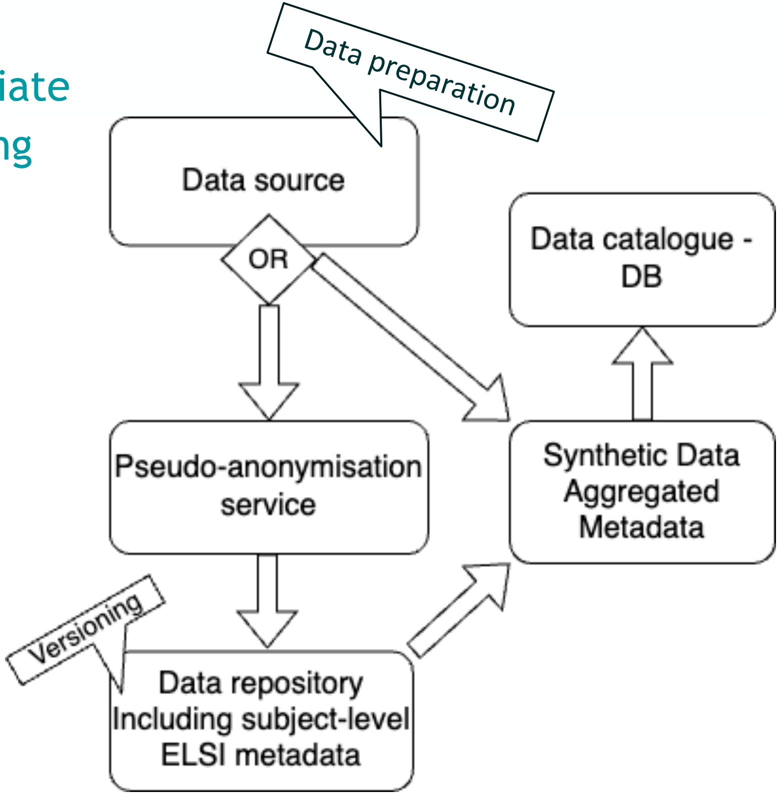
- Data holders
  - Inform data subjects on intended 1+MG data use and obtain consent where appropriate
  - Create ELSI metadata (also subject-specific) coding

    Scope of permitted data use purposes

    Vulnerabilities

    Data subject preferences for being contacted

    Transfer data into 1+MG data models
  - Generate information for data catalogue
  - Generate synthetic datasets
  - Pseudonymise data
  - Transmit (where necessary) to repository
- 1+MG
  - Check quality and completeness
  - Announce data in data catalogue



# Data Inclusion - IT requirements

- **Data transfer tools** that are **suitable for biomedical researchers** and clinical people that allow **secure transfers** to the 1+MG IT infrastructure.
- **Data integrity check** after transfer has to be ensured (e.g. use of checksums, data scrubbing)
- **Data format validation** (e.g. for genomics data, associated metadata)
- **To be discussed** if sufficiently relevant: **Secondary pseudonymisation of genome and phenotypic data separately** and/or **physical separation** of genomic and phenotypic data on different servers
- Collection / uploading of **structured metadata into suitable tools** relevant for ingestion into the data management system (e.g. ELSI metadata relevant for data use) and for display in the data catalogue

# Data inclusion - information requirement

- Data holder administrative data
- Data use conditions [subject-level information]
  - Categories of purposes for data use with recipients
  - Limitations on territorial scope (only outside EEA can be limited)
  - Data subject’s consent that data holder can disclose additional information where needed by data user
  - Information on data subjects where applicable, e.g. relevant vulnerabilities such as minors and anything that may have implications on processing
- Information around recontacting subjects [subject-level information]
  - Purposes for re-contacting with communication channels (including [permanent] contacts) and language

    Participation in research studies

    Obtaining consent / active information on new purposes

    General information provision

    (Incidental findings)
  - General “no return” flag (initially collection level but will be specified on the level of data subjects)

