YARETA

Yareta
Research Data
Preservation Policy

UNIVERSITÉ DE GENÈVE

CONTENT

INTRODUCTION

The Yareta Preservation Policy (the Policy) serves as an overarching general policy framework for Yareta, a generic repository service for managing, preserving and sharing research data suitable to all researchers and Higher Education Institutions (HEI) in the Canton of Geneva, Switzerland.

The Policy is based upon the following guiding documents in the field of research data management and digital preservation:

- The Open Archival Information System (OAIS) reference model (ISO 14721) developed by the Consultative Committee for Space Data Systems (CCSDS): https://public.ccsds.org/pubs/650x0m2.pdf
- The FAIR principles defined by a consortium of scientists and organizations: https://www.go-fair.org/fair-principles
- The model for a digital preservation policy framework prepared by the Inter-university Consortium for Political and Social Research (ICPSR): https://www.icpsr.umich.edu/files/ICPSR/curation/preservation/policies/dp-policy-outline.pdf
- The Research Data Management Policy Template prepared by the Data Life-Cycle Management (DLCM) project, mandated by swissuniversities: https://www.dlcm.ch/resources/dlcm-policy

The purpose of this Policy is to ensure that research data in Yareta is actively managed for reproducibility and reuse by others, well into the future, in line with data management and digital preservation standards and best practices.

The organization of the Policy reflects the seven attributes of a trusted digital repository:

1. OAIS compliance
2. Administrative responsibility
3. Organizational viability
4. Financial sustainability
5. Technological and procedural suitability
6. Systems security
7. Procedural accountability

Version	1.2
Prepared by:	Lydie Echernier, Research Advisor and Service Manager for Yareta, e-Research Team, UNIGE David Gold, Developer and Service Manager for Yareta, e-Research Team, UNIGE Mathieu Vonlanthen, Development Engineer for Yareta, e-Research Team, UNIGE
Validated by:	Pierre-Yves Burgi, Head of the Scientific Information and Research Department (RISE), IT Division, UNIGE, and Representative for Yareta Hugues Cazeaux, Head of the e-Research Team, UNIGE, and Solution Architect for Yareta
Creation date:	24 October 2022
Last revision date:	31 October 2022
Publication date:	31 October 2022

1 OAIS COMPLIANCE

Yareta and the underlying Data Life Cycle Management (DLCM) technology were designed in compliance with the OAIS reference model (ISO 14721) and various open and international standards to facilitate their integration with other information systems and to avoid technology lock-in. Its architecture supports the acquisition, validation, ingestion, storage, management, preservation, and dissemination of all types of research data following current best practices of preservation.

Besides the OAIS reference model used to ensure the long term preservation of data, Yareta was designed in respect of the FAIR principles, a set of guiding principles that govern research data management best practices to make data Findable, Accessible, Interoperable and Reusable. It also implements open technologies and standards such as:

- The Library of Congress' BagIt hierarchical file packaging format, Metadata Encoding & Transmission Standard (METS) and Preservation Metadata Maintenance Activity (PREMIS) to produce high-quality, standards-compliant Archival Information Packages.
- The Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH) to facilitate the exchange of metadata.
- The DataCite Metadata Schema and the Digital Object Identifier (DOI) to ensure the correct and unique identification of resources.
- The File Information Tool Set (FITS) from Harvard, the PUID scheme from the UK National Archives and the Recommended Formats Statement of the Library of Congress, to encourage the use of open standards while keeping the solution open to the broadest range of file formats in use in the different scientific disciplines.
- The LOCKSS (Lots of Copies Keep Stuff Safe) program from Stanford University to collectively ensure data preservation, integrity checking and repair on a distributed and decentralised network.
- The Open Researcher and Contributor ID (ORCID), Research Organization Registry ID (ROR ID) and Software Package Data Exchange (SPDX-ID) to enhance data visibility and attribution.

2 ADMINISTRATIVE RESPONSIBILITIES

2.1 PURPOSE

Yareta aims to enable the FAIRness of all research data in the Canton of Geneva by utilizing and combining open technologies, standards and principles aligned with European policies and international developments, to create sustainable, transparent and auditable data services that facilitate the management, preservation and sharing of research data following the OAIS reference model (ISO 14721).

2.2 MANDATE

Yareta is a generic research data repository offered to all researchers and Higher Education Institutions (HEI) located in the Canton of Geneva as part of the University of Geneva (UNIGE)'s mission to make research data as widely available as possible, particularly when public funds have been received for the realization of the research project (§14 of UNIGE Institutional Policy on Management of Research Data).

Yareta is also the institutional data repository for the UNIGE. As such, this document is accompanied by:

- UNIGE Guidelines on Integrity in Scientific Research
  Aim at guaranteeing integrity in scientific research, promoting high quality research, ensuring a coherent framework for integrity, sensitizing researchers to the risks stemming from conflicts of interest, and defining a procedure for alleged integrity violations.
- UNIGE Institutional Policy on Management of Research Data
  Affirms the importance of sound management of research data and reminds the stakeholders of their respective duties.
- UNIGE Charter for Open Science
  Marks the commitment of the institution and the academic community to the sharing of scientific knowledge, and affirms the University's adherence to the principles of open science, which aim to ensure free and easy access to scientific publications, research data and the methodologies used to generate this data.
- Guidelines on application at the University of the law on public information, access to documents, and protection of personal data (LIPAD)
  Aim at informing members of the University community of the law on public information, access to documents, and protection of personal data (LIPAD), and to promote its consistent application.

