ESCO Session at ESIP Winter 2021

- Session Information
    - Session Abstract
    - ESIP Session Links
    - Attendee list
- Agenda
    - Welcome
    - Presentations
        - Speakers
            - Ignacio Zuleta (ARD Zone)
            - Steven Labahn (USGS, CEOS Land Surface Imaging Virtual Constellation (LSI-VC) Co-Lead)
            - Chris Lynnes (NASA):
        - Questions to the speakers (captured from the Slido tool and the Zoom chat)
    - Breakout Sessions
        - Room 1 - leader: Ed Armstrong
        - Room 2 - leaders: Steve Olding, Chris Lynnes
        - Room 3 - leader: Shannon Leslie
        - Room 4 - leader: Allan Doyle
    - Closing Plenary

Session Information

Title: Analysis Ready Data in Science and Industry

Part of the 2021 ESIP Winter Meeting (Jan. 26th-29th, 2021)
https://2021esipwintermeeting.sched.com/

Held on January 27, 11:00 EST

The meeting was conducted via Qiqochat, using Zoom for presentations and breakout sessions.
Slido was used to capture audience questions and to conduct the two polls.

Session Abstract

Interest in the subject and implementation of Analysis Ready Data (ARD), especially for remote sensing products, continues to build in the domain of science data producers and private industry, and their user communities.

In this session we will explore and hear about the landscape of ARD activities from science data producers, private industry stakeholders, and international coordination activities like the Committee on Earth Observation Satellites (CEOS) and others. We will solicit presentations on ARD definitions and assessment, implementation, and practical examples of ARD datasets and their applications highlighting both the successes and challenges.

One of the potential outcomes of this session is to build momentum toward more harmonization of diverse ARD activities and definitions.

ESIP Session Links

- Session page on ESIP Sched
- ESIP Session notes page (live) (pdf)

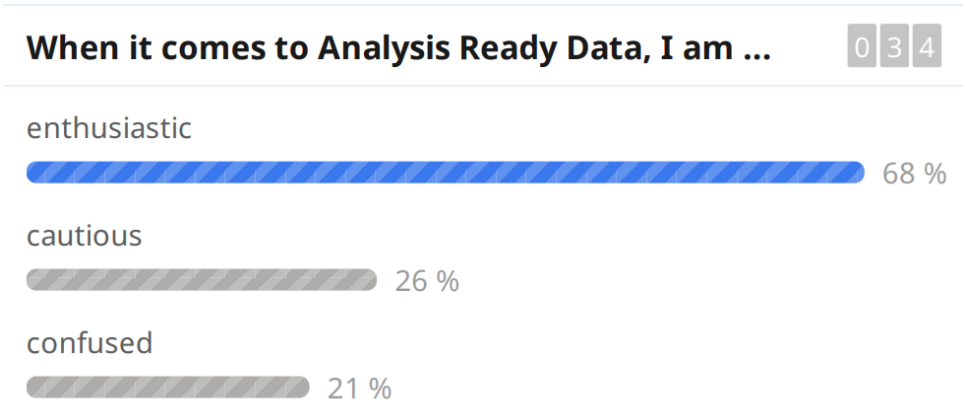
Attendee list

- Available on the ESIP session notes page

Agenda

Welcome

- Ed Armstrong (LP.DAAC, ESCO): *Introductory presentation*
- Audience poll #1 on ARD:



- Audience poll #2 on ARD:

Presentations

Speakers

Ignacio Zuleta (ARD Zone)

- *Virtual constellations, ARD and sensor fusion: the future of earth observation* [Missed presentation, Technical Issues]

Steven Labahn (USGS, CEOS Land Surface Imaging Virtual Constellation (LSI-VC) Co-Lead)

- *CEOS Analysis Ready Data (ARD)*

Chris Lynnes (NASA):

- *Analysis Ready Satellite Data in NASA's EOSDIS*

Questions to the speakers (captured from the Slido tool and the Zoom chat)

