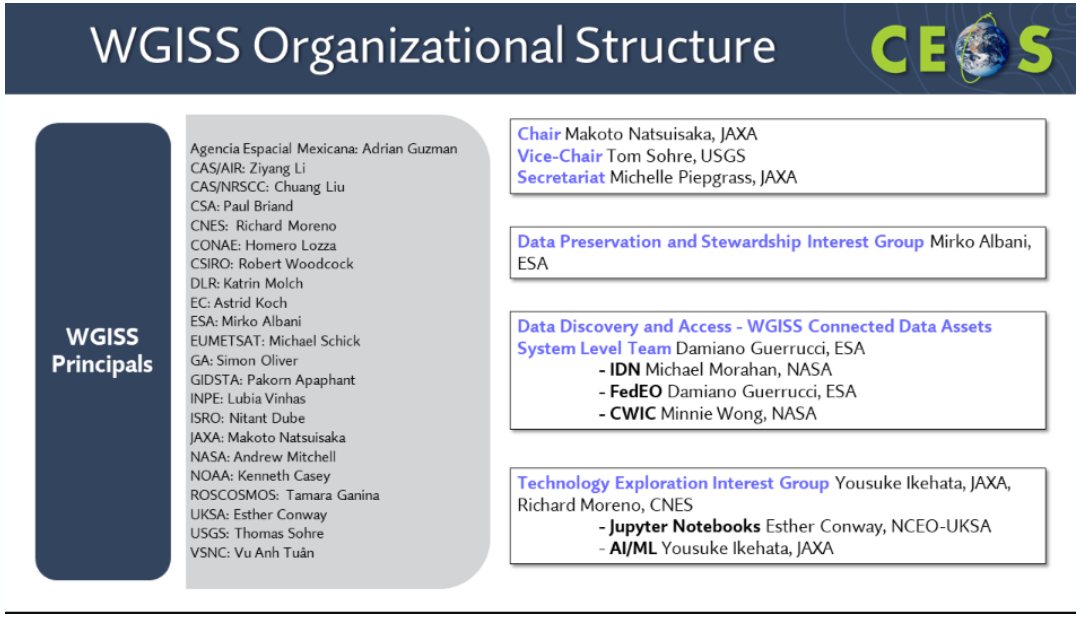
CEOS Best Practice for Long Term Data Preservation and Curation

I. Maggio (Rhea for ESA) PV 2023 May 2–4, 2023 - CERN

WGISS Background and Scope

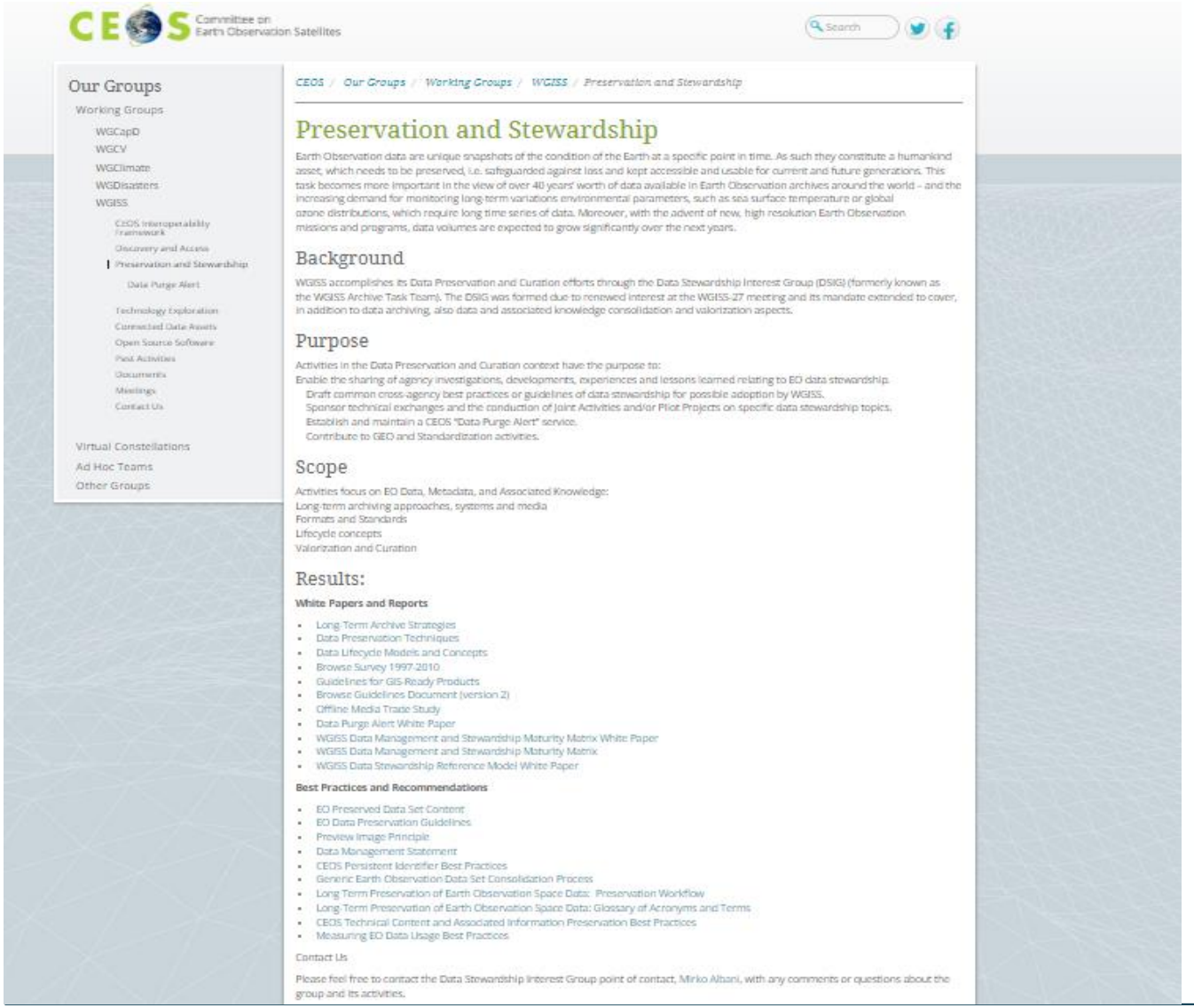
WGISS (The Working Group on Information Systems and Services) is a subsidiary body supporting CEOS.



