Minutes of the 59th Meeting of the CEOS Working Group on Information Systems & Services (WGISS)

Chaired by USGS, Vice-Chaired by ISRO
Bangkok, Thailand | March, 2025

Table of Contents

Day 1: Monday 24th March, 2025

Session 1: Opening Session

1.1 - Welcome & Review of Agenda

Tom Sohre (USGS, WGISS Chair) reported [*slides*]:

- Welcomed all to the meeting, and apologised for not being able to attend in person. Thanked Nitant Dube (ISRO, WGISS Vice-Chair) for stepping up to chair the meeting on Tom's behalf.
- Pakorn Apaphant (GISTDA) welcomed WGISS to Bangkok, and is proud to support CEOS activities. Pakorn started his CEOS work within WGISS, being active in the group for almost ten years, before changing roles within GISTDA.
- Participants were invited to introduce themselves. The full list of participants can be found in Appendix A.

1.2 - Keynote from GISTDA

Tanita Suepa (GISTDA) reported [*slides*]:

- GISTDA works across the entire space value chain, including upstream and downstream development, as well as capacity building. GISTDA has a strong focus on developing human resources in Thailand for EO technologies.
- GISTDA's series of high-resolution optical satellites include THEOS-1 (Thaichote, launched in 2008), THEOS-2 (launched in 2023) and THEOS-2A (to be launched in 2025).
- Operate an open platform (AWAGAD) to provide access to both Thai and international satellite data. Accessibility, availability and affordability is key.
- Sphere is GISTDA's cloud-based platform, providing data on the FAIR principles. Provides dataset access and tools, as well as the ability to develop custom applications.
- GISTDA's Digital Twin platform uses 3D City Model and mapping data. From the base map, developed a solar rooftop system application to identify and monitor rooftop solar panels.
- Satellite data, both optical and SAR, is used to understand the impact of natural disasters, including droughts, floods and forest fires.
- GISTDA is also supporting environmental finance, including carbon credit initiatives.

- Dragonfly is a mobile application used by farmers for farm management, including yield estimation.
- Global collaboration and innovation is key to GISTDA's future, including through CEOS and WGISS.
- Thailand Space Expo 2025 will be held in October, bringing together space industry leaders, and connecting the community.

Discussion

- Tanita noted the team is in the process of linking the Digital Twin platform with ground station and IoT data, to improve the real time accuracy.
- Sphere only includes data for Thailand, but the platform is open for anyone to use. It was built by the government, but is available as a prototyping and testing space for academic and commercial applications.
- Pakorn Apaphant (GISTDA) recognised it could be helpful for the Sphere team to participate in WGISS. GISTDA would like the platform to meet CEOS standards where possible.
- Launch for THEOS-2A is expected in a few months, once issues with the launcher have been resolved.

1.3 - CEOS Executive Officer Report

Steven Ramage (CEOS Executive Officer) reported [*slides*]:

- UKSA is the current CEOS Chair. Their themes are focused on getting EO data closer to the users. One aspect is the CEOS in Schools programme, which brings together school age children from around the world to explore how EO can be used to tackle global challenges.
- The CEOS 2025-2027 Work Plan is currently under review, and includes over 100 deliverables. Final review and virtual endorsement is scheduled for after SIT-40. WGISS has seven open and new deliverables:
    - DATA-24-01: White Paper on EO Data collections management and governance
    - DATA-24-02: White paper on Software preservation
    - DATA-24-03: CEOS Interoperability Handbook 2.0
    - DATA-25-01: White paper on federated access
    - DATA-25-02: CEOS Interoperability Maturity Matrix

  - DATA-25-03: White Paper on EO for Digital Twins
  - DATA-25-04: EO Data Citation Guidelines
- The following action was recorded at SIT Technical Workshop 2024:
  - SIT-TW-2024-03: *CEOS Executive Officer will support the WGISS Interoperability team in securing the appropriate engagement from across CEOS, as well as identify potential approaches and opportunities for CEOS to consider for engagement with the means and opportunities for commercial sector engagement. WGISS is to report on progress regularly at CEOS Secretariat meetings.*
- The Executive Officer Team is conscious that they haven't done much to support this activity. WGISS should let Steven know where support is required.
- WGISS has some ongoing and upcoming activities which may be to engage with the commercial sector on. CEOS recently formalised guidelines for commercial sector engagement, in that they are welcome to participate at a technical (i.e. Working Group) level.
- The Biodiversity Study Team have been meeting regularly to explore a post-2024 strategy for CEOS and Biodiversity. They are working directly with stakeholders to complete a table to understand requirements for EO for different stakeholders.
- 40th CEOS Strategic Implementation Team Meeting (SIT-40) will be held 8-10 April, 2025, in Fukuoka, Japan. WGISS will present the draft interoperability handbook.
- Funding is the major challenge for the Group on Earth Observation (GEO), and will dictate the size and shape of the GEO Secretariat going forward. The Post-2025 GEO Work Programme will be adopted at the upcoming GEO Global Forum (5 - 9 May, 2025; Rome, Italy).

Discussion

- Tom Sohre (USGS, WGISS Chair) asked whether there has been any conversations regarding commercial organisations formally becoming members. Steven noted CEOS leadership haven't discussed this, however some of the potential new members have asked. Steven may follow this up regarding the New Space elements in the CEOS Work Plan.
- Nitant Dube (ISRO, WGISS Vice-Chair) recognised that WGISS should coordinate with the GEO Data & Knowledge Working Group. Steven noted that the SEO is participating in the GEOSS Infrastructure Development Task Team (GIDTT).

WGISS-59-01	WGISS Chair/Vice-Chair to share the draft Interoperability Handbook with Paola De Salvo for GEO's review and input.	**Due: Q2 2025**

1.4 - Data Preservation & Stewardship Interest Group (DSIG) Work Plan

Mirko Albani (ESA) reported [*slides*]:

- DSIG is currently working on two deliverables: Data collection management and governance (DATA-24-01), and software preservation (DATA-24-02). An additional deliverable is proposed for 2026, related to citation of EO data (DATA-25-04).
- Ongoing/future deliverables will address:
  - Data associated content and information preservation
  - Long term data preservation processes
  - Data provenance and citation
- WGISS DSIG contributed to the paper Harmonizing Quality Measures of FAIRness Assessment Towards Machine-Actionable Quality Information, led by the Information Quality Cluster (IQC). ICQ won the ESIP collaboration of the year award for 2024.

WGISS-59-02	WGISS Secretariat to link the Harmonizing Quality Measures of FAIRness Assessment Towards Machine-Actionable Quality Information under 'External Documents of Relevance'.	**Due: ASAP**

1.5 - Data Discovery & Access Interest Group (DAIG) Work Plan

Damiano Guerrucci (ESA) reported [*slides*]:

- The STAC Best Practices deliverable (DATA-22-05) is now closed.
- Now starting to work on a white paper on Federated Access (DATA-25-01).
- Connected Data Assets and Data Discovery is an ongoing activity. Encourage all to reach out to connect their assets to FedEO and/or IDN.
- DAIG has also made significant contributions to the Interoperability Handbook.
- Future activities may include:
  - Interacting more with standards organisations such as OGC and ISO
  - Cloud infrastructure
  - Data formats

1.6 - Technology Exploration Interest Group (TEIG) Work Plan

Yousuke Ikehata (JAXA) reported [*slides*]:

- Two deliverables from TEIG were closed last year: AI/ML white paper (DATA-23-01), and Jupyter Notebook Best Practices (DATA-22-01). AI/ML white paper will be continually updated with the latest information.
- TEIG is collaborating with DAIG for DATA-25-01 (White paper on federated access).
- TEIG has opened one new deliverable for 2025, a white paper on EO Data for Digital Twins. This will be finished by WGISS-62 (September/October 2026).

Discussion

- Nitant Dube (ISRO, WGISS Vice-Chair) suggested that the AI/ML white paper suggest mechanisms for publication and discoverability of AI/ML datasets.
- Damiano Guerrucci (ESA) recognised this is related to traceability and data replication, a topic WGISS should consider in the future.
- Yousuke noted that ESIP's AI/ML Ready Data checklist, which TEIG has reviewed, includes requirements for findability and accessibility.

1.7 - Data Interoperability & Use Interest Group (DIIG) Work Plan

Nitant Dube (ISRO, WGISS Vice-Chair) reported [*slides (a)*, *slides (b)*]:

- Currently working on the Interoperability Handbook v2.0, but also considering other activities to support this.
- After discussions this week, the draft (available on GitHub) will be available for community comment. Will present the first draft to SIT-40 (April 2025), and the final draft to SIT Technical Workshop (September 2025). The document will be presented for endorsement at CEOS Plenary (November 2025).
- A total of 93 recommendations are currently included in the Handbook, distributed across the five factors: Vocabulary (14), Architecture (40), Interface (16), Quality (5) and Policy (8).
- The Interoperability Maturity Matrix will be developed over 2026, building off the existing Maturity Matrices from WGISS and WGCV.
- 2027 will focus on Interoperability Demonstrators, with the development of Earth Observation Plug and Play (EOPnP) modules, as well as the CEOS Common Online Dictionary.

1.8 - Select Action Review

Libby Rose (WGISS Secretariat) reported [*slides*]:

- A total of 30 actions from previous meetings remain open, with 19 in progress. Good progress has been made towards WGISS-58 and WGISS-WGCV joint meeting actions, with half closed. Only seven actions from WGISS-57 remain open.
- WGISS-58-01: *Interest Group leads to review the GEO Open Data Open Knowledge (ODOK) actions to identify opportunities for collaboration.*
    - It was agreed there is nothing specific to follow up on this topic, and the action can be closed.
- WGISS-58-29: *WGISS members should ensure they have a GitHub account, and are a member of the CEOS Organisational account. Contact Brian Terry (SEO) for assistance if needed.*
    - A final reminder to all to complete this action. The action will now be closed.
- WGISS-55-05: *WGISS (Yves Coene, Damiano Guerrucci) to invite WGDisasters to provide tools and software developed by WGDisasters so they can be made discoverable to global users through the service discovery.*
    - WGDisasters was contacted in September 2023, however no response was received. It was agreed the action can be closed.
- WGISS-58-30: *WGISS Secretariat to put together guidance for Interest Groups about surveying CEOS Agencies on various topics.*
    - Some points to consider include target audience & goals, number of responses expected, and the effort required to respond.
    - If a survey is determined to be the best way forward, it is recommended that questions be reviewed by Exec before the survey is distributed.

Session 2: Agency Reports

2.1 - UKSA

Robert Fletcher reported [*slides*]:

- UKSA partners with UK industry and the wider European market on space assets. Upcoming missions include Biomass (launching in April 2025), MicroCarb (launching in Summer of 2025), and TRUTHS (launching in 2030).

- TRUTHS will be a calibration laboratory in space, providing a benchmark reference of the optical radiation state of the planet.
- SSTL, a start-up built out of the University of Surrey, is currently building ten satellites. This includes the HotSat series and THEOS-2A (for Thailand).
- UK EO Data Hub is a joint programme between a number of EO-related organisations in the UK, and funded by UKSA. It is a pathfinder project, where an EO data platform has been built to access various free and open datasets, as well as commercial data. The pathfinder concluded in March 2025, and funding is now being secured for the next phase.
- UKSA is CEOS Chair for 2025, and is working on four key areas:
    - Share knowledge and best practice for bridging the gap between EO data and public services
    - Unlock EO data for the UN Framework Convention on Climate Change global stocktakes
    - Champion the joint UK-NASA led work to develop best practices for the space-based measurement of methane emissions – and push for their adoption in an update to the Global Methane Pledge at COP30 in Brazil.
    - Inspire the next generation with 'CEOS in Schools'.

Discussion

- TRUTHS will use solar and lunar calibration techniques to provide highly accurate measurements on the ground, for better calibration.
- Tom Sohre (USGS, WGISS Chair) recognised the UK EO Data Hub would be replicating data. Did the pathfinder project develop mechanisms to pull data from Copernicus Data Space Ecosystem (CDSE), ensuring the data is kept in sync?
- Robert noted the EO Data Hub pulls directly from source. For example, Sentinel data comes directly from the CEDA archive, and is fed through daily. Currently, mainly UK Sentinel data is hosted, but a future evolution could be to integrate the EO Data Hub with the CDSE to pull other data as required.

2.2 - CSIRO

Matt Paget reported [*slides*]:

- Have implemented significant upgrades to the EASI (Earth Analytics Science and Innovation) platform recently, to reduce maintenance overhead and increase scaling efficiency, improving the overall efficiency of the platform.
- The CEOS Analytics Lab (CAL) is also built on EASI architecture, so the same updates will be rolled out to CAL shortly.
- A workflow in EASI has been built to perform an intercomparison exercise for various Sentinel-1 archives. The archives from Alaska Satellite Facility (ASF) and Geoscience Australia have been compared, and CSIRO is discussing with the NASA JPL NISAR and OPERA teams to include their datasets as well. While CEOS-ARD for SAR is well specified, there are choices to be made in the processing chain which means the resulting products are not equivalent. Results will be published shortly.
- NovaSAR, a partnership between the UK, Australia, Philippines and India, is able to image in HH polarisation, making it useful for flood detection.
- EASI has been used for a number of applications, including a Land Cover Classification System and AquaWatch Australia.
- Australia is also working to unlock their AVHRR archives, and will present their work on atmospheric correction at LPS 2025.
- CSIRO also processes Himawari data, developing a TOA workflow, picking up the data from an AWS bucket managed by NOAA. The current workflow includes unpacking and combining the data, which is not ideal.
- Support for hyperspectral data is being added into EASI, using new capabilities in Open Data Cube (ODC) 1.9. The new version improves flexibility and efficiency for search and load, with wavelength as a dimension.
- CSIRO would like to implement CEOS Interoperability Handbook recommendations into their services, including EASI.

Discussion

- For Himawari, the number of bands means the data volumes are quite large. The workflows are still being developed, and ZARR and Kerchunk will be explored to help tackle the challenge.

- Robert Fletcher (UKSA) met with Alex Held (CSIRO) regarding AquaWatch last week, and noted a version of EASI has been deployed for AquaWatch.
- Matt highlighted the importance of a governance programme to support these data cube technologies.
- Alex Leith (SEO) has been running cloud native geospatial workshops for the SEO, which includes using the ODC STAC implementation. The workshops will be run at the Cloud Native Geospatial Conference (May 2025; Utah, USA), and IGARSS 2025 (August 2025; Brisbane, Australia).
- Matt recognised there is organisational learning that could be shared with WGISS around managing and exploiting large datasets.
- Alex noted specific training on how to scale applications for large datasets can be very specific, however the notebook applications side can be very transferrable.

WGISS-59-03	Organise a session at WGISS-60 on lessons learned from working with large volumes of data (big data). Explore the concept of compute next to data.	**Due: WGISS-60**

2.3 - GISTDA

Prayot Puangjaktha and Pawarin Kuha reported [*slides*]:

- GISTDA operates a ground station in Chon Buri, Thailand.
- THEOS-1, launched in 2008, remains active, with a resolution up to 2 m. THEOS-2, launched in 2023 achieves a resolution up to 0.5 m.
- Actively developing THEOS-3, a microsatellite VNIR/SWIR constellation planned for launch from 2028. THEOS-4 is planned to be a constellation of SAR satellites, with two X-band and two C-band platforms planned to launch from 2031. THEOS-5 will be a pair of very-high resolution optical imagers, planned to launch from 2031.
- GISTDA also downlink and distribute data from global partners. Data is stored in the National Space Data System (NSDC).
- GISTDA is working on achieving compliance with the CEOS-ARD specifications for THEOS-1 and THEOS-2 data.
- Have completed the Data Management and Stewardship Maturity Matrix (DMSMM) self assessment, and found that most elements are partially managed. Working on improving the online access of the satellite products.

Areas	Components	THEOS-1	THEOS-2
Discoverability	MMP1: Metadata for Discovery	★ ★	★ ★
Accessibility	MMP2: Online Access	★ ★ ★	★ ★ ★
Usability	MMP3: Data encoding	★ ★	★ ★
	MMP4: Data Documentation	★ ★	★ ★
	MMP5: Data Traceability	★ ★	★ ★
	MMP6: Data Validation	★ ★ ★	★ ★ ★
	MMP7: Data Metrology	★ ★	★ ★
	MMP8: Data Quality Control	★ ★	★ ★
Preservation	MMP9: Data Preservation	★ ★	★ ★
	MMP10: Data Verification	★ ★	★ ★
Curation	MMP11: Data Processing/Reprocessing	★ ★	★ ★
	MMP12: Persistent & Resolvable Identifier	★ ★ ★	★ ★ ★

- GISTDA is planning for their corner reflectors to join CEOS SARCalNet, to improve data sharing.
- GISTDA participates in Sentinel Asia, to support disaster monitoring and response efforts across the Asia-Pacific region.