# Data inclusion - information requirement

- Information on data retention criteria / requirements / time
- Legal basis for inclusion
- Scientific metadata for all 1+MG data types as defined in 1+MG
- Metadata on changes in datasets (e.g. exercise of data subjects rights)
- Information on additional data beyond the 1+MG defined datasets or samples
- Integrity check information (after data transfer/periodically during storage)
- Time stamp of upload
- Registration / versioning of datasets (creation of new dataset version 1.0 or adding to existing datasets creating new version 2.0)
- A universal unique identifier (UUID) of the same data and version (identical copy) assigned and used in all federated systems

# General requirements IT IS: data in 1+MG

- Storage that is secured against external and internal attacks /accidental disclosure (e.g. no direct internet access);
- Protection against data loss (e.g. back-ups and disaster recovery)
- Encryption of data where the IT environment allows fast decryption / encryption on the fly [to be discussed for relevance / efficiency / cost as data are almost constantly in use]
- Access restriction: default is no access, fine-grained access control lists (ACLs)
- Data management with possibility to send automated reminders in view of upcoming deadlines (end of access, end of retention time for a dataset)
- Possibility for record-level metadata: version, ELSI information, data use and retention time
- Ability to remove data from the active system (i.e. no longer accessible)
- Ability to restrict processing to certain users, purposes and subsets of data
- Ability to remove data from back-ups or restrict their restoring

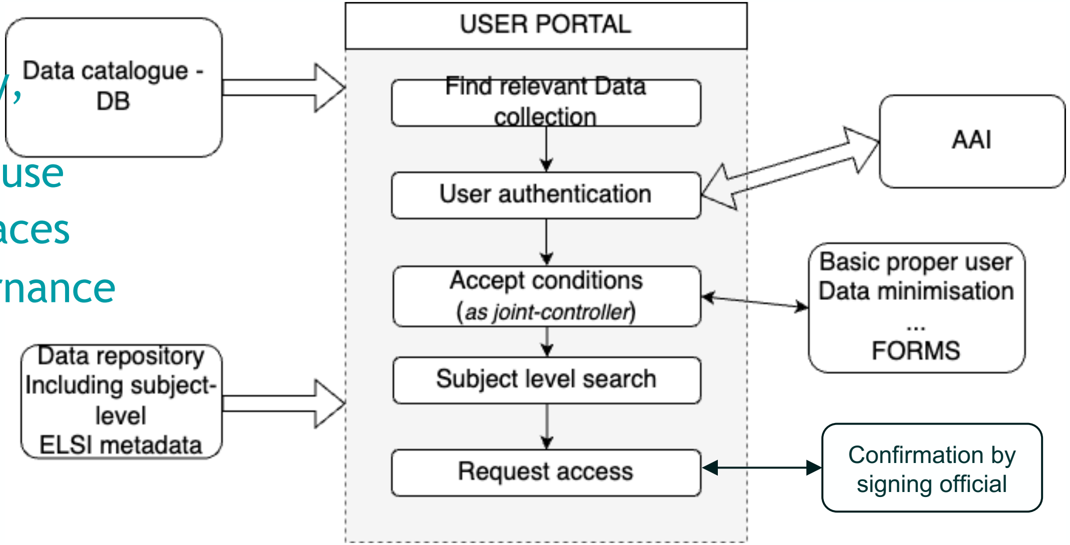
# General requirement information on data

- Information on data location of the different versions
- Information on retention for individual dataset versions
- Data archived for project xy; duration of retention
- Data under delete request (to be deleted at the end of use or archiving; data not to be restored from backup)
- Information on changes of dataset on subject level
- Versioning following update of data for new data points
- Versioning of datasets for changes (rectified)
- Dataset current version “no more use for N / at all”
- Pseudonym matching table (if applicable)
- Controller for hosting
- Legal basis for hosting (if not defined by 1+MG)
- Encryption keys

# Data access request

- Data user
  - Investigates on webpage the data available in 1+MG and the principal data use conditions
  - Is interested and wants to see if 1+MG has data matching his inclusion criteria
- 1+MG
  - Offers a tool to allow subject level data discovery
  - Has a form where for data discovery, the user must specify his principal purpose and agrees to the terms of use
  - Has access request type user interfaces
  - One form for all; agreed joint governance
  - Requires employer-confirmed user authentication
- Data user
  - Selects relevant datasets
  - Requests access

