

Official Report on the 2021 Computational and Autonomous Workflows Workshop (CAW 2021)



Sean Wilkinson
Katie Knight
Olga Kuchar
Kshitij Mehta
Arjun Shankar
Matthew Wolf

March 2022



DOCUMENT AVAILABILITY

Reports produced after January 1, 1996, are generally available free via OSTI.GOV.

Website www.osti.gov

Reports produced before January 1, 1996, may be purchased by members of the public from the following source:

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
Telephone 703-605-6000 (1-800-553-6847)
TDD 703-487-4639
Fax 703-605-6900
E-mail info@ntis.gov
Website <http://classic.ntis.gov/>

Reports are available to DOE employees, DOE contractors, Energy Technology Data Exchange representatives, and International Nuclear Information System representatives from the following source:

Office of Scientific and Technical Information
PO Box 62
Oak Ridge, TN 37831
Telephone 865-576-8401
Fax 865-576-5728
E-mail reports@osti.gov
Website <https://www.osti.gov/>

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

National Center for Computational Sciences

**OFFICIAL REPORT ON THE 2021 COMPUTATIONAL AND AUTONOMOUS WORKFLOWS
WORKSHOP (CAW 2021)**

Sean Wilkinson
Katie Knight
Olga Kuchar
Kshitij Mehta
Arjun Shankar
Matthew Wolf

March 2022

Prepared by
OAK RIDGE NATIONAL LABORATORY
Oak Ridge, TN 37831
managed by
UT-BATTELLE LLC
for the
US DEPARTMENT OF ENERGY
under contract DE-AC05-00OR22725

CONTENTS

LIST OF FIGURES	v
ACRONYMS AND ABBREVIATIONS	vi
DEFINITIONS.....	viii
EXECUTIVE SUMMARY.....	x
ACKNOWLEDGMENTS	xii
1. INTRODUCTION	1
1.1 BRIEF EXPLANATION OF THE FAIR PRINCIPLES.....	1
1.2 FAIR AS A MOVEMENT	2
1.3 SIMILAR INITIATIVES.....	3
1.4 WORKFLOWS.....	4
1.5 FAIR WORKFLOWS.....	4
1.6 ORNL FAIR WORKFLOW ACTIVITIES.....	5
2. WORKSHOP STRUCTURE, GOALS, AND FINDINGS	6
2.1 WORKSHOP STRUCTURE.....	6
2.2 GOALS	7
2.3 SUMMARIES OF KEYNOTE PRESENTATIONS	7
2.3.1 FAIR Workflows: A step closer to the Scientific Paper of the Future (D. Garijo).....	7
2.3.2 The FAIR+ World According to Me (C. Kirkpatrick).....	8
2.4 LIGHTNING TALKS.....	9
2.5 WORKFLOW SHOWCASE FOR FAIR COLLABORATION	9
2.6 BREAKOUT SESSIONS	10
2.6.1 Streaming Workflows	10
2.6.2 Reusability and Interoperability.....	12
2.6.3 Workflow Lifecycle	14
2.6.4 Workflow Services.....	17
2.6.5 Metrics for FAIR.....	18
3. RESEARCH CHALLENGES	22
3.1 COMMON UNDERSTANDING.....	22
3.2 RICH METADATA FOR WORKFLOWS	23
3.3 FAIR WORKFLOW DEVELOPMENT	24
3.4 PATTERNS AND POLICIES	24
3.5 AUTOMATABLE METADATA.....	25
3.6 WORKFLOW REPOSITORIES	26
4. ORNL CHALLENGES	28
4.1 FAIR EDUCATION	28
4.2 INFRASTRUCTURE	29
4.3 CULTURE	30
5. OLCF AS THE HUB FOR FAIR WORKFLOWS	32
5.1 LEADING DOE EFFORTS.....	32
5.2 SERVING THE ORNL COMMUNITY.....	33
6. CONCLUSIONS	35
6.1 RECOMMENDATIONS FOR DOE	35
6.2 RECOMMENDATIONS FOR ORNL	36
6.3 RECOMMENDATIONS FOR OLCF	37
6.4 NEXT STEPS	37
REFERENCES.....	38
APPENDIX A. LIST OF PARTICIPANTS	A-1
APPENDIX B. WORKSHOP AGENDA.....	B-1

APPENDIX C. LIGHTNING TALK ABSTRACTS	C-1
APPENDIX D. SLIDE IMAGES	D-1

LIST OF FIGURES

Figure 1. FAIR principles and sub-principles [5].	2
Figure 2. Concrete examples of FAIR Digital Objects at ORNL (credit: K. Knight lightning talk).....	28

ACRONYMS AND ABBREVIATIONS

API	Application Programming Interface
ARM	Atmospheric Radiation Measurement
ASCR	Advanced Scientific Computing Research
ATLAS	A Toroidal LHC ApparatuS
BESSD	Biological and Environmental Systems Science Directorate
CARE	Collective benefit, Authority to control, Responsibility, and Ethics
CAW	Computational and Autonomous Workflows
CCSD	Computing and Computational Sciences Directorate
CERN	European Organization for Nuclear Research
CI	Cyberinfrastructure
CI/CD	Continuous Integration / Continuous Delivery (or Continuous Deployment)
CMR	Common Metadata Repository
CNMS	Center for Nanophase Materials Science
CODATA	National Academies of Sciences' U.S. National Committee for the Committee on Data
CSV	Comma-Separated Values
CWL	Common Workflow Language
DAAC	Distributed Active Archive Center
DAG	Directed Acyclic Graph
DKL	Deep Kernel Learning
DOE	United States Department of Energy
DOI	Digital Object Identifier
ECO	EarthCube Office
EELS	Electron Energy Loss Spectroscopy
ELIT	Ensemble Learning Iterative Training
EOSC	European Open Science Cloud
EOSDIS	Earth Observing Data and Information Systems
ESDIS	Earth Science Data and Information System
ESTD	Energy Science and Technology Directorate
FAIR	Findable, Accessible, Interoperable, Reusable
FAIR4RS	FAIR for Research Software
FDO	FAIR Digital Object
FIP	FAIR Implementation Profile
GFISCO	GO FAIR International Support and Coordination Office
GPU	Graphics Processing Unit
HFIR	High Flux Isotope Reactor
HPC	High Performance Computing
IN	Implementation Network
INTERSECT	Interconnected Science Ecosystem
iRF-LOOP	Iterative Random Forest Leave One Out Prediction
KBase	DOE Systems Biology Knowledgebase
LDRD	Laboratory Directed Research and Development
LHC	Large Hadron Collider
ML	Machine Learning
MRA	Microbiology Resource Announcement
NASA	National Aeronautics and Space Administration
NCCS	National Center for Computational Sciences
NScD	Neutron Sciences Directorate
NSSD	National Security Sciences Directorate

OLCF	Oak Ridge Leadership Computing Facility
ORNL	Oak Ridge National Laboratory
OSTI	Office of Scientific and Technical Information
PFM	Piezoresponse Force Microscopy
PI	Principal Investigator
PSD	Physical Sciences Directorate
PuRe	Publicly Reusable
RDA	Research Data Alliance
REP	Reproducibility Enhancement Principles
RFP	Request for Proposal
SDSC	San Diego Supercomputer Center
Shift-VAE	Shift-invariant variational autoencoder
SNS	Spallation Neutron Source
STEM	Scanning Transmission Electron Microscope
TAB	Technical Advisory Board
TOP	Transparency and Openness Promotion
UCSD	University of California San Diego
UPM	Universidad Politécnica de Madrid
URL	Uniform Resource Locator
VISTA	Visual Informatics for Science and Technology Advances
VODAN	Virus Outbreak Data Network
W3C	World Wide Web Consortium
WfMS	Workflow Management System
X-AI	Explainable Artificial Intelligence

DEFINITIONS

accessible	a property ascribed to data when the following are satisfied: (meta)data are retrievable by their identifier using a standardized communications protocol; the protocol is open, free, and universally implementable; the protocol allows for an authentication and authorization procedure, where necessary; and metadata are accessible, even when the data are no longer available.
algorithm	a set or sequence of rules followed to solve a problem, especially in calculations by a computer.
automation	the process of removing the human from the loop and delegating to machines.
autonomous workflow	also known as an “autonomic workflow”; a workflow which can modify itself to react to opportunities or anomalies generated at run-time [1].
digital object	an abstraction that can refer to any kind of digital information that can be treated as a single object composed of bit sequences.
DOI	an acronym for Digital Object Identifier; is a string of numbers, letters, and symbols used to uniquely identify an article, document, or data, and to provide it with a permanent web address (URL); to easily locate a digital object from your citation.
FAIR	Findable, Accessible, Interoperable, and Reusable data principles.
findable	a property ascribed to data when the following are satisfied: (meta)data are assigned a globally unique and persistent identifier; data are described with rich metadata; metadata clearly and explicitly include the identifier of the data it describes; and (meta)data are registered or indexed in a searchable resource.
human-in-the-loop	describes a workflow which requires human interaction and/or decision making, often in a way which halts execution while waiting for human input.
interoperable	a property ascribed to data when the following are satisfied: (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation; (meta)data use vocabularies that follow FAIR principles; and (meta)data include qualified references to other (meta)data.
metadata	information about the data, such as the protocol used to create the data.
pipeline	a set of processing elements connected in series.
reusable	a property ascribed to data when the following are satisfied: meta(data) are richly described with a plurality of accurate and relevant attributes; (meta)data are released with a clear and accessible data usage license;

(meta)data are associated with detailed provenance; and (meta)data meet domain-relevant community standards.

scalability

a concept for describing a change in performance with respect to a change in resources; in HPC, there are multiple definitions for scaling, including strong scaling (total problem size is fixed) and weak scaling (problem size per CPU core is fixed) [2].

streaming data

also known as event stream data; data which is generated continuously by different sources.

workflow

a sequence of transformations applied to inputs to produce a set of outputs; in computing, workflows are typically associated with automation.

workflow lifecycle

the sequence of stages of a workflow's development.

EXECUTIVE SUMMARY

In July 2021, the Oak Ridge National Laboratory (ORNL) convened the second Computational and Autonomous Workflows (CAW) workshop. CAW was a virtual workshop held over two half-days that brought together 62 scientists and engineers from six laboratory directorates to discuss how the Findable, Accessible, Interoperable, and Reusable (FAIR) data principles for scientific data management and stewardship apply to computational and autonomous workflows. Together, we explored the workshop's theme of "FAIR workflows" from different perspectives, diving into topics such as: metrics surrounding FAIR workflows; defining a workflow lifecycle; challenges of streaming workflows with FAIR; and what it means for workflows to be interoperable and reusable.

From our discussions, the following challenge questions were identified for the United States Department of Energy (DOE):

- What are the community-accepted definitions, terms, and vocabulary for workflows? *Recommendation:* develop and publish a scientific workflow reference that defines, describes, and provides examples to build upon new abstractions for interoperability and reusability.
- What should "publishing a workflow" mean, with associated quality metrics, provenance, and an understanding of longevity? *Recommendation:* 1) collect real-world examples of successfully sharing and using workflows; 2) research a new model for the lifecycle of workflows, including guidance on versioning, metrics for quality, and guidelines for preservation.
- Why is searching for published scientific workflows so difficult? *Recommendation:* 1) fund a working group and publish a workflow metadata schema to address "F" - Findability - in FAIR; and 2) fund domain-specific workflow working groups to develop an understanding of metadata "richness" within their communities' workflow usages.
- Are there common design patterns and policies for DOE science workflows? *Recommendation:* 1) define workflow characteristics and maintain a list of workflow tools and their capabilities based on these characteristics; 2) using workflow patterns, develop a workflow framework / ecosystem for plug-and-play components; 3) research new approaches in using and evaluating policy-driven workflow executions.
- Can we improve metadata collection and make it more useful? *Recommendation:* research, evaluate, and provide metrics for automating the creation of metadata that is both human-readable and machine-actionable.

Additionally, several challenge questions were discussed that specifically apply to ORNL:

- How do we improve literacy on data, workflows, and FAIR? *Recommendation:* 1) create tutorials and other educational materials to educate ORNL staff; 2) develop an internal workflows website to help our scientists and engineers find appropriate tools and resources across the lab; 3) regular onsite training on existing workflow tools, systems, and use cases.
- What existing infrastructure is available that supports science workflows at ORNL? *Recommendation:* 1) build a workflows community that helps others repurpose our own workflows; 2) build a workflow repository.

- How do we support a culture of open science? *Recommendation:* as an institution, invest in building a science culture committed to FAIR.

Finally, the Oak Ridge Leadership Computing Facility (OLCF) provides the world's most advanced computing systems for the open science community, including industry, and engages the scientific and engineering communities to advance science and technology research in the United States. Some key cross-cutting challenge questions were discussed:

- How can OLCF lead DOE efforts in FAIR workflows? *Recommendation:* facilitate solving the DOE challenges by coordinating the diverse OLCF user base.
- What role can OLCF play for the ORNL workflows community? *Recommendation:* serve as the hub for FAIR and workflows at ORNL.

The remainder of this report is organized into 6 sections.

Section 1 is an introduction that provides background material for this workshop, by outlining the FAIR principles and the movement and initiatives surrounding them, as well as providing comparison with other data initiatives like PuRe and CARE. It quickly summarizes workflows, applying FAIR to workflows, and FAIR workflows at ORNL.

Section 2 is an overview of the content presented at the workshop. It includes the workshop structure and goals as well as summaries of the 2 keynote presentations by external speakers, the 13 lightning talks by ORNL speakers, an extended “workflow showcase” talk, and the five breakout session tracks: streaming workflows, reusability and interoperability, workflow lifecycle, workflow services, and metrics for FAIR.

Section 3 summarizes challenges in the research community that DOE Advanced Scientific Computing Research (ASCR) may want to address. It details some of the most important research challenges that were identified in the workshop discussions.

Section 4 summarizes challenges specific to ORNL regarding FAIR workflows.

Section 5 addresses the cross-cutting challenges for OLCF based on recommendations for DOE and ORNL.

Section 6 concludes this report with a summary of the key highlights and recommendations.

ACKNOWLEDGMENTS

This research used resources of the Oak Ridge Leadership Computing Facility at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725.

We would also like to acknowledge and thank Swen Boehm, Jong Choi, Yan Liu, Ketan Maheshwari, and Bruce Wilson for reviewing and providing feedback for this report.

1. INTRODUCTION

In 2016, Mark Wilkinson et al. published a paper, “The FAIR Guiding Principles for scientific data management and stewardship”, that gave a memorable acronym to decades of research and community efforts [3]. “FAIR” stands for “Findable, Accessible, Interoperable, Reusable”, which are principles that place emphasis not only on supporting data reuse by humans, but also on enhancing the ability of machines to find and use the data automatically. The FAIR principles have been well-received globally, and adoption is increasing, but the computational and autonomous workflows which find and use the data automatically have remained very difficult to construct. Additionally, although the FAIR principles are explicitly intended to be applied to workflows as well as data and other digital objects related to research in Wilkinson’s paper, open questions exist about how to apply FAIR to workflows. For example, many workflow components can be expressed as data, but workflows are more than just data – they have a process component. As descriptions of processes, workflows inherit properties of FAIR data, but as executable processes, they also inherit properties of software [4]. This makes applying the FAIR principles directly to workflows more complicated than applying them to data.

The remainder of this section begins by outlining the FAIR principles, the FAIR movement, and FAIR initiatives, as well as comparing with other data initiatives like PuRe and CARE. Then, this section handles the topics of workflows, applying FAIR to workflows, and FAIR workflows at ORNL.

1.1 BRIEF EXPLANATION OF THE FAIR PRINCIPLES

The FAIR Guiding Principles describe expected behaviors from digital objects (e.g., datasets, workflows) that make them Findable, Accessible, Interoperable and Reusable by machines. These principles offer a foundational layer for data management but avoid formal definitions about how to fulfill each element, leaving it up to the data provider as to how best to implement (or not), based on domain needs and constraints.

- *Findable*: in short, data and metadata should be discoverable. The principles assert that data/metadata should clearly and explicitly include a globally unique identifier. The data should also be described with rich metadata that is registered or indexed in a searchable resource. "Rich metadata" is understandably (and necessarily) vague, as are some of the other findability requirements, as various domain stakeholders will have different needs and requirements.
- *Accessible*: metadata are perpetually retrievable via the globally unique identifier using some standardized communications protocol that is also open, free, universally implementable and allows for an authentication and authorization procedure.
- *Interoperable*: metadata uses a formal, accessible, shared, and broadly acceptable language for knowledge retrieval. Metadata vocabularies follow the FAIR principles and include qualified references to other metadata.
- *Reusable*: metadata are richly described according to domain-relevant community standards with a plurality of accurate/relevant attributes (see also “Findable”), including a clear data usage license and detailed provenance information.

Although FAIR has only 4 letters, there are actually 15 principles that comprise FAIR, and these are treated as sub-principles for the main 4 principles, as illustrated in Figure 1. The FAIR principles specifically emphasize enhancing the ability of machines to automatically find and use data; thus, they are designed as guidelines for scientific data management and stewardship, and enabling automation by machines has major implications for workflows, too. The FAIR principles intentionally avoid any strict definition of how "FAIRness" should be achieved, as this is necessarily domain-specific [3]. Digital resource features will be variously Findable, Accessible, Interoperable, and Reusable according to the needs and restrictions of a domain, and becoming fully FAIR should be taken as aspirational, not prescriptive.



Figure 1. FAIR principles and sub-principles [5].

1.2 FAIR AS A MOVEMENT

The GO FAIR initiative [6] is a self-governed and member-driven community that aims to help individuals and organizations implement the FAIR data principles by working together through various Implementation Networks (INs). These INs are composed of three activity pillars: GO BUILD, GO CHANGE, and GO TRAIN. Furthermore, GO FAIR has an International Support and Coordination Office (GFISCO), internationally operating at three sites (Paris, Hamburg, and Leiden), which is funded by the Ministries of Science of France, Germany, and the Netherlands, as well as National Support and Coordination Offices (e.g., GO FAIR US [7]).

GO BUILD focuses on technological aspects of FAIR. This pillar supports and coordinates community activities around improving adoption of globally unique and persistent identifiers, agreeing on common metadata representation formats, agreeing on a minimal set of generic metadata content, defining domain-relevant community standards, and designing and providing reference implementation for various services (e.g., model or software repositories, data search engines, and other FAIR-compliant services).

GO CHANGE intends to optimize coordination between existing and new FAIR initiatives and activities. This pillar is more of a “top down” or “horizontal” governance, helping to trace and steer progress, such as helping to prevent fragmentation, promote culture change and advocacy, and advocate for data management plans (conceptualization, composition, evaluation, and monitoring). Overall, GO CHANGE provides assistance and guidance in achieving these goals, focusing on culture change via partnerships, funding, assessment, and best practices.

GO TRAIN focuses on FAIR awareness and skills development training. It aims to provide a platform for FAIR conversations featuring the latest developments from the community, provide examples of pragmatic how-to-approaches and use-cases addressing FAIR from researchers and research support staff perspectives, capture and share the discussion outcomes in discoverable educational format i.e., bite-sized FAIR lesson material, help make FAIR more accessible to broader audiences, and explore FAIR practices from a global perspective.

1.3 SIMILAR INITIATIVES

The Office of Science’s Publicly Reusable (PuRe) data [8] consists of various office-sponsored repositories and analytics platforms that intend to make data both public and reusable. For resources to be classified as PuRe, they must be sponsored by an Office of Science program office, considered to be an authoritative provider of data or capabilities in the area, publicly available, regularly peer-reviewed, and include a data management plan that provides preservation and stewardship planning. Not all FAIR data principles are required for a resource to be designated PuRe.

The CARE Principles for Indigenous Data Governance [9] represent another related movement in which the FAIR principles are involved, but the problem to be solved is very different. Where FAIR is more focused on open data and open science, CARE is more focused on the data rights of indigenous peoples and their interests. CARE is complementary to FAIR, but its core principles – Collective benefit, Authority to control, Responsibility, and Ethics – are more closely related to social issues than scientific ones.

The Guidelines for Transparency and Openness Promotion (TOP) were designed as author guidelines for scientific journals that could help to promote transparency, openness, and reproducibility, all three of which are recognized as vital features of science [10]. The authors stated that most scientists embrace these features as disciplinary norms and values, but that a growing body of evidence suggests that the valued features are not routine in daily practice. Thus, the TOP guidelines can be viewed as related to the FAIR principles, but with more of a focus on reforming scientific journals.

The Reproducibility Enhancement Principles (REP) were presented as a novel set of principles targeting disclosure challenges involving computation. These recommendations, which built upon more general proposals from the TOP guidelines as well as recommendations for field data, emerged from workshop discussions among funding agencies, publishers and journal editors, industry participants, and researchers representing a broad range of domains. The main idea is that data, code, and workflows should be available and cited to “move toward ameliorating irreproducibility in computational research” [11].

1.4 WORKFLOWS

The word “workflow” means many things to different people, depending on where they first learned the word. The layperson’s definition of a workflow is often closely associated with business processes and flowcharts, but in scientific computing, workflows are most closely associated with automation. Even in computing, however, “workflow” can be an overloaded term.

