Early Adopter NAME - Use case template

Early Adopter's Name: CRG / CNAG (Centre for Genomic Regulation / Centro Nacional de Análisis Genómico)

Organisation type:
Research institution

Organisation size:
Please refer to the classification from Eurostat: https://ec.europa.eu/eurostat/web/structural-business-statistics/structural-business-statistics/sme

Large organisation

Organisation Research Field(s):
See *"Fields of R&D classification" as indicated in p.59 of the Frascati Manual 2015: Guidelines for Collecting and Reporting Data on Research and Experimental Development, Published by the OECD* https://warwick.ac.uk/services/ris/systems/dataquality/frascati/frascati_manual_2015.pdf

- Social Sciences :
- Natural Sciences : 1.2, 1.6
- Engineering and Technology :2.6, 2.9, 2.10
- Medical and Health Sciences : 3.3
- Agricultural Sciences: 4.1, 4.2
- Humanities

Organisation Profile:
Please give us a short profile of your organisation and your stakeholder community

The Centre for Genomic Regulation (CRG) is an international biomedical research institute of excellence, created in December 2000. It is a non-profit foundation funded by the Catalan Government through the Department of Business & Knowledge and the Department of Health, the Spanish Ministry of Science & Innovation, the "la Caixa" Banking Foundation, and includes the participation of Pompeu Fabra University.

The mission of the CRG is to discover and advance knowledge for the benefit of society, public health and economic prosperity.

The CRG believes that the medicine of the future depends on the groundbreaking science of today. This requires an interdisciplinary scientific team focused on understanding the complexity of life from the genome to the cell to a whole organism and its interaction with the environment, offering an integrated view of genetic diseases.

The CNAG-CRG is a non-profit organization funded by the Spanish Ministry of Economics Affairs & Digital Transformation and the Catalan Government through the Economy and Knowledge Department and the Health Department. Competitive grants and contractual research with the private sector provide additional funds. From the 1st July 2015, the CNAG was integrated into the CRG.

The CNAG-CRG was created in 2009 with the mission to carry out projects in DNA sequencing and analysis in collaboration with researchers from Catalonia, Spain and from the international research community in order to ensure the competitiveness of our country in the strategic area of genomics. It started operations in March 2010 with twelve last-generation sequencing systems, which has enabled the center to build a sequencing capacity of over 1000 Gbases/day, the equivalent of completely sequencing ten human genomes every 24 hours. This capacity positions the CNAG-CRG as one of the largest European centers in terms of sequencing capacity. The Center has a staff of highly qualified individuals, 50% of which hold PhD degrees. The bioinformatics team together with our outstanding computing infrastructure (9 petabytes of data storage and over 3000 cores of computing) also positions the CNAG-CRG as a center of excellence in data analysis.

The CNAG-CRG takes part in genome sequencing and analysis projects in areas as diverse as cancer genetics, rare disorders, host-pathogen interactions, the preservation of endangered species, evolutionary studies and the improvement of species of agricultural interest, in collaboration with scientists from universities, hospitals, research centers and companies in the sector of biotechnology and pharma.

Organisation website URL: https://www.crg.eu https://www.cnag.crg.eu

Suggested Use case title: Archiving Genomic and Imaging Data

Problem definition:

The CRG and CNAG generate large amounts of biological data using state of the art instrumentation including next generation sequencers, high resolution microscopes and mass spectrometers. The data is used for cutting edge research in the following fields

- Bioinformatics and Genomics
- Cell and Developmental Biology
- Gene Regulation, Stem Cells and Cancer
- Systems Biology
- Rare Diseases

As well as generating data for the institutes' researchers, the following core facilities provide services to external stakeholders and collaborators:

- Advanced Light Microscopy Unit
- Biomolecular Screening and & Protein Technologies Unit
- Proteomics Unit
- Tissue Engineering Unit
- Flow Cytometry Analysis and Cell Sorting Unit
- Bioinformatics Unit
- Genomics Unit

Various types of sequencing services are offered :

- DNA Sequencing
- Long Read Sequencing
- Transcriptome Sequencing
- Epigenetic Sequencing
- Single Cell Sequencing
- 3D Genomics

Currently we have over 14 petabytes of storage in various systems at the institute. Data must be

stored securely for long periods of time so that re-analysis can be done as new methods and techniques are developed. As well as pure research, the instruments at the institute are being applied in the clinical domain necessitating the encryption of data and compliance with various regulatory requirements such as the GDPR.

