Minutes v1.0
LSI-VC-10 Teleconference #4: Further Topics, Discussion Time and Wrap-up

20 May 2021

Participants

Catalyst (PCI):	Wolfgang Lueck
CEO:	Marie-Claire Greening
CSA:	Paul Briand
ESA:	Ferran Gascon, Frank Martin Seifert
EC:	Zolti Szantoi, Peter Strobl
GA:	Adam Lewis
IEEE:	Chris Durell, Brandon Russell
JAXA:	Takeo Tadono, Ake Rosenqvist
LSI-VC Sec:	Matt Steventon, Libby Rose, Stephen Ward
NASA:	Bradley Doorn, Jim Irons, Diane Davies
NOAA:	Kevin Gallo
SEO:	Brian Killough
UK Catapult for UKSA:	Electra Panagoulia
USGS:	Steve Labahn, Chris Barber, Chris Barnes, Tim Stryker, Thomas Cecere, Steven Covington

The presentation slides compiled for this meeting are here and attached in Appendix A.

Introduction

Matt Steventon (LSI-VC Secretariat) welcomed everyone to the final teleconference of the virtual LSI-VC-10 meeting. This call provides time for further topics, discussions and wrap-up. We will hear about the lessons USGS learned regarding Landsat Collection 2 implementation in the cloud, before opening the floor for discussions on requirements and global grids – with a scene-setting presentation from Peter Strobl (EC).

USGS Lessons Learned Regarding Landsat Collection 2 Implementation in the Cloud

Steve Labhan (USGS, LSI-VC Co-lead) presented on the Landsat Cloud project and some of the lessons learned that may be useful to other agencies attempting similar projects. The main goal of the project was to modernize processing, access and distribution of the full Landsat data archive, by establishing an enterprise cloud environment for Landsat.

The Cloud Optimized GeoTIFF (COG) format was selected for the project, and the team worked closely with the visualisation team to standardise the COG parameters. The SpatioTemporal Asset Catalog (STAC) is a new collaborative standard for managing access metadata.

The processing of Landsat Collection 2 (Level 1 and Level 2) took just one month, with an average of 293,000 scenes per day. This is in comparison to the 18 months to process Collection 1. On-premises production for Landsat 8 would typically have a maximum rate of 25,000 scenes per day.

Steve also demonstrated the capabilities of the Landsat STAC Browser, a web-based GUI that allows users to browse and review STAC records and thumbnails for Landsat data in the cloud. Another platform from USGS is SAT API, which allows users to programmatically search the Landsat data in the cloud for products within their spatial and temporal AOIs, and returns links to objects that the user can then incorporate into their processing workflow.

Some of the lessons learned by USGS through this process were:

- On-demand database scaling allowed for increasing read replicas and instance sizes during scale testing in production;
- Migration from Dynamo DB to Aurora Database architecture and design (scenes_db, catalog_db) made a big difference in performance and cost;
- Redshift tool helped with post processing data validation efforts for Collection 2;
- Separate accounts for development, system test, and production helped to organise the efforts;
- Continuous Integration Continuous Deployment (CICD) processes work well;
- GitLab was invaluable as a repository (EROS Enterprise Asset);
- AWS cloud storage infrastructure works very well for data distribution;
- Security groups used to control/limit access to services and applications worked well;
- Ability to scale processing has been proven to be cost effective;
- Batch processing has limitations – AWS recommends migrating to Kubernetes;
- Contention with resources in the same account (batch, step, spot instances shared between processing and inventory management);
- Multiple accounts need an enterprise approach to managing the accounts, push out updates, monitoring, etc.;
- Separate processing and distribution into separate accounts;
- Two teams independently developing systems/processes and competing for AWS service limits.

Steve noted that processing for Landsat collection 2 was 300-350k USD. Storing data costs approximately 30-40k USD a year.

Discussion

- Adam Lewis (GA, LSI-VC Co-lead) asked about the team's approach to avoiding vendor lock in. Steve responded that a tool-by-tool analysis was performed to address this question. The team tried to avoid writing any code that would be specific to AWS wherever possible (meaning most should be transferable to other platforms, e.g., Google, Azure). They found that customising around Amazon's tools was more cost effective than trying to be 100% vendor agnostic. The task was a balancing act, but one of the lessons learned was that a vendor agnostic approach is costly. However, the biggest risk area is the visualisation area, where the team needed to take advantage of the Amazon tools. Other areas may need rearchitechting if they might be moved in the future.
- Adam asked if USGS keep tapes of the Collection 1 and 2 data. Steve believes this is also kept at EROS, along with the raw data backups.