2.3 OBJECTIVES

Yareta supports high-quality research data management, preservation and sharing across all Geneva's HEI and provides them with foundations for data stewardship – collecting, documenting, sharing and publishing – by:

- Developing and maintaining a high-quality infrastructure to facilitate a **responsible management, preservation and publication of research data** in compliance with legal provisions, publishers and funders requirements and the FAIR principles (Findable, Accessible, Interoperable and Reusable).
- Guaranteeing secure access to data while **promoting good practices and international standards**, for an inclusive approach to data management, preservation and sharing.
- Collaborating with researchers to develop practical solutions for the preservation and publication of large or sensitive data to **ensure the reproducibility of scientific arguments and results** over the long term, respecting the different working practices and processes within HEI, faculties, departments and research groups.
- **Highlighting the value of research data as independent research results** by making visible the contributions to research projects of UNIGE and Canton of Geneva researchers.

3 ORGANIZATIONAL VIABILITY

3.1 SCOPE

Yareta is a generic data repository suitable to all researchers and institutions in the Canton of Geneva.

This Policy applies to:

- All research data material collected and held for the purpose of long-term preservation and sharing by Yareta.
- Preservation metadata and persistent identifiers (DOIs) associated with data archives and collections.
- Distributed storage network, replicated across two UNIGE institutional nodes.

What Yareta does (following the OAIS reference model):

- Portal: Provides a user interface to transfer data and metadata from researchers.
- Pre-Ingest module: Performs file identification and assessments, including virus scanning, file format identification, data compliance level assessment and file filtering.
- Ingest module: Prepares data for archiving using best practice standards and technologies.
- Archival Storage module: Manages processing and storage transfer of archives for long-term preservation and ensures secure data storage.
- Data Management module: Makes the archives findable and registers DOIs to data registries.
- Access module: Makes the archives accessible for reuse, consistent with access rights set for the data by researchers.
- Preservation Planning module: Manages archive and collection retention periods and policies.
- Administration module: Monitors the technical environment to plan and implements any measures necessary to mitigate risks.

What Yareta doesn't do:

- Provide collaborative working spaces for active data management (short-term storage).
- Accept submissions that consist solely of secondary data (publication, data paper, documentation).
- Perform file format migration or normalization, data derivation and data transformation.
- Ensure data compliance with ethical, legal, copyright and Intellectual Property Rights issues.

It is emphasized that researchers remain responsible for the content of the archives and collections they create on Yareta. Researchers may be asked to make necessary changes and/or additions to their data or metadata to comply with this Policy. In case of non-compliance, the UNIGE may take certain measures and draw consequences at the discretion of the UNIGE. The UNIGE Guidelines on Integrity in Scientific Research and the Guidelines on application at the University of the Law on public information, access to documents, and protection of personal data (LIPAD) state the restrictive measures (sanctions) in force.

3.2 OPERATING PRINCIPLES

Yareta is committed to developing and implementing procedures that are appropriate, proportionate, evidence-based, practical, cost-effective and sustainable, and in the best interest of enhancing its mission. To do so, it operates in accordance with the following principles:

- Compliance with internationally recognized standards and practices, such as OAIS and FAIR principles.
- Technology alignment and integration for enhanced compliance with the requirements from statutory, academic, funding and publishing bodies.

- Development and maintenance of a reliable architecture and services to ensure access to archives and collections.
- Adoption of technical and metadata standards for interoperability that enable data and metadata exchange.
- Promotion of quality standards for data and metadata to increase the value of data and facilitate data discovery, access and reuse.
- Observation of the research community, in order to follow the evolution of the requirements for functionalities that meet their needs for data management, preservation and sharing.

3.3 ROLES AND RESPONSIBILITIES

Yareta is operated within the UNIGE by the Scientific Information and Research Department (RISe) of the IT Division, headed by Pierre-Yves Burgi – acting as Representative for Yareta – and developed and maintained by the e-Research team, headed by Hugues Cazeaux – acting as Solution Architect for Yareta. The e-Research team currently includes 6 developers and 2 service managers, and operates in close collaboration with the UNIGE Library, the University's Open Research Data Committee – which includes representatives from different University services listed below – and researchers in the UNIGE Faculties and other HEI of the Geneva Canton.

The following parties play an essential role in fulfilling Yareta's mission to protect research data assets in the Canton of Geneva.

UNIGE IT Division is expected to:

- Coordinate the creation and development of high-quality services on data preservation and sharing.
- Provide a quality infrastructure to facilitate good data management, preservation and publication, including:
  1. Standard services to assist with the management of large and sensitive data.
  2. Custom infrastructure and services where necessary.
- Provide secure access management to data according to international standards and IT security guidelines.

UNIGE Library is expected to:

- Coordinate a network of Data Specialists embedded within the faculties.
- Work with faculties to provide support and training in data management, preservation and publication.