2.4 - JAXA

Makoto Natsuisaka reported [*slides*]:

- EarthCARE, a joint ESA-JAXA mission, was launched on May 29, 2024. JAXA provides the Cloud Profiling Radar (CPR) onboard EarthCARE. Level-1b products were released in January 2025, and Level 2 products were released in March 2025.
- ALOS-4 was launched on July 1, 2024. The initial calibration and validation has now been completed and data products are available through two different commercial providers.
- GOSAT-GW is expected to launch in the 2025 Japanese Financial Year (JFY2025), carrying the AMSR3 sensor, which will extend the 22+ year continuous archive of microwave observations. GOSAT-GW also carries TANSO-3, led by the Japanese Ministry of the Environment (MOE), which will improve observation capability of greenhouse gases from GOSAT-2/TANSO-2.
- G−Portal (https://gportal.jaxa.jp/) is a data dissemination system for JAXA EO products, enabling search & download of products via HTTPS and SFTP and providing products, documents, tools etc. G-Portal is connected with FedEO and IDN, and the CEOS OpenSearch server and WebAPI server have been updated.

- JAXA is investigating the implementation of STAC for G-Portal, and is starting the migration to the cloud.
- Makoto participated in the CGMS WGIV Cloud Technology workshop in February 2025.

2.5 - ISRO

Sai Kalpana reported [*slides*]:

- ISRO's current EO missions include RESOURCESAT-2A, EOS-04, CARTOSAT-2E/3, EOS-06, SARAl and INSAT-3DR/-3DS. Upcoming missions include NISAR, EOS-09, OCEANSAT-3A and TRISHNA.
- Bhoonidhi is ISRO's EO Data Hub, which includes ISRO and partner datasets. New products include those from EOS-04, RESOURCESAT-2A and Sentinel-2C. The products can all be accessed via the STAC-enabled API.
- MOSDAC now includes data from INSAT-3DS, the newest geostationary meteorological satellite positioned over India.
- ISRO is working with DSIG and ESA to restore the AVHRR data held by ISRO, to develop Level 1B and Level 1C products. The products are now available on Bhoondihi for dissemination.
- Bhuvan provides geospatial data and services. Night time light data from 2012-2023 has been produced for the entirity of India, showing the rapid development of the country. New optical and SAR satellite data layers are also now included.
- VEDAS is a visualisation tool for geospatial analysis.
- The National Information System for Climate and Environment Studies (NICES) has developed EO based methodologies for Essential Climate Variables (ECVs). 12 ECVs and over 60 geophysical products are available with a time span of 5 to 30 years.

Discussion

- Libby Rose (WGISS Secretariat) noted the potential connection between NICES and WGClimate's ECV Inventory.
- Yves Coene (ESA/Spacebel) added that ECVs are included as keywords in FedEO. WGISS should connect with WGClimate to understand the discovery of ECV datasets, and improve their representation in FedEO and IDN.
- Mirko Albani (ESA) recognised more coordination on ECVs datasets would be beneficial, in particular when it comes to standards and specifications.

- Matt Paget (CSIRO) suggested ECVs could be considered as a case study for interoperability.

WGISS-59-04	WGISS Chair to connect with WGClimate for a discussion about ECV specifications, discoverability and archival. Consider how ECVs/CDRs could be used as a demonstrator for interoperability.	**Due: WGISS-60**

Day 2: Tuesday 25th March, 2025

Session 1 & 2: Opening Session & Agency Reports (cont.)

1.9 - System Engineering Office Report

Dave Borges (NASA, SEO) reported [*slides*]:

- CEOS Analytics Lab is a cloud native analytics platform, which evolved out of the Earth Analytics Interoperability Laboratory (EAIL). It is a CEOS-specific platform for technical collaboration within CEOS. Primarily based on AWS, and supported by SEO, CSIRO and Chilean partners.
- Regarding WGCV-WGISS-24-12 (*SEO to discuss with WGISS DAIG regarding connecting IDN & FedEO to CEOS Analytics Lab*), SEO has reached out to the DAIG team, but haven't had a dedicated conversation yet.
- SEO has created an Organisational GitHub for CEOS, on which there has been a slow but steady increase with repositories and participation. Encourage all to request membership.
- Historically CEOS has done very little on the topic of licensing, which has its challenges due to the fact that CEOS is not a legal entity. Trial licensing at repository level has been implemented to provide open-source licensing where necessary but maintain copyright ownership by individual contributors (agencies).
- SEO is using with Apache 2.0 for software, and CC-BY 4.0 for data & content.
- CEOS-ARD repository has some contributor guidelines, and SEO is working to develop standardised guidelines for the broader account. Will share the draft with WGISS when it is ready.
- Regarding WGCV-WGISS-24-11 (*SEO to discuss with Paolo Castracane & Peter Strobl (and perhaps NOAA Knowledge Graph team) about transferring the current Cal/Val Portal Terms*

& Definition Wiki to GitHub KCEO repository, considering the appropriate format), SEO is expecting WGCV to lead this, but remains ready to support.

- SEO is considering closing down the CEOS COVE Portal as use metrics are quite low. Intent would be to make the individual tools available as open source through GitHub.
- SEO is working on developing CEOS Liaison for Earth Observation (CLEO), a natural language chat query tool, specifically tuned to CEOS content (website, documents, MIM Database). The tool is almost ready for beta testing, and WGISS members are welcome to join the beta testing group.
- Also developing EO-GPT, which will provide decomposable Earth Observation analysis capability, with the ability to have natural language conversations to conduct complex, Earth observation analysis and research.
- WGISS-58-21 (*TEIG to investigate with DAIG and SEO the available technologies for federated authentication, and which would fit with the objectives of CEOS*) will remain open. SEO remains interested in learning more about that topic, and understanding the feasibility of federated authentication for CEOS infrastructure.

WGISS-59-05	WGISS members to contact the SEO to be involved in beta testing of either CEOS Liaison for Earth Observations (CLEO) or EO-GPT.	**Due: Q2 2025**

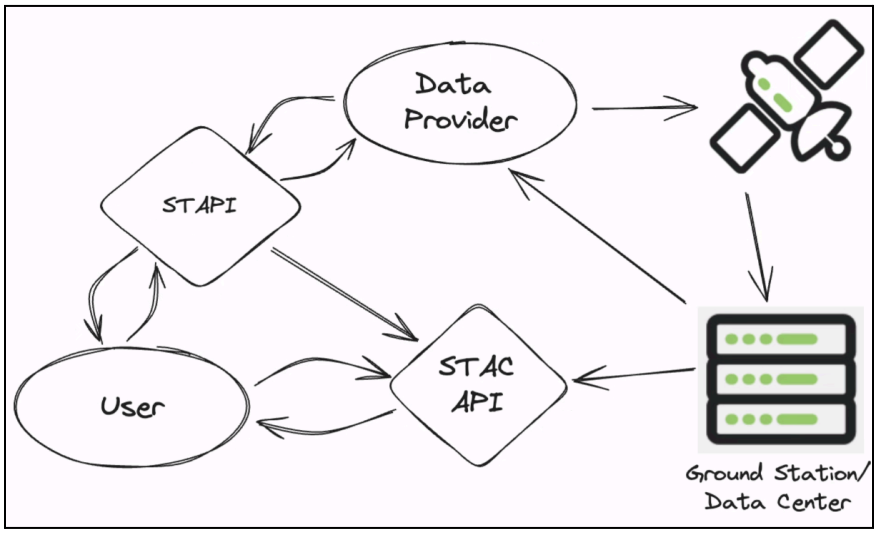
Discussion

- Nitant Dube (ISRO, WGISS Vice-Chair) noted there is a plan to use the CEOS Analytics Lab for interoperability demonstrators, including the EO plug and play modules. Likely won't be this year, but in the future.
- Regarding licensing, Nitant noted the Interoperability Handbook has been licensed as CC-BY 4.0. This also relates to the legal interoperability in the Policy factor of the handbook.
- Matt Paget (CSIRO) recognised the exceptional capabilities of COVE, however with the low usage statistics, agrees it would be best to move the code to GitHub. It would be a shame to lose the tool completely.

1.10 - Satellite Tasking API (STAPI)

Matt Hansen (Element84) reported [*slides*]:

- Matt has been involved in the open source community, and helped develop STAC. STAC was developed through 'guerilla standards' - an open source way to develop standards and drive adoption.
- For commercial satellites, the real value is in the new and future data. Developing better ways to leverage and order data is key.
- The Satellite Tasking API (STAPI) focuses on standardising and improving the efficiency of interactions between a user and a data provider.
- STAPI aims to define an interoperable solution for viewing data opportunities and ordering data. STAC API would work together with STAPI for the discovery of data.
- Allows users to look for opportunities across providers, and automate the process.



- Element84 has hosted a number of Sprints to develop this concept, starting in 2022. At the end of the 2023 sprint, a working prototype was developed.
- Have worked to engage international partners as well, with the April 2024 sprint hosted in Berlin.
- Aim to involve both established data providers (for which APIs may exist, but interoperability is a problem), as well as start ups who haven't launched yet.
- The specification is available on GitHub.

- Three main entities in STAPI: product (description of data to be collected), opportunity (description of geospatial data that may be collected in the future) and order (similar to an opportunity, but data has been ordered).
- Product descriptions could be aligned with the CEOS-ARD specifications.
- The opportunities endpoint is optional, users can go straight to ordering.
- Constraints are limitations on the collection of the data, supplied to Opportunities and Order endpoints. Order parameters are options that apply to the delivery of the data, and are just supplied to Orders.
- Both Opportunities and Orders are GeoJSON, for compatibility with GIS software.
- Another sprint will be held in Lisbon, Portugal, in April 2025.

Discussion

- Sai Kalapana (ISRO) noted that ISRO provides future tasking services, and is developing an API. They will look at incorporating the STAPI specifications into their work. Can searching the archive with STAC and ordering future products with STAPI be combined?
- Matt noted Umbra set up their API to search both in their archives and for future opportunities. There are some issues with validation, so there is a bit more work to do on this topic.
- The team are currently focused on driving adoption, but hope that eventually it could become an OGC Community Standard like STAC.
- Yves Coene (ESA/Spacebel) recalled the OCG standard for Tasking / Feasibility interface for EO based on SPS (Sensor Planning Service). Matt confirmed the STAPI team were aware of this, but adoption appeared to be minimal.

1.11 - IDN Report

Michael Morahan (NASA/KBR) reported [*slides*]:

- Over 500 collections are provided through FedEO to IDN.
- JAXA granules are collected through JAXA's API, which will come back online shortly.
- ISRO, USGS, and NRSCC collections are provided through CWIC, which contains over 200 collections.
- GISTDA collections include three from THEOS, however they need to be reviewed and updated. GISTDA might also consider updating to a STAC API.

- Working with USGS EROS to add their STAC collections to IDN. This includes Level-1 and Level-2 Collection 2 products.
- Currently 31 INPE STAC collections connect to IDN, CMR and FedEO. By adding the STAC Collection URL endpoints to the Collection Metadata record, CEOS Agencies can identify the STAC collection-granule connection in the IDN metadata entry.
- WGISS Connected Data Assets Client Partner Guide has been added to GitHub as a Jupyter Notebook, to simplify the review and testing of OpenSearch APIs. This also helps to identify issues with the CMR (IDN) OpenSearch API.
- GCMD Keywords releases are being updated regularly. GCMD Keyword Forum lets GCMD metadata providers request new or updates to the GCMD Keywords and allows users to provide feedback and comments.

WGISS-59-06	Prayot Puangjaktha is the new nominated GISTDA representative for DAIG, and will review the GISTDA datasets in IDN. WGISS Secretariat to add Prayot to the DAIG mailing list.	**Due: ASAP**

2.7 - NASA ESDIS STAC Update

Doug Newman (NASA) reported [*slides*]:

- In 2024, NASA ESDIS implemented the collection search extension for CMR's STAC API.
- A virtual 'All' provided was added to CMR STAC to allow for searching across all providers.
- For CWIC, the INPE and USGS EROS collection records in CMR/IDN were updated to point to agency STAC APIs. Using CMR STAC a user can now discover collections at CMR/IDN and navigate to granules at both INPE and USGS EROS.
- The STAC browser will be modified to support collection search for collections that do not support granule search.
- Would like to add more STAC API enabled collections from other agencies, noting that ISRO has an API but no collections have been updated.
- Once the migration has been completed, OpenSearch implementation will be replaced with STAC.
- The CEOS STAC Collection & Granule discovery BP has been released as v1.0.1

Discussion

- Nitant Dube (ISRO, WGISS Vice-Chair) noted ISRO will work with Michael to update their collections.

2.6 - NASA Agency Report

Doug Newman reported [*slides*]:

- NISAR is scheduled to be launched in May 2025, and will generate a large amount of data. The total archive volume is currently 128.6 PB, and growing at an average of 147.3 TB/Day.
- The primary goal of cloud migration is to provide user access to data, adjacent to high performance compute.
- 67 PB are hosted in AWS S3, in US-WEST-2, accessed through S3 API or HTTPS. Discovery is through STAC and CMR APIs, or EarthData Search User Interface.
- Working on migrating the remaining 61 PB of on prem data, plus all new data acquisitions. Expect to have the migration complete by the end of 2026.
- Openscapes supports scientists using data from NASA Earthdata served from the DAACs as they migrate workflows to the cloud, easing the discovery, access and usage of NASA Earth Science data.
- Harmony is NASA ESDIS' cloud based open-source data transformation platform. It provides unified access to data products and services across various Earth science disciplines.
- Worldview is a visualization user interface for over 1000 global, full-resolution satellite imagery layers, allowing users to visualize and then download the underlying data.
- The Fire Information for Resource Management System (FIRMS) distributes Near Real-Time (NRT) active fire data as part of NASA's Land, Atmosphere Near real-time Capability for Earth observation (LANCE). Globally these data are available within 3 hours of satellite observation, and for the US and Canada active fire detections are available in real-time. During the January 2025 California Wildfires, the page received more visits than any other NASA webpage.
- To maintain services under a reduced budget, ESDIS is working on consolidating assets to reduce infrastructure and maintenance costs. This includes web unification and service de-duplication.

1.12 - FedEO Report

Yves Coene (ESA/Spacebel) reported [*slides*]:

- FedEO currently contains 234 million granules and 2698 collections. INPE collections have increased from 35 to 63 collections (through the STAC API), and ESA has increased from 38 to 40 million granules. CNES (REFLECS) will be reconnected to FedEO shortly.
- Working with JAXA on the Integration in FedEO of new JAXA collection search endpoint implementation, from the previous G-Portal catalogue. The new search endpoint will be implemented in April 2025.
- ESA/DLR wish to retain their separate portals (clients), so FedEO is currently experimenting to identify issues and limitations, and identify and group duplicate products.
- FedEO is working to connect to the CDSE STAC API, which will replace the previous integration in FedEO based on the legacy OpenSearch API. 17 collections are now available on FedEO, with data access via HTTPS or S3. Improvements are needed to "inherit" global and collection specific queryables as OpenSearch parameters.
- STAC WebMapLink Extension allows defining bindings with various visualization services. There is a prototype STAC implementation in ESA Catalogue / ESA WMTS Service serving quicklooks, and work is ongoing.
- The online STAC documentation for FedEO has been updated, with the Pystac_client collection-search examples added as a Jupyter Notebook.
- DAIG have been investigating existing STAC Validation Tools, to understand whether a CEOS tool is needed. Two examples are the stacindex.org tool, and Copernicus CDSE / EOF Conformance Test Suite. Three tools were investigated: STAC API Validator, STAC Validator and STAC Node Validator.
- A CEOS STAC Validator would ideally be an easy to use web interface, available from the CEOS Website and GitHub. It would combine checks from multiple libraries to cover STAC API and STAC objects. Some modifications would be needed in the libraries to simplify the compute for larger collections.
- An initial version to combine existing batch tools is in development. This could be extended later with additional tests for selected CEOS Best Practices. A first version is expected before WGISS-60.

Session 3: Interoperability: Policy Factor

3.1 - GEO Data Licensing Guidance

Nitant Dube (ISRO, WGISS Vice-Chair) reported [*slides*]:

- The GEO guidelines for Open Data Licensing were developed by the GEO Law and Policy Subgroup (LP-SG), and approved by the GEO Data Working Group.
- Legal interoperability is important to ensure data, tools and services can be aggregated by end users.
- Data providers often use the terms 'Full and Open Access', however this does not provide sufficient legal certainty. The meaning of this phrase could differ between organisations.
- End User Licence Agreements (EULAs) often include substantive restrictions, and require legal review.
- Treatment of value added products, including AI/ML and other cloud computing technologies, can complicate the legal issues.
- GEO recommends use of Standard Open Data License, for example Creative Commons Zero 1.0 Universal Public Domain Dedication (CCo), Open Data Commons Public Domain Dedication and License (PDDL) or Creative Commons Attribution 4.0 International (CC BY 4.0). Each has different attribution, copyright and use requirements.
- A recommendation has been included in the Interoperability Handbook for this topic.
- Should perhaps also include licensing for software (e.g. Apache 2.0) in these recommendations.

Discussion

- Mirko Albani (ESA) recognised that the GEO group that developed these guidelines may not have included space agencies. Mirko had the three recommended licenses reviewed by ESA's legal team a couple of years ago, who identified some drawbacks from the space agency perspective.
- Tom Sohre (USGS, WGISS Chair) noted that it will be challenging for all members to conduct a legal analysis. Suggested the recommendation be more general.