- ✓ Promotes collaboration in the development of systems and services that manage and supply Earth Observation data;
- ✓ Creates and demonstrates prototypes supporting CEOS and Group on Earth Observation (GEO) requirements;
- ✓ Addresses the internal management of EO data, the creation of information systems and the delivery of interoperable services.

The activities and expertise of WGISS span the full range of the information life cycle from the requirements and metadata definition for the initial ingestion of satellite data into archives through to the incorporation of derived information into end-user applications.

DSIG Background and Scope



- Enable the sharing of agency investigations, developments, experiences and lessons learned relating to EO data stewardship.
- Draft common cross-agency best practices or guidelines of data stewardship for possible adoption by WGISS.
- Sponsor technical exchanges and the conduction of Joint Activities and/or Pilot Projects on specific data stewardship topics.
- Establish and maintain a CEOS “Data Purge Alert” service.
- Contribute to GEO and Standardization activities.
- Activities focus on EO Data, Metadata, and Associated Information.
- Long-term archiving approaches, systems and media.
- Data Formats and Standards.
- Preservation Lifecycle concepts.
- Data Valorization and Curation.

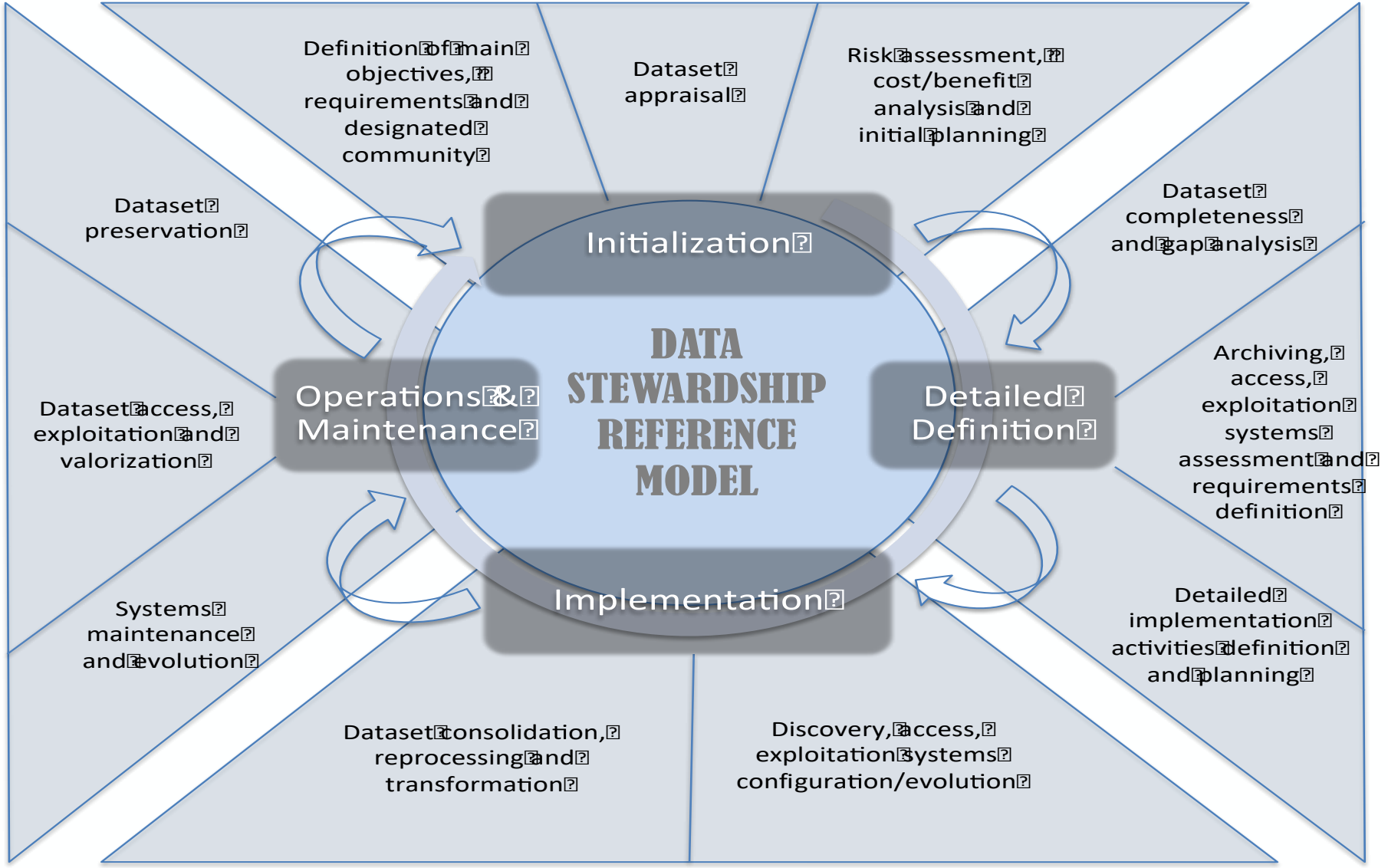
BIG DATA & LTDP: Challenges

- ✓ Capture
- ✓ Curation
- ✓ Storage
- ✓ Search
- ✓ Sharing
- ✓ Transfer
- ✓ Analysis
- ✓ Visualization

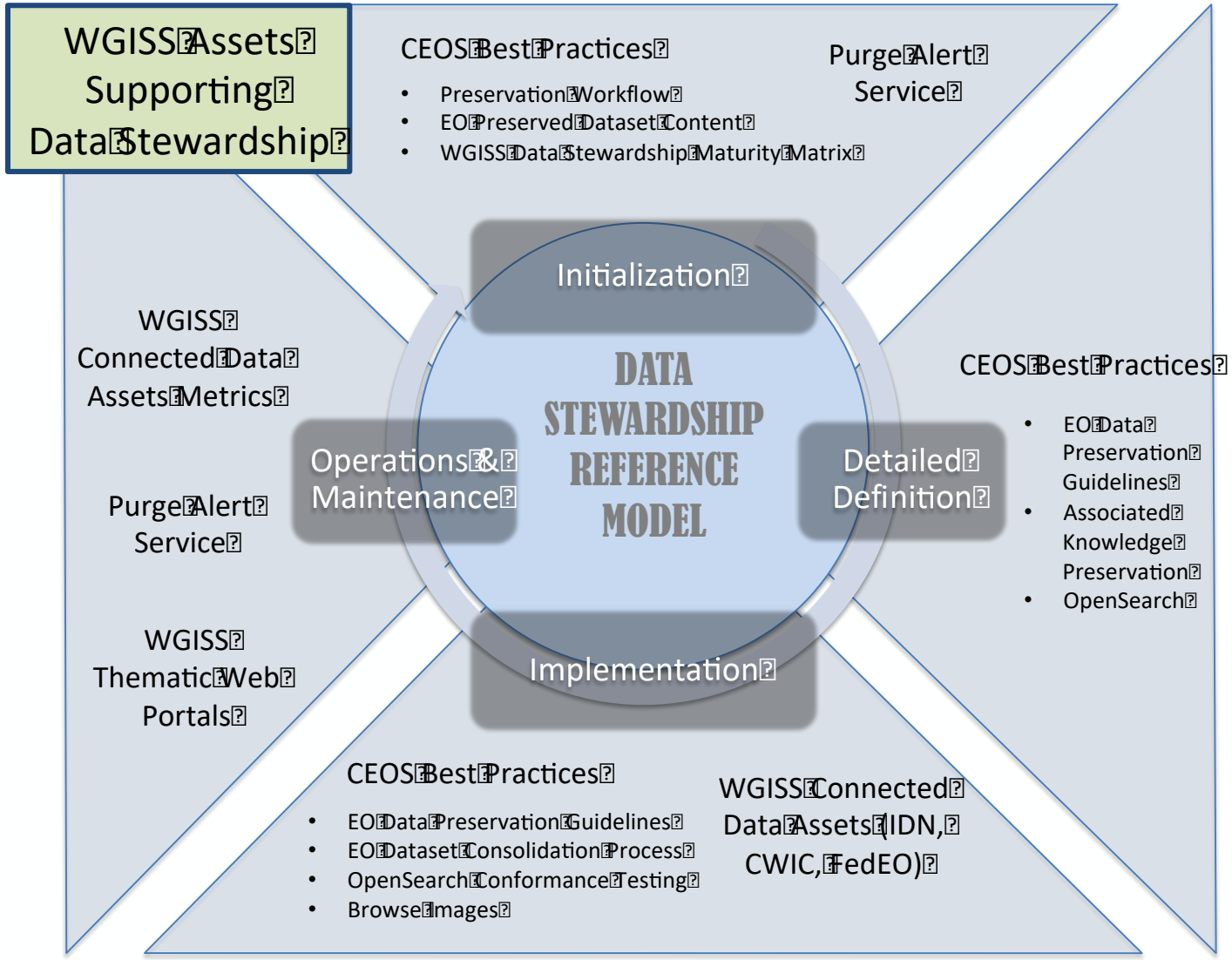
A large proportion of users are not domain experts anymore, therefore data discovery tools, documentation and support are also needed.

