RE*solution*

innovative medicines initiative

101034439 - RE*solution*

Medical Genetic Solutions for RESOLUTE

WP3 – Resource of variant-specific data on SLC genes with medical annotation

D3.3 Data Management Plan 2

Lead contributor	Andrea Garofoli (1- CeMM)
	AGarofoli@cemm.oeaw.ac.at
Other contributors	Robert Stanton (8- Pfizer) robert.stanton@pfizer.com
	Tabea Wiedmer (1- CeMM) TWiedmer@cemm.oeaw.ac.at
	Ulrich Goldmann (1- CeMM) UGoldmann@cemm.oeaw.ac.at
	Barbara Steurer (1- CeMM) BSteurer@cemm.oeaw.ac.at
	Evandro Ferrada (1- CeMM) EFerrada@cemm.oeaw.ac.at
	Álvaro Inglés-Prieto (1- CeMM) AInglesprieto@cemm.oeaw.ac.at

Document History

Version	Date	Description
V1.0	15 Nov 2021	First draft
V2.0	16 May 2022	Mid-term draft

efpia

Abstract

Data plays a central role in RE*solution*. In the preliminary stages of the project the retrieval of publicly accessible medical genetics information on human Solute Carriers (SLC) is critical to drive the decision-making process. Later, a broad range of data is generated by the joint efforts between experimental and computational analyses. Therefore, a thorough planning of all the pivotal aspects of data management, such as curation, integration and sharing strategies, is imperative.

The aim of this Data Management Plan (DMP) is to outline a set of guidelines on how the information, either collected from open access sources or produced throughout the project, will be handled to achieve the best possible research quality. Said guidelines are strongly influenced by FAIR's principles (See *FAIR*) to promote best practice in research and enhance the openness of data.

The collaboration between RE*solution* and RESOLUTE is essential for data management. The infrastructure developed in the past years by RESOLUTE lays the foundation for the RE*solution* computational platform, therefore the data management Work Packages (WP) from both the consortia will work alongside to update and improve this infrastructure to accommodate the new information and to adapt to the evolving demands of the projects. Data collected and generated by RE*solution* will, respectively, be integrated within the Knowledgebase (Task 3.2) and the Database (Task 3.3) (See *Data Repository*). The RESOLUTE web portal (https://re-solute.eu) will act as the main hub for the direct access and the visualization of the information. Since data for both RESOLUTE and RE*solution* is stored on the same platform, data access separation will be ensured by implementing different authentication privileges based on project affiliations of each consortium partner.

The DMP is not only the result of the combined efforts between RESOLUTE and RE*solution* but also acts as a living document, expected to be updated throughout the years to meet the new needs both projects will face during their lifetime. The first version of this document was submitted after 6 months from the start of the project (Deliverable 3.1). This document (Deliverable 3.3) is a mid-term version of the original deliverable, and it includes all the changes and the updated acquired during the project's first year. The third and final version will be presented at the end of the second year (Deliverable 3.4).

Data types

Data of various types will be produced and/or used in RE*solution*.

This section explains the different types envisioned to have a significant role in the project and, for each dataset, a set of characteristics that will describe the data structure, standards, source, expected use, and more. For data generated by RE*solution*, the ownership is determined by the institutions stated in the tables and the involved authors.

Genetic variants data

Data Description	SLC genetic variants identified from NGS variant calling analyses
Data Level	Derived (processed)
Data Format	Tables, tab separated values (open, human readable format)
Data Source	gnomAD (publicly accessible datasets)
Meta Data	- Description of the collected information - Database version / date of data collection
Quality Control	Information retrieved from a curated source, no further QC needed
Data Origin	- Task 1.1 (Data mining for WP1 deliverables) - Task 3.2 (Data mining for development of Knowledgebase)
Data Author	Publicly accessible sources. Collected by all consortium members
Data Relation & Utility	- Used to support the decision-making process in Task 1.2, Task 1.3 and Task 1.4

	- Data will be included in the RESOLUTE/RE*solution* Knowledgebase, Task 3.2 - Used to support the analyses in Task 4.1 and Task 4.2
Data Volume	Less than 1 GB
Data Sustainability	Data can be collected from the following public repositories: - Genome Aggregation Database (gnomAD)