3.2 - Review of Policy Factor

Nitant Dube (ISRO, WGISS Vice-Chair) reviewed the Policy factor recommendations [*slides*]:

- Need to consider commercial providers as well as space agencies in this chapter. Commercial providers should be in the position to consider and apply the recommendations.
- The policy chapter currently has eight recommendations:
  - POL#1: Identify policies in your organization/country related to data and services and conduct periodic check/audit for compliance to these policies. Identify policies which may be barriers to interoperability of data and services and flag them for resolution.
    - Is it also important to ensure the user community knows the policies as well.
  - POL#2: CEOS MIM Database: Publish and periodically update Information about present and planned Earth observation Satellites in CEOS MIM Database .This will help in planning and overall coordination among different EO stakeholders.
    - The CEOS MIM database only includes CEOS Agency missions. How can this cater to non-CEOS organisations? This was generalised to encourage regular publishing of mission details in an online database.
    - It would be good to be able to have all capabilities in one central source. Machine readability of the CEOS MIM Database could help with this.
  - POL#3: Open Standards: Ensure your organizations implement open standards such as those published by the OGC for data and services in preference to drafting new specifications for data formats, metadata formats and service APIs.
    - STAC was developed as a 'guerilla standard' - by the community, not by OCG/ISO/IEEE.
    - CEOS-ARD is also not a formal standard.
  - POL#4: Open Data: Organizations should ensure that the non-commercial Earth observation data is proactively made freely available for use, reuse and redistribution to users in human and machine readable form.
  - POL#5 Open Source Software: Where possible, share software applications as open source software, enabling others to use the same tools as are used internally to process or transform data products or to demonstrate the use of standards to access your data and services. Jupyter Notebooks can be developed and provided

following CEOS Jupyter Notebooks Best Practices. Organisation should ensure their software is licensed with open source licenses, preferably Apache 2.0.

- Jupyter Notebooks is likely not relevant here.

- POL#6 Open Science: Promote the concept of open science for collaborative development. Open science ensures availability of the state-of-the-art algorithms and software providing consistent products from different data providers and supporting reproducibility.
  - Could review against open science statements from agencies.

WGISS-59-07	DIIG to compare the Open Science recommendation from the Policy Chapter of the Interoperability Handbook against the NASA Open Science policy, and other CEOS Agency Open Science policies.	**Due: WGISS-60**

- POL#7 Data Licensing: Organizations sharing open and unrestricted data should license the data using an open source license, consistent with Agency policies. A Custom license can restrict access for users. The GEO data licensing Guidance can be referenced for examples, which includes, Creative Commons Zero 1.0 Universal Public Domain Dedication (CCo), Open Data Commons Public Domain Dedication and License (PDDL) v1.0, or Creative Commons Attribution 4.0 International (CC BY 4.0). CC BY 4.0 is preferred.
- POL#8 Data Procurement from third party: Organizations planning to procure/outsource Earth Observation data, to possible extent should ensure that the data being procured are CEOS-ARD compliant
  - This is a tough recommendation to meet.

- All changes made during the discussion are recorded in this Pull Request in the GitHub repository.

Discussion

- Dave Borges (NASA, SEO) recognised that commercial data will generally not be free & open, however we need to work towards ensuring interoperability between their datasets and CEOS data regardless. Interoperability regardless of cost.

Session 4: Interoperability: Interface Factor

4.1 - Data Discovery

Damiano Guerrucci (ESA) reviewed the Data Discovery recommendations:

- DISC#1: Collection and granule discovery interfaces should comply with the CEOS STAC Collection and Granule Discovery Best Practices (preferred) or CEOS OpenSearch Best Practices.
- DISC#2: Service and tool discovery interfaces should comply with CEOS Service Discovery Best Practice.
- DISC#3: Collection and granule metadata obtained via the discovery interfaces should advertise the existence of the corresponding file-level online data access and subfile or pixel-based access services and endpoints (e.g. OGC WCS, WMTS, WCPS, OGC API Maps, OGC API Tiles, etc.).
- DISC#4: Granule metadata obtained via the discovery interfaces should include the online data access URL to the granule in native format and to a low resolution representation (i.e. quick look or thumbnail) in Web-friendly format, e.g. JPEG or PNG. The low-resolution representation may be a static file or an OGC WMS/WMTS or API Maps/Tiles response.
    - ‘Native format’ may be a bit ambiguous. ‘High resolution’ may be clearer.
- DISC#5: Collection and granule metadata obtained via the discovery interfaces should advertise the existence of the corresponding authentication endpoint for access to the data (if any).
    - Added the clarification for human & machine access.
- DISC#6: Resource metadata including keywords should link each keyword to its HTTP URI and to the appropriate thesaurus (i.e. controlled vocabularies).
    - This could be linked to the Vocabulary factor.
- DISC#7: Keywords from controlled vocabularies that allow lookup of keyword information via Linked Data principles, e.g. HTTP URI dereferencing or SPARQL interfaces are preferred. The use of GCMD controlled keywords is encouraged.
    - Haven’t talked about GCMD in the Vocabulary recommendations. Need to have an interface between the keywords and definition.
- DISC#8: Resource metadata shall include the Persistent Identifier (e.g. DOI) of the corresponding resource (if available). See also CEOS Persistent Identifier Best Practices.

    - Perhaps there should be something in the architecture factor to require that DOIs are included.
    - The term 'resource' here was included to cover collections, granules, etc.
- DISC#9: Collection metadata should refer to quality information about the collection, e.g. expressed as WGISS Data Management and Stewardship Maturity Matrix.
    - Not referring to any quality information in the archival section.
    - 'Quality information' is vague. The intent was focused on the matrix factors.
    - Shouldn't define here what quality information is. If there is a resource that can help in the discovery of data, it should be included in the metadata.
- DISC#10: Discovery interfaces should be accessible and return responses without requiring authentication.
- DISC#11: For facilitating discovery and access, data shall be organised in collections according to the principles outlined in the forthcoming WGISS Data Collections Management Practices White Paper.
    - Should be moved to the start of the list.
    - This is not directly related to interface, and is also covered in Architecture.
    - Need discovery interface for collections, and discovery interface for granules.
- All changes made during the discussion are recorded in this Pull Request in the GitHub repository.

4.2 - Data Access

Damiano Guerrucci (ESA) reviewed the Data Access recommendations:

- DACC#1: Granule data stored in the cloud should be accessible via the S3 (Simple Storage Service) and HTTP(S) protocols.
    - Agreed that this recommendation should be less prescriptive.
- DACC#2: Granule data stored in the cloud should preferably be accessible in cloud-optimized formats, e.g. Zarr or Cloud-Optimized GeoTIFF (COG).
- DACC#3: Data access should support file-level access and subfile or pixel-based access.
- DACC#4: Data download interfaces over HTTP should support Range Requests to allow clients to request a portion of a file. Typical use case: access to a portion of a Cloud-Optimized GeoTIFF (COG) file.

    - This was combined with #3.
- DACC#5: In case a granule consists of many individual assets (files), it shall be possible to access each asset individually or access all subcomponents of a granule with a single request.
- All changes made during the discussion are recorded in this Pull Request in the GitHub repository.

4.3 - Authentication and Authorization

Damiano Guerrucci (ESA) reviewed the Authentication and Authorization recommendations:

- AUTH#1: Authentication interfaces should support the OpenID Connect Protocol.
    - For machine to machine with S3, OpenID Connect isn't used.
- AUTH#2: HTTP requests (e.g. for data access) subject to authorization shall include the user token with claims in JWT format returned by the Authentication interface with every API request as a bearer token in the HTTP authorization header.
    - This is a bit technical. It is the mechanism used by Coperncius, but may be too specific on the implementation.
    - Have already said in DISC#5 regarding the corresponding authentication access point.
    - Lots of different authorisation methods. Is a bearer token the preferred method? Avoiding basic authentication.
- Added a new recommendation for file level access.
- The data should be presented in any format that allows pixel-level access.
- Data providers could document the API (interfaces) consumed by their users (Discovery, Data Access etc.) via executable Jupyter Notebooks. This can be discussed in the supporting tools and technologies chapter to be added to the handbook.
- The alignment of Jupyter Notebooks with the WGISS best practices on this topic should be also recommended in the supporting tools and technologies chapter.
- The Architecture section should recommend that interfaces are based on RESTful principles and prefer JSON payloads over XML payloads.
- All changes made during the discussion are recorded in this Pull Request in the GitHub repository.

Session 5: Interoperability: Vocabulary and Quality Factors

5.1 - Review of Vocabulary Recommendations

Peter Strobl (EC-JRC) reviewed the Vocabulary recommendations [*slides*]:

- Terminology in earth system sciences has been an issue for a while. Terminology work in CEOS started in WGCV, with participation from WGISS as well. The current interoperability work started soon after, following discussions around CEOS-ARD.
- The terminology group published a peer reviewed paper last year: *Lost in Translation: The Need for Common Vocabularies and an Interoperable Thesaurus in Earth Observation Sciences*
- CEOS is a good place to establish a formal thesaurus, consolidating glossaries from various publications and resources.
- A good thesaurus is consistent, interrelated, understandable, educational and updateable.
- The European Commission's KCEO has started implementing these concepts in a Git based glossary. The plan is to make this a joint CEOS-KCEO Glossary. Terms from the WGISS Glossary and the WGCV CalVal Portal Wiki have been imported.
- The chapter in the handbook is broken up into Semantics and Thesaurus recommendations.
- SEM#1: Initiate the development of a joint open Earth Observation thesaurus, as an immediate measure by migrating the merged WGCV/WGISS Terms and Definitions wiki to a GitHub repository by either joining it with the KCEO pilot glossary creating a CEOS/KCEO thesaurus or by at least maintaining it as a compatible branch on GitHub.
    - This is more directed at CEOS than data providers, and is written as an action.
    - It should be more general, and instruct data providers to use a common thesaurus.
- SEM#2: Invite and encourage all CEOS actors to participate in cleaning-up and harmonising the joint glossary, to achieve consistency and compliance with the rules (see below) as soon as possible. Provide capability enabling public comment and discussion and functionality for adding new terms.
    - This should also be more general, to encourage participation from outside CEOS.
- SEM#3: Enable version control and change management at the individual term level and link to historical and alternative definitions

- SEM#4: Abandon and ban the practice of entity, project, or document specific vocabularies e.g. in the form of 'terms and definitions' chapters. Source (via weblink), maintain, and develop all terms that serve or might serve in more than one context in the new online, shared repository
    - This is very strong language, and was softened to discourage the use of such glossaries.
- SEM#5: Seek formal collaboration with ISO/TC 211 (who may perhaps be asked to lead), OGC, the WMO task team on WIGOS metadata and any other stakeholder in Earth System Sciences interested to join and strive for domain wide adoption of a common thesaurus
- SEM#6: Set-up common online repositories for abbreviations and acronyms as well as e.g. agreed metadata fields with unified and binding lists of options.
- THES#1: The terms defined in the thesaurus vocabulary should be consistent and divided into classes such as Base, Core, Controversial and High Impact (tbc). The 'Base Terms' should have cross community agreement and should not have circular or ambiguous definitions. The 'Core Term' should be using the 'Base Term' consistently and can be allowed to have minor tweaks with approval from the identified committee. The 'Controversial Term' should have qualifiers attached to them with links to discussions, which led to the association of the qualifier. The 'High Impact Term' should be approved by a specialist committee and should be linked to a document providing details of the term.
- THES#2: The definition of a term may not contain the term itself nor other circular definitions (e.g., where term A is defined using term B and term B is defined using term A). The development of a clear set of base terms will make this easier to achieve.
- THES3#3: The terms used in the thesaurus should be interrelated with clear and mappable relationships between other terms (parent, sibling, child) avoiding, in particular, circular (child becomes parent) relations. Overlaps between terms that are supposed to delineate more generic concepts (siblings) should be avoided or minimized
- THES#4: Definitions have to be kept unambiguous and short, and written in a form such that they can replace the term in a sentence
- THES#5: Examples and explanations may only be given in 'Notes' complementing the definition not as part of the main definition.
    - It was clarified that 'Notes' and 'Examples' are seperate sections to accompany the definition.

- THES#6: Every definition shall list all source documents it is based on, wherever possible as weblinks
- THES#7: Thesaurus terms should be version controlled at the individual term level.
- THES#8: Where a term is deemed 'controversial', then contradictory definitions can be provided, but only with clear links to alternative definitions and explanations as to what context a term is used in.
- DISC#8: Keywords from controlled vocabularies that allow lookup of keyword information via Linked Data principles, e.g. HTTP URI dereferencing or SPARQL interfaces are preferred. The use of GCMD controlled keywords is encouraged.
    - The recommendation from Architecture was combined with SEM#6, but needs further review.
- All changes made during the discussion are recorded in this Pull Request in the GitHub repository.

Day 3: Wednesday 26th March, 2025

Session 5: Interoperability: Vocabulary and Quality Factors (cont.)

5.2 - Review of Calibration and Validation Recommendations

Cody Anderson (USGS, WGCV Chair) reviewed the Quality recommendations [*slides*]:

- WGCV have refined the definition for the Quality factor, deciding on: *Indicators (parameters, metrics, etc.) for informing users of the trustworthiness (accuracy, uncertainty, consistency, etc.) of the data provided (measurands, measurements, observations, etc.)*
- Some of the terms in the definition are not well defined (e.g. trustworthiness), but have been kept vague on purpose. Examples have been provided to explain the intent.
- Explaining the algorithms/processing done on the dataset is key for users to understand the quality of the data.
- Information exchange is critical for quality interoperability, such as through community cal/val groups (such as CEOS WGCV). This should be included as a specific recommendation in the handbook, and an overarching recommendation in the Policy factor.
- CALVAL#1: The Measurand and Uncertainty of stated values within products are key to communicating and understanding data quality.

- CALVAL#2: All products should have associated quality indicators, traceable to reference standards to allows users to access usability of the data for their applications
- CALVAL#3: CEOS ARD Framework should be used as a starting point for development of Interoperable and Analysis Ready Data.
    - There are currently no quality requirements in CEOS-ARD framework, however there is intent to extend CEOS-ARD to quality as well.
    - Need to consider how CEOS-ARD is included across the handbook.
- CALVAL#4: Post-launch, Level-1 products should be calibrated using CEOS Fiducial Reference Measurements (CEOS-FRM).
    - A reference should be provided for this.
- CALVAL#5: CEOS endorsed Cal/Val sites and reference network such as RadCalNet and SARCalNet should be used for satellite cross-comparison
- CALVAL#6: The Quality Assurance Framework for Earth Observation QA4EO developed by Group on Earth Observations (GEO) and endorsed by CEOS should be followed to enable interoperability and quality assessment of earth observation data
- CALVAL#7: The Joint Agency Commercial Imagery Evaluation (JACIE) Best Practices document can be used as a guideline for standard calibration and validation activities to be performed
- CALVAL#8: The ESA Earth Data Assessment Project (EDAP) process provides multiple reporting metrics related to quality
- CALVAL#9: CEOS CAL/VAL portal can be used as a reference site for accessing agreed good practices and CAL/VAL protocols for interoperability for Earth observation calibration and validation activities
- All changes made during the discussion are recorded in this Pull Request in the GitHub repository.

Discussion

- Nitant Dube (ISRO, WGISS Vice-Chair) asked whether lunar/solar calibration recommendations should be included. This is specific to optical imagers, and hence likely too detailed for these recommendations. It is also included in the reference documents.
- Something general on vicarious calibration could also be good.

WGISS-59-08	WGCV to consider adding a general recommendation for vicarious calibration in the Quality Factor of the Interoperability Handbook.	**Due: Q2 2025**

Session 7: Federated Authentication and Authorisation

7.1 - NASA Federated Authentication & Authorisation

Doug Newman (NASA) reported [*slides*]:

- NASA is embarking on a Holistic Identity Management project. Goals for data consumers is to reduce the barriers to accessing NASA data and services, while improving security.
- The backbone of NASA's architecture is Keycloak technology which provides single sign-on, identity brokering, social login and user federation.
- Applications can specify what identity providers are allowed for their application. For example, apps allowing access to sensitive information would allow NASA Launchpad multi-factor authentication only.
- Google and NASA Launchpad login is planned for roll-out in Q2 2025. In Q3 2025, additional social logins will be added (AWS Cognito and id.gov), followed by Facebook and Microsoft in Q1 2026.

Discussion

- Nitant Dube (ISRO, WGISS Vice-Chair) asked about tracking user behaviour when using social login.
- Doug recognised tracking users is essential to NESDIS operations. Following the user choosing to login with social networks, the system asks for additional information as per a regular NASA EarthData account.
- Machine-to-machine workflows will remain unchanged through EarthData logins and Vera tokens.
- Tom Sohre (USGS, WGISS Chair) asked whether NASA studied different social login providers. What were the pros and cons?
- Doug noted that a trade study was done, and can like be shared with WGISS. It remains unclear how social logins could be used to federate between agencies. If different agencies use e.g. Google, this could create a more seamless workflow for users, but may also need some integration at the service provider (e.g. Keycloak) level.

WGISS-59-09	Doug Newman to check if the trade study done by NASA for the various social login providers for their Federated Authentication and Authorisation work can be shared with WGISS.	**Due: WGISS-60**

- Alastair McKinstry (ICHEC) noted the use of eduroam and eduGAIN across the education sector (in particular in Europe). Is there potential for that to be added as a provider?
- Doug noted this is not currently on NASA's roadmap, but if WGISS feels that would be a good value add, and would enable federation, it could be considered in future.