LTDP Best Practice Landscape

Data Stewardship Reference Lifecycle



ASSETS Supporting Data Stewardship



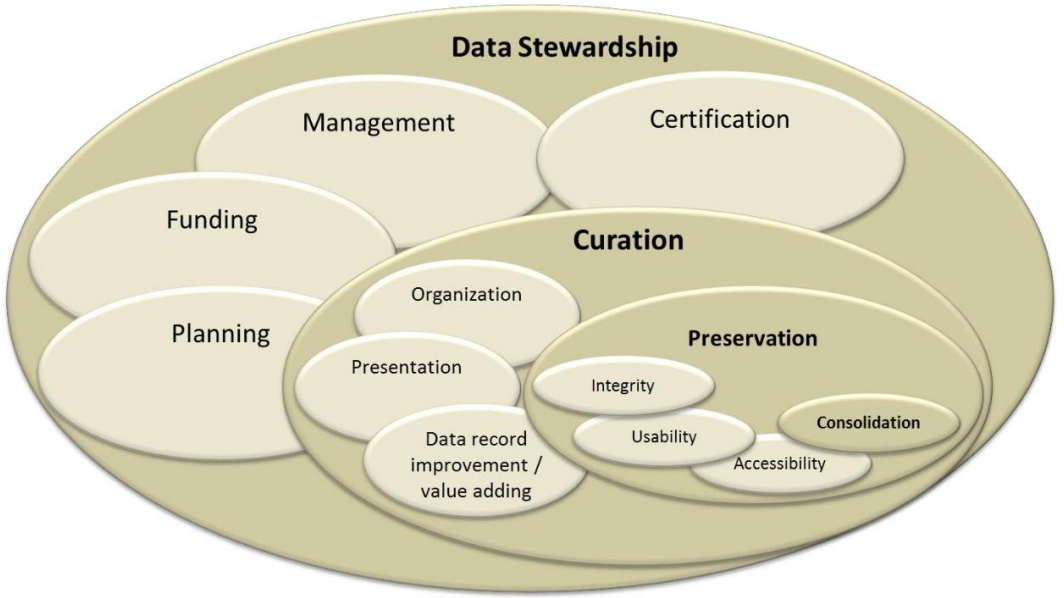
WGISS ASSETS	Initialization	Detailed Definition	Implementation	Operations & Maintenance
Best Practices and White Paper				
EO Data Glossary of Acronyms and Terms	▲			▲
Preservation Workflow	▲			▲
EO Data Preservation Guidelines	▲			
Preserved Data Set Content	▲			
Generic EO Data Set Consolidation Process	▲			▲
CEOS Persistent Identifiers Best Practices				▲
CEOS Associated knowledge Preservation Best Practices	▲	▲		▲
EO Data Purge Alert Procedure				▲
WGISS Data Management and Stewardship Maturity Matrix	▲	▲	▲	▲
Measuring of Earth Observation Data Usage	▲	▲		
WGISS DSIG Browse Guidelines Document images	▲	▲		
CEOS OpenSearch Documentation		▲		
CWIC Documentation		▲		
IDN Documentation		▲		
FedEO Documentation		▲		
FDA Documentation	▲	▲		
WGISS Connected Data Assets Metrics		▲		

Data Stewardship Concept

Data stewardship is the responsibility for planning, management, certification, and adequate funding for EO data sets throughout the mission phases and data life cycle.

Data curation consists of value adding, organization, presentation and preservation activities, which aim at establishing and increasing the value of EO data sets over their life cycle, at favoring their exploitation, possibly through the combination with other data records, and at extending the communities which are using the data sets.

Data preservation consists of actions on individual or multi-mission EO data sets with the goal to ensure their integrity over time, their discoverability and accessibility, and to facilitate their (re)-use in the long term.



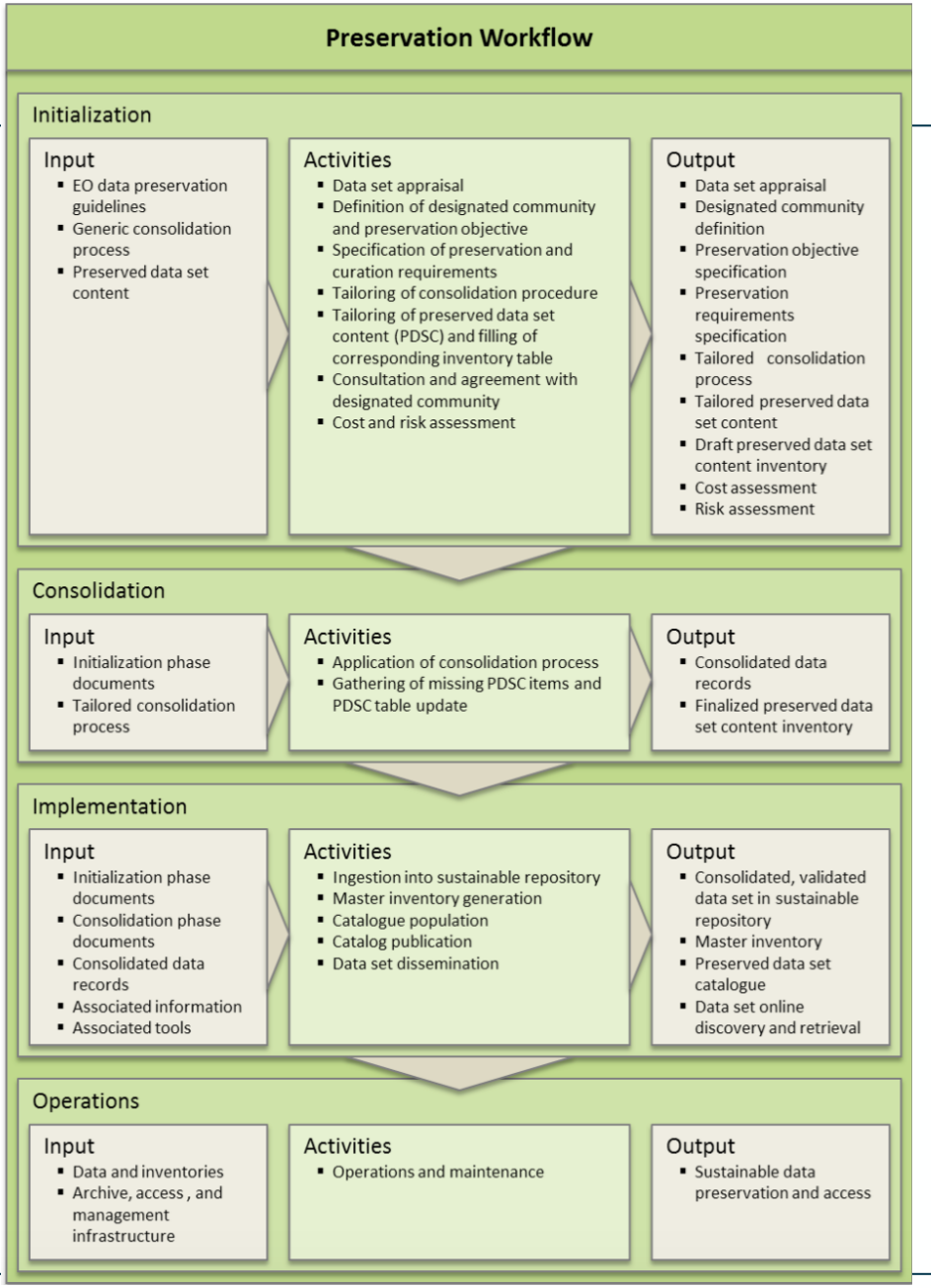
The core target of the Data Stewardship lifecycle is the preserved data set, composed of consolidated:

- ***Data records***: these include raw data, Level 0 data and higher-level products, browse images, auxiliary and ancillary data, calibration and validation data sets, metadata and descriptive information, in-situ data.
- ***Technical Content and Associated Information***: this includes all the processing software used in the product generation, quality control, the product visualisation and value adding tools, and documentation needed to make the data records understandable to the designated community. This includes information on the mission operation concept, product specifications, instrument characteristics, algorithm descriptions, Cal/Val procedures, mission/instrument performance reports, quality related information, etc. technical content and associated information is necessary to ensure data remains understandable and usable in time.

Preservation Workflow Best Practice

Preservation Workflow

The preservation workflow defines a procedure recommended to be applied to digital data for their preservation with the objective to optimize their reuse in the long term.



Preservation Guidelines Best Practice

Preservation Guidelines

The Preservation Guidelines document covers the planning and implementation steps of the CEOS Preservation Workflow. The guidelines should be applied to historic, current and future Earth Observation data sets. The document addresses technical and organizational aspects for the long-term preservation of EO data. It includes security, accessibility and usability aspects, recommending applicable standards and procedures.
Eight main “themes” consisting of “guiding principles” and a set of “guidelines” that should be applied to guarantee the preservation, accessibility, and usability of EO space data in the long term.

- ✓ Preserved data set content definition and appraisal
- ✓ Archive operations and organization
- ✓ Archive security
- ✓ Data ingestion
- ✓ Archive maintenance
- ✓ Data access and interoperability
- ✓ Data exploitation and re-processing
- ✓ Data purge prevention

Adherence Level	Description	Condition for adherence
Level A	Basic data security, integrity and access.	Implementation of all high priority guidelines.
Level B	Medium-level data security, integrity, access and interoperability.	Implementation of all high and medium priority guidelines.
Level C	Top-level data security, integrity, access, and interoperability.	Implementation of all guidelines.

Preservation DataSet Content Best Practice

Preserved DataSet Content

SCOPE:
This document identifies the EO mission data set assets content that should be preserved to ensure long-term usability and exploitation of Earth Science data.
The PDSC should be tailored appropriately for each mission/instrument. The tailoring of the PDSC should involve mission experts (e.g. instrument designers, quality working groups) and the designated user communities.

The tailored document should have a defined owner and should be kept under review throughout the mission lifecycle, at the end of each mission phase/stage.

The PDSC Best Practice has been also used as input to the definition of the ***"ISO/DIS 19165-2 Geographic information — Preservation of digital data and metadata — Part 2: Content specifications for earth observation data and derived digital products"*** standard.

Preserved DataSet Content

The PDSC BP follows the stages of the Earth Observation missions:

- **Mission Concept (MC):** Defines the mission to a sufficient level to show the scientific value and technical feasibility.
- **Mission Definition (MD):** This stage is concerned with the mission scientific requirements detailed definition and the selection of technical solutions for system concept.
- **Mission Implementation (MI):** According to Mission Definition results, this stage produces the detailed definition and implementation of the mission system and components.
- **Mission Operations (MO):** This stage identifies the operational timeframe of the mission being the period during which data are captured, algorithms are revised and improved, activities concerned with input analysis, calibration and validation of sensor/instrument as well as activities concerned with qualification of processed data are performed.
- **Post Mission (PM):** This represents the Post-Operations and Preservation stages which mainly focus on the archived data to accommodate the need to preserve them in the long term for further reuse and exploitation.

