swissuniversities

Swiss Open Research Data Grants (CHORD): Track B
List of Approved Projects

Call Deadline: 01.07.2022

Decision by the Delegation Open Science: 09.12.2022

Overview Approved Projects Track B

Submissions: 25 Approved projects: 16, of which 3 with conditions Funding rate: 64%

Short Title	Full Title	Leading Institution	Partner Institution(s)	Project Manager	Total Costs (CHF)	Funds Requested (CHF)
ModelArchive	*Open Research Data (ORD) best practices for computational macromolecular models*	University of Basel (UniBS)	USI, UNIL, EPFL and SIB	Torsten Schwede	1'080'000	540'000
OPEM	*Open EM Data Network*	University of Geneva (UNIGE)	UNIL, UniBE and UniBS	Robbie Loewith	920'000	460'000
Swiss DBGI-KM	*Knowledge Management in the Swiss Digital Botanical Gardens Initiative*	University of Fribourg (UniFR)	UniNE, ETHZ and SIB	Pierre-Marie Allard	1'393'322	689'079
OBELIS	*Open Elite Data Project*	University of Lausanne (UNIL)		Felix Bühlmann	710'000	355'000
CRISPR4ALL	*Lentiviral arrays for the indexed activation, deletion and silencing of all human genes*	University of Zurich (UZH)		Adriano Aguzzi	993'653	496'827
UpLORD	*Upgrading the linguistic ORD-ecosystem*	University of Zurich (UZH)		Noah Bubenhofer	1'398'500	612'000
scFAIR	*scFAIR: Standardization and stewardship of single-cell metadata*	University of Lausanne (UNIL)	EPFL and SIB	Marc Robinson-Rechavi	1'412'732	673'384
AFFORD	*A Framework for Avoiding the Open Research Data Dump*	University of Zurich (UZH)		Vartan Kurtcuoglu	1'029'580	500'000
enviPath	*Fostering interdisciplinary sharing and usage of chemical biodegradation data*	University of Zurich (UZH)		Kathrin Fenner	774'000	387'000
SIMBioData	*Standardized integration of multi-omics biomedical data*	University of Basel (UniBS)	EPFL	Mihaela Zavolan	1'498'663	749'200
RE2VITAL	*REuse and REproducibility of intraVITAL microscopy via open data practices in image-driven immunological research*	Università della Svizzera Italiana (USI)		Santiago González	806'064	324'800
ODTPR-SMS	*An Open Digital Twin Platform for Research on the Swiss Mobility System*	Swiss Federal Institute of Technology Zurich (ETH Zurich)		Gloria Romera	1'500'000	750'000

ORDEA	*Open Research Data Environments for the Arts*	University of Zurich (UZH)	ETHZ, UniBS and UniBE	Tristan Weddigen	1’336’000	653’000
Swiss-AL	*Swiss-AL: Linguistic ORD Practices for Applied Sciences*	Zurich University of Applied Sciences (ZHAW)		Julia Krasselt	1’058’104	461’527
LOD4HSS	*Linked Open Data for the Humanities and Social Sciences: Developing ORD Best Practices, Communities and Sustainable Services with Geovistory.*	University of Bern (UniBE)	UniFR, UZH, UniNE, UniBS, HEP-Vaud, USI, LARHRA, KleioLab, DaSCH and JMU	Tobias Hodel	660’160	291’080
AstroORDAS	*Online Open Research Data Analysis Services for Astrophysics and Multi-Messenger Astronomy*	University of Geneva (UNIGE)	EPFL and FHNW	Stéphane Paltani	1’309’130	619’177

Short summaries of the projects

Abstracts by the applicants:

ModelArchive

Open Research Data (ORD) best practices for computational macromolecular models

Proteins, DNA, and RNA are essential for all biological processes, and their functions are intertwined with their 3D structure. Traditionally, structures are determined experimentally, mainly with X-ray crystallography, NMR, and cryo-EM techniques, but recently computational methods have made impressive progress in accurate 3D protein structure prediction. In fact, the journal Nature has nominated protein structure prediction as “Method of the Year 2021”. The structural biology community has pioneered open research data principles, as exemplified by the Protein Data Bank (PDB), the global de facto standard archive of experimentally-determined macromolecular structures. However, the PDB does not archive structures determined through computational modelling, resulting in computational models stored in undefined locations, in incompatible formats, and lacking essential metadata. Following recommendations from an international community workshop, we have developed an archive for computed macromolecular structures, ModelArchive (https://modelarchive.org), and an extension of the mmCIF data format to store model metadata. However, data standards and best practices are not yet established for complex computational models involving proteins, DNA, RNA and/or small molecules, different conformational states of the same macromolecule, and synthetic proteins constructed through design methodologies. With the technical infrastructure of ModelArchive now established, we are in a good position to further develop ORD practices in our community. This includes defining and promoting best practices for data and metadata standards, establishing deposition policies with publishers and funding agencies, improving usefulness of protein models through linking accuracy estimates and accompanying metadata, and connecting to other ORD resources to make models easily findable, accessible and reusable.

OPEM

Open EM Data Network

Electron Microscopy (EM) recently experienced several technological breakthroughs that transformed this technique from being a specialized tool, used by relatively few experts, to a core technique desired by many researchers, particularly those in life sciences. Democratizing access to EM, however, is challenging as the microscopes and ancillary machines are incredibly expensive, and the volume of data generated is enormous. To meet these challenges, campuses across Switzerland, both in the swissuniversities domain (including the partners of this grant) and the ETH domain, have made major EM hardware investments including the creation of the Dubochet Center for Imaging, an EM consortium involving Lausanne, Geneva and soon Bern). These massive efforts will hopefully be additionally consolidated and augmented by the Swiss Roadmap for Research Infrastructures program, which has also recognized the strategic importance of unhindered access to EM techniques for Swiss researchers. The purpose of this current grant is to establish for the Swiss EM community open EM data management workflows, which are fully compatible with the FAIR principles as well as internally established data repositories. Managing EM data lifecycles and ensuring compliance with ORD policies is currently primarily the responsibility of the individual researcher and consequently varies immensely between universities and individual research groups. Here, we propose to create a Swiss-wide network wherein facility managers, through improved data acquisition protocols, and scientists, through better research data management, adherent to the FAIR principles and compatible with international open repositories, will foster open research and streamline the generation of open data publications. An analogous proposal has been submitted through the parallel ETH domain call ensuring that the Open EM Data Network - OPEM is being established at all relevant research institutions throughout Switzerland (Please see the appendix for the ETH domain application).

Swiss DBGI-KM

Knowledge Management in the Swiss Digital Botanical Gardens Initiative

The Digital Botanical Gardens Initiative (DBGI) is a recently launched initiative exploring innovative Open Science solutions for collecting, managing, and sharing digital information acquired from living botanical collections. Focus is on large-scale chemodiversity characterization using mass spectrometry. The DBGI will initially leverage collections of the Swiss botanical gardens to establish robust and scalable chemodiversity digitization workflows. The ultimate goal is to apply these approaches globally in wild ecosystems. The knowledge gathered will then be used to orient biodiversity conservation projects. Knowledge management is a central component of the DBGI. For this, the proposed approach consists of the systematic collection of data at the biological and chemical levels and associated relevant metadata. Physical and digital objects will be linked through a dedicated sample tracking system. Semantic web technologies will then be used to combine chemical, spectral and phylogenetic data in a tailored knowledge graph. This graph will benefit from connection to relevant ontologies (e.g., CheBI, Plant Trait Ontology) and knowledge bases such as Wikidata. Following the Open Notebook Science guidelines, we will share research outputs early in the project. With the *"Knowledge Management in the Swiss Digital Botanical Gardens Initiative"* proposal, we are setting up a team of complementary expertises (UniFr, UniNe, ETHZ-SIS and SIB) that aim at harmonizing and automating the many steps of the data life cycle in the DBGI. This initiative will be a place to establish and share with the community optimized workflows for the digitization of the chemical information of large biodiverse ensembles.