The workflows community uses a wide range of definitions for workflows [12]. Workflows may involve data movement operations over different types of computer networks, protocols, and connectivity; orchestrating computational tasks across heterogeneous resources such as High Performance Computing (HPC), cloud, and edge; and interacting with a wide range of distributed services. Workflows are often defined as sets of computational tasks expressed as directed acyclic graphs with data dependencies. Another view of workflows defines them with higher-level abstractions which describe sets of interactions between services and/or entities, such as computing centers and scientific facilities. As a result, computational workflows may be simply defined in Linux Bash scripts, complex systems that execute entire scientific campaigns across globally distributed resources, and everything in between.

Our interest in workflows research at ORNL comes from the increasing complexity involved in correctly performing the scientific method with modern considerations like data publishing and reproducibility, along with the sheer difficulties in modern data-intensive science that Big Data presents. Automating as many of these practices as possible ensures that ORNL’s domain scientists can focus their energies on the original pillars of science: theory and experiment.

1.5 FAIR WORKFLOWS

Goble et. al [4] propose two areas in which the FAIR principles apply to workflows: 1) FAIR data both for and from workflows, where workflows will include descriptive metadata about the data produced as well as metadata that helps trace that data’s provenance; and 2) criteria for FAIR digital objects, where a workflow is seen as an “object” that describes methodology that may be subsequently distributed, used, cited, and modified.

The authors note that “[d]etermining whether the data produced by a workflow is FAIR is not straightforward and requires concrete criteria, which should be provided by both the FAIR indicators and the workflow specification”, so there is still research needed to understand what criteria or metrics are necessary to produce FAIR data from any given workflow. Data identifiers, access, and licensing agreements were marked as particular challenges for workflows as data may be variously combined and generated throughout the workflow process.

Regarding workflows as FAIR Digital Objects, they write that, “[a]s *descriptions* of processes workflows inherit properties of FAIR data, but as *executable* processes they inherit properties of *software*. Workflows as processes challenge the FAIR principles by their structure, forms, versioning, executability, and reuse”.

Thus, following Goble’s numbering, our interest in FAIR workflows means that we are interested in applying the FAIR principles to 1) the data being handled within the workflows and 2) the workflows themselves, considered as digital objects that represent processes. We know that the FAIR principles provide positive returns on investment when applied to data, and we argue that the FAIR principles would provide similar returns when applied to the workflows themselves. Some specific examples include citations by other consumers of published workflows, bug reports and fixes from the open-source community which improve the workflows themselves, and reduced redundant development effort.

1.6 ORNL FAIR WORKFLOW ACTIVITIES

Matthew Wolf et al. have started an investigation of richer, automatable metadata for FAIR workflows [13]. A key observation from their paper is that reusability is particularly relevant for workflows. Although dataset reusability is frequently restricted to questions about licensing and allowable/auditable reuse, this work contends that by the very nature of workflows, this definition is insufficient for the “open science” intentions of FAIR. Workflows represent the connections between input datasets, computational components, and potentially large numbers of intermediate datasets and routing decisions. Reuse of that connection context requires understanding more than just its license.

Moving FAIR metadata from human-focused auditing to automation is argued as a natural and important evolution of the FAIR standards. Just as most data is generated by computers for computers, so too should most metadata be generated by and for the computational workflows and runtime environments. Autonomous workflows, based on an open ecosystem of tools and services, offer advantages to individual users, but more importantly serve as tools for improving the community-oriented software engineering, management, and sharing of workflow concepts. The paper discusses some initial work on identifying categories of automatable and semi-automatable metadata for reusability, but it also identifies the opportunity for significant further community research and refinement.

2. WORKSHOP STRUCTURE, GOALS, AND FINDINGS

CAW 2021 was a virtual workshop event with a theme of FAIR workflows, held over two half-days during the mornings of July 20 and 21, 2021. This workshop convened many members of the ORNL workflows community – 62 scientists from 6 directorates at ORNL – as well as two renowned external keynote speakers. In addition to the keynote presentations, there were 13 lightning talks and one extended talk (a “workflow showcase”) by members of the ORNL workflows community which helped set context and seed discussion for the 5 breakout session tracks.

In this section of the report, a summary of the workshop is presented, including the structure, goals, summary of keynote presentations and lightning talks, and summaries of the five breakout sessions.

2.1 WORKSHOP STRUCTURE

The CAW 2021 virtual workshop convened 62 scientists from ORNL over two half-days via Zoom teleconferencing software [14]. These experts represented six directorates at ORNL: Biological and Environmental Systems Science (BESSD), Computing and Computational Sciences (CCSD), Energy Science and Technology (ESTD), National Security Sciences (NSSD), Neutron Sciences (NScD), and Physical Sciences (PSD). Workshop participants are listed in APPENDIX A.

The first day of the workshop opened with an introduction from ORNL’s Gina Tourassi, the Director of the National Center of Computational Sciences (NCCS) and the Oak Ridge Leadership Computing Facility (OLCF) at ORNL. This was followed by a keynote delivered by invited external speaker Daniel Garijo of Universidad Politécnica de Madrid (UPM). Dr. Garijo’s talk is summarized in Section 2.3.1. Then, six lightning talks were delivered by our ORNL colleagues, selected specifically from submitted abstracts in order to present FAIR Workflows from the perspective of computer science. Details about these lightning talks are available in Section 2.4. These lightning talks were also chosen to set the stage for the day’s five breakout sessions, which are detailed in Section 2.6. Following the breakout sessions, participants returned to a plenary session for out briefing before the adjourning for the day.

The second day of the workshop opened with a brief summary of the previous day’s events before proceeding to seven more lightning talks which focused more closely on the perspective of the domain scientist working with scientific applications. Details about these lightning talks are available in Section 2.4. These talks were followed by an extended talk by ORNL’s Bruce Wilson, outlined in Section 2.5, which was intended to showcase a specific set of workflows for FAIR collaboration, namely those from the Earth and Environmental Sciences. Then, participants headed to breakout sessions again which had the same topics as the first day, with the stipulation that they should each head to a different session on the second day. The idea was to repeat the session topics each day with different participants in order to expand and elaborate on the topics as further talks and discussion helped thoughts evolve. These breakout sessions are summarized in Section 2.6. Following the breakout sessions, participants returned to a plenary session for outbriefing before the adjourning for the day. Finally, the workshop closed with a keynote delivered by invited external speaker Christine Kirkpatrick of the San Diego Supercomputer Center, summarized in Section 2.3.2.

Discussions in the breakouts and final plenary session were highly interactive. Participants were encouraged to form small groups for brainstorming and exploring ideas in depth. The facilitators of the breakout sessions set expectations, guided conversations, and kept the participants on schedule. The discussions were mediated with the help of the freely available Google Docs web application [15] for collaborative online word processing. Organizers and scribes took care to record the conversation artifacts.

The full agenda is included in APPENDIX B.

2.2 GOALS

One goal of this workshop was to bring together the ORNL workflows community for the purpose of discussing the FAIR principles, computational and autonomous workflows, and how to improve ORNL research specifically by applying the FAIR principles to workflows. Another goal of convening the community was to unite the community, especially in a way that would lead to the establishment of a community of practice as well as a working group that would meet periodically to discuss and apply the FAIR principles to group members' workflows.

The workshop was designed to be inclusive and accessible for all research-minded personnel at ORNL, regardless of the level of experience or specific expertise in topics like the FAIR principles or workflows. The FAIR principles and workflows, as topics, are generally applicable to all data-driven science, and thus the organizers assumed that attendees would be interested in exposure to ideas and concepts that might not be familiar to them.

2.3 SUMMARIES OF KEYNOTE PRESENTATIONS

This section summarizes the presentations of our keynote speakers, Daniel Garijo and Christine Kirkpatrick. Both speakers were invited from external institutions in order to provide new perspective that would inform and inspire our ORNL colleagues for the breakout sessions.

2.3.1 FAIR Workflows: A step closer to the Scientific Paper of the Future (D. Garijo)

Daniel Garijo Verdejo is a distinguished researcher at the Universidad Politécnica de Madrid, where he also completed his PhD. Previously, he held a Research Computer Scientist position at the Information Sciences Institute of the University of Southern California, in Los Angeles. Dr. Garijo's line of research focuses on the area of e-Science, in particular on capturing the context and metadata of scientific software and computational experiments to foster their (re)usability. To do so, all the elements belonging to an experiment are converted to Knowledge Graphs and exposed as Linked Data, in both human-readable and machine-readable manner. During his career, Dr. Garijo has participated in community initiatives critical for establishing or promoting the Findable, Accessible, Interoperable and Reusable (FAIR) principles, such as W3C [16] standardization processes (e.g., Provenance Incubator Group [17], Provenance Working Group [18]), the Dublin Core Metadata Initiative [19] and the Research Data Alliance FAIR for Research Software (FAIR4RS) [20].

Dr. Garijo's talk focused primarily on reproducibility, its importance to the practice of the scientific method, and current research to improve reproducibility. He began by explaining his own personal experience and difficulties in trying to reproduce results from the scientific literature, and then he provided an overview of changes in the scientific community that would have helped, like movements towards open data, open-source software, and open access publications. After a quick review of the FAIR principles, he proceeded to break down the "anatomy of a workflow" and explain how the FAIR principles could be applied to each piece. For example, data is a crucial part of a workflow – be it input data, output data, or intermediate results – but software is important, too, and it cannot simply be treated as data, as he wrote with Carole Goble [4]. During most of the talk, he presented problems for which active research is underway, but for which solutions are not yet available.

He also presented material about two more efforts toward fully reproducible science, which were RO-Crate [21] and the Scientific Paper of the Future Initiative [22]. RO-Crate is a community effort to establish a lightweight approach to packaging research data with their metadata. The Scientific Paper of

the Future Initiative was inspired by a geosciences paper with a similar title [23] which specifies guidelines for making reusable data, reusable software, and results which are documented with their computational provenance.

The full talk is available online as a video [24].

2.3.2 The FAIR+ World According to Me (C. Kirkpatrick)

Christine Kirkpatrick is Division Director of San Diego Supercomputer Center’s (SDSC) Research Data Services, which manages infrastructure, networking, and services for research projects of regional and national scope. Ms. Kirkpatrick’s expertise is in the implementation of research computing services and operational cyberinfrastructure (CI) at scale. Ms. Kirkpatrick founded the US GO FAIR Office, which is hosted at SDSC, and she is also on the leadership of the West Big Data Innovation Hub, the Open Storage Network, and the EarthCube Office (ECO), for which she serves as Principal Investigator. She serves on the Technical Advisory Board (TAB) for the Research Data Alliance (RDA), the external Advisory Board for the European Open Science Cloud (EOSC) Nordic, and the National Academies of Sciences’ U.S. National Committee for the Committee on Data (CODATA).

She began by explaining that she titled her talk “FAIR+” because she intended to speak not only about the FAIR principles but also to sneak in a little bit about reproducibility.

Where Dr. Garijo had spoken more from a workflows perspective, Ms. Kirkpatrick spoke from more of a data stewardship perspective. She described the “Fourth Paradigm of Science”, which is data-intensive science [25], with the other three being experimental, theoretical, and computational science. She highlighted the difficulties of data-intensive science by relating that data scientists reportedly devote 79% of their time to obtaining and cleaning data [26] and that the reality of data science is that it is a lot of work.

Ms. Kirkpatrick advised that she personally does not focus on the four letters that comprise the FAIR acronym; she focuses on the 15 principles behind them. These principles can be used to create FAIR maturity indicators. She emphasized explanations of what FAIR is as well as what FAIR is not, because FAIR is not all-encompassing. For example, FAIR has nothing to do with data validation or data quality, because, as she said, “you can have really FAIR data that is still garbage.” FAIR also does not encompass reproducibility, and FAIR does not imply or require Open Science or Open Data; many who are active in the FAIR community work in healthcare fields with patient data.

Ms. Kirkpatrick also gave a brief history of data-intensive science before FAIR, from the Semantic Web (Berners-Lee 2001) to Lightweight Linked Data [27] and Schema.org [28] to the current efforts towards FAIR Digital Objects [29]. She summarized how FAIR is being adopted and practiced all over the world, such as by Horizon 2020 [30] and the European Open Science Cloud (EOSC) [31] in the European Union and the Virus Outbreak Data Network (VODAN) [32] in Africa and Asia. She also summarized a “who’s who” list in Data Consortia, including the Research Data Alliance [33], CODATA [34], GO FAIR [6], and the World Data System [35], which she mentioned will be hosted in Tennessee for the next five years. She also publicly recognized several members of ORNL for being luminaries in the FAIR community, including Stanton Martin, Giri Prakash, and Katie Knight.

The full talk is available online as a video [36].

2.4 LIGHTNING TALKS

A total of 13 lightning talks were given, and all speakers were all from ORNL. The format for the lightning talks consisted of a 5-minute presentation with a target of 5 slides, the last of which would specifically apply the information to ORNL. The talks were organized into two sets: computer science and domain science.

On the first day, the six talks in the computer science set were given. The titles and speakers for the talks were:

- “Putting Reusability First in FAIR Workflows” — Matthew Wolf
- “Why Workflow Management Is Not the Problem” — David Rogers
- “Toward Fair Digital Objects for Workflows” — Katie Knight
- “A Critical Take on Workflows and FAIR” — Ketan Maheshwari
- “In-line vs. In-transit In Situ: Which Technique to Use at Scale?” — James Kress
- “Scientific Visualization as a Service” — David Pugmire

On the second day, the seven talks in the domain science set were given. The titles and speakers for the talks were:

- “FAIR Workflows in the DOE Systems Biology Knowledgebase” — Zach Crockett
- “DOE User Facilities Implementing FAIR Data Principles: ARM Data Center Example” — Ranjeet Devarakonda
- “Improving Reusability and Accessibility of iRF-LOOP using the Cheetah-Savanna Suite of Workflow Tools” — Angelica M. Walker
- “A workflow for neutron scattering data analysis” — Yongqiang Cheng
- “From Microscopic Images to Simulations via Deep Learning: A Comprehensive Workflow to Develop Physics-Based Understandings” — Ayana Ghosh
- “Experimental discovery of structure-property relationships in ferroelectric materials via active learning” — Yongtao Liu
- “Automated Discovery of Physics in the Electron Microscope” — Kevin Roccapiore

Video of the lightning talks from Day 1 is available online [37], and video of the lightning talks from Day 2 is also available online [38]. The original abstracts can be found in APPENDIX C, and the original slides can be found in APPENDIX D.

2.5 WORKFLOW SHOWCASE FOR FAIR COLLABORATION

On the second half-day of the workshop, following the breakout sessions, Bruce Wilson delivered a “workflow showcase” on the topic of “FAIR and Workflows in the Earth and Environmental Sciences.” The full talk is available online as a video [39].

The biggest motivation for this talk is the idea that ORNL would form a Community of Practice revolving around the FAIR principles, and this talk would provide exposition for the first workflow that the community would collaborate to “FAIRify” in periodic meetings following the workshop. Over time, the

plan is for the community to analyze workflows shared by other members of the community to provide practice and support in identifying and implementing FAIR practices within workflows. Bruce Wilson graciously agreed to provide an example for the first showcase.

This talk began with the ORNL Distributed Active Archive Center (ORNL DAAC) and how it seeks to observe and implement the FAIR principles. The ORNL DAAC is one of twelve geographically dispersed DAACs which make up NASA’s Earth Observing System Data and Information Systems (EOSDIS). The Earth Science Data and Information System (ESDIS) project that runs the information system EOSDIS. The different DAACs have different areas of expertise, and the ORNL DAAC specializes in biogeochemical dynamics, ecological data, and environmental processes. The speaker related that FAIR is a goal and a journey, but that no one can ever get to a “perfectly FAIR” state because there is always a little more that can be done. He also stressed that machine accessibility has always been the priority of FAIR, with human accessibility viewed as a collateral benefit. ESDIS engages in frequent self-assessments that inform decision-making about investing additional resources in FAIR efficiently. Their most recent self-assessment concluded that their biggest gap is the *Interoperability* criterion that metadata include references to other metadata. Additionally, FAIRness for automated use lags behind FAIRness for human interactive use. EOSDIS is migrating to Earthdata Cloud, implemented in Amazon Web Services. Unfortunately, this comes with a host of interesting issues, including unpredictable billing that is incurred when users download data from the cloud.

Next, the speaker shifted focus from hosting a data repository to a project, Daymet [40] in which ORNL is a data producer. He detailed several important datasets, including one which contains 40 years of daily surface weather data in the United States, taken at 1 km resolution. He also described a variety of tools and capabilities that Daymet provides for its users to help make the data useful, as well as emerging access methods which are still experimental. He also mentioned the NASA Common Metadata Repository (CMR) [41] and the steps taken toward focusing not only on the FAIRness of data, but also on the workflows that interact with the data.

In summary, Bruce Wilson’s talk communicated that the Earth Sciences have a long tradition of open data sharing, and many organizations are committed to FAIR and open data. ORNL is a part of that history in many ways. For example, ORNL has collaborated with NASA for decades not only in hosting data through the DAAC, but also in acting as the producer and distributor for Daymet. NASA Earth Observation data is moving to the cloud, both to enable analysis in-place by sending compute to the data, as well as to standardize data access methods.

2.6 BREAKOUT SESSIONS

There were five focus areas for breakout sessions: streaming workflows, reusability and interoperability, workflow lifecycle, workflow services, and metrics for FAIR. Attendees voted these topics higher than all others before the workshop from a list provided by the workshop organizers as part of the workshop registration process. The reason for allowing the topics to be, in a sense, “crowdsourced”, was because this would maximize the interest in the topics and thereby encourage attendees to participate actively; moderating this process by allowing selection from a list also allowed organizers to ensure breakout topics would be focused on FAIR and workflows. This desire to focus strongly on topics of both interest and relevance also led to the decision to hold breakout sessions on consecutive days with the same topics but different attendees.

2.6.1 Streaming Workflows

The breakout sessions about streaming workflows were led by Matthew Wolf. These discussions focused on the following questions:

- What does “streaming workflows” mean for your science/research/practice?
- What is the current state of the practice (at ORNL and other places, as well, if known)?
- What are the requirements?
- What are the gaps?
- What are the next steps? (prototypes, software projects that are envisioned, RFP, et al.)

The remainder of this section summarizes our discussions and highlights key points.

Discussion

The attendees focused on the topics of what constitutes a streaming workflow, how it is different from other types of workflows, and some of the issues that these differences cause for FAIR. Attendees had a variety of experiences with streaming data, from limited exposure at the very end of a streaming pipeline to being intimately involved in real-time requirements for low-level signal processing. Correspondingly, the nature of streams varied considerably, from something that looked more like a publish/subscribe delivery system on the one end to continuous streams of raw binary data in software-defined radio experimental apparatuses on the other end.

As such, this raises questions about FAIR – the qualitative and quantitative results of a streaming system can be tied not only to the raw data that is input but also the specific ordering and time of insertion. Further, streaming workflows are frequently defined by the fact that the data is not stored as part of the processing, making it unclear what the “dataset” is at all. This adds a layer of requirements on creating a FAIR streaming dataset that may be difficult to achieve. This opens the question of whether it is possible to have a FAIR workflow that works over non-FAIR data, and the consensus across the two days seemed to be that it was both possible and useful to have such FAIR workflows, but the definition of the FAIR workflow metadata would need to be adapted for such streams.

As a result, the discussions addressed how to think about reuse and reproducibility for streaming workflows, and how closely the FAIR principles can elevate practitioners toward some larger goal of “open science”. The discussion did not achieve a breakthrough here, other than to point to the need for developing a community of practice at ORNL that could help refine what was useful and had sufficient commonality to be usable across the diverse types of streaming workflows.

Among the stream types that the breakout members had worked with were the following unordered list:

- Software radio
- Radar
- Spallation Neutron Source (SNS) experimental data output
- Smart grid situational awareness
- DOE Atmospheric Radiation Measurement (ARM)
- In situ analysis pipelines
- Distributed Active Archive Center (DAAC) meteorology sources
- Human-in-the-loop scalable Machine Learning (ML) algorithms
- A Toroidal LHC Apparatus (ATLAS) instrumental data from European Organization for Nuclear Research (CERN)
- Radioastronomy

The consensus was that, although this is only a small subset of the streaming workflow applications at the lab, there was no real coherence or shared knowledge of approaches, techniques, and platforms within ORNL.

Highlights

FAIR streaming workflows help sharpen the definition between “reusable” and “reproducible”, as there are many event streams in science that are fundamentally unable to be replayed. The goals of FAIR can be met for the workflow context, even if the data stream is never the same twice.

Although there are many examples of streaming or event-based science workflows at ORNL, they are not approachable for research and development activities. The equipment involved is too specialized to allow for casual utilization, the security restrictions make it impossible to experiment with, or the stream was too embedded into production environments to allow third party exploration. Creating a reference environment for shared development would be a great help.

Summary

- Streaming and event-based workflows are an important case to consider for FAIR.
- The lab has a vested interest in being more systematic in developing a community of practice around streaming and event-based workflows.
- A workflow might be FAIR even if the data and software that are used in it are not fully FAIR.

