101172693 – VICT3R

Developing and implementing Virtual Control Groups to reduce animal use in Toxicology Research

WP 1 – Project coordination and management

D1.2 Initial Data Management Plan (DMP)

Lead contributor:	Manuel Pastor (1-UPF) Juan Manuel Ramírez Anguita (1-UPF) Jorge Lemos Portela (1-UPF)
Other contributors:	Ferran Sanz (1-UPF)
Due date:	28/02/2025
Delivery date:	23/04/2025
Dissemination level:	Public (PU)

Contents

Document History

Version	Date	Description
V0.1	25 Feb 2025	First internal draft
V0.2	27 Feb 2025	Second internal draft
V0.3	3 March 2025	Third internal draft, including comments from FS
V0.4	5 March 2025	Version for formal review
V0.5	6 March 2025	Version including comments and suggestions from internal reviewers. The English grammar and spelling were reviewed, correcting minor errors.
V0.6	10 March 2025	Version for consortium review
V1.0	23 April 2025	Final version integrating feedback from partners

Definitions

Participants of the VICT3R Consortium are referred to herein according to the following codes:

1. UNIVERSIDAD POMPEU FABRA (UPF)
2. SYNAPSE RESEARCH MANAGEMENT PARTNERS SL (SYNAPSE)
3. FRAUNHOFER GESELLSCHAFT ZUR FORDERUNG DER ANGEWANDTEN FORSCHUNG EV (ITEM)
4. GRITSYSTEMS AS (grit42)
5. TECHNISCHE UNIVERSITAT DORTMUND (TUDO)
6. PHUSE CUIDEACHTA FAOI THEORAINN RATHAIOCHTA (PHUSE)
7. MEDBIOINFORMATICS SOLUTIONS SL (MBIS)
8. BARCELONA SUPERCOMPUTING CENTER CENTRO NACIONAL DE SUPERCOMPUTACION (BSC CNS)
9. UNIVERSITAT KONSTANZ (UKON)
10. DECIPHEX LIMITED (Deciphex)
11. ORGANON SRL (Organon)
12. SYNCWORK AKTIENGESELLSCHAFT (Syncwork)
13. VRIJE UNIVERSITEIT BRUSSEL (VUB)
14. BAYER AKTIENGESELLSCHAFT (Bayer)
15. MERCK KOMMANDITGESELLSCHAFT AUF AKTIEN (MKDG)
16. AMGEN RESEARCH (MUNICH) GMBH (Amgen)
17. ASTRAZENECA UK LIMITED (AZ)
18. BASF SE (BASF SE)
19. BOEHRINGER INGELHEIM INTERNATIONALGMBH (BII)
20. BRISTOL-MYERS SQUIBB COMPANY CORP (BMS)
21. INCYTE BIOSCIENCES DISTRIBUTION B.V (Incyte BD)
22. INSTEM SCIENTIFIC LIMITED (Instem)
23. IPSEN INNOVATION SAS (IPSEN)
24. JANSSEN PHARMACEUTICA NV (Janssen)
25. NOVARTIS PHARMA AG (Novartis)
26. NOVO NORDISK A/S (Novo)
27. ORION OYJ (Orion)
28. PFIZER INC (Pfizer)
29. F. HOFFMANN-LA ROCHE AG (ROCHE)
30. SANOFI-AVENTIS DEUTSCHLAND GMBH (Sanofi)
31. TAKEDA PHARMACEUTICALS INTERNATIONAL AG (TPIZ)
32. UCB BIOPHARMA (UCB)
33. ABBVIE INC (AbbVie)

Grant Agreement	The agreement signed between the beneficiaries and the IHI for the undertaking of the VICT3R project (101172693).
Project	The sum of all activities carried out in the framework of the Grant Agreement.
Consortium	The VICT3R Consortium, comprising the above-mentioned legal entities.
Consortium Agreement	Agreement concluded amongst VICT3R beneficiaries for the implementation of the Grant Agreement (GA). Such an agreement shall not affect the parties' obligations to the Community and/or to one another arising from the GA.