- Egress cost for the whole AWS Collection was in the order of low millions (USD).
- The figure on Slide 12 indicates the trend towards downloading data from the cloud. Data is also collected on how many people are accessing and manipulating data in the cloud, in comparison to just those who download data. Most use cloud direct access (where customer pays for the egress limiter) rather than platforms like EarthExplorer, however the data can still be accessed through those platforms. The egress limiter put in place by USGS could limit the performance when accessing data through mirror platforms, hence direct access is faster and often preferred. Steven Covington (USGS) noted that initial issues with high-volume access by machine-to-machine have now been resolved. Direct access is faster, because of the egress limiter put in place by USGS. Platforms such as Google Earth Engine, Microsoft Azure, etc. likely negotiate special rates for direct access making it more affordable.
- There has been a trend towards people moving from Collection 1 data to Collection 2 data. With Collection 2, options for download have changed (COGS, partial scenes, split by bands, etc.), which changes the standard metrics used in the past. Hence, doing direct comparisons between the download volume now and in the past is complicated.
- Paul Briand (CSA) asked whether the number of users has increased since the move to cloud. Steve noted that it is tracked, however the team wasn't certain on the trend. Likely to not be a big increase in users, if there was one. The main metric of interest is the data volume downloaded.
- Adam questioned to what extent knowledge on these systems and approaches being taken is being routinely shared by members of groups (USGS, GA, DE Africa, etc). Steve noted that this is probably not happening in a structured way at the moment, and is something the teams should do. In WGISS there is some sharing of information, but more technical discussions are probably not happening. There were some hoops that USGS had to jump through regarding the commercial tools provided by Amazon compared to the customised ones their team had to use. This is something that other agencies might face when working in a commercial cloud environment, and this could be good information to share for awareness.
- Raw data to Level 0 processing is done on premises at USGS EROS. Everything done in the cloud is Collection focused. Level 0 processing only has to be done once, and reprocessing is not likely to be needed.
- The move to the cloud was done to keep up with accelerating data volumes and scope of analyses to be undertaken (global, deep time series, etc.). It is no longer feasible for users to download data and process on premises.
- Cloud processing is much cheaper than on premises processing. There would also be an initial outlay required for hardware that would only be used for this task and not again for 3-5 years. This would come with various issues with space and buildings that would have been hurdles as well. Using the cloud solved all these issues.
- Maintaining stability for users as cloud platforms change and minimising disruption could be a challenge. USGS has set things up to be transferable from vendor to vendor, however they can't have the data over multiple platforms due to contract restraints. Hopefully in the future more flexible contracts can be negotiated.
- Ake Rosenqvist (JAXA) commented in chat: "*On STAC - the STAC community have developed CARD4L extensions for STAC, one for optical (SR and ST) and one for SAR (NRB and POL). They have sought*

contact with the CARD4L SAR team (via Fang Yuan/GA) asking for help to refine the specs. More info here: https://github.com/stac-extensions/card4l The SAR team will be working with these people to help refine extensions for NRB and Polarimetric CARD4L products, but they are also interested in following up on the optical products. USGS is already participating on Landsat specifications."

- It was agreed to put STAC matters on future LSI-VC agendas.

LSI-VC-10-17	Ake to share the email he has received from the STAC development community. LSI-VC team to consider potential contributions. Matt to include a STAC item on future LSI-VC teleconference agendas.	**ASAP**

- The switch to COG increased the archive volume by about 10-15%.
- Ferran Gascon (ESA) commented that ESA is currently considering cloud distribution options for the archive of Sentinel data. They currently have a hybrid option for data storage. Targeting completion of a reprocessing of the Sentinel-2 data archive in 6 months (called 'Collection 1'). ESA is considering COG and STAC for their 'Collection 2'. For Collection 1, they will stick to JPEG2000 due to data volume savings (40-50%). The timeline for Collection 2 is not yet defined, but could be 3-4 years, and is very much dependent on cost. If the cost keeps going down then Collections could be processed more frequently. He noted that this is a topic for discussion with the EC.

Requirements Open Discussion

- Brad Doorn (NASA) commented that GEOGLAM will continue to develop their statement of user requirements and will be expanding to include accuracy measures. They will also include AFOLU in this expansion. The timeline can be found in the GEOGLAM Work Plan.
- Steve noted that through the AFOLU work, Sylvia Wilson has been reaching out (in-country) and looking for input on satellite data requirements. There is an active process set up by the Obama Administration (and continued under Trump & Biden Administrations) for passing on user needs to agencies. NASA has devoted significant expertise and time to this interagency process. The NISAR mission acquisition plan was adjusted based on interagency feedback.
- Jim Irons (NASA) noted the above interagency process is referred to as the Satellite Needs Working Group (SNWG). USGS leads the collection of requirements/desires from U.S. Federal Agencies for remote sensing EO (every other year). NASA completed a survey in 2020. In 2021, NASA has been conducting interviews with each of the agencies that responded with 133 requirements. The team is now producing a report addressing each one of those and highlighting the sources of data. The reports back to the federal agencies have been completed in the last month. NASA, with funding from OMB (Office of Management and Budget), have initiated efforts to address some of the needs that could not be addressed, such as the Landsat-Sentinel 2 harmonisation effort. They also would like to use NISAR to work toward reducing data latency.
- Brad added that there are lots of land imaging requirements out there, through SNWG, Decadal Survey, etc. NASA is starting to determine what the key issues are. One common issue is latency, and this is being prioritised when considering new missions.

- Tim Stryker (USGS) noted that this aligns nicely with the LSI-VC purpose (requirements driven by users' needs). The team should be asking how we can better address the types and timing of data that users are asking for.
- Matt asked whether there is a way the team can remain more aware of the progress of the SNWG. Jim suggested that in the timeframe of the next LSI-VC biannual meeting (September), Jim could organise for someone from NASA to present to LSI-VC.
- USGS has a website that reports on the work of the SNWG.

LSI-VC-10-18	Matt to include an item on the agenda for LSI-VC-11 or the regular teleconference in September for an SNWG report from Jim Irons / other NASA colleagues.	**September**

- Steve noted that it is good to have a collection of inputs of user needs. A challenge is to adjust it to make something useful from this information, and to make meaning out of it. Across the whole LSI-VC group that is a big challenge, and part of the group's job is to work this out. It is a good start to collect the data, but it needs to be organised in a way that is usable and helpful.

Global Grid Open Discussion

Peter Strobl presented regarding global gridding. He noted that, to be an ARD-compliant product, the data must be processed to a geo-referenced projection to enable position identification within the data product. If "geo-referenced projection" is to be understood as being transformed to a "geo-referenced grid" then this requires re-sampling.

The OGC Data Cube Community of Practice says that all layers in a Data Cube need to share the same grid to allow interoperability between layers. However, the current interoperability guidelines do not require two Data Cubes (or ARD datasets) to share the same grid to be considered interoperable.

To avoid repeated re-sampling in complex multi-source environments there are essentially two options: the "Point Cloud" approach, which suffers from high processing effort, only end-to-end processing and low reusability, and the "Grid System" approach, which has a limited number of representations and lacks user acceptance.

The JRC-INSPIRE GRG Workshop in 2017 found that the attitude of the participants towards a common grid system was strongly positive, and such a system has the potential to largely boost global data sharing.

