Minute Madness poster session schedule

Pilar de Teodoro (recording submitted)	Enabling data discovery in big datasets
Petiton Julien (recording submitted)	Using the REGARDS Framework for the renewal of the CNES archive system
Kajal Haria	ERS-1/2 SAR and ENVISAT ASAR CEOS-ARD NRB Product Development Project
Benjamin Branch (no poster submitted)	Workforce data curation training with AI considerations in Federated Data Repositories by netcdf-empowered community informatics at OKD edge site
Poppy Townsend	Scoping Net Zero Research Computing
Di Xian (no poster submitted)	Fengyun meteorological satellite global observation and applications
Molly MacRae	Implementing FAIR principles in the IPCC context to support open science and provide a citable platform to acknowledge the work of authors.
Charlotte Wehn	Getting Out The Data - Fighting The Latency Dragon
David Giaretta	Preserving an endangered society: the case of Maldives
David Giaretta	Developing an OAIS based Interoperability Framework
Aurèle Nicolet	Towards a sustainable data and knowledge preservation
Adithya Thaduri (no poster submitted)	Data standardization and integration for maintenance of Critical Infrastructure
Baptiste Cecconi (no poster submitted)	TFCat (Time-Frequency Catalogue): JSON Implementation and Python library
Zhe Xu (no poster submitted)	Design and Implementation of FENGYUN Meteorological Satellite Archive Data Migration Mission
Liv Toonen	CryoSat Ice Baseline-E: Operational & Reprocessed Data Quality Control
Joey Mukherjee	Converting a Traditional Space Science Operations Center to a Cloud-Based Architecture
David Giaretta	Petabyte scale, OAIS/ISO 16363 conformant archive
Jordi Andilla (no poster submitted)	Data structure and long-term preservation of sequential images from optical microscopy
Bernard Pruin	In-Situ and IoT data inventorisation in support of earth observation data analysis
Fay Done	Adding Value to ESA Heritage and Third-Party Mission Archived Datasets via Reprocessing: ATSR and ALOS
Aurèle Nicolet	Computational appraisal for enhancing data value discovery: recent researches and lessons learned for scientific data governance
Deborah Agarwal	Focusing on Scalable Citations to Improve Data Usability and FAIRness
Marcelo Garcia (remote)	Developing a Prototype of Library Dataset supporting Services

Enabling data discovery in big datasets

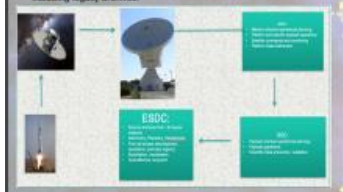
Pilar de Teodoro, AURORA BV for ESA
Sara Nieto, Rhea group for ESA
Monica Fernández, Rhea group for ESA
Hector Pérez, Rhea group for ESA
Christophe Arviset, ESA
ESAC Science Data Centre (ESDC), European Space Astronomy Centre (ESAC), Madrid, Spain

Abstract

The ESAC Science Data Centre (ESDC) is handling the archive data for several astronomy, solar and planetary missions. We started with some gigabytes of information, currently in the hundreds of terabytes and not so far in the future we will handle petabytes. An important fraction of them reside in database systems which allows to analyse the data, structured, semi and unstructured, directly in the ESDC systems using VO protocols. How to store this data in a database to give the users the ability to query easily the contents of a space mission?. How do we choose a solution that will handle some small data to one that scales better for big data?. May this be a nightmare? One does not fit all, but maybe in the future it well may happen. We will review the evolution of database solutions for big data space projects with special focus on the ones that we have already tested (PostgreSQL, CitusDB, PostgresXL, Greenplum) with specific implementation for the Gaia DR3 release, the European JWST archive, the Euclid Science Archive and the future PLATO Data Archive

How ESDC gets the data

The ESDC is the library for ESA missions catalogues. It allows the searching on its data for all scientific missions in all its phases from development to operations phase. It is distributed into helio, human and robotic exploration, observatory, survey and planetary missions including legacy archives.



ESAC archives

The main webpage of the ESDC archives can be found in http://archives.esac.esa.int. From that page you can navigate to any ESAC archive.

Gaia Data downloaded

The Gaia monthly statistics clearly show an increasing trend in the downloads associated to the data releases.

Gaia Archive Users

The amount of users is growing with each release for Gaia. We can analyze accesses to the archives, showing why and what was important to the users. Statistics are measured for all the archives.

System Stack

When the scientific data is ready it will be ingested in our databases(mainly metadata and/or catalogues) and if there is a data repository it will be stored in a petascale data storage. Then, using VO protocols, the data will be able to be retrieved via a web portal, Jupyter notebooks or under the science exploitation platform.

Distributed databases

We have compared parallel databases based on PostgreSQL for typical data base server requirements and specific astronomical use cases:

- Scalability queries
- Aggregation queries
- Geometrical Queries
- Xmatch queries

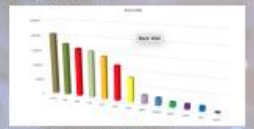
We determined that Greenplum was the best solution to use with IVOA TAP but also Spark may be used for cut-off products.

Gaia Archive New features

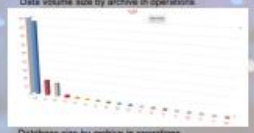
VO protocols are used in the forms to query the archives and also through astroquery. Gaia Archive is also adding new features such as the single object web page.

September 2022 archives size

The data on the archives is stored directly in the database or in filesystem. The first option allows to run queries directly on the data using a database management system, while the second option, will store only the metadata into the database and will locate the data in the filesystem to retrieve it. Examples of the first option is Gaia and Euclid where all the catalogues are stored inside the database.



Data volume size by archive in operations



Database size by archive in operations

Greenplum

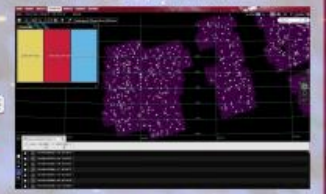
Postgres-XL

CitusDB

Spark

Euclid Archive Visualization

Euclid visualization is based on ESASky and will allow the users to search for specific data, postcards, run cutoffs on the images...



Pilar de Teodoro
pilar.teodoro@sst.esa.int

PV2023@CERN
02-04 May 2023

Using the REGARDS Framework for the renewal of the CNES archive system

Julien Petiton
CNES, 18 av E. Belin, 31401 Toulouse Cedex 9, France

Context

For the long-term preservation of space mission data, CNES has developed the STAF service. The STAF service was introduced in 1995 and has been steadily improved since then. STAF infrastructure is currently in version 3 ("STAF v3").

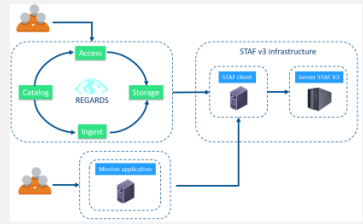
Unlike STAF v3, which uses dedicated infrastructure and software, the STAF redesign ("STAF v4" project) relies on components made available to the Mission Centers or Data Centers, to adapt to their needs in terms of data storage or catalogs. These components are the Datalake object storage infrastructure and the REGARDS access catalog framework (Open Source project available on Github).

STAF v4 project is also an opportunity to improve governance of the CNES archives.