Executive Summary

This document presents the VICT3R project's Initial Data Management Plan, which outlines the actions directed at managing the project data in accordance with the FAIR principles and within the constraints imposed by the involvement of proprietary data and the conditions laid down in the Contract Agreement.

1. Data summary

VICT3R is a knowledge-based project that will use previously generated data describing the characteristics of non-treated animals in preclinical studies (controls) to generate virtual equivalents (virtual controls) with the potential to replace real animals in future studies. Therefore, data is at the project's core, and appropriate data handling is critical for the project's successful completion.

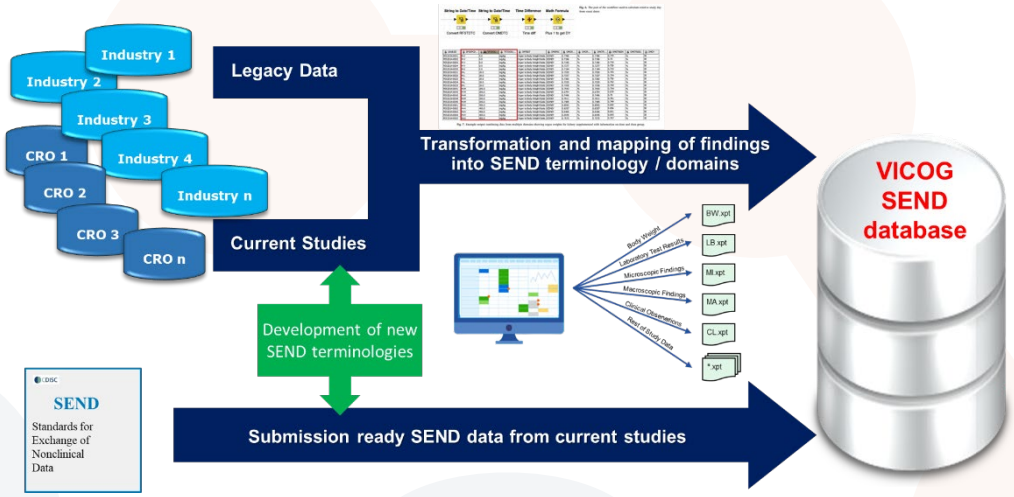

Figure 1. Workflow for data extraction and transfer to the VICOG DB, modified from Golden E et al., 2023.

The VICT3R project used as a head start the preliminary work done in building a control animals database (named VICOG DB) started by the end of the IMI eTRANSAFE project [1]. This database was continued by a subgroup of eTRANSAFE partners, who are also involved in the VICT3R project. Figure 1 illustrates how this database has been constructed. The VICOG DB [2], which included a significant set of curated data, was made available to the VICT3R consortium right at the start of the project.

All the data used in the project will be collected either from previously performed preclinical studies or from routine studies carried out during the project life. These data will be, therefore, donated by the pharmaceutical companies participating in the project as partners (data-donor partners). Due to regulatory requirements, the data obtained in preclinical studies is usually collected using the CDISC SEND format [3]. All donated data will in principle be integrated in the VICT3R database. Therefore, we intend to maintain the same format and organize the data import and database construction around

the original definition of SEND domains. In cases where the original SEND format fails to represent the data appropriately, we will extend the SEND standard in collaboration with PHUSE and CDISC.

Regarding the volume of new data generated, according to the Consortium Agreement, we foresee collecting control data from 3739 legacy studies and 187 new studies. This amounts to control data from nearly 4 thousand studies (3926). We will collect, curate, and store data for multiple SEND domains at the individual animal level for each study.

The data collected and curated during the project's life will constitute the core of a software platform that will be used to generate virtual control data for future preclinical studies. If properly maintained, this system can be used in the future by any pharmaceutical, biomedical company, or academic institution carrying out animal studies, allowing them to reduce the use of real animals in research significantly. This means that the data is highly valuable since it will support a methodology for replacing a significant number of animals. The final model of accessibility to the platform after the end of the project is still to be decided. This work is being carried out by the Sustainability Taskforce in the WP10 Dissemination and Sustainability framework. Opening access to third parties under a commercial agreement will be considered.