- Steven Labahn (USGS, CEOS Land Surface Imaging Virtual Constellation (LSI-VC) Co-Lead): *CEOS Analysis Ready Data (ARD)*
  - Reply: Ease of use - guaranteed that product is met with a certain kind of quality and standard. Showed scorecard summary tables - level of assurance around what a data product contains
  - Does CEOS provide more granularity on what 'use with ease and confidence' means for a new user?
  - Are all data transformations (data and code), from original to ARD, saved? What system is used to save this history?
  - ARD is a powerful "holy grail" and CARD4L is so well-defined, is there a danger that people might think ARD is a solved problem
    - Reply: CARD4L is not trying to define all of ARD
  - Why did you decide not to specify the metadata format & data packaging? That seems like a key usability requirement.
  - In STAC, we are currently working on creating CARD4L extensions, and there are two open PRs for SAR and Optical.
  - Can you say a few more words about COAST?
- Chris Lynnes (NASA): *Analysis Ready Satellite Data in NASA's EOSDIS*
  - One of the challenges is that the "processing level" definitions appear to be different between land and atmosphere... where in land Level-2 is gridded/ground referenced whereas that's Level-3 in the atmosphere domain
  - pixels expand near the pole simply because of the projection used. the FOVs are function of off-nadir angle only
  - How might the ESIP endorsed Analytics definition be updated with ARD concepts? https://datascience.codata.org/articles/10.5334/dsj-2017-006/
  - Confounding factors?
  - In terms of "easy to used spatial characteristics", have you considered using a standardized hierarchical spatial indexing system like H3? see H3geo.org.
  - How did you create AIRS L2G time-aggregated data before you run PANDAS on it? Did you store everything in memory or save it in an intermediate format?

Breakout Sessions

Breakout into 4 Zoom rooms - each room had about 10-12 participants. There were 3 questions to help focus the conversation:

- For ARD producers: What is your (organization's) approach to producing ARD?
- For ARD users: What use cases does ARD fulfill for you?
- Should ESIP get involved in ARD? How?

Room 1 - leader: Ed Armstrong

- For ARD producers: What is your (organization's) approach to producing ARD?
- For ARD users: What use cases does ARD fulfill for you?
    - Ed: Interdisciplinary use cases important
    - Douglas: Toolkit and services important for on the fly generation
    - Rob R (Space Scientist): Uncertainty documentation important ! Machine consumable. Ready for data assimilation.
    - Byron P: GIS applications are important. Issue quality control for data preparation and data understanding.
    - Discussion about Pangeo…..
- Should ESIP get involved in ARD? How?
    - ESIP cloud computing cluster is one possibility: P Quinn working on data chunking…..
    - Annie: How can ESIP facilitate? What are the resources needed?. Define the problem space.
    - Rich: Training on resources and tools important. Its bifurcated, different material at different institutions.
    - Alek: Follow OGC Testbed model. Sponsered orgs provide common data. Like a test bed that people could evaluate.
    - Chris L: OGC is time consuming. Better HDF "Zoo", where data can be collocated with tools

Room 2 - leaders: Steve Olding, Chris Lynnes

- For ARD producers: What is your (organization's) approach to producing ARD?
- For ARD users: What use cases does ARD fulfill for you?
- Should ESIP get involved in ARD? How?
- General ARD discussion
    - Many ways to make data ready for analysis - e.g. make tools work better, put out more information with the data, providing training or how-to guides.
    - Analysis ready may mean different things to different users / domains.
    - Useful to look at extreme cases? May have some data that are useless for analysis. First step identify data that can be made analysis ready - ease of upgrading with automated tools. Maybe able to design thresholds - from unusable > usable with manual work > usable automatically. Non-binary forms of analysis ready data e.g. ready for certain types of analysis.
    - On-the-fly ARD provides an option to defer some decisions e.g. what projection to use.
    - Pre-process steps to get to what a user would consider AR. Difference seems to be whether those steps can be automated or manual. Seems like the same destination is desired but the approach is different.
    - ARD need to be collated and formatted in a similar way.
    - Tiling needs to be addressed in ARD.
    - Data formats and metadata also important.
    - Someone has to pay for the processing to get to ARD.
    - Users want predictability and repeatability.
    - Performance issues with on-the-fly ARD. Will diminish with move to cloud and less expensive compute power.
    - Different needs. NASA primarily serves the research community vs. application or commercial community. Would it be useful to Survey users?