Peter defined the criteria for (spatial) discretisation as:

- Assessable: based on ellipsoidal Earth model;
- Unambiguous: every point on the surface belongs to a cell;
- Gap-free: no point on the surface belongs to more than one cell;
- Hierarchical: grids can be refined from coarse to finer levels following mathematical rules (cell refinement);
- Nested: finer level cells do not overlap coarser cells;

- Intrinsic: the grid is a product of a mathematical tessellation of the ellipsoid, a cell is only determined by location;
- Instantaneous: the grid is defined for any point in time.

Discussion

- Adam has advocated for discrete grids a lot, funding activities in the past. He noted the announcement in the recent Australian budget for the funding of ATLAS ($10M/year) as a significant investment into the national mapping work of GA. Under this project, there might be some contributions from Australia to DGGS forthcoming. Simon Costello leads the branch in GA.
- The global grid issue is a problem across all spatial resolutions.
- Brian Killough (SEO) commented that CEOS ARD does not require gridding – the data provider communicates what their gridding is, but the specifications do not specify what it should be. Adam added that the data isn't always forced into a gridding.
- Jim noted that, looking at the PFS, it doesn't seem to require gridding, just geolocation, but Open Data Cubes require a grid. The challenge has always been, when dealing with interoperability: how to combine datasets using different systems – one or more will need to be re-gridded. He also noted that the polls (in Peter's presentation) indicated that the grids people want to use are based on heritage, but respondents asked for a global grid.
- The way providers currently distribute their data is very much driven by considerations of that provider and their heritage, rather than what users desire. They usually lack a thorough analysis of what can be done, but on the other hand, the grids with which the datasets are provided are of course shaping the way users use data, as users avoid resampling. If data providers changed to something more suitable, users would likely follow – according to the poll results shown.
- To increase interoperability, adding a common gridding requirement to CARD4L could work into the long-term requirements. However, Brian suggested this could be taking the role of CARD4L too far.
- Steven Covington (USGS) noted the interoperability continuum, where CARD4L is the starting point – the minimum to set off on this continuum. Steve Labahn added that there are techniques introduced into the framework, and suggested something like this could be added to Target requirements, as these are aspirational and could drive change. This could also be a good topic for an Advisory Note. There is a need to have these discussions amongst the primary data providers, as there is enough evidence that users are encouraging and expecting us to do so. Having common gridding as a Target requirement avoids limiting the inclusivity of the CARD4L specifications.
- Of the two options suggested by Steve, Adam supports the Advisory Note approach, if this topic were to be incorporated into CARD4L at all. The group decided this needs to be discussed further in LSI-VC.
- Adam questioned whether CARD4L should be the vehicle to discuss whether we should have a global gridding system – or should it just inherit the global grid system when it is used by data providers?
- Jim noted that NASA looks to the Landsat Science Team for guidance, and the Landsat ARD products are referred to as gridded products. However, it is hard to reach consensus on the best grid approach.

- Steve wondered whether OGC might be the right forum for this discussion. Peter is already active there.
- Zolti Szantoi (EC/JRC, LSI-VC Co-lead) suggested starting by seeing what the agencies responsible for major land surface imaging programs (e.g., USGS, ESA, NASA) have to say, as they will be the drivers of change.
- An interesting study would be for these agencies to start from Level 0 and move up to Level 2 DGGS, then compare results and evaluate what the actual interoperability differences would be for end users. Does it make a big difference? This would then provide some tangible evidence on what DGGS can offer. Ferran supported the idea to start with some experiments like this. Adam noted that there will need to be some agreement from the agencies on what model to go with, and that decision is tough. Ferran noted the team would need guidance on DGGS to start, particularly a recommendation of which DGGS is most promising.
- For the point cloud approach, Level-1B is the jumping off point. Orthorectification is the first massive resampling, and some domains don't want the data to be orthorectified, as they will perform analysis first then grid downstream. Brian asked if Peter thinks it's worth taking L1B from multiple providers, choosing DGGS, and performing the analysis noted above. Peter advocates for Level-1B data to be the input to Data Cubes, to avoid gridding and re-gridding, which is computationally costly. OGC Testbed-16 had some work in this direction.
- HLS is on a common grid but not DGGS, and not global.
- Another issue was raised regarding working with Level-1B sensor geometry, it was suggested to test what are the advantages of starting from Level-1B.
- Adam asked in chat: "*How about using Sentinel-3 as a test case? They are geophysical measurements that are geolocated but not gridded?"*

Closing

- Zoltan Szantoi (EC/JRC, LSI-VC Co-Lead) thanked everyone for their attendance and very valuable contributions to the discussion over all four LSI-VC-10 calls.

Appendix A: Meeting Presentation Slides

Further Topics, Discussion Time, and Wrap-up

LSI-VC-10 Teleconference #4

1

Overview

- USGS lessons learned regarding Landsat Collection 2 implementation in the cloud [for information]
- Requirements (e.g., for AFOLU / Global Stocktake, GEOGLAM, others? USGEO?) [open discussion]
- Global grid [open discussion] (Peter Strobl)
- Further discussion time [open discussion]
- Action review and wrap-up