2. FAIR data

The FAIR principles for scientific data management and stewardship [4] will be applied in building and managing the VICT3R data corpus, in line with incorporating these principles in industrial pharmaceutical R&D [5]. The goal of making the data FAIR will be achieved by using relevant data standards (e.g., SEND), annotating the data with all the metadata required, assuring the preservation of the data during and beyond the project execution, and defining and publishing clear and fair rules for the access to the data by interested users from inside and outside the VICT3R consortium.

2.1 Making data findable, including provisions for metadata

The VICT3R database is a dynamic resource that will grow with the incorporation of data from the partners during the life of the project and beyond. As such, it cannot be associated with a persistent identifier. However, a whole set of dissemination and outreach activities are being planned by WP10 and will be carried out to raise awareness of the existence of this resource to any relevant stakeholder. In particular, to enable these activities, a Sustainability Task Force was set up last February and will meet monthly. Further details can be found in the D10.2 First plan for exploitation and sustainability activities, which will be submitted shortly.

No widely accepted metadata standards could be applied to the database's contents. However, the database software used in the project (grit42) facilitates the annotation of each data point with the study design characteristics, the animal, the facility, and any other potential covariate of relevance. The platform used to store the data has been designed to facilitate data querying and extraction using these variables to generate subsets of control data that are reusable in different contexts.

2.2 Making data accessible

Repository:

The VICT3R database will be made available and preserved through the resources provided by the project's coordinator (UPF). The main database will use software developed by partner grit42, which will allow secure online access to the data and be connected to the software tools in charge of generating the virtual control groups.

A dedicated IT infrastructure is already in place at UPF's premises, which will host a data lake for raw data uploading and storage, the database and all the required software. The technical details of this infrastructure have been documented in VICT3R-ITdoc-17.docx (document accessible only to project partners).

Open access to VICT3R research output will be an important aim of project. Restricted access is implemented to safeguard sensitive or proprietary data contributed by private partners, aligning with legal and contractual obligations. This ensures compliance with the project's Grant Agreement and Consortium Agreement.

Accessibility to the data will be guaranteed by using standard protocols (REST API) for web access and by using a standard format (CDISC SEND). No specific or proprietary software will be required to access the data. Software tools used to extract data (R and Python scripts and libraries) will be provided as open source and made accessible at the project's GitHub repository (https://github.com/VICT3Rproject) whenever possible to make the procedure easy and reproducible. Other relevant software will be provided under open-source licenses to ensure accessibility and encourage widespread adoption.

The data access endpoints will be secured using user tokens. Access will require authentication through personal credentials, assigned and managed by UPF to authorized members of the Consortium. Database access includes multi-factor authentication processes to enhance access security.

A specific committee, DARC (Data Access Review Committee), has been established to ensure fair access to the results and enforce continuous commitment to data sharing by the VICT3R partners. The DARC will review the quantity and quality of the data contribution of industry partners and make decisions regarding the provision of access to the VICT3R database and further project assets by the VICT3R partners. The transfer of data from partner contributors will be facilitated by putting all the required technical resources in place and UPF providing all legal guarantees of data confidentiality and secure data handling.

As outlined in the consortium agreement, the data in the VICT3R database will remain available and findable for at least five years beyond the project duration, with sustainability plans in place ensuring updates and accessibility.

2.3 Making data interoperable

The joint exploitation of large amounts of data, which is usually attained by means of the integration of data from different sources and organizations, requires the standardization of the other source terminologies and data models to make them interoperable. Two main types of data will be exploited in the generation of Virtual Control Groups (VCGs): 1) Alphanumeric data from a) historical, high quality and standardised data collected from the control animals routinely used in systemic toxicity studies b) from other types of studies such as developmental and reproductive toxicology (DART), chronic/carcinogenicity studies, or certain types of pharmacological studies after checking feasibility for their use in the generation of VCGs and 2) Digital histopathological images and metadata from control animals.