OBELIS

Open Elite Data Project

The « Swiss Elite Observatory » (OBELIS) has developed an internationally unique social science database, which has collected extensive information on over 35’000 members of political, economic, cultural and academic elites from 1890 to 2020. It includes data on social origins, education, career, military rankings, positions in committees and associations, prizes and family relationships of Swiss elites. It is connected to other national and international data platforms (Dodis, Swiss historical dictionary, metagrid.ch) and serves as the basis of a large, international and diverse number of research endeavours. In average, the Swiss Elites database is visited by 1300 individual users per week and is also part of a large “citizen science” project. The objective of this proposal is to: 1) adopt the existing database and to render it fully compatible with the FAIR and open data principles. 2) enlarge and strengthen the scientific communities studying elites with an active and outreaching community management and the organization of regular ORD training offers (summer school, course modules, etc.) 3) roll out to and merge its data architecture with a wider project, the World Elite Database (WED), the leading international database on elites and the super-rich. Through this project we will set the standards for ORD data management and ORD data use in large prosopographical social science databases. It will allow a large international research community to understand the political, economic and environmental challenges by studying elites, their decisions and their influence and become a critical resource for many journalistic initiatives and non-governmental organizations.

CRISPR4ALL

Lentiviral arrays for the indexed activation, deletion and silencing of all human genes

Phenotypic CRISPR screens are instrumental many fundamental biological phenomena, and arrayed CRISPR libraries extend the screening territory to cell-nonautonomous, biochemical and morphological phenotypes. We have recently completed the generation of two human genome-wide arrayed plasmid libraries termed “T.spiezzo” (gene ablation, 19,936 plasmids) and “T.gonfio” (gene activation and epigenetic silencing, 22,442 plasmids). These tools allow for the individual manipulation of every human gene, are broadly applicable to all human biology, and are attracting strong interest from biological and medical researchers world-wide. However, while these plasmid collections are suitable to studying immortalized transfectable cell lines, their most transformative utilization involves primary human cells and human induced pluripotent stem cells, which require that plasmid libraries be repackaged into lentiviral vectors. Here we propose to enable the access of the global research community to these toolsets and to the research data generated therefrom by 1) generating lentiviruses from our library plasmids and providing them to the research community; 2) by providing practical training, protocols and standards to interested users, and 3) by creating a community of users adopting robust FAIR-compliant computational tools for collaborative research and data sharing. We have described the tools produced thus far in a bioRxiv preprint, and the feedback from the community makes us confident that our library resources, our innovative high-throughput cloning methods and our shared-data protocols will become valuable resources and may even be adopted as standards in a broad range of fields of biology.

UpLORD

Upgrading the linguistic ORD-ecosystem

Since 2018 a consortium of partners has been working on building a national ecosystem of infrastructures, which covers the whole linguistic data lifecycle according to ORD requirements (FAIR principles) from data generating, processing and analyzing to data sharing and archiving. This ecosystem includes the national technology platform LiRI and the national repository for publishing and archiving linguistic data (SWISSUbase) as service providers, a database of Swiss media texts and a platform for hosting of and searching in large text and audio/video corpora. In tight collaboration with two target scientific communities (CLARIN-CH and the NCCR “Evolving Language”), the services upgraded through this ORD proposal will significatively facilitate access to and reuse of linguistic research data as they will be provided to the national and European scientific community. In addition, through the large array of partners represented by the CLARIN-CH consortium, this ORD project will help the establishment of numerous collaborations at different levels: (i) within the target national communities, (ii) between the two service providers and the target scientific communities, (iii) with the applicants of two other ORD proposals. The project focuses on upgrading workflows and interoperability of existing infrastructure services, establishing working groups on the national level, documenting and promoting best practices, raising awareness and training about ORD practices in the context of teaching, research and publishing, and building a robust practice of data curation. In the long-term, this project will significatively contribute to developing a strong foundation for a sustainable ORD strategy for linguistic data in Switzerland.

scFAIR

scFAIR: Standardization and stewardship of single-cell metadata

Single-cell functional genomics is a novel field of biology, which is bringing major insight into the life sciences. Single-cell data are rapidly increasing both in quantity and in diversity, but lack method and metadata standardization. While some large projects have clear standards of reporting, most datasets in biological databases have partial or non-standardized metadata. This leads to multiple non-compatible standards across datasets, and limits reusability, which in turn presents challenges to make these data useful to an in-creasing community of specialists and non-specialists. Therefore, there is a need for a centralized, standardized repository where researchers can collaboratively upload, annotate, or access single-cell metadata.