STAF v3 architecture

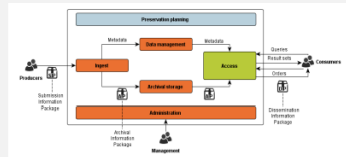
The service needs to evolve for the following reasons:
- To handle a growing archive volume
- End of maintenance for SL8500 libraries
- Production of T10K cassettes SOLARIS stopped
- Software obsolescence (STAR client)
- The STAF is only managed by STAR client in command line (no user interface, no access right managed)

Assessment of 25 years of use:
- 4 petabytes (~57 million files)
- No data loss
- Archive of exclusive data
- Data archived but not referenced in a catalog
- Archive managed by REGARDS and by other applications (resulting unreferenced and non-usable data)



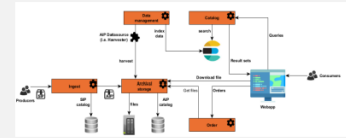
Framework REGARDS

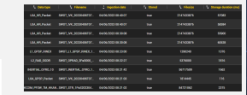
- **Generic** software (Open Source available on Github)
- Development started in 2015
- Using for CNES archives & Mission Center (SWOT)
- Implement the **FAIR** principles
- Implement **OAIS** functionnal model (CCSDS)
- Composed of a back end and a front end
- Easily **adaptable** & **configurable** to various space projects



[Back End]
- Microservices architecture
- Each microservice matches an elementary REGARDS function
- Plugin mechanism to extend functions of microservices and web interface
- Each microservice exposes a REST or AMQP API

[Front End]
- Provides a user and an administration web interface
- Provides enhanced IHM configuration capabilities





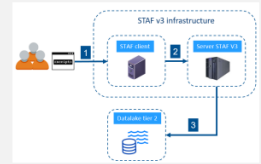
Capture of REGARDS IHM with storage value

STAF v4 architecture

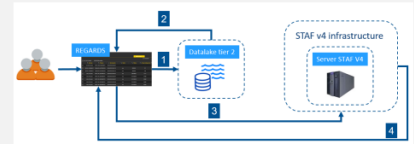
- STAF v4 is a componant of the CNES **datalake** infrastructure (tier 3)
- Capacity of **25 petabytes** (~400 million files)
- Object storage (**S3**)
- Implement user interface & access right managed by REGARDS
- Only accessible through **REGARDS** (interface & REST API)
- Infrastructure evolution transparent for the business
- Build a metadata catalog "storage" similar to that of the datalake infrastructure
- Implement an overview of the archive by REGARDS catalog
- The renewal of the service allows to make an **inventory of the obsolete archive**

Use case migration

- [1] Request to STAR client to extract files
- [2] Star client sends a request to STAF v3 to extract files from the target tape
- [3] Extraction & copy on the datalake tier 2 (disk)

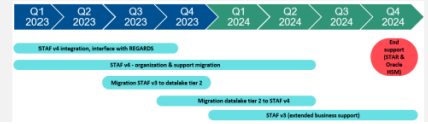


- [1] REGARDS sends request to the tier 2 datalake (disk)
- [2] Data referenced in REGARDS (building metadata)
- [3] REGARDS request STAF v4 to copy files
- [4] File path updated in the catalog (with integrity control)



Planning

- Migration of only valid files after inventory
- Migration is realized by businesses with the support of SERAD



IDEAS-QA4EO

PV 2023 CONFERENCE

esa

ERS-1/2 SAR and ENVISAT ASAR CEOS-ARD NRB Product Development Project

*This project fulfils the needs of ESA's **Heritage Space Programme** to generate a **CEOS-ARD SAR** product aligned to Sentinel-1 and Sentinel-2 ARD outputs, for the historic ERS-1/2 and ENVISAT missions.*

Baseline specification:

CEOS-ARD specification for SAR **N**ormalised **R**adar **B**ackscatter (NRB) PFS **v5.5**

Development approach:

Closely follows Sentinel-1 approach to support:

- **Immediate analysis and facilitation of data use**
- **Interoperability**
- **Cloud computation capability**
- **Open science compliance**

Current status:

- **Builds Upon:** PyroSAR and SNAP
- **DEM:** EEA-10 / GLO-30 or GLO-90
- **RTC:** Flattening Gamma: Radiometric Terrain Correction for SAR Imagery
- **Gridding:** Aligned to MGRS, geometry of each tile read from a Sentinel-2 reference KML file
- **CEOS Requirement Traceability:**
  - 30/30 *Threshold* requirements (**100% compliance**)
  - 7/14 *Target* requirements (**50% compliance**)

TELESPAZIO a LEONARDO and THALES company

Kajal Haria[(1)], Michael Williams[(1)], Garin Smith[(1)], Ricardo Noguera[(1)], Clément Albinet[(2)], Mirko Albani[(2)], Antonio Valentino[(3)], Sabrina Pinori[(4)]

[(1)]*Telespazio UK (TPZ UK), UK;* [(2)] *European Space Agency (ESA)/ESRIN, Italy;* [(3)] *RHEA Group, Italy,* [(4)] *Serco, Italy*

serco

- Workforce data curation training with AI considerations in Federated Data Repositories by netcdf-empowered community informatics at OKD edge site
- Benjamin Branch

Scoping Net Zero Research Computing

Towards a UKRI roadmap to deliver carbon neutral digital research infrastructure by 2040 or sooner

Core project team: Martin Juckes[1], Charlotte Pascoe[1], Ag Stephens[1], Poppy Townsend[1], Jennifer Bulpett[1], Katie Cartmell[1], Miranda MacFarlane[2]
1. National Centre for Atmospheric Science, UK, 2. Kings College London, UK

What we set out to do

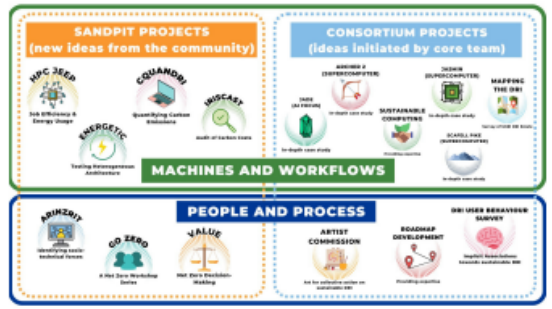
The **UKRI Net Zero Digital Research Infrastructure Scoping project** is a large interdisciplinary project with three key objectives:

- Collect evidence to inform UKRI digital research infrastructure investment decisions.
- Provide recommendations for UKRI and their community with an outline roadmap for achieving carbon neutrality across all digital research infrastructure by 2040 or sooner.
- Enable UKRI to play a positive and leading role in the national and global transition to a sustainable economy.

Gathering evidence

Two broad subject areas:

- **machines and workflows** the hardware and software infrastructure that sits at the centre of the digital research infrastructure
- **people and process** the expert staff, the systems and institutions that frame their work, and the scientific user community delivering the UKRI programme of research and innovation.



These projects have produced over 100 detailed recommendations

Vision for 2040

A vision for UKRI DRI in 2040

- Facilities have a **five-star sustainability** status, with everything from the tea bags in the staff canteens to the racks of servers in the data centres covered by a comprehensive life-cycle analysis.
- **Virtual and augmented realities** transform our interactions with data and with each other, reshaping our notions of space and time and shattering existing barriers to understanding.
- Experts provide a **resource of digital excellence** supporting a transformed national economy.
- The UK DRI reputation for environmental excellence and its leading role in promoting productivity through **Open Science** policies and workflows attracts leading researchers from all over the world.

