DALTON LUNGA, SILVIA ULLO, UJJWAL VERMA, GEORGE PERCIVALL,
FABIO PACIFICI, AND RONNY HÄNSCH

Analysis-Ready Data and FAIR-AI—Standardization of Research Collaboration and Transparency Across Earth-Observation Communities

The IEEE Geoscience and Remote Sensing Society (GRSS) Image Analysis and Data Fusion Technical Committee (IADF TC) serves as a global, multidisciplinary network for geospatial image analysis, e.g., machine learning (ML), image and signal processing, and computer vision (CV). The IADF is also responsible for defining the directions of the data fusion contests while paying attention to remote sensing (RS) data's multisensor, multiscale, and multitemporal integration challenges.

Among recent activities, the IADF is collaborating with the GRSS TC on Standards for Earth Observation (GSEO) and other groups to promote two complementary initiatives: 1) reducing the overhead cost of preprocessing raw data and 2) improving infrastructure to support the community reuse of Earth-observation (EO) data and artificial intelligence (AI) tools. The EO community has engaged the aforementioned 1) via a series of workshops on analysis-ready data (ARD) [1], which has laid bare that current best practices are provider specific [2]. Engagements in developing the aforementioned 2) are in the early stages. They lack a guiding framework similar to findable, accessible, interoperable, reusable (FAIR) [3], outlining standardized principles for best data stewardship and governance practices.

Developing templates and tools for consistently formatting/preprocessing data within a discipline is becoming common in many domains. It is a practice that is helping to increase research transparency and collaboration. Although not broadly adopted in EO, such a practice could enable data and derived AI tools become easily accessible and reusable. However, the immense diversity of modalities and sensing instruments across EO/RS makes research development and adoption challenging.

This article provides a first outlook on guidelines for the EO/RS communities to create/adapt ARD data formats that integrate with various AI workflows. Such best practices have the potential to expand the impacts of image analysis and data fusion with AI and make it simpler for data providers to provide data that are more interoperable and reusable in cross-modal applications.

WHY STANDARD ARD, FAIR DATA, AND AI EO SERVICES?

Poor best practices and the lack of standardized templates can present barriers to advancing scientific research and knowledge generation in EO. For example, synthesizing cross-modal data can be extremely time consuming and carries huge overhead costs when preprocessing raw data. In addition, with AI tools being increasingly pervasive across humanitarian applications, the time needed to generate insights is becoming critical, and reusability is preferred. At the same time, duplication of efforts is costly and hinders fast progress.

Digital Object Identifier 10.1109/MGRS.2023.3267904
Date of current version: 30 June 2023

Standards for data have been proposed as essential elements to advance EO sciences. The Open Geospatial Consortium's Sensor Observation Service standard defines a web service interface that allows pulling observations, sensor metadata, and representations of observed features [4]. Such accredited standards help outline broad governing protocols, but can take longer to build governing processes and consensus. In contrast, grassroots efforts [5], [6] can foster efficient adaptation of best practices to harmonize cross-modal/-sensor data, creating datasheets and model cards for AI tool types, with working groups helping to maintain cross pollination of taxonomies.

ARD

With the increased availability of observations from multiple EO missions, merging these observations allows for better temporal coverage and higher spatial resolutions. Specifically, ARD can be created from observations across modalities. The ARD could be processed to ensure easy utilization for AI-based EO applications. Standard ARD components include atmospheric compensation, orthorectification, pansharpening, color balancing, bundle block adjustment, and grid alignment. In its advancement, future ARD processes could see radiometric and geometric adjustments applied to data across modalities to create "harmonized" data [7], enabling study of the evolution of a given location through time using information from multiple modalities. Figure 1 shows the envisioned standardization process informed by ARD and FAIR principles [8]. For example, suitable cross-modal data formats could be established to create AI-ready datasets compatible with open source ML frameworks, accelerating the path from image analysis and data fusion research prototyping to production deployment. As FAIR principles continue to pave the way for the state of practice in other scientific domains, the EO community could benefit by following suit and introducing FAIR-EO definitions to guide research transparency and collaboration.

TOWARD CROSS-MODAL ARD AND FAIR DATA PRINCIPLES

The aforementioned shortcomings present an opportunity to collaborate toward a concise and measurable set of cross-modal FAIR ARD and FAIR model principles, ultimately advancing image analysis and data fusion algorithmic impacts at scale. The overarching goal is to harness best-practice ARD developments to minimize user burden by harmonizing heterogeneous imagery data and promoting FAIR principles for both EO data and AI products.

A joint IADF-GSEO paper submitted to the 2023 International Geoscience and Remote Sensing Symposium revisits current best practices and outlines guidelines for advancing EO data and derivative AI products for broader community use.