Other University services including the:

Rectorate,
Research and Grants Office,
Legal Affairs Department,
Committee for Ethical Research in Geneva (CUREG),
Data Science Competences Center (CCSD),
Technology Transfer Office (UNITEC),

Are expected to:

- Provide expert input on policy and practical issues related to data protection and ownership.
- Ensure that an understanding of good research data practices is recognized as part of the academic culture.
- Devise strategies to deal with the economic aspects of long-term data archiving.

Principal Investigators or those in equivalent roles in the UNIGE Faculties and other HEI of the Geneva Canton are expected to:

- Request the creation of an "Organizational Units", representing their research team, project, department, laboratory, etc., that matches the specific context of their research environment.

- Be the primary contact for all financial/billing issues and budget the costs of research data preservation at the application stage of the Organizational Unit.
- Ensure that research data management requirements imposed by statutory, academic, funding and publishing bodies are met, including the appropriate handling of personal and confidential data as required by the Law on Public Information, Access to Documents, and Protection of Personal Data (LIPAD), the General Data Protection Regulation (GDPR), and the Federal Act on Research involving Human Beings (HRA), as applicable – consult the Cantonal Commission for Ethics and Research (CCER) or the University of Geneva's Committee for Ethical Research (CUREG 2.0) for more information.
- Understand who owns the data resulting from their research and the implications for data management, preservation and publication, and ensure that contractual agreements with third parties on data confidentiality are always respected.
- Monitor access requests and notifications, and ensure adequate management and control over their archives, including quality and consistency of research data and associated metadata.

Researchers in the UNIGE Faculties and other HEI of the Geneva Canton are expected to:

- Follow the research data management strategy and practices of their HEI, faculties, departments and research groups, and attend data management training when needed.
- Ensure that research data is appropriately documented in accordance with the FAIR principles (Findable, Accessible, Interoperable and Reusable).
- Ensure that metadata related to data is relevant and useful even when the data is not public, with a statement explaining why access is restricted or closed, who can use the data and under what circumstances.

3.4 SELECTION AND ACQUISITION

Yareta is a generalist, subject-agnostic repository designed for the preservation of any kind of file formats and open to researchers from all HEI in the Canton of Geneva. It accepts original data generated and recorded by researchers as well as pre-existing data collected from other sources. Researchers need to evaluate and select the data to keep for the long term according to policies or legal requirements they are subject to, and to choose within a set of submission, preservation and dissemination policies accordingly. Yareta ensures that all data is accompanied by sufficient administrative, descriptive, structural and technical metadata by performing automated file assessments and control checks to ensure coherence with policies. All data that fail the automated quality assessment and/or checks are returned to the researcher for further review, resulting in either improvement (e.g. improved metadata) and resubmission, or disposal. While all file formats are accepted, Yareta includes a star system informing researchers of the level of compliance of each data file with recommended file formats for long-term preservation, making them aware of the risks associated with obsolescence. At the end of the data retention period, a disposal process assists researchers in determining the fate of their dataset, resulting in either an extension of the retention period or the disposal of the data, while preserving its identifier (DOI) and associated metadata as a "tombstone".

The following lists detail the minimum set of requirements, customizable features, curation features and support available to researchers during the selection and acquisition process in Yareta.

MINIMAL REQUIREMENTS

Yareta strives to facilitate the work of researchers with minimal requirements:

- Researchers request the creation of an Organizational Unit for themselves or their team, or request to be assigned to one or more existing Organizational Units to manage deposits, archives, metadata, contributors, policies, validation and access.
- Researchers provide a minimum set of metadata required by the DataCite open descriptive metadata schema to assign a persistent identifier to the archive.

- A Digital Object Identifier (DOI) is automatically assigned to a data archive and can be pre-reserved by researchers prior to the processing and storage of the archives for long-term preservation.
- Researchers set a level of access to their archives, either public, restricted or closed, with an optional embargo period.
- For archives in public access, researchers choose a license, with a default choice of Creative Commons licenses (e.g. CC0 or CC-BY, depending on the configuration).

CUSTOMIZABLE FEATURES

Customizable features allow researchers to continue their work as usual with the workflow that suits them:

- Upon creation of their Organizational Unit, researchers set the policies governing:
  - The submission of data, with an optional validation workflow.
  - The retention of archives, with a guaranteed preservation timeline ranging from 5 years to forever, or customizable to the researchers' needs.
  - The default license, including the Creative Commons suite of CC0, CC-BY, CC-BY-SA, CC-BY-NA, and individual, customizable licenses.
  - The dissemination of archives customized on demand (to the IIIF protocol, a standard to disseminate and share digital objects).
- Optional additional rich metadata allowing for better findability and reuse are available, including:
  - Additional descriptive metadata with complementary fields from the pool of DataCite metadata.
  - The ability to associate a custom metadata schema in XML and JSON, with automated validation of the file type (format) and content (structure) against the schema.
  - Research-domain tags based on Swiss and international registries: research areas and disciplines from the Swiss National Science Foundation (SNSF), Re3data subjects, or any other registry upon request.
  - Keyword tags to categorize and filter data.
  - Archive types (based on the DataCite resourceType attribute) allowing for the preservation of different types of digital objects.
  - Custom graphic elements, such as logos, avatars and thumbnails, to improve the portal's user-friendliness and the readability of the content.
- Well-documented RESTful APIs allow Yareta to be integrated with other applications so that it fits directly into research workflows – here are some integration examples:
  - Laboratory instruments (Laboratory Information Systems and Electronic Laboratory Notebooks).
  - Visualization tools, analytics, or any other application to automatically extract and synchronize data (e.g. Jupyter Notebooks).
  - External datacenters to keep a copy of the data to comply with a local policy.