About

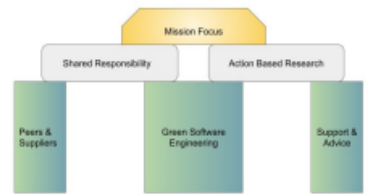
- £1.86 million project funded by United Kingdom Research and Innovation (UKRI), a non-departmental public body sponsored by the UK government Department for Science, Innovation and Technology
- Administered by the Natural Environment Research Council (NERC)
- Based within the Centre for Environmental Data Analysis (CEDA) and the National Centre for Atmospheric Science (NCAS)
- Project partners (~40 researchers) from 20 different institutions
- The 19 month project is due to finish in June 2023

Project scope

Digital research infrastructure (DRI) ranges from high performance computers (HPCs) to university server rooms and everything in between.
The project scope covers UKRI owned and majority funded DRI, and the impacts associated with DRI research outputs and procurement.

Strategic recommendations

Six thematic recommendation areas:



Two key recommended actions:
Firstly a Net Zero DRI Delivery Service to provide support to decision makers, establish and disseminate best practice as it applies to the UKRI DRI, maintain community cohesion through meetings and communication activities, and maintain a map of the UKRI DRI and its carbon footprint.
Secondly a portfolio of projects which allow the community to develop and deploy Net Zero solutions, including the creation of a national resource of green software engineers.

Contact: support@ceda.ac.uk
Website: net-zero-dri.ceda.ac.uk

UK Research and Innovation · Natural Environment Research Council · Centre for Environmental Data Analysis · National Centre for Atmospheric Science · King's College London

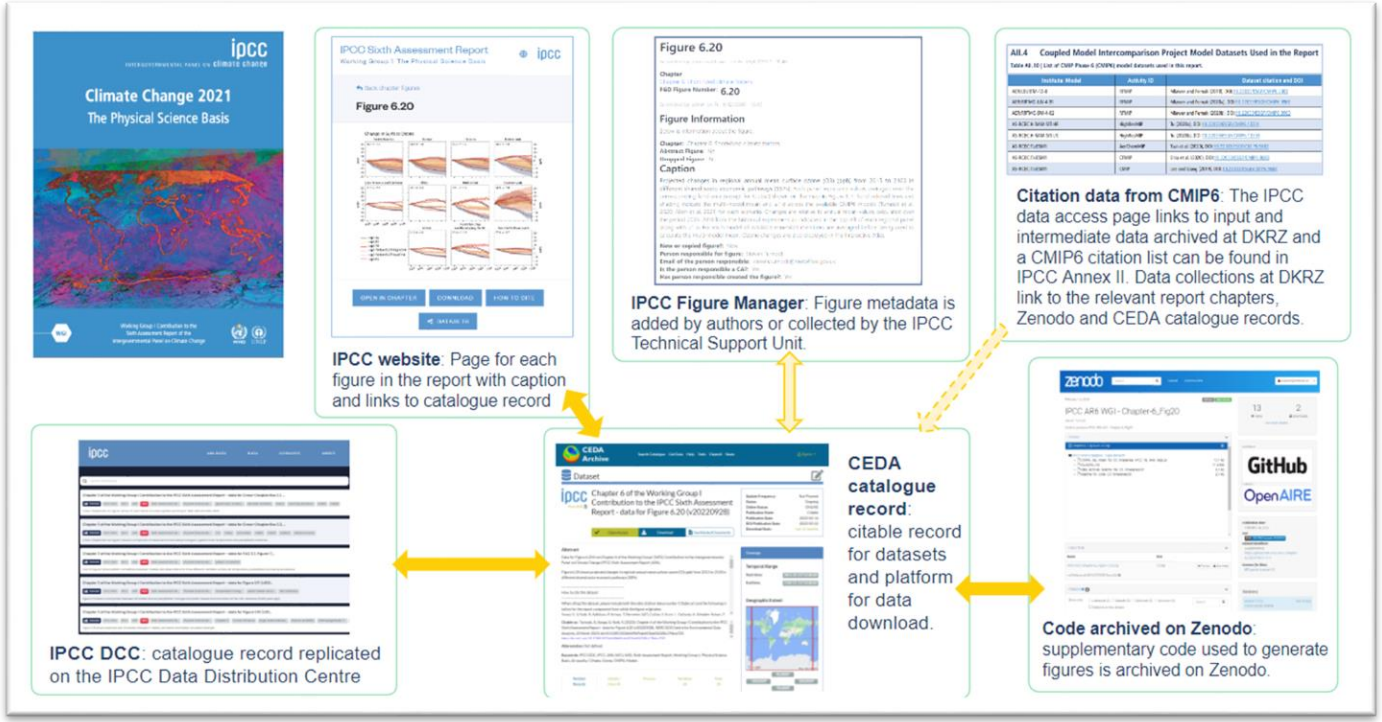
- Fengyun meteorological satellite global observation and applications
- Di Xian

Making IPCC Data FAIR

The reality of implementing FAIR principles in the IPCC context to support open science and provide a citable platform to acknowledge the work of authors.

Molly MacRae, Emily Anderson, Diego Cammarano, Charlotte Pascoe, Anna Pirani, Lina Sitz, Martina Stockhause

Come and find out about our: Data Publication Ecosystem, Workflow of Data, FAIR implementation, Decision Making Trees and Recommendations for AR7



Find out how the catalogue record is FAIR!

Getting Out The Data - Fighting The Latency Dragon

Charlotte Wehn

Deutsches Zentrum für Luft- und Raumfahrt (DLR), Earth Observation Center, Oberpfaffenhofen
charlotte.wehn@dlr.de

The D-SDA Long Term Archive

D-SDA is the German Satellite Data Archive, a long term archive for earth observation data established in the 1990s.

- Operated by DLR (German Aerospace Center) Earth Observation Center
- More than 40 national, European and international missions
- More than 40 PB of data

D-SDA Archival overview

Long term storage on tape

- Cost: tape is cheaper than disk
- Energy efficiency: idle tape does not use energy
- Reliability: tape is designed for longer life than disk
- Security: it's harder to delete or encrypt data on tape

Putting the Data to Work

The D-SDA is an active archive, with data being used for scientific and third party projects and demand growing with new ideas.

Time series, e.g. for

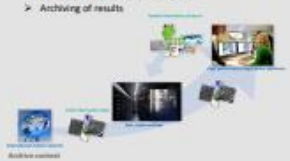
- Urban growth
- Land use
- Ozone hole changes
- Climate change

New exploitation platforms for data access and processing

- Processing services on platforms instead of downloads ("user to the data")
- Large amounts of data online, usually more recent data
- Processing optimized data formats (analysis ready data)
- Often very high processing power

Platforms interact with archives:

- Initial load from archive to platform
- Gap filling
- Reload of historical or evicted data
- Archiving of results



Archive content

Higher processing power and algorithmic maturity enable new ways of working with earth observation data

- **More data is needed ...**
- **... faster**

The Problem

Latency!

Tape speed increases, but latency is hard to reduce and makes overall access slow.

Startup time for tape reads

- Tape mounts
- Tape positioning → can take a very long time

Implications:

- Serial reads of many files, same write and read order → OK
- Random reads from many tapes → Slow

Write and read patterns differ frequently.

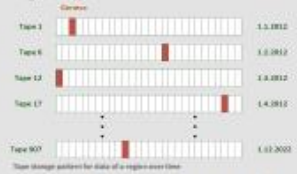
Exploitation Platforms

- Initial load: certain types of data or regions (e.g. Europe) → selective reads over large number of tapes
- Reloads: by user demand → random reads

Time series over a certain region

- Writes by archiving time
- Reads by location

Example: Time series for images taken over Geneva, using monthly image over 10 years.



Tape storage pattern for data of a region over time

With large amounts of data, reading data like this takes forever!

