GHRSST DAS TAG session

Propositions for changes

outline

- Revisions to GHRSST implementation
  - GDS format
  - List of metadata (NetCDF attributes)
  - Missing format specifications for some product levels – geostationary datasets
- Organisation : a new R/GTS

New metadata netCDF attributes

Issues with attributes

- Some standards have been enriched with new attributes that improve product content (**ACDD)** description: https://podaac.jpl.nasa.gov/PO.DAAC_DataManagementPractices#Metadata Conventions
- Part of some agencies requirements and brings more consistency with other (non-GHRSST) datasets

Constraints

- The new attributes **should not be mandatory** (to preserve validity of current products), just **recommended**
- The new attributes should not break anything in current GDS, os that files in revised format can be read without changing anything to current software code:
  - No change in definition or content of existing attributes
  - No removal of existing attributes
  - Only exception may be some obsolete and redundant attributes (but should be assessed carefully)
- « **add** » rather than « **delete** » or « **replace** »

New proposal for attributes (Ed Armstrong, PO.DAAC)

Proposed metadata revisions

Placeholder for Excel Spreadsheet
for metadata revisions (ACDD_VersionChange_v6
_tmp.xlsx)

GDS product format

Is it L2 or L3 or L4 ? Ambiguities on some datasets

- **L2 with downgraded resolution** (ex : VIIRS with 1500m resolution)
  - Resampling/regridding could semantically be L3 but here we are still in swath projection
  - Possible confusion with fixed grid L3U / L3C / L3S products
  - **Suggestion** : **to keep L2 for any LEO along-track dataset as swath products are more associated with « L2 » concept from user perspective**

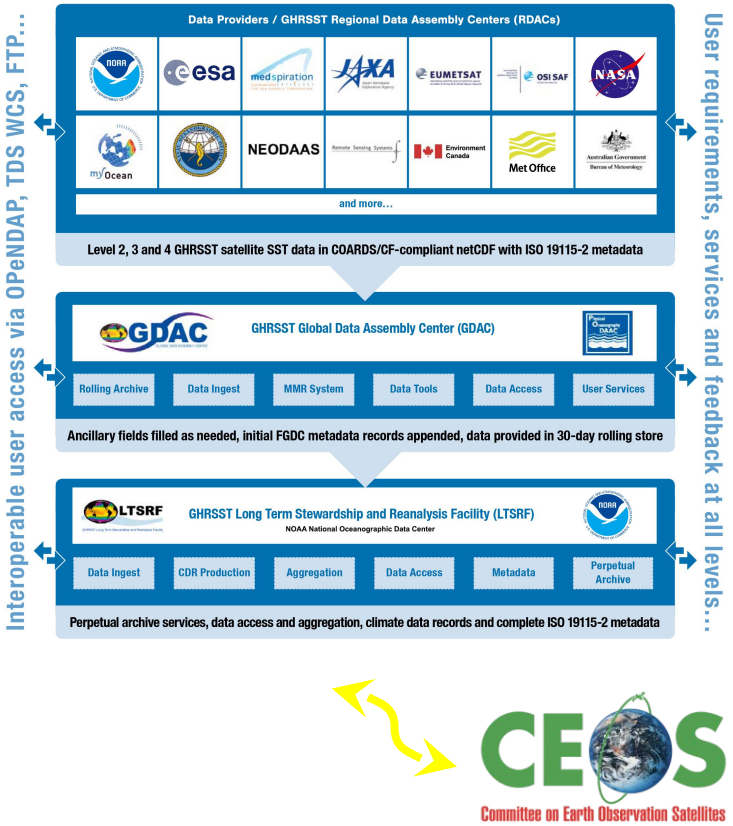
- **L2 with gap filling**
  - Interpolation or any method used for filling cloud covered pixels could semantically be called L4
  - Risk of confusion with fixed grid L4 analysed products
  - **Suggestion** : **to keep L2 for any LEO along-track dataset**

- **Geostationary products : L2 or L3 ?**
  - GDS is ambiguous on whether geostationry products wil L2P type of content are L2 or L3
  - **Un-collated (L3U): L2** data granules remapped to a space grid without combining any observations from overlapping orbits
    - Definition matches single geostationary snapshot (full temporal resolution, one image) of geostationary except for « L2 » (as processing is done from L1)
  - **Collated (L3C):** observations combined from a single instrument into a space-time grid
    - Definition matches composite geostationary products (ex : hourly with some merging, downgraded temporal resolution)
    - Possible confusion with some L3C products (from L2)

  - **Suggestion** : geostationary products, being gridded product on a fixed grid, are from user perspective more associated with « L3 » concept. Use L3U and L3C product type for geostationary products depending if they are single or coposite images.