CURATION FEATURES

Curation features assist researchers with quality assurance and control.

Automated curation features:

- A pre-ingest module capable of extracting, listing and testing compressed data from a ZIP file, while retaining the folder hierarchy and data structure.
- An automated evaluation of the file by the system, including:
  - Data File format identification (FITS).
  - Review of file formats and names to automatically sort files that should be excluded from an archive (e.g. system files) or that require manual validation (e.g. file names containing "old").
  - Data compliance level assessment using the PUID scheme from the UK National Archives.
  - Library of Congress Format Statement recommendation, to encourage the use of open standards while keeping the solution open to all file types.
  - Control involving a range of different functionalities (e.g. consistency of levels of access and levels of sensitivity).

- Fine-grained filtering to allow researchers to validate, ignore or exclude files based on a specific status.

Human curation:

- Researchers can choose to include an optional validation workflow with enhanced curation performed by a Data Steward (i.e., a domain expert, archivist, librarian, etc.) enabling manual review with the option to approve, reject or send datasets back to edition.
- Researchers benefit from a system-integrated DataTag model for sharing and using sensitive data in a standardized and responsible way.
- Researchers can edit an archive's metadata and link to related research outputs:
    - Indicating that an archive is referenced by another archive or by a publication (DOI referenced by)
    - Indicating that an archive is identical to another one (DOI identical to)
    - Adding the DOI of another archive that replaces the current one (DOI obsoleted by)
- Researchers can create collections interlinking archives with qualified referencing in DataCite metadata.
- Researchers are provided with dedicated support to help them plan the preservation of their data, taking into account the cost of long-term data archiving and the requirements from the Swiss (SNSF) and European (Horizon) research funding programs.

SUPPORT

Support is available to help researchers get familiarized with the Yareta portal and its many features

- Extended documentation with a "Getting started" guide, a full "User guide", FAQs, a series of video clips, tooltips and guided tours – offered in the form of tutorials – is available on a self-service basis.
- Institutional resources are allocated to data management support in each of the HEI of the Geneva Canton, with training opportunities both onsite and online.
- Custom support is available for the submission and management of large and/or sensitive datasets.

Yareta also offers additional features, such as multi-language support and visualizations of documents, images, audio, video, etc., including chemical molecules in 3D.

3.5 ACCESS AND USE

Yareta focusses on keeping the data archives discoverable and accessible to researchers and re-users. This includes presenting descriptive metadata through a custom search interface and through public APIs, and tracking changes in technology, standards and requirements of the research community to continue to meet its needs. Yareta's users include, but are not limited to, researchers from all HEI in the Canton of Geneva and their partners in Switzerland and abroad, UNIGE faculties and students, practitioners, policymakers, journalists, and anyone with a desire to access research data archives. ORCID, ROR ID and SPDX-ID support allow for researcher name, institution and license disambiguation and gives more visibility to their data by connecting their publications across different platforms.

Yareta maintains public, persistent and reliable access to the descriptive metadata of the archive, independent of the archive's content, accompanied by metrics to track the use and impact of data, as well as a data citation feature to ensure proper attribution, improve reproducibility, discoverability of the data, and contribute to its recognition as a research output. Yareta also devotes significant resources to developing and implementing features ensuring confidentiality and protection of personal or sensitive data. Access to archive content is governed by a role-based access control and archive access level, with public archives available for download without login wall, while restricted or closed archives remain available on demand after login, as appropriate.

Yareta uses current technology and standards to provide a range of access services:

- Yareta offers discoverability through indexation and harvesting of DataCite descriptive metadata by the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

- Open and well documented RESTful APIs allow for integration with existing IT infrastructures.
- Traditional search techniques are augmented with a customizable faceted navigation system.
- Metric display completes each archive with its number of files, number of views, number of downloads and the possibility to rate the quality and usefulness of each archive.
- Authentication and role-based access control allow researchers to grant appropriate permissions to their team members based on their workflows, and beyond their own team to facilitate sharing and collaboration within their partners in the international community.
- Access Control Lists (ACL) assign permissions to individual researchers or to groups, and a notification system enables researchers to process access requests on a case-by-case basis (to ask, grant or deny access).
- Data citations can be generated automatically following common standards formats (APA 6th edition, MLA 7th edition and Chicago 16th edition).

Yareta enables users to download data associated with an archive through various methods:

- Direct download of all data associated with a public archive: the data is available for download immediately and files are accessed from a link displayed in the browser.
- Group order: several archives' data are added to a 'shopping cart' feature and are made available for download as an order after the Yareta has finished processing the request.
- Request for access for data in an archive under a restricted or closed access level: data is available through direct download or grouped order if and when access is granted by the archive's owner.

