Early Adopter: European Institute of Oncology

Early Adopter's Name: European Institute of Oncology (IEO)

Organisation type: *IRCCS: Scientific Institute for Research, Hospitalization and Healthcare.*

Organisation size: *large enterprise.*

Organisation Research Field(s): Medical and Health Sciences

Organisation Profile:

The IEO is a comprehensive cancer centre dedicated to adult oncology, which integrates prevention, diagnosis, treatment and research with a multidisciplinary approach. At IEO, a complete integration exists between clinical and research activities in order to translate scientific results into therapy, as quickly as possible. Basic and translational research takes place at the Department of Experimental Oncology (DEO), in a which is also home to the European School of Medicine Molecular (SEMM), Center for Genomic Science of the Italian Institute of Technology and FIRC Institute of Molecular Oncology. IEO's DEO is composed of about 300 scientists, whose research activities are aimed at discovering the molecular mechanisms involved in the development of cancer. Specifically, research conducted at DEO is founded on five principles: Independent research, Strong interaction with clinicians, Cutting-edge technology (including Technological Units with state-of-the-art equipment and expertise, and Clinical Technoshots for the dissemination of technologies that may favor specific translational-research projects), Open, collaborative and participatory research environment and Intense education activity.

Organisation website URL: https://www.ieo.it/, https://www.research.ieo.it

Suggested Use case title: Archival and accessibility of omics data.

Problem definition:

Our institute is dealing with an increasing amount of omics data generated in our laboratories or by external collaborators. It includes for instance genomics, epigenomics, metabolomics, proteomics, imaging, clinical data.

The data generated can be used more than once:

- The data is first analyzed for a given project. After that first analyses, the data can usually be archived.
- Later, the data can be reused for meta-analyzes, for bioinformatics validation, etc. In such event, it is often reused together with other datasets, by different groups or researchers.

In this context, our institute is confronted to two problems: computational resources and control of data access.

1. Resources

ARCHIVER - Archiving and Preservation for Research Environments project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824516.

The amount of data generated is growing exponentially. We are therefore looking toward solutions to increase our storage (as well as computation) capacity, either on premise, in the cloud, or with a hybrid solution. Many data can be archived on a cold environment, some of it may never be accessed again, while other data may need to be retrieved either for legal or scientific reasons (re-analyses, meta-analyzes or data integration). The time for which the data need to be archived depends on the same scientific and legal requirements (e.g. patient data, founding agencies).

2. Data access

We need a configurable user/group access control to the data. Although we promote the usage of the data by all researchers, the access should be approved by a legal and/or scientific committee and it should be possible to monitor the data usage. A typical workflow would be:

- Create “projects” to which researchers are associated
- Each data generated is associated to a project A, and therefore accessible to the associated members,
- An external researcher asks for accessing the data for a project B. Once the request is accepted, all researchers associated to project B have access to the data.
- At any time, the organization knows who is accessing each data, and to which project it is associated.

The solution adopted would also facilitate the sharing of data with external collaborators, who should be able to access it through the same workflow.

Finally, it should be possible to bill the single groups or units for the usage of the resources.

Is this use case new for your organisation?

Up to now, our organization relied on an on-premise infrastructure (HPC cluster with associated on premise storage). A new solution should be identified and implemented within the next 12 months.

Data and metadata Characteristics:

During the last 10 years, almost 1 PB of data has been generated. With the adoption of new technologies, we estimate a production of ~250 TB/year starting in 2020.

Our organization generates omics data, both from cellular lines and patients. In particular:

- **Genomics** data (RNA-seq, ChIP-seq, WGS, WES, single cell sequencing, single molecule sequencing): previous technologies (NGS) generates 5-50 GB per sample. Novel technologies (single molecules), generates several 300 GB per sample. As of today, we have accumulated ~100TB of primary sequencing data (20TB/year in the last year). This should increase to 120TB/year. In the context of technological projects, the researchers may which to access large groups of data (up to 100 TB of archived data),
- **Imaging data:** ~50TB of imaging data is generated each year (~20GB/experiment). This number may increase with novel technologies under development
- **Proteomics**: ~150GB per experiment, ~5TB/year.

The numbers provided are rough estimates. Other datatypes could be integrated, for instance radiomics data. For each dataset, additional space (up to the same size), is occupied by processed data.

Cost requirements:
The estimated cost requirements will be defined taking into account the cost of the extension of an on-premise solution, and third-party resources offered by national and international organizations.

ARCHIVER - Archiving and Preservation for Research Environments project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 824516.

The solution should provide cost effective long-term storage of the data, still allowing the access of large datasets at reasonable price.

Benefits and expected impact:

The ARCHIVER solution would have the following impact:

1) Providing a cost-effective solution for storing omics data, as an alternative or in complement to our on-premise infrastructure
2) Facilitate the management of data access.

Such solution would not only benefit to the researchers in our organization: many collaborations are bringing together researchers from different institutes and hospitals, who need to access and process the same data.

Contact person & details:

Arnaud Ceol
https://www.linkedin.com/in/arnaudceol/
@arnaudceolpro

ARCHIVER - Archiving and Preservation for Research Environments project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 824516.