Workarounds

Optimize access patterns

Data extraction from the archive can be optimized within processing chains, done e.g. for the DLR TIMELINE project (see PV presentation by M. Wolfmüller [1]).

- Start extracting data before processing
- Build a pipeline reading data from tape, extracting from the archive and processing data in parallel bulks

Drawbacks:

- Need to know data needs in advance
- Needs sufficient amount of disk cache / processing cache
- Considerable effort to adjust processing chains, cache sizes and processing resources
- With more processing power, archive resources still remain the bottleneck
- Hard for external systems (e.g. platforms): no knowledge about archiving environment
- User driven / non-systematic data requirements not predictable

Keep everything online

With larger disks and falling disk prices, more data can be kept online.

Drawbacks:

- Data volume growing exponentially
- Cost + energy consumption of disk still higher than tape
- For reliable archive: 2 disk copies in different locations, needs twice the disk capacity

Our Solution

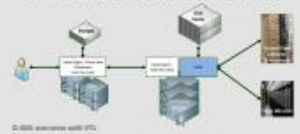
We now store our primary copy on a virtual tape library, which emulates a tape library, drives and tapes.

Virtual tape library (VTL): FastLTA Silent Brick System

- Disk based → **Fast positioning, less latency**
- Virtual tape = Disk shelf with erasure coding
  4 out of 12 disks **redundancy → data safety**
- Powers down disks for "unmounted tapes"
  Max. number of active disks controlled by virtual drives
  → Limit and control energy consumption!

Integration in archiving environment

- Configured with 26 disk shelves / 2 virtual drives per VTL
- Integrated as "tape library" with existing hierarchical storage management software (Oracle HSM)
  → Established mechanisms for data steering
  → No change for archiving application
- Keep one real tape copy in different location (building)



D-SDA overview with VTL

Experience

FastLTA Silent Brick Systems have been in production since 2021.

Operations

- Mostly stable
- Occasional disk failures and firmware updates
- Two unplanned downtimes of an entire system (central part failure), less than five unplanned downtimes for a single Shelf
- No data loss

Energy consumption

- Vendor information: typical usage for one system with 2 active Shelves: ~1,2kW
- Power measurement currently only possible for entire room including other servers and storage systems
- Live measurement for traditional tape library: ~ 5kW
- Systematic measurements still outstanding

Performance

- Power-on time: slightly faster than tape mounts
- "Spooling" time: none
- Bandwidth higher than our current tape drives
- Overall read times reduced, especially long reads

Read times extracted from archiving application logs for data deliveries for AC-SAF project:

Application read times: tape vs. virtual tape library

References

1. M. Wolfmüller, S. Holzwarth, S. Asam, S. Kiemle, D. Krause, A. Scherbachenko: Optimized Data Access from and to a Long-term Archive for the Processing of Time Series, PV 2023, CERN, Geneva, Switzerland

DLR

PRESERVING AN ENDANGERED SOCIETY
THE CASE OF MALDIVES

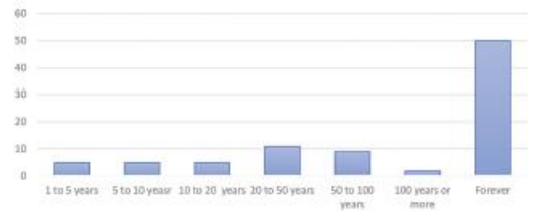
Maldives has the lowest terrain of any country in the world making it very vulnerable to sea-level rise. Much of it is likely be uninhabitable by 2050.

The population would have move, for example, to artificial floating platforms or elevated land purchased in other countries.

Governmental, legal, societal, commercial and personal information, currently in many organisations, which is the lifeblood of Maldives society, must be preserved for use in the new location.

Survey to find requirements for preservation system

some results:

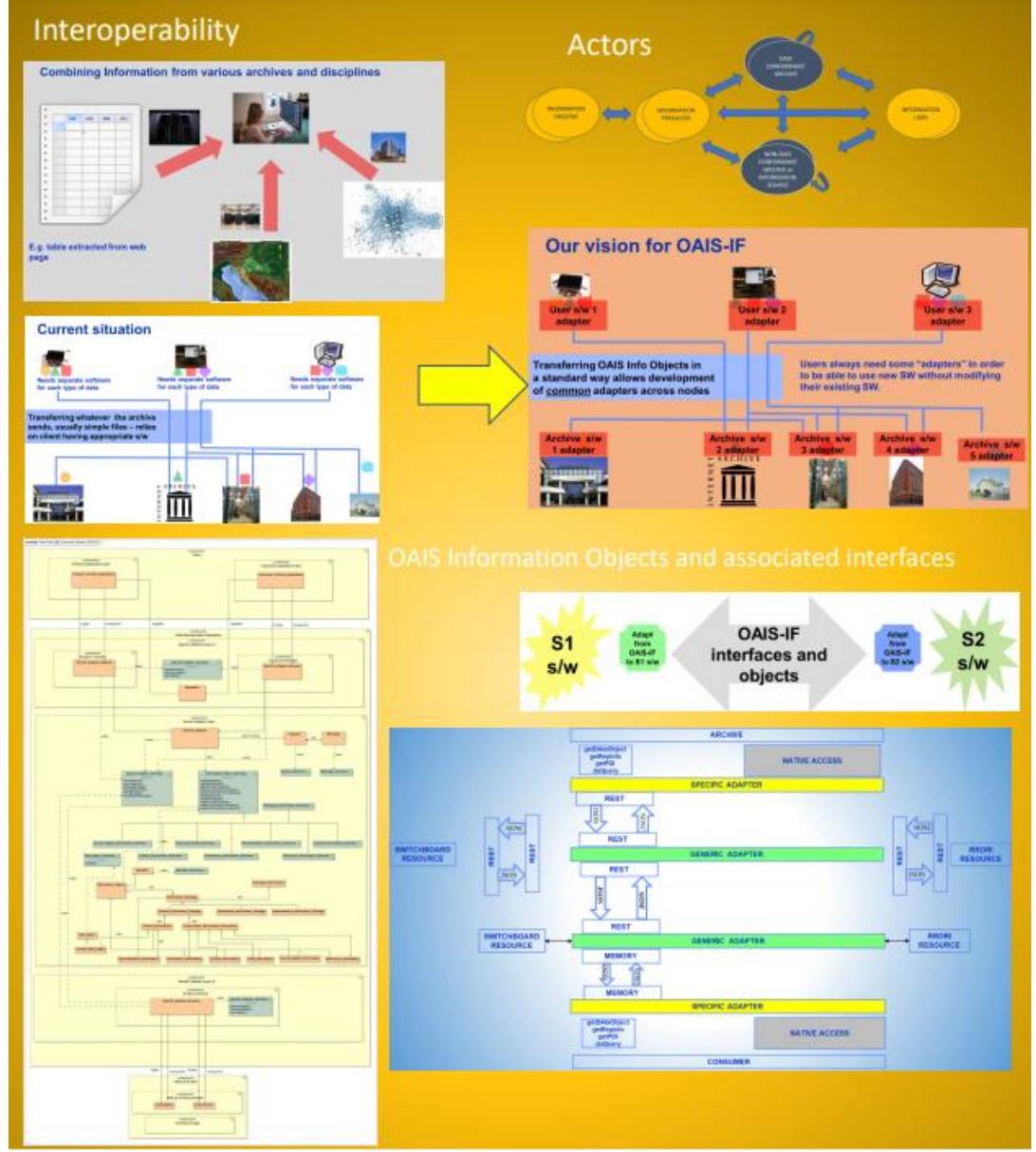


- **Need massively scalable OAIS and ISO 16363 conformant archive system**
- **Must justify resources needed for archive by adding value by enabling re-use and combining**
- **Must also engender culture changes throughout to capture information needed**