GDS needs to be cleaned and clarified with respect to the changes we will decide.

GHRSST R-GTS Framework: Today



All “official” data flow from RDAC to GDAC to LTSRF

Data is accessible at all levels

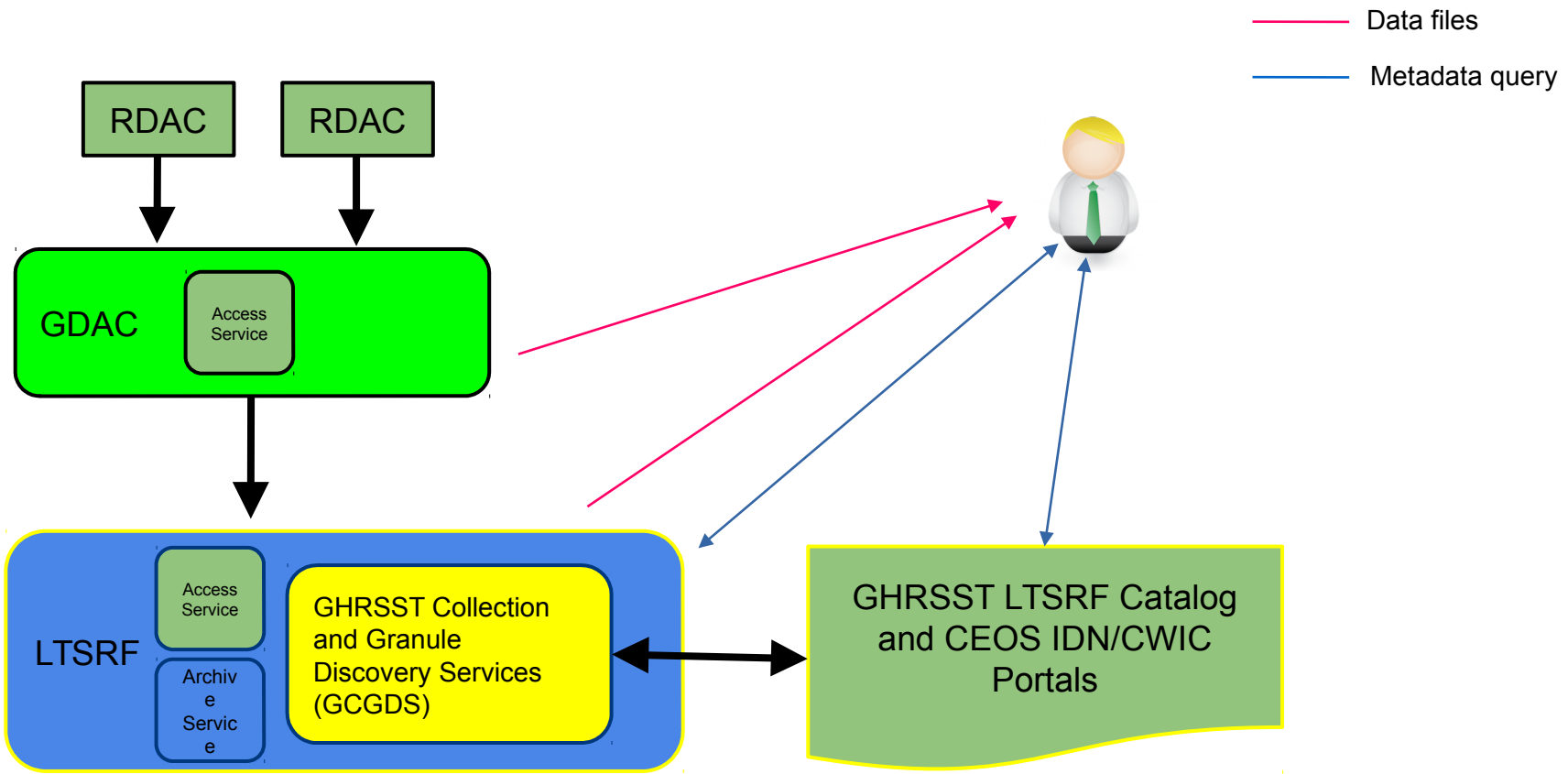
RDACs free to do whatever they like, as long as they submit GDS-compliant data to GDAC