1+MG: Obtains confirmation by signing official or white-lists in place


USER PORTAL
Data catalogue - DB
Find relevant Data collection
AAI
User authentication
Accept conditions (*as joint-controller*)
Basic proper user
Data minimisation
...
FORMS
Data repository Including subject-level ELSI metadata
Subject level search
Request access
Confirmation by signing official

# Data discovery - IT requirements

- Central data catalogue that seamlessly comprises information from national data catalogues where applicable
- Data catalogue should offer sufficient characterisation of datasets for requesters to allow assessing the usability of the data for their purposes; ideally, also synthetic datasets are made available.
- Option for data requesters to specify their type of access request (research, healthcare, policy development - still to be defined based on use case input) -> such specification influences user/requester interface offered
- Possibility for requesters to identify relevant data subjects based on genotype, phenotype and/or other features such as ethnic origin
- For subject level data discovery, the requester must also specify the area of research based on controlled vocabulary (e.g. disease specific, legal basis)
- Authentication and authorisation infrastructure (AAI) that allows authentication of requester by home institution and/or role (relevant roles are researcher, healthcare professional etc.).
- The AAI must be technically interoperable to allow for federated identity management.

# Data discovery - IT requirements

- Data requesters must be able to log into one AAI implementation that gives access to data across all 1+MG sites.
- Possibility to block persons from access if they appear on a “blocked list”
- Offer suitable “terms of use” that need to be confirmed before the requester can start the search on a subject-level (i.e. data on individuals)
- Data management system that allows to limit data searches to relevant data only (offer only datasets for search that can be processed under the legal basis of the requester, by the category of requester for the purposes of the project)
- Ensure that such focussed search cannot be bypassed and/or that incorrect behaviour is monitored / flagged
- Data discovery tool that allows subject-level searches for genotype and/or phenotype without releasing information that can be used to identify subjects and/or reconstruct the data by single search or combination of multiple searches by built in protective features
- Logging of subject level searches
- Possibility to download / view synthetic datasets of relevant data collections

# Data access request - IT requirements

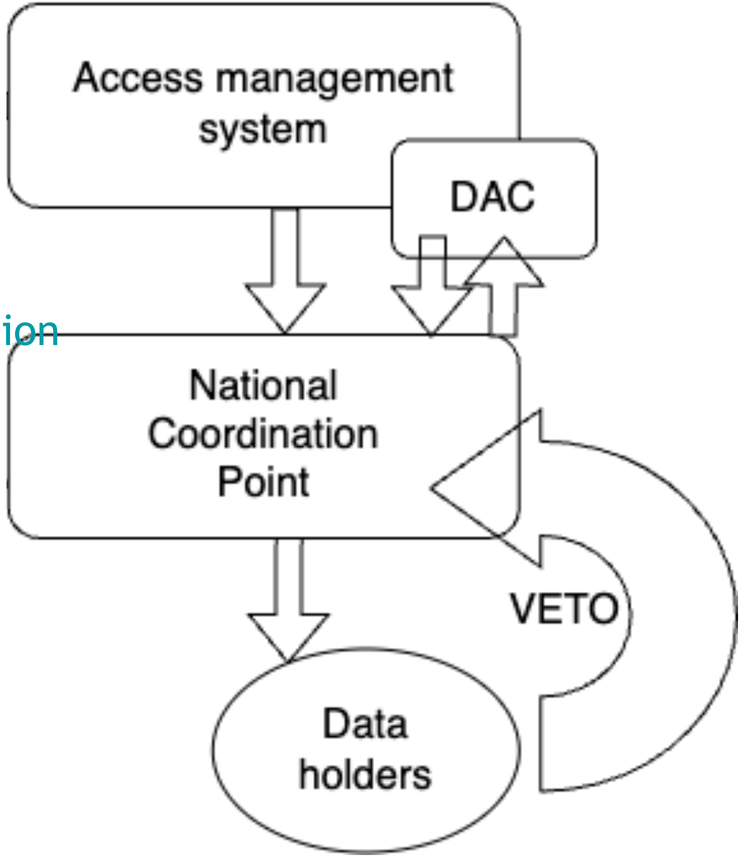
- Data user portal that allows seamless transfer from data discovery to access request on the selected datasets, i.e. launching access request on selection
- Web form for the collection of relevant information for access review and approval, depending on access request type. Characterisation of access request based on controlled vocabulary matching the ELSI metadata
- Possibility to launch same access request for several institutions (joint controllers)
- Ideally: electronic signature for registered organisations’ signing official
- Ideally: possibility for  users to be flagged as mandated