Session 6: Interoperability: Architecture Factor

6.1 - Data Architecture (Archival and Preservation)

Mirko Albani (ESA) reviewed the Data Preservation recommendations [*slides*]:

- DPRES#1: Use a common glossary of terms and definitions applicable to data archiving.
  - This will be removed from this chapter, as it is covered by the vocabulary chapter.
- DPRES#2: Archival systems should comply with the Reference Model for an Open Archival Information System (OAIS) and with the forthcoming "OAIS-Interoperability Framework" to facilitate interoperability between archives.
- DPRES#3: Archived data should be appraised and properly documented before ingestion in the archives following the forthcoming CEOS Data Appraisal Procedure.
  - The documentation will be developed relatively quickly by DSIG.
  - Data Appraisal is something that should be done before the data is archived, but it is important to include the documentation in the archives.

WGISS-59-10	DSIG to document EO Data Collection Appraisal procedures.	**Due: WGISS-60**

- DPRES#4: Data and information ingestion into archives should follow internationally recognised standards and best practices (e.g. https://public.ccsds.org/Pubs/651x0m1.pdf) with documented tailoring derived from the generic activities described therein.
- DPRES#5: Data and Information archiving and preservation should follow internationally recognised standards and best practices (e.g. those produced by WGISS) with documented tailoring derived from the generic activities described therein.

- #4 is specific to the ingestion step, and #5 is more general. Ingestion could be included in the general statement in #5.
- Best practices are general, and providers need to document the specific process used.

- DPRES#6: Periodically perform archival systems upgrade to the most adequate proven technology to ensure data and information long term preservation. Ensure migration of archived data from old to new systems.
  - Media is generally considered part of the system. 'Most adequate proven technology' captures all types of media etc.
  - This is related to software preservation and maintenance. A new recommendation was added for software preservation.
- DPRES#7: An harmonised approach should be pursued for the development and use of archiving systems, interfaces and formats to improve compatibility of services provided by different organisations.
  - Done through information sharing in WGISS.
  - #9 covers the formats, etc., while #7 is a bit broader.
  - Agreed to remove for simplicity, noting general collaboration is covered also in the Policy section.
- DPRES#8: Maintain formal descriptions of data and information archiving formats.
  - Archive formats can sometimes be different to operational data.
  - Describing the format should include how to interpret the bytes within the file.
  - Agreed to remove this as it is mostly covered by #10.
- DPRES#9: Perform archived data and information repackaging and/or reformatting to comply with new standard formats and/or exchange formats. Properly document change made to the archived data.
  - Extended this to ensure the data isn't changed during reformatting.
- DPRES#10: Periodically verify the integrity of the archive collection/content through integrity check on a representative set of the archived data.
- DPRES#11: Ensure that the content of the archived data and associated information remains unchanged and, if changes are made, that these are documented and that this documentation is preserved and made available as well (provenance information).

    - Removed this, covered by #9.
- DPRES#12: Manage evolution of archived data collections according to the Shared Collection Lifecycle Management Principles for EO Data best practice.
- DPRES#13: Assign a Persistent Identifier to data archived and published to users and ensure the availability of all associated information in the relevant Landing Page following the CEOS Persistent Identifiers Best Practice.
    - DOIs are assigned to data when it is ready for dissemination, but the same data is also archived.
    - This may belong elsewhere as it deals with access, disseminating, archiving, etc.
- DPRES#14: Maintain at least two copies of at least the core data products, in two different geographic locations.
    - This doesn't deal with interoperability directly. Important for archives, but not necessary for interoperability.
    - Agreed to remove.
- DPRES#15: Keep archives equipment (hardware and software) up to date and in conformance with vendor recommendations.
    - Not be specific for interoperability.
    - OAIS Interoperability Framework shows use cases of archive interoperability. If archives are interoperable, an organisation can rely on others to keep the necessary copies.
    - Would need to make sure different archives are both up to date to ensure interoperability.
    - There are also differing perspectives on what an 'archive' means. In some cases, the 'archive' is not publicly accessible.
    - Rephrased such that two organisations wishing to have interoperable archives need to have compatible hardware and software.
- All changes made during the discussion are recorded in this Pull Request in the GitHub repository.

6.2 - Data and Metadata Architecture

Alex Leith (SEO) reviewed the Data and Metadata Architecture recommendations:

- The Architecture section is currently broken into six subchapters:
  - Data Architecture
  - Data Preservation
  - Metadata Architecture
  - Cloud Architecture
  - Publishing Architecture
  - Technologies
- Question around where CEOS-ARD fits within. Perhaps Quality, but CEOS-ARD has no quality requirements.
- Would CEOS-ARD be referenceable in the metadata architecture section? CEOS-ARD describes a process to get to a product - it is more than the metadata.
- There might be other elements regarding data formats to include in Data Architecture.
- It was agreed to merge the metadata and data recommendations.
- Agreed early on in developing the handbook that WGISS should not be emphasising technologies, as they evolve and change. Agreed that a separate technologies section could be included, where current technologies are discussed (but not necessarily recommended).
- DATA#1: CEOS-ARD Framework should be used as a starting point for development of Analysis Ready Data.
- DATA#2: CEOS-ARD Product Family Specifications (PFS) should be used for development and assessment of ARD products, including both self-assessments and peer review.
  - Combined with previous DATA#3.
- DATA#4: The ISO 19115 series of standards should be used to produce geospatial metadata at the collection level.
  - What is meant by collection? USGS' understanding that ISO 19115 would be at the product level.
  - Removed the last phrase for generality.

- DATA#5: A Collection of data should have all granules packaged consistently and produced with consistent quality.
    - In the interface section, only collection and granule are discussed. Should ensure the entire document is referring to collections/products/granules consistently.
- DATA#6: Collection-specific metadata formats may be used, but packaging must include STAC documents at the Collection and Granule/Item level.
    - This is very specific, but, in Alex's opinion, STAC is the most important technology for interoperability. A reference to the STAC Best Practices was added.
- DATA#7: Hashes for all files in a packaged granule should be available, so that integrity checks can be run.
    - Changed to checksums for generalisation.
- DATA#8: Pixel-level metadata such as scene quality masks should be clearly documented with a reference to lookup tables.
    - Pixel-level metadata is usually only provided with level 2 and higher products, and isn't necessarily increasing interoperability.
    - Sort of covered in the quality factor. But the key here is lookup tables - there should be some standards for lookup tables.
- DATA#9: File names and folder or path structures should be consistent and include appropriate information to distinguish the specific granule.
    - Maybe too specific, but was trying to capture how files are named.
    - Versioning should also be considered.
- DATA#10: Assign a Persistent Identifier to data archived and published to users and ensure the availability of all associated information in the relevant Landing Page following the CEOS Persistent Identifiers Best Practice.
- All changes made during the discussion are recorded in this Pull Request in the GitHub repository.

6.3 - Cloud Architecture

Alex Leith (SEO) reviewed the Cloud Architecture recommendations:

- CLOUD#1: The data to be shared through cloud should be converted to cloud optimized formats for faster and interoperable access across multiple applications
    - Comes back to the range requests in the Data Access section of Interface.

    - Replaced with the text from DACC#2: Granule data stored in the cloud should preferably be accessible in cloud-optimized formats, e.g. Zarr or Cloud-Optimized GeoTIFF (COG).
- CLOUD#2: Granules must not be zipped when shared via the cloud, so that cloud optimized data formats can be leveraged.
    - Have requirements for some missions to be able to download multiple granules, which are zipped on the fly. Changed to stored on the cloud, rather than shared.
    - Efficiency of dissemination should also be considered but outside the scope of this document.
- CLOUD#3: Where possible, cloud providers' standard as-a-service offerings should be used in preference to self-developed solutions, enabling interoperability of tools that work with that cloud provider.
    - For example, using AWS tools means others can use it too, rather than a custom solution.
- CLOUD#5: Intermediation of a cloud provider's object store is discouraged, meaning that if possible, direct access using the service's APIs should be available.
    - Agreed to removed, as it is covered by #3.
- CLOUD#6 When possible, internal tooling and web services that are not a security of privacy risk should be made open source and openly accessible
    - This is covered in the Policy factor.
- Agreed to move the Cloud Architecture recommendations to Publishing, as the cloud is one type of publishing.
- All changes made during the discussion are recorded in this Pull Request in the GitHub repository.

6.4 - Publishing Architecture/Analytics Architecture

Alex Leith (SEO) reviewed the Publishing Architecture recommendations:

- PUBLISH#1: Granules should be available immediately and not through an on-demand service.
    - There are cases for data visualisation where on-demand can impact interoperability, however higher level products are often produced on-demand.
    - This was removed.

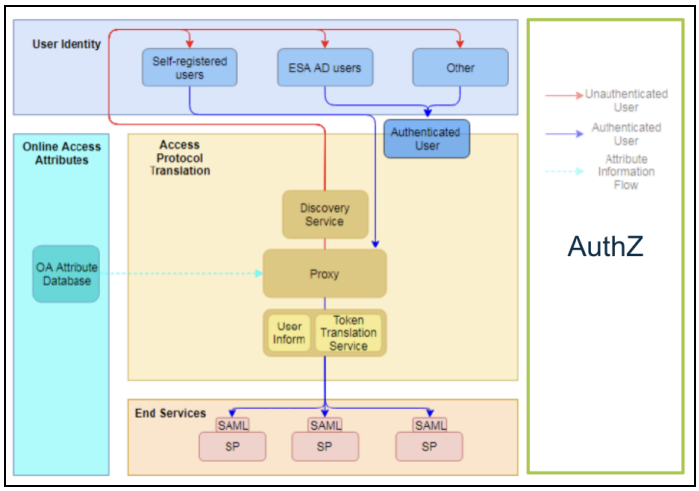
- PUBLISH#2: Each data collection that is published as a publicly-accessible product should include a public granule-level notification including for when it is added, updated or deleted/archived. This should be able to be filtered by location.
    - This is important for replicas of collections.
    - The filtering by location was removed.
- PUBLISH#3: A collection should have a full listing of all available granules in a standard format, preferably cloud optimized. STAC-geoparquet is used by some providers currently
- PUBLISH#4: Data licensing should be clear and adhere to national policies, with a strong preference for standard open licenses such as Creative Commons
    - This is covered by the Policy factor, and was hence removed.
- All changes made during the discussion are recorded in this Pull Request in the GitHub repository.

Session 7: Federated Authentication and Authorisation (cont.)

7.2 - Federated AuthN/AuthZ Use Cases

Sofia Marzo (ESA) and Saskia Brose (ESA) reported [*slides*]:

- ESA deals with various mission requirements and user access needs, which has led to a decentralised and fragmented authorisation framework. ESA would like to federate the systems for simplification within ESA, but also for interoperability between agencies.
- IAM (Identity and Access Management) refers to the framework of policies, processes, and technologies used to manage digital identities and control access to resources within an organization.
- An identity provider (IdP) is a system that creates, maintains, and manages identity information while providing authentication services to applications. A service provider (SP) is a system entity that receives and accepts authentication assertions from the IdP.
- OpenID Connect (OIDC) is used for authentication, and OAuth 2.0 is for authorization. By combining OAuth 2.0 and OIDC, apps get both authentication and authorization in a single, secure flow.
- ESA has designed the EO-IAM system, for Authorisation and Authentication.



- The ESA-NASA MAAP Project provides joint access to data, via either NASA EarthData or ESA logins. There are two instances of MAAP (one NASA and another for ESA), which can be cross-accessed.
- Destination Earth (DestinE) initiative of the European Commission has a federated identity provider, and a federated service.
- eduGAIN is a service that interconnects identity federations around the world, simplifying access to content, services, and resources for the global research and education community.

Discussion

- Yousuke Ikehata (JAXA) suggested best practices for using federated authentication could be considered, instead of a white paper. They could be based on the experiences of ESA and other agencies in implementing OpenID Connect.
- Damiano Guerrucci (ESA) recommended that MAAP be presented as a use case in the white paper, as there is still much more to understand regarding the implementation of federated authentication and authorization. Likely not at the right point to develop best practices yet.
- Sofia noted eduGAIN is mainly for research institutions, which does include ESA. In principle, it was targeted at universities.

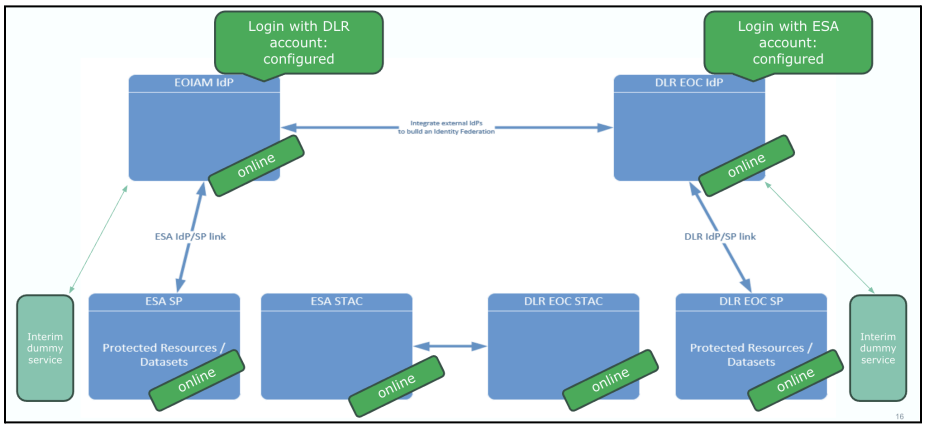
7.3 - Bilateral ESA-DLR Federation Use Cases

Mario Winkler (DLR), Sofia Marzo (ESA) and Iolanda Maggio (ESA/Starion) reported [*slides*]:

- The objective is for ESA and DLR to improve the federation of institutional EO data repositories hosted by each agency. A small demonstrator was developed to discover, download, process, manage EO data across repositories.
- Use cases consider both users and data managers.

Use Cases	Description	Status
UC1.1	A user can access a client and discover (free and open) data from online repositories hosted by different organizations and can download a granule from a free and open data set. ------ Consider Federation and Discovery	Ongoing
UC1.2	A user can access a client and discover (free and open) data from online repositories hosted by different organizations and can download a granule from a free and open data set. ------ Consider Authorization	Ongoing
UC2	A user can access a client and discover (restrained) data from online repositories hosted by different organizations and can ask for authorisation to download restrained data and can download a granule from a restrained data set.	Ongoing
UC3	A user can access a client and discover (free and open or restrained) data located in near-line/offline repositories (archived data) hosted by different organizations and can download a granule (--> deferred access).	Not started
UC4.1	A data manager can access a client and obtain information about data hosted in online, near-line, and offline repositories of different organizations and decide whether to remove data in own repository or whether to copy data across from external repository. ----- Consider the Near line and online repository	Not started
UC4.2	A data manager can access a client and obtain information about data hosted in online, near-line, and offline repositories of different organizations and decide whether to remove data in own repository or whether to copy data across from external repository. ----- Consider the offline repository	Ongoing
UC5	A user can access a collaborative platform and discover, visualize, and process data hosted by different repositories (online, near-line, offline).	Not started
UC6	A mechanism to exchange statistics about usage of data by users of another organisation shall be defined, together with a harmonised identification of metrics from the different sources.	Not started

- For Use Case 1.1, multi-mission STAC catalogs need to be accessible from external network collections. The multi-organisation STAC catalog includes both ESA and DLR collections, and can be accessed from either ESA or DLR side. Preliminary results are already available, but some technical issues remain to be resolved.
- For Use Case 2, a federated user profile will contain a number of attributes, including username, email address, name, organization/institution, and country of residence. Profile enrichment solutions can be applied, such that local identity provider's user profiles might contain more attributes.



- An authorisation policy will need to be created jointly between DLR and ESA to agree which features should be granted to which roles.
- Use Case 4 looks at two tasks: gap filling and data exchange, and content inventory verification. Two tests have been successfully completed, and the prototype will be available on GitHub.
- The received feedback from these use cases will provide inputs to the OAIS-Interoperability Framework specifications.

Discussion

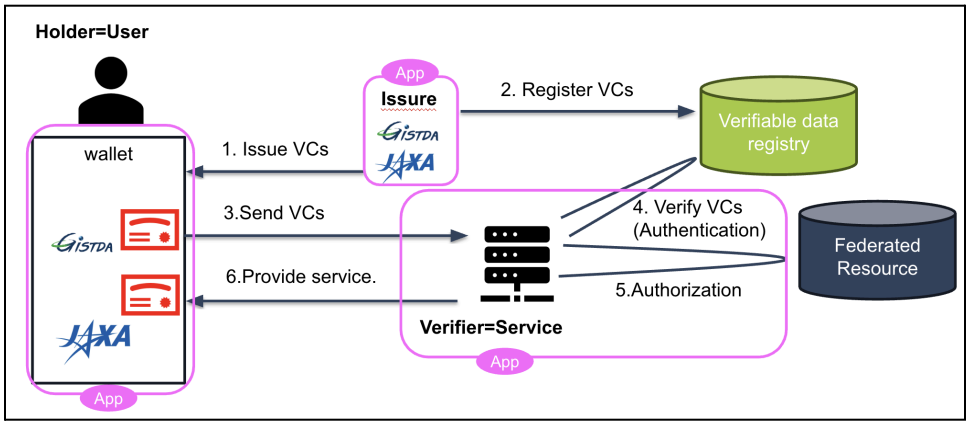
- Iolanda noted that dummy data was created to test the search for missing data, with specific data placed to be found. A matching exercise could be done in future.
- Tom Sohre (USGS, WGISS Chair) asked about the policy side of federation of two organisations, noting there must be some level of trust for another agency to authenticate on your behalf.
- Sofia recognised ESA and DLR haven't gone through the formal process yet, but it is one of the goals. This was just a technical demonstration, but to go into operations, a formal agreement would need to be put into place. There will also need to be agreements and policies around the sharing of personal information.