Metadata “grows as it flows” from one level to the next

Considered highly successful, and nothing is “broken”

LTSRF publishes collection metadata to CEOS IDN, and supports CWIC granule searches via CSW and OpenSearch

Existing GHRSST R-GTS Framework



In the existing R/GTS framework, users can access GHRSST data from RDACs, GDAC, and LTSRF. GDAC has the most comprehensive metadata catalog. LTSRF's catalog is close, less the most recent 30 days for most products. LTSRF has the most comprehensive store of data files. Note GDAC at JPL also provides catalog to CEOS via the NASA CMR.

R/GTS revision

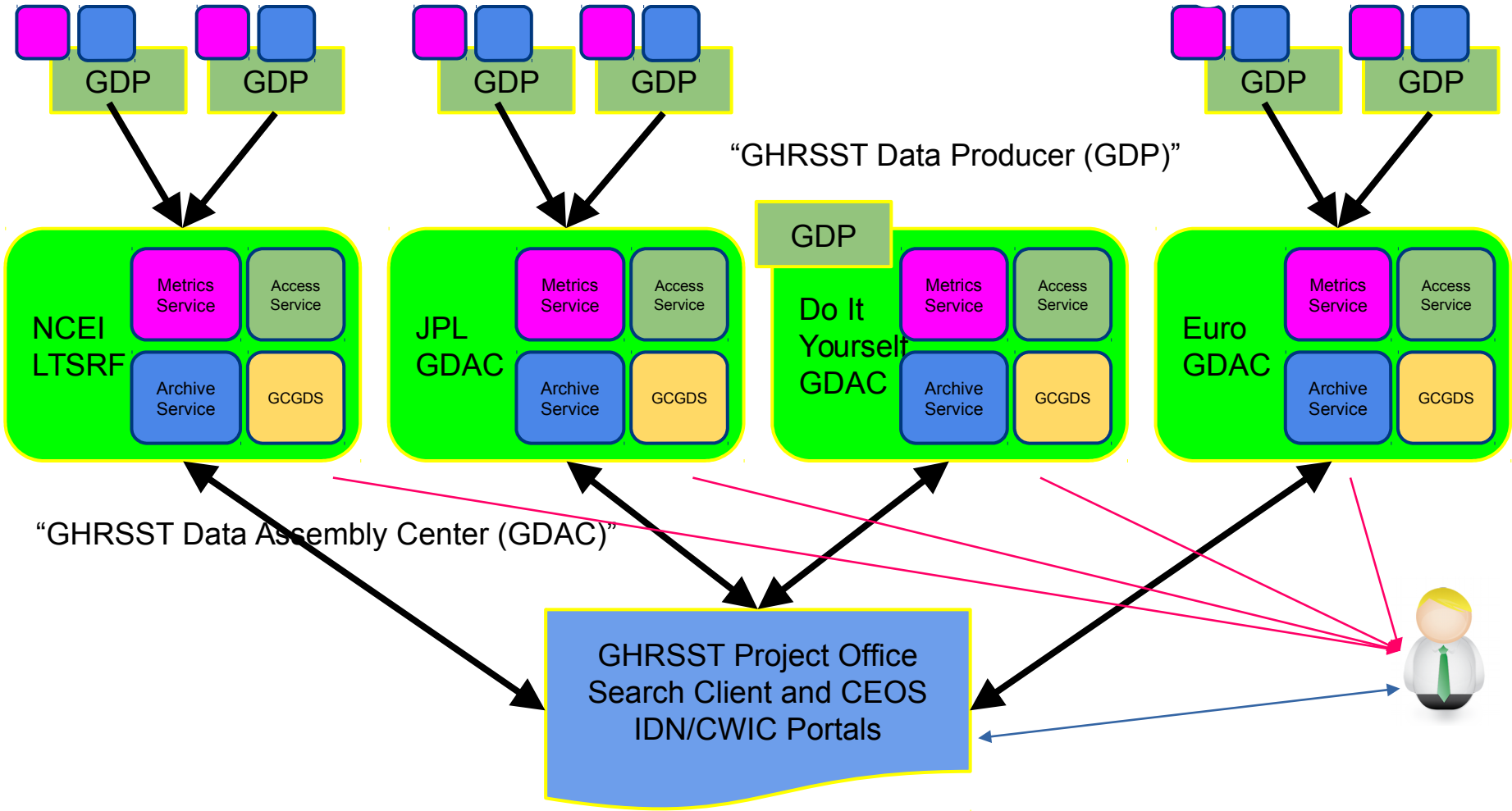
Issues with current GHRSST system

- Current GDACs or LTSRF don’t archive all GHRSST products or won’t in the future
  - Infrastructure and cost limitations
  - Focus on dataset of interest for their own usage
  - Complex data exchange procedures or monitoring requiring frequent interactions with data providers
  - Some agencies may wish to be sole access point and archive for their products (data policy, user registration,...)
- Difficult for central repository help desks to support users on products they are not responsible for
- Conceptual weakness as it is dependent on funding of the agencies performing this central services
- « Physical » central repository of products concept may be somewhat obsolete or not bringing much advantages – single access point for users is the main benefit and can be « logical »

Suggested changes

- More distributed system where different “GHRSST DACs” or data producers themselves (GDPs) ensure data dissemination and archiving functions
- Central portal for data discovery and inventory search

Proposed Future R-GTS Framework



GCGDS = GHRSST Collection and Granule Discovery Services

ISO 19115-2 collection level record
OpenSearch for granules

Archive Service

Long term preservation at an established facility conforming to tenets of the OAIS Reference Model

Access Service

Minimum access would include ftp/http, plus TDS. TDS possibly optional.

Metrics Service

User access metrics. Consider DataONE API as candidate.

Proposed Future R-GTS Framework