2

USGS Collection 2 Cloud Implementation Lessons Learned

Steve Labahn
USGS

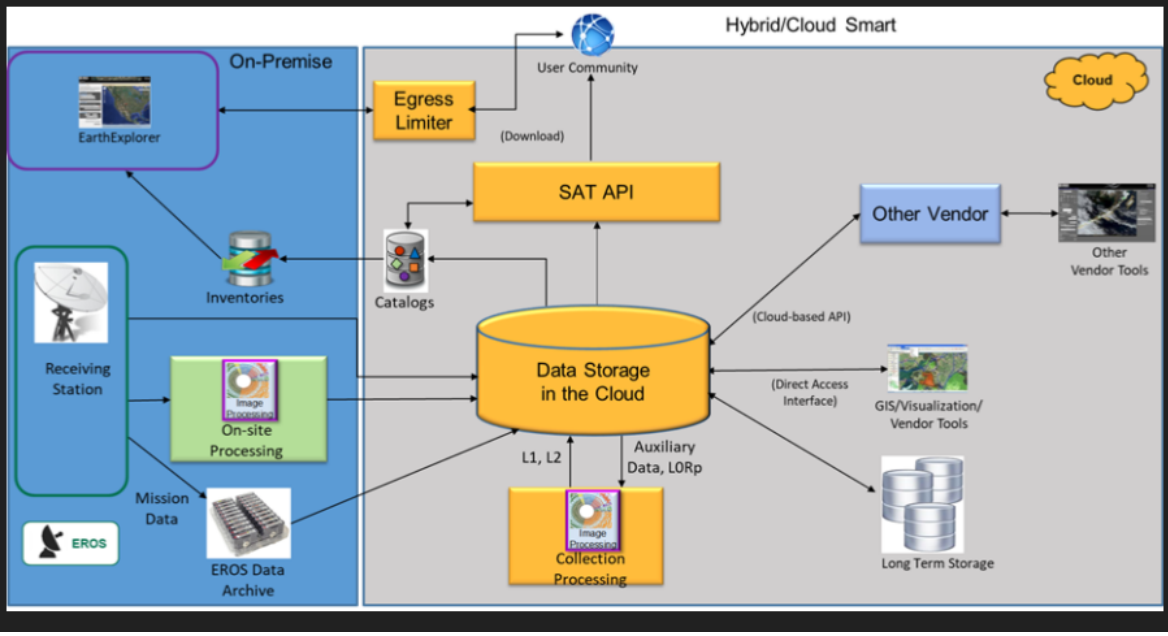
3

Landsat Cloud Project Scope

- Modernize Processing, Access, and Distribution of Landsat Data
  - Change from a primary business model of downloads to enabling access to the full archive
  - Enable users to interact with the data in an integrated environment
  - Ensure provenance and data stewardship
- Key Project Objectives:
  - Establish an enterprise cloud environment for Landsat
  - Enable access to Collection 1 Level-1 and Level-2 in the cloud
    - Replicate Collection 1 Level-1 and establish operational data management procedures
    - Demonstrate global scale production of Landsat data in the cloud through production of Level-2 products using a cloud framework
    - Process Landsat archive in 1-2 months rather than 18 months
  - Establish modern access and visualization tools to access data
  - Establish an Environment and System to Produce and Enable Landsat Collection 2 in the cloud
  - Demonstrate key science use cases exploiting Landsat data

4

Landsat Cloud Operational Concept View



5

AWS Tools & Services

- **Compute:**
  - AWS VPC, EC2, ECR, ECS, Spot Fleets, AWS Elastic Load Balancing, Batch, Lambda, Step functions
- **Network**
  - Route 53, CloudFront, API Gateway
- **Storage**
  - AWS S3, ElasticCache, EBS, EFS, Glacier
- **Database**
  - AWS Aurora PostgreSQL, Redshift, DynamoDB
- **Messaging**
  - AWS SQS, SNS
- **Analytics**
  - Athena, Glue
- **Management Tools**
  - AWS CloudWatch, Cloud Trail, CloudFormation, IAM, Secrets Manager
- **Developer Tools**
  - AWS Code Deploy
- **Other non-AWS tools:**
  - JIRA, Confluence, Jenkins, Git, Python, C, Docker, CloudCheckr

6

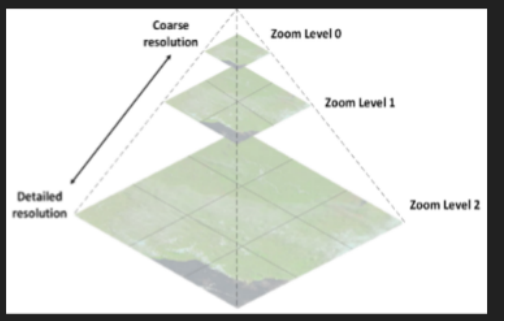
Web-Enabled to Cloud-Enabled

- Next revolution of Landsat data transitioned from a web-enabled approach to a “Smart Cloud” implementation
- USGS historical free and open data policy that enabled great progress in the last transformation to becoming web-enabled was continued into this new cloud environment
- New paradigm enables opportunity for users to access the data directly allowing:
  - Execution of algorithms directly on only the data they need
  - Selective data usage (e.g., specific bands for use)
  - Reduced need for permanent IT infrastructure (e.g., serverless compute)

7

Cloud Optimized GeoTIFF (COG) Format

- Conducted trade study on cloud formats, which resulted in the selection of Cloud Optimized GeoTIFFs (COGs)
- An enhanced GeoTIFF with tiling and overviews
  - Uses internal tiling instead of lines to speed access and support better remote reading
  - Downsampled overviews are generated when lower resolution data is acceptable
  - No changes to the underlying pixels
  - Stored in an unbundled format
  - Data is internally compressed
  - Enables HTTP Get Range requests
- Worked closely with Visualization team to standardize the COG parameters
- Delivered prototype COGs for use in LandsatLook development
- COG formatting built into Landsat Product Generation System (LPGS)



8

SpatioTemporal Asset Catalog (STAC)

- New collaborative standard for managing access metadata
  - Open-source on GitHub, hosted by Radiant Earth, includes Landsat extension
  - Flexibility to support many types of geospatial data (satellite, drone, radar, etc.)
  - Allows for interoperability between satellite metadata (e.g., Landsat 8 + Sentinel 2)
  - Lives alongside product-level metadata (MTL, XML)
- Exposes data in a common, machine-readable JSON format for both end users and internal processes
- Includes direct links to S3 objects
- Can be exploited through Jupyter Notebooks by end users to read data directly from the cloud without downloading
- Gaining wide adoption by the remote sensing community
  - i.e., Government, International, Commercial, Academia