2.6.2 Reusability and Interoperability

The breakout sessions about reusability and interoperability were led by Kshitij Mehta. These discussions focused on the following questions:

- What existing workflow systems/tools have you used before?
- What does “reusability and interoperability” mean for your science/research/practice?
- What is the current state of the practice (at ORNL and other places, as well, if known)?
- What are the requirements?
- What are the gaps?
- What are the next steps? (prototypes, software projects that are envisioned, RFP, et al.)

The remainder of this section summarizes our discussions and highlights key points.

Discussion

Attendees reported limited experience with Workflow Management Systems (WfMS) that are used across DOE laboratories and academia. Some of the systems that attendees reported having used include Cheetah [42], Fireworks [43], Pegasus [44], and RADICAL-Pilot [45]. However, tools used in the Python

community that include Dask [46] and Jupyter [47] were beginning to gain popularity amongst scientists. Ad-hoc solutions using Python and shell scripting were the dominant model for running jobs on HPC systems. A significant reason behind the limited adoption of WfMS as opposed to tools used in the Python community was the limited community support, education, and proliferation of workflow tools in the community.

An important part of the discussions was focused around defining reusability and interoperability of workflows. While the scientific community has a better understanding of reusability and interoperability of *data*, defining the same for workflows is more challenging. It is important to define and distinguish between reusing one's own workflow and being able to reuse workflows created by the science community. The distinction between repeatability, repurposability, and reusability needs to be clarified.

Another part of the discussions was focused around how to gauge the reusability of a workflow. In loosely defined terms, it would be helpful to have a metric that quantifies human involvement in a workflow. A workflow must have sufficient metadata to perform a “handover” to a different user without tedious manual involvement. Capturing rich metadata can enable operations such as versioning and *diff* on workflows. We need to investigate technologies such as containers more deeply for their suitability and shortcomings for HPC.

From an architectural standpoint, a core aspect of reusable and interoperable workflows is designing abstractions for different layers in a software hierarchy to create more modular designs and service-driven architectures that allow for swapping underlying technologies. *Deviating from a static architecture to an ecosystem of tools can help create interoperable workflows.* Rich abstractions that can *transparently* run a workflow on different hardware and systems are important for building reusable and interoperable workflows. Portability is an important requirement for developing reusable workflows. Additionally, as data management forms a core component of science, support for efficient data movement, streaming, and storage is important.

The workflows community also needs to identify common workflow patterns across science applications and use cases to develop such an ecosystem of tools that can provide performant implementations for different patterns. Currently, different tools provide support for subsets of patterns. Metadata and provenance for *workflows* along with data needs to be captured and managed.

Finally, a community of workflow tools developers and users needs to be formed for better proliferation of workflow technologies across science teams at the lab.

Highlights

Support for workflow tools from facilities such as OLCF/NCCS can drive wider adoption of workflow tools. To do so, focused education and training for users is necessary. The workflows community needs to identify common workflow patterns and characterize different tools according to the patterns they support.

A community-driven definition and scope of reusability and interoperability for workflows, along with standardized taxonomy for different workflows is necessary for building FAIR workflows. An interoperable ecosystem of portable tools that support different workflow patterns can prove to be an impactful architectural pattern for FAIR workflows.

Standardized efforts to recognize metadata and taxonomy for workflows, along with the association of vocabulary with domain-specific metadata, is vital for constructing the right metadata for reusability.

Developing infrastructure that can automatically extract such rich metadata, possibly by A.I., can be an important step towards this goal.

Summary

- HPC facilities can help drive the wider adoption of workflow systems through support and education for applications and science teams.
- An ecosystem of tools as opposed to support for individual tools can help create interoperable workflows.
- Infrastructure that supports creating rich metadata and taxonomy for workflows in a portable way is necessary for creating reusable and interoperable workflows.
- The workflows community needs to identify common workflow patterns to bridge the gap between applications and workflow systems.

2.6.3 Workflow Lifecycle

The breakout sessions about workflow lifecycle were led by Olga Kuchar. These discussions focused on the following questions:

- What is a workflow?
- What does “workflow lifecycle” mean for your science/research/practice?
- What is the current state of the practice (at ORNL and other places, as well, if known)?
- What are the requirements?
- What are the gaps / challenges?
- What are the next steps? (prototypes, software projects that are envisioned, RFP, et al.)

The remainder of this section summarizes our discussions and highlights key points.

Discussions

Participants wrestled with the abstractions surrounding workflow definitions. Related concepts of automation, algorithms, task definitions, and pipelines were teased apart to extract consensus for their definitions. By understanding what defines the boundaries of these different concepts, the participants set the ground rules for the building blocks in which to communicate and further refine ideas that would affect workflow lifecycles.

Definitions from the breakout session:

Algorithm: a set or sequence of rules followed for problem-solving operations

Workflow: a sequence of operations performed on inputs to produce a set of outputs

Pipeline: a set of processing elements connected in series

Automation: the process of removing the human from the loop and delegating to machines

As technology progresses, so too does the cognitive overhead. Humans introduce error into systems reducing reproducibility. For complex data management, automation reduces cognitive overhead on information workers enabling them to accomplish more at the creative problem-solving level than with the low-level details of the ever-evolving technology stack. Similarly, humans have a discrete amount of time that make it difficult for them to master new technologies. Workflows offer the possibility of dividing up information work among groups of people to build reusable components that further automation. Each component created that can operate over the greatest collection of problems factors out work that other humans would have had to do, resulting in a reproducible work. As such, information work can be crowdsourced where symbiotic communities build out collective components, workflows, and design patterns to aid the collective with managing big data problems. Each step taken to create a robust facet to the workflow problem space improves our ability to produce quality research and work at an ever-increasing speed while enabling us to grapple with larger and more abstract problems.

One problem with this line of thinking, however, is discovery and management of trust and quality. If we build a collective community around the construction of workflows and their components, how do we find them, version them, index, and build trust? An example of this may be seen in Docker Hub [48]. Docker images and descriptors for the construction of Docker images are all available in a centralized website that may be searched for functionality. These Docker images may be composed into collections of services that could represent workflows. How do we know that the compositions are valid and reproducible and trustworthy? And using this tangible example and other tangible examples, what design patterns can we extract to empower future workflows?

And with any technology, there is a lifecycle. All that begins ends. What does that mean for a workflow lifecycle? How long is a workflow viable? Is it dependent on the quality of its components structure? For example, to use a car as a metaphor, does the car wear out if its fuel injectors go bad? Or is it a simple replacement of the deprecated part? When does a workflow need to be destroyed? For example, suppose we use videocassettes as a metaphor. Do the factories that produced videocassettes still need to exist?

From a design perspective, we typically reason about workflows through a requirements gathering process understanding the preconditions, post-conditions, and some high-level logical method that must take place to convert inputs into outputs. These representations are often thought of as directed acyclic graphs (DAGs) or flowcharts. Despite the planning efforts upfront, this view is often limited to a per-problem perspective, treating the construction of every workflow as a one-off solution. Instead of viewing the set of logical transforms to solve a single tangible problem, what are the collection of requirements needed for FAIR workflows? This means that there may be general cross-cutting design patterns across all families of workflow problems that can be codified and rendered into a collection of best practices.

Highlights

The participants of the workshop identified the following requirements for FAIR workflows:

- *Workflow repository*: workflows must be stored in a centralized location where they can be versioned and assigned DOIs to reference tagged releases
- *Reproducibility*: workflows and corresponding components must track metadata for managing external data and hardware configurations to assert reproducibility
- *Metadata*: reproducibility metadata and licensing metadata among others must be supported by a workflow repository
- *Security*: workflows and corresponding components must have a means of verifying authenticity to assert trust
- *Monitoring*: workflows must have a means in which to track metrics such as interactions between compute nodes and data generation, usage tracking, and many others.
- *Reuse*: existing workflows must be composed of interchangeable parts to enable pre-existing works to be recomposed into derived works.

Summary

- Creation of a workflow is difficult to know ahead of time; it lends itself to an iterative process of planning, development, research, and testing.
- Since workflows have cycles, directed acyclic graphs may not be appropriate representations.
- Long, manual entry of metadata is a tedious task for the scientist and represents a significant barrier to entry.
- Easier dataset referencing would aid construction of workflows.
- Unique naming of workflows since there is a desire to continually edit and improve upon an existing workflow.
- Introducing third-party APIs introduces instability both in terms of reproducibility and compatibility.
- Workflow indexing and longevity need to be understood, and maybe refreshed like a domain name.
- Workflow literacy is needed for both developers and domain scientists.
- Discovery of workflow tools is needed.

2.6.4 Workflow Services

The breakout sessions about workflow services were led by Arjun Shankar. These discussions focused on the following questions:

- What does “workflow services” mean for your science/research/practice?
- What is the current state of the practice (at ORNL and other places, as well, if known)?
- What are the requirements?
- What are the gaps?
- What are the next steps? (prototypes, software projects that are envisioned, RFP, et al.)

The remainder of this section summarizes our discussions and highlights key points.

Discussion

Participants began by defining what workflow services meant for their research and practice. Discussions centered around specific constituent aspects of workflows that applied to data such as meta-data services, cataloging services, natural language processing services, data transfer services, etc. as parts of a workflow.

As particular examples of the state-of-the-practice at ORNL, one of the participants highlighted how workflow services are pulled together from a loose collection of scripts. Another participant described a parameter exploration workflow service which went through steps (which would be services to compose into workflows) of generating data, spectrum analysis, fastidious documentation, abstract analysis; with analysis and observation as distinct phases of the experiment.

In the context of diverse workflows, breakout participants began listing requirements of workflow services as functional capabilities needed by the science. Examples included parsing to extract metadata, drag and drop services, automation of scripting, and identification of jobs and storage structures. The need for increased automation was a clear requirement of workflow services.

A significant gap is the lack of a common shared understanding of what a workflow must accomplish - a characteristic that should be common to most workflows. In addition to such an overarching requirement, repeated smaller functional components (parsing, drag-and-drop, well accepted naming conventions, etc.) will help improve the adoption of workflows for science. The system architecture would ideally make it easy to “compose” the end-to-end service made up of constituent workflow services. Examples of workflow services that will lead to an improved understanding and adoption of services by scientists include:

- automated naming,
- data cataloging,
- parsing of experiment specifications, and
- data movement

A set of higher-order composable services for visualization, orchestration, and Continuous Integration/Continuous Delivery (CI/CD) mechanisms would significantly facilitate the composition of the services into an offering which would simplify the scientist’s journey of discovery.

These higher-level functional services would rely on underlying technology supported by:

- Natural Language Processing
- Publish/subscribe tools
- Registry services and persistent state manager
- Visual composition,
- Drag-and-drop data management
- Top-down descriptive language or human-expressed glue logic.
- Interactive API for transfers between services.

Highlights

The challenges in providing a common fabric of workflows that allows intuitive composition and use by end-users will need both an incremental progression towards improved capabilities, as well as a “network effect” which will lead to a multiplicative proliferation of use. It could be initiated by DOE laboratories to establish common guidelines for principal investigators (PIs) to upload their data, use common protocols, allow open exchange of data and structures. A hypothesis is that workflows and programming languages are going to converge. Just as we have seen a proliferation of programming languages before they converge on a limited set, we currently have a proliferation of workflow tools. This can be intimidating, especially to newcomers. We expect our emphasis going forward to be on the following efforts:

- Interoperable and standard efforts for workflow services.
- Focus on both the interactive user experience to simplify the scientist’s discovery journey as well as the automation and composition of the services.

Summary

- Workflow services for science vary across science domains but continue to lack a shared understanding for providers and users.
- Scientists usually need specific functional components. System designers must provide these and mechanisms to compose them.
- Progress needs both incremental offerings, and top-down directives (to facilitate progress).

2.6.5 Metrics for FAIR

The breakout sessions about metrics for FAIR were led by Katie Knight. These discussions focused on the following questions:

- What does “metrics for FAIR” mean for your science/research/practice?
- What is the current state of the practice (at ORNL and other places, as well, if known)?
- What are the requirements?

- What are the gaps?
- What are the next steps? (prototypes, software projects that are envisioned, RFP, et al.)

The remainder of this section summarizes our discussions and highlights key points.

Discussion

FAIR metrics are not well understood by the workflows community, at least as they were represented in the breakout sessions for this two day workshop. Those in attendance thought of metrics as a way to “grade” FAIRness, as opposed to measure the degree to which something is FAIR; that is, they saw it as a “good/bad” set of measurements instead of a report on the current state of affairs regarding FAIR for a given digital object.

One significant outcome of these breakouts is the concept of the FAIR Fingerprint, where instead of using the word “metric”, which has connotations of being rated or ranked, the Fingerprint illustrates the degree of sameness or difference between different objects regarding how many of the FAIR principles are either applicable or currently implemented. This may overlap with the FAIR community’s proposal of a FAIR Implementation Profile (FIP) [49], which provides a means of describing the current state of FAIRness for any given digital object(s) by capturing the comprehensive set of implementation choices made at the discretion of individual communities of practice, resulting in a collection of community-specific FIPs to compose an online resource called the FIP Convergence Matrix [50]. This matrix tracks the evolving landscape of FAIR implementations to inform reuse and interoperability.

Another participant intuited the point made by Goble [4] regarding the difficulty of tracking provenance in data produced by a workflow. They remarked that both knowing who or what produced the data is important, as is the need to provide versioning for the metadata itself, as it may change. Additional questions raised were how effective automation/population of the “rich metadata” can be achieved (a principle under F2 [3]), as well as any subsequent quality assurance for metadata accuracy, all of which will affect how compliant a workflow (or data related to a workflow) is, if measured by a FAIR metric. For rich metadata, the group discussed that a balance should be struck between the amount of required “bare minimum” and “extra” metadata when describing a workflow to ensure that the community avoids overwhelming users with potentially irrelevant metadata requirements. Automated metadata harvesting was suggested, such as the collection of structural metadata (e.g., “What did the workflow actually do?”).

Yet another challenge raised was the “cost” of FAIRness, in that compliance (with anything) takes time away from something else; say, research. FAIR Data Stewardship may be the answer, where different domains employ Data Stewards to assist with compliance for both FAIR data and workflows.

FAIR metrics have proved useful for self-assessment. For an assessment done by ESDIS, a \$170 million NASA science systems project across 14 sites involving approximately 500 people, FAIR was a useful assessment tool for DAAC managers to determine the average opinion that the various DAACs have of themselves, how that matches with the program managers, where the largest improvements might be made, the level of agreement where these improvements ought to be made, and to what degree that is consistent across managers and other elements.

Metrics as they related to FAIR Digital Objects were discussed, where the FAIR Digital Object may describe different workflow components. One option may be to describe different design patterns for workflows as well as snippets that can be connected to make bespoke workflows (as it was acknowledged that all individual workflows are going to be bespoke to a certain extent). These design patterns might be

assembled into a cookbook of workflow “recipes” for facilitating assembling tailored workflows, each with their associated FAIR metrics according to how said “recipe” may be able to comply with the FAIR requirements.

Of tools currently providing users with a means to achieve FAIRness, KBase, an open software and data platform for biology, was discussed. This tool provides metadata that OSTI requests for issuing a DOI, but regarding FAIR’s principle of “rich metadata,” metadata related to users running applications or jobs (where they get a job ID hash) is lost, as a user or data manager would have to dig through each application run to see / find this metadata. Ongoing research is underway to aid KBase users with FAIR adoption: there is a need to show users what is and is not FAIR, as well as what is the benefit of making things FAIR. Most of KBase users are not aware of FAIR, and there was discussion that pushing these users to use FAIR would be beneficial, possibly by something similar to conveying the FAIR fingerprint.

Regarding interoperability, it is unclear as to what this means for workflows. For instance, if a user wishes to work between domains and per-domain templates are implemented, are mappings between domains and/or compatibility checks then also necessary? There was also discussion about “reproducibility”, which seemed to be confused with “reusability”, and indicates that there may need to be some additional education/clarification for the workflows community (or redefinition altogether) around the “R” in FAIR, as well as a clear definition around what FAIR means across different workflow domains. Even still, one participant pointed out that publishers want reproducibility. While data and code might be published together, there is still the question as to how much of a workflow would be useful for reproducibility: would a workflow designed for the Summit supercomputer at OLCF be useful to someone who does not have access to Summit? The counterpoint to this is that a workflow might function as a form of documentation or template. In this manner, that workflow may later be generalized to run on other systems.

Highlights

A FAIR Fingerprint would illustrate the degree of sameness or difference between different objects regarding how many of the FAIR principles are either applicable or currently implemented, and it would avoid using the word “metric”, which has connotations of being rated or ranked. There may also be future research potential related to FAIR Digital Objects or Research Objects to help convey design patterns or workflow components in a more standardized way.

FAIR metrics have proved useful for self-assessment for various tools, and they could be a means for the community to assess data management/workflow milestones and needed areas for funding.

Workflows are particularly tricky in terms of provenance (the R in FAIR) and interoperability. A balance should be struck between the amount of required “bare minimum” and “extra” metadata when describing a workflow to ensure that the community avoids overwhelming users with potentially irrelevant metadata requirements.

Summary

- Scientists need FAIR training: details about what FAIR is/is not, and what the metrics are “measuring”.
- FAIR Fingerprint may be a way forward for “measuring” FAIR without implying that a resource is “unFAIR”.
- FAIR Metrics have helped with self-assessment of various tools (e.g., repository software).
- Existing tools like KBase are already helping users implement FAIR; more research is needed to develop these tools and help users see the value of FAIR and help with communicating in a way that is both simple and actionable.
- FAIR Digital Objects/Research Objects may be a research direction for helping make workflows FAIR.

3. RESEARCH CHALLENGES

The attendees and speakers highlighted several research challenges facing FAIR, workflows, and the workflow research community. Some of these challenges are deceptively simple, such as using the same shared vocabulary, but in that example, vocabulary by itself proved to be a major stumbling block in the breakout sessions.

This section details some of the most important research challenges that were identified in the workshop discussions, beginning with the need for a common understanding and vocabulary for workflows. Next, a set of problems for necessary and sufficient metadata are described along with a potential solution, FAIR Digital Objects. After that, this section covers two approaches to execution of FAIR workflows, which are driven by abstract workflow patterns and by user-declared policies, respectively. The next challenge outlined is the need for “automatable metadata”, where computers create metadata alongside data, for consumption by both humans and computers. Finally, this section presents challenges facing workflow repositories’ construction and curation.

3.1 COMMON UNDERSTANDING

The term “workflow” is overloaded generally, and it has several domain- and context-specific meanings in the scientific communities. Unfortunately, these definitions tend to be limiting and mutually contradictory. Thus, there is a significant challenge in building a better scientific workflows community of developers and users just from a perspective of agreeing to common terms, interfaces, and layers.

At its simplest level, a workflow is exactly what the compound word means – a flow of work. The flow part implies that there is some continuous process, either of introduction of new things to be worked upon, or of multiple types of work actions that need to be orchestrated, or (frequently) both. The term in the scientific context generally implies some degree of automation or delegated control, and all the examples we explored in the workshop ranged between fully automated and partially automated, with some “human-in-the-loop” components.

Beyond this basic level of description, though, it becomes difficult to find agreement. Some scientific workflows focus on the bulk movement of data, whether that be from an instrument out to a hosted environment where it can be processed and disseminated, or between major compute facilities to support multi-site campaigns, scientific archive management, or shared repositories. Some are process-focused, where each data file created by an experiment or simulation run gets processed in the same way to clean, analyze, visualize, etc. in a consistent way. Yet others focus on continuous transfers of data between coordinated processes, also known as stream-based workflows, such as the federated connections between experiments and digital twins or other live computational control modules that run concurrently.

Even in this limited set of examples pulled from the many discussed at the workshop highlight the difficulties that achieving common semantics for workflow layers may have. For instance, previous efforts in the scientific workflows community have tried to leverage shared terminology like “Work Item” [51]. However, it is not clear what that term could or should denote in the context of a streaming process where each data frame or event may only be interpretable in the context of the whole stream – what then is the work item? Is it each frame? Is it the stream in its entirety? Is it some window over the stream?

The great opportunity in FAIR approaches applied to workflows correspondingly imposes some constraints on how we construct a shared way of understanding how to interoperate and reuse one team’s definition of a workflow within the context of another. There is also great opportunity in defining workflows broadly while trying to resolve these issues, as the workshop highlighted some intriguing corners of FAIR in science. For example, FAIR when applied to a scientific streaming workflow could

have a few different implications. Some streams are ephemeral by nature, as some observations just can never be made again. Can the stream of input data be labeled as FAIR in that case? Could the software and workflow associated with the stream processing be considered FAIR if the data that it works upon is not FAIR? Or does the definition of a FAIR stream not really refer to a byte-by-byte reconstruction of its previous state?

The answers to these questions represent opportunities for building not only better terminology for describing the processes of creating scientific workflows, but also for forging new abstractions and community semantics around how an individual investigator’s contributions fit into an understanding of the scientific workflow lifecycle. Current practice forces many scientists into idiosyncratic ways of constructing workflows based on snippets of Bash scripts, Python files, and bits of framework borrowed from across the computing spectrum. Many practitioners who do daily workflow construction and execution, therefore, frequently do not even think of themselves as such. They are simply focused on the flow of their work. The promise of FAIR workflows means that we need to embrace this diversity and all the semantic complexity that it implies.