There is also a need for standards in the way single-cell data are stored and annotated, especially for cell type and other associated information. Indeed, metadata is critical to the capacity to use these large and potentially very informative datasets. It includes protocols, which constrain which transcripts were accessible or which normalizations are relevant, the association be-tween barcodes and annotations, or the methods used to identify cell types. Existing ontologies and controlled vocabularies are not used systematically, even when information is reported.

The goal of this proposal is to build a collaborative platform supporting and disseminating ORD practices for the single-cell genomics community, both for sharing datasets and their metadata, and for standardizing the way data are shared across datasets. In addition to the applicants' single-cell community network, we also propose to promote communication and outreach, to increase relevance and implementation of these standards.

AFFORD

A Framework for Avoiding the Open Research Data Dump

Requiring open access to research data is a necessary step towards credible, reliable, and reproducible research, but it is not sufficient. Unless tools and infrastructure for proper data cura-tion and structuring are provided, the requirement is bound to produce data dumps where data are freely accessible, but of limited utility. Publishing data in a form that allows for further use is highly resource-intensive, which is the key reason why it is rarely done. We aim to establish a sustainable support framework that lowers the barriers to publishing data and other forms of research output in an accessible form by bundling know-how, workflows, and tools under the umbrella of one organizational entity of the respective Swiss university. We further propose to accompany a complex reference research project through the full cycle of experiment planning, research output creation, analysis, curation, and publishing. We will thereby quantify the resources needed for the support framework, field-test the framework, and optimize it. This project-based, data-driven approach will improve the sustainability of the support framework by providing reliable resource requirement estimates to decision makers at the university level. It will further accelerate acceptance by the research community by ensuring that the support framework has reached a sufficient level of maturity before it is made available to all university researchers.

enviPath

Fostering interdisciplinary sharing and usage of chemical biodegradation data

Understanding and predicting the fate of synthetic chemicals in the environment is crucial to evaluate their potential risk for humans and ecosystems. The persistence of a chemical depends on how fast it is enzymatically transformed by microbes, which in turn is determined by its molecular structure and environmental conditions. Biodegradation data is available from scientific publications and regulatory reports, but a systematic storage of such data in terms of standardized input/output formats is required to make the data amenable to further scientific exploitation to ultimately advance our understanding of persistence and to accelerate the shift towards a greener chemical industry. To this end, the online platform enviPath has been developed. It systematically stores biodegradation pathways, rates, and experimental conditions, and it predicts pathways using expert and machine-learning systems. enviPath is the only provider of comprehensive biodegradation data internationally and is increasingly used as source of information, data sharing and storage space, integral workflow part, and teaching resource. This project will develop and implement tools to expand enviPath's capabilities for reciprocally linking into two key research communities, namely systems biology and analytical chemistry. Concomitantly, standard formats and documentation for data input/output will be further developed to facilitate submission and exploitation of biodegradation data in the fields of environmental chemistry and cheminformatics. User workshops and newly developed training materials will further attract a growing community from different scientific fields to access and use enviPath as a central platform to store, share, and exploit biodegradation data for inter-disciplinary research.

SIMBioData

Standardized integration of multi-omics biomedical data

As technological advances enable the collection of vast datasets of biomedical measurements, many ongoing studies attempt to decipher various aspects of human health from such data. Although the focus has been primarily on genetic information, other data modalities, such as the abundances of RNAs and proteins within cells and tissues, relate more directly to phenotypes. However, these latter modalities raise significantly more data analysis challenges, and so far, the emphasis in large consortia has been almost exclusively on data production, curation and storage. Efforts to standardize analysis methods so as to allow application on a large scale without the need for subjective choices, are virtually non-existent. Moreover, while measures have been put in place to ensure that the data generated in scientific studies satisfies the FAIR principles, FAIRification of methods does not help in addressing issues of data quality, internal consistency, and interpretability of analysis outputs. We propose that to really harness the potential of the wealth of omics data for biomedical research, it is essential to establish a standardized, sustainable and evolvable method infrastructure for extracting biophysically-meaningful quantities and underlying regulatory information. In particular, only by providing standardized methods that extract biophysically-meaningful quantities in a transparent manner, will it become possible to quantitatively compare and integrate results from omics data across different modalities and experimental approaches. In addition, we feel that our project will provide an ideal prototype for the analysis component of the SwissBioData initiative, which is scheduled to start after the completion of our project.