3.6 CHALLENGES AND RISKS

Yareta faces the following challenges related to digital preservation:

TECHNOLOGICAL EVOLUTION

The multiplication of file formats and the institutional choice to not perform file format migration or normalization implies that existing data files preserved in Yareta face the risk of obsolescence. Yareta must therefore adapt to a constantly evolving technology to guarantee the continuity of access to these files. It must also maintain scalable systems and workflows (i.e. staffing, technology, infrastructure, practices) to handle the rapid and evolving growth in data volume.

AWARENESS AND TRAINING

The majority of UNIGE and HEI staff in Geneva do not have digital preservation and data management as an explicit part of their mandate. Yareta is committed to raising awareness about digital preservation and data management issues and to providing appropriate training to the UNIGE staff and to the broader community of researchers, information managers, archivists, and data users. It also strives to increase collaboration between UNIGE divisions and IT departments in order to make more efficient use of available staff expertise and resources and to promote consistent system administration.

MANAGEMENT AND RESOURCES

UNIGE is expanding the types of digital content collected to include collections and assets from galleries, libraries, archives, and museums (GLAM), which means that existing technologies, policies, procedures and practices will need to be revised. Yareta also requires sustainable and ongoing financial support, including the allocation of technological, administrative, and personnel resources to meet all the development needs of current and future users.

4 FINANCIAL SUSTAINABILITY

4.1 INSTITUTIONAL COMMITMENT

Yareta is hosted at the University of Geneva (UNIGE), founded in 1559. UNIGE is a public institution endowed with a moral status, placed under the supervision of the Geneva State Council, which exercises this supervision through the Department of Public Education. UNIGE organizes itself, sets its priorities and methods of action and is responsible for its management within the framework of the Loi sur l'université du 13 juin 2008 (effective March 17, 2009), which also regulates its funding.

Preservation of data in Yareta is a joint responsibility between Yareta and the depositing researchers and institutions. If the UNIGE terminates Yareta for any reason, the UNIGE will return the archives and preservation metadata to the depositing researchers and institutions on mutually agreed terms. At the time of transfer, the UNIGE will ensure the transfer media and the dissemination format will be relevant and compatible with current best practices and standards.

4.2 COOPERATION AND COLLABORATION

The solution was developed as part of two larger projects aiming to develop various services for data management for researchers and HEI in Switzerland:

PL 12146

The Geneva cantonal bill "Digital Infrastructure for Research" (PL 12146) 2018-2024:

The State of Geneva, UNIGE and the University of Applied Sciences and Arts of Western Switzerland (HES-SO) in Geneva have joined forces to create a framework convention on digital technology. This agreement aims to facilitate and encourage collaboration in the fields of teaching, continuing education, applied research and services related to digital transformation. This collaboration allows for joint activities such as the provision of internships, the organization of digital events, collaborations with students on these topics, and collaboration on cantonal, national and international research projects. The State of Geneva has granted UNIGE a subsidy of more than CHF 15 million, running from 2018 to 2022 and extended to 2024, to finance the development of research infrastructure and the implementation of associated digital services, particularly in the fields of high-performance computing and research data management.

This project is aimed at all Geneva HEI and intends to:

1. Strengthen and develop the existing digital infrastructure.
2. Develop new digital solutions and services for research.
3. Use state-of-the-art technologies in line with national and international research policies.
4. Ensure the quality of information, documentation and support associated with the services.
5. Listen, support and advise researchers on digital issues.

To strengthen the teams, several staff members have been hired for the duration of the project. They amount to 6 staff members in the e-Research team responsible for the Yareta development and service management at UNIGE. The 3 remaining team members hold permanent contracts to ensure the continuity of services.

DLCM

The national Data Life Cycle Management project (dlcm.ch) 2015-2021:

The DLCM technology powering Yareta is the result of a coordinated effort by UNIGE together with nine partners – eight Swiss HEI, namely EPFL, ETHZ, HEG, HES-SO, UNIBAS, UNIL and UNIZH from 2015, and ZHAW from 2018, and one service

provider, SWITCH – aiming at proposing national-level solutions covering the whole Data Life-Cycle Management (DLCM) tailored for both researchers and institutions, to help them safely manage, publish and preserve their data over the very long term (20 years and over) while dealing with most tools and formats across all scientific disciplines.

In 2014, as a foundation to the DLCM project, 49 interviews were conducted in 30 different disciplines of 6 HEI to identify researchers' primary needs and existing solutions in place. During the DLCM 2015-2018 period, the project identified the basic components that contributed to the development of a generic data repository suitable to all researchers and institutions across Switzerland. With one additional partner involved, the project then conducted a number of user-centred pilots during the pursuant phase in 2019 and 2020 to refine the basic set of services with additional features. In 2021, the project resulted in the provision of FAIR- and OAIS-compliant data repositories integrated to partner solutions, and of training and consulting services from a national network of experts.

Yareta, the Geneva instance of the DLCM solution, was launched on June 26, 2019. Its name comes from a cushion plant in the family Apiaceae native to South America, which is known to be one of the oldest living organisms on earth – over 3'000 years old according to some estimations. Yareta is registered with Re3data, BASE (Bielefeld Academic Search Engine), OpenAIRE and OpenDOAR (Directory of Open Access Repositories), and is eligible for funding for data upload to researchers under a SNSF grant.