A key to this overall idea is that users would be directed to the central GPO or CEOS catalogs, where all GHRSST data, no matter where it resides, could be discovered. When access is initiated, those central catalogs provide the granule data access links to the data files at the appropriate repository

The GPO would need to establish a verification capability to ensure all components provide reliable services

User metrics services would be very simple at first, focused on data volumes, files, and numbers of users. GPO would aggregate these numbers.

Services to be implemented

Metrics Service	• For System: GPO uses https://statuschecker.fgdc.gov to monitor set of agreed-upon endpoints at GDACs • Each GDAC provides monthly aggregated number of files, unique IP addresses, and volumes served per access method	Access Service	• Http (https) • Data Access Protocol (DAP) • Ftp (sftp) • WMS and WCS for L3 and L4 data
Archive Service	• Each GDAC (Archive Services) provides a written response to a short document template about how they meet OAIS Reference Model responsibilities functional entity areas (Ingest, Archival Storage, Access, etc.)	GCGDS	• ISO 19115-2 collection level record for each GHRSST product, submitted to CEOS IDN • Granule search endpoint meeting OpenSearch CWIC specifications

New list of GDACs and related GDPs

Copernius S-3 GDAC/GDP (EUMETSAT)

JAXA GDAC/GDP

CMEMS GDAC (CNR)

- UKMO
- GOS
- DMI
- METNO
- Ifremer
- Meteo-France

GHRSST GDAC (JPL PO.DAAC)

- REMSS
- JPL
- JPL_OUROCEAN
- NAVO
- JAXA
- CMC

GHRSST LTSRF (NOAA NCEI)

- OSPO
- STAR (future RDAC)
- NCEI
- ABOM
- UFRJ

GHRSST GDAC (IFREMER)

- OSISAF
- Medspiration
- RMSS/NAVO/OSPO (*)

(*) Note: each GDAC could also acquire whatever other data they need.

R/GTS : do we agree on… access services for NRT products ?

- **Mandatory** : http access to data folders
  - Allowed alternatives : https
- **Strongly recommended** : FTP
  - Allowed alternatives : SFTP
- **Strongly recommended** : DAP (OPeNDAP)
  - Allowed implementations : Hyrax, Thredds
- **Recommended** : WMS/WCS for L3 and L4 products

R/GTS : do we agree on… archiving services ?