Human omics association data

Data Description	Genetic human variants-disease associations
Data Level	Interpreted (analysed)
Data Format	Tables, tab separated values (open, human readable format)
Data Source	Publicly accessible datasets
Meta Data	- Description of the collected information - Description of scoring system (if applicable) - Database version / date of data collection
Quality Control	- Individually defined by each dataset. - Quantification of variants' impact - Filtering based on trait/phenotype specified in each association
Data Origin	- Task 1.1 (Data mining for WP1 deliverables) - Task 3.2 (Data mining for Knowledgebase development) - Task 4.2 (Data mining for WP4 deliverables)
Data Author	Publicly accessible sources. Collected by all consortium members
Data Relation & Utility	- Used to support the decision-making process in Task 1.2, Task 1.3 and Task 1.4 - Data will be included in RE*solution* Knowledgebase, Task 3.2 - Used to benchmark and train variant effect predictors (VEPs) in task 4.2
Data Volume	Less than 1 GB
Data Sustainability	Data can be collected from the following public repositories: - ClinVar - Genebass - IEU open GWAS project - LitVar - Open Targets - Orphanet - Uniprot - Priority Index - ENSEMBL

Variant functional characterization

Data Description	Characterization of the impact of genetic variants on SLC function
Data Level	Primary (raw), Derived (processed), Interpreted (analysed)
Data Format	Based on each assay protocol (relative fluorescence unit, IC50 value, response curve, etc.)

	In general: Tables, tab separated values (open, human readable format)
Data Source	Experiments performed in RE*solution WP2*; SLCs and variants are decided throughout RE*solution WP1*; Raw data (e.g., reagents, cell lines, etc.) is produced in Task 2.1
Meta Data	- Assay description and protocol - Readout format, quantification and normalization - Cell line libraries - List of reagents (including seller, shipping coordinates, quantities, etc.)
Quality Control	For generated cell lines: IF co-localization with organelle markers, western blot. For assays: standardized quality control parameters individually defined by each assay (Z' score, stable uptake, etc.)
Data Origin	Task 2.1 and 2.3 (Functional characterization of selected SLC genetic variants)
Data Author	AXXAM, CeMM
Data Relation & Utility	- The list of SLCs is provided by Task 1.2, variants are chosen based on Task 1.3. Variants are also provided from DMS analysis, from Task 2.3 - Used to select the SLC and variants to further investigate in Task 2.4 - Variants' impact will be included in the RE*solution* Database, Task 3.3 - Used to benchmark and train VEPs in task 4.2
Data Volume	< 1 TB (expected)
Data Sustainability	- Image Data Resource (IDR) - PubChem Bioassay

Deep Mutational Scanning (DMS)

Data Description	One by one, each position of a protein amino acid chain is systematically mutated in each of the other 19 amino acids.
Data Level	Derived (processed)
Data Format	Based on DMS chosen screening assay (e.g.: assays based on survival screening, including transport-mediated drug toxicity or substrate depletion, dyes, or transcriptional pathways activation).
Data Source	SLCs selected in Task 1.4; Raw data is produced in Task 2.1
Meta Data	- Library preparation protocol - Screening protocol - Cell line libraries - List of reagents (including seller, shipping coordinates, quantities, etc.)
Quality Control	- Quality assessment based on the chosen assays - Comparison of results with literature
Data Origin	Task 2.2 (Deep mutational scanning of selected SLCs)
Data Author	CeMM
Data Relation & Utility	- SLC that will be analysed are chosen in Task 1.4 - Based on the screening outcome, a selected number of variants will be analysed in Task 2.3 - Used to benchmark and train VEPs in task 4.2
Data Volume	< 1 TB (expected)
Data Sustainability	- Used to benchmark and train VEPs in task 4.2