OLOS.swiss, the national instance of the DLCM solution, offers similar services to all Swiss researchers and HEI outside of the Canton of Geneva.

RDM SUPPORT NETWORK

The national Research Data Management support network (RDM support):

Yareta team members take an active part in a national support network born from the DLCM project and other local or national projects funded by national research funding agencies and the State Secretariat for Education, Research and Innvation (SERI). It gathers stakeholders from research data management support teams within universities, universities of applied sciences, universities of teacher education, and research institutes. The network meets twice a year on issues related to:

1. Support with Data Management Planning (DMP).
2. Research data management training.
3. Consultations on legal and ethical issues related to research data management.
4. Research data management infrastructures, services and tools.

5 TECHNOLOGICAL AND PROCEDURAL SUITABILITY

In considering what technological and procedural preservation actions are appropriate, Yareta intends to retain data in its original file format within archives, and to enable their dissemination.

Preservation actions are guided by the following principles:

- Maintain existing technologies and evaluate evolving technologies for future developments.
- Monitor and adapt to new technologies in line with the evolution of best practices for digital preservation.
- Normalize technology and workflows across applications with similar preservation requirements.
- Follow current data management best practices and advise researchers and HEI accordingly.

Yareta is committed to preserving published archives in accordance with this Policy, the procedures described on the UNIGE e-Research website, and the Yareta specifications which lists the standards, protocols, registries, technologies and open-source applications used by Yareta features. Yareta provides access to research data through the maintenance of archives, descriptive metadata records and associated identifiers, and provides citations and access metrics, which include page-views, downloads, and ratings. Yareta is committed to this level of preservation, access, and reporting for a period of at least 5 years for all archives, and for at least 20 years for archives set to be preserved "forever", from their date of publication on Yareta. At the end of the retention period, Yareta informs researchers to give them a choice between:

- Extension of the retention period: Yareta will retain the data associated with their published archives beyond the initial commitment for a period of at least 5 years.
- Disposal: Yareta will securely destroy the data associated with their published archives while retaining its related descriptive metadata record and identifier.

In the event of no response from researchers, Yareta will endeavor to retain published archives beyond the initial commitment but reserves the right to assess the long-term viability of the archive to determine whether to retain the archive beyond the initial commitment. Yareta may also withdraw data from the archive files using the disposal process for routine maintenance reasons (e.g., if an archive is discovered to be a duplicate, or a high-resolution version replacing a lower-quality file) or as a result of a reassessment of an archive's value in light of research-defined selection criteria, while retaining the corresponding descriptive metadata record and identifier and maintaining a reference to the new archive that replaces the old one (DOI obsoleted by).

Yareta's preservation, access and reporting actions are supported by dedicated security features, which also enable visibility, reuse and reproducibility of research results:

- Open standards for data preservation for reliable storage and transfer (PREMIS, METS and BagIt) to prevent lock-in and support exit strategies.
- Security measures such as checksums, redundancy and geographic separation of copies (by default, data remains in Geneva, but archive nodes can be configured at any other storage provider upon request).
- Audit trails capture and record all user and system activities made to files and transactions – logged activities include actions on Organizational Units (changelog with names and timestamps), archives (creation, automated checks, validation, approval, submission, technical validation, metadata edition, completion, cleaning) and individual files (creation, processing, virus scan, file format identification).
- Connectors to multiple archival storage systems such as file system, S3 protocol and tapes.
- Containerization (Docker, Kubernetes) for security improvement, costs reduction, and cloud portability.
- Technical measures supporting a "Privacy by Default and by Design" compliant with the Swiss and European legislations, such as the Law on Public Information, Access to Documents, and Protection of Personal Data (LIPAD) and the General Data Protection Regulation (GDPR).

6 SYSTEM SECURITY

Assurances for data integrity and the continuity of access to data archived in Yareta are guaranteed by the following items:

1. Technical infrastructure
2. Data management
3. Backup policies
4. Storage facilities and security
5. Disaster prevention and recovery

TECHNICAL INFRASTRUCTURE

Yareta is powered by the DLCM technology, which is open-source and runs on open-source technologies. The DLCM technology is released under a GNU General Public License v2.0 or later license. Development and maintenance actions performed by the e-Research team of UNIGE use the agile Scrum methodology.

The DLCM technology was developed with the Solidify framework, developed by UNIGE, extended from Spring Boot.

- Archival system: DLCM Solution (Java, Spring Boot, MariaDB, Elasticsearch, ActiveMQ, ClamAV, FITS)
- Frontend: Yareta Portal (Angular)
- Operating system: Debian Linux

Yareta is running three independent instances:

- The production system.
- A development quality assurance system, ensuring that changes can be tested and validated prior to being applied to the training quality assurance system.
- A training quality assurance system, enabling researchers and staff members responsible for training to test and validate new developments prior to being deployed to the production system.

DATA MANAGEMENT

Data management in Yareta resides on asynchronous replication, involving sharing data between two redundant resources to ensure consistency, improve reliability and fault-tolerance. Each data file is associated with:

- A documented fixity check.
- A virus check at the time of initial ingestion into Yareta.
- A documented object verification procedure.

For each data file, we store independent MD5, SHA1 and SHA256 checksums.