7.4 - SSI - Decentralised Authentication / JAXA

Yousuke Ikehata (JAXA) reported [*slides*]:

- Self-Sovereign Identity (SSI) is an approach to digital identity that gives individuals control over the information they use to prove who they are to websites, services, and applications across the web.
- For Lightweight Directory Access Protocol (LDAP) authentication, usernames and passwords are stored in a local database. OpenID Connect is a more modern authentication method, using existing platform authentication (e.g. Google, Facebook).
- Verifiable Credentials (VCs) are digital credentials which follow the relevant World Wide Web Consortium open standards. They can represent information found in physical credentials, such as a passport or license, as well as new things that have no physical equivalent, such as ownership of a bank account. They have numerous advantages over physical credentials, most notably that they're digitally signed, which makes them tamper-resistant and instantaneously verifiable.

- A Decentralized IDentifier (DID) is a type of globally unique identifier that enables an entity to be identified in a manner that is verifiable, persistent (as long as the DID controller desires), and does not require the use of a centralized registry.
- Verifiable Credentials (VC) can be provided as a json document with an electronic signature. In SSI, the user sends their VCs to the web service, which are verified by DIDs.



- SSI can enhance privacy by blockchain and encryption, and decrease fraudulent identity. However, they aren't commonly used, and must be used through a dedicated application on a smartphone. Standardisation of VCs is also needed.

7.5 - Open Discussion

- DAIG and TEIG will be starting a joint white paper on this topic, using inputs from WGISS-58 and -59 as a starting point. Best practices may also be considered in the future.
- There are some existing open actions to understand the feasibility at the CEOS level.
- Alastair McKinstry (ICHEC) found the level of detail presented today useful. The white paper should make sure to be clear on definitions to avoid any confusion.
- Damiano Guerrucci (ESA) recognised the importance of the policy aspects, and the need to effectively communicate to agency management the benefits, but also the risks. The procedure to achieve federation, beyond the technical, should also be discussed in the white paper.
- Tom Sohre (USGS, WGISS Chair) noted that federated data has also been discussed previously. Will this be covered in the white paper? Damiano noted data federation is a separate topic and should be discussed after authentication and authorization.

WGISS-59-11	DAIG and TEIG to consider whether a separate white paper on Data Federation is needed, alongside the current work on User Federation.	**Due: WGISS-60**

- Nitant Dube (ISRO, WGISS Vice-Chair) recognised the Copernicus Regional Data Hubs could be another use case for user federation.

WGISS-59-12	DAIG to discuss with the Copernicus Regional Data Hubs regarding their approach to federated authentication and authorisation, as well as data federation. Invite to present at WGISS-60 if appropriate.	**Due: WGISS-60**

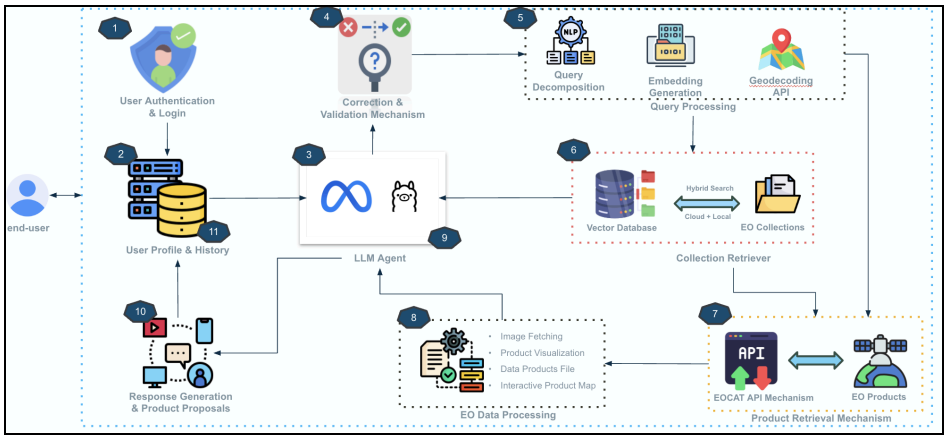
- Libby Rose (WGISS Secretariat) suggested the target audience for the white paper could be policy and management within agencies. The white paper should aim to communicate the benefits and risks of federation.
- Damiano noted the need to share data between agencies to cope with data volumes. There are also issues with multiple copies of the same data, and communicating to users the differences between copies.
- Alastair volunteered to be involved in testing and use case development for federated authentication and authorisation.

Session 8: Technology Exploration

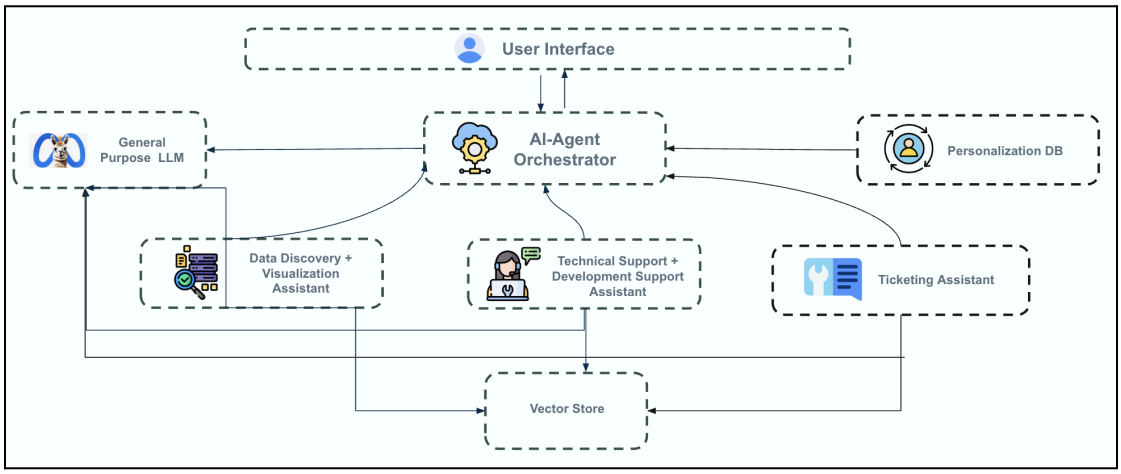
8.1 - EO Advanced AI Assistant

Hayret Abdula Keary (ESA) reported [*slides*]:

- ESA Common Services provides data access for 80 ESA and Third Party Missions. Undertook two demonstrator projects to take advantage of Large Language Models (LLMs) and other Artificial Intelligence (AI) techniques.
- The EO AI Technical Assistant uses summarised information from 3.7K web pages and 4.7K documents, and data from the TellUs ticketing system (52K tickets) to train a LLM which can support user community queries regarding ESA data.
- A thorough data anonymization procedure was performed using LLMs combined with scripting methods. A sentiment analysis and classification of TellUs data was also performed, using the Roberta model. A GUI was developed on Streamlit.
- The EO AI Data Discovery Assistant aims to facilitate the discovery of Earth Observation data products and collections, and simplify the process for finding optimal products.



- Used a pre-trained Roberta model to correct and validify the query such that it could be passed to the API.
- The tool provides the top 100 products which would be relevant to the user's query.
- The EO AI Technical Assistant and EO AI Data Discovery Assistant are being combined into a single EO AI Advanced Assistant.



- The AI-Agent Orchestrator ensures query management and routing.
- The Ticketing Assistant would be a new element, aiming to automate the ticket population for technical support and restrained data access. It will be integrated with the ESA TellUs system and EO IAM (Identity and Access Management) to provide personalised recommendations based on past activity.
- In the transition to operations, there are a number things to consider, including scalability by supporting an increasing number of parallel users, system and component-based monitoring, integration with external operational systems, benchmarking, and compliance with Data Protection and Security rules.

Trustworthiness must also be closely considered, to ensure the information provided is consistently accurate.

- The team is hoping to have a beta version of the EO AI Advanced Assistant ready by September 2025.

Discussion

- The project uses the llama 3.1 LLM. RAG (Retrieval-Augmented Generation) is used for retrieving information from the vector database. ESA's PhiLab is developing a version of llama fine-tuned for EO, which will hopefully be made publicly available by June 2025.
- Yousuke Ikehata (JAXA) requested that this example be included in the AI/ML white paper.

WGISS-59-13	Hayret Abdula Keary (ESA) to write up their EO AI Advanced Assistant as a use case for LLM in the AI/ML White Paper.	**Due: Q2 2025**

- Hayret added that the project uses resources within ESRIN, and hence does not have to rely on external providers which is helping with the cost. The biggest cost was in relation to purchasing the papers that the models were trained on.

8.2 - Jupyter Notebook Best Practices

Esther Conway (SFTC/UK) / Yousuke Ikehata (JAXA) reported [*slides*]:

- The Jupyter Notebooks work came out of a collaboration with WGCapD.
- Held a Jupyter Notebooks for Capacity Development Webinar, reaching about 500 people in two different sessions.
- EUMETSAT's guiding principles were a key input document for the Best Practices.
- Throughout developing the best practices, the understanding of dependencies changed. A lot of different examples are included in the document.
- Citation and Access is important, with not just DOIs but location and structure for third party distributors.
- Open Source Software licensing is also key. A common archive needs to allow for different levels of licensing. Creative commons licences are recommended.
- There are also differing perspectives around preservation of notebooks. A two tier archive, with active and legacy notebooks, could be considered, or a positive pressure / automatic retirement model.

- Have distributed the best practices to UK government bodies, and the recommendations appear to have held up across domains.

Discussion

- Alex Leith (SEO) noted that for licensing, most repositories of notebooks use software licences e.g. Apache 2.0, rather than Creative Commons, which is usually used for text. However, Jupyter Notebooks do straddle the line between code and text.
- Esther noted this recommendation came from NASA, as training notebooks was the key focus. In most of the training notebooks, the code isn't unique or outstanding, and the text is the unique element. Using creative commons can simplify the process, as agencies can have complex requirements around software licensing.

Day 4: Thursday 27th March, 2025

Session 9: Collections Management

9.1 - NASA EarthData Cloud Archive Recommendations

Doug Newman (NASA) reported [*slides*]:

- With traditional archives, there are a number of risks which need to be mitigated, such as accidental deletion/corruption, malicious deletion/corruption, hardware viability, hardware cost and natural disasters. Natural disasters are increasing in their frequency, and hence have an increasing likelihood of impacting data archives.
- NASA ESDIS has traditionally adopted the 3-2-1 approach to mitigate these risks: 3 copies of all data, 2 different media types, 1 copy at a different physical location.
- For cloud archives, media types are not under the agency's control, and physical distance between copies cannot be controlled.
- All three major providers that ESDIS deals with (AWS, Azure, Google Cloud) have a durability of at least eleven 9s. However, if data is lost, vendors only provide a refund and the data cannot be recovered. To date, ESDIS hasn't had any problems with loss of data in AWS S3.
- When moving data to the cloud, risks associated with hardware viability and cost are replaced by cloud vendor viability (such as going out of business, change in services), and cloud vendor cost.

- The growth of the archive will exponentially increase the cost. ESDIS will not be able to afford to have multiple copies of the 140PB of data that NISAR will generate over its lifespan.
- ESDIS has observed a 66% cost saving on their cloud archive by implementing S3 Intelligent Tiering.
- Using cold storage can reduce storage costs, which mitigates some of the cost risk but adds risk to data access.
- Can also reduce costs with S3 versioning, that is when data deletion is attempted, it isn't deleted immediately, but rather marked as deleted for a period of time, after which it is actually deleted. This could mitigate the accidental/malicious deletion risk.
- ESDIS is currently designing and testing their solutions, which involve two tiers of data:
    - Tier 1: Data which does not exist somewhere else, and cannot be derived from other data.
    - Tier 2: Data which exists at another agency or entity, or can be reconstructed from data in the Tier 1 category. Data is 'low use' - i.e. hasn't been distributed in 5 years.
- The following strategies are subject to change:

Tier 1 strategy

Risk	Mitigation
Disasters	multi-AZ cloud storage* 'S3 … redundantly store objects on multiple devices across a minimum of 3 Availability Zones in an AWS Region'
Cloud Vendor viability	Two cloud copies with different vendors
Cloud Vendor cost	Two cloud copies with different vendors
Accidental deletion/corruption	S3 versioning + MFA deletion
Malicious deletion/corruption	S3 versioning + MFA deletion

Tier 2 strategy

Risk	Mitigation
Disasters	multi-AZ cloud storage* 'S3 … redundantly store objects on multiple devices across a minimum of 3 Availability Zones in an AWS Region'
Cloud Vendor viability	Retrieval/reconstruction of data
Cloud Vendor cost	Retrieval/reconstruction of data
Accidental deletion/corruption	S3 versioning + MFA deletion
Malicious deletion/corruption	S3 versioning + MFA deletion

Discussion

- Nitant Dube (ISRO, WGISS Vice-Chair) asked whether data is also archived within the national archives.
- Doug noted that only data from completed missions, where the data was no longer needed by NASA, was transferred to the national archives. However, they only store data for a certain amount of time and old data is now being shipped back to NASA.
- Tom Sohre (USGS, WGISS Chair) added that, at USGS, it depends on the record schedule for the data. The national archives are also struggling with the volume of data, so USGS have been storing data for longer.

- Doug recognised the process of migrating data is still ongoing, and is scheduled to be completed by 2026. Currently have 67 PB in the archive, and 61PB still on premises. All new data (147 TB per day) is going straight to the cloud. This includes all data that is public, including some Level-0 but not all.
- Mirko Albani (ESA) noted that ESA keeps all Level-0 data, as it is the most important copy for potential future reprocessing. However, with all the levels of data kept, around 9 copies of the dataset were being stored. ESA is trying to reduce the number of copies, and will discuss the two-tier approach presented by NASA.
- Nitant Dube (ISRO, WGISS Vice-Chair) recognised that making copies of data isn't the only way to mitigate a risk.

Session 8: Technology Exploration (cont.)

8.3 - Session Objectives

Yousuke Ikehata (JAXA) reported [*slides*]:

- Maral Bayaraa (UKSA) has joined Yousuke as co-lead of Technology Exploration Interest Group (TEIG).
- TEIG is working on two ongoing activities, including artificial intelligence and cloud computing, with two new topics to be explored: digital twins and cloud compression.

8.4 - AI/ML White Paper: Discussion & Next Steps

Yousuke Ikehata (JAXA) reported [*slides*]:

- ESIP defined an 'AI/ML Ready Data' checklist, which involves four categories: Data Preparation, Data Quality, Data Documentation and Data Access.
- The ESIP checklist has been compared with the CEOS-ARD specifications:
  AI/ML ready data VS ARD
- ESIP's checklist includes AI oriented requirements, e.g. have targets been identified and labeled, have measures been taken to reduce bias. This is not covered in CEOS-ARD.
- CEOS-ARD only defines specifics for data, while ESIP's checklist also asks for additional aspects such as data access and data documentation.
- If CEOS-ARD certified data has an appropriate landing page (e.g. DOI), the data could be considered 'AI/ML' ready.

- There are some additional questions posed by ESIP which may need to be clarified due to ambiguity of terms:
  - Is this essentially raw data or a derived / processed data product?
  - Is this observational data or simulation / model output?
  - Has it been downsampled to reduce resolution or is it raw? If so, are the raw data available?
- This topic will be discussed at LSI-VC-17 in April, 2025.
- To gather an understanding of these requirements, TEIG has shared a survey with WGISS members. More responses to the survey are needed.

Discussion

- Tom Sohre (USGS, WGISS Chair) asked for clarification on the survey's purpose.
- Alex Leith (SEO) recognised that marine domains often use very heterogeneous data, and probably the datatypes ESIP was targeting with the checklist. CEOS-ARD specifications can help ensure AI/ML ready specifications for EO datasets.
- Nitant Dube (ISRO, WGISS Vice-Chair) noted that for training, ground truth data is also needed for some models. Does this need to be provided alongside EO data for it to be considered 'AI/ML ready'?
- Alastair McKinstry (ICHEC) recognised that AI/ML ready EO data should specify whether the dataset has enough information included to not need expertise in EO data. This is particularly important for the metadata.
- Alex recalled the Clay project (https://madewithclay.org/) extracted chips of Sentinel-1, Sentinel-2 and elevation datasets to build a large earth model.
- Alastair recognised the documentation needs to be prepared to cater for the needs of an AI training expert. That is the goal of the AIREO project - preparing datasets for non-EO experts. The CEOS-ARD specifications are mostly already good for AI training data, however the main shortfalls are around labelling of data and training datasets for object classification.
- Alex participated in an AI/ML workshop at Radiant Earth in 2020. The report produced would likely provide good context.