SLC homology models

Data Description	3D homology modelling of SLC proteins
Data Level	derived (processed)
Data Format	PDB (community accepted, open, human readable format)
Data Source	Computed; collected from public repositories (PDB, HGNC, AlphaFold)
Meta Data	- Sequence alignment - Modelling script / parameters
Quality Control	- Accuracy of the template/target alignment - Structure evaluation scores (PROCHECK and PROQM) - Enrichment calculations to assess the models' utility for virtual screening
Data Origin	Task 4.1 (Structural and functional interpretation of SLC variants)
Data Author	CeMM, UniVie
Data Relation & Utility	- Used in Task 4.1 to map variants (either collected from literature or found in DMS analyses) in the structure of SLCs - Used in conjunction with predictions obtained from Task 4.2 to derive hypotheses on the effect of variants - Integrated within the SLC variant compendium produced by Task 4.3
Data Volume	
Data Sustainability	Publicly accessible models can be collected from the following repositories: - PDB - HGNC - AlphaFold Selected structural models will be submitted to the following public repositories: - ModelArchive (https://modelarchive.org)

Variant effect prediction

Data Description	Prediction of pathogenicity of SLC variants
Data Level	Interpreted (analysed)
Data Format	Tables, tab separated values (open, human readable format).
Data Source	Computed; collected from public databases (dbNSFP); collected from VEP-specific databases
Meta Data	- Source VEP - VEP class (unsupervised, supervised, empirical, metapredictor) - Training data (if supervised) - Classifier method (e.g., random forest, logistic regression, etc.) - Scoring algorithm
Quality Control	- Benchmark of source VEP on known pathogenic variants
Data Origin	Task 4.2 (Variant effect prediction using machine learning)
Data Author	CeMM, Pfizer, Bayer
Data Relation & Utility	- Integrated within the SLC variant compendium produced by Task 4.3
Data Volume	< 10 GB (expected)

Data Sustainability	- GitHub

As previously stated, the information outlined herein can potentially change in the next versions of the document, just like the rest of the DMP, based on the needs of the project. The tables are expected to be modified on regular basis, both in terms of number and content, as the project evolves over the months.

FAIR

RE*solution* wants to promote a culture of openness and actively join forces with all the research community.

The FAIR principles were first defined and published in 2016. Their goal is to provide a set of guidelines focused on the harmonization of research data curation, to ultimately enhance a programmatic and open use of all the information. The FAIRification of data is based on 4 concepts:

- Data must be ***F****indable*
- Data must be ***A****ccessible*
- Data must be ***I****nteroperable*
- Data must be ***R****e-usable*

RE*solution* is actively and prospectively making its data *FAIR*. An accurate annotation of all data involved in the project is provided by the version-controlled *metadata* files generated for each dataset and stored in the RESOLUTE/RE*solution* database. We will establish the compliance for each of these guidelines:

Findability

Findability will be ensured using identifiers (IDs), used to tag and quickly retrieve individual entities within a dataset. In order to achieve this, IDs must be persistent, unique and must follow a well-defined structure. For the sake of consistency, RE*solution*-exclusive data reflects the ID schema developed in RESOLUTE, with one additional field to indicate the association with the project (See *Data Identifiers*). Furthermore, the Open Research and Contributors IDs (ORCIDs) will identify the members actively involved in the consortium.

Naming conventions are also important to enhance the findability of information. For each dataset, it will be evaluated whether a naming convention already established in the research community which can be applied within RE*solution* exists or not. If it does not, a novel convention will be created within the project for the specific data type. Individual repositories, including the internally developed RESOLUTE/RE*solution* web portal (See *Data Repository*), will provide keyword-based search options to the research community.

Metadata files encompassing all the coordinates to easily localize the data will be used to annotate each dataset. As for the naming conventions, metadata standards established in the research community (Dublin Core, ISA, MIBBI, etc.) will be used, whenever applicable. Versioning numbers will also be provided to identify specific instances of data. Established semantics standards will be adopted for each data type.

Data Identifiers

The use of identifiers is imperative to produce findable and accessible data. In order to avoid discontinuity of information, RE*solution* will recreate the data ID schema adopted by RESOLUTE, based on principles and recommendations broadly accepted in life sciences research.

A RESOLUTE ID starts with two letters, specifying the type of entity (e.g., CE for cell line). The main part consists of four alphanumeric characters specifying the actual entity. In the end, an optional checksum character

adds redundancy and allows spotting typos. The main part encodes an incremental number, using the digits 0-9 and the letters A-Z with the exception of D, I, J, L, O, Q, U and Y because of their similarity to other digits or letters. This results in an alphabet of 28 symbols and thus a total of up to 614,656 words using four characters. The checksum character is taken from an extended alphabet (including Q, U and Y) and is calculated as modulo 31 (i.e., the next prime number). RE*solution*-specific data will introduce an additional field, to identify its belonging to the project and easily tell it apart from RESOLUTE. The ID is not case sensitive.