3.2 RICH METADATA FOR WORKFLOWS

Thirteen of the 15 sub-principles of FAIR focus specifically on metadata, which represents a challenge by itself because metadata is time-consuming and expensive. One of the FAIR sub-principles, F2, specifies that “data are described with rich metadata,” which is intentionally vague; the definition of “richness” varies from one domain to the next, or even from one dataset to the next. In this vein, there are numerous challenges specific to FAIR in the computer science domain – especially challenges related to data sharing, data availability, and data reuse [52].

For data sharing, FAIR explicitly states that data should be shared using a standardized protocol but does not specify how this can or should be achieved. This rolls into issues related to accessibility and interoperability for workflows, as a lack of consistent sharing protocols will complicate metadata definitions/consistency and may interfere with workflows that span multiple systems and datasets.

Data availability raises questions about clear and easily shareable access controls. Metadata should clearly reflect what data is available and how that data is available, and ideally be readable by a machine, especially if this data is meant to be used by an automated workflow or some other means of analysis that does not necessarily include vetting by a human.

Data reuse also poses challenges, as mechanisms to ensure that data is both appropriate for the task at hand (assuming that the data is ingested by a machine or algorithm and may need to be “vetted” automatically) as well as some means of tracking provenance as data moves through a workflow or some other process: how has the data been transformed? Has appropriate attribution been provided for the data creator?

For each of these challenges, all roll into an overarching problem of necessary and sufficient metadata, which will vary from one domain to the next in both scope (metadata vocabularies are not consistent across domains, or even always present) and origin (metadata may come from an instrument and be proprietary or be supplied by a human). This pertains to F2, “data are described with rich metadata,” where “richness” will be relative to a domain in both vocabulary and origin and dependent on whether the metadata has been vetted in some way for accuracy, consistency, and completeness.

These challenges are not trivial, but there may be a way forward: FAIR Digital Objects (FDO). An “FDO is a stable actionable unit that bundles sufficient information to allow the reliable interpretation and processing of the data contained in it” [53].

While the creation of FDOs will likely require humans (data stewards, perhaps), these objects are designed to assist in providing machine-readable context to anything operating on that object (workflow, algorithm, etc.). It is also possible that a workflow could be designed to “mint” these objects (or aggregate related FDOs together as a new “object”), assisting with ongoing data stewardship efforts as the data landscape changes and inevitably grows.

3.3 FAIR WORKFLOW DEVELOPMENT

This workshop raised some important research questions that remain open for additional investigation by the broader research community, and they will require investigation and testing by different science domains. These investigations and tested implementations will be significant in helping to draw boundaries around FAIR in practice: where it works, where it fails, and where it perhaps is not applicable for a given domain.

For example, can workflows be assessed for various FAIR attributes, independent of the data the workflows use? More specifically, can a workflow be FAIR if the data involved are not FAIR? For instance, if a workflow operates on data that has little or no findability, accessibility, interoperability, or reusability, is the workflow itself still FAIR? Some investigation into how FAIRness for workflows can (or should) be decoupled from the data that the workflow operates on is necessary.

FAIR is a set of principles that, together, are meant to increase the value of digital objects by ensuring that they are both reusable and reused. The lack of technical requirements is intentional: domains will necessarily have different needs, technologies, and, therefore, different implementations. This also means that FAIR is a means to begin important conversations around what we talk about when we talk about FAIR—not just for data, but for the workflows that produce this data, and the scientists who (re)use these workflows. The idea for a FAIR “fingerprint” that originated in this workshop, as opposed to meaningless metrics that cannot apply across fields, is a novel idea that deserves exploration.

Streaming data is also a puzzle for FAIRness, which makes applying the FAIR principles to streaming workflows even more challenging. It is unclear, for example, how to assign DOIs to data streams, for similar reasons that make it unclear whether to apply the FAIR principles to the radio station as a URI or to the raw radio signals; it would seem easier to record raw radio signals into files and apply the principles to the recordings, but would those still be considered streaming data? These problems are additionally compounded with reproducibility considerations for the scientific method itself. Furthermore, infrastructure for streaming systems may have an extra layer to capture for FAIR.

3.4 PATTERNS AND POLICIES

As mentioned previously in Section 3.1, the scientific workflows community faces several challenges in agreeing on terminology and the semantics of describing workflow needs and implementations. Even in today’s fractured workflow environment, however, there are substantial challenges in constructing approaches and user interfaces for composing scientific workflows. Individualized, ad hoc combinations of tools (e.g., utilizing Pegasus for the macro workflow along with numerous Bash scripts and Python scripts for internal workflow components) mean that there are significant difficulties in composing workflows at the outset, as well as being able to audit and revisit those workflows in the future to understand what occurred. We cannot expect a single magical standard to replace all these existing workflow systems, so there is a need to understand the shared platforms, services, and functions from which one could construct a FAIR representation of the workflow.

The developers of Common Workflow Language (CWL) [54] maintain a list [55] of more than 300 different workflow tools and systems used across industry and national laboratories for computational and

data analysis workflows. While several tools provide ad hoc solutions to different science problems, a significant number of these systems provide overlapping capabilities for general workflows. A detailed study of the similarities and differences between tools would be impactful for science users, albeit highly tedious. While there has been work done in characterizing workflow systems [56], the absence of community-wide adoption of ways to characterize workflow systems makes this a challenging task. As a result, it is difficult for those in the workflow developer community to recommend workflow tools or systems which would be best suited for a particular science problem.

The fundamental nature of software development runs risks with respect to long-term sustainability, which is development and maintenance of a software product in the long term. Understandably, science teams can thus be hesitant to adopt a workflow system because they want to avoid being restricted to it. On the contrary, it is more amenable to adopt patterns than actual technologies or implementations. As a simple example, users of database technologies can benefit from being able to switch between different implementations or database vendors, as opposed to being constrained to a specific implementation. It would be highly beneficial for science workflows to be able to use workflow abstractions instead of workflow systems. An important step in developing the FAIR workflows methodology, especially interoperability, is to fill this gap between science applications and workflow systems through an abstraction layer or standard that provides the core workflow functionalities.

The workflows community has identified several common workflow patterns such as pipelined execution, bag-of-tasks, high-performance distributed execution, and ensemble runs, to name a few. Emerging patterns driven by the field of ML applications and edge computing include dynamic task and resource management. One approach towards providing a standardized set of workflow functionalities is to develop an ecosystem of tools that provide different aspects of an end-to-end workflow. Such an ecosystem or framework focused approach would enable “plugging in” existing workflow tools, and thereby avoid creation of yet another workflow system. Hypothetically, can such an ecosystem also provide a recommendation system that outputs a set of workflow patterns applicable to a science use depending on its desired features?

A different approach towards FAIR workflows is using policy-driven workflow execution as a main mechanism to express workflows. Can application use cases express end goals and policies to describe their workflows? This would include dynamic control of running applications to tune for changes in the state of an application or a system, for performance optimization, and for automated data management to adhere to predefined constraints around completeness and performance. Users can express high-level policies, which can then be actuated transparently by workflow systems and library implementations. Policies can become complex, but in general, they have the advantage of focusing on the end goal instead of the underlying technology.

This sort of policy-driven semantic can be much more compelling for end users, and it lends itself to declarative semantics for constructing workflows that simplify the end user’s cognitive burden of constructing high performance scientific workflows; however, in order to generate FAIR repositories of the data and workflow, it must be addressed not only how that policy was stated, but also what the particular policy evaluation was at the time that it was used. Maintaining an auditable/discoverable record of the runtime impact of the policy controls, and indeed determining what sort of record is required later for reusability or reproducibility requirements, are substantial research needs that come out of the workflow challenges identified at the workshop.

3.5 AUTOMATABLE METADATA

In the beginning, data were created for computers by humans, and then for many years, most data have been created by computers for computers. Most metadata have been created for humans but not

necessarily for computers, however, which causes substantial difficulties when constructing computational autonomous workflows. Without proper metadata, a computer must be programmed to “discover” important details about the data such as format, content, and history, which not only adds complexity to workflows but also renders them prone to failure for “bad data” with unknown formats, for example, or malformed contents. Therefore, there is a need for data science to progress to a point where computers create metadata alongside data, for consumption by both humans and computers.

FAIR is not a recipe or “magic”: in fact, “the lack of uniformity in data models from one repository to another, and in the richness and availability of metadata descriptions, makes integration and analysis of these data a manual, time-consuming task with no scalability” [57]. FAIRness, as it exists now, is a set of behaviors that depend on the end-user to do a lot of extra work, not conceptually but in practice. The community standards have not moved to a point where there are consistent incentives to use FAIR in daily life. Yet, this same paper provides an example of what a Web-based implementation that is fully FAIR-compliant would look like. While not necessarily a recommendation, this interoperability infrastructure does provide an example of FAIR in practice. Other implementations such as this are necessary to provide more examples—and workflows—to the broader science community.

Thus, we contend that the common practice of FAIR does not scale; FAIR data sets need something more. The basic behaviors with which FAIR started are not enough to achieve the FAIR goals, which are themselves a bit slippery to define.

3.6 WORKFLOW REPOSITORIES

Given that repositories for code and data have existed for years, the idea of repositories for workflows seems like a natural extension, but some of these efforts, such as WorkflowHub [58] are still so new that no publications exist to detail them. Workflows “age” in the same way that components of workflows like data and code age. Thus, workflows face lifecycle problems that are similar to those facing data and code. There are numerous reasons why this problem of publishing workflows for others to consume is difficult, as well as the related problem of consuming published workflows.

Some of these problems are related to the richness of the metadata surrounding the workflows, as detailed in Section 3.2, but when publishing a workflow, scientists encounter basic problems like unique naming, too. Inevitably, a repository will contain many similar workflows, and they will likely be similarly named. Subsequent versions of the same workflow should be stored, too, and thus identifiers – or at least identifying metadata of some kind – will be crucial so that workflow components can be referenced consistently from other workflows. Some ways that this process has been managed in the data and code communities are by using namespaces, hashes, and semantic versioning.

Findability becomes a major problem within workflow repositories, too, so that users can find the right tool for their job. Tools for indexing and searching the workflows will need to consider methods for suggesting appropriate or relevant workflows for users, and this opens a lot of questions for incorporating ML and rich metadata for workflows. How can a user’s cognitive overhead be reduced, given a library of well-curated workflow components and some unknown input data source? It seems necessary to be able to construct a mapping of some kind. Could this be in the realm of machine learning to use curated metadata from workflow components to map to a subset of input data in order to select the “best” workflow component? Could an approach like machine learning adapt and generalize to an ever-changing landscape of workflow components that spans domains from chemistry to biology and even finance, all toward the goal of reducing a researcher’s overhead for getting work done quickly and reproducibly? How could metrics for workflows be defined to assess quality, so that suggestions for the “best” (most relevant) workflow also take into account how “good” (highest quality) the workflow is?

Amidst all this workflow indexing, there is also a problem of lifespan. DOIs are maintained for approximately 20 years. What is the appropriate duration of time for workflow indexing? Since the concept of a workflow may cover transient one-off activities or consistent activities such as financial reporting, there is a wide range of existence durations. Would it make more sense to have a workflow index identifier that must be renewed at particular intervals like a lease?

There are also questions regarding third parties. A workflow author may want to share the minimal amount of metadata about their workflow, keeping most of the details of their work secret. The institution they work for may want to monitor and report against the workflow, however, and this might represent a competing interest. Another question would be how to handle workflows that depend on third-party APIs, which introduces uncertainty due to the dependence. Instability can result, both in terms of reproducibility and compatibility. The Linux operating system manages dependency trees for installing different applications. Could an approach like this be used for managing third-party APIs? Would third-party APIs then need to provide some sort of trusted versioning and release system themselves if they want to be integrated?

Thus, there are many subtle details surrounding open questions that will need to be solved for workflow repositories so they are developed in the way that code and data repositories have developed.

4. ORNL CHALLENGES

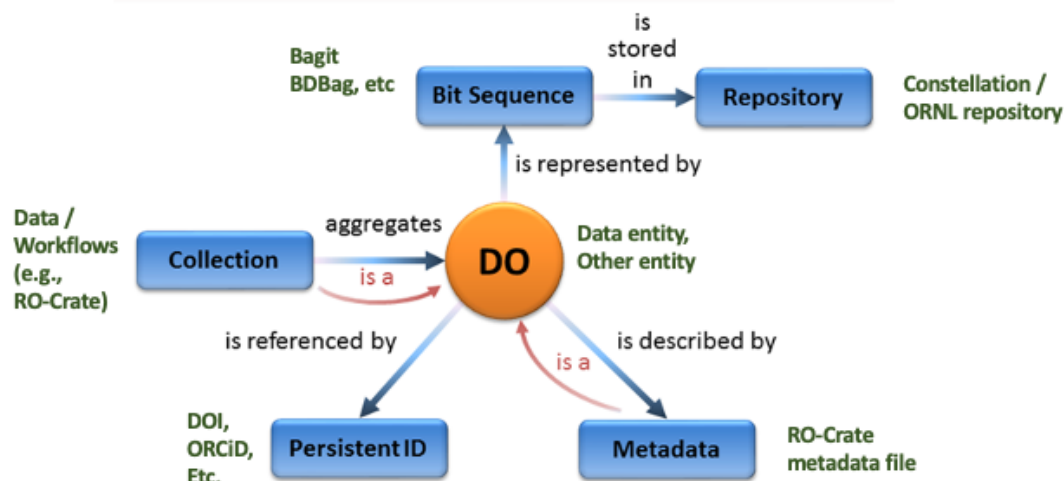
In addition to the many research challenges and opportunities for FAIR workflows, ORNL scientists also face challenges that are unique to their own lab. These challenges range from very simple problems such as confusion over what the “R” in “FAIR” stands for, to very complicated problems such as implementing incentives that encourage/force adoption of the FAIR principles for data and workflows.

This section is devoted to challenges identified during the workshop which are specific to ORNL, including FAIR education, infrastructure, and culture.

4.1 FAIR EDUCATION

General knowledge of the FAIR principles at ORNL varies widely between individual scientists and disciplines. Many standard practices at ORNL naturally coincide with the FAIR, and some scientists are both aware of the FAIR principles and consciously observing them, actively striving to improve the FAIRness of their projects. On the other hand, there are many scientists at ORNL who are not especially aware of the FAIR principles, even though their work unconsciously observes some of the principles anyway. We believe that increased awareness of the FAIR principles would lead to better science at ORNL, and thus we believe that this represents an opportunity for workforce development through a sustained program of training and engagement. As shown in Figure 2, ORNL clearly has the capabilities in place to work with FAIR Digital Objects (see Section 3.2), but many attendees did not know this.

FAIR Digital Objects at ORNL



Adapted from Schwardmann, U. (2020).

Figure 2. Concrete examples of FAIR Digital Objects at ORNL (credit: K. Knight lightning talk).

As another example, there was a great deal of confusion among attendees about the letter “R” in FAIR, especially regarding other related concepts which share the same first letter. Many attendees repeatedly confused “reproducibility” with “reusability”, probably because reproducibility was the focus of our opening keynote by Daniel Garijo (Section 2.3.1). Reproducibility is a crucially important concept in the scientific method and therefore in automating scientific practice with workflows, but its meaning is

stronger than what is implied in the FAIR principles. Several breakout sessions also reported discussions on other related “R” words, including refactorability, repeatability, replicability, and repurposability. There is ample opportunity here for ORNL to create a report for explaining the distinctions and relationships between these important concepts from a workflow perspective.

Practical, hands-on FAIR training is needed to address some of these problems. It is clear from the breakouts that attendees did not know what FAIR really means (namely, that the principles are guidelines, not a set of rules), but the sense from the community is that there is a need to show what FAIR does, not what it is supposed to do. ORNL has many connections to the FAIR community that will aid in this endeavor; notably, ORNL’s Katie Knight is the chair of GO TRAIN for GO FAIR US (see Section 1.2).

This is also an opportune time for leadership at ORNL to consider how to encourage and support the lab’s researchers to implement the FAIR principles in their day-to-day scientific practice. Without significant short-term benefits to offset the costs, researchers at ORNL may decide that implementing FAIR adds too much complexity to warrant adoption, despite the long-term benefits that FAIR provides. ORNL needs to provide encouragement, education, and support for the FAIR principles to be adopted widely.

4.2 INFRASTRUCTURE

There are infrastructure challenges for both FAIR and workflow activities at ORNL. For example, ensuring interoperability of workflows with other systems is very difficult when the systems are very different from the original. ORNL boasts leadership-class computing resources in the form of Summit and the upcoming Frontier system, but workflows designed for these systems can be very hard to generalize for other computing resources because ORNL’s resources are truly world-leading. This interoperability problem also emphasizes the need for a cohesive and unified ORNL workflows community that can consume and repurpose its own workflows to achieve the maximum return on producing them.

Infrastructure is needed for the ORNL community to store workflow lifecycle data alongside metadata like provenance data, as well. Resources such as test systems are often required for software development as well as experiments, validation, and testing of solutions; attendees raised concerns that there are not enough such resources at ORNL. Attendees also stressed that infrastructures for working with very unusual systems, such as streaming systems, may require advanced solutions with extra layers to capture for FAIR.

Many attendees also suggested that a centralized place to store workflows, as well as related metadata describing the workflow, would increase the quality of the science performed at ORNL. Such repositories accumulate the knowledge and artifacts of the entire community to avoid the waste of redundant effort. This would be especially helpful at ORNL, where infrastructure is often world-leading but uncommon enough that off-the-shelf workflows require modification; for this reason, workflows known to work here at ORNL are especially useful to other ORNL researchers. Unfortunately, ORNL’s workflows are not currently collected in any single place, much less a location that can be searched. A workflow repository for ORNL would accelerate research across the lab, and its construction would also solve numerous research problems facing the global community, as outlined in Section 3.6.

A major goal for those in the workflows community is to avoid bespoke solutions that “reinvent the wheel” by implementing new tools when widely adopted tools were already available. During this workshop, attendees identified several challenges related to the lifespan of a workflow. Every workflow is subject to “bit rot”, but what is the right lifetime for a workflow? How long can a workflow reasonably be expected to execute correctly without maintenance? Attendees raised the point that using third-party APIs in workflows forces long-term dependence on third parties, as well as the difficulty in onboarding new

team members for existing workflows. There are numerous opportunities for ORNL to define standards for workflow lifespans and collaboration.

Finally, there are numerous opportunities to create tools and infrastructure which would make observance of the FAIR principles easier for researchers at ORNL. FAIR is intimately connected to metadata, but long, manual entry of metadata is a tedious task for the scientist and represents a significant barrier to entry. Datasets need to be easier to reference, too, as more and more data migrate to the cloud. Infrastructure that supports creating rich metadata and taxonomy for workflows in a portable way is necessary for creating reusable and interoperable workflows, but it would also make life much easier and more convenient for scientific researchers at ORNL, which would accelerate adoption of FAIR.

4.3 CULTURE

As with any community, there are strengths and weaknesses in ORNL's culture. A key observation that arose from the discussions was that "Science is better when it adopts FAIR", and therefore, ORNL can continue to improve its science culture through a commitment to FAIR in its workflows, software, and data sets. By building on our existing partnerships to organizations like GO FAIR US [7] and the Research Data Alliance [33] that have been championing FAIR for data assets, ORNL has the opportunity to carve out an identity in applying and becoming a living laboratory for FAIR scientific workflows.

There are cultural differences within ORNL that lead to division within the workflows community, as well. Some attendees observed that younger scientists seem to prefer working with Jupyter notebooks, which use a Mathematica-style interface [59] to encapsulate workflows which most commonly use Python code, over the more traditional approach that commonly uses command-line interfaces and Bash scripts to glue together polyglot applications. Other attendees noted that there is a symbiosis between domain and workflow scientists which ORNL needs to take better advantage. Domain scientists should not be writing their own workflow management systems, if only because it takes time and resources away from performing their domain science; similarly, difficulty in determining the correct granularity of a workflow without domain-specific knowledge means that workflow scientists should not be writing workflows without domain scientists. In short, domain and workflow scientists need to collaborate more closely at ORNL to push science forward for the entire laboratory.

This common theme of working together and pooling resources is perhaps the biggest problem facing the ORNL workflows community. Some cynical attendees noted that ORNL's workflows community only exists in the abstract sense; aside from this CAW workshop, there are no laboratory-wide efforts to convene workflow producers, workflow consumers, and other interested parties. The lack of a unified, active workflows community at ORNL, or even a coherent view of workflow activities at the lab, significantly retards the progress of scientific discovery. For one thing, the lack of a unified community for workflows immediately leads to tool proliferation and reinventing the wheel; this is an extremely common problem in workflows, and failure to converge on common tools leads not only to waste in the form of redundant efforts, but also it leads to lesser quality tools that are less well-tested and debugged. Another problem can be missed opportunities for collaboration. As a concrete example, researchers working on problems related to data streams would benefit from a coherent view of workflow activities at the lab because they might be able to collaborate with others who have already solved problems with streaming workflows; in lieu of a coherent view of streaming activities themselves, which ORNL also lacks, collaborators would still find each other through the workflows community.