What is archiving service ?

- Long-time preservation of datasets (redundant backup) ?
- User access to complete time series ?
- Main requirements of OAIS reference model : (from https://en.wikipedia.org/wiki/Open_Archival_Information_System)
  - Key components:
    - Archival Storage
    - Preservation Planning

All GHRSST datasets have to be preserved for long term

- All versions ? Latest ?
- Long-term backup ? who's doing what ?
  - Is it GDAC or GDP responsibility or some resp. shared between both ?
  - A data producer (GDP) has to commit to the archiving of its GHRSST datasets
  - If it does not have this capacity, it should be done in relation with an attached GDAC

Access to complete time series must be implemented following same data access requirements as for near real time products (GDAC resp.)

What is a OpenSearch queryt

Open Search Queries examples

The [query] in the open search URI will follow the same syntax used in the full text search. Here below we provide some examples.

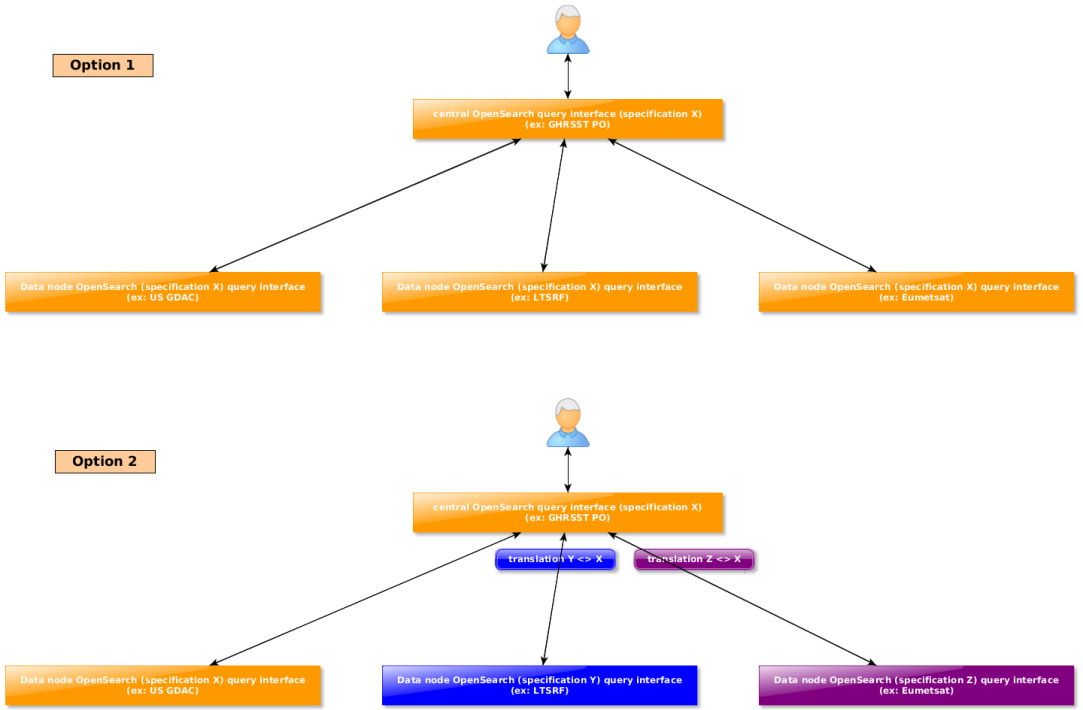
Example	Open Search
Searches every product with OL_1_EFR___ product type or products containing the string "OL_1_EFR___" in the metadata.	https://coda.eumetsat.int/search?q=OL_1_EFR___
Search every products ingested in the last day	https://coda.eumetsat.int/search?q=ingestionDate:[NOW-1DAYS TO NOW]
Search every products ingested in the last month	https://coda.eumetsat.int/search?q=ingestionDate:[NOW-30DAYS TO NOW]
Search every products ingested in the last hour	https://coda.eumetsat.int/search?q=ingestionDate:[NOW-1HOUR TO NOW]
Search every products having sensing in the last three months	https://coda.eumetsat.int/search?q=beginPosition:[NOW-3MONTHS TO NOW] AND endPosition: [NOW-3MONTHS TO NOW]
Search every products delimitated by the polygon vertices: -4.53 29.85, 26.75 29.85, 26.75 46.80,-4.53 46.80,-4.53 29.85	https://coda.eumetsat.int/search?q=footprint:"Intersects(POLYGON((-4.53 29.85, 26.75 29.85, 26.75 46.80,-4.53 46.8 4.53 29.85)))"

https://coda.
/search?q=fo
4.53%2046.80

Do we agree on … GCGDS services ?