# Data access request - information needs

- Type of access request
    - Information provided by requester: research / healthcare / policy dev
    - Information required for respective type of access request (see list; handling files, structured data and free text)
- Log information of subject-level searches by requester
- Information on (institutionally) mandated and blocked researchers
- Information on signing officials for each institution
- Association of records with defined access requests, i.e. the possibility to keep an auditable track of which records are / were used for which access
- Recommendation by central DAC, where applicable: feedback of National Coordination Points, final decision on request (per collection / subset)
- Objections by individual data subjects
- Logging all interactions until approval and access provision (who, when, what)
- (Contract)

# Data access request review (1)

- 1+MG
  - Channels the access request to internal DAC and National Coordination Point (NCP)
  - DAC reviews access request and sends justified decision recommendation to NCP
- NCP
  - Provides relevant decision makers on national level with information on request and recommendation
- Data holder / permit authority
  - Reviews request and recommendation
  - Exercise a (justified) veto within a certain time frame if not agreeing with recommendation
- 1+MG
  - Integrates feedback into one coherent reply
  - Drives consensus process in case of conflicting replies

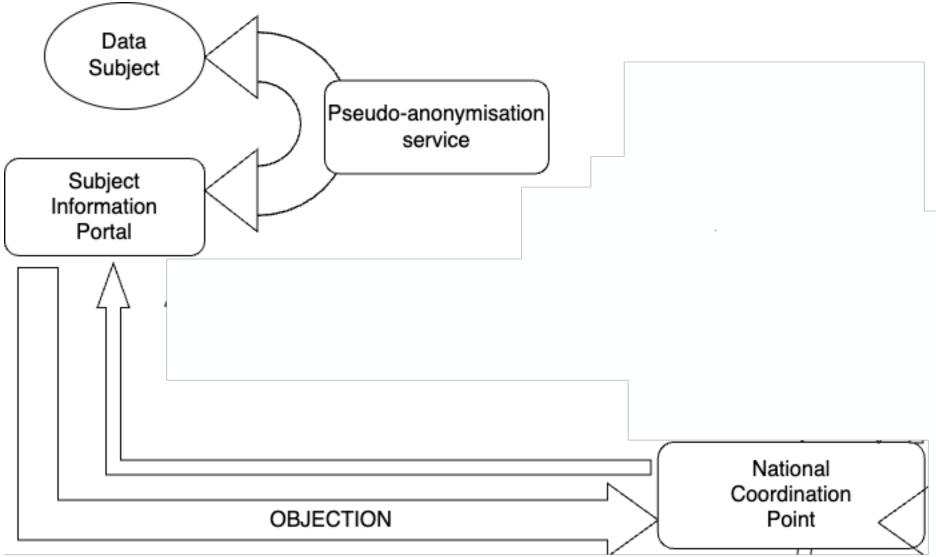


# Access request review - IT requirements

- Access request management system that acquires the necessary access request information (from the user portal)
- Notifications to the relevant actors
- Ideally validation of submissions by detecting obvious mistakes and automatic rejection in such cases or flagging of such faults at minimum.
- Allows communication between actors; documented exchange of messages and consensus process
- API to retrieve information stored in the system.
- Possibility for documented approvals / rejections with justification and exchanges on request in the system.