RE2VITAL

REuse and REproducibility of intraVITAL microscopy via open data practices in image-driven immunological research

The immune system involves complex biological processes with a strong dynamic component, such as cell migration and cell-to-cell interaction. Imaging technologies are becoming pivotal to study these processes. Of note, in the last two decades, intravital microscopy (IVM) allowed to unravel unprecedented details on the cellular dynamics of the immune system. IVM generates multidimensional data (3d videos with multiple acquisition channels), which are analyzed by performing cell tracking and by computing measures of cell motility/interaction. Unfortunately, IVM data, metadata, and tracks typically remain stored in private servers with restricted access; only part of it being published as supplementary materials along with manuscripts, mainly for visualization purposes. This hampers data reusability, reproducibility, and further analyses to answer additional questions outside the original goal. Moreover, the application of data mining methods is particularly affected, as large datasets are typically required. To overcome these limitations, thanks to a SNSF funding and an international network of 15 world-class laboratories, we created IMMUNEMAP, an Open Data platform for IVM data dedicated to immunological processes. Currently, IMMUNEMAP counts more than 350 IVM videos and 24’000 single-cell trajectories. Here, we aim at advancing and fully anchoring Open Data practices with IMMUNEMAP by:

a) Consolidating the functionalities of IMMUNEMAP including findability and interoperability;

b) Promoting data exchange between immunology and computer vision communities, thus expanding the network of partners;

c) Educating researchers to adopt open protocols that safeguard reproducibility. These advancements will foster the availability of IVM data under the FAIR principles and allow data-driven immunological studies.

ODTPR-SMS

An Open Digital Twin Platform for Research on the Swiss Mobility System

Digital Twins (DTs) are the newest buzzword for Information and Communication Technology (ICT) and follow the Internet of Things (IoT) as a Gartner’s hyped technology (Tao et al., 2018). While there is no consensus on the meaning of DTs (Grübel et al., 2022), they consist of geo-referenced data to represent relevant elements of a system for analysis and decision-making (Grieves & Vickers, 2017). However, the model simplicity – a Physical Twin and a Digital Twin (Grübel et al., 2022) – that subsumes implementation complexity is alluring across industry, governance, and research. New DTs are constantly developed (Grübel et al., 2022). Nonetheless, each DT is a unique ungeneralised artefact and there is still no open format for designing DTs and exchanging data between DTs (Roest, 2019). In the mobility context, DTs would describe individual or household spatial behavior with a temporal resolution anywhere from annual to real-time data. There is a wide array of ORD practices such as NADIM and Renku that could be combined into DTs. In this project, we close this gap by joining existing open standards to coordinate them as a DT open standard. For instance, together with the Swiss Data Science Center (SDSC), we will integrate the Renku platform into ORD practises in mobility research. The DT format allows for practices of easier communication of ideas, results, and methods as the data is represented from collection over analysis to presentation. Thus, we realize the potential of ORD practices in mobility research and beyond through the DT approach.

ORDEA

Open Research Data Environments for the Arts

Despite the rapidly growing number of visual artefacts being made available from archives, collections, and museums (GLAM) and the strong support for digital research components fostered by national and international funding agencies, the research community still faces fundamental challenges in leveraging the full potential for digitised archival resources. The lack of accessibility, of joint ontological foundations and of easy-to-use toolkits prevents the efficient reusability and interoperability of holding and research data. This proposal introduces a framework of four Action Fields for [A] shared ontological frameworks for research data, [B] tooling and practices for referencing and interlinking data, [C] tooling and practices for research and holding data integration, and [D] tooling and practices for interlinking visual with semantic data and practices for semantic data analysis. These are achieved by integrating acknowledged yet extendable international standards for data and data exchange and ontologies. Furthermore, the project combines the in-volved partners' toolkits - both those already existing and those currently being developed - into a comprehensive and domain-specific framework for producing, analysing, visualising, and reusing visual resources from the GLAM sector for further scholarly research. The proposal seeks to advance ORD practices in art history and related disciplines in accordance with the Swiss National Open Research Data Strategy, as well as to enhance computer vision methods and knowledge graph analytics for cultural heritage resources.