FAIR is something that must be intentional; it does not "just happen". Attendees noted in different breakout sessions that ORNL and the greater scientific establishment seldom reward research into FAIR activities, and similarly, they noted that implementing FAIR requires extra time and effort that are seldom compensated through funding. As a result, even the most conscientious FAIR advocates must make hard

decisions that weigh costs and benefits; those final decisions do not always increase FAIRness. In short, things need to be organized in a way that lets FAIR be natural, or else it will be very difficult to increase adoption.

FAIR is also different for different communities. The FAIR principles are a set of guidelines, but they are not prescriptive because they cannot be – data problems differ widely. ORNL science is mostly conducted with moderate sensitive data, but some ORNL researchers work with health data or nuclear reactor data which require much stricter policies to be observed. As a result, implementing FAIR principles is much different for researchers who use highly sensitive data than it is for researchers who are able, by virtue of their lesser data constraints, to perform fully Open Science. This presents an interesting challenge for leadership at ORNL when considering laboratory-wide FAIR policies, not least because it is extremely difficult to measure adherence across communities using a common set of metrics.

The organizers of CAW 2021 envisioned the formation, after the workshop, of a working group specifically focused on FAIR workflows. This working group would meet periodically for seminars and hackathons in which a group member volunteered their workflow to be “FAIRified”. This practice would provide experience and new ideas to all group members, and obviously the volunteered workflow would improve, too.

There are numerous advantages to the establishment of a community of practice. For example, workflow literacy is needed for both software developers and domain scientists. Software developers who are new to workflows may be overwhelmed to find that there are more than 300 known workflow management systems available [55], but a healthy community will easily help them navigate to the right tool. Domain scientists would benefit from learning how to organize their workflows in ways that are suitable for automation. The lab also has a vested interest in developing a community of practice around streaming and event-based workflows.

FAIR can also be woven into existing institutional commitments through ORNL’s Laboratory Directed Research and Development (LDRD) initiatives such as Interconnected Science Ecosystem (INTERSECT), the Artificial Intelligence (AI) initiative, Carbon Lifecycle, and others. Leveraging existing organizational opportunities like the Data Asset Council or the Visual Informatics for Science and Technology Advances (VISTA) community, the lab’s leadership can incentivize a focus on reusability, interoperability, accessibility, and findability of the techniques, datasets, and approaches.

Further, leveraging key partners, ORNL should take the opportunity to organize national and international meetings to establish our prominence in this intersection point. Whether through workshop additions to existing high-profile conferences and events (e.g., Supercomputing, Smoky Mountain Conference) or as stand-alone invitational events like a Dagstuhl seminar, ORNL is positioned to be a leading global influence in this key area of open science. Finding institutional resources and building upon this opportunity to establish a new funding profile for FAIR workflows will allow us as a community to be better at sharing as one ORNL.

5. OLCF AS THE HUB FOR FAIR WORKFLOWS

The Oak Ridge Leadership Computing Facility (OLCF) is a national computing resource for the DOE, and it is located on the main ORNL campus. OLCF develops and applies computational science capabilities that address some of the world’s most pressing concerns, and it has a history of hosting record-breaking supercomputers. OLCF is not specifically an ORNL resource, but many members of the ORNL community are users of OLCF. Because of these intersecting interests, many of the challenges summarized in Sections 3 and 4 are relevant to OLCF.

This section is devoted to challenges identified during the workshop for OLCF, including how OLCF can lead DOE efforts with respect to FAIR workflows, and what role OLCF can play for the ORNL workflows community.

5.1 LEADING DOE EFFORTS

OLCF is, first and foremost, a designated DOE user facility for leadership computing. OLCF partners with academia and industry to design and build world-leading computational resources for science, as well as the tools and algorithms needed to make the most of those resources. Although the focus in the past was on HPC resources for computational science – especially for modeling and simulation – data-intensive science is now considered as a pillar of science, too [25]. Thus, leadership computing must now consider both computational and data-intensive science, and therefore must lead efforts to design and build world-leading resources for it, along with the tools and algorithms to make use of the resources.

Because of its leading role in DOE computing, OLCF needs to lead research and development in FAIR, workflows, and FAIR workflows for the DOE. The FAIR principles provide guidelines for scientific data management and stewardship, and automation in the form of workflows is critical for the ever-increasing volumes of data being analyzed by modern science. This workshop identified several ways for OLCF to take leadership; for example, the lack of a common vocabulary for simple terms like “workflow”, detailed in Section 3.1, complicates communication between OLCF users as well as staff, as we experienced during the workshop. Nearly all attendees were OLCF users, and a lot of time was wasted in the breakout sessions due to debating on what the words meant. If OLCF were to lead a community effort to define these terms even just for its own users, then it would have downstream effects on other events within the DOE space, such as this CAW workshop.

Since OLCF users span all DOE domain sciences, it would be suitable as a coordination hub for efforts such as defining the richness of metadata for different domain sciences, as discussed in Section 3.2. The different domains can provide examples of different kinds of workflows, especially ones that will likely influence their respective domains; some examples of questions these may answer are discussed in Section 3.3. OLCF can also play an instrumental role in identifying workflow patterns, as discussed in Section 3.4. Its vast trove of profiles, application logs, and other information on applications that run on its computing resources can help identify and characterize science workflows. This can help build a portfolio of the complete workflow lifecycle of various applications that includes hardware and software utilization. By working with its users, OLCF can help achieve numerous challenging objectives.

OLCF can also help lead DOE efforts in FAIR workflows by helping to coordinate and run training events and tutorials for educating scientists. It already does this to some extent by educating its own users, who are often leaders in their own fields, but the topics are almost exclusively related to computational science. Even holding training events and tutorials about FAIR workflows only for OLCF’s users will have a downstream effect on DOE science, and there is no reason to restrict the attendance of such events to OLCF users.

Perhaps the most obvious way for OLCF to lead efforts in FAIR workflows for the DOE would be to provide a centralized workflow repository for the DOE, which is a challenge that is discussed in Section 3.6. No such repository currently exists. The hypothetical benefits of a centralized workflow repository are tantalizing, but there are many open questions as to what such a repository would even look like. Efforts such as WorkflowHub [58] are still in their infancy; if OLCF built and hosted such a repository in a manner that garnered adoption throughout the DOE and possibly the entire world, it would instantly elevate OLCF as a hub for FAIR workflows.

5.2 SERVING THE ORNL COMMUNITY

OLCF is located on ORNL's main campus, and OLCF staff members are ORNL employees. Additionally, many ORNL researchers are also OLCF users. Thus, even though OLCF is not specifically an ORNL resource for institutional computing, it is deeply embedded into the ORNL community for all things that are related to computing and data.

Because ORNL is a DOE lab, the research challenges discussed in Section 3 are all relevant to ORNL, too. The ways for OLCF to lead research efforts in FAIR, workflows, and FAIR workflows have been discussed previously in Section 5.1. This section focuses on the role OLCF plays for solving the challenges discussed in Section 4 which are specific to ORNL.

General knowledge of the FAIR principles at ORNL varies widely between individual scientists and domains, even in the self-selecting crowd of CAW 2021 attendees. This topic is discussed in greater detail in Section 4.1, including the reasons that education and training events for FAIR are needed at ORNL. Who would host these events? Ideally, such events would be independent of specific scientific domains because of their broad applicability to all scientific domains at ORNL. One possible answer, then, would be the Research Library & Information Services Group, but we argue here that the OLCF is the most natural answer. OLCF provides resources like Constellation [60] for working with digital objects, as shown in Figure 2, so it makes sense that it should provide training for those resources, too.

Similarly, the infrastructure challenges to FAIR and workflow activities described in Section 4.2 should also be domain-independent, and the natural answer to provide this infrastructure at ORNL is OLCF. The leadership-class computing resources referenced in that section are OLCF resources, and interoperability of workflows is already being studied within OLCF simply to help support its own users – many of whom are ORNL researchers. Some of the existing infrastructure needed to store workflow lifecycle data alongside metadata like provenance data is already available through services like Constellation and DataFed [61]. Test systems for improving software development as well as experiments, validation, and testing of solutions cannot be solely provided by OLCF, because test systems must mimic the real systems, but better test systems for OLCF resources may prove useful.

Another major opportunity for OLCF to meet a challenge facing ORNL that was identified at this workshop overlaps with a challenge facing the entire DOE – designing, building, and deploying a centralized workflow repository. The problem itself has been discussed previously in Section 3.6, and the potential for OLCF to solve this problem for the DOE is discussed in Section 5.1. Thus, this has been identified as a challenge for ORNL and for the DOE, which suggests that it can be solved first by OLCF on behalf of ORNL before being rolled out for the entire DOE.

There are numerous opportunities for OLCF to create tools and services which would make observance of the FAIR principles easier for researchers at ORNL. As mentioned in Section 4.2, the long, manual process of entering metadata is a tedious task for the scientist and represents a significant barrier to entry. This workshop also found that datasets need to be easier to reference, too, as more and more data migrate

to the cloud. Making life much easier and more convenient for ORNL researchers through tools and services is a way for OLCF to accelerate the adoption of FAIR at ORNL.

Finally, the organizers of CAW 2021 envisioned the formation, after the workshop, of a working group specifically focused on FAIR workflows, as mentioned in Section 4.3. Such a working group would meet periodically for seminars and hackathons, and group members would volunteer their workflows to be “FAIRified” by the rest of the group. Establishing and nurturing a working group and/or a community of practice for FAIR workflows at ORNL is an opportunity for OLCF to lead the way and serve the ORNL community.

6. CONCLUSIONS

The CAW workshop was the only laboratory-wide effort in 2021 to convene the workflows community, and it achieved each of its primary goals. It brought together workflow producers, workflow consumers, FAIR advocates, and other parties with shared interests to discuss the FAIR principles, computational and autonomous workflows, and how to improve ORNL research specifically by applying the FAIR principles to workflows. The workshop succeeded in its goal to be inclusive and accessible for all research-minded personnel at ORNL, regardless of the level of experience or specific expertise in FAIR or workflows. The workshop received valuable contributions from graduate students and early career scientists as well as senior members of the lab, and it was well-attended. In fact, six directorates were represented: Biological and Environmental Systems Science (BESSD), Computing and Computational Sciences (CCSD), Energy Science and Technology (ESTD), National Security Sciences (NSSD), Neutron Sciences (NScD), and Physical Sciences (PSD). This attendance stands as a testament to the wide interest and appeal across the lab for FAIR workflows.

An important observation throughout the workshop was that “Science is better when it adopts FAIR”. Among the many other findings are specific areas of opportunity for researchers, the DOE, ORNL, and OLCF to take global leadership roles in data and workflow science. The rest of this section summarizes the highlights and takeaways from this workshop, especially as they relate to the challenge questions for DOE, ORNL, and OLCF that were identified in the workshop’s discussions.

6.1 RECOMMENDATIONS FOR DOE

The large number of abstract concepts in the realm of scientific workflows makes it imperative to establish a common language. What are the community-accepted definitions, terms, and vocabulary for workflows? In response to this question, we recommend developing and publishing a scientific workflow reference that defines, describes, and provides examples to build upon new abstractions for interoperability and reusability. It seems shocking that rigorous definitions for terms like “workflow” and “scalability” do not exist, but this lack of agreement represents an opportunity to develop a common language.

There are more than 300 workflow management systems in existence [55], and they power countless scientific workflows that publish code, data, and literature. What should “publishing a workflow” mean? In response to this question, we have two recommendations. First, we recommend collecting real-world examples where workflows are successfully used and shared. Second, we recommend researching a new model for the lifecycle of workflows that includes a thorough understanding of quality metrics, provenance, versioning, and longevity. Currently, the state of the art for publishing a workflow involves publishing code, data, and accompanying documentation to describe how to reconstruct it from its parts, but there is little or no information about the quality of a published workflow, whether its parts still work correctly, or whether they will be maintained in the future.

Consider again the countless scientific workflows that result in publications. Why is searching for published scientific workflows so difficult? In response to this question, we have two recommendations. First, we recommend funding a working group and publishing a workflow metadata schema to address this Findability problem. Second, we recommend funding domain-specific workflow working groups which will develop an understanding of metadata “richness” for their respective communities’ workflow usages. The definition of “richness” varies from one domain to the next, but within domains, such working groups could determine how best to specialize a generic workflow metadata schema for their communities’ needs.

Suppose now that these countless workflows, or at least those involving DOE science, are findable. Are there common design patterns and policies for DOE science workflows? In response to this question, we have three recommendations. First, we recommend defining workflow characteristics and maintaining a list of workflow tools and their capabilities with respect to these characteristics. Second, we recommend developing a workflow ecosystem of plug-and-play components based on common workflow design patterns. Finally, we recommend researching new approaches in using and evaluating policy-driven workflow executions. Policies can become complex, but because they focus on the end goal instead of the underlying technology, they can be much more compelling for end users. Policies lend themselves well to declarative semantics for constructing workflows that simplify the end user's cognitive burden.

Finally, most metadata have been created for humans but not necessarily for computers, and this subtlety causes substantial difficulties when constructing computational autonomous workflows. Can we improve metadata collection and make it more useful? In response to this question, we recommend pursuing research, evaluation, and metrics for automating the creation of metadata that are both human-readable and machine-actionable. An autonomous workflow can modify itself to react to opportunities or anomalies generated at run-time [1], and this adaptability is greatly aided by the availability of machine-actionable metadata. Thus, improving automated metadata collection will enable the construction of workflows with advanced capabilities well beyond what is currently possible.

6.2 RECOMMENDATIONS FOR ORNL

The evolution of modern science has increased the cognitive burden by adding computational and data-driven concerns to the traditional concerns of theory and experiments [25]; it is very difficult for scientists to master all these things at once. How do we improve literacy on data, workflows, and FAIR? In response to this question, we have three recommendations. First, we recommend creating tutorials and other educational materials to educate ORNL staff. Second, we recommend developing an internal workflows website to help our scientists and engineers find appropriate tools and resources across the lab. Finally, we recommend regular onsite training on existing workflow tools, systems, and use cases. We need to ease the cognitive burden on ORNL's domain scientists, and we see many opportunities for workforce development through education and training in these areas.

There are many producers and consumers of workflows at ORNL, and the fact that 6 directorates were represented by attendees evidences the widespread interest and potential impact of workflows at ORNL. What existing infrastructure is available that supports science workflows at ORNL? In response to this question, we have two recommendations. First, we recommend nurturing and growing our workflows community so that members help share and repurpose our own workflows. Second, we recommend building a workflow repository. The workflows community at ORNL has demonstrated expertise and leadership, but, it is not well organized or cohesive. There is great opportunity to accelerate science at ORNL through better combining forces across disciplines, but supporting infrastructure needs to be improved.

Finally, although a small percentage of ORNL's work is classified or restricted under its growing national security programs, the missions of ORNL are primarily for open science that is published in open literature. How do we support a culture of open science? In response to this question, we recommend investing in building a science culture which is committed to FAIR. The FAIR principles are deliberately vague in some of their definitions because FAIR applies equally well to open, classified, and restricted science; it differs only in how it is applied. As we have observed repeatedly, "science is better when it adopts FAIR", and thus ORNL should support a culture of open science by making a commitment to the FAIR principles as an institution.

6.3 RECOMMENDATIONS FOR OLCF

Although the focus in the past was on HPC resources for computational science – especially for modeling and simulation – data-intensive science is now considered as a pillar of science, too [25]. Now that leadership computing must consider both computational and data-intensive science, OLCF must lead efforts to design and build world-leading resources for computational and data-intensive science, along with the tools and algorithms to make use of the resources. How can OLCF lead DOE efforts in FAIR workflows? In response to this question, we recommend facilitating solving the DOE-specific challenges described in this report by coordinating the diverse OLCF user base. OLCF’s users span all of DOE domain sciences, which makes OLCF suitable as a coordination hub for efforts such as defining the richness of metadata for different domain sciences. Other ways for OLCF to lead DOE efforts for FAIR workflows include hosting training events and tutorials and providing a centralized workflow repository. There is great potential for OLCF to help solve the DOE challenges by coordinating its user base and organizing communal efforts.

OLCF is deeply embedded into the ORNL community for all things that are related to computing and data, even though OLCF is not specifically an ORNL resource. What role can OLCF play for the ORNL workflows community? In response to this question, we recommend that OLCF serve as the hub for FAIR and workflows at ORNL. All OLCF staff and many OLCF users are members of ORNL, and OLCF provides resources like Constellation [60] and DataFed [61] for the same kinds of problems that FAIR workflows are working to address. One major way that OLCF can serve as the hub would be to establish the community of practice in which members take turns “FAIRifying” each other’s workflows. Another major way that OLCF can play a major role would be to provide a workflow repository. This was mentioned in the previous paragraph as something that would also help solve DOE challenges, too, and indeed it would help both; ideally, these would be evolutionary stages of a centralized repository, rather than different repositories. Overall, the lack of organization or cohesiveness within the ORNL workflows community represents a great opportunity for OLCF to assume the role of hub for FAIR and workflows at ORNL.

6.4 NEXT STEPS

Finally, we conclude that the most natural next step for the Computational and Autonomous Workflows (CAW) series of workshops is to continue the series, and for this reason we recommend hosting another installment in 2022 to build upon the work of the first two. Continuing to convene our workflows community will help it to grow and to become more cohesive, and it will also help us continue working toward an ultimate goal of hosting CAW as an international event in the future, supporting DOE, ORNL, and OLCF as clearly recognized global leaders in workflow science and scientific automation.

REFERENCES

- [1] G. Tretola and E. Zimeo, "Autonomic Internet-Scale Workflows," in *Proceedings of the 3rd International Workshop on Monitoring, Adaptation and Beyond*, New York, NY, USA, 2010.
- [2] Y. Babuji, A. Woodard, Z. Li, D. S. Katz, B. Clifford, R. Kumar, L. Lacinski, R. Chard, J. M. Wozniak, I. Foster, M. Wilde and K. Chard, "Parsl: Pervasive Parallel Programming in Python," in *Proceedings of the 28th International Symposium on High-Performance Parallel and Distributed Computing*, Phoenix, AZ, USA, 2019.
- [3] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, J. Bouwman, A. J. Brookes and T. Clark, "The FAIR Guiding Principles for scientific data management and stewardship," *Scientific Data*, p. doi:<https://doi.org/10.1038/sdata.2016.18>, 2016.
- [4] C. Goble, S. Cohen-Boulakia, S. Soiland-Reyes, D. Garijo, Y. Gil, M. R. Crusoe, K. Peters and D. Schober, "FAIR Computational Workflows," *Data Intelligence*, pp. 108-121. doi:10.1162/dint_a_00033, 2020.
- [5] Australian National Data Service, "FAIR Data image map," [Online]. Available: https://www.ands.org.au/_data/assets/image/0011/1416098/FAIR-Data-image-map-graphic-v2-721px.png. [Accessed 10 March 2022].
- [6] "GO FAIR Initiative - GO FAIR," [Online]. Available: <https://www.go-fair.org/go-fair-initiative/>. [Accessed 8 March 2022].
- [7] "Advancing FAIR in the US," [Online]. Available: <https://gofair.us/>. [Accessed 8 March 2022].
- [8] "PuRe Data Resources at a Glance," [Online]. Available: <https://science.osti.gov/Initiatives/PuRe-Data/Resources-at-a-Glance>. [Accessed 8 March 2022].
- [9] "CARE Principles of Indigenous Data Governance — Global Indigenous Data Alliance," [Online]. Available: <https://www.gida-global.org/care>. [Accessed 8 March 2022].
- [10] B. A. Nosek, G. Alter, G. C. Banks, D. Borsboom, S. D. Bowman, S. J. Breckler, S. Buck, C. D. Chambers, G. Chin, G. Christensen, M. Contestabile, A. Dafoe, E. Eich, J. Freese, R. Glennerster, D. Goroff, D. P. Green, B. Hesse and Humphreys, "Promoting an open research culture," *Science*, vol. 348, pp. 1422-1425. <http://doi.org/10.1126/science.aab2374>, 2015.
- [11] V. Stodden, M. McNutt, D. H. Bailey, E. Deelman, Y. Gil, B. Hanson, M. A. Heroux, J. P. A. Ioannidis and M. Taufer, "Enhancing reproducibility for computational methods," *Science*, vol. 354, pp. 1240-1241. doi:10.1126/science.aah6168, 2016.
- [12] R. Ferreira da Silva, K. Chard, H. Casanova, D. Laney, D. Ahn, S. Jha, W. E. Allcock, G. Bauer, D. Duplyakin, B. Enders, T. M. Heer, E. Lançon, S. Sanielevici and K. Sayers, "Workflows Community Summit: Tightening the Integration between Computing Facilities and Scientific Workflows," Zenodo, 2022.
- [13] M. Wolf, J. Logan, K. Mehta, D. Jacobson, M. Cashman, A. M. Walker, G. Eisenhauer, P. Widener and A. Cliff, "Reusability First: Toward FAIR Workflows," in *2021 IEEE International Conference on Cluster Computing (CLUSTER)*, Los Alamitos, CA, USA, 2021.
- [14] "Video Conferencing, Cloud Phone, Webinars, Chat, Virtual Events | Zoom," [Online]. Available: <https://zoom.us/>. [Accessed 8 March 2022].
- [15] "Google Docs: Free Online Document Editor | Google Workspace," [Online]. Available: <https://www.google.com/docs/about/>. [Accessed 8 March 2022].
- [16] "World Wide Web Consortium (W3C)," [Online]. Available: <https://www.w3.org/>. [Accessed 8 March 2022].