ID	Type	Identification	Description	Quality Information	Notes
MC 1.1	Doc	Scientific Scenario and User Communities	Defines scientific scenario and expected goals. Also list Principal Investigator, designated user communities and third party actors.	Required uncertainty for services and applications, lifetime, data availability, data accuracy, data latency, revisit time, geographical coverage, spatial resolution.	
MC 1.2	Doc	Mission Requirement Document	Defines scientific mission and sensor requirements, processing methods, qualification, methods	Calibration plan and quality assessment plan for the mission. Uncertainty requirements for instrument product (e.g. radiometric/geometric uncertainty, coverage, revisit time, etc.) Justification for the design decisions (e.g. band selection)	Most information should be contained in the mission documentation, e.g. the Mission Requirement Document (MRD), Mission Operations Concept Document (MOCD) and Mission Description Document (MDD) according to ECSS or equivalent standards.
MC 1.3	Doc	Mission Operation Plan	Defines the plan on how the mission will be conducted	Plan for handling quality information	

Technical Content and Associated Information Preservation Best Practice

Technical Content and Information Preservation

SCOPE:

Long-term accessibility and exploitability of Earth Science data requires that not only sensed data, but also technical content and associated information, needs to be properly preserved and made accessible.

Information technology is changing rapidly and this change also affects digital data from Earth Observation missions. On the other hand, insufficient documentation regarding the data, the inability to discover the data, or service compatibility can also prevent their re-use. Digital objects need a hardware and software environment in order to be managed.

This document aims to provide recommendations and best practices for the preservation of Space Technical Content and Associated Information.

Technical Content and Information Preservation

CONTENT:

Technical Content and Associated Information: this includes all the Tools used in the Data Records generation, quality control, visualization and valorisation, and all the Information needed to make the Data Records understandable and usable by the Designated Community.

To follow an excerpt:

The existing and possible preservation format of Information Preservation are listed below:

- ✓ Text documents (often MS Word, Excel Files, txt, etc.) can be preserved as:
  - PostScript, PDF/A, DSSSL, RTF, ASCII, SGML, TIFF, CGM
- ✓ Images (JPEG, TIFF, PNG, FITS, etc.) can be preserved as:
  - Loss of Quality JPEG, JPEG2000
  - Lossless compression TIFF, PBM, PNG, FITS.
- ✓ Metadata (ASCII, XML, SGML, etc.) can be preserved as:
  - ASCII, the most durable format for metadata because it is widespread, backwards compatible when used with Unicode (superset of ASCII), and utilizes human-readable characters, not numeric codes.
  - For higher functionality, SGML or XML should be used.
- ✓ Multimedia (AVI, QuickTime, MPEG, WMV, MJ2) can be preserved as:
  - MJ2, MPEG-A, Mp4.
- ✓ Physical (3D) can be preserved as:
  - Scan 3D format.

Generic EO Data Set Consolidation Process

Generic EO Data Consolidation Process

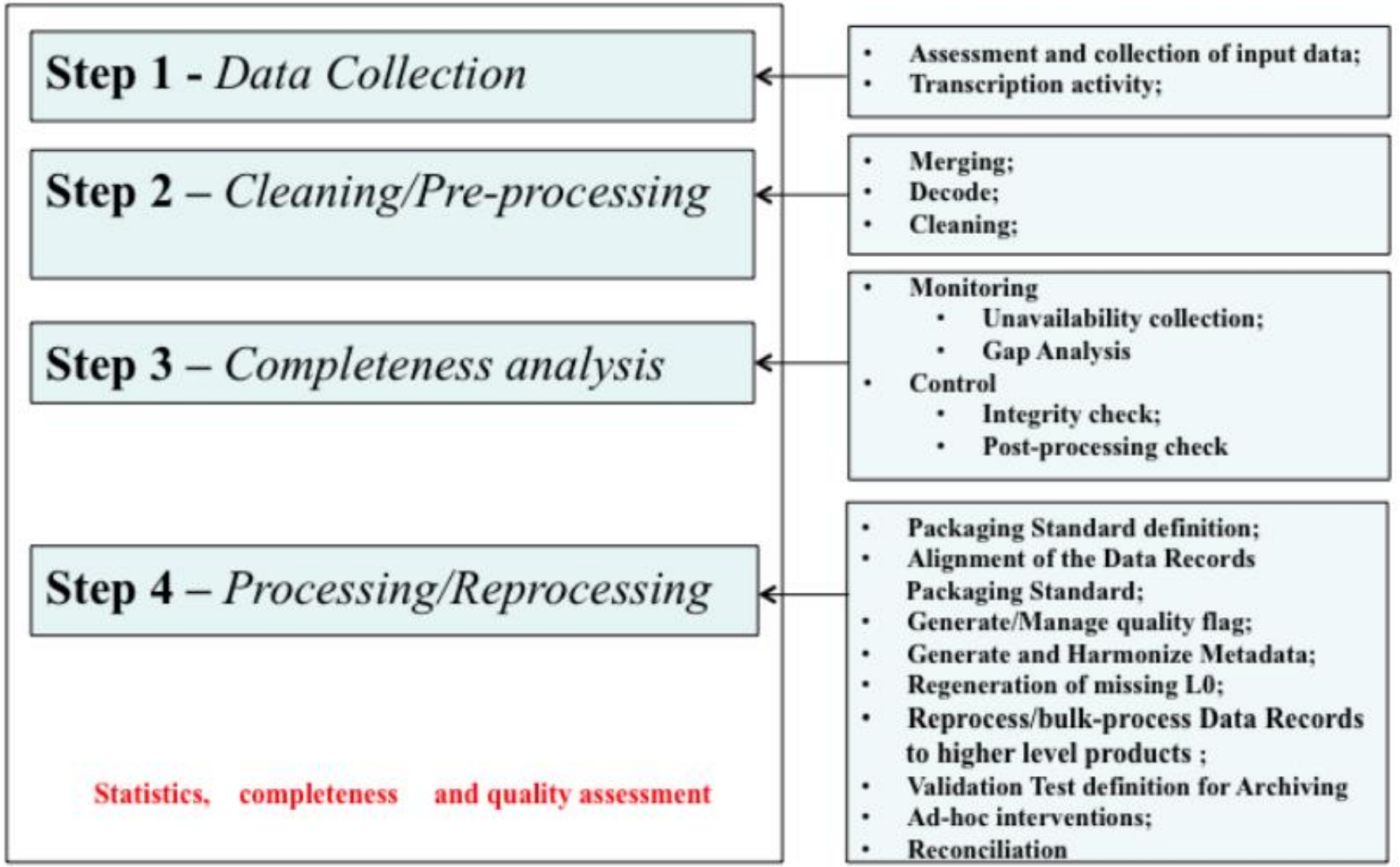
SCOPE:

This document represents the Generic EO Consolidation Process and it is intended to be used as input to a process of the Preservation Workflow BP, to produce the mission-specific consolidation process.

It consists of a series of recommendations and advice focused on the implementation of actions for the *consolidation* of the Data Records and their Associated Information, for a given mission.

These recommendations are meant to be used as guidance for the mission requirement definition, ground system implementation, data centres operations services, for the preservation of their data holdings.

Generic EO Data Consolidation Process



Persistent Identifier Best Practice

Persistent Identifier

SCOPE:

The main purpose of the Persistent Identifiers Best Practice is to help data providers in ensuring unique identification of their datasets (with related benefits in terms of data integrity and provenance), and users in citing and finding specific datasets. In this context the Earth Sciences and Earth Observation mission data identified objectives and needs are listed below:

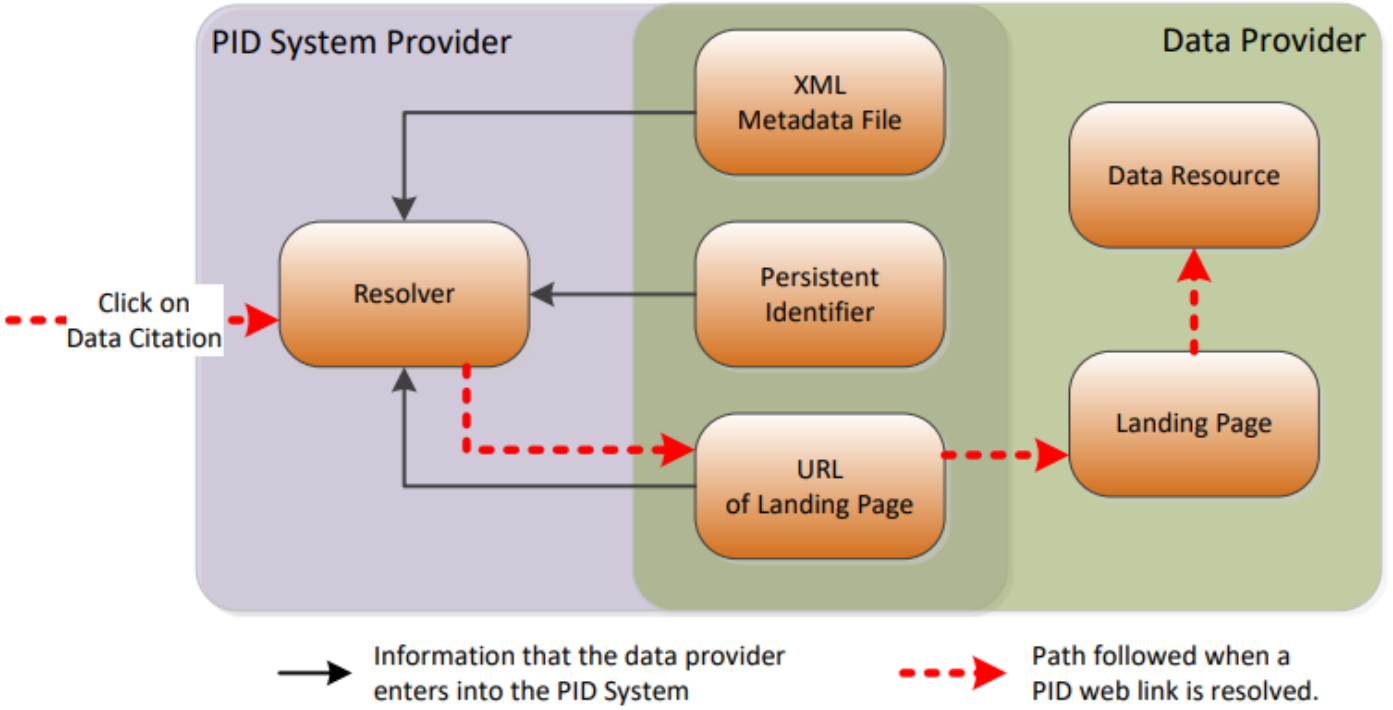
Objectives & Needs

- Globally unique, unambiguous and permanent identification of a digital object for locating and accessing over time
- Improve discoverability and accessibility
- Enable users to retrieve objects without knowing their location
- Enable repositories to change the location of objects internally
- Enable repositories to share objects with other services where appropriate
- Enable researchers to cite digital objects consistently over time, which also benefits data holders
- Increase data visibility and use
- Increase credibility and value of data holdings.