# Data access request review (2)

- NCP
  - Submits information on data access request to subject information portal or similar service to reach out to data subjects
- National level
  - Has pseudonymisation service that allows linking between pseudonym and identity
- Data subject
  - May object to the use of data in specific research project
- NCP
  - Feeds information on objection to 1+MG
- 1+MG
  - Makes sure that only records of data subjects having not objected are made available

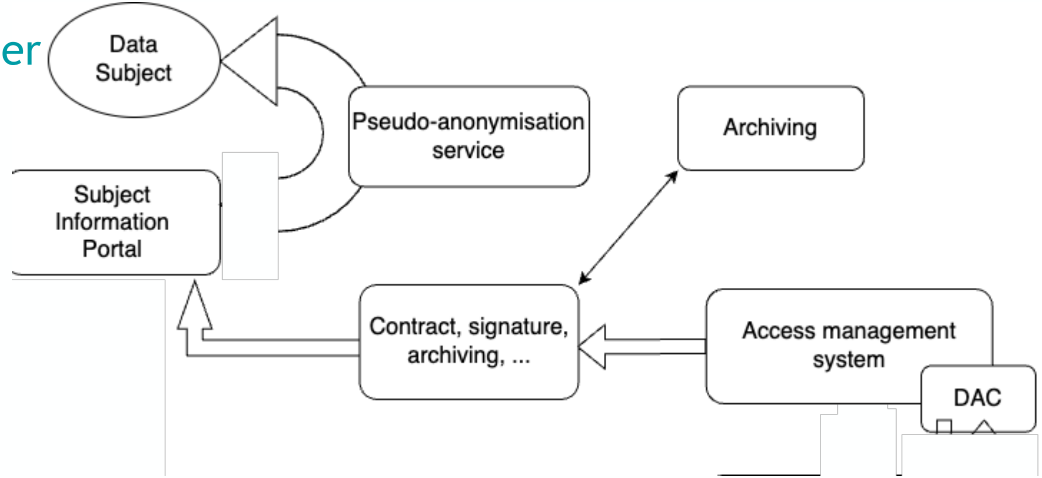

Data Subject
Pseudo-anonymisation service
Subject Information Portal
OBJECTION
National Coordination Point

# Access request review - IT requirements

- Possibility to inform data subjects in advance where positive decision is likely
- Requires automatic linking between data subject identity and pseudonym
- Ideally: information portal where data subjects can find information and set the parameters for their information
  - How to be informed
  - When to be informed
- Possibility to object to the data disclosure / processing for the research project
- Ability of the IT infrastructure to handle individual objections (per data subject  per project)

# Data access request review (3)

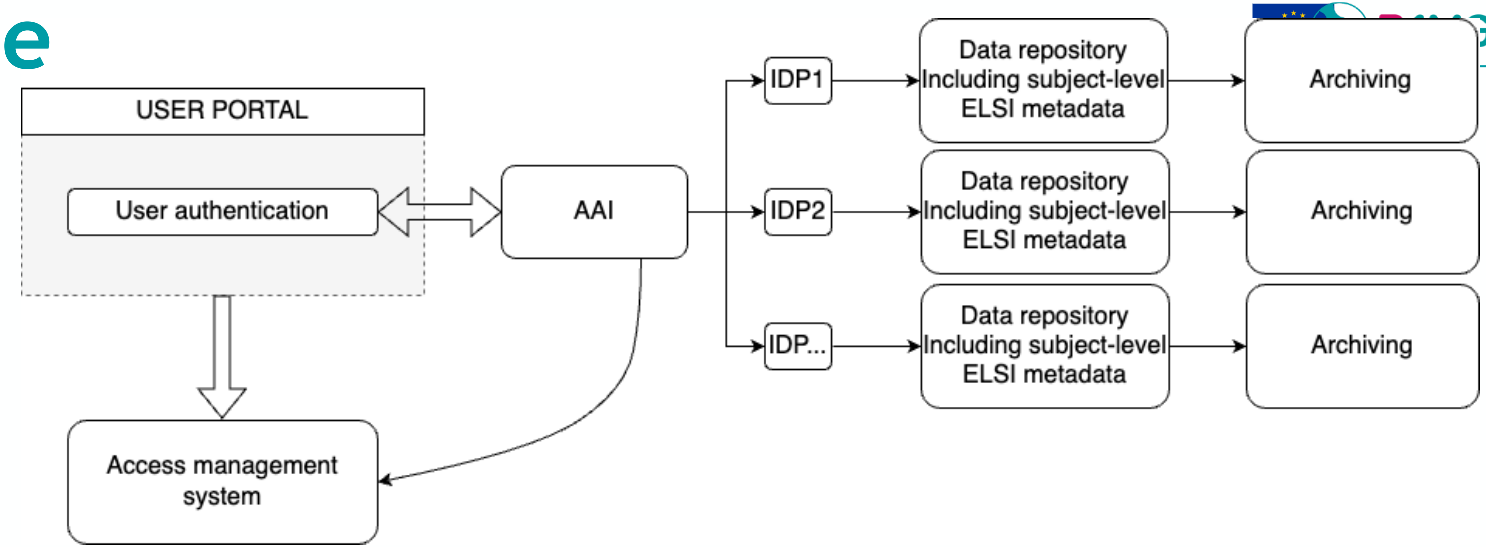
- 1+MG
  - In case of negative access decision, informs the user including justification of rejection
  - In case of positive access decision, produces contract and sends it to user
- User institution
  - Signs contract
- 1+MG
  - Archives access request, review feedback and contract
  - Submits information to subject information portal for those data subjects that are included in the decision
- National level
  - Allows that data subjects can find this information in the portal (or get notification where required)