- [17] "W3C Provenance Incubator Group Wiki - XG Provenance Wiki," [Online]. Available: https://www.w3.org/2005/Incubator/prov/wiki/W3C_Provenance_Incubator_Group_Wiki. [Accessed 8 March 2022].
- [18] "Provenance Working Group," [Online]. Available: <https://www.w3.org/groups/wg/prov>. [Accessed 8 March 2022].
- [19] "DCMI: Home," [Online]. Available: <https://www.dublincore.org/>. [Accessed 8 March 2022].
- [20] "FAIR for Research Software (FAIR4RS) WG | RDA," [Online]. Available: <https://www.rd-alliance.org/groups/fair-research-software-fair4rs-wg>. [Accessed 8 March 2022].
- [21] S. Soiland-Reyes, P. Sefton, M. Crosas, L. J. Castro, F. Coppens, J. M. Fernández, D. Garijo, B. Grüning, M. La Rosa, S. Leo, E. ÓCarragáin, M. Portier and A. Trisovic, "Packaging research artefacts with RO-Crate," *Data Science*, pp. 1-42. doi:10.3233/DS-210053, 2022.
- [22] "Scientific Paper of the Future," [Online]. Available: <https://scientificpaperofthefuture.org/>. [Accessed 8 March 2022].
- [23] Y. Gil, C. H. David, I. Demir, B. T. Essawy, R. W. Fulweiler, J. L. Goodall, L. Karlstrom, H. Lee, H. J. Mills, J.-H. Oh, S. A. Pierce, A. Pope, M. W. Tzeng, S. R. Villamizar and X. Yu, "Toward the Geoscience Paper of the Future: Best practices for documenting and sharing research from data to software to provenance," *Earth and Space Science*, vol. 3, pp. 388-415. doi:10.1002/2015EA000136, 2016.
- [24] "CAW 2021 Workshop Day 1: Keynote - Daniel Garijo, Universidad Politécnica de Madrid on Vimeo," [Online]. Available: <https://vimeo.com/577331362>. [Accessed 8 March 2022].
- [25] T. Hey, S. Tinsley and K. Tolle, *The Fourth Paradigm: Data-Intensive Scientific Discovery*, Microsoft Research, 2009.
- [26] CrowdFlower, "2016 Data Science Report," 2016.
- [27] E. Wilde and Y. Liu, "Lightweight Linked Data," in *2008 IEEE International Conference on Information Reuse and Integration*, Las Vegas, NV, USA, 2008.
- [28] "Schema.org - Schema.org," [Online]. Available: <https://schema.org/>. [Accessed 8 March 2022].
- [29] U. Schwardmann, "Digital Objects – FAIR Digital Objects: Which Services Are Required?," *Data Science Journal*, vol. 19, no. 1, pp. 15. doi: <http://doi.org/10.5334/dsj-2020-015>, 2020.
- [30] "Horizon 2020 | European Commission," [Online]. Available: https://ec.europa.eu/info/research-and-innovation/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-2020_en. [Accessed 8 March 2022].
- [31] "EOSC Hub," [Online]. Available: <https://eosc-hub.eu/>. [Accessed 8 March 2022].
- [32] "VODAN Home | VODAN Africa & Asia," [Online]. Available: <https://www.vodan-totafrica.info/>. [Accessed 8 March 2022].
- [33] "RDA | Research Data Sharing without barriers," [Online]. Available: <https://www.rd-alliance.org/>. [Accessed 8 March 2022].
- [34] "Home - CODATA, The Committee on Data for Science and Technology," [Online]. Available: <https://codata.org/>. [Accessed 8 March 2022].
- [35] "Home — World Data System: Trusted Data Services for Global Science," [Online]. Available: <https://www.worlddatasystem.org/>. [Accessed 8 March 2022].
- [36] "CAW 2021 Workshop Day 2: Keynote - Christine Kirkpatrick on Vimeo," [Online]. Available: <https://vimeo.com/579511199>. [Accessed 8 March 2022].
- [37] "CAW 2021 Workshop Day 1: Lightning Talks on Vimeo," [Online]. Available: <https://vimeo.com/577331227>. [Accessed 8 March 2022].
- [38] "CAW 2021 Workshop Day 2: Lightning Talks on Vimeo," [Online]. Available: <https://vimeo.com/579515941>. [Accessed 8 March 2022].

- [39] "CAW 2021 Workshop Day 2: Bruce Wilson Presentation on Vimeo," [Online]. Available: <https://vimeo.com/579501438>. [Accessed 8 March 2022].
- [40] "Daymet," [Online]. Available: <https://daymet.ornl.gov/>. [Accessed 8 March 2022].
- [41] "Common Metadata Repository (CMR) | Earthdata," [Online]. Available: <https://earthdata.nasa.gov/eosdis/science-system-description/eosdis-components/cmr>. [Accessed 8 March 2022].
- [42] K. Mehta, B. Allen, M. Wolf, J. Logan, E. Suchyta, J. Choi, K. Takahashi, I. Yakushin, T. Munson, I. Foster and S. Klasky, "A Codesign Framework for Online Data Analysis and Reduction," in *2019 IEEE/ACM Workflows in Support of Large-Scale Science (WORKS)*, 2019.
- [43] A. Jain, S. P. Ong, W. Chen, B. Medasani, X. Qu, M. Kocher, M. Brafman, P. Guido, G.-M. Rignanes, G. Hautier, D. Gunter and K. A. Persson, "FireWorks: a dynamic workflow system designed for high-throughput applications," *Concurrency and Computation: Practice and Experience*, vol. 27, pp. 5037-5059. doi:10.1002/cpe.3505, 2015.
- [44] E. Deelman, G. Singh, M.-H. Su, J. Blythe, Y. Gil, C. Kesselman, G. Mehta, K. Vahi, G. B. Berriman, J. Good, A. Laity, J. C. Jacob and D. S. Katz, "Pegasus: A Framework for Mapping Complex Scientific Workflows onto Distributed Systems," *Scientific Programming*, vol. 13, pp. 219-237. doi:10.1155/2005/128026, 2005.
- [45] A. Merzky, M. Turilli, M. Maldonado, M. Santcroos and S. Jha, "Using Pilot Systems to Execute Many Task Workloads on Supercomputers," in *JSSPP: Workshop on Job Scheduling Strategies for Parallel Processing*, 2015.
- [46] M. Rocklin, "Dask: Parallel Computation with Blocked algorithms and Task Scheduling," in *Proceedings of the 14th Python in Science Conference*, 2015.
- [47] T. Kluyver, B. Ragan-Kelley, F. Pérez, B. Granger, M. Bussonnier, J. Frederic, K. Kelley, J. Hamrick, J. Grout, S. Corlay, P. Ivanov, D. Avila, S. Abdalla and C. Willing, "Jupyter Notebooks – a publishing format for reproducible computational workflows," in *Positioning and Power in Academic Publishing: Players, Agents and Agendas*, IOS Press, 2016, pp. 87-90. doi:10.3233/978-1-61499-649-1-87.
- [48] "Docker Hub Container Image Library | App Containerization," [Online]. Available: <https://hub.docker.com/>. [Accessed 8 March 2022].
- [49] E. Schultes, B. Magagna, K. M. Hettne, R. Pergl, M. Suchánek and T. Kuhn, "Reusable FAIR Implementation Profiles as Accelerators of FAIR Convergence," in *ER 2020: Advances in Conceptual Modeling*, 2020.
- [50] H. P. Sustkova, K. M. Hettne, P. Wittenburg, A. Jacobsen, T. Kuhn, R. Pergl, J. Slifka, P. McQuilton, B. Magagna, S.-A. Sansone, M. Stocker, M. Imming, L. Lannom, M. Musen and Schult, "FAIR Convergence Matrix: Optimizing the Reuse of Existing FAIR-Related Resources," *Data Intelligence*, vol. 2, pp. 158-170. doi:10.1162/dint_a_00038, 2020.
- [51] A. Kumar, W. M. Van Der Aalst and E. M. Verbeek, "Dynamic Work Distribution in Workflow Management Systems: How to Balance Quality and Performance," *Journal of Management Information Systems*, vol. 18, no. 3, pp. 157-193. <https://doi.org/10.1080/07421222.2002.11045693>, 2002.
- [52] K. A. Anjaria, "Computational implementation and formalism of FAIR data stewardship principles," *Data Technologies and Applications*, pp. 193-214. doi:10.1108/DTA-09-2019-0164, 2020.
- [53] K. De Smedt, D. Koureas and P. Wittenburg, "FAIR Digital Objects for Science: From Data Pieces to Actionable Knowledge Units," *Publications*, vol. 8, no. 2, p. <http://doi.org/10.3390/publications8020021>, 2020.

- [54] M. R. Crusoe, S. Abeln, A. Iosup, P. Amstutz, J. Chilton, N. Tijanić, H. Ménager, S. Soiland-Reyes, B. Gavrilovic, C. Goble and The CWL Community, "Methods Included: Standardizing Computational Reuse and Portability with the Common Workflow Language," arXiv, 2021.
- [55] P. Amstutz, M. Mikheev, M. R. Crusoe, N. Tijanić and S. Lampa, "Existing Workflow Systems," [Online]. Available: <https://s.apache.org/existing-workflow-systems>. [Accessed December 2021].
- [56] G. Juve, A. Chervenak, E. Deelman, S. Bharathi, G. Mehta and K. Vahi, "Characterizing and profiling scientific workflows," *Future Generation Computer Systems*, vol. 29, pp. 682-692. doi:10.1016/j.future.2012.08.015, 2013.
- [57] M. D. Wilkinson, R. Verborgh, L. O. Bonino da Silva Santos, T. Clark, M. A. Swertz, F. D. Kelpin, A. J. Gray, E. A. Schultes, E. M. van Mulligen, P. Ciccarese, A. Kuzniar, A. Gavai, M. Thompson and Kal, "Interoperability and FAIRness through a novel combination of Web technologies," *PeerJ Computer Science*, pp. <https://doi.org/10.7717/peerj-cs.110>, 2017.
- [58] "The WorkflowHub," [Online]. Available: <https://workflowhub.eu/>. [Accessed 8 March 2022].
- [59] S. Wolfram, "The Mathematica Book (3rd ed.)," *Assembly Automation*, vol. 19, pp. 77-77. doi:<https://doi.org/10.1108/aa.1999.19.1.77.1>, 1999.
- [60] S. S. Vazhkudai, J. Harney, R. Gunasekaran, D. Stansberry, S.-H. Lim, T. Barron, A. Nash and A. Ramanathan, "Constellation: A science graph network for scalable data and knowledge discovery in extreme-scale scientific collaborations," in *2016 IEEE International Conference on Big Data (Big Data)*, 2016.
- [61] D. Stansberry, S. S. Somnath, J. B. Breet, G. S. Shutt and M. Shankar, "DataFed: Towards Reproducible Research via Federated Data Management," in *2019 International Conference on Computational Science and Computational Intelligence (CSCI)*, Los Alamitos, CA, USA, 2019.

APPENDIX A. LIST OF PARTICIPANTS

APPENDIX A. LIST OF PARTICIPANTS

External (keynote) speakers

Christine Kirkpatrick, San Diego Supercomputer Center, University of California San Diego
Daniel Garijo Verdejo, Universidad Politécnica de Madrid

Registered attendees and speakers from ORNL

Welcome and Opening Remarks: Georgia (Gina) Tourassi, Director of the National Center of Computational Sciences and the Oak Ridge Leadership Computing Facility, Oak Ridge National Laboratory

Asterisks denote speakers for the lightning talks and workflow showcase.

Subil Abraham	Olga Kuchar
Valentine Anantharaj	Rajeev Kumar
Elke Arenholz	Chris Lindsley
Temitope Benson	Yan Liu
Debsindhu Bhowmik	Yongtao Liu*
Swen Boehm	Jeremy Logan
Michael Brim	Ketan Maheshwari*
Mikaela Cashman	Marshall McDonnell
Yongqiang Cheng*	Kshitij Mehta
Jong Choi	John Miller
Ashley Cliff	Kenneth Moreland
Madison Clubb	Debangshu Mukherjee
Mark Coletti	Thomas Naughton
Bob Cottingham	David Pugmire*
Zach Crockett*	Viktor Reshniak
Ranjeet Devarakonda*	Kevin Roccapriore*
Abhijeet Dhakane	David Rogers*
Wael Elwasif	Ashi Savara
Christian Engelmann	Arjun Shankar
Kat Engstrom	Matt Sieger
Charles Finney	Alka Singh
Sudershan Gangrade	Suhas Somnath
Ayana Ghosh*	Sarat Sreepathi
Antigoni Georgiadou	Rohit Srivastava
Muralikrishnan Gopalakrishnan Meena	Michele Thornton
Seth Hitefield	Angelica M. Walker*
Daniel Jacobson	Sean Wilkinson
Gustav Jansen	Jerrold Williams
Wayne Joubert	Bruce Wilson*
Katie Knight*	Matthew Wolf*
James Kress*	

APPENDIX B. WORKSHOP AGENDA

APPENDIX B. WORKSHOP AGENDA

Day 1: Tuesday, July 20, 2021

9:00 a.m. – 9:30 a.m. Welcome: Sean Wilkinson, [Gina Tourassi](#)

9:30 a.m. – 10:15 a.m. Opening Keynote:
[“FAIR Workflows: A step closer to the Scientific Paper of the Future”](#)
Daniel Garijo, Universidad Politécnica de Madrid
(Slides are also available for [download](#).)

10:15 a.m. – 10:25 a.m. Questions/discussion; Transition to Lightning Talks

10:25 a.m. – 11:00 a.m. [Lightning Talks](#):

- “Putting Reusability First in FAIR Workflows” — Matthew Wolf
- “Why Workflow Management Is Not the Problem” — David Rogers
- “Toward Fair Digital Objects for Workflows” — Katie Knight
- “A Critical Take on Workflows and FAIR” — Ketan Maheshwari
- “In-line vs. In-transit In Situ: Which Technique to Use at Scale?” — James Kress
- “Scientific Visualization as a Service” — David Pugmire

11:00 a.m. – 11:10 a.m. Move to Breakout Rooms

11:10 a.m. – 12:10 p.m. Breakout Groups:

- Streaming workflows (Lead: Matthew Wolf, Scribe: Jong Choi)
- Reusability & Interoperability (Lead: Kshitij Mehta, Scribe: Ayana Ghosh)
- Workflow lifecycle (Lead: Olga Kuchar, Scribe: David Rogers)
- Workflow services (Lead: Arjun Shankar, Scribe: Mark Coletti)
- Metrics for FAIR (Lead: Katie Knight, Scribe: Suhas Somnath)

12:10 p.m. – 1:00 p.m. Return to main room; Outbriefs from each Breakout; Adjourn day 1

Day 2: Wednesday, July 21, 2021

9:00 a.m. – 9:10 a.m. Welcome/Overview: [Kshitij Mehta](#)

9:10 a.m. – 9:50 a.m. [Lightning Talks](#):

- “FAIR Workflows in the DOE Systems Biology Knowledgebase” — Zach Crockett
- “DOE User Facilities Implementing FAIR Data Principles: ARM Data Center Example” — Ranjeet Devarakonda
- “Improving Reusability and Accessibility of iRF-LOOP using the Cheetah-Savanna Suite of Workflow Tools” — Angelica M. Walker
- “A workflow for neutron scattering data analysis” — Yongqiang Cheng
- “From Microscopic Images to Simulations via Deep Learning: A Comprehensive Workflow to Develop Physics-Based Understandings” — Ayana Ghosh
- “Experimental discovery of structure-property relationships in ferroelectric materials via active learning” — Yongtao Liu
- “Automated Discovery of Physics in the Electron Microscope” — Kevin Roccapiore

9:50 a.m. – 10:00 a.m. Move to Breakout Rooms

10:00 a.m. – 11:00 a.m. Breakout Groups:

- Streaming workflows (Lead: Matthew Wolf, Scribe: Rajeev Kumar)
- Reusability & Interoperability (Lead: Kshitij Mehta, Scribe: Ketan Maheshwari)
- Workflow lifecycle (Lead: Olga Kuchar, Scribe: Debangshu Mukherjee)
- Workflow services and patterns (Lead: Arjun Shankar, Scribe: James Kress)
- Metrics for FAIR (Lead: Katie Knight, Scribe: Mark Coletti)

11:00 a.m. – 11:20 a.m. Workflow Showcase for FAIR Collaboration:

[“FAIR and Workflows in the Earth and Environmental Sciences”](#)

Bruce Wilson

11:20 a.m. – 11:30 a.m. Return to main room; Set up for Breakout Summaries

11:30 a.m. – 12:00 p.m. [Plenary Discussion and Summary](#)

12:00 p.m. – 12:45 p.m. Closing Keynote:

[“The FAIR+ World According to Me”](#)

Christine Kirkpatrick, University of California San Diego

12:45 p.m. – 1:00 p.m. [Wrap-up and Next Steps](#)

APPENDIX C. LIGHTNING TALK ABSTRACTS

APPENDIX C. ABSTRACTS AND PRESENTATION SLIDES

Day 1: Tuesday, July 20

Welcome: Sean Wilkinson, [Gina Tourassi](#)

Opening Keynote:

[“FAIR Workflows: A step closer to the Scientific Paper of the Future”](#)

Daniel Garijo, Universidad Politécnica de Madrid

(Slides are also available for [download](#).)

Lightning Talks:

Additionally, video of the lightning talks from Day 1 is available at <https://vimeo.com/577331227>.

Video of the lightning talks from Day 2 is available for viewing online consecutively as one video recording which can be found at <https://vimeo.com/579515941>.

Putting Reusability First in FAIR Workflows (Matthew Wolf)

Matthew Wolf delivered the first talk on Day 1, “Putting Reusability First in FAIR Workflows”. The full talk begins at 0:40 in the video recording. The talk’s original abstract follows:

The FAIR (Findable, Accessible, Interoperable, and Reusable) principles have had a significant impact on framing scientific data management policies, from the individual investigator to the national funding agencies. For data, not only does this ordering make a nice acronym, but it also makes sense to make findability primary. However, workflows (and to an extent software) are different in that they are inherently defined by how you use them. As such, we contend that reusability and interoperability are key to understanding a way toward FAIR usage of workflows and that the priority should be in putting reusability first. In this talk, I will lay out some key observations we have made about the connection between reusability in FAIR and technical debt in software engineering, and how that has served as the basis for current and future work in building more reusable workflow services and frameworks.

Why Workflow Management Is Not the Problem (David Rogers)

David Rogers delivered the second talk on Day 1, “Why Workflow Management Is Not the Problem”. The full talk begins at 5:59 in the video recording. The talk’s original abstract follows:

Lack of an appropriate workflow manager or task scheduler is often cited as a problem motivating development of tools for automating computational data generation and analysis. Key research topics include: representing and documenting workflows at a high-level, scheduling by matching tasks to available resources, automating transfer and archival of the source and result data, and curation, annotation, and discovery of data results. This talk provides an alternative perspective from which none of the above research topics are challenging. In this perspective, everything is a dataset. All datasets are static, versioned, and globally addressable. Code that transforms the data is also a first-class dataset, and has a dual representation both as source code and execution result. We see evidence of the emergence of this new paradigm in github, pypi, distributed filesharing, and cloud "bucket" storage.

Toward Fair Digital Objects for Workflows (Katie Knight)

Katie Knight delivered the third talk on Day 1, “Toward Fair Digital Objects for Workflows”. The full talk begins at 11:42 in the video recording. The talk’s original abstract follows:

FAIR Digital Objects (FDOs) are essential when connecting heterogeneous data from different communities: but what about workflows? This lightning talk will provide a brief overview of the proposed requirements for a semantic model for FDOs as proposed by the FDO Forum, a new group that seeks to advance the specification and application of FDOs. Then, the remainder will focus on the proposed FDO for workflows as described by the Canonical Workflow Framework for Research Working Group (<https://fairdo.org/wg/fdo-cwfr/>).

A Critical Take on Workflows and FAIR (Ketan Maheshwari)

Ketan Maheshwari delivered the fourth talk on Day 1, “A Critical Take on Workflows and FAIR”. The full talk begins at 17:10 in the video recording. The talk’s original abstract follows:

FAIR principles, requirements and research topics as generally stated are not entirely suitable or applicable to workflows. While some of them such as Findability and Accessibility are already solved by search engines and public repositories such as Github (so, not a research topic anymore), others such as Interoperability is a pointless exercise with little practical value and great cost. In this lightening talk, I will present some of my thoughts that are rather critical of the topics surrounding FAIR and workflows. In particular, I will argue why interoperability is an absurd endeavor to undertake and why only Reusability is a worthwhile research to pursue. Workflows Interoperability, I believe, is a symptom akin to bloatware and feature creep that is generally seen in software. The root cause is arguably the culture surrounding the research software development practices. To address this, I plan to read out an "open letter" in a humorous vein, addressed to a typical Workflow Management System, time permitting.