Swiss-AL

Swiss-AL: Linguistic ORD Practices for Applied Sciences

The ZHAW School of Applied Linguistics develops Swiss-AL (Applied Linguistics), an extensive collection of text data in all Swiss national languages, a linguistic processing pipeline, and a browser-based analysis workbench, enabling researchers to explore copyright-protected data on public language use. Swiss-AL is part of the national linguistic ecosystem CLARIN-CH and has become an essential component of the national ecosystem of language resources and linguistic infrastructures. It is operated by the ZHAW Digital Discourse Lab (ZHAW, 2022b) where it is applied in inter- and transdisciplinary research projects on analyzing how language is used in publicly relevant discourses. Swiss-AL already applies FAIR principles by making data accessible and findable on a dedicated workbench, by processing language data (e.g., websites, social media) in an interoperable, standardized fashion and by fostering reusability for different purposes. To take these practices to the next level, a scientific panel of different academic disciplines will evaluate the special disciplinary requirements for linguistic ORD. The overarching goal is to integrate good practices for language-related ORD matching disciplinary research routines manifested by the target scientific communities of Swiss-AL: applied sciences in general, applied linguistics in particular and the CLARIN-CH and the European CLARIN communities. The project innovatively promotes competences in the use of linguistic ORD and explores the value of such data outside linguistic disciplines. This will be reached through implementing standards of sustainable documentation, legal issues/privacy, data circulation, and research data management in accordance with FAIR principles.

LOD4HSS

Linked Open Data for the Humanities and Social Sciences: Developing ORD Best Practices, Communities and Sustainable Services with Geovistory.

While academic research in the Humanities and Social Sciences (HSS) collects high-quality information in order to gain more in-depth knowledge about past and present societies, there exist limited ORD practices allowing to reuse the significant amount of data produced. New virtual research environments (VRE) are being adopted with different potential to make data reusable, ranging from silos with specific data models to collaboratively enriched, open, high-quality knowledge graphs. The latter approach was adopted to create Geovistory, an easy-to-use VRE and Linked Open Data publication platform designed to enable and foster ORD practices in HSS. Geovistory is already used as a production and publication platform by different SNF and ANR projects, as well as in doctoral research and academic teaching. To develop its full potential, it is essential to broaden the community of users and supporting institutions. This project aims to validate and propagate ORD workflows using Geovistory according to different use-cases and in line with the needs of different domains in the HSS (WP2), to strengthen the share of best practices in a community of data producers and consumers (WP3) and define a clear governance and financial model closely interlinked with academia (WP4). As a result, ORD blueprints using Geovistory will be available for Swiss and international HSS researchers, as well as an engaged community and documentation of best practices. This project thus contributes to methodological innovation of ORD practices and makes the Geovistory VRE sustainable from a community perspective as ORD infrastructure meeting the specificities of HSS research.

AstroORDAS

Online Open Research Data Analysis Services for Astrophysics and Multi-Messenger Astronomy

We propose to develop an ecosystem of cloud-based services and technologies to provide added value to data from science data centers for astronomy, astroparticle and cosmology projects in which we are involved, with the aim to implement the Findable Accessible Interoperable Reusable principles for Open Research Data (ORD) collection, elaboration, and dissemination, building upon our Multi-Messenger Online Data Analysis (MMODA) platform. Our collaboration hosts data centers for several European Space Agency's missions, in particular INTEGRAL, Gaia and Euclid, and is participating in the data centers of the Cherenkov Telescope Array and the Square Kilometer Array. We will extend MMODA to new data analysis types and to integrate MMODA with an emerging world-wide network of online astronomical ORD analysis services (ORDAS) that will be accessible from scientific publications, including refereed journal articles and short messaging services supporting ORD. Integration of ORDAS with publications will ensure full reproducibility of published results and their traceability to raw observational data and will allow the development of in-built intelligence of MMODA services. This will ultimately foster information exchange for studies of astrophysical sources. Our development will drive astrophysical data centers into a new era of cloudnative openly reusable and reproducible scientific operations. The experience we will gain will be useful to other science domains with diverse and vast ORD. This will be demonstrated by integrating our publication-linked ORDAS technology within the multidisciplinary EuroScienceGateway project of the European Open Science Cloud.