WGISS-59-14	Alex Leith to check with Jed Sundwall to get a copy of the report from the 2020 AI/ML EO workshop, for input into the AI/ML White Paper.	**Due: ASAP**

- Tom asked for clarity around the scope of this task. Should we develop specifications for what we consider to be 'AI/ML Ready Data'? Or just explore the topic? It is important to consider resources available.

WGISS-59-15	TEIG to document the scope of the AI/ML ready data definition task, and who we would be doing it for.	**Due: Q2 2025**

8.5 - Digital Twins White Paper

Yousuke Ikehata (JAXA) reported [*slides*]:

- A number of CEOS Agencies and other organisations are creating digital twins using EO data. TEIG is starting an activity to document and discuss these activities, and the approaches used.
- A number of presentations have been made at previous WGISS meetings on this topic.
- The key differences between a true digital twin and any other digital model are in its ability to update itself automatically to the conditions of the physical twin, as the physical twin changes and vice versa.
- Data from EO and other sources are integrated together to generate the necessary insights for updating the state of the digital twin.
- The real value of digital twins is in its ability to forecast and simulate alternative future scenarios, so that recommendations for decision making can be derived. This is a significant task and may be feasible only with the promise of data-driven approaches, such as Deep Learning, at its core.

Discussion

- Tom Sohre (USGS, WGISS Chair) asked for clarity on the scope of the activity. Will recommendations be provided?

WGISS-59-16	TEIG to document the scope of the Digital Twins white paper.	**Due: ASAP**

- Yousuke noted the primary goal is information sharing, and a showcase of capabilities. The first step in the white paper will be to define what a Digital Twin is.
- Matt Paget (CSIRO) has some experience with digital twins, noting the temporal element to support forecasting and simulation is the key. EO data must be used alongside other data sources to address specific questions.

- Alastair McKinstry (ICHEC) agreed that the temporal element is a key feature. Digital Twins include a wide range of scales, from full Earth models to smaller geographical scales with small features. The data must be prepared appropriately such that it can be continually updated in the model.
- Nitant Dube (ISRO, WGISS Vice-Chair) noted it would be good to understand the technologies required to build Digital Twins, including examples.

WGISS-59-17	TEIG to include Matt Paget and Alastair McKinstry in the Digital Twins white paper team. GISTDA and ESA to nominate someone to join the team.	**Due: ASAP**

Session 9: Collections Management

9.2 - Session Introduction & Status of Collections Management Work

Mirko Albani (ESA) reported [*slides*]:

- This activity aims to address some challenges with data collection management, including integrity, authenticity, replica management, reproducibility, citation and archive policy.
- Have shared experiences and lessons learned on these challenges at previous WGISS meetings, and a white paper has been compiled.

9.3 - Collection Management and Governance White Paper

Iolanda Maggio (ESA/Starion) reported [*slides*]:

- The purpose of this document is to describe a set of core Earth observation (EO) Collection lifecycle management principles shared by CEOS Agencies.
- Well-managed data collections should provide consistent data accessibility, usability, and long-term preservation, and follow FAIR (Findable, Accessible, Interoperable, and Reproducible) scientific practices. Data collections should facilitate the reanalysis of satellite measurements.
- The activity started with the Data Stewardship Reference Model defined by WGISS.
- The Data Collection Stewardship process starts during the initialization phase and continues until the Operations and Maintenance phase. For each new data reprocessing the process needs to be restarted. A number of principles for each stage were defined.

- In the initialisation stage:
    1. Products should be consistent within Collections
    2. Collection Upgrades should be deliberate, meaningful and well considered
    3. At conception of a Collection Upgrade, cost storage, computation requirements, open-sourcing of production code and end-of-life preservation costs should be considered
    4. Consider opportunities for interoperability and alignment with products in other organization's collections
- In the definition stage:
    5. Provisional products are part of the development process and their management and control are expected to differ from those of standard products within a collection
- In the implementation stage:
    6. a) Ensure Collection infrastructure providers enable disaster recovery, b) Ensure Collection data integrity
    7. Confirm archive or disposal at end-of-life of an old Collection at publication stage of the new one
    8. Do not delete or modify Collections without appropriate review, authorisation, and end user communication
    9. Assignment of a Persistent Identifier at collection level
    10. Maintain accessible records of product files that have been removed or replaced within a collection
    11. Consider approaches that authenticate replicas of collections
- In the operations & maintenance stage:
    12. Maintain accessible metadata and documentation for decommissioned Collections.
    13. Communicate early and as required with stakeholders on pending decommission to the Collection
    14. Source, ancillary, custom auxiliary data, algorithms and production software, and associated documentation should be retained indefinitely
    15. Obsolete collections should be disposed of in line with relevant entity and government regulations, policies or procedures
    16. Only one version of a baseline data product should be discoverable within a collection

- The final version was distributed to WGISS Exec for review, and no further comments were received.

Discussion

- Yves Coene (ESA/Spacebel) asked about the definition of 'products', noting WGISS should be consistent across various documentation.
- Iolanda noted these terms are defined in the document's glossary, but agreed consistency across WGISS documentation is important.
- Tom Sohre (USGS, WGISS Chair) recognised these terms and definitions should be included in the CEOS Common Dictionary exercise.
- It was agreed to approve the document, following a brief review of the glossary.

WGISS-59-18	DSIG to check the glossary from the Collection Management and Governance White Paper with other WGISS documents. Once complete, send to Peter Strobl to ensure the definitions and terms are included in the CEOS Common Dictionary.	**Due: ASAP**
WGISS-59-19	The Collection Management and Governance White Paper was approved. Following the completion of WGISS-59-18, WGISS Secretariat to publish the document online and distribute to WGISS members.	**Due: ASAP**

9.4 - EO Collection Data Citation Guidelines

Iolanda Maggio (ESA/Starion) reported [*slides*]:

- This is a new activity as a follow on for the Collection Management and Governance White Paper.
- Invite all WGISS members to share their methods for citing data, to produce general guidelines.
- Data citation refers to how to reference data, in the same way researchers routinely provide a bibliographic reference to other scholarly resources. Scientific papers should contain information about the data used in its creation for reproducibility, or to note the creator and draw some conclusions about the dataset's reliability.
- Data should be cited for data attribution, data findability and access, data reuse, transparency, reproducibility, data recognition, data impact and data sharing.

- Citations can help data providers know how much their data is used for scientific research, and derive insights about how it is used.
- A data citation must include the core components of author, title, year of publication, version, data publisher and persistent identifier. There are additional components that can be included if desired.
- The activity will:
    - Perform a Data Citation guidelines survey
    - Share agency experiences and lessons learned in data citation
    - Confirm the core component of data citation for EO needs
    - Select additional components of a data citation
    - Draft the EO Collection Data Citation Guidelines

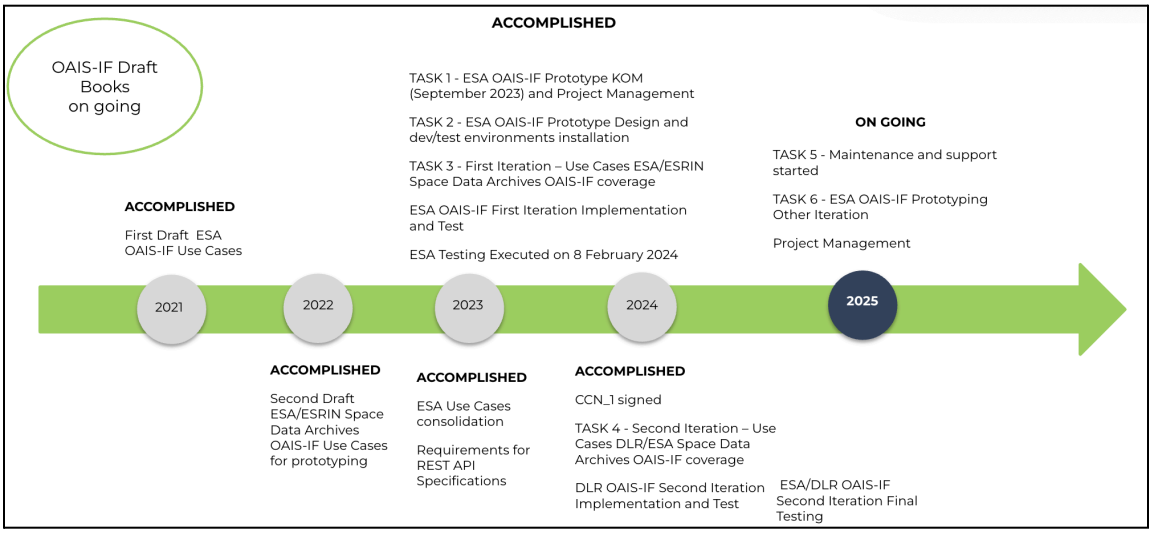
Discussion

- Nitant Dube (ISRO, WGISS Vice-Chair) recognised a collection will have a citation in the form of a DOI etc., however, within a collection, there could be additional DOIs available for each product.
- Mirko Albani (ESA) noted that citations should be defined for whatever is being used, and DOIs should be available for whatever is being cited. The DOI aspect has been covered already in existing best practices, this new activity will just cover the citation.
- Yves Coene (ESA/Spacebel) recognised that what matters most is what information is provided in the DOI, as citation styles are often prescribed by the journal or publisher. Tools such as citation.doi.org allow any style once the DOI exists.
- Iolanda noted confusion in citation could arise when the data is owned and disseminated by different organisations.
- Tom Sohre (USGS, WGISS Chair) suggested this could be more related to data replication, rather than citation. For example, the DOI for Landsat data on Google Earth Engine points back to the USGS website. Tom Sohre (USGS, WGISS Chair) would expect that people reference the USGS DOI. But if the provider changes the data in any way, they should provide a new DOI.
- Yousuke Ikehata (JAXA) noted that JAXA assigns a single DOI for multiple different product versions, to avoid having too many landing pages.
- Matt Paget (CSIRO) suggested that the guidelines recommend that data disseminators refer to the original DOI provided by the data provider.

9.5 - OAIS-Interoperability Framework Prototypes

Roberta Svanetti (ESA/Deda Tech) reported [*slides*]:

- OAIS-Interoperability Framework (OAIS-IF) is a supplement of the OAIS standards under draft from the Consultative Committee for Space Data Systems (CCSDS) Mission Operations and Information Management Services (MOIMS) Area Data Archive Interoperability (DAI) Working Group.
- OAIS-IF adds capabilities for system interoperability between users and archives, as well as between archives. It is applicable to any archive that complies with that OAIS standard as well as any Archive (non-OAIS conformant) that wishes to interoperate using the standard.
- OAIS-IF aims to define an implementable architectural framework for digital repositories and describe interoperable protocol and interface specifications based on standards, with broad applicability across a spectrum of use cases.
- Three prototypes are ongoing to demonstrate the draft framework. The prototypes will validate the OAIS-IF Core Specifications, assess applicability in the Agency context, test technical feasibility and gather feedback from developers



- Swagger UI was used to manually test the interactions with the REST APIs.
- A Specific Adapter was used to establish a prototype connection to external archives with the appropriate protocols implemented.
- The test bed for the first iteration was built as a set of Maven, hosted in a shared component repository for automated build and dependency management between

application stacks, and use of automated test tools. Postman will be used for the second iteration.

- Have successfully demonstrated an implementable architectural framework for ESA Space Data and DLR Long Term Archive repositories.
- A generic adaptor was implemented, starting from the OAIS-IF core specifications, based on OAIS family standards. Specific adapters are implementable in the user system.

Session 11: Heritage Datasets Recovery

This item was moved from later in the day.

11.2 - Status of AVHRR data recovery work

Mirko Albani (ESA) reported [*slides*]:

- DSIG has been working to recover heritage datasets for climate and other applications.
- Working to recover AVHRR local area coverage (LAC) data, gathering inputs from regional archives from around the world from before 2008. The objective is to build a complete global dataset of 1km resolution data from 1978.
- All data from the USGS network is now available at ESA. A software converter from the stitched format to HRTP format has been developed. The goal is to open access to all users by the end of 2025, providing access to global 1992-1995 data.
- The European dataset includes data from University of Bern, Dundee Station and ESA holdings, for a total of 260,000 harmonised products. Processing to Level-1C is complete, and the data is now available via ESA.
- A gap analysis tool was developed to understand the number of available products at specific locations for each time period. This can be used to trigger or support data retrieval activities. Currently only for ESA holdings, but will be extended to other CEOS Agency archives.
- A new project was started in Q1 2024 at ESA to reprocess European AVHRR LAC products series (NOAA & MetOp missions) to generate a L1C Fundamental Data Record (FDR). The project will be completed by mid 2026.
- Contacted the Polish Institute of Geodesy and Cartography Department of Remote Sensing, who hold data from NOAA satellites from 1996. The 350GB dataset has been

collected by ESA, and will be processed into Level-1B and Level-1C, with open access to users by Q1 2026.

- ESA has downloaded the full Dundee AVHRR Level-0 archive (8.9TB, 235,541 L0 products). Some data were reprocessed and included in the European dataset. Remaining data will be integrated to further extend geographical coverage towards Greenland and extend in time back to 1978.
- USGS and NRCan/CCMEO will be invited to make a presentation at WGISS-60 to update on their archive availability.
- Argentina is transferring their data to ESA (740GB archive), transcribing data from CD-ROMs.
- Data transfer of INPE's archive over Brazil to ESA has been completed.
- University of Hawaii shipped their archive (on exabyte tapes) to ESA and a transcription chain is being assembled. The ESRIN laboratory has manufactured a special device to roll/unroll the tapes at very slow speed and clean the tape surface from moisture before transcription.
- For Africa, have received data from SANSA and ASI (Kenya).
- For China, ESA has received access to NSMC-CMA's archive, but it is unclear how to download the data. Would like to invite NSMC-CMA to present at a future meeting.
- ISRO has processed the raw data to Level-1A, and will now process the data to Level-1C, using the same software as ESA. Will then make the data available and accessible through ISRO's Bhoonidhi platform.
- Have established contact with Mongolia. It is unclear what data they have, but will be hopefully shared with ESA soon.
- CSIRO is planning to reprocess all their data into higher levels and provide access.
- Inventory of existing national/regional HRPT and LAC data archives Version 3 has been produced in Q1 2024.
- ESA is compiling a list of Meteorological offices and other organizations around the world who might have archived AVHRR LAC data.

Discussion

- Matt Paget (CSIRO) recalled that CSIRO may also have some AVHRR data for Antarctica.

WGISS-59-20	CSIRO to check on the Australian archive of AVHRR data on Antarctica. Connect with the Digital Earth Antarctica team as they may have interest in this archive too.	**Due: WGISS-60**

- Tom Sohre (USGS, WGISS Chair) recognised the significance of this project in terms of CEOS collaboration. This project should be better highlighted within CEOS communications. Access to these datasets should also be better promoted.

WGISS-59-21	WGISS Secretariat to draft some communications about the AVHRR data recovery work, highlighting the outcomes from the effective WGISS collaboration.	**Due: WGISS-60**

Session 10: Software Preservation

10.1 - Concept and Session Introduction

Iolanda Maggio (ESA/Starion) reported [*slides*]:

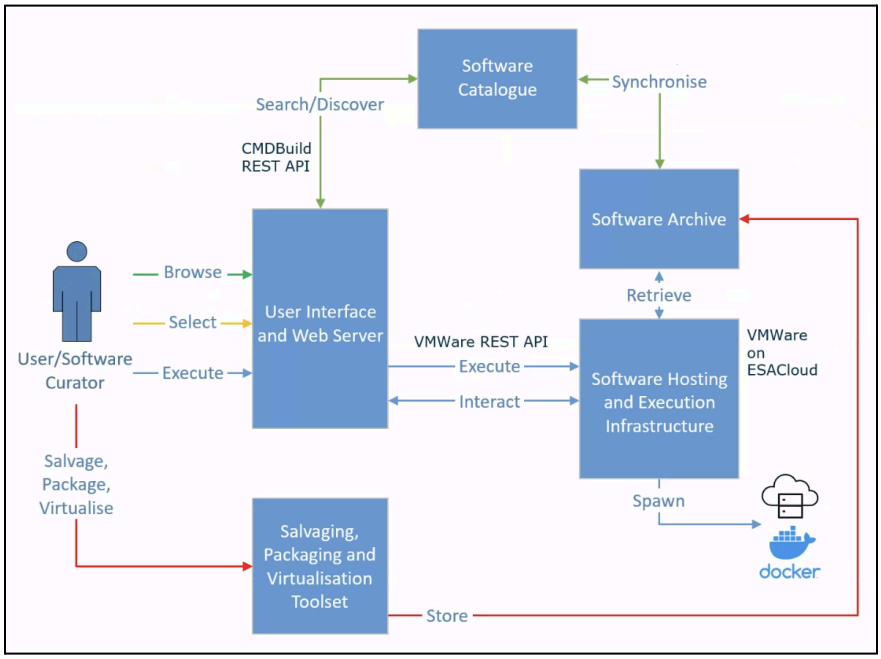
- The Software Preservation White Paper will assist data and software managers in the Earth observation (EO) domain with the task of ensuring the long-term preservation of software relating to EO missions and data, thus improving data accessibility and usability for current and potential future users.
- The intended audience should comprise data and software providers, decision makers and scientists, and data managers/stewards for data centres and repositories.
- Details of the main principles of software preservation will be provided, as well as brief descriptions of the primary strategies that may be implemented by data managers, together with challenges.
- Software preservation is important to maintain the maximum value of the satellite data, including for re-analysis.
- Three principles will be investigated:
    - Collection of software, documents and knowledge
    - Preservation, in a complete and robust manner, ensuring redundancy
    - Sharing software in an accessible, open and persistent way.
- Will cover a number of techniques of different software preservation strategies, including the pros and cons of each.