In-line vs. In-transit In Situ: Which Technique to Use at Scale? (James Kress)

James Kress delivered the fifth talk on Day 1, “In-line vs. In-transit In Situ: Which Technique to Use at Scale?”. The full talk begins at 22:28 in the video recording. The talk’s original abstract follows:

The DOE is experiencing a massive increase in the data generated by computing, experimental, and observational facilities. In situ processing is commonly used to address this challenge because it enables analysis at a lower cost. However, the best ways to carry out in situ processing in a reusable and workflow friendly approach is far from a solved problem. This work starts address a major challenge to pervasive in situ: the complexity of in situ costs and the reusability of visualization workflows across the DOE ecosystem. This complexity stems from the number and diversity of factors involved: algorithm selection, in situ placement strategy, data (complexity, diversity, and reduction), and heterogeneous architectures (from HPC to edge). Further, this problem is very important. Wrong choices for in situ algorithms or placement can lead to exorbitant costs. On the other hand, overly conservative choices, i.e., not analyzing data frequently enough, can lead to the loss of scientific information. In situ methodologies are varied and have different resource requirements that impact the timeliness and cost for both HPC and streaming/federated environments. These methodologies include running algorithms in a synchronous on-line manner, an asynchronous manner (same node, separate nodes, or federated resources), or a hybrid of the two. The cost for in situ algorithms on each methodology can vary dramatically. This presentation will discuss recent work that has begun the work of characterizing costs to create a cost model for different classes of in situ algorithms, heterogeneous architectures, and levels of resource concurrency. We will show how this model allowed for increased performance and decreased cost over traditional synchronous methods. Furthermore, we will discuss the timeliness of visualization workflows, and show how

asynchronous visualization allows for greater control over the timeliness of visualization and how this relates to dedicated resource requirements.

Scientific Visualization as a Service (David Pugmire)

David Pugmire delivered the sixth talk on Day 1, “Scientific Visualization as a Service”. The full talk begins at 28:04 in the video recording. The talk’s original abstract follows:

One of the primary challenges facing scientists is extracting understanding from the large amounts of data produced by simulations, experiments, and observational facilities. The use of data across the entire lifetime ranging from real-time to post-hoc analysis is complex and varied, typically requiring a collaborative effort across multiple teams of scientists. Over time, three sets of tools have emerged: one set for analysis, another for visualization, and a final set for orchestrating the tasks. This trifurcated tool set often results in the manual assembly of analysis and visualization workflows, which are one-off solutions that are often fragile and difficult to generalize. To address these challenges, we propose a serviced-based paradigm and a set of abstractions to guide its design. These abstractions allow for the creation of services that can access and interpret data, and enable interoperability for intelligent scheduling of workflow systems. This work results from a codesign process over analysis, visualization, and workflow tools to provide the flexibility required for production use. This talk will describe a forward-looking research and development plan that centers on the concept of visualization and analysis technology as reusable services, and also describes several real-world use cases that implement these concepts.

FAIR Workflows in the DOE Systems Biology Knowledgebase (Zach Crockett)

Zach Crockett delivered the first talk on Day 2, “FAIR Workflows in the DOE Systems Biology Knowledgebase”. The full talk begins at 1:38 in the video recording. The talk’s original abstract follows:

The Department of Energy Systems Biology Knowledgebase (KBase) is an open source, online bioinformatics platform that allows users to construct custom analysis workflows that can be copied, shared, and reused. KBase contains over 200 Apps for various analyses and the platform integrates many types of data from multiple public resources. Users can perform analyses based on their individual needs or leverage shared workflows to replicate analyses with their own data. The data model integrates different data types into an internal representation that allows interoperability in a unified environment. Workflows are built as Narratives in KBase, which are based on reproducible Jupyter Notebooks. The flexibility of Narratives allows users to modify, extend, or combine workflows for greater scientific discovery, while also supporting findable, accessible, interoperable, and reusable (FAIR) data management best practices. Narratives also support markdown cells that enable users to add rich documentation, and code cells that run user-created scripts alongside the KBase Apps. Details on data source, App runs, and downstream data products in a Narrative are captured to ensure provenance, critical for reproducible science. KBase also allows users to create a static Narrative, which is a publicly-available snapshot of a Narrative. These snapshots capture the current state of the Narrative for sharing and publication, while also allowing further analysis to be done inside the original Narrative. KBase can register DOIs for these static Narratives (upon request) with the Office of Science and Technical Information (OSTI). Static Narratives are indexed by search engines so that they can be found through searches or through data portals such as Google’s dataset search. These data are accessible to viewers who can interact with and reuse the Narrative within the KBase environment. As these features offer built-in advantages to facilitate reproducible, shareable workflows that can be reused and modified, publishers are interested in partnering with KBase to create standardized workflows to accompany publication. In particular, the American Society of Microbiology’s Microbiology Resource Announcement (MRA) journal sees a large number of KBase users organically publishing genome

announcements using work done in KBase Narratives. MRA editors are working with KBase to create template Narratives that can be copied by users and run with their own data.

DOE User Facilities Implementing FAIR Data Principles: ARM Data Center Example (Ranjeet Devarakonda)

Ranjeet Devarakonda delivered the second talk on Day 2, “DOE User Facilities Implementing FAIR Data Principles: ARM Data Center Example”. The full talk begins at 6:46 in the video recording. The talk’s original abstract follows:

Atmospheric Radiation Measurement (ARM) is a multi-laboratory/multi-institutional, US Department of Energy Office of Science National User Facility. ARM's data is currently hosted at the ARM Data Center (ADC) in Oak Ridge, Tennessee. The ADC holds more than 12,000 data products, with a total holding of more than 3 PB of data that dates back to 1992. This includes data from instruments, value-added products, model outputs, field campaigns, and principal investigator contributed data. In this paper, we discuss how big federal scientific data centers, such as ARM, that use modern and scalable architecture apply findable, accessible, interoperable, and reusable (FAIR) data principles to improve overall efficiency. These principles mainly emphasize machine-to-machine interactions that are directly applicable to ARM because of their data volume.

Improving Reusability and Accessibility of iRF-LOOP using the Cheetah-Savanna Suite of Workflow Tools (Angelica M. Walker)

Angelica M. Walker delivered the third talk on Day 2, “Improving Reusability and Accessibility of iRF-LOOP using the Cheetah-Savanna Suite of Workflow Tools”. The full talk begins at 13:18 in the video recording. The talk’s original abstract follows:

FAIR workflows can be applied to an explainable-AI (X-AI) method such as iterative Random Forest Leave One Out Prediction (iRF-LOOP) to efficiently generate high throughput predictive models. iRF-LOOP is used on a variety of data sets to create all-to-all association networks. For every dependent variable a single iRF model is created, resulting in n runs for a matrix of size n features and m samples. As the size of the feature sets increase, so does the number of individual models. Due to differing requirements on each High Performance Computing (HPC) system, the process of producing submission scripts is manual to account for resource allocation and runtime parameters. Estimating run times for the individual models is difficult unless one is familiar with the algorithm, limiting the reusability of iRF-LOOP. Additionally, iRF-LOOP runs are submitted manually in sets corresponding to the number of nodes, i.e., for 5 nodes only 5 individual iRF models can be run at once, and the following set is submitted once all 5 models are completed. If one model takes significantly longer than the remaining models, then node hours are wasted waiting for the remaining node to be freed. Following a full run of iRF-LOOP, a list of failed models is generated if necessary and the resubmission process is repeated, involving another step of human intervention. These issues driven by human intervention and resulting idle nodes can be eliminated using the Cheetah and Savanna suite of tools. Prior to integration with Cheetah-Savanna, in one 2-hour allocation of 20 nodes an average of 53.11 individual iRF models were completed, compared to 280 individual iRF models using Cheetah-Savanna integration. This over 5-fold improvement in total runtime reduces the number of wasted node hours, while also applying the principles of FAIR workflows by reducing the amount of human intervention in the submission, error handling, and resubmission process. This reusable workflow will allow the efficient, reproducible use of iRF-LOOP by many different people and groups on a variety of HPC architectures.

A workflow for neutron scattering data analysis (Yongqiang Cheng)

Yongqiang Cheng delivered the fourth talk on Day 2, “A workflow for neutron scattering data analysis”. The full talk begins at 18:13 in the video recording. The talk’s original abstract follows:

Two of the world’s most powerful neutron sources hosted at ORNL, the High Flux Isotope Reactor (HFIR) and the Spallation Neutron Source (SNS), are DOE’s premier user facilities for materials research. Neutron scattering technique has unique advantages in “seeing” where atoms are and what they do, thus providing fundamental insight for understanding the material’s behavior and leading to rational design of novel materials. Atomistic modeling is a highly relevant and complementary approach in analyzing and interpreting the experimental data, and it often holds the key for the efficiency and productivity of many beamlines. In this presentation, I will discuss the methods and workflows that are being used and developed at SNS to integrate atomistic modeling and neutron scattering to achieve automated and streamlined neutron data analysis. Such a “digital twin” for neutron scattering experiments is expected to play an even greater role when combined with advanced data analytics such as data mining and machine learning.

From Microscopic Images to Simulations via Deep Learning: A Comprehensive Workflow to Develop Physics-Based Understandings (Ayana Ghosh)

Ayana Ghosh delivered the fifth talk on Day 2, “From Microscopic Images to Simulations via Deep Learning: A Comprehensive Workflow to Develop Physics-Based Understandings”. The full talk begins at 24:32 in the video recording. The talk’s original abstract follows:

Over the last decades, electron and scanning probe microscopies along with computational simulations performed at several time and length scales have evolved as the primary schemes to study systems in the domain of physical and life sciences on the atomic and mesoscale levels. High flux of data is generated in both fields that includes a range of structural, spectral insights and information on materials properties. In recent years, deep learning has paved ways in these avenues to accelerate understandings of physical phenomena. However, studies exploiting the nuances of all three to establish a bridge between such learnings is still in its infancy. In this work, we show how deep learning can bridge together the knowledge learned from microscopic images and first-principles simulations to develop a comprehensive understanding of the physics of the materials of interest. Here we focus on how deep convolutional neural networks can be employed to identify atomic features (type and position) in graphene. We utilize an ensemble learning iterative training (ELIT) framework to consider out-of-distribution effects levitated from variations in microscopic measurements. The initial set of models in this framework uses simulations-based data to reconstruct partial features once applied to microscopic images, followed by refining the training set and iteratively training ensemble of models, yielding robust feature finding and pixel-wise uncertainty quantifications. The detected features (atoms or defects) are then used to construct simulation objects to perform first-principles simulations to find optimized geometry of the structures followed by studying temperature-dependent dynamics of system evolutions with ad-atoms and defects. The results along with associated uncertainties in predictions at various levels as obtained employing this framework may be used to evaluate and modify experimental conditions and regions of interest. We aim to further expand this approach to incorporate edge-computing involving direct transfer of image-based data from microscopes via Jetson AGX Xavier and then analyze, train deep neural networks using a GPU-based platform followed by performing simulations using CPU-based high-performance computing resources and feed back to the human in the loop, altogether on-the-fly, to better guide experiments while learning from theoretical models.

Experimental discovery of structure-property relationships in ferroelectric materials via active learning (Yongtao Liu)

Yongtao Liu delivered the sixth talk on Day 2, “Experimental discovery of structure-property relationships in ferroelectric materials via active learning”. The full talk begins at 30:17 in the video recording. The talk’s original abstract follows:

Domain wall dynamics in ferroelectric materials underpins multiple applications ranging from actuators to information technology devices. Many of the properties of domain walls have been understood by various imaging methods including scanning probe microscopy and scanning electron microscopy. Despite the advances in microscopy techniques over the past decades, the imaging process of various microscopy techniques is still mostly based on rectangular grid scanning. This is time-consuming when performing a spectrum characterization of materials (e.g., polarization switching hysteresis in Piezoresponse Force Microscopy (PFM)). In the meantime, it is well understood that the information of interest is tied to the spatial microstructure, such as topography or domain wall structure. Here, we report the development and implementation of a workflow based on the Deep Kernel Learning (DKL) for PFM that allows problem-specific tuning of workflow and operation in real-time, allowing to actively learn the relationship between polarization switching and domain structure in ferroelectric materials during the experiment. The structure of the DKL provides insight into the physics of the process. To solve a universal problem in microscopy measurements—image drift, we embedded a drift correction algorithm into the workflow to eliminate the drift during in-situ measurement. This drift correction method is based on a recently developed shift-invariant variational autoencoder (shift-VAE). The developed approach can also be adapted to other imaging and spectroscopy methods, such as electron microscopy, optical microscopy, and chemical imaging.

Automated Discovery of Physics in the Electron Microscope (Kevin Roccapiore)

Kevin Roccapiore delivered the seventh talk on Day 2, “Automated Discovery of Physics in the Electron Microscope”. The full talk begins at 35:52 in the video recording. The talk’s original abstract follows:

Fundamental new opportunities for exploring the physics of nanoscale and atomic systems have recently begun to surface with the emergence of advanced electron and scanning probe microscopy techniques. Particularly, in the scanning transmission electron microscope (STEM), the latest technological breakthroughs with electron monochromation, aberration correction, and fast pixelated detectors have enabled an unprecedented degree of exploration into physical phenomena such as charge density waves, vibrational and plasmonic excitations, and superconducting systems. With the STEM, the change in energy of transmitted electrons is detected with electron energy loss spectroscopy (EELS) which provides insight into the local dielectric function and plasmonic behavior and is key to understanding the physics of collective excitations in nanoscale systems. Many classes of quantum devices rely on the understanding of plasmonic and phononic properties on the nanometer level. To probe these systems using STEM-EELS – like with most other physical imaging modalities – the experimentalist decides the regions for spectral measurements, whether a single point spectroscopy or a grid of points. Experimental discovery of new physical relationships can therefore often be impeded by operator bias. This has prompted much interest in automated and autonomous experimentation to accelerate and promote the rate of physical discovery. Bayesian optimization and Gaussian processes have been at the forefront of automation in X-Ray scattering, synthesis, and recently even the STEM. Practically however, these approaches offer limited flexibility due to a lack of prior information. Here we combine the power of Bayesian optimization with deep convolutional neural networks, forming deep kernel learning (DKL-BO), which enables automated physical discovery in the STEM. We demonstrate that almost an unlimited degree of physics discovery is possible using this method, where a physics-based criteria for exploration is specified. This method allows for extended periods of automation by incorporating automatic drift

compensation. An associated workflow is developed, its performance on pre-acquired data of various nanoscale systems is quantified, and finally we present the deployment on an operational microscope in which edge plasmon functionality in MnPS₃, a metal thiophosphate, is discovered and further explored. Other pathways using different measurement techniques in the STEM are additionally considered.

Workflow Showcase for FAIR Collaboration

[“FAIR and Workflows in the Earth and Environmental Sciences”](#)

Bruce Wilson

Closing Keynote

[“The FAIR+ World According to Me”](#)

Christine Kirkpatrick, University of California San Diego

[Wrap-up and Next Steps](#)



APPENDIX D. SLIDE IMAGES

APPENDIX D. SLIDE IMAGES

Putting Reusability First in FAIR Workflows
Matthew Wolf

Based on "Reusability First: Toward FAIR Workflows", M. Wolf, J. Logan, K. Mehta, D. Jacobson, M. Cashman, A. Walker, G. Eisenhauer, P. Widener, and A. Cliff. To appear in Cluster 2021

ORNL is managed by UT-Battelle LLC for the US Department of Energy






Reusable First



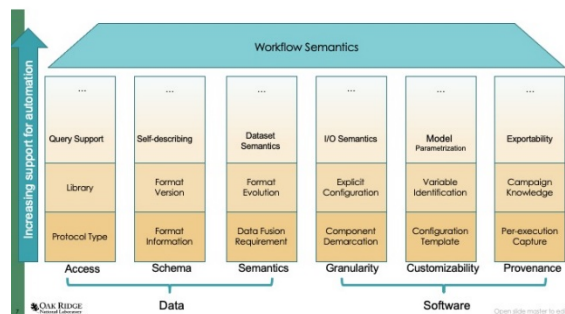
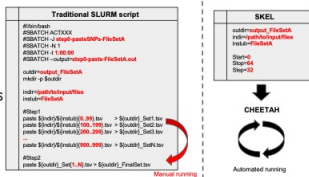
- FAIR (Findable, Accessible, ...) is typically the ordering for data sets
- Workflows aren't data, and they aren't software
 - Workflows are defined by how data and software are connected**
- Reusing someone else's data is about using their information in your existing computational connections
 - Correspondingly, your first concerns tend to be about finding that data
- Reusing someone else's workflow demands more** and can't be an afterthought

Core Idea: Reusability is the Inverse of Technical Debt

- Workflow reuse doesn't require tools
 - Get a student/intern/postdoc
- Anything that automation doesn't support becomes a human cost of reuse
 - Human commitment -> servicing the technical debt**
- The FAIR metadata for reusable workflows needs to be focused on **automation**
 - Switch from auditing  to automating 

Improving Reusability

- Focus metadata on system-actionable quantities
- Leverage existing tools to extract metadata with minimal human intervention
- Our Demonstration
 - Model-driven code generation (Skel)
 - Self-describing data formats (ADIOS)
 - Campaign-oriented workflow management (Cheetah/Savanna)
- But how to assess progress toward reusability?



FAIR Workflows and Data at ORNL

- We're leveraging experience working with a number of ORNL science applications
 - Particularly computational biology – see Ashley's talk tomorrow!
 - But if you're interested, let us know!
 - Biology, materials, fusion, climate, ...
- Our implementation builds on experiences with ORNL and ECP software
 - Skel, Cheetah/Savanna, and ADIOS are ORNL products

Toward Fair Digital Objects for Workflows

Katie Knight

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

ENERGY

National Center for Computational Sciences
Scalable Protected Data Facilities (SPDF)

Abstract

- FAIR Digital Objects (FDOs) are essential when connecting heterogeneous data from different communities: but what about workflows?
- This lightning talk will provide a brief overview of the proposed requirements for a semantic model for FDOs as proposed by the FDO Forum, a new group that seeks to advance the specification and application of FDOs.
- Then, the remainder will focus on the proposed FDO for workflows as described by the Canonical Workflow Framework for Research Working Group (<https://fairdo.org/wg/fdo-cwfr/>)

OAK RIDGE National Laboratory National Center for Computational Sciences Computing and Computational Sciences Directorate

What's This All About?

FAIR Digital Objects - Canonical Framework for Workflow Research (FDO-CFWR)

• <https://fairdo.org/>

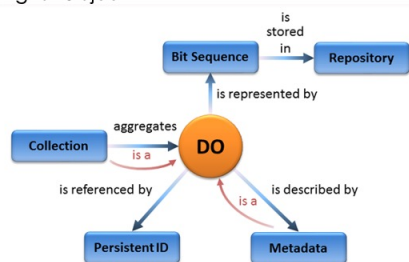
Need to introduce libraries of canonical components which are close to researcher practices and allow researchers to easily orchestrate automatic workflows.

Such workflow frameworks that need to be attractive to researchers to adopt (the researcher should not be bothered by all FAIR requirements).

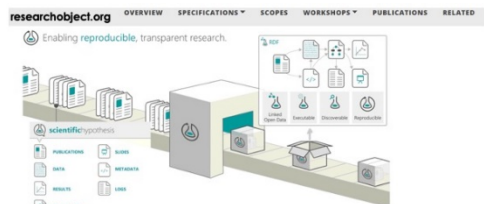
The best way to achieve this is to request that all canonical components should support the concept of FAIR Digital Objects (FDO).

OAK RIDGE National Laboratory National Center for Computational Sciences Computing and Computational Sciences Directorate

FAIR Digital Object



OAK RIDGE National Laboratory National Center for Computational Sciences Computing and Computational Sciences Directorate



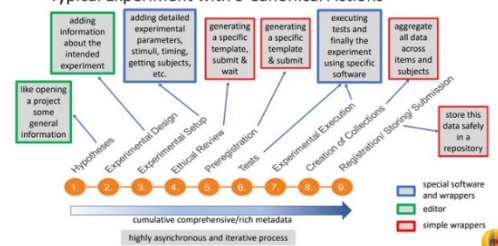
Research Object Crate

RO-Crate has been developed as a schema.org-based JSON lightweight approach to the next generation Research Object serialization.

OAK RIDGE National Laboratory National Center for Computational Sciences Computing and Computational Sciences Directorate

Slide from Peter Wittenburg, Max Planck Computing & Data Facility

Typical Experiment with 9 Canonical Actions



OAK RIDGE National Laboratory National Center for Computational Sciences Computing and Computational Sciences Directorate

Examples

Using Jupyter for reproducible scientific workflows

Marjan Beg, Juliette Taka, Thomas Khuyter, Alexander Kononov, Min Ragan-Kelley, Nicolas M. Théry, Hans Fangohr

Literate computing has emerged as an important tool for computational studies and open science, with growing follow-up of best practices. In this work, we report two case studies - one in computational magnetism and another in computational mathematics - where domain-specific software was exposed to the Jupyter environment. This enables high-level control of simulations and computation, interactive exploration of computational results, batch processing on HPC resources, and reproducible workflow documentation in Jupyter notebooks. In the first study, Uleming drives existing computational micromagnetics software through a domain-specific language embedded in Python. In the second study, a dedicated Jupyter kernel interfaces with the GAP system for computational discrete algebra and its dedicated programming language. In light of these case studies, we discuss the benefits of this approach, including progress toward more reproducible and reusable research results and outputs, notably through the use of infrastructure such as JupyterHub and Binder.