David Giaretta
PTAB - Primary Trustworthy Digital Repository Authorisation Body Ltd
Dorset, UK
david@giaretta.org
http://www.iso16363.org/

Ahmed Asim
National Archives of Maldives
H. Keleethia, Hakuraa Goalhi
Malé 20097, Maldives
ahmed.asim@archives.gov.mv
https://archives.gov.mv/

iPRES 2023 CONFERENCE



Developing an OAIS based Interoperability Framework

D. Giaretta[1], M. Kearney[2], J. Garrett[3], and S. Hughes[4], Terry Longstreth[5], Roberta Svanetti[6], Robert Rovetto[7]
Members of CCSDS Data Archive Interoperability Working Group https://cwe.ccsds.org/moims/default.aspx#_MOIMS-DAI
[1] PTAB Ltd, and Giaretta Associates Ltd, Dorset, UK, david@giaretta.org
[2] Sponsored by Google, Huntsville AL USA, KearneySolutions@gmail.com
[3] Garrett Software, Columbia, MD, USA, garrett@his.com
[4] Jet Propulsion Laboratory, California Institute of Technology, Pasadena CA USA, J.Steven.Hughes@jpl.nasa.gov
[5] Consultant, Data and Information Standards, longstreth@acm.org
[6] Deda Cloud Srl, Roberta.Svanetti@dedagroup.it on behalf of ESA
[7] Independent consultant, New York, USA; & Europe, ontologos@yahoo.com, https://purl.org/space-ontology

Towards a sustainable data and knowledge preservation

Aurèle NICOLET*, Basma MAKHLOUF SHABOU*

(*) Information science, Geneva School of Business Administration (HEG-GE), University of Applied Sciences and Arts Western Switzerland (HES-SO)

Digital general environmental impacts

Research data's environmental issues

Collecting data

Collection instruments (sensors, scanners...)

Huge amounts of data that affect its processing and storage

Processing and analysing data

Software is not necessarily designed to optimize energy consumption.

Code decay for which one of the reasons is the change of hardware

Preserving data

Environmental control (temperature difference, humidity control) requires electrically powered devices

File format and hardware migration
- regular hardware changes (media obsolescence)
- checksum calculation[2]

Redundant backups to ensure retention and accessibility requires hardware for storage

Arch'Eco Project

Main objectives

- Prospect for good practices in data and records management
- Characterize ecological issues in data and records acquisition and processing
- Identify and quantify the environmental impacts of data processing via the life cycle of data and archives
- Define the functional and technical specifications of a tool to manage and control environmental costs
- Recommend resources to better support ecological sustainability to decision makers

Research program

- inventory of practices and mapping of processes and tasks applied to the acquisition and processing of data and archives
- review of systematic literature and case studies to identify good practices (methods, tools, metrics) in sustainable data and records management
- development of a conceptual model with an overview of archival functions, tasks and tools
- mapping of metrics, stakeholder consultation sessions to collect academic and professional validation

Contact

Haute ecole de gestion de Geneve
Campus de Batelle - Bâtiment B
Rue de la Tambourine 17, 1227
Carouge
basma.makhlouf-shabou@hesge.ch
aurele.nicolet@hesge.ch

References

[1] BORDAGE, Frédéric, 2019. *Empreinte environnementale du numérique mondial* [Online]. GreenIT. September 2019. Retrieved from: https://www.greenit.fr/wp-content/uploads/2019/10/2019-10-GREENIT-etude_EENM-rapport-accessible-VF_.pdf

[2] KINNAMAN, Alex and MUNSHOWER, Alan, 2022. Green GeekS with Anything: Decreasing Environmental Impact of Digital Libraries at Virginia Tech. In: *iPres 2022: The 18th International Conference on Digital Preservation*. Online. Glasgow Scotland UK. 13 September 2022. Retrieved from: http://hdl.handle.net/10919/112382

Icons from Noun Project by
- Ali Nur Rohman
- Annette Spithoven
- Art shop
- Creative Stall
- Eucalyp
- Gung Yoga
- Jan Niklas Prause
- JunGSA
- Lluisa Iborra
- Made X Made
- Mohamed MB
- Muhammad Sukirman
- Vectors Market
- Vectors Point

- Data standardization and integration for maintenance of Critical Infrastructure
- Adithya Thaduri

- TFCat (Time-Frequency Catalogue): JSON Implementation and Python library
- Baptiste Cecconi

- Design and Implementation of FENGYUN Meteorological Satellite Archive Data Migration Mission
- Zhe Xu

CryoSat Ice Baseline-E: Operational & Reprocessed Data Quality Control

L. Toonen[1], M. Williams[1], E. Turner[1], A. Di Bella[2]
[1] Telespazio UK Ltd (UK) e-mail: Erica.Turner@telespazio.com; [2] SERCO c/o ESA/ESRIN (Italy)

IDEAS-QA4EO

TELESPAZIO a LEONARDO and THALES company

CryoSat Mission

- Launched in 2010
- ESA's dedicated ice mission
- SAR Interferometric Radar Altimeter (SIRAL) can measure high-resolution geophysical parameters over all ocean and ice environments.

CryoSat Users!
We need your feedback!
Visit our poster, find out more about our QC activities, and fill in the user survey.

What do we do?

- **Operational Data Quality Control**
  We add **value** by:
  - Informing users about the quality and completeness of the data.
  - Investigating unexpected data gaps, failures or missing input files to maintain data quality and availabity of operational data.

- **Reprocessed Data Quality Control**
  We add **value** by:
  - Supporting the selection of requirements for new processor evolution.
  - Testing and verification of the new processors.
  - Investigation of anomalies and failures.
  - Detailed documenting of guidelines to transfer knowledge for upcoming processor evolutions.

PV2023: Adding value (to) and preserving Scientific & Technical data: 2nd – 4th May 2023, poster ID 3178

Converting a Traditional Space Science Operations Center to a Cloud-Based Architecture

J. Mukherjee (joey.mukherjee@swri.org), C. Gonzalez, K. Pickens, U. Salman, S. Ybarra

ABSTRACT

Over the course of the last few years, few technologies have become as ubiquitous as cloud technologies. The building block of this revolution are the "containers" using tools such as Docker. Building on top of containers are orchestration tools such as Docker Swarm or Kubernetes. Science Operations Centers (SOCs) are what a space science project uses to allow members of a team to collaborate more easily in doing research, as well as to manage the operations of an instrument or spacecraft. They are also typically responsible for data processing and archiving of space science data. These SOCs can take advantage of the technological advances in cloud technology to improve both user experience and data integrity, as well as enhance the ability of a SOC to process and archive data for use while the mission is active or even after the mission is complete.

Converting to a cloud-based infrastructure can be a challenging endeavor and is normally accomplished by the "lift-and-shift" approach, (i.e., moving an existing code base from on-premises systems to the cloud), redoing everything to take advantage of the scalable nature of the cloud, optimizing for the main cost drivers of the cloud (e.g., ingress and egress of data), or some combination of the two.

Our approach was to convert our existing SOCs to a more cloud-based infrastructure without having to rewrite the entirety of our SOC, but to redo certain pieces such that they work more efficiently with our local resources and moving targeted components to the cloud as necessary. In our case, we moved our front-end web application, data processing, image analysis, and data visualization software to the cloud.