The remaining work needs to start by revisiting common ARD essentials and aim to forge their evolution with FAIR principles to support cross-modal-based ML and CV opportunities emerging as central aspects for solving complex EO challenges. An integrated framework (presented as a general scheme in Figure 1) that combines ARD and FAIR for modernizing the state of practice in AI for EO must

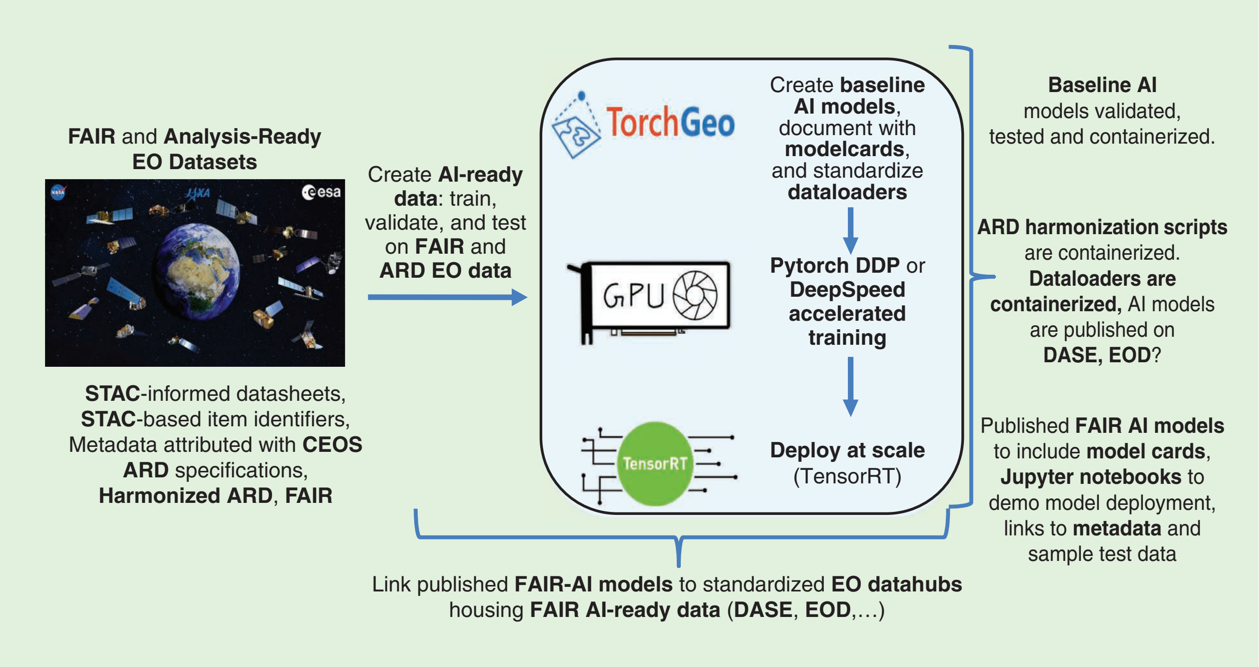
FAIR and Analysis-Ready EO Datasets
esa
JAXA
STAC-informed datasheets, STAC-based item identifiers, Metadata attributed with CEOS ARD specifications, Harmonized ARD, FAIR
Create AI-ready data: train, validate, and test on FAIR and ARD EO data
TorchGeo
Create baseline AI models, document with modelcards, and standardize dataloaders
GPU
Pytorch DDP or DeepSpeed accelerated training
TensorRT
Deploy at scale (TensorRT)
Baseline AI models validated, tested and containerized.
ARD harmonization scripts are containerized. Dataloaders are containerized, AI models are published on DASE, EOD?
Published FAIR AI models to include model cards, Jupyter notebooks to demo model deployment, links to metadata and sample test data
Link published FAIR-AI models to standardized EO datahubs housing FAIR AI-ready data (DASE, EOD,...)

FIGURE 1. ARD-motivated FAIR EO data and FAIR-AI model principles integrated into a common EO process. STAC: SpatioTemporal Asset Catalog; DASE: Data and Algorithm Standard Evaluation; EOD: Earth observation database; DPP: Distributed Data Parallel.

be contextualized. The framework will depend on several building blocks, including software scripts that demonstrate data harmonization, creation of datasheets for AI-ready datasets, creation of model cards for FAIR-AI models, measurement/validation metrics, and standardized environments for model deployment.