Comments: 11 pages, 3 figures
Subjects: Mathematical Software (cs.MS); Numerical Analysis (math.NA); Computational Physics (physics.comp-ph)
Journal-reference: Computing in Science & Engineering 23, 36-46 (2021)
DOI: 10.1109/MCSE.2021.3052161
arXiv:2102.09562 [cs.MS] (or arXiv:2102.09562v1 [cs.MS] for this version)

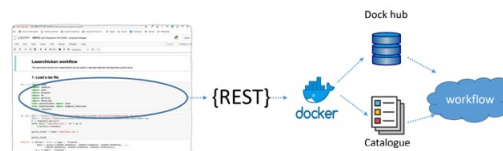
<https://arxiv.org/ftp/arxiv/papers/2102/2102.09562.pdf>

OAK RIDGE National Laboratory National Center for Computational Sciences Computing and Computational Sciences Directorate

FAIR-Cells: turn python code into RESTful services and Docker container

<https://pypi.org/project/FAIR-Cells/>

- Extract notebook cells as RESTful services
- Enable service based workflow in cloud

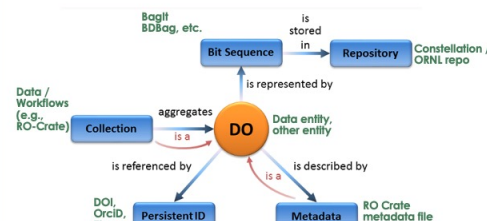


21 Krüger W., et al., FAIR-Cells: an interactive tool for enabling the FAIRness of code fragments in Jupyter notebooks, in the proceedings of International Conference on High-Performance Computing and Simulation (HPCCS) 2020, 34-46.

OAK RIDGE National Laboratory National Center for Computational Sciences Computing and Computational Sciences Directorate

FAIR Workflows and Data at ORNL

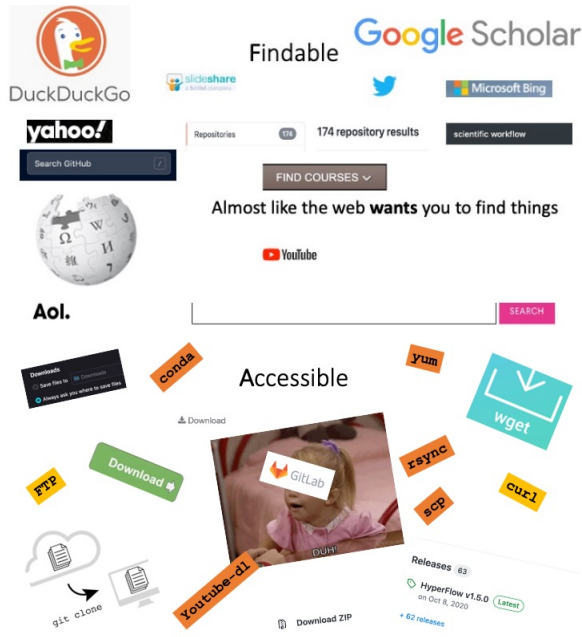
- Self-described objects (data, instruments, workflows)



OAK RIDGE National Laboratory National Center for Computational Sciences Computing and Computational Sciences Directorate

Workflows and FAIR: A Critical Take

Ketan M
Oak Ridge National Laboratory

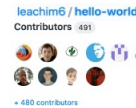


Interoperable



Reusable

- Follows F and A
- Worthwhile and achievable
- A collection of community curated workflows / templates in a Github repo.



EbookFoundation / free-programming-books



Summary

- FAIR not as applicable to workflows as it is to Data.
- For workflows, perhaps a better parallel is Software.
- Simple, Lean, Lightweight and Non-invasive Workflow Management Systems desired.

**In-line vs. In-transit In Situ:
Which Technique to Use at Scale?**

James Kress
University of Oregon &
Oak Ridge National Laboratory

Matthew Lattin
Lawrence Livermore
National Laboratory

Hank Childs
University of Oregon

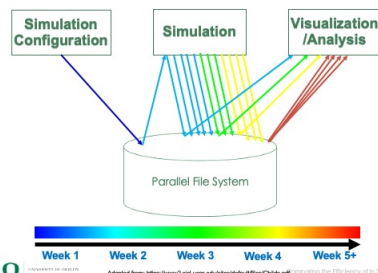
Jong Choi, Mark Kim, Matthew Wolf, Norbert Padonaski, Scott Klasky, David Pugmire
Oak Ridge National Laboratory

July 20, 2021

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

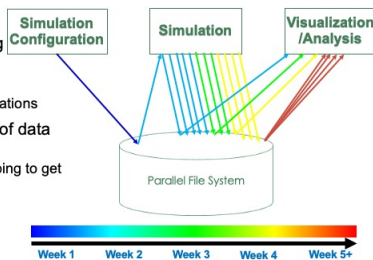
kressjm@ornl.gov

Post-hoc Analysis



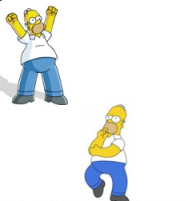
Problems with the Post-hoc Paradigm

- Potentially missing discovery
 - Missing data
 - Infrequent visualizations
- Time and amount of data written
 - IO bottleneck is going to get worse



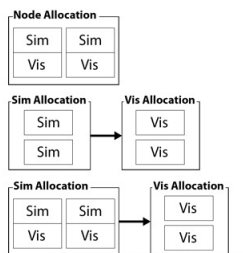
In Situ to the Rescue

- What is in Situ?
 - Produce visualization & analysis during an active simulation
 - Multiple ways in situ can be accomplished
- Application developer response to in situ:
 - YES!
 - Less data to disk and we have lots of compute power
 - High data fidelity
 - I'm not sure...
 - COST**
 - Reusability**
 - Some a priori knowledge needed
 - Limited time to view data
 - Resilience**



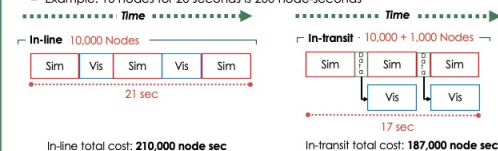
Common In Situ Configurations

- In-line (Co-processing, tightly coupled)
 - Simulation and visualization run in the same process using the same resources (on-node)
- In-transit (Concurrent-processing, loosely coupled)
 - Simulation transfers data to a separate set of visualization nodes (off-node)
- Hybrid coupling
 - Utilizing aspects of both in-line and in-transit



What does total time and cost mean?

- Time measured in seconds
- Cost measured in node-seconds: (# of nodes) X (# of seconds)
 - Example: 10 nodes for 20 seconds is 200 node-seconds



Scalability

- Translating simulation results to visualization
 - Critical Point:** If scalability is variable for each code/situation, how then can in situ be reusable?



Cost Models for Reusable In Situ

Visualization Cost Efficiency Factor (VCEF)

Measure of how much more efficiently an algorithm runs on a smaller resource

$$\frac{\text{In-line Vis Time} \times \# \text{ Nodes}}{\text{In-transit Vis Time} \times \# \text{ Nodes}} \quad \text{e.g.} \quad \frac{1 \times 1,000}{50 \times 10} = \frac{1,000}{500} = 2$$

In-line costs 2X that of in-transit

- Total cost
 - In-line: More cost effective for computation heavy algorithms at low concurrency
 - In-transit: More cost effective with communication heavy workloads, and computation heavy algorithms at large scale
- Total Time
 - Communication-bound algorithm: in-transit was faster in all test cases
 - Computation-bound algorithm: in-line was faster in most cases with a short simulation cycle time

FAIR Workflows and Data at ORNL

- Reusability
 - This work focuses on cost models for in situ visualization that will help enable workflows to schedule visualization cost effectively, saving time and resources, while also giving a measure of timeliness for when visualization results will be ready
- Interoperability
 - This work makes use VTK-m, ADIOS, and FIDES. These technologies allow for data from different simulations/sources to easily be utilized, while also allowing the visualization to be portable and performant on a wide array of compute systems.

Visualization as a Service for Scientific Data

David Pugmire¹

James Kress¹, Jieyang Chen¹, Hank Childs², Jong Choi¹, Dmitry Ganyushin¹, Berk Geveci², Scott Kasky¹, Xin Liang¹, Jeremy Logan¹, Nicole Marsaglia², Kshitij Menal¹, Norbert Podhorszki¹, Caitlin Ross², Chuck Atkins², Eric Suchytal¹, Nick Thompson¹, Steven Walton², Lipeng Wan¹, Matthew Wolf¹

¹ Oak Ridge National Laboratory ² Kivware, Inc. ³ University of Oregon

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

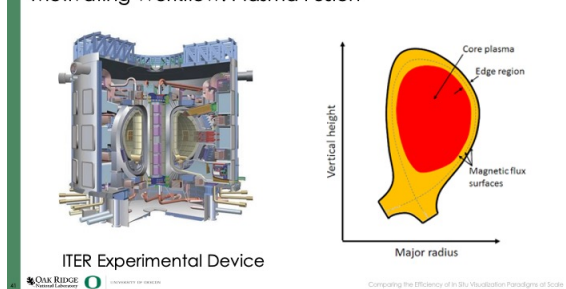


ENERGY

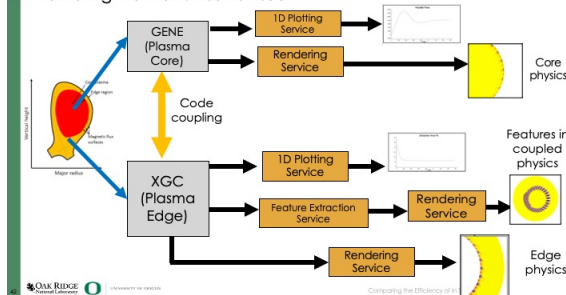
Challenges of Visualization in the Present and Future

- Gaining insight from data has always been hard
- In the past, the process was well understood
 - Low barrier to entry:** Data saved to disk and easily shared
 - Collaboration:** Each stakeholder can access at convenience
 - Standardization:** Common file formats, access patterns (e.g. from disk), analysis and visualization tools
- Challenges:**
 - Big Data:** Saving all data to disk is infeasible. In situ is becoming the norm
 - Complex hardware:** Heterogenous computing, deep memory hierarchy and the rising costs of moving data
 - Automated workflows:** Requires analysis and visualization to be more tightly integrated
 - Lack of cohesive solutions:** Existing in situ efforts are often "one off" solutions that are brittle, difficult to generalize

Motivating Workflow: Plasma Fusion



Motivating Workflow: Plasma Fusion



Motivating Workflow: Plasma Fusion

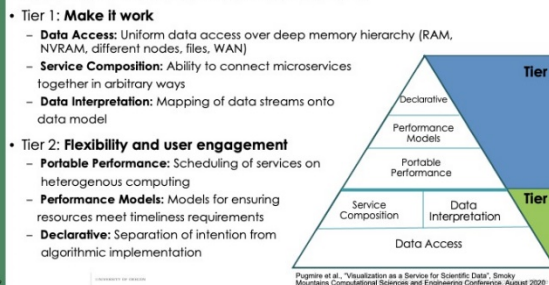
- It would be easy to hack this together for a workflow
- But....
 - Workflows change
 - Codes change
 - Using in other domains
- Things are going to get a lot more complex
 - Example: Federated Systems

Visualization as a Service

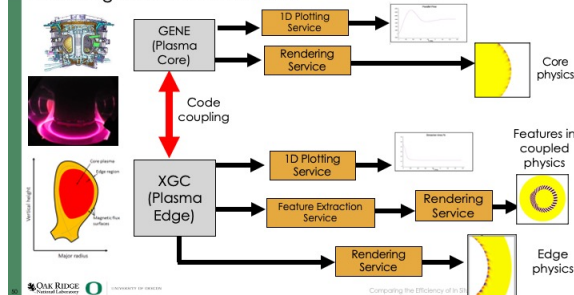
- A move away from monolithic systems will mitigate challenges of streaming data and complex computing
- Service Oriented Architectures (SOA) address these challenges
 - SOA: self-contained black box with well-defined feature set
 - Examples include Infrastructure as a Service (IaaS), Software as a Service (SaaS), and Microservices
- We need to work together
 - Our own area is complicated enough
 - Invent new wheels?
 - Or leveraging other expert-made wheels to create scalable solutions



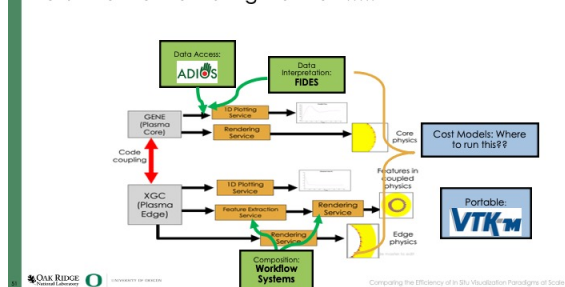
Visualization as a Service Abstractions



Motivating Workflow: Plasma Fusion



Return to the motivating workflow.....




FAIR Workflows and Data at ORNL

- Vis and analysis is hard and getting harder
 - Working together is critical
- Vis as a Service approach:
 - Avoid inventing new wheels
 - Interfaces to leverage other expertly-made wheels
- FAIR principles:
 - Accessible: Data models
 - Interoperable: Portability, cost models
 - Reusable: Dynamic



INTEGRATION and
MODELING for
PREDICTIVE
BIOLOGY




KBase

PREDICTIVE BIOLOGY


DOE Systems Biology Knowledgebase

FAIR Workflows in KBase

Zach Crockett,
Computational and Predictive Biology



Office of Science
Office of Biological and Environmental Research



What is KBase?

- Open, FAIR biological data science platform that empowers scientists to collaboratively drive discovery, prediction, and design of function in plants, microbes, and their communities.
- KBase Narratives are custom, reproducible workflows that hold data, analyses, and knowledge.
- Public Narratives can be copied freely and reproduced by re-running the apps with the same data or new data
- Narratives can be "published" - static versions made available outside the login, with DOI (by request).



4

Microbial Resource Announcements (MRA)



- Organic workflow created independently by KBase users in more than 50 publications
- Example Workflow:
 - Isolate an organism
 - Sequence Genome
 - Classification
 - Evaluate metabolic capabilities



5

MRA Template and FAIR Workflows

- Enhances FAIRness of workflows:
 - **F**: Static Narratives are findable through search engines and Google Dataset Search (if it has a DOI).
 - **A**: KBase is free to access and requires no software downloads.
 - **I**: The internal data representation allows data interoperability inside KBase and can be exported using data-specific standard file types.
 - **R**: Narratives can be copied and re-run with new or the same data. Templates allow the same workflows on different data.
- Model for other publishers



6

FAIR Data Workflows and Data at ORNL

- ORNL is the most represented laboratory in the published datasets
- Plant Microbe Interfaces - published genomes through MRA:

Published Genome (full citation at DOI)	Associated Static Narrative
<i>Terrioglobus albidus</i> Strain ORNL (doi:10.25982/44746.21/1635640)	https://kbase.us/n/44746/21/
<i>Starkyea</i> sp. Strain ORNL1 (doi:10.25982/55377.22/1637507)	https://kbase.us/n/55377/22/
<i>Roseimicrobium</i> sp. Strain ORNL1 (doi:10.25982/56021.26/1637512)	https://kbase.us/n/56021/26/
<i>Larkineella</i> sp. Strain BK230 (doi:10.25982/54100.27/1635639)	https://kbase.us/n/54100/27/



7

DOE User Facilities Implementing FAIR Data Principles: ARM Data Center Example

Ranjeet Devarakonda and Team

Presenter: Ranjeet Devarakonda
Group Leader, ARM Data Science & Integration
devarakonda@ornl.gov

Oak Ridge National Laboratory

About Atmospheric Radiation Measurement

ARM is a multi-lab/multi-institute User facility and a key contributor to national and international climate research efforts.

- Provides high quality, research data products for atmospheric and climate sciences.
- Primary focus on measurements needed to advance the understanding of clouds and radiative feedbacks.
- To use this understanding to improve the performance of climate models.

The ARM Data Center (ADC) was established in 1994 to collect, manage, and distribute data produced by the Atmospheric Radiation Measurement User Facility (ARM).

FAIR Data Principles

F

A

I

R

Findable

Accessible

Interoperable

Reusable

FAIR: ARM Data Discovery Interface

Findability: Data and supplementary materials have rich metadata and have a globally unique and persistent identifier

FAIR: Meta(data) sharing in External Portals

Improves the visibility of ARM data products in external data clearinghouses and relevant scientific portals.

FAIR: Data Retrieval, Packaging, and Delivery

Accessibility: Metadata and data are understandable and accessible to both humans and machines. Data is archived in a trusted repository.

FAIR: Web Services

Popular with timeseries data

Directly download data into users workspace

Saves time and repeatable

Captures Metrics per download

<https://adc.arm.gov/armlive/>

FAIR : Data formats and usage

- ARM data and metadata are stored in non-proprietary, community defined, standards based structured format.
Example: data is in NetCDF and Metadata is stored in RDBMS, but can be exported to ISO-19115-2, JSON-LD, FGDC- CSDGM.
- Follows 20-year rule: The metadata accompanying a data set should be written for a user 20 years into the future.
- ARM meta(data) is comprehensive in a way that it captures data quality information and lineage.
- ARM data is free to use, distribute, remix, and built upon. CC license by 4.0

Interoperability: Metadata use a structured, accessible, shared, and broadly applicable language for knowledge representation.

Reusability: Data and supplemental data have clear usage licenses and provide accurate information on provenance.

FAIR Workflows and Data at ORNL

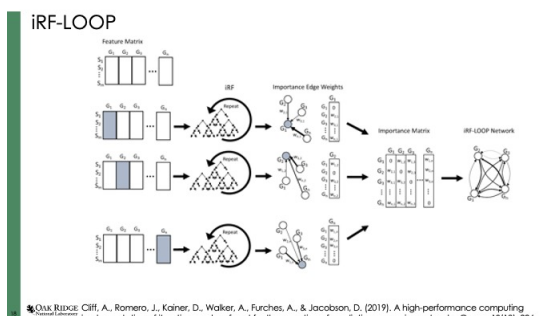
Thank you!

Ranjeet Devarakonda
devarakonda@ornl.gov

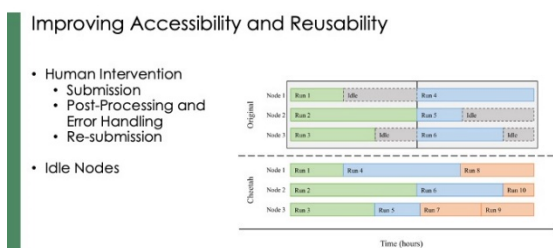
Improving Reusability and Accessibility of iRF-LOOP using the Cheetah-Savanna Suite of Workflow Tools

Angelica M. Walker
Graduate Research Assistant

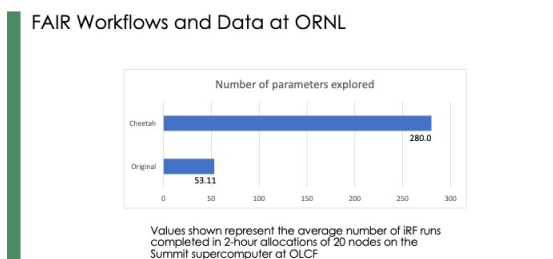
ORNL is managed by UT-Battelle LLC for the US Department of Energy



Oak Ridge: Cliff, A., Romero, J., Kainer, D., Walker, A., Furches, A., & Jacobson, D. (2019). A high-performance computing implementation of iterative random forest for the creation of predictive expression networks. *Genes*, 10(12), 996.



Oak Ridge: Wolf, M., Logan, J., Mehta, K., Jacobson, D., Cashman, M., Walker, A.M., Eisenhauer, G., Widener, P., Cliff, A. (2021, September) Reusability First: Toward FAIR Workflows [Paper Presentation]. IEEE Cluster 2021, Portland, OR.



Oak Ridge: Wolf, M., Logan, J., Mehta, K., Jacobson, D., Cashman, M., Walker, A.M., Eisenhauer, G., Widener, P., Cliff, A. (2021, September) Reusability First: Toward FAIR Workflows [Paper Presentation]. IEEE Cluster 2021, Portland, OR.

Acknowledgements

ORNL	Georgia Institute of Technology
Matthew Wolf	Greg Eisenhauer
Jeremy Logan	Sandia National Laboratories
Kshitij Mehta	Patrick Widener
Mikaela Cashman	
Ashley Cliff	
Daniel Jacobson	
JAIL	

Oak Ridge

A workflow for neutron scattering data analysis

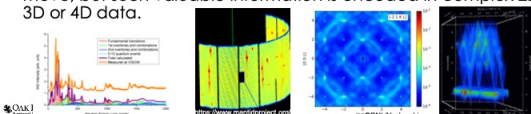
YQ Cheng
Computational Instrument Scientist
Spectroscopy Group
Neutron Scattering Division, ORNL

CAW 