After data collection into a centralized data lake, pipelines granting data standardization into the SEND (Standard for Exchange of Nonclinical Data) format will be implemented. Alphanumeric data shared within this project's scope are in SEND format or will be mapped into a SEND-like format.

SEND is a standardized machine-readable format developed by the Clinical Data Interchange Standards Consortium (CDISC) for organizing and submitting nonclinical (preclinical) study data to regulatory agencies. This standard aligns with several FAIR data management principles—Findable, Accessible, Interoperable, and Reusable. In SEND, data is organized domain-wise so that every tabulation element (e.g., CSV file, SQL table) corresponds to just one domain in the specification (SEND implementation guide- SENDIG). SEND-controlled terminology required for semantic standardization of data is maintained by CDISC, as well as relationships between elements from different domains. All in all, the use of SEND will enable massive reusability and interoperability. VICT3Rs “SENDified” data might be usable for other purposes.

In turn, digital histopathology whole slide images (WSI) and related metadata of control animals will be collected, curated, and standardized to be linked to the alphanumerical data of the respective control animals. The standardization process for these data consists of the creation of the “Digital WSI database” (DWDB) to host control digital WSI and a metadata interface connecting relevant metadata to DWDB. This metadata will be of paramount importance in the generation of VCGs as relevant metadata available for each image will be the link with records in the VICT3R database. This metadata must cover the following aspects: animal identifier available in the SENDified alphanumeric data, specimen information (type of tissue and pre-existing conditions), information on sample collection and processing, imaging modality in the capture of the histopathology image and image acquisition parameters, additional image metadata, annotations, and associated interpretations by pathologists with associated findings recorded in the microscopic MI domain, and Information about the institution where the image was acquired. This infrastructure for control animals will set the standard for easy addition of digital WSI to VCGs in prospective studies.

Unfortunately, not all the parameters required for the successful implementation of the VCG approach have been incorporated into SEND so far. Development of new SEND domains for certain targeted study types is currently ongoing. The contribution of the VICT3R consortium to the development of the SEND standard will be focused on the creation of controlled terminologies for specific variables, such as those focused on “vehicle constituents” and the beta testing of new domains not yet included

in the current SENDIG such as the Nervous System (NV) Domain. This terminology gap will be addressed with the help of the VICT3R partner PHUSE.

2.4 Increase data re-use

Most of the VICT3R data will use data standards (CDISC-SEND) that are well established in the project field, which means that most potential users already have all the software tools needed to import and use it.

VICT3R will also apply the FAIR principles in the design, development, implementation and distribution of any software tool used for data import, curation, and exploitation. The software developed in the project may also to an important extent be made available under an open-source license and will be deposited in an accessible software repository (e.g., GitHub).

During the project's life, access to the VICT3R database will be limited to the project partners. The conditions for accessing the data by third parties are still to be determined by the sustainability work package.

Regarding data provenance, every data point in the VICT3R database is annotated with respect of the preclinical study from which it was obtained. In this sense, the identification of the source is exhaustive and will require no further documentation.

The data quality will be controlled by applying a detailed data curation workflow, as described in the grant application (see Figure 2). Animals' data must be provided in SEND or SEND-like format. A secure data lake for data upload, storage, and download has been made available. Real-time monitoring of data donations will be provided through a dashboard.