- Lack of documentation and expert knowledge is a challenge often encountered when trying to preserve software.
- Even if all the required expertise is available, preservation is still a time-consuming process that requires significant effort and commitment from the personnel involved

10.2 - ESA - Heritage Software Hub (HESH)

Bryan Keary (ESA/Solenix) reported [*slides*]:

- Heritage Software Hub (HESH) is ESA ESRIN's hub for preserving software, where the software is still accessible while being preserved.
- HESH is a platform for collection and preservation of Earth observation (EO) software in a functional environment to enable on-demand use of preserved software.
- Software is packaged in containers to eliminate or reduce the need to preserve specific hardware and legacy OS components.



- Users can manage and access the software via Virtual Machines.
- Encountered a number of challenges, including lack of documentation and expert knowledge, as well as lack of familiarity with the software by data managers/preservationists. Risk of partial/full obsolescence prior to beginning the preservation process poses a critical threat.
- These risks can be mitigated in future by good preservation planning and starting preservation activities as soon as possible.

- Dependency on other software, e.g. libraries, is also a risk.
- ESRIN has developed the concept of the EO Software Preservation Module, which is the collection of everything required to adequately preserve the software for future use. This includes test datasets.
- ESRIN is now working to convert Level-0 ERS and Envisat SAR data to higher level products, using preserved software.

Discussion

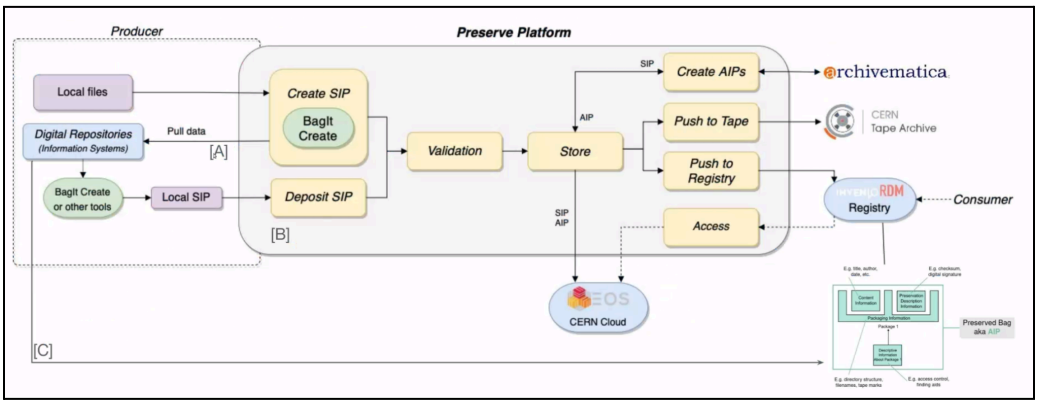
- Bryan clarified that it is the source code that is preserved, such that the software can be re-built from scratch. The docker containers are used for convenience for running the code on HESH. VMWare is also not critical for the archiving process, and is just used for executing.
- Mirko Albani (ESA) added that the software is archived in HESH without making modifications. A copy is then made available on HESH for executing, for which modifications may need to be made.
- Nitant Dube (ISRO, WGISS Vice-Chair) recognised that during the active lifecycle of the software, it goes through various versions. All versions should be preserved.
- Paulo Sacramento (ESA/Solenix) noted that VMWare, and other commercial vendors, support Open Virtualization Format (OVF), which is vendor independent.

10.3 - CERN Preservation Strategy Framework

Jean Yves Le Meur (CERN) reported [*slides*]:

- CERN is looking closely at what is being done by space agencies regarding their digital preservation strategies.
- Aiming to align with the EU Long Term Preservation recommendations.
- CERN is participating in a number of community activities on this topic, including the Data Preservation Coalition (DPC), Archiver EU project, Data Retention Task Force, and EOSC EDEN project.
- CERN is working to get their Information Systems certified as a Trusted Repository, from both the policy and technical sides.
- The current CERN Archiving Policy is under revision. The new version will redefine the perimeter (scope), access and responsibilities.
- The perimeter will be flexible, and will describe the process to enter in the scope of preservation.

- For information technology, CERN will develop ‘preservation as a service’, with infrastructure to help CERN data repositories to preserve content at best.



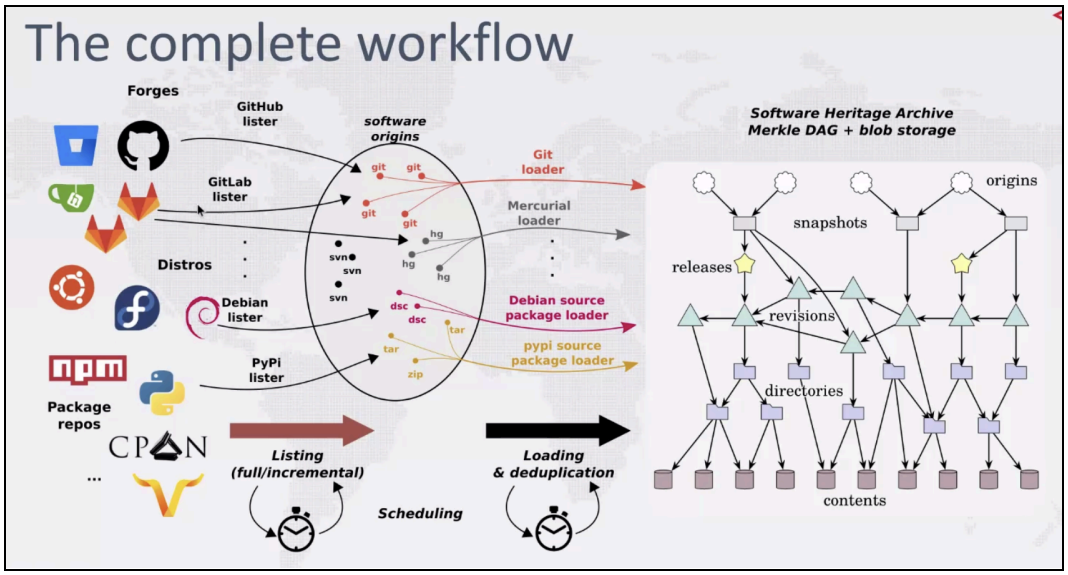
Discussion

- Mirko Albani (ESA) asked about the benefits of participating in the Digital Preservation Coalition. Jean Yves noted that it is a large consortium with many members, containing a wide array of knowledge about various tools. Webinars are hosted on various topics relating to Digital Preservation.
- Jean Yves added that CERN isn’t the only organisation to have developed a ‘preservation as a service’ system, however there aren’t any standards covering this topic yet.

10.4 - Software Preservation

Morane Gruenpeter (Software Heritage) reported [*slides*]:

- Software Heritage is a non-profit, multi-stakeholder and open infrastructure to collect, preserve and share all source codes. More than 23 billion source files are currently included in the archive. The project is connected with the French National Archive (HAL).
- Software artefacts are the executables and source codes, and the environments around it. Software Heritage aims to archive the full history of the software, preserving the source code.
- Software is a key pillar in academia, alongside data and articles.
- Software Heritage has four key services: deposit, crawling, save code now, and rescue operations.



- CodeMeta is an academic community and a tool to describe software with intrinsic metadata, and is a subset of schema.org. Software Heritage have developed a CodeMeta tool to encourage developers to include a JSON metadata file in their source code.
- Software Hash Identifiers (SWHAS) are intrinsic identifiers which are intimately bound to the designated object, in a decentralised and cryptographically strong way.
- Software preservation requires a global and coordinated effort.

Discussion

- Morane added that all software is converted to be git-compatible, and different packaged systems and legacy systems are converted into more modern systems.
- Iolanda Maggio (ESA/Starion) asked whether all surrounding software, e.g. libraries, are also preserved alongside the software. Morane noted that it is the responsibility of the person requesting the archive to preserve everything required. Dependencies in a git are not searched for.

Session 11: Heritage Datasets Recovery (cont.)

11.3 - Updated differences between AVHRR L1b/1c satellites products

Mirko Albani (ESA) reported [*slides*]:

- DSIG is pursuing harmonization/alignment of AVHRR Level-1B and -1C products available in different archives from ESA, NOAA, EUMETSAT and CSIRO.

- Having aligned AVHRR Level-1B and Level-1C products worldwide can facilitate exploitation and use of AVHRR long time data series.

	ESA	NOAA	EUMETSAT	CSIRO
GAC L1b	N.A.	NOAA format	EPS or HDF 5 format	TBD
GAC L1c	N.A.	N.A.	netCDF format	TBD
LAC L1b	NOAA format	NOAA format	EPS or HDF 5 format	TBD
LAC L1c	netCDF format	N.A.	N.A.	TBD

- CSIRO is still working on finding the details, but the formats are L1b HRTP format, L1c netCDF. CSIRO will create a proper landing page for the data, and make visible the relevant details.

WGISS-59-22	CSIRO to create a landing page for their Australian AVHRR collection, which should include information about the processing and data formats.	**Due: WGISS-60**

- The next steps are to understand the difference from a user perspective on the different formats. Extend the analysis and investigate potential approaches for GAC/LAC data products alignment.

Discussion

- Sai Kalapana (ISRO) asked that ISRO products be included in this table as well.

WGISS-59-23	DSIG to add ISRO's data formats to the comparison of AVHRR collections.	**Due: WGISS-60**

- Nitant Dube (ISRO, WGISS Vice-Chair) recognised the importance for consistency between the level definitions.

Session 12: Other Topics

This item was moved from Friday morning.

12.2 - Communications

Libby Rose (WGISS Secretariat) reported [*slides*]:

- The new Collaborations page on the WGISS website replaces the previous 'Past Activities' page. Information from the three previous subpages (Recovery Observatory, Water Portal, GA.4.Disasters) were consolidated into one page.
- A new page has been added under Data Stewardship and Preservation for the Heritage Dataset Recovery project.
- A new section has been added to the Best Practices & Guides page to highlight External Documents of Relevance.
- A WGISS Overview Slide Deck was started last year, however the usage and purpose for this remains unclear.
- The CEOS Communications Team is also working on an article focused on the WGISS Connected Data Assets, plus some social media communications about the Jupyter Notebooks Best Practices.
- Future topics for communications could include the Interoperability Handbook, to solicit community feedback and encourage use, and Heritage Datasets as discussed previously.
- CEOS will have an exhibition booth at Living Planet Symposium 2025, where WGISS can provide physical or digital materials.

Discussion

- Key activities of WGISS, such as the Connected Data Assets and Data Purge Alert, should be visible at a higher level within the website. Activities should be featured on the WGISS homepage.

WGISS-59-24	WGISS Secretariat to redesign the WGISS website home page, to better highlight current and ongoing activities. Add 'Interest Groups' to the headings in the website menu.	**Due: Q2 2025**

- Tom Sohre (USGS, WGISS Chair) recognised the usefulness of the FedEO and IDN data access tables, and suggested these should be also more visible. A combined table would also be beneficial for simplicity.

WGISS-59-25	DAIG to try to merge the two data access tables for FedEO and IDN. This combined table should then be featured at a higher level on the website, either on the Connected Data Assets or Data Discovery and Access page.	**Due: Q2 2025**

- The other opportunity on the CEOS website is the 'Our Work' page, where tools and services are listed in a visible way. Purge Alert could potentially be added to this page.

WGISS-59-26	WGISS Secretariat to ask SEO whether Purge Alert should be added to the CEOS Website 'Our Work' page.	**Due: Q2 2025**

- Matt Paget (CSIRO) noted the overview slide deck is helpful for various occasions, and can be updated as needed. Tom suggested it should be updated at least annually. The slide deck should contain the scope of the deliverables, but not necessarily detailed status.
- Interest group leads should be responsible for updating their activities in the slide deck.

WGISS-59-27	WGISS Secretariat to update the WGISS Overview slide deck, and share with Interest Group leads to provide their updates.	**Due: Q2 2025**

Action Review

Libby Rose (WGISS Secretariat) reviewed the actions recorded throughout the meeting. The full list of actions can be found in Appendix B.

Day 5: Friday 28th March, 2025

Session 12: Other Topics

12.1 - CEOS-ARD Strategy: Discussion on items for WGISS to support

Matt Steventon (LSI-VC Secretariat) reported [*slides*]:

- CEOS-ARD Strategy 2024 was endorsed at the CEOS Plenary 2024, as an update to previous versions from 2019 and 2021, to reflect on progress to date and take stock of future directions and needs.
- The overarching goal is a broad portfolio of CEOS-ARD that is easily discovered, accessed and utilised.

- WGISS can primarily support the discovery, access, utilisation and interoperability aspects of the CEOS-ARD Strategy.
- Have identified the need for a future evolution of CEOS-ARD, in particular around higher requirements for data 'quality', among other things. 2025 will focus on gathering community inputs towards these concepts, with events planned for ESA LPS, IGARSS, IAC and others.

The following items from the Strategy were raised for discussion:

- 1.4: CEOS-ARD at mission inception and planning stages and in archive reprocessing plans.
  - Tom Sohre (USGS, WGISS Chair) is not sure whether WGISS tracks these reprocessing plans centrally, but members are likely aware of these events.
  - Nitant Dube (ISRO, WGISS Vice-Chair) recognised linkage to the heritage data aspects.
  - Can/should WGISS take a stocktake of these types of events? The primary goal is to make sure those in charge of the reprocessing plans are aware of CEOS-ARD specifications.
  - Matt noted awareness could also be raised in conjunction with the CEOS MIM Database annual survey of Agency programmes and plans.
- 2.4: Replica datasets, authenticity, traceability, and inherited CEOS-ARD compliance.
  - Tom noted WGISS hasn't done much work to date regarding replicas and authentication, but the topic has come up in other contexts. Something WGISS can look at in future.
  - The activity would need to consider what 'replica' means, and how to check whether a replica is a true replica.
- 3.1: Embrace the CEOS Interoperability Framework.
  - Looking to the handbook to help guide CEOS-ARD to become more interoperable.
  - Tom recommended the CEOS-ARD team review all the recommendations in the Interoperability Handbook.

WGISS-59-28	LSI-VC to review the Interoperability Handbook recommendations, considering alignment with the CEOS-ARD framework.	**Due: Q3 2025**

- Following the handbook's completion, WGISS will do some demonstrator projects to demonstrate the benefits of interoperability and test the recommendation.
- WGISS is also planning to develop a maturity matrix for self-assessment against those recommendations.
- LSI-VC would be interested in collaborating on an interoperability demonstrator focused on CEOS-ARD.
- The handbook will also not be a static document, and will be continually updated based on new information.

- 3.2: Alignment of CEOS-ARD and STAC.
  - Have been moving the CEOS-ARD specifications to GitHub, breaking down the specifications into building blocks. One goal with this is to make the mapping to STAC extensions.
  - Is there anything from the STAC best practices to consider in this effort?
  - Yves Coene (ESA/Spacebel) noted the current best practices focus on the discovery aspects. However, in FedEO and IDN, where CEOS-ARD datasets exist, there are no relevant keywords to support discovery.
- 3.3: Cloud native approaches and machine-to-machine access and utilization
  - This is related to the AI/ML Ready Data work done by WGISS TEIG.
  - Nitant noted that adding labels to data would make CEOS-ARD essentially ready to be used for AI/ML training.
  - Alex Leith (SEO) noted there are a couple of points in the architecture and interface sections of the Interoperability Handbook which focus on the cloud aspect. An 'interoperability checklist' could be included as an annex to the CEOS-ARD specifications.

WGISS-59-29	Alex Leith to pull out the specific recommendations from the Interoperability Handbook which could support cloud native approaches for CEOS-ARD, for discussion with LSI-VC. A checklist for interoperability could be included as an annex to the CEOS-ARD specifications.	**Due: Q2 2025**

- 3.4: CEOS-ARD discoverability and branding
  - Yves noted that proper keywords can help with the discoverability in the WGISS Connected Data Assets.

WGISS-59-30	CEOS-ARD Oversight Group to work with CEOS MIM Database team and LSI-VC to identify and develop 'CEOS-ARD' keywords to be provided to DAIG for FedEO and IDN, to support discovery of the datasets. CEOS-ARD Oversight Group should consider how to recommend the use of these keywords to support better discovery of CEOS-ARD datasets.	**Due: WGISS-60**

- 3.5: CEOS-ARD in the commercial cloud.
  - Clearly recommending how datasets should be tagged (e.g. with 'CEOS-ARD') would help discoverability in online data hubs. Tagging is currently inconsistent.
  - Including the logo and links through to the specifications would also be nice.
- 6.1: CEOS-Industry ARD workshops alongside key CEOS meetings
  - LSI-VC has been running CEOS-ARD industry workshops alongside LSI-VC meetings. There is an opportunity for WGISS to hold similar sessions at their future meetings. Matt is also encouraging WGCV to do the same at their upcoming meeting in India.
  - The goals for these workshops is primarily information sharing for both sides.
  - Nitant would prefer a CEOS-wide workshop with a broader perspective.