FAIR-AI MODELS

Recent developments from the ML community [9], [10] could provide initial building blocks to advance metadata standardization, but for EO applications [11], [12]. The ideas of developing datasheets [9] for data and model cards [10] for models have been introduced as mechanisms to organize the essential facts about datasets and ML models in a structured way. Model cards are short documents accompanying trained ML models that provide a benchmarked evaluation in various conditions. For EO/RS, such conditions could include different cultural and geographic locations, seasonality, sensor resolution, and object feature types relevant to the intended application domain. The RS community could aim to develop model cards to catalog model performance characteristics, intended use cases, potential pitfalls, or other information to help users evaluate suitability or compose detailed queries to match their application contexts. Similarly, each data source should be developed with a datasheet documenting its motivation, composition, collection process, recommended uses, and models generated from the data.

STANDARDS FOR EVALUATION

Evaluation metrics provide an effective tool for assessing the performance of AI models. Most of the evaluation metrics for CV-based EO applications are adapted from traditional CV tasks (such as image classification, semantic segmentation, and so on). These traditional CV metrics were designed for natural images. In addition, different evaluation metrics are sensitive to different types of errors [13]. Focusing on only one metric will result in a biased AI model. EO applications need community agreed-upon holistic evaluation metrics to develop a path for characterizing research/operational progress, and limits for AI-ready EO datasets and AI models when deployed in real-world applications.

WHAT THE GRSS COMMUNITY IS GOING TO DO

The first step is to propose a framework for cross-modal ARD processing, and to provide definitions of what FAIR means for EO datasets and AI models. The following briefly summarizes the corresponding details and definitions of FAIR elements for EO:

- *Findable*:
  - RS training and validation image metadata should first be standardized through the SpatioTemporal Asset Catalog (STAC) [12] family of specifications to create structured datasheets and easy-to-query formats.
  - Cross-modal datasheets that provide a detailed description of the datasets, including resolution and the number of channels, sensor type, and collection date.
  - Metadata contain STAC-based item identifiers that enable other users to search for data.
  - Datasheets that contain machine-readable keywords, with metadata that are easy for humans and machines to find.
  - Dataset metadata should be written using RS-based attributes similar to Committee on Earth Observation Satellites (CEOS) ARD specifications [14].
  - Develop open-consensus, GRSS-based ARD standards that advance harmonization of ARD datasets by vendors and imagery providers. Coordination can be established to consider CEOS ARD specifications and include broader EO expert contributions, e.g., the GRSS and industry.
- *Accessible*:
  - ARD datasets, including EO benchmarks, should be made available and shared through searchable public repositories, with the data retrievable by standardized interfaces [(application programming interfaces (APIs)], including access by identifier.
  - Human–computer-interaction-searchable repositories with tools (such as, e.g., an Earth observation database [15]) that search cross-modal datasets should be developed.
  - Using STAC specifications, datasheets should be discoverable by humans and machines.
- *Interoperable*:
  - Benchmark EO datasets should be in common formats and standardized through shared ARD best practices.
  - Datasheets and model cards that contain references to training data, hyperparameter settings, validation metrics, and hardware platforms used in experiments.
  - Data fusion experiments should be published in containerized environments to encourage interoperability and reproducibility across computing platforms.
  - Conduct experiments on data fusion using ARD data from multiple sensors accessed using open APIs. The experiments should involve multiple software applications that identify best practices and needed harmonization to lessen analysts' burden in creating fusion workflows.
- *Reusable*:
  - Publishing of data, datasheets, model cards, and model weights should be supported by Jupyter notebooks for quick human interaction to understand and test models on new data.
  - Datasheets that include details in machine-readable format on how data were collected.
  - Establish domain-relevant community standards for AI-based data fusion evaluation methods based on the GRSS IADF and data and algorithm standard evaluation results [16].

After that, standard ARD components must be highlighted while focusing on emerging needs at the nexus of cross-modal, cross-sensor EO/RS; image analysis; and data fusion technologies. Essential components must be identified to establish standards for systematic advancement and provision of image analysis and data fusion methods in the era of AI and big EO data.

The GRSS and, in particular, the IADF and the GSEO will continue to work toward these goals. However, standards do not exist in isolation. There is an application-related context that needs to be respected, existing work that needs to be incorporated, best practices that should be adapted, and communities that need to validate the proposed principles by adhering to and using them. Thus, we actively reach out and invite other groups working toward similar goals to focus our efforts and collaborate as a new standard needs to be created by the community for the community to be successful.

AUTHOR INFORMATION

Dalton Lunga (lungadd@ornl.gov) is with the Oak Ridge National Laboratory, Oak Ridge, TN 37830 USA, and is the IEEE Geoscience and Remote Sensing Society Image Analysis and Data Fusion Working Group on Machine/Deep Learning for Image Analysis lead. He is a Senior Member of IEEE.

Silvia Ullo (ullo@unisannio.it) is with the University of Sannio, 82100, Benevento, Italy, and is the IEEE Geoscience and Remote Sensing Society Image Analysis and Data Fusion Working Group on Machine/Deep Learning for Image Analysis co-lead. She is a Senior Member of IEEE.