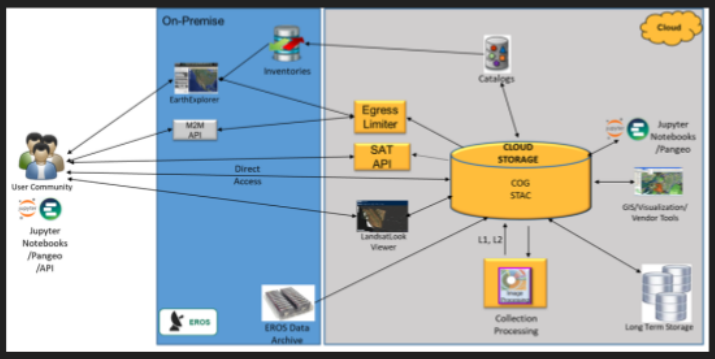
9

Processing Metrics

- **Collection 2 processing took one month to complete**
  - 8.8 million scenes to Level 1 and Level 2
  - Dates of processing: 19 August – 19 September 2020
  - Average per day: 293,000/day
  - Kept initial numbers lower as we were gaining experience with running at-scale cloud processing
  - On-premises production for Landsat 8 would typically top out at ~25,000/day
- **Previous Collection 1 processing (only Level 1) took over 18 months (with some downtime between individual missions while code changes were completed)**
- **Docker images were used to run the Image Processing containers**
- **AWS Spot EC2 types:**
  - General purpose w/SSD (m5d) 4xlarge to 24xlarge
  - Memory Optimized w/SSD (r5d) 4xlarge to 24xlarge
- **AWS Batch and Step Functions were used for scheduling the jobs**
  - Overall – worked okay, but we did run into limits in that environment
  - Spot terminations caused some issues in the processing runs
- **CloudWatch costs were not expected**
  - We had logs turned on for the processing runs, which caused unexpected costs

10

USGS Landsat Collection 2 Access Architecture



Traditional Access:

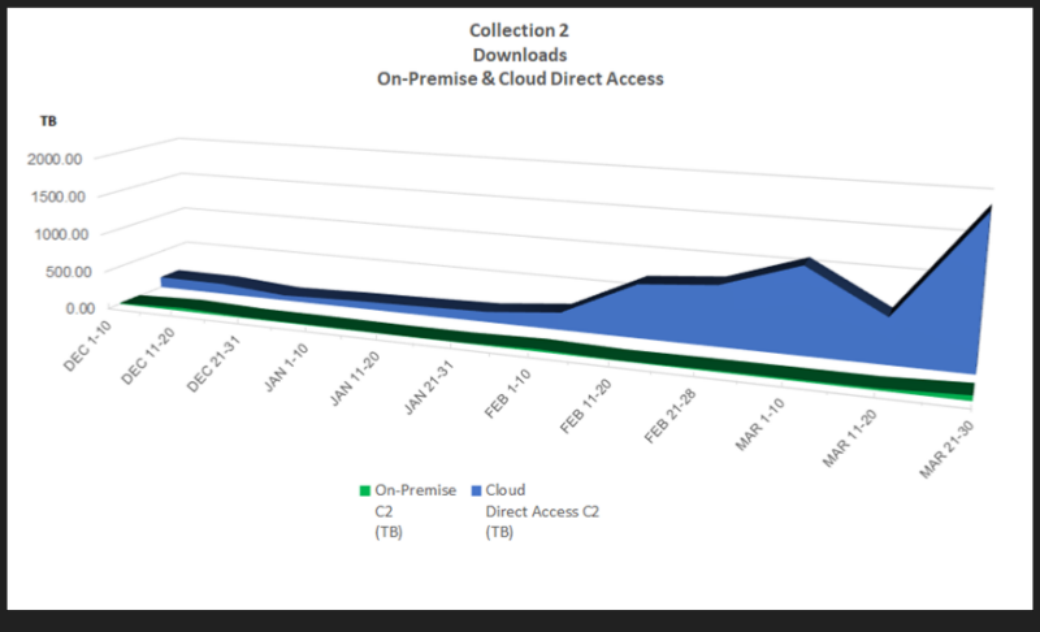
- EarthExplorer (usgs.gov)
- Machine-to-Machine (M2M) API

Landsat Commercial Cloud Direct Access User Guide

- AWS Command Line Interface (CLI)
  - Powerful tool for batch processing and script-based workflows
- STAC Browser
  - Web-based GUI that allows users to browse and review STAC records and thumbnails for Landsat data in the cloud
- SAT API
  - Allows users to programmatically search the Landsat data in the cloud for products within their spatial and temporal AOIs
  - SAT API returns links to objects that the user can then incorporate into their processing workflow

11

Distribution Metrics



12

Lessons Learned

- **What went well:**
  - On-demand database scaling
    - Increasing read replicas and instance sizes during scale testing in production
  - Aurora Database Architecture and design (scenes_db, catalog_db)
  - Migration from Dynamo DB to Aurora
  - Redshift post processing data validation efforts for Collection 2
  - Separate accounts for Development, System Test, Production
    - Separation of Development, System Test, Production was invaluable!
  - Continuous Integration Continuous Deployment (CICD) processes work well
    - CloudFormation Templates used for all deployments
    - Packer used for creating machine images
    - Ansible Playbooks
    - Jenkins Pipelines
    - Signoff process between environments
    - GitLab runners
  - GitLab was invaluable as a repository (EROS Enterprise Asset)
  - AWS cloud storage infrastructure works very well for data distribution
  - Security groups used to control/limit access to services and applications
  - Ability to scale processing has been proven to be cost effective

13

Lessons Learned

- **Opportunities for improvement:**
  - Batch processing has limitations; AWS recommends migrating to Kubernetes
  - Contention with resources in the same account
    - Batch
    - Step
    - Spot instances shared between processing and inventory management
  - Multiple accounts need an enterprise approach to managing the accounts, push out updates, monitoring, etc.
  - Separate processing and distribution into separate accounts
    - Two teams independently developing systems/processes and competing for AWS service limits

14

Global Grids for CARD

indispensable or unnecessary?