- The asynchronous replication is performed daily.
- The fixity check is performed weekly.
- A replication check is performed monthly.
- The object verification is performed annually.

Metadata is indexed in an Elasticsearch cluster for fast and powerful searching. All changes to metadata records on Yareta are versioned.

BACKUP POLICIES

Additionally, each file copy has two back-up replicas located on different disk servers. Back-ups are created and maintained locally and off-site according to the following guidelines:

1. New and edited data and metadata in Yareta are immediately and redundantly logged to disk.
2. Daily back-ups are maintained for two months and stored locally.

3. Weekly back-ups are maintained for two years and stored locally and off-site.
4. Yearly back-ups to non-disk media (tapes) are maintained for 10 years and stored off-site.

STORAGE FACILITIES AND SECURITY

Physically, Yareta technical infrastructure is located on UNIGE premises:

- Locally at Uni Dufour, Rue Général-Dufour 24, 1204 Geneva, Switzerland.
- Off-site at Campus Biotech, Chemin des Mines 9, 1202 Geneva, Switzerland.

Storage media include HDD, NVME, Tapes, managed by Hitachi HCP and IBM Spectrum Protect.

Yareta follows policies established by the UNIGE IT Division for maintaining security of the technical environment in compliance with applicable laws and regulations governing data security. These policies are set forth in the UNIGE Politique de sécurité du système d'information (PSSI). UNIGE employs staff members solely dedicated to network and system security. Only authorized staff members involved with IT maintenance have access to the archiving system. Datacenters are climate-controlled and protected by a fire protection system, and access privileges to datacenters are limited by a staff badge access system.

Yareta also uses the following security procedures:

- Authentication: SWITCHaai and Shibboleth
- Authorization: OAuth2 using Spring Authorization Server
- Antivirus: ClamAV
- VPN: FortiGate
- Security scanner: Nessus
- Code analysis: SonarQube

DISASTER PREVENTION AND RECOVERY

All systems and services delivered by UNIGE, including Yareta, are subject to risk and vulnerability analysis on a yearly basis. Dedicated policy documents address specific aspects of the long-term protection of each of UNIGE's digital assets.

Yareta, as the UNIGE's digital preservation infrastructure, is considered mission-critical and has appropriate levels of monitoring, response and recovery. A dedicated Disaster Recovery Plan for Yareta covers the technical and administrative procedures to ensure consistent and systematic data security, including system security requirements, operational requirements and regular audits and reviews.

7 PROCEDURAL ACCOUNTABILITY

7.1 AUDIT AND TRANSPARENCY

Preservation actions are reviewed periodically, in 12-month cycles. The duration of the review cycle depends on changes in the technological environment, on risk assessment, on available financial and human resources and on the needs of researchers and HEI.

Yareta uses two maturity models for digital preservation:

The National Digital Stewardship Alliance's (NDSA) levels of digital preservation (LoP), which focus on 5 aspects of preservation:

1. Storage and Geographic Location
2. File Fixity and Data Integrity
3. Information Security
4. Metadata
5. File Formats

The Digital Preservation Coalition (DPC) Rapid Assessment Model (RAM), a tool designed to enable rapid benchmarking of digital preservation capabilities, which focuses on 6 organizational capabilities and 5 service capabilities:

1. Organizational viability
2. Policy and strategy
3. Legal basis
4. IT capability
5. Continuous Improvement
6. Community
7. Acquisition, Transfer and Ingest
8. Bitstream Preservation
9. Content Preservation
10. Metadata Management
11. Discovery and Access

7.2 POLICY FRAMEWORK ADMINISTRATION

The Yareta Research Data Preservation Policy was completed in October 2022, approved by the UNIGE IT Division's Head of the Scientific Information and Research Department (RISe) and Representative for Yareta, and by the Head of e-Research Team and Solution Architect for Yareta, on October 31, 2022. The recommended changes will be brought to the attention of the staff members responsible for Yareta within the UNIGE IT Division for validation and to the UNIGE Rectorate for information. Each new version of the Policy will carry a version number and a date stamp.

7.3 DEFINITIONS

ORGANIZATIONAL UNIT

An Organizational Unit is a logical entity that can represent a structural unit, such as your laboratory, institute or department, or your research group or project. It allows for the management of team members and their permissions to deposit and archive research data.

DEPOSIT

A deposit is where users upload their datasets and associate metadata to it. It is fully editable until it is submitted

ARCHIVE

Once submitted, a deposit becomes an archive. The archive is published on Yareta and is linked with a DOI. Only its metadata can be edited. Its data files are immutable.

COLLECTION

Yareta allows users to combine several archives into a collection which has its own DOI

ARCHIVE TOMBSTONE

The tombstone is what is left of an archive after it has been disposed of. Data files associated with the published archive are securely destroyed. The archive's descriptive metadata record and DOI remain preserved

ORGANIZATIONAL UNIT ROLES

Users can have different roles within an organizational until.
These roles are as follows:

- Manager
- Steward
- Approver
- Creator
- Visitor

Information about these roles can be found in the Yareta Integration Guide.

ORDER

Yareta allows users to add several archives to an "order" which will download all selected archives at once when submitted.