Ujjwal Verma (ujjwal.verma@manipal.edu) is with the Department of Electronics and Communication Engineering, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal 576104, India, and is the IEEE Geoscience and Remote Sensing Society Image Analysis and Data Fusion Working Group on Machine/Deep Learning for Image Analysis co-lead. He is a Senior Member of IEEE.

George Percivall (percivall@ieee.org) is with GeoRoundtable, Annapolis, MD 21114 USA, and is the IEEE Geoscience and Remote Sensing Society Technical Committee on Standards for Earth Observation cochair. He is a Senior Member of IEEE.

Fabio Pacifici is with Maxar Technologies Inc, Westminster, CO 80234 USA, and is the IEEE Geoscience and Remote Sensing Society vice president of technical activities. He is a Senior Member of IEEE.

Ronny Hänsch (ronny.haensch@dlr.de) is with the German Aerospace Center, 82234 Weßling, Germany, and is the IEEE Geoscience and Remote Sensing Society Image Analysis and Data Fusion chair. He is a Senior Member of IEEE.

REFERENCES

[1] Z. Ignacio. "Analysis ready data workshops." ARD.Zone. Accessed: Dec. 1, 2022. [Online]. Available: https://www.ard.zone

[2] J. L. Dwyer, D. P. Roy, B. Sauer, C. B. Jenkerson, H. K. Zhang, and L. Lymburner, "Analysis ready data: Enabling analysis of the Landsat archive," *Remote Sens.*, vol. 10, no. 9, Aug. 2018, Art. no. 1363, doi: 10.3390/rs10091363. [Online]. Available: https://www.mdpi.com/2072-4292/10/9/1363

[3] M. D. Wilkinson et al., "The FAIR guiding principles for scientific data management and stewardship," *Scientific Data*, vol. 3, no. 1, Mar. 2016, Art. no. 160018, doi: 10.1038/sdata.2016.18.

[4] "Sensor observation service," Open Geospatial Consortium, Arlington, VA, USA, 2023. [Online]. Available: https://www.ogc.org/standard/sos/

[5] "Standards working groups," Open Geospatial Consortium, Arlington, VA, USA, 2023. [Online]. Available: https://www.ogc.org/about-ogc/committees/swg/

[6] "Data readiness," Earth Science Information Partners, Severna Park, MD, USA, 2023. [Online]. Available: https://wiki.esipfed.org/Data_Readiness

[7] M. Claverie et al., "The harmonized Landsat and sentinel-2 surface reflectance data set," *Remote Sens. Environ.*, vol. 219, pp. 145–161, Dec. 2018, doi: 10.1016/j.rse.2018.09.002. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425718304139

[8] "FAIR for machine learning (FAIR4ML) IG," Research Data Alliance, USA, Australia, Europe, 2023. [Online]. Available: https://www.rd-alliance.org/groups/fair-machine-learning-fair4ml-ig

[9] T. Gebru et al., "Datasheets for datasets," 2018. [Online]. Available: https://arxiv.org/abs/1803.09010

[10] M. Mitchell et al., "Model cards for model reporting," in *Proc. Conf. Fairness, Accountability, Transparency*, Jan. 2019, pp. 220–229, doi: 10.1145/3287560.3287596.

[11] D. Lunga and P. Dias, "Advancing data fusion in earth sciences," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2022, pp. 5077–5080, doi: 10.1109/IGARSS46834.2022.9883176.

[12] J. Rincione and M. Hanson, "CMR SpatioTemporal Asset Catalog (CMR-STAC) documentation," NASA Earth Science, Nat. Aeronaut. Space Admin., Washington, DC, USA, 2021. [Online]. Available: https://wiki.earthdata.nasa.gov/display/ED/CMR+SpatioTemporal+Asset+Catalog+%28CMR-STAC%29+Documentation

[13] B. Cheng, R. Girshick, P. Dollár, A. C. Berg, and A. Kirillov, "Boundary IOU: Improving object-centric image segmentation evaluation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2021, pp. 15,329–15,337, doi: 10.1109/CVPR46437.2021.01508.

[14] "CEOS analysis ready data," Committee on Earth Observation Satellites, France, Canada, USA, Thailand, 2022. [Online]. Available: https://ceos.org/ard/

[15] M. Schmitt, P. Ghamisi, N. Yokoya, and R. Hänsch, "EOD: The IEEE GRSS earth observation database," in *Proc. IEEE Int. Geosci. Remote Sens. Symp. (IGARSS)*, 2022, pp. 5365–5368, doi: 10.1109/IGARSS46834.2022.9884725.

[16] G. I. T. Committee. "GRSS data and algorithm standard evaluation." GRSS DASE. Accessed: May 3, 2023. [Online]. Available: http://dase.grss-ieee.org/