Peter Strobl, EC-JRC

CEOS LSI-VC-10 virtual meeting, 20 May 2021

European Commission

Analysis Ready Data (ARD)

An Analysis Ready Data (ARD) product is generated from raw data and processed so that it can be used **without the need for further processing** to be applied by users.

… minimum processing requirement to be an ARD-compliant product: the data **must be processed to a geo-referenced projection** to enable the position identification within the data product. …

- If “geo-referenced projection“ is to be understood as being transformed to a “***georeferenced grid***” than this requires **re-sampling**!

European Commission

CEOS Analysis Ready Data (ARD)

CEOS Analysis Ready Data (CARD) are satellite data that have been processed to a minimum set of requirements and organized into a form that allows immediate analysis with a minimum of additional user effort and **interoperability** both through time and **with other datasets**.

- It is fair to assume that “other datasets” means those who also meet the CARD requirements, i.e. they are “gridded” (and thus re-sampled).

CARD Interoperability

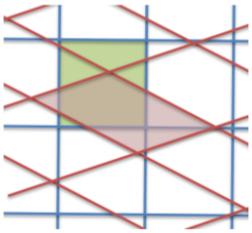
Interoperable *Products* refers to a set of two or more ARD products which are sufficiently documented **to enable processing** across a continuum of geometric and/or radiometric standards **to permit direct** quantitative **comparison**.”

- Not the *interoperable products* themselves are able to ‘interoperate’ (i.e. be compared or analysed together) but their derivates.
- ‘Interoperability’ here means a ‘can’ not an ’is’ and, if the underlying references are not the same, it requires adaptation (processing)!
- How ready is ‘ready’ in ARD?

Changing references

Observations need to relate to the same standards or references to be comparable:

- scaled data (based on points)
  need to be **re-scaled** (e.g. ºF → ºC)
- gridded data (based on intervals)
  need to be **re-gridded** (or resampled)
  (e.g. 1” WGS84 /Pseudo Mercator EPSG:3857
  → 30m Mollweide EPSG:54009)



 Sounds doable for continuous parameters which can be interpolated (e.g. radiance, temperature), but categorical data (e.g. masks/flags)?

ARD and Datacubes

- ARD are meant to build data cubes!
- OGC data cube Community Practise* says:

All layers in a data cube need to share the same grid
to allow interoperability between layers

 Co-gridding is an important element of interoperability WITHIN a data cube

*https://portal.ogc.org/files/18-095r7

Datacube Interoperability

BUT

Two data cubes (or ARD datasets) do not need to share the same grid to be considered interoperable(?)

- If so, is interoperability restricted to a 'one way' road?
  (i.e. a specific cube can only be involved once during an analysis workflow)
- And then, how is reproducibility being secured?
  (e.g. for Cubes A,B routing A→B gives a *slightly* different result as B→A, repeating exchange and involving more Cubes worsens things considerably!)

Consequences of re-sampling for interoperability

(unless the volume of data is largely amplified each time)

- always entails an interpolation of data
- always diminishes data accuracy or entail data loss
- always is irreversible
- is (more or less) computer-intense
- **accumulates these effects when repeated!**

- How compatible is this with FAIR principles?

Big Geospatial Data Analysis Strategies

To avoid repeated re-sampling in complex multi-source environments there are essentially two options:

- "Point Cloud" approach:
  - Store all observations with their locations (as n-tupels)
  - Resample (all input data) to a user selected grid only at the point of analysis
  - High processing effort, only end-to-end processing, low re-usability

- "Grid System" approach:
  - Discretise (re-sample) all observations to common grid system (not only in spatial dimension!)
  - Limited number of (spatial) representations, lack of user acceptance

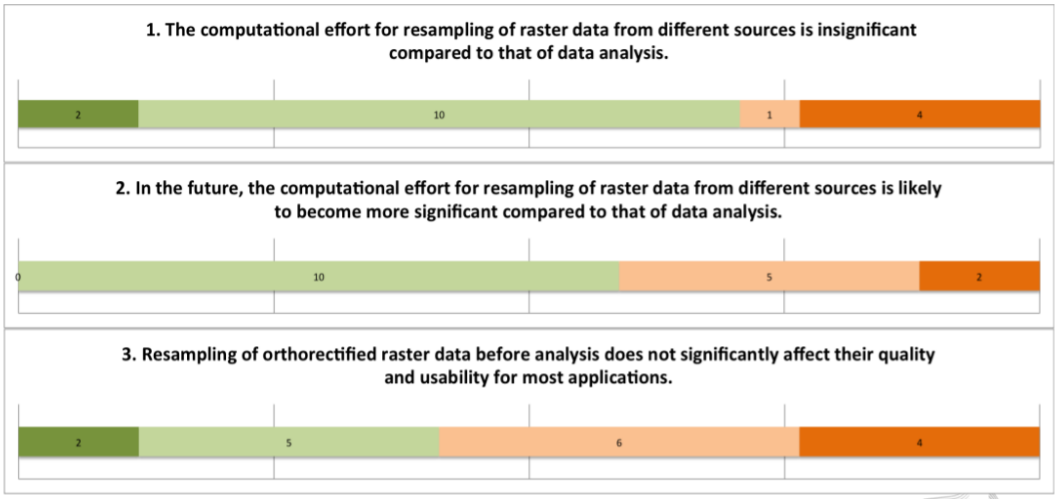
INSPIRE* wisdom

"… it would be highly desirable that all the themes with similar needs make use of the same geographical grid system in order to maintain their coherence."