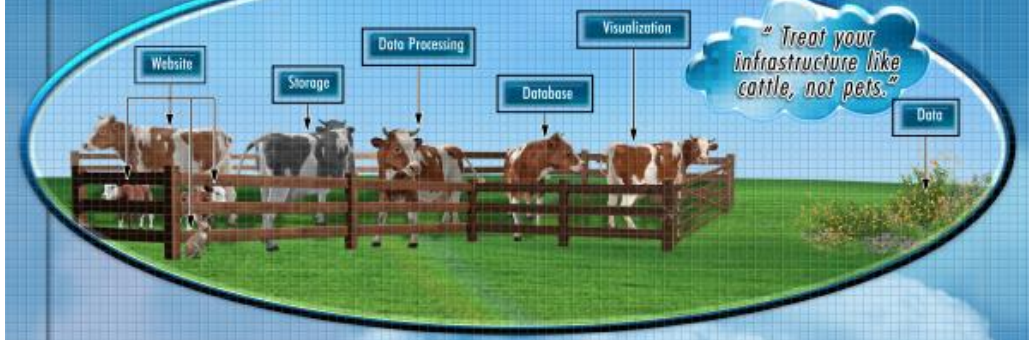
Furthermore, by having the data on the cloud, interested parties can have access to the data more easily by moving their software to the cloud rather than the traditional approach of downloading the data to their local systems. This paradigm shift of moving software to the data will be fundamental as data volumes continue to grow in the future.

PROBLEM

Data volumes are growing out of control and although possible to store the data locally, it is getting increasingly unwieldy. Storing data on the cloud makes sense, but getting software to work with the data on the cloud without having to download will work best with the technology of containerization.

SOLUTION

Move software to the data!



FUTURE WORK

The containers should be ephemeral and there is work to be done in this space. As an example, databases cannot be shut down and restarted as easily as a webserver so migrating to a cloud friendly database such as CockroachDB is one alternative. The cloud computing world has so many plugins and components that more time should be taken to learn how these can help.

Technology Stack

- Rancher
- Docker
- Kubernetes
- CakePHP
- RabbitMQ
- MySQL

Petabyte scale, OAIS/ISO 16363 conformant archive

David Giaretta [1], Antonio Guillermo Martínez Largo [2], María Fuertes [3], Teo Redondo 4

PetaByte Scale

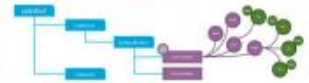
LABDRIVE has been tested to receive +610 million files and 15PB in a month (500TB/day data rate), scaling itself to more than 6500 Kubernetes pods to process the workload.

Core Capabilities:

Scalability:

Supports the whole data lifecycle:

Data Containers:



Types of storage:

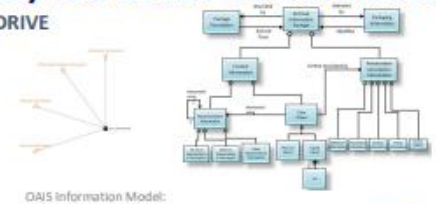
Configuration to support OAIS / ISO 16363 Conformance

OAIS Preservation Operations - supported by LABDRIVE

Data Object may be:

1. Kept by archive, unchanged
2. Kept by archive but may be changed
3. Not kept by archive – handed over

Case 1 – Archive adds Representation Information
Case 2 – Transform the Data Object
Case 3 – Create and hand over complete AIPs



OAIS Information Model:

Representation Information Network:

ARCHIVER Project Framework

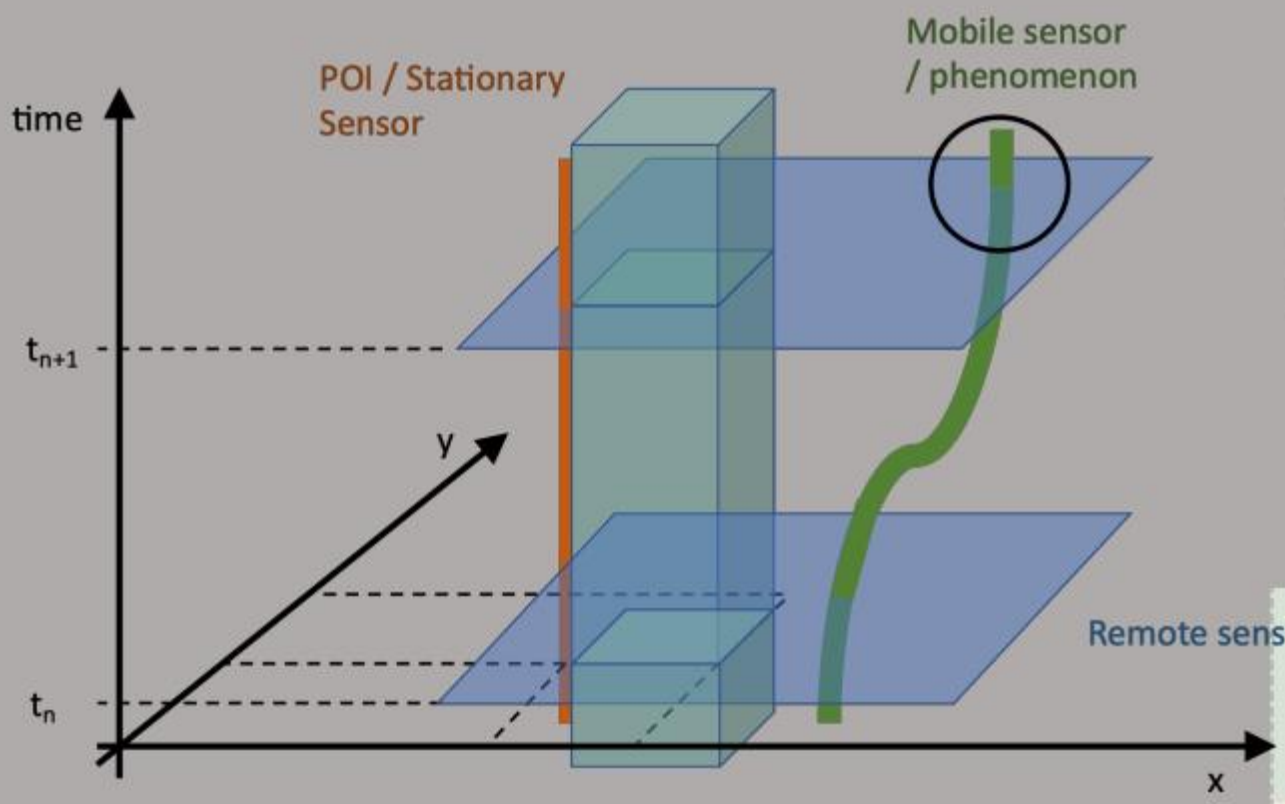
LIBNOVA CONSORTIUM (ARCHIVER Project' winner):
LABDRIVE PLATFORM is a Research Data Management and Preservation platform resulting of a joint effort and intense R&D.
With it, Researchers can **do more** while Organizations **reduce risks and costs**.

