ARCHIVER

ARCHIVING AND PRESERVATION FOR RESEARCH ENVIRONMENTS

Project Abstract: Phase 1

LIBNOVA Consortium (LIBNOVA – CSIC-IFCA – University of Barcelona – Giaretta Associates)

Short description of the company or consortium

The consortium is composed of the following institutions:

LIBNOVA – a company centered around digital preservation technology, with more than 10-years history and customers all over the world. LIBNOVA's portfolio covers all the needs to curate and preserve both cultural heritage (libraries, archives, museums) as well as scientific datasets (universities and research institutions). LIBNOVA Research Labs (2017) manages all research initiatives for the company.

CSIC – IFCA – the Spanish National Research Council (in this project through IFCA – Instituto de Física de Cantabria) is the main agent of the Spanish System for Science, Technology and Innovation. CSIC generates knowledge through scientific and technical research, transfers results from research (to boost and create technology-based enterprises), and manages large facilities and unique scientific and technical infrastructures.

University of Barcelona – is the foremost public institution of higher education in Catalonia, and also the principal centre of university research in Spain. A reference European institution for research activity, both in terms of the number of research programmes and the excellence achieved.

Giaretta Associates – David Giaretta has worked in digital preservation since 1990, and was chairman of the OAIS Reference Model (ISO 14721), the "de facto" standard for building digital archives. He leads the group which produced the ISO standard for audit and certification of trustworthy digital repositories (ISO 16363).

Description of the proposed solution and how it meets the R&D objectives identified in the Contract

The solution we are proposing is built on pre-existing digital preservation platforms already in use by many leading organizations across the world. It proposes a solution for the whole organization and for the whole data life-cycle, completely aligned with OAIS, ISO16363, FAIR and TRUST principles, with powerful and really innovative capabilities in all functionality layers.

The Research, Management and Preservation Platform will combine existing technologies and new components, to solve obstacles for research dataset management (including preservation) identified in the Archiver project.

Five areas comprise the architecture:

- Containers – keep content accessible with several protocols, organized and protected. These containers keep metadata, data and code together to ensure usability (OAIS-aligned).
- Dynamic Insights – help users when dealing with personal information, digital preservation and emissions reduction, with the following components: Data Policies Assistant, GDPR Assistant, Emissions Optimizer, Digital Preservation
- Budget assistant – helps users to plan and follow expenditures
- Content gateway – connects the platform with repositories for discovery solutions, such as Invenio or Dataverse
- Digital Preservation, OAIS and FAIR conformance – as support for the OAIS Information Model and for the Mandatory Responsibilities, and the results will fully support repositories in OAIS conformance. The focus on usability is also critical for the "Interoperability" and "Reusability" required by the FAIR principles.

A brief summary of areas of work follows:

- Not only the type of media to preserve but also the data center, location, etc. and the Quality of Service (QoS).
- A key functionality will be to manage the full Representation Information Network required by the OAIS Information Model. Where possible the import of complete metadata schema automatically will be supported
- The API will support search, retrieve data and metadata, as well as all functional capabilities.
- Connection for data publishing in repositories (i.e. Zenodo - CERN).
- Provenance, reproducibility or processing are also areas of great importance for scientific data

The resulting product used within a repository will be ISO 16363 audited.

R&D for Phase 1

Phase 1 focuses on the creation of the detailed Design of the solution, including the system architecture and technical design of all components that will fulfill the functional requirements of the Buyers' Group. Consequently all R&D ideas and activities planned for the project will find their way into the description of the solution, which will include the following:

- Scalability (sustained high throughput in the 100s of PBs range)
- Digital Preservation Best Practices (OAIS, ISO 16363, PSC-Preservation Storage Criteria, Best Practices recommendations and implementations – OAIS Information Model including Representation Information and Preservation Description Information components, problem detection such as duplicates, hidden encryption - , format migration/evolution, exit strategy)
- Metadata management (import/creation and preservation), following OAIS
- Data integrity management (integrity chain, integrity at rest)
- FAIR principles (F: containers, customized metadata, structured hierarchy, … A: multiprotocol access, public sharing, discovery solutions, … I: Data policies, research data Representation Information, … R: Integrated active integrity control, Representation Information and Provenance Information)
- Cost efficiency (flexibility on deployment, several computation/storage options)

How the objectives will be achieved (technical and organizational measures to carry out the R&D and achieve the stated objectives)

The Consortium members have organized themselves in the following manner to achieve the planned R&D objectives for Phase 1 (as all this needs to be consolidated into the Solution Design documentation):

LIBNOVA

- Project management
- Solution Architecture (including analysis on how to capture and encode Representation Information and Provenance)
- CEPH-related research

Giaretta Associates

- OAIS Alignment
- ISO 16363 (and other standards) alignment
- ISO 16363 self-certification following the complete ISO 16363/ISO 16919 standards.

CSIC (Spanish National Research Council)

- Research for design and implementation options

- Infrastructure (networking, processing, local storage) architecture, including CEPH deployment and optimization
- HA/Disaster recovery architecture
- Research lifecycle consultancy

University of Barcelona

- Identify regulatory requirements for markets/use cases/needs (in addition to GDPR and other already identified EU requirements, including anonymization of datasets)
- Legal compliance needs, and IPR management
- FAIR Principles alignment

Any steps that will be taken to prepare for commercialization of the solution after the PCP

Although it is too early in Phase 1 to plan so far ahead, some ideas could be put forward. The market for scientific data is gaining more and more volume and interest every year, and can be clearly separated into private institutions and those with some type of public backing. There are also notable geographical differences in the approach, whether it refers to American institutions on one side and European institutions on the other side.

We envision a dual approach (from which customers may choose), now common in the industry:

- Solution as-a-service in the cloud
- Solution on-premises “as a license”

In either case may opt for a one-time license charge (with yearly renewals and maintenance) or a subscription-based model, from the start.

Consulting services will be available for necessary customizations.

Exact details will be produced in Phase 2 and finalized in Phase 3.