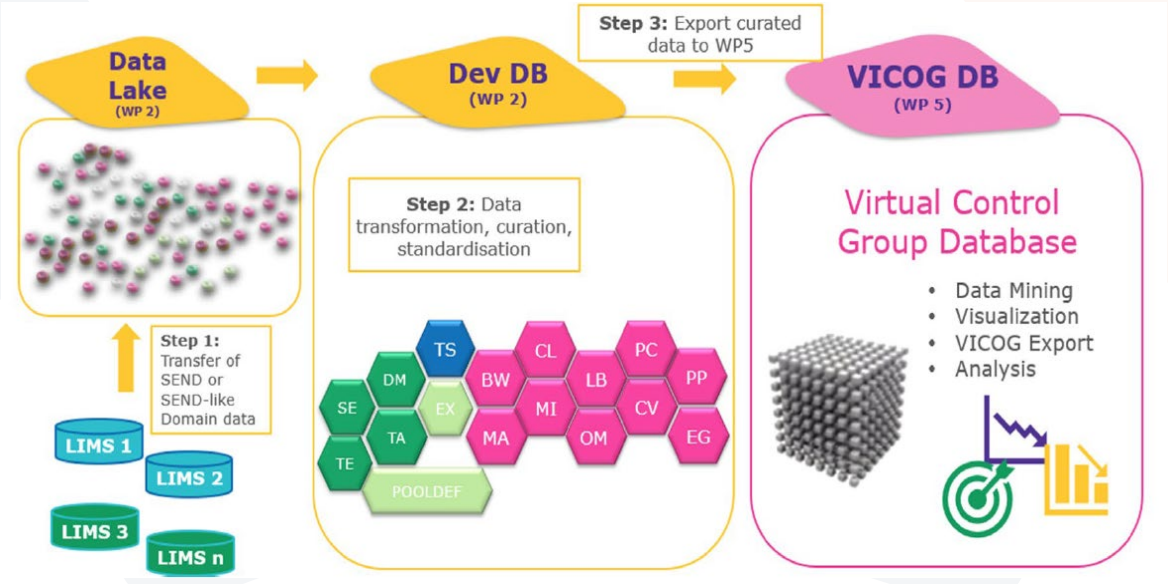

Figure 2: Data workflow of the VICT3R project (extracted from the VICT3R Grant Application).

Data uploaded to the VICT3R data lake was thoroughly revised, and semi-automated procedures were implemented by WP2 for data curation. Data donors and subject matter experts, in close collaboration with histopathology experts, participate in internal review meetings on standardized terms regarding the microscopic findings (MI) domain. Specific procedures for demographics (DE), laboratory test results (LB), and those SEND domains related to weight measurements are being developed.

The development database will be subjected to different quality protocols involving gap analyses, plausibility checks, and data density measures. Feedback on these analysis results will be given to data providers for improvement in an iterative fashion. Data contributions failing the established quality criteria will be considered inadequate for inclusion in the production version of the VICT3R.

The plan of the VICT3R consortium is to keep research outputs publicly available to an important extent. Scientific and professional publications produced will comply with the open access policy, provided that it might not compromise the IPR of the institutions sharing this data. In this regard, UPF will be in charge of maintaining the infrastructure needed for storage and will manage preservation, availability, and accessibility to the datasets.

3. Other research outputs

Apart from the control data stored in the VICT3R database, other digital resources, like software and workflows, will be shared under an open-source license, to an important extent, and made available adequately documented in an open-access repository like GitHub. A specific section for Public IT tools has also been created on the project website: https://www.vict3r.eu/public-it-tools/. Other relevant results will be used to produce public deliverables like D6.4 (End-to-end pipeline to generate VCGs combining matching and statistical analysis methods) or D9.4 (Validation reports and certificates for applied software solutions). A comprehensive list of all public deliverables can be found in the Grant Agreement.

Special attention will be paid to documenting how the VICT3R database and software platform can be used to generate virtual control groups in practice, generating user guides and training material as well as documenting examples of application in real or simulated situations, with the aim of facilitating as much as possible the use of the project outcomes by different actors (researchers in the academy, industry, regulators, etc.). All public training materials will be made available in the Training section of the VICT3R website: https://www.vict3r.eu/training/.

The project will also apply the FAIR principles to any research output. Scientific articles will be published under open access models, in accordance with the Commission policies, allowing online access to scientific information that is free of charge to the user and is reusable. For other relevant digital resources and documents, a copy will be uploaded to public repositories like Zenodo (https://zenodo.org), labelled appropriately with metadata, and assigned a unique digital identifier (DOI). Only standard formats will be applied to these resources to facilitate their reusability.