Data Subject
Pseudo-anonymisation service
Archiving
Subject Information Portal
Contract, signature, archiving, ...
Access management system
DAC

# Access request review

- In case of positive decision
- Automated generation of contract
  - Including data use conditions
  - Including other relevant requirements that may be derived from ELSI metadata
- Obtaining signature
- Archiving of contract and corresponding access request procedure

# Data use


USER PORTAL
User authentication
AAI
IDP1
IDP2
IDP...
Data repository Including subject-level ELSI metadata
Data repository Including subject-level ELSI metadata
Data repository Including subject-level ELSI metadata
Archiving
Archiving
Archiving
Access management system

- 1+MG
  - Offers portal where users can authenticate themselves
  - Gives authorisation to user to process those records that were approved for access
  - Offers federated environment that appears to the user as one virtual machine rather
- User
  - Performs analysis
- National IT infrastructure
  - Logs access and stores access logs in relation to user and user’s access permission
  - Archives dataset or information on dataset version used for analysis

# Data use - IT infrastructure requirements

- Offering non-personal data to users for their purposes where reasonably possible (e.g. existing knowledge on variants; synthetic data)
- Possibility to issue, store and consume data permits: these apply to data in stored and remaining in different places
- Link approved access request to relevant data collection(s) or relevant data types and data subjects (individual records); includes handling of persistent identifiers for datasets
- Allow user operation in a virtual environment that allows operations over a distributed system or offer temporary data pooling as alternative
- User separation
- Strong (multi-factor) authentication ensuring verification of user identity, contribute link to EU eIdenties where these become available
- Only personal, no shared accounts
- Allow several users from the same controller (e.g. via user groups) as well as the possibility to include processors in the permit

# Data use - IT infrastructure requirements

- Possibility to give access only to the relevant records including only the datatypes needed of the relevant data subjects
- Ideally linking of genomic data and phenotypic data on the fly for data processing (from data on separate servers and/or different pseudonym), also decryption / encryption on the fly only when access is needed
- Possibility that external data, which is not held in 1+MG with the data holder or another data holder (e.g. public registry) can be linked to the respective records held internally for joint use.
- Possibility for collaborative approaches for data analysis where several users from different legal entities work on the same project
- Offer different user interfaces depending on the access request (healthcare professionals different to researcher; different rights / possibility to operate on the data)
- Open for discussion: virtual desktop interface vs “algorithm to the data” only
- Open for discussion: allow 1+MG held IT infrastructure only or include commercial cloud solutions