Since we provide services and collaborate with a large community external to the institute, we also require means to share data securely with them and allow performant access.

Is this use case new for your organisation?
If so, what is the envisaged timeline for implementation?
If not, how is it currently implemented?
Suggested length: 300 words

Currently, our data is stored in various disk and tape based systems. We maintain replicas in 2 data-centres in Barcelona separated by a distance of 7km. Keeping up with the data generation rate, and the multitude of different storage systems presents a challenge for the IT team. We are looking at ways to consolidate and rationalise these systems to simplify management and to improve cost-efficiency.

As well as needing to solve the problem of providing safe, secure and cost-effective storage we are actively looking at ways to provide FAIR access to data and to allow secure and performant access to external collaborators. We need a system that can cater to all of the following elements:

Safe storage : guarding against silent data corruption, multiple replicas across multiple locations, snapshotting, checksumming,
Secure storage : flexible access controls, federated identity management, encryption, regulatory compliance
FAIR storage - linking data with rich metadata allowing complex querying
Accessible storage : allowing efficient bulk transfer of data utilising fast protocols such as gridftp.
Cost-effective storage : Low cost per terabyte per month and minimal additional costs (e.g. transfer costs, api usage costs). It would be preferable to have everything amortized into a single figure rather than having to deal with multiple line items to estimate costs.

Based on these very complex problems and the amount of work needed to transition from current systems to a new system we would estimate a timeline of around 2 years for development and implementation.

Data and metadata Characteristics:
Please describe the type of data subject to your use case, including the size range
Suggested length: 200 words

The greatest proportion of the data is in the form of standard NGS formats such as fastq and bam files. We are increasingly dealing with large amounts of fast5 files generated by nanopore sequencing. There is also a large amount of imaging data in the form of tif, jpg, nd2 and raw files.

In total we have over 1.5 billion files ranging in size from tens of kilobytes up to a few hundred gigabytes. Large amounts of small files provide difficult technical challenges.

We are currently growing at a rate of around 2 petabytes a year but this is likely to increase substantially over the coming years.

There are various LIMS systems and database systems in use for storing metadata associated with the data generation pipelines.

Cost requirements:
Please explain what your organisation's requirements in terms of costs for the solution developed. Suggested length: 300 words
We have carried out detailed analysis of the cost for storage which has been audited so that we can accurately do both internal and external billing for cost-recovery. Since we currently host our own data, we do not incur any transfer costs or api costs.

Benefits and expected impact:
How would the community that you are addressing benefit from the ARCHIVER solutions? Please describe which would be the foreseen impact of this activity
Suggested length: 600 words

The main tangible benefits would be

- Consolidation and rationalisation of storage systems
- Reduction in costs of storage on a euro-per-terabyte-per-month basis
- Integration of data and metadata according to FAIR principles
- Providing a means of secure and performant access to external stakeholders

In terms of intangible benefits to our researchers and our wider community, we would expect the storage to become less of a worry allowing them to concentrate on the science instead of having to deal with the problems of 'storage juggling'. Having greater visibility of what is stored would also allow more collaboration and possibly create new links between groups that would not have been possible previously. Finally having access to a secure and regulatory compliant storage solution would allow us to increase the cooperation we have with local hospitals and to be able to take on more clinically significant projects.

Contact person & details:
Emyr James
https://www.linkedin.com/in/emyr-james-0071192/

Disclaimer: this form aims to gather further information on your use case for promotional purposes. It does not guarantee the deployment of your use case on the solutions resulting from the ARCHIVER project. As a reminder, the Buyers Group will approve the Early Adopters use cases that will be deployed to test the pilot solutions.