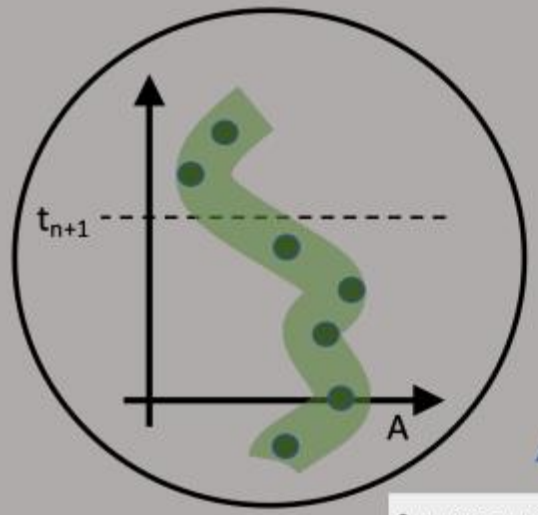
ARCHIVER PROJECT:
Archiving and Data Preservation services for Research Environments for PB-scale datasets using commercial cloud services via the EOSC.
https://archiver-project.eu/

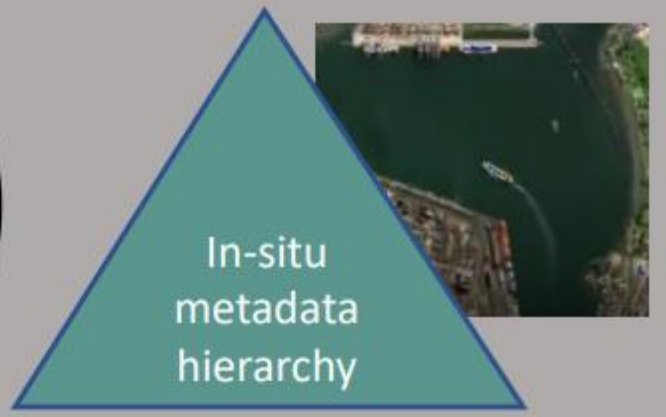
[1] PTAB Ltd, Dorset, UK, david@giaretta.org | [2] LIBNOVA, Spain, a.guillermo@libnova.com |
[3] LIBNOVA, Spain, mfuertes@libnova.com | [4] Formerly at LIBNOVA, now at Bidaidea, Spain, teofilo.redondo@bidaidea.com

- Data structure and long-term preservation of sequential images from optical microscopy
- Jordi Andilla

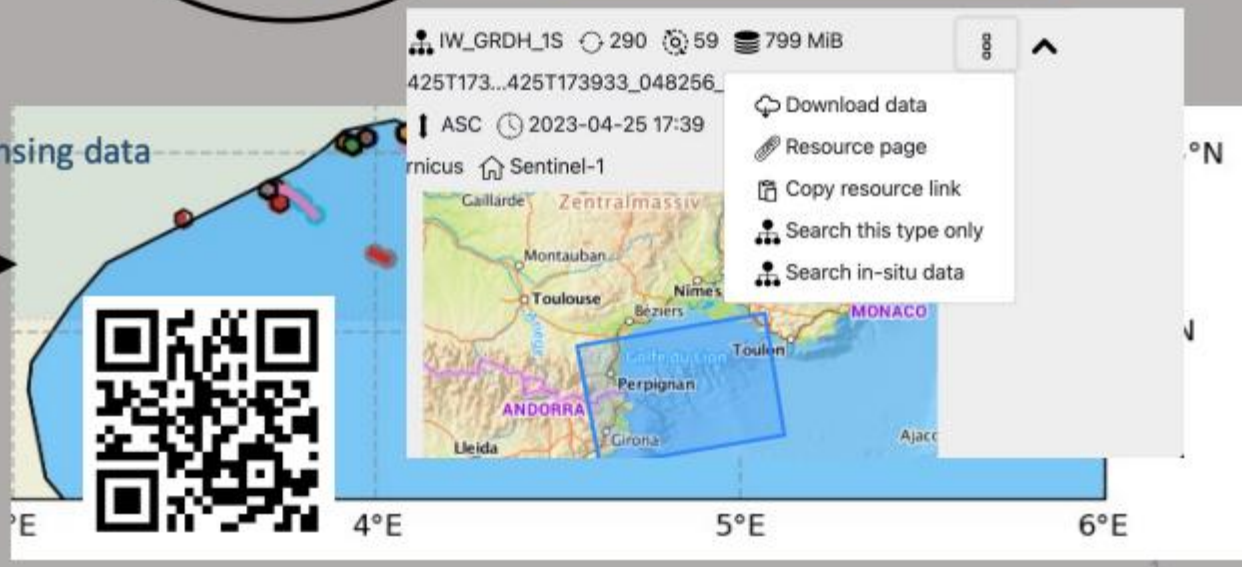
In-Situ and IoT data inventory for EO data







Remote sensing data



analysis and validation

Two Decades of ATSR data from three instruments!!

ESA Heritage Mission

Level 0 – Level 1B reprocessing

Level 1B ATSR dataset improvements feed into *Level 2* datasets in: Sea-surface temperature, Land-surface temperature, Aerosol, Cloud and many more!

ALOS Third Party Mission – AVNIR-2 and PRISM

Level 0 – 1C reprocessing

Enhanced datasets lead to improved cartographic and vegetation mapping!

Can be used synergistically with other ESA/Copernicus data, Sentinel-2, Landsat

Science and Technology Facilities Council

RAL Space

PV 2023 CONFERENCE

TELESPAZIO a LEONARDO and THALES company

serco

"Adding Value to ESA Heritage and Third-Party Mission Archived Datasets via Reprocessing: ATSR and ALOS"

Fay Done, Pauline Cocevar, Michael Williams, Sabrina Pinori, Philippe Goryl and Roberto Biasutti

Computational appraisal for enhancing data value measurement :
recent researches and lessons learned for scientific data governance
Recap of main tested ideas

Basma Makhlouf Shabou* **Aurèle Nicolet***
(*) Information science, Geneva School of Business Administration (HEG-GE), University of Applied Sciences and Arts Western

QADEPs (2012-2013)

Quality of Public Electronic Data and Documents aims to identify and measure the qualities of digital archives in order to better systematize their assessment.

Appraisal Systematization

Outputs : set of variables for the operationalization of the appraisal criteria and a tested method for their application

Infonomics (2013-2016)

Infonomics sought to appraise and value Information Assets (IA) with a multidisciplinary approach, based on three dimensions: data values, costs and risks. Infonomics sought to appraise and value Information Assets (IA) with a multidisciplinary approach, based on three dimensions: data values, costs and risks.

Outputs : A conceptual and operational framework for the evaluation of IAs, use cases, a maturity model and appraisal metrics derived from the three dimensions and a methodology for evaluating and pricing IAs

Appraisal Automation

Archiselect (2017-2018)

Archiselect aims to provide an aid for the process of macro- and micro-appraisal of both, structured digital files and unstructured data sets in the context of public or private administrations.

Outputs : An algorithmic approaches and data mining methods were then tested to propose measurable and computational data appraisal metrics

Appraisal Smartification

Maturity Assessment for Appraisal in the AI Age (2023 – 2025)

Assessing the maturity of appraisal processes and tools will allow us to identify the archival, technical, technological, cultural, and strategical barriers and facilitators to effectively apply AI tools for appraisal processes.

InterPARES Trust AI - Artificial Intelligence

It will addresses the following questions:
1) how defensible are current appraisal decisions?
2) How stable, coherent, and reproducible are appraisal practices?
3) What are the prerequisite conditions of AI integration to a given appraisal practice?
4) How are data, records and archives prepared to be appraised automatically and or "smartly"?
5) What are the complementary actions to upgrade appraisal practices for the AI facilities?