Source: INSPIRE D2.8.II.1 Data Specification on Elevation – Technical Guidelines (2013)

*INSPIRE is the EU initiative to establish an infrastructure for spatial information in Europe that will help to make spatial or geographical information more accessible and interoperable for a wide range of purposes supporting sustainable development. https://inspire.ec.europa.eu/

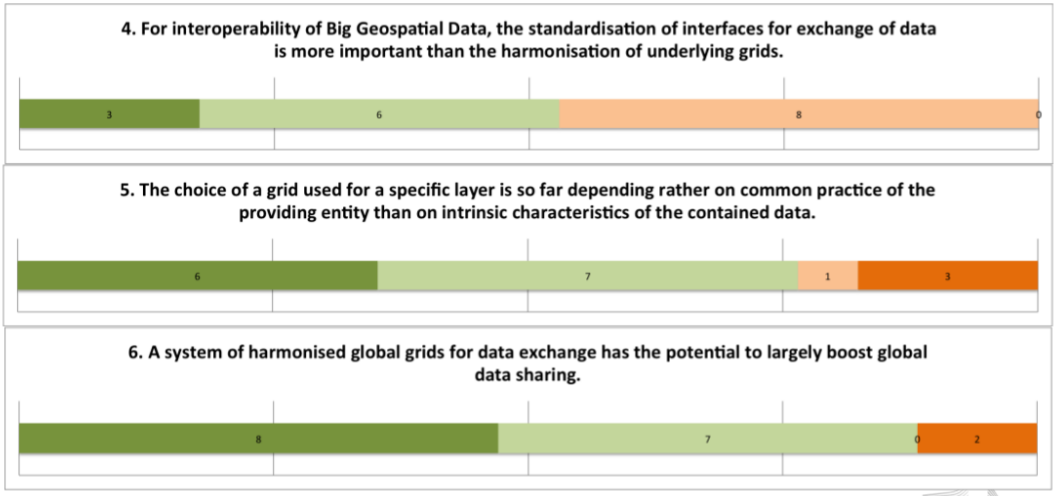
JRC-INSPIRE GRG Workshop 2017



11

European Commission

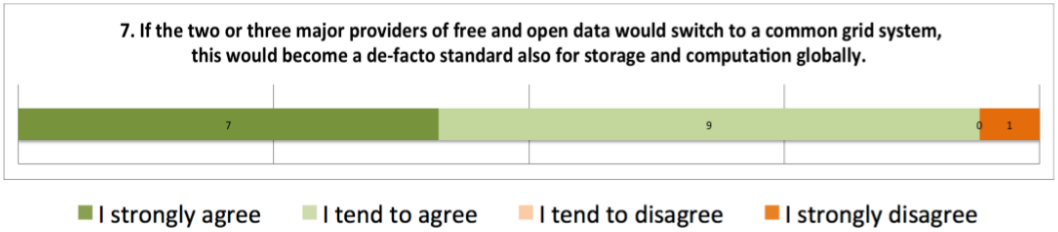
JRC-INSPIRE Workshop 2017 Questionnaire



12

European Commission

JRC-INSPIRE Workshop 2017 Questionnaire



13

European Commission

Geodata representation in the 21st century

Questions for the leading global EO data providers:

Continuous (point-clouds) or discrete (grids)?

If grids then:

Global or continental,
mono-resolution or hierarchical?

Which criteria for 'good' global grids?
(e.g. Goodchild/Kimerling)

Main candidate global grid(system)s?

Separate or together?

M. Goodchild (2019):
So, in the final analysis, the big-picture question for DGGSs remains the same as it has been for more than two decades: *how do we use the compelling arguments for these multi-resolution systems to persuade the larger scientific community to adopt them, in preference to the distorted representations of digital maps?*
https://doi.org/10.3138/cart.54.1.preface

14

European Commission

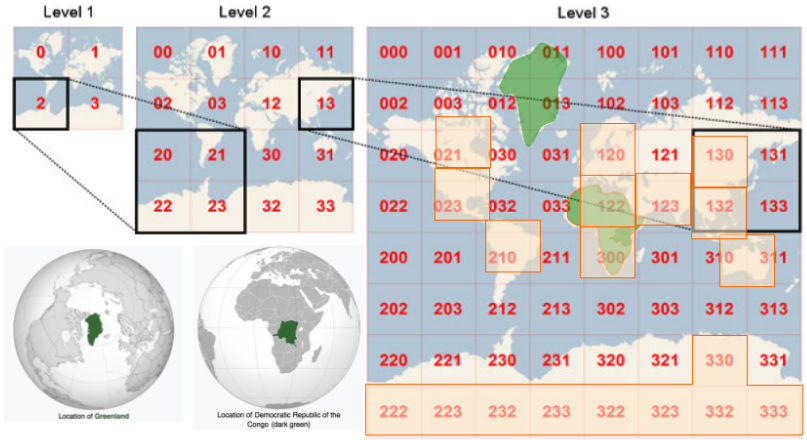
Criteria for (spatial) discretisation

- assessable: based on ellipsoidal Earth model
- unambiguous: every point on the surface belongs to a cell
- gap free: no point on the surface belongs to more than one cell
- hierarchical: grids can be refined from coarser to finer levels following mathematical rules (cell refinement)
- nested: finer level cells do not overlap coarser cells
- intrinsic: the grid is a product of a mathematical tessellation of the ellipsoid, a cell is only determined by location
- instantaneous: the grid is defined for any point in time

15

European Commission

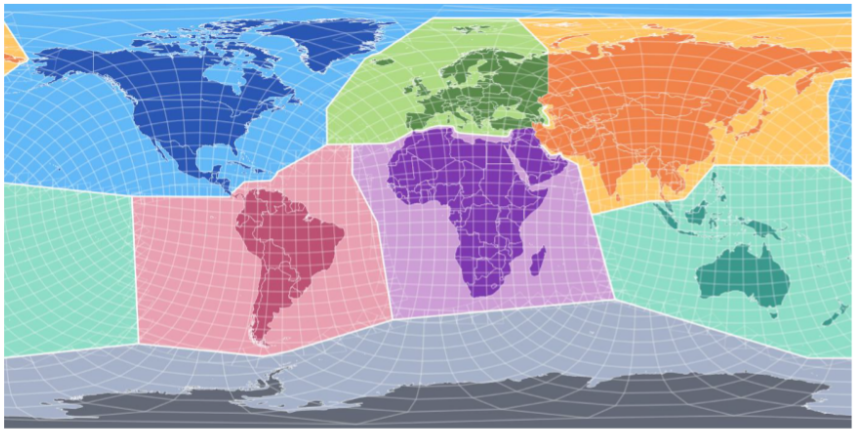
The WMTS standard (base EPSG:3857)