8.6 - Data compression

Robert Fletcher (UKSA) reported [*slides*]:

- Data compression is the process of reducing the size of a data file by reducing the number of bits needed to represent the data file.
- Data compression involves various algorithms and standards that determine how data is compressed. These different algorithms dictate the methods and rules for reducing the size of files or data streams.
- New data volumes from EO satellites will create challenges around storage cost, duplication of data, data management, efficient access to large data volumes and operational efficiency.
- Data deduplication is a data storage technology that takes redundant and repeated parameters and/or files and replaces them with a single identifier or pointer whilst saving only one copy of the data, therefore significantly reducing the size of the data being stored.

- Lossy data compression methods sacrifice some data to achieve higher compression ratios, however at the expense of quality.
- Lossless compression algorithms retain all the original data, and retain data integrity. The original data can be fully reconstructed from the compressed version without any loss of information.
- Challenges in EO data compression include preserving critical details, adapting to diverse sensors, real-time processing needs and computational complexity.
- AWS have conducted a demonstrator to deploy a low cost solution to compress Worldview-3 satellite imagery with Geospatial Data Abstraction Library (GDAL), using AWS Lambda, Docker, S3 and S3 Glacier storage. The solution archives the original images, and keeps the compressed images online. Data quality is kept at 95%, while reducing the file sizes by 96%.
- AI based data compression brings adaptability and learning capabilities to the compression process, rather than fixed rules in traditional algorithms.
- Vito has used deep learning models to achieve high compression ratios while maintaining image quality, to allow for complex data analysis. Developed CORSA, which is a revolutionary AI Driven Data Compression Algorithm for EO data. CORSA has proven compression rates of 100x on Sentinel-1, Sentinel-2 and PRISMA data, without compromising data quality. CORSA is described as 'nearly lossless'.
- DestinEStreamer (within Copernicus' Destination Earth) has achieved a compression ratio of input data of 1:34, with less than 1% difference in quality compared to the original datasets.

Discussion

- Tom Sohre (USGS, WGISS Chair) commended Robert on a great overview presentation of this topic, to help WGISS understand where to go next. The use cases presented were helpful to understand different approaches.
- Alex Leith (SEO) recognised the opposite aspect where 'super resolution' datasets are made from Sentinel-2 or Landsat-8 data, adding information to provide higher resolution data. It would be interesting to see how that is impacting the data quality.
- Alastair McKinstry (ICHEC) highlighted the importance of considering the use case when comparing compression techniques. In general, for scientific analysis, the original data is needed. Furthermore, the hardware support for compression and compression algorithms can impact the speed and efficiency of uplink and downlink from satellites.

WGISS-59-31	TEIG to invite compression experts to WGISS-60 to present their experiences, perhaps from Destination Earth DestinEStreamer (Copernicus), or VITO Remote Sensing's CORSA. Other suggestions for presentations to this session are welcomed.	**Due: WGISS-60**

Session 13: Closing Session

13.0 - Future WGISS Activities

Tom Sohre (USGS, WGISS Chair) invited all to use the extra time to discuss future WGISS activities. The following questions were raised:

- How should WGISS tackle the topic of replica datasets and authenticity?
    - This topic has been touched on a number of times in the past. Should WGISS produce some documentation to provide guidance on this topic, beyond information sharing?
    - The interoperability handbook, specifically the architecture/publishing sections, should be used as a starting point.
    - Could develop a concept of a "Certified replica/repository of a CEOS-ARD Collection", with a set of quality metrics that need to be met, such as 99.999% completeness in terms of granules.
    - With whom the responsibility lies to ensure the quality of the replicated data should also be considered. How do users know that replicas are complete and accurate?
    - Most checks used are at a file/granule level, and are not applicable for pixel level access.
    - OCG 24-033 could be a key reference.
    - The granule (or collection) metadata could become a W3C verifiable credential (with DID) and includes the multihash of the referenced assets.
    - Multi-URLs could be indicated in STAC or ISO-19115.
    - The signature of the "issuer" is present within the DID, and can also be identified in the replica granules and verified by the users. DID can be included in the discovery metadata.
    - WGISS could also explore technologies that enable efficient replication of datasets.

- WGISS should discuss the way Agencies would expect to see duplicated/replica dataset presented to users, and identify approach(s) to authenticate data that has been replicated.
- WGISS could also explore tamper or blue seal technologies.
- Commercial sector input would likely be useful for this topic.
- The Collections Management White Paper Principle 11 states '*Consider approaches that authenticate replicas of collections*'. Suggested approaches are given, but no single method is recommended. It is also not explored in detail, and could be expanded on.

WGISS-59-32	DSIG (and perhaps DAIG) to consider whether WGISS should develop Best Practices for replica datasets, in particular approach(s) to authenticate data that has been replicated. Consider providing guidance on metadata fields for source and tracing of data production. Scope out the effort, and discuss further at WGISS-60.	**Due: WGISS-60**

- How would we like to engage with commercial organizations (who and how)?
  - A workshop could be defined alongside a future WGISS meeting to discuss the topic of replica datasets with the commercial sector.
  - CEOS has recently formalised language online regarding engagement with the commercial sector. No specific guidance was given as to how engagement should be conducted such that it is fair and balanced.
  - Using CEOS-ARD in AI/ML Applications may be a good engagement topic for the commercial sector.
  - The presentation earlier in the week regarding Satellite Tasking API was a good way to engage with the commercial sector.
  - WGISS should look for specific opportunities to engage with the commercial sector as they arise.
- What is the WGISS role with STAC beyond what has already been done?
  - STAC and STAC API should be used as a central part of the interoperability demonstrators. The benefits of STAC should be communicated widely to the community.
  - WGISS should remain engaged in this topic as STAC is increasing in uptake and use across the sector.

- WGISS Connected Data Assets will complete the migration from OpenSearch to STAC, and develop a guide for agencies wishing to transition to STAC.
- An integrated handbook across all the different aspects of STAC could be helpful, covering discovery, CEOS-ARD, data descriptions for data loading, and more.
- Better alignment between which STAC extensions should be used by various implementations would be beneficial.

- Are there specific CEOS entities that WGISS should develop stronger relationships with (and what would that look like)?
  - The concept of 'Decision Ready Data' would be a topic to discuss with WGDisasters. Automated tasking (such as STAPI) could also be of interest to WGDisasters.

WGISS-59-32	WGISS Chair/Vice-Chair to arrange a telecon with WGDisasters to discuss potential topics of mutual interest. Consider whether a joint meeting in 2026 would be worthwhile.	**Due: Q2 2025**

  - LSI-VC could be another group to coordinate closely with.
  - The joint meeting with WGCV in October 2024 was successful, due to the many overlapping topics.
- When we explore technologies (AI/ML, Cloud, etc), what is the unique added value that WGISS provides, above what is already being done within Agencies?
  - Discovery of foundation models for EO could be explored by CEOS. Some background on this topic could be shared at a future meeting, as an information sharing exercise.
  - The multilateral nature of WGISS allows for the development of interoperable cloud solutions, considering challenges with varied providers and various legislation.
  - WGISS is a platform to bring together the collective minds of many subject matter experts from across many agencies.
  - It is useful to know how individual organisations are aligned with others.
- What other topics should WGISS tackle in the future?
  - ESA is proposing to use Zarr for storage of data in the future, which could be explored within WGISS.
  - WGISS could explore implementing event-driven notification services using a publish/subscribe architecture and advancing data/metadata as a service models to

improve real-time data delivery, interoperability, and user access to Earth observation resources.

13.1 - Action Review

The remaining actions were reviewed offline. The full list of actions can be found in Appendix B.

13.2 - Closing Remarks

Tom Sohre (USGS, WGISS Chair) reported [*slides*, *WGISS-60 information*]:

- Thanked GISTDA for hosting the meeting.
- The Agency Reports are a valuable part of WGISS meetings, and provide a great opportunity to share information about WGISS related processes within agencies.
- Interoperability Handbook sessions showed good progress to develop and refine the recommendations.
- The AVHRR data recovery work has shown a great example of collaboration.
- WGISS-60 will be held at Oberpfaffenhofen, Germany, hosted by DLR, on October 13-17, 2025. Hotel information will be provided by mid-April, 2025.
- Damiano Guerrucci (ESA) has nominated to take on the role of WGISS Vice-Chair from November 2025 - November 2027, followed by Chair for two years.

Discussion

- Damiano Guerrucci (ESA) commended WGISS for the work done this week regarding the Interoperability Handbook. The handbook summarises a lot of information in a relatively small number of recommendations, which was only possible due to previous work. The handbook showcases the end to end completeness of WGISS activities.

Appendix A: List of Participants

In-person participants

Affiliation	Name
CSIRO	Matt Paget
ESA	Mirko Albani
ESA/Solenix	Filippo Marchesi
ESA/Spacebel	Yves Coene
ESA/Starion	Iolanda Maggio
GISTDA	Mr. Boonchoob Boongthong
GISTDA	Natthawat Hongkarnjanakul
GISTDA	Pakorn Apaphant
GISTDA	Pawarin Kuha
GISTDA	Phasaporn Aroonjaroensuk
GISTDA	Poramet Thuwakham
GISTDA	Prayot Puangjaktha
GISTDA	Tanita Suepa
GISTDA	Viphada Boonlerd
GISTDA	Worathan Wichakoon
ISRO	Nitant Dube
JAXA	Yousuke Ikehata
SEO/Auspatious	Alex Leith
UKSA	Robert Fletcher
VNSC	Nguyen Tien Cong
WGISS Secretariat	Libby Rose

Virtual Participants

Affiliation	Name
CEOS Executive Officer	Steven Ramage
CERN	Jean Yves Le Meur
CGI	John Milton-Worssell
CGI UK Ltd	Richard Doyle
Deda Tech	Roberta Svanetti
DLR	Christopher Reck
DLR	Mario Winkler
ESA	Damiano Guerrucci
ESA	Hayret Abdula Keary
ESA	Salvatore Pinto
ESA	Saskia Brose
ESA	Sofia Marzo
ESA/Solenix	Bryan Keary
ESA/Solenix	Paolo Sacromento
ESA/Starion	Daniele Iozzino
ESA/Starion	Sergio Folco
Geoscience Australia	Michael Wellington
ICHEC	Alastair McKinstry
ISRO/NRSC	Narendran J
ISRO/NRSC	Sai Kalpana
JAXA	Makoto Natsuisaka
LSI-VC	Matt Steventon
NASA/EED-3/KBR	Michael Morahan
Software Heritage	Morane Gruenpeter
STFC	Esther Conway
UKSA	Maral Bayaraa
USGS	Tom Sohre
USGS/KBR	Chris Barnes

Appendix B: Actions

WGISS-59-01	WGISS Chair/Vice-Chair to share the draft Interoperability Handbook with Paola De Salvo for GEO's review and input.	**Due: Q2 2025**
WGISS-59-02	WGISS Secretariat to link the Harmonizing Quality Measures of FAIRness Assessment Towards Machine-Actionable Quality Information under 'External Documents of Relevance'.	**Due: ASAP**
WGISS-59-03	Organise a session at WGISS-60 on lessons learned from working with large volumes of data (big data). Explore the concept of compute next to data.	**Due: WGISS-60**
WGISS-59-04	WGISS Chair to connect with WGClimate for a discussion about ECV specifications, discoverability and archival. Consider how ECVs/CDRs could be used as a demonstrator for interoperability.	**Due: WGISS-60**
WGISS-59-05	WGISS members to contact the SEO to be involved in beta testing of either CEOS Liaison for Earth Observations (CLEO) or EO-GPT.	**Due: Q2 2025**
WGISS-59-06	Prayot Puangjaktha is the new nominated GISTDA representative for DAIG, and will review the GISTDA datasets in IDN. WGISS Secretariat to add Prayot to the DAIG mailing list.	**Due: ASAP**
WGISS-59-07	DIIG to compare the Open Science recommendation from the Policy Chapter of the Interoperability Handbook against the NASA Open Science policy, and other CEOS Agency Open Science policies.	**Due: WGISS-60**
WGISS-59-08	WGCV to consider adding a general recommendation for vicarious calibration in the Quality Factor of the Interoperability Handbook.	**Due: Q2 2025**
WGISS-59-09	Doug Newman to check if the trade study done by NASA for the various social login providers for their Federated Authentication and Authorisation work can be shared with WGISS.	**Due: WGISS-60**
WGISS-59-10	DSIG to document EO Data Collection Appraisal procedures.	**Due: WGISS-60**

WGISS-59-11	DAIG and TEIG to consider whether a separate white paper on Data Federation is needed, alongside the current work on User Federation.	**Due: WGISS-60**
WGISS-59-12	DAIG to discuss with the Copernicus Regional Data Hubs regarding their approach to federated authentication and authorisation, as well as data federation. Invite to present at WGISS-60 if appropriate.	**Due: WGISS-60**
WGISS-59-13	Hayret Abdula Keary (ESA) to write up their EO AI Advanced Assistant as a use case for LLM in the AI/ML White Paper.	**Due: Q2 2025**
WGISS-59-14	Alex Leith to check with Jed Sundwall to get a copy of the report from the 2020 AI/ML EO workshop, for input into the AI/ML White Paper.	**Due: ASAP**
WGISS-59-15	TEIG to document the scope of the AI/ML ready data definition task, and who we would be doing it for.	**Due: Q2 2025**
WGISS-59-16	TEIG to document the scope of the Digital Twins white paper.	**Due: ASAP**
WGISS-59-17	TEIG to include Matt Paget and Alastair McKinstry in the Digital Twins white paper team. GISTDA and ESA to nominate someone to join the team.	**Due: ASAP**
WGISS-59-18	DSIG to check the glossary from the Collection Management and Governance White Paper with other WGISS documents. Once complete, send to Peter Strobl to ensure the definitions and terms are included in the CEOS Common Dictionary.	**Due: ASAP**
WGISS-59-19	The Collection Management and Governance White Paper was approved. Following the completion of WGISS-59-17, WGISS Secretariat to publish the document online and distribute to WGISS members.	**Due: ASAP**
WGISS-59-20	CSIRO to check on the Australian archive of AVHRR data on Antarctica. Connect with the Digital Earth Antarctica team as they may have interest in this archive too.	**Due: WGISS-60**

WGISS-59-21	WGISS Secretariat to draft some communications about the AVHRR data recovery work, highlighting the outcomes from the effective WGISS collaboration.	**Due: WGISS-60**
WGISS-59-22	CSIRO to create a landing page for their Australian AVHRR collection, which should include information about the processing and data formats.	**Due: WGISS-60**
WGISS-59-23	DSIG to add ISRO's data formats to the comparison of AVHRR collections.	**Due: WGISS-60**
WGISS-59-24	WGISS Secretariat to redesign the WGISS website home page, to better highlight current and ongoing activities. Add 'Interest Groups' to the headings in the website menu.	**Due: Q2 2025**
WGISS-59-25	DAIG to try to merge the two data access tables for FedEO and IDN. This combined table should then be featured at a higher level on the website, either on the Connected Data Assets or Data Discovery and Access page.	**Due: Q2 2025**
WGISS-59-26	WGISS Secretariat to ask SEO whether Purge Alert should be added to the CEOS Website 'Our Work' page.	**Due: Q2 2025**
WGISS-59-27	WGISS Secretariat to update the WGISS Overview slide deck, and share with Interest Group leads to provide their updates.	**Due: Q2 2025**
WGISS-59-28	LSI-VC to review the Interoperability Handbook recommendations, considering alignment with the CEOS-ARD framework.	**Due: Q3 2025**
WGISS-59-29	Alex Leith to pull out the specific recommendations from the Interoperability Handbook which could support cloud native approaches for CEOS-ARD, for discussion with LSI-VC. A checklist for interoperability could be included as an annex to the CEOS-ARD specifications.	**Due: Q2 2025**
WGISS-59-30	CEOS-ARD Oversight Group to work with CEOS MIM Database team and LSI-VC to identify and develop 'CEOS-ARD' keywords to be provided to DAIG for FedEO and IDN, to support discovery of the datasets. CEOS-ARD Oversight Group should consider how to recommend the use of these keywords to support better discovery of CEOS-ARD datasets.	**Due: WGISS-60**

WGISS-59-31	TEIG to invite compression experts to WGISS-60 to present their experiences, perhaps from Destination Earth DestinEStreamer (Copernicus), or VITO Remote Sensing's CORSA. Other suggestions for presentations to this session are welcomed.	**Due: WGISS-60**
WGISS-59-32	DSIG (and perhaps DAIG) to consider whether WGISS should develop Best Practices for replica datasets, in particular approach(s) to authenticate data that has been replicated. Consider providing guidance on metadata fields for source and tracing of data production. Scope out the effort, and discuss further at WGISS-60.	**Due: WGISS-60**
WGISS-59-32	WGISS Chair/Vice-Chair to arrange a telecon with WGDisasters to discuss potential topics of mutual interest. Consider whether a joint meeting in 2026 would be worthwhile.	**Due: Q2 2025**