Accessibility

Accessibility will be ensured by a data sharing framework able to connect the userbase with the information. The web portal serves as the main access hub and provides several navigation tools to ease data exploitation; publicly accessible data will be freely accessible by anyone from the Knowledgebase pages (https://re-solute.eu/knowledgebase), while data produced throughout RE*solution* will be provided through the Database pages (https://re-solute.eu/database) (See *Data Repository*). An Application Programming Interface (API) will be provided to allow programmatic access to the information, directly from the Knowledgebase/Database (Task 3.5, Deliverable 3.3). To enforce the separation between RESOLUTE and RE*solution*, an authentication system will be implemented to prevent the access to RE*solution* restricted data to members which are not part of the consortium (See Data Separation).

In the later stages of the project, published data will be also submitted to domain-specific, public repositories to increase its accessibility, sustainability, and visibility. Peer-reviewed articles published by RE*solution*, together with the related links, result data dashboards and reports will also be available to the general userbase from the web portal's open access pages. The release of all the data will be the result of a dissemination procedure where all the consortium members will be involved in the review and the approval of the information (*See Data Dissemination*). Before the official release, data will only be accessible to the consortium members (See *Secure Access*).

Metadata will outline detailed information regarding access rights, related repositories, and potential licensing information.

Data Repository: Knowledgebase and Database

Data collected, generated and processed throughout RE*solution* physically resides on the CeMM storage server and the Microsoft Azure Cloud. The storage infrastructure was developed by RESOLUTE and is based on the compartmentation of information between a Knowledgebase and a Database.

The RE*solution* **Database** (Task 3.3) includes all the data processed within the project and classified either as *primary*, *derived* and *interpreted*. It will also incorporate non-data-specific notions (e.g., WP1's SLC selection process and outcome, assays used in WP2, etc.), materials used (e.g., reagents, cell lines, etc.) and events (e.g., experimental workflows, shipping details, measurements parameters, etc.). Both the Database and the stored data are designed to enhance the findability, accessibility, interoperability, and re-usability (See *FAIR*) of all the information. Every entry stored in the Database is only available to RE*solution* consortium members, in line with the agreements on confidentiality obligations. More information regarding knowledge management and protection of results can be found in the RE*solution* Consortium Agreement (CA).

The RE*solution* **Knowledgebase** (Task 3.2) principally includes information tied to human genetic profiling, structural-activity relationship, and disease association for all human SLCs. Data collected and cross-referenced from publicly available resources is integrated into the Knowledgebase already established by RESOLUTE to offer an expanded portrait for each solute carrier. Moreover, the Knowledgebase will progressively incorporate data from the RE*solution* Database and make it available to the public through the release of scientific publications and the development of a comprehensive compendium of *interpreted* data (e.g., pathogenicity of genetic variants, biochemical insights, potential drug targets, etc.) (Task 4.3, Deliverable 4.4). The Knowledgebase is freely accessible to the public via the web portal and available for programmatic data mining through its API (Task 3.5).

Whilst there is a clear separation between these two storages, a tight interaction between the information they incorporate is required. To ease this interplay, the data is cross linked and cross referenced through the respective metadata.

Data Dissemination

All data collected, produced or processed within the project must undergo quality control (See *Data Quality Assessment*) and annotation (See *Data Annotation*) procedures in order to be disseminated. The standard procedure requires that two representative of each consortium members are notified, by the data authors, about the intention to release the newly produced dataset. The representatives will proceed with the reviewing of both the data and the respective documentation. This reviewing step can be rejected by each individual reviewer in case of conflict of interest.

More details on the process will be provided by the *Dissemination and Exploitation Plan* (DEP, Deliverable 5.2). One of the goals of the DEP is to define the internal procedures of notification and agreement for efficient sharing of data and resources produced throughout the project. Moreover, the document also encompasses the efforts focused on the visibility enhancement of the project.

- **data release:** Any participant can propose data for public release by submitting a written *data release proposal* to the Work Package 3 team, specifying the following details:
    - an abstract on the data,
    - the motivation for publication,
    - the data identifiers,
    - the annotation status,
    - the standards compliance,
    - assessment of quality control,
    - license to use for publication.

  Work Package 3 team will collect these proposals for further discussion in *data annotation and release workshops*, taking place regularly at consortium meetings or possibly also as tele conferences in between. Participation at the workshop is voluntary, and all participants will receive the collected data release proposals two weeks before the workshop. At the workshop, the proposals will be discussed in detail with the aim of curating data, extending and refining annotation and strengthening quality control, but there will be no immediate decision taken on publication. Within a week after the workshop, Work Package 3 team together with the proposal's authors will send out a refined version of the data release proposal to the *publication and dissemination officers* of all consortium partners. In case there are no objections within a 30 days review period, the data will be published and openly accessible in the RE*solution* Database. Minor objections can be directly resolved with the objecting participant, while major objections require an adaption of the proposal, effectively restarting the review period.
- **supporting a scientific publication:** Data might also be disseminated as integral part or supplemental material to a scientific, peer-reviewed publication. For this route, the same *data release proposal* as described above has to be submitted to Work Package 3 team, but a data annotation and release workshop is not required. All scientific publications will follow the *gold* or *hybrid* open access policy by submitting the manuscripts to open access journals or by paying for this option in the subscription-only journals. For more details on dissemination in the form of scientific publications, please refer to the *Dissemination and Exploitation Plan* (Deliverable 5.2).
- **end of project:** At the end of the project (i.e. June 2023) we intend to release and make openly accessible a vast majority of all data within the RE*solution* Database that was not published before.

Data Sharing

All the consortium beneficiaries can freely share all data collected, produced, and processed through the Knowledgebase/Database developed for RE*solution*. Moreover, the web portal will act as the user interface for the retrieval of the information deposited in the data storage, also programmatically accessible using the API.

The main means of communication and information sharing are the RE*solution* SharePoint (for documents, protocols, presentations, etc.), emails and teleconferences (mostly focused on WP progress updates). RE*solution* will implement Message Digest 5 (MD5) checksums and regular backups as safe measures to ensure the consistency of information and to prevent accidental information loss.

Details on the sharing process with the research community is explained in the *Dissemination and Exploitation Plan* (DEP) (see *Data Dissemination*).

Interoperability

Interoperability will be ensured by careful planning during data preparation. All data processed throughout RE*solution* will adopt de facto standards approved by the research community, both in terms of data formats and data structures. This will ease information exchange between different WPs and allows the use of a wide range of analysis technologies. In the perspective of openness of information, this philosophy makes it possible for RE*solution* to be a polestar for SLC research even after its conclusion by providing rich and comprehensive SLC knowledge to the research community.

Metadata will provide a description of the chosen standards and links to the respective manuals.

Data Annotation

Data documentation is essential for the findability, interoperability and the re-usability of information. Whenever it is possible, RE*solution* will adopt community derived standards for the annotation. If no standard can be found for a dataset, authors will also provide additional documentation for the used annotation structure.

When data undergoes the submission process, annotation and metadata files of the respective dataset are reviewed and, possibly, further curated by the consortium members.

Re-Usability

Re-Usability will be ensured adopting the Creative Commons ‘Attribution’ (CC BY 4.0) open access license model, allowing the community to re-use the published data and, potentially, recreate all the experiments. Protocols and workflows used by the researchers of RE*solution* will include automated data quality control phases and downstream manual curation to establish the integrity of the information (See *Data Quality Assessment*).

Open access to data will follow the dissemination procedure envisioned for the project and its preservation will be ensured for at least 5 years after the official end of the consortium (See *Data Dissemination* and *Data Sustainability*).

Metadata will include details concerning data preparation and the data processing workflows, to guide through the reproduction of data analyses.

Data Lifecycle

The first phase of data lifecycle encompasses the collection or the production of knowledge. The DMP provides additional details regarding the means of information retrieval, its format and the list of the authors (See Data Types). The same information will be included in the dataset-specific metadata.

The inclusion in the Knowledgebase/Database can only occur if data satisfies an initial quality control evaluation, proving it meets the quality standards defined for RE*solution* (See *Data Quality Assessment*). The storage server hosted by CeMM and the Microsoft Azure Cloud will be the internal repositories of the project (See Data Sharing). Microsoft Sharepoint is an alternative information sharing platform, more suitable for small files used for communication (e.g., documentation, protocols, etc.).

The contingent release of data to the general public and the upload to public repositories are guided by a dissemination procedure (See *Data Dissemintion*). *Data release workshops*, either held during consortium conferences or through online meetings, will offer the opportunity to all the partners to review, further curate and agree on the public release of data.

Data Quality Assessment

In order to achieve the high-quality research standards RE*solution* strives for, the use of a 3 layers data quality control assessment is enforced.

1. **Quality score**: data-specific and community established quality control thresholds for each dataset generated or collected (see *Data Types*). Only the entries that reach said thresholds will be used for downstream analyses.

2. **Internal review**: for the data generated internally within the project, the public release is initiated with a *data release proposal*, where the authors annotate and present to the rest of the consortium the results of the internal quality assessment (see *Data Dissemination*).

3. **External review**: only in case of data submitted to public repositories due to the involvement in a scientific publication, the peer-review process offers an additional level of quality control.

Computational analyses

A strategy for the ETL (extract-transform-load) computational analyses has been outlined in the Data Modelling document (Deliverable 3.2). The proposed approach is based on pipelines developed using a Workflow Manager (Nextflow), executed using Cloud Computing (Azure Cloud), and operated by Containerized software (Docker). The interaction between these three elements will ensure the *FAIR*ness and the reproducibility of computationally produced data. Please refer to the Data Modelling document for additional details.

Data Sustainability

All data entering the RE*solution* Database are stored for the course of the project and will be actively supported by CeMM for at least 5 more years after its conclusion. Long-term data preservation of information produced during the project's lifespan will be ensured by the submission to domain-specific public repositories established within the research community (e.g., Gene Expression Omnibus for functional genomics data, ChEMBL for bioactive molecules, Cell Image Library for cell-specific imaging data, etc.). Details regarding said public resources are listed in the *Data Type* paragraph (See *Data Types*).

Eventual novel techniques, protocols and standard operation procedures developed within the project will either be published in open access (gold) journals or described in the respective dataset's metadata and deposited in the RESOLUTE website.

Allocation of Resources

The data volume expected to be integrated in the RE*solution* storage is a total of up to 4 TB (with a production rate of 1 TB per semester). CeMM hosts the server infrastructure, which will be integrated with the Microsoft Azure Cloud. Storage costs, combined with the High-Performance Computing (HPC) systems used throughout the project, are estimated to be about 24,000 EUR. This includes all the operational costs needed to ensure data's FAIRness (See *FAIR*).

RE*solution*'s WP3 is responsible for data management. Monthly online meetings will be hosted, in alternance with RESOLUTE's WP8, to inform members of both the consortia about all the progresses related to data management in RESOLUTE and RE*solution*.

Data Security

Whilst production or use of sensitive data (e.g., non-anonymized patient information) is not anticipated in RE*solution*, careful planning of the data infrastructure's security is crucial for the proper pursuance of the project's goals.

Secure Storage

RE*solution* will adapt the same strategies implemented by RESOLUTE. Documents like Standard Operating Procedures (SOP), protocols, reports, or meeting minutes, are collaboratively written and stored in a Microsoft SharePoint environment via a cloud-hosted service (Microsoft Office 365).

Data (primary, derived and interpreted data) are stored at local IT infrastructure of CeMM and also in the Microsoft Azure Cloud environment (Azure Blob Containers) which fulfils the rules of the EU General Data Protection Regulation (GDPR). Both resources are managed by the CeMM IT department. Via inclusion on Microsoft Azure Cloud storage for primary (raw) data, the full scale of estimated required storage volume for RE*solution* (see ***Error! Reference source not found.***) is available.

To minimize the risk of data loss, a local backup strategy as well as a cloud backup strategy is in place. The local backup runs incrementally and writes to magnetic tapes (IBM Backup system). Data in the Microsoft Azure Cloud is deposited in zone-redundant storage, i.e., replicated to three separate physical locations with independent power, cooling, and networking. To rule out data manipulation or corruption we employ MD5 checksums.

Access to the CeMM storage server is strictly restricted and available only from within the CeMM intranet, while access to the Microsoft Azure Cloud storage is controlled via a carefully maintained user list at an Azure Active Directory, allowing participants to submit or review internal data. On the other hand, Shared Access Signatures (SAS) that are created on-demand provide a secure access for public users to view and download published data.

Secure Access

Participants of the RE*solution* consortium have access to all the documents and information in SharePoint via a personalized RE*solution* account, managed via the Microsoft Azure Active Directory service. Implementing an Open Authorization 2.0 (OAuth2) workflow, the same RE*solution* account is used to securely access data and information in the internal sections of the RESOLUTE and the RE*solution* Database and Knowledgebase (see *Data Repository*).

Authenticated as well as unauthenticated (i.e., public) access to data and information in the RESOLUTE web portal, the RE*solution* Knowledgebase, and the RE*solution* Database is protected by implementation of transport layer security using the secure communication protocol (HTTPS) and on-demand generation of SAS codes to Microsoft Azure Storage blobs.

Secure Transfer

Data is produced by individual participants locally but managed centrally at the CeMM. Therefore, fast and secure data transfer is required. To allow remote access to data sets and transfer via the internet to the CeMM, we employ three strategies:

1. As initial solution, we used the Microsoft SharePoint environment, which provides an easy-to-use interface for data down- and upload. However, transfer of large files or transfer of a bulk of files turned out to be cumbersome, and data at the SharePoint has to be moved manually from or to the CeMM storage server.

2. As additional solution, we set up a SFTP server with one dedicated user account per participant, allowing for easy programmatic transfer of many and/or large files. Again, data at the Secure File

Transfer Protocol (SFTP) server must be moved manually from or to the CeMM storage server on demand.

3. The Microsoft Azure Cloud is securely accessible from every participant and data transfer can be automated programmatically. Also, direct download links can be generated to be used in the RESOLUTE web portal, the RESOLUTE and RE*solution* Database or Knowledgebase.

Data separation

RE*solution* will use the storage infrastructure developed for RESOLUTE. Conditional access based on the Azure Active Directory authentication system will be implemented to separate the data spaces for the two projects. Access to the RE*solution* Database will be only possible for the RESOLUTE members that are involved in both the projects. However, access to the Knowledgebase is possible for every member of either the consortia (See *Data Repository*).

Ethical and Legal Aspects

RE*solution* requires careful reviews for ethical and legal compliance for all the data retrieved from public repositories (e.g., for the inclusion of disease association data in the RE*solution* Knowledgebase) or generated throughout the project.

Descriptions of the ethical aspects are individually covered by each deliverable, while the general legal considerations are outlined in the CA.

Appendix

Abbreviations

API	Application Programming Interface
CA	Consortium Agreement
CC	Creative Commons
DEP	Dissemination and Exploitation Plan
DMP	Data Management Plan
FAIR	Findability, Accessibility, Interoperability, and Reusability
GDPR	General Data Protection Regulation
HTTPS	Hyper Text Transfer Protocol Secure
HPC	High-Performance Computing
ID	Identifier
IMI	Innovative Medicine Initiative
MD5	Message Digest algorithm; used for calculating file checksums
OAuth2	Open Authorization 2.0
ORCID	Open Researcher and Contributor Identifier
RESOLUTE	Research Empowerment on Solute Carriers
SFTP	Secure File Transfer Protocol
SAS	Shared Access Signatures
SLC	Solute Carrier

SOP	Standard Operating Procedure
W3C	World Wide Web Consortium
WP	Work Package

Consortium Members

short name	**full name**
CeMM	CeMM Research Center for Molecular Medicine of the Austrian Academy of Sciences
UoX	University of Oxford
UoL	University of Liverpool
AXXAM	Axxam SpA
ULei	Universiteit Leiden
MPIMR	Max-Planck Institut für medizinische Forschung
UniVie	Universität Wien
Pfizer	Pfizer Ltd.
Bayer	Bayer AG