16

European Commission

The EQUI7 grid (TU Vienna)

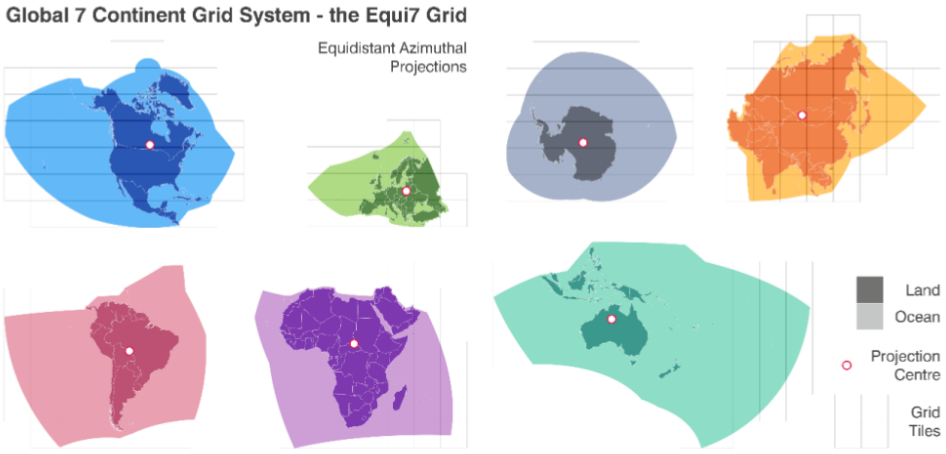


B. Bauer-Marschallinger, Optimisationof global grids for high-resolution remote sensing data, Computers & Geosciences, 2014 doi:10.1016/j.cageo.2014.07.005

17

European Commission

The EQUI7 grid (TU Vienna)



18

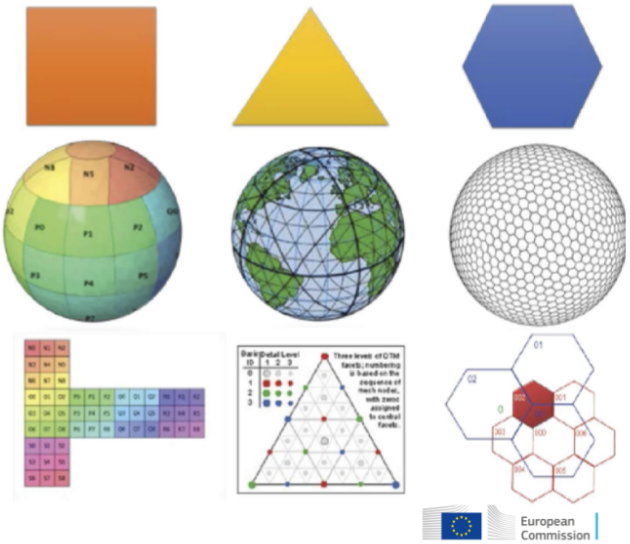
European Commission

Discrete Global Grid Systems

"…spatial reference system uses hierarchical tessellation of cells to partition and address the globe.

"DGGS are characterized by properties of cell structure, geo-encoding, quantization strategy and associated mathematical functions."

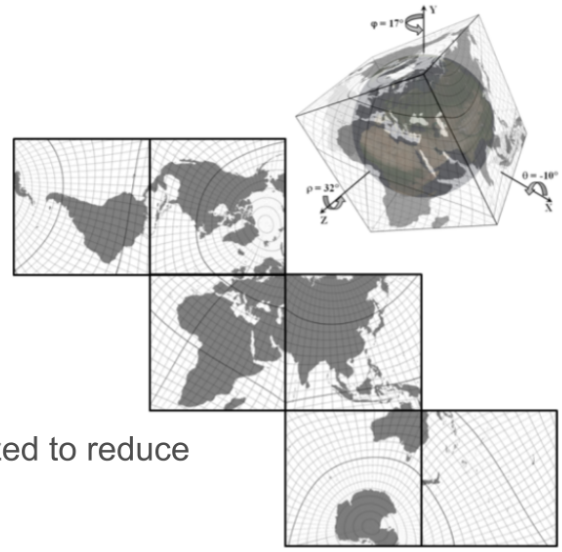
ISO 19170-1 Geographic information — Discrete Global Grid Systems Specifications —Part 1:Core Reference System and Operations, and Equal Area Earth Reference System
https://www.iso.org/standard/32588.html



19

DGGS optimization

- After the main choices, i.e.:
  - shape ("square")
  - Refinement ratio (4 - quadtree)
  - "cube-sphere mapping"
- Cube sphere mapping can be optimized to reduce distortions e.g. over land masses



Dimirijevic A., Strobl P., *Continuous 2D Maps Based on Spherical Cube Datasets,* Proc. 55th International Scientific Conference on Information, Communication and Energy Systems and Technologies (ICEST), doi:10.1109/ICEST49890.2020.9232678, 2020

20

Discussion

- Does CARD require gridding?
- Is interoperability (in a processing chain restricted to a ‘one way’ road?
- How is reproducibility being secured when data are resampled?
- How compatible is repeated resampling with FAIR principles?
- Is there room for a “common global CARD grid system”
- Which are the top priority criteria for a CARD global grid?

European Commission

Thank you!
Any questions?

peter.strobl@ec.europa.eu

European Commission


Action Review and Wrap-up
LSI-VC Sec & Leads
15