PRIMARY DATA

Data collected directly as opposed to data collected from previously done research. Primary data on Yareta are categorized as follows:

- Observational
- Experimental
- Reference
- Simulation
- Derived
- Digitized data

SECONDARY DATA

Data collected from previously done research. Secondary data on Yareta are categorized as follows:

- Publication
- Data paper
- Documentation

7.4 REFERENCES

The following links are displayed in order of appearance.

- OAIS reference model (ISO 14721): https://public.ccsds.org/pubs/650x0m2.pdf
- FAIR principles: https://www.go-fair.org/fair-principles
- Inter-university Consortium for Political and Social Research (ICPSR) digital preservation policy framework: https://www.icpsr.umich.edu/files/ICPSR/curation/preservation/policies/dp-policy-outline.pdf
- Research Data Management Policy Template prepared by the DLCM project: https://www.dlcm.ch/resources/dlcm-policy
- The Library of Congress' BagIt hierarchical file packaging format: https://www.rfc-editor.org/rfc/rfc8493
- Metadata Encoding & Transmission Standard (METS): http://www.loc.gov/standards/mets/
- Open Archive Initiative Protocol for Metadata Harvesting (OAI-PMH): https://www.openarchives.org/pmh/
- DataCite Metadata Schema: https://schema.datacite.org/
- Digital Object Identifier (DOI): https://www.iso.org/standard/43506.html
- File Information Tool Set (FITS): https://projects.iq.harvard.edu/fits/home
- PUID scheme from the UK National Archive: https://www.nationalarchives.gov.uk/aboutapps/pronom/puid.htm
- Recommended Formats Statement of the Library of Congress: https://www.loc.gov/preservation/resources/rfs/
- LOCKSS (Lots of Copies Keep Stuff Safe): https://www.lockss.org/
- Open Researcher and Contributor ID (ORCID): https://orcid.org/
- Research Organization Registry (ID ROR ID): https://ror.org/
- Software Package Data Exchange (SPDX-ID): https://spdx.dev/
- UNIGE Institutional Policy on Management of Research Data: https://memento.unige.ch/doc/0320
- UNIGE Guidelines on Integrity in Scientific Research: https://memento.unige.ch/doc/0003
- UNIGE Charter for Open Science: https://www.unige.ch/openscience/en/open-science-2
- Guidelines on application at the University of the law on public information, access to documents, and protection of personal data (LIPAD): https://memento.unige.ch/doc/0160
- Law on Public Information, Access to Documents, and Protection of Personal Data (LIPAD): https://silgeneve.ch/legis/index.aspx
- General Data Protection Regulation (GDPR): https://gdpr-info.eu/
- Federal Act on Research involving Human Beings (HRA): https://www.fedlex.admin.ch/eli/cc/2013/617/en
- Cantonal Commission for Ethics and Research (CCER): https://www.ge.ch/organisation/commission-cantonale-ethique-recherche
- UNIGE Committee for Ethical Research (CUREG 2.0): https://cureg.unige.ch/
- Yareta in Re3data: https://www.re3data.org/repository/r3d100013423
- Yareta in BASE (Bielefeld Academic Search Engine): https://www.base-search.net/Search/Results?q=dccoll:ftugeneveyareta
- Yareta in OpenAIRE: https://explore.openaire.eu/search/dataprovider?datasourceId=re3data_____::263180f806632244a29e88306d7eed5f
- Yareta in OpenDOAR: https://v2.sherpa.ac.uk/id/repository/10042
- Swiss National Science Foundation (SNSF): https://www.snf.ch/en
- OLOS.swiss: https://access.olos.swiss/portal/
- Swiss national Research Data Management support network (RDM support): https://forschungsdaten.info/fdm-im-deutschsprachigen-raum/schweiz/rdm-support-stakeholders/
- DataTag model: http://datatags.org/datatags-compliant
- UNIGE Information System Security Policy: https://memento.unige.ch/doc/0174
- UNIGE e-Research website: https://www.unige.ch/eresearch/en/services/yareta/
- DLCM gitlab page: https://gitlab.unige.ch/dlcm/community/dlcm-backend
- UNIGE Loi sur l'université du 13 juin 2008: https://silgeneve.ch/legis/index.aspx
- Geneva cantonal bill "Digital Infrastructure for Research" (PL 12146) 2018-2024: https://ge.ch/grandconseil/memorial/seances/010409/48/3/

- Yareta specification: https://www.unige.ch/eresearch/en/services/yareta/specifications/
- Yareta user guide: https://yareta.unige.ch/doc/Yareta-QuickStartGuide.html
- Yareta integration guide: https://admin.yareta.unige.ch/administration/docs/DLCM-IntegrationGuide.html
- Yareta RESTful APIs documentation: https://admin.yareta.unige.ch/administration/docs/DLCM-APIs.html
- Yareta tools guide: https://admin.yareta.unige.ch/administration/docs/DLCM-ToolsGuide.html
- SWITCHaai: https://www.switch.ch/en/aai/

Photographic credit: UNIGE, Jacques Erard, 2018 https://phototheque.unige.ch/unige:121641

Contact:

e-Research
Université de Genève
Division du système et des technologies de l'information et de la communication / IT Division
24 rue du Général-Dufour
1211 Genève 4
eresearch@unige.ch

October 2022