Persistent Identifier

CONTENT:

This document aims to provide recommendations and best practices on the use of Persistent Identifiers (PIDs) to Earth Observation mission data, allowing globally unique, unambiguous, and permanent identification of a digital object. Few relevant use cases are also in the best practice.



Measuring EO Data Usage Best Best Practice

Measuring EO Data Usage

SCOPE:

Metrics and indicators have been historically collected by data owners/providers to gather relevant information on data usage, to generate statistics, stimulate user engagement, and to monitor processes and services. In the past, data providers were performing this independently, without coordination.

Today, the evolving landscape in Earth Observation (EO) data usage, with the arrival of new technologies and the Big Data paradigm (e.g. bringing users to the data as a complementary approach to data download) allows for more powerful statistics and analysis.

This document provides recommended parameters/metrics/indicators to be used, together with relevant information to be collected by data providers.

Metric Code	Description	Parameters to be captured	Difficult Rating	Rationale
MET_EODO_01	Mission/Sensor/Product Level size of data downloaded	Size of data downloaded per Mission/ Sensor/Product Level	★	• User needs analysis • User community interest in the data offer • Verification and validation of data if none is downloaded anymore • New reprocessing campaigns for data with few downloads • Top ten missions and sensors data requested
MET_EODO_02	Mission/Sensor/Product Level number of files downloaded	Number of files Downloaded per Mission/ Sensor/Product Level	★	• User needs analysis • User community interest in the data offer Verification and validation of data if none is downloaded anymore • New reprocessing campaigns for data with few downloads • Top ten missions and sensors data requested
MET_EODO_03	Temporal distribution of missions and sensors data	Number and/or size of mission/sensors data	★	• Top ten data

Glossary of Acronyms and Terms

Glossary of Acronyms and Terms

SCOPE:

This document provides a list of definitions for frequently used acronyms and terms in the field of Earth observation data stewardship. The main goal is to align the Glossary and terms between various sources and Agencies.

Term	Definition	OAIS Equivalent
Access *(noun)*	Services and functions which make the stored information holdings accessible to users by providing data search, discovery, retrieval, and dissemination functions. Access can refer to either the functionality, the services providing the functionality, or the entity providing the corresponding services. • Off-Line - Access to information by mail, telephone, facsimile, or other non- direct interface. • Near-line - On-line access to information or data with system related time delays, e.g. resulting from data retrieval from tape library and / or CPU-intensive 'on-the-fly' product generation. • On-Line - Access to information by direct interface to an information data base via electronic networking. Access includes data search, discovery, and retrieval	The OAIS functional entity that contains the services and functions which make the archival information holdings and related services visible to Consumers.
Acquisition *(noun)*	Acquisition describes the complete process from optional on-board recording, downlink and reception, up to the reconstruction of instrument source packets on ground. See also acquisition planning.	
Acquisition Planning *(noun)*	Computation of a non-conflicting timeline of activities for the space segments and for corresponding reception activities of the stations. The planned activities comprise recording, downlink and reception. Planning has to take into	

Conclusions

Data holdings are growing exponentially in Earth Science data archives worldwide. Only a systematic and standardised approach to data preservation, management and curation during the entire data lifecycle, coordinated between data holders and user communities, will ensure that these data sets will be accessible and useable to current and future generations of users for monitoring long-term variations in environmental parameters as a basis for objectively assessing and predicting the effects of global change.

Preservation, management, and curation of Earth observation data and information acquired from space, should be addressed during all phases of Earth observation missions – from the initial mission planning, throughout the entire mission lifetime, and during the post-mission phase. In the frame of CEOS WGISS the main objective of the Data Preservation and Stewardship Interest Group is to draft common cross-agency best practices or guidelines of data stewardship for possible adoption by the interested organisation.

All best practices can be accessed freely and openly on the CEOS web site (https://ceos.org/ourwork/workinggroups/wgiss/preservation/)


thank you
danke
謝謝
ngiyabonga
teşekkür ederim
gracias
merci
obrigado
dziękuję
bedankt
go raibh maith agat
sukriya
kop khun krap
terima kasih
grazie
arigatô
dakujem
tapadh leat
mochchakkeram
dank je
спасибо
рахмат
Баярлалаа
faafetai lava
vinaka
спасибі
blagodaram
misaotra
matondo
paldies
grazzi
mahalo
хвала
asante
manana
obrigada
tenki
murakoze
chokrane
mamnun
дякую
djiere dieuf
tau
akun
dankon
aëitü
chnorakaloutioun
gratias ago
gràcies
sulpáy
taiku
takk
dakujem
trugarez
мерси
merci
shukriya
dhanyavadagalu
diolch
xièxie
감사합니다
ευχαριστώ
tanemirt
rahmet
najis tuke
ありがとう
didi madloba
kam sah hamnida
তোমাকে ধন্যবাদ
rahmat
sagolun
mèsi
sobodi
dēkuji
enkosi
bayarlalaa
gràcie
hvala
mauruuru
köszönöm
kiitos
dankie
dhanyavad
nanni
nandri
barka
welalin
tack
mersi
kia ora
spas