# Data use - IT infrastructure requirements

- Only properly documented, versioned and secure software can be offered for users; a comprehensive set of analysis tools / libraries has to be offered.
- Users should be allowed to run analyses from a list of approved workflows/pipelines. [suitable also for clinicians – easy visualization]
- Where users bring in own software, checked for malware/ security vulnerability
- There should be functions and processes  to validate the results if they stem from algorithms brought-in by the users.
- Establish a central container registry and the possibility to archive containers.
- Allow user management of access: suspend and resume work
- Possibility to archive and restore the project specific work environment during times of suspended activities
- Secure transfer mechanisms and safe delete in case of temporary pooling
- Logging of data operations
- Monitoring for suspicious behavior
- Possibility of data versioning in case of data deletes or rectification or updates with additional data points

# Results - IT infrastructure requirements

- Airlocked system, only supervised methods for taking aggregated data / software / results out of the protected environment
- Possibility to archive processed datasets and allow referencing for publications
- Archiving of versioned datasets with information which projects rely on which version to allow reproducibility together with archived containers
- Information portal to provide information on all data use for general public
- Automated search for publications based on mandatory reference to 1+MG
- API to allow countries to display only projects relevant for the data collection in the country

# Data use - information requirements

- Store permits / contract including expiry dates
- Information on data use
  - For projects / use application: uses version xy of dataset
  - For Data: booked or used for project / use application xy (on record level)
- Documentation on any changes to permit with time-stamped versioning
  - Regarding datasets
  - Regarding controllers and personnel with access
  - Regarding duration
- Research results
  - Information on publication linked to use request (and due to that, data subject)
  - Information on communication on publication (where applicable)
  - Information on patents / products / services derived from use request
- Healthcare results [still subject to healthcare reuse data governance]
  - Feedback on outcome

# Data use - information requirements

- Policy development results [still subject to policy development data governance]
  - Outcome report (if or if not intended outcome was achieved)
  - Link to established policies (where applicable)
- “No return” flag for incidental findings
- Log information for data use
- Documentation of exports
- Archiving requirements
- Data stability requirements by data user; needs to be complemented with a written justification
- Versions of datasets
- Reference numbers of workflows in a workflow registry where detailed information on software, data used and IT systems are stored
- Information on any data transmissions (where, when, why)

# Other elements: incidental findings

- Information management that allows messaging between users and data holders, e.g. for reporting incidental findings, requests for additional data / biosamples or recruitment of data subjects in new studies (including automatic management of “no contact” information)
- Requires correspondence between user and NCP as well as NCP to relevant bodies on national level (flexible setup – each country with their own specificities)

# General requirements - IT infrastructure

- A clear assignment of which considerations (requirements) should be covered during the software development phase and which are to be handled at deployment/operation phase (e.g. via SOPs).
- Implementation in certified compliance with ISO 27001 and ISO 27701.
- At minimum external audit, centres pooling data also certification
- Regular testing of effectiveness of safeguards
- Staff must be trained in data protection and IT security
- Data protection policy framework must be in place
- Joint agreement on minimum policies and requirements of “secure systems”
- GDPR compliant log information management (how long are logs kept, how are they stored, who has access)
- Management of time-dependent processing (data, metadata, log data) including safe delete at the end of the retention time
- Information management system that can link access requests to contracts to log files, users, containers, versioned dataset used etc. that is auditable

# General requirements - information

- Information on audits / extension of certification
- Documentation of the testing of the efficiency of safeguards
- Documentation of system changes / upgrades / incidents
- Documentation of data breaches
- Documentation of data and metadata definitions
- Recording of decisions on the different levels
- Documentation of DPIAs
- Documentation of exchanges with DPO and CISO
- Documentation of trainings
- Documentation policy framework with versioning

# How to implement?

- Who is responsible for what?
- Where are independent solutions possible?
- Where is interoperability required?
- Where are 1+MG solutions required?
- Additional requirements beyond DPbDD:
  - Where is interoperability with EHDS required? [to be an authorised participant in the EHDS]
  - Where is reuse with EHDS possible / desired?

B1MG

Beyond one Million Genomes

+ B1MG has received funding from the European Union’s Horizon 2020 Research and Innovation programme under grant agreement No 951724