Room 3 - leader: Shannon Leslie

- For ARD producers: What is your (organization's) approach to producing ARD?
- For ARD users: What use cases does ARD fulfill for you?
- Should ESIP get involved in ARD? How?
    - If ESIP had a group that helped with terminology, collecting info. How should people understand the different terms?
- Cloud-focused discussion:
    - ESIP cloud cluster has talked about ARD; there is a document with this info
    - Pangeo project has a draft standard for metadata that includes chunk locations in files you are pointing to; provides universal way of making any data format cloud-optimized
    - Metadata becomes very important (CF conventions) with cloud optimized data
        - Can't hide metadata through some layer of service
    - Expose grid in same way; need to interpret metadata in there and present same way
    - Question: Are cloud-ready and ARD synonymous?
        - Not necessarily, but all (cloud) workflows include the re-chunking step.
- Other
    - Need to expand ARD definition beyond satellite
    - ARD becomes more important as access methods expand
    - Seems to be emphasis on time series (currently)
    - People generating time series are maybe going away from COG, going to zarr...
    - Want to expand ARD to climate data, other types of data
    - Forecasters traditionally look at 2D maps, but now using 3D
    - Broaden definition of ARD in future; ARDs depend on users - e.g., Netflix users want Earthdata in video. Alexa users want Earthdata in voice. Minecraft users want Earthdata in 3D blocks.
    - ARD is a format that users can understand and put into their tools
    - Liked the idea of improving the tooling and helping people learn how to do
    - ERDDAP: https://coastwatch.pfeg.noaa.gov/erddap/index.html

Room 4 - leader: Allan Doyle

- For ARD producers: What is your (organization's) approach to producing ARD?
  - Pulling data out of HDF files, providing it in other formats (e.g., csv) for ML pipelines, etc
    - Search by variables (e.g., NO2), find the data, and the CSV data is displayed (Lingua Logica): https://nasamadesimple.net
    - Developed a way of reading metadata
  - Created a basic streaming interface (as csv, json, binary) w/applications in heliophysics; not yet migrated to the cloud
  - Very basic (programmer focused) web site: hapi-server.org
    - There are clients and server codes also in github
    - The specification is available here:
      - https://github.com/hapi-server/data-specification
- For ARD users: What use cases does ARD fulfill for you?
  - Data format (how to read conventions in a standardized manner)
    - Need it for ML and Deep Learning
    - Which data structure do we prefer to read our data? NumPy
    - Challenge: no standardization, VIIIRS vs Helio, cmip6 repository - registered themselves with that particular organization - ESFG
    - Provided filters and which organizations you want data from (NASA…)
    - Data come from models, not real observations
  - Has anyone worked with CMIP6 data repositories? Though it comprises model/simulated data, a similar repo is needed for observational data
- Should ESIP get involved in ARD? How?
  - Update the endorsed definition of analytics (versus analysis?)
    - Analytics: already done analysis
  - Not all the same (for some, gridded data)
- Other:
  - Distinction between analysis ready data and tools that help users (e.g., for viz): not by preparing the data
  - Dataset interoperability
  - Standardization - data volumes are much larger in earth science.
- High-level takeaways
  - ARD on-the-fly can help tailor data for specific analysis needs
    - ARD Services
    - ARD tools
  - ARD can feed AI/ML
    - See AI-themed sessions during rest of ESIP meeting
  - It's hard for users to talk about ARD without talking about "format", i.e. data format or format of data stream from a service

Closing Plenary

- Reports from each of the four breakout rooms
- Possible next steps
  - "Test bed" approach with datasets (OGC model - data + recipes) sponsored by ESIP providing data in a standard way (but labor intensive)
    - Zoo of analysis ready data - how well is this tool working with ARD?
  - Related discussions ongoing in Cloud Computing Cluster (not so much metadata content, but with Analysis Ready Cloud-Optimized formats- ARCO!)