InterPARES Trust AI

CONTACT
basma.makhlouf-shabou@hesge.ch

h e g Haute école de gestion Genève

Hes·so Genève University of Applied Sciences and Arts Western Switzerland

Focusing on Scalable Citations to Improve Data Usability & FAIRness

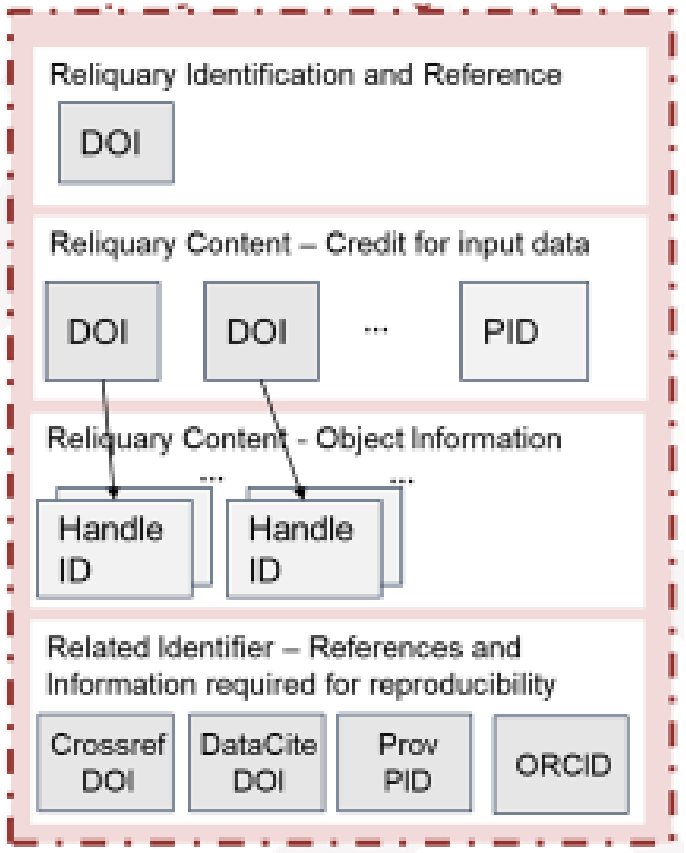
Challenges in data citation:

- Citing many datasets across repositories
- Citing subsets of a larger dataset
- Tracing results back to their origins

How to include in reference lists?

How to gather all references?

How to include in credit systems?



Object Collections:

- Information on citing the collection
- Information on contained citable objects
- Optional information on relations, persons, provenance, project, etc.

Next Steps:

- Gather use cases
- Utilize existing PID collection approaches
- Engage with stakeholders of the credit system (crossref, DataCite)
- Involve researchers

Engage:

Join our RDA Complex Citation WG:
https://www.rd-alliance.org/groups/complex-citations-working-group

Tell us about your use cases:
https://docs.google.com/spreadsheets/d/1GSNLw7Pg1qN1A7KT6raTeGs31JGNc43q0Zw4qfyNnFg

Exporting Institutional Repository Metadata as Dataset

Marcelo Garcia Daryl Grenz
Mohamed Ba-Essa

University Library, King Abdullah University of Science and Technology (KAUST), Thuwal, Saudi Arabia

Motivation

Datasets play a crucial role in scientific research, and it is expected of researchers to publish their datasets. Although we generally consider datasets to be the outcome of an experiment or a sensor, a university library also holds a dataset. This dataset is comprised of the university's publications. Interestingly, it seems that institutional repositories don't publish their metadata as a bulk dataset, possibly on the assumption that making the information available for harvesting through interfaces such as OAI-PMH is sufficient to support reuse.

We think a separate bulk dataset in a commonly used format may be useful to potential users who are not able or interested in working through OAI-PMH. For example, such a dataset could be used by administrators of the university for reporting purposes, or by students who need a sample of thousands of files to train their models. Finally, having a list of the university's outputs in an easy to use format may have uses we can't imagine yet.

The inspiration for this work came from the CORD-19 project lead by the Allen Institute for AI (AI2). Where they aggregated information about research papers related to COVID-19 and published an archive of these papers in a machine-readable format so the AI community could use them.

Based on this example, we decided to explore the idea of exporting the content of our institutional repository as a dataset. The outcome is a CSV file, a format that is easy to use in data science or AI workflows.

Methodology

We query our local publication database for selected fields and use the result to generate a CSV file that is publicly posted in a dataset record in our institutional repository. We chose fields that are common to many kinds of outputs, like author, abstract, publication date, and type of publication. We also include a link to an output's full text PDF (if available) and to its extracted text to facilitate the process of tokenization. At this stage we didn't include the text directly in the dataset. It was necessary to change the separator of the authors field to make easier to export to VOSviewer.

Suggested Use

Figure 1. Open Access Dashboard

The idea of this dataset was to invite the KAUST community to explore the academic output of the university. We hope that our community will find other uses, and come back with feedback, for example about which other fields should be included. We already have some ideas of potential uses:

- As the data source for the library's open access dashboard (1). This will require additional open access and departmental information be included.
- To allow users to quickly get a copy of metadata for thousands of KAUST affiliated research outputs that they can sort or filter as they desire without being limited by the search or export functionality of the repository software.
- Allow users to quickly assess the overlap between this list and lists in other databases, such as Scopus, Web of Science or Lens.org by including all of the known external ids in the dataset.

Usage

As an example we created a Jupyter notebook as part of a Github repository, the repository includes the notebook itself and a sample CSV file for development. The full dataset is available on KAUST's institutional repository. The links are below:

- Dataset: `https://repository.kaust.edu.sa/handle/10754/691065`
- Notebook: `https://github.com/kaust-library/RepoDataset`

Using the full dataset, we make simple queries like the percentage of articles, presentations, preprints, etc. See Fig (2a). Or, from the articles, the percentage across the top 10 publishers. See Fig (2b)

(a) Types of publications (b) Top Publishers

Figure 2. Types and Publishers

Another example of usage would be to load into an exploration tool like VOSviewer, to see the relation between elements. We used one of the notebook in GitHub to create a graph in VOSviewer of the elements in the field *abstract*, see Fig (3).

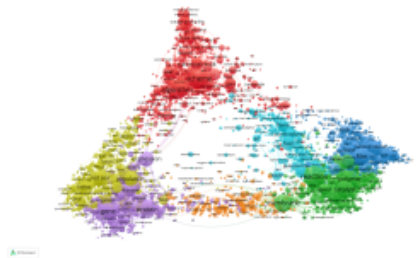


Figure 3. Visualization of items in the abstract.

Conclusion

The idea of this project was to explore another venue for the library to engage with our community, and more specifically with data scientists and AI practitioners. We hope the community will use the dataset and provide feedback on how they are using the dataset and how we can improve the service.

References

[1] Nees Jan van Eck and Ludo Waltman.
VOSviewer visualizing scientific landscape.
`https://www.vosviewer.com/`.

[2] Lucy Lu Wang, Kyle Lo, Yoganand Chandrasekhar, Russell Reas, Jiangjiang Yang, Doug Burdick, Darrin Eide, Kathryn Funk, Yannis Katsis, Rodney Michael Kinney, Yunyao Li, Ziyang Liu, William Merrill, Paul Mooney, Dewey A. Murdick, Devvret Rishi, Jerry Sheehan, Zhihong Shen, Brandon Stilson, Alex D. Wade, Kuansan Wang, Nancy Xin Ru Wang, Christopher Wilhelm, Boya Xie, Douglas M. Raymond, Daniel S. Weld, Oren Etzioni, and Sebastian Kohlmeier.
CORD-19: The COVID-19 open research dataset.
In *Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020*, Online, July 2020. Association for Computational Linguistics.