- CSW : for dataset discovery
  - A lot of existing services in place : how interoperable are they ? Is ISO 19115/ISO 19139 compliance enough ?
  - Do we need GHRSST specification requirements ?
  - Or « translation » services ?
- OpenSearch : for granule inventory crawling
  - OpenSearch does not normalize vocabulary and interoperability is not ensured by nature
    - but recommended OpenSearch protocols from CEOS and ESIP exists
  - Several implementation existing out there with no straight forward interoperability
  - Some on-the-shelf and/or open source software available to implement this
  - Choice of a standard for queries and result format should consider availability of software for implementation
  - Other alternative than OpenSearch ?
    - OpenData (seems to have the same issues)
    - THREDDS inventory capability ?

How do we implement GCGDS ?



Option 1, 2, 3 : federated query system

A single access portal for search & metadata queries

Each query rooted to equivalent service at connected GDACs

Results assembled at GHRSST-PO portal and returned to user

implementation	Pros and cons
Option 1 GHRSST-PO is central portal for federated queries Every GDAC implements the exact same CSW and OpenSearch services	CSW and OpenSearch services are already operated at different GDACs These services may not be interoperable and to make this work, redundant services (complying to GHRSST requirements) must be implemented. This solution does not take advantage of existing services already in place.
Option 2 GHRSST-PO is central portal for federated queries GHRSST-PO query system implements the interoperability layer and ensures the translation to each GDAC CSW and OpenSearch services	Takes advantage of services in place Continuous implementation effort required at GHRSST-PO level to maintain overall system interoperability GHRSST-PO needs to backed by supporting agency for hosting the service
Option 3 Same but central service implemented at NASA	Already implemented for GHRSST data that flows to JPL GDAC and NOAA LTSRF Commitment to implement gateways to other non US services ? On which resources ? Which limitations ?

How do we implement GCGDS ?

implementation	Pros and cons
Option 1 GHRSST-PO is central portal for federated queries Every GDAC implements the exact same CSW and OpenSearch services	CSW and OpenSearch services are already operated at different GDACs These services may not be interoperable and to make this work, redundant services (complying to GHRSST requirements) must be implemented. This solution does not take advantage of existing services already in place.
Option 2 GHRSST-PO is central portal for federated queries GHRSST-PO query system implements the interoperability layer and ensures the translation to each GDAC CSW and OpenSearch services	Takes advantage of services in place Continuous implementation effort required at GHRSST-PO level to maintain overall system interoperability Not easy task for GHRSST-PO unless backed by some agency
Option 3 Same but central service implemented at NASA	Already implemented for GHRSST data managed by JPL GDAC and LTSRF Commitment to implement gateways to other non US services ? On which resources ? Which limitations ?
Option 4 Not a federated system : just a central repository for dataset and granule level metadata with search/discovery services (CSW and OpenSearch)	Probably more robust and responsive solution from user perspective Simpler implementation Operation requires more monitoring and interaction with GDACs to ensure the metadata continuously flow toward the central repository.

Do we agree on … metrics services

- What metrics ?
  - Status of access services
    - What does that mean ? Responding URL ? Still populated with data ? ….
  - Data usage metrics
    - Number of unique IP, data volume, number of files
      - Not easy to monitor for all services : OK for HTTP/FTP but what about THREDDS, OpenDAP, WMS/WCS services
      - Implementation may be complex
      - Probably numerous ways to do that (ex : number of unique Ips) : difficult to maintain consistency
- Suggested checker : https://statuschecker.fgdc.gov
  - Only a handful of services monitored (and not FTP or HTTP, OpenSearch)
  - Dependency on third party service
- I think it is possibily time and resource consuming, may be not mature enough, and should be later priority (effort should be on other access/archive/search services)