4. Allocation of resources

Since VICT3R is a data-based project, the data collection, curation, and storage resources have been allocated from the beginning. In the Grant Agreement, we declared our deep commitment to the Open Science principles, and for this reason, all the resources required to implement these principles were contemplated in the Grant Agreement.

In the Grant Agreement there is already a work package devoted to this task: WP2 Data management, curation and quality control (RDT). The leader of T2.1 (UPF) will host the data and implement all the technical measures for guaranteeing the security of the data access, recovery, and backup plans. In the Consortium Agreement UPF committed to maintain the VICT3R database accessible for at least five years, after the end of the project.

However, the sustainability plans should make possible the maintenance of the most important project results for a much longer period, including involving formulas for their commercial exploitation, and the Consortium is working towards reaching sustainable solutions for project outputs after the project end. On the one hand, the resources required for the hosting and the hardware maintenance are expected to be relatively low: UPF is an academic institution that owns the hardware used for the hardware, so there are no periodic hosting fees to be covered. On the other hand, if the project succeeds in providing solutions for generating VCGs that are able to replace concurrent controls, particularly for regulatory purposes, their use will be in high demand. This would facilitate obtaining maintenance fees from project partners and access fees from third parties. However, as only the first 6 months of the project have lapsed, it is too early to provide further details on possible exploitation models or the characteristics of the data government body. The definition of sustainability approaches is being carried out by the Sustainability Taskforce and the WP10, as described above.

5. Data security

The partner in charge of the database hosting (UPF) has put in place appropriate measures to guarantee secure access to the VICT3R database and other digital resources. There is also a plan for data backup and recovery, which has been documented in VICT3R-ITdoc-17.docx (accessible to project partners). The collection of control data from the data donating partners has been organized by UPF, who has set up a secure data lake and a data uploader application developed ad-hoc for this purpose, described in D2.1 IHI_VICT3R_D2.1_Data Lake for secure data exchange_v1.0.pdf (accessible to IHI and project partners).

The bulk of the data generated in the project will be safely hosted during the project's life at UPF's premises, with an extension of at least five years post-project.

6. Ethics

The data donated by the partners and stored in the VICT3R database was obtained from experimental animals belonging to untreated (control) groups of routine preclinical studies not directly linked to the project. No human data nor personal data of any kind is being used by the project, thus making non-applicable considerations related to informed consent or data sharing.

7. Open Science

This document mainly focuses on delineating the DMP and applying the FAIR principles for data stewardship. However, the project's commitment to open science is beyond data management. At the core of the project is the convincement that sharing increases the value of data, even for data that has traditionally been stored in corporate silos. For this reason, the deep commitment of the VICT3R project to open science will be reflected in other initiatives, for example:

- Implementing appropriate communication and dissemination measures to maximize outreach.
- Opening up the Consortium for new collaborations, including by organizing public webinars.
- Involving the European Open Science Cloud in research output and data management.
- Integrating the European Open Science Policy in the ESR initiatives.

As described above, some of these activities will be carried out by the Sustainability Taskforce and the WP10.

8. References

[1] Sanz F, Pognan F, Steger-Hartmann T, Díaz C, Asakura S, Amberg A, et al. eTRANSAFE: data science to empower translational safety assessment. Nat Rev Drug Discov. 2023;22:605-606. doi:10.1038/d41573-023-00099-5

[2] Golden E, Allen D, Amberg A, Anger LT, Baker E, Baran SW, et al. Toward implementing virtual control groups in nonclinical safety studies. ALTEX. Published online 2023. doi:10.14573/altex.2310041

[3] CDISC SEND format

[4] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3:160018. doi:10.1038/sdata.2016.18

[5] Wise J, de Barron AG, Splendiani A, Balali-Mood B, Vasant D, Little E, Mellino G, Harrow I, Smith I, Taubert J, van Bochove K, Romacker M, Walgemoed P, Jimenez RC, Winnenburg R, Plasterer T, Gupta V, Hedley V. Implementation and relevance of FAIR data principles in biopharmaceutical R&D. Drug Discov Today. 2019;24(4):933-938. doi: 10.1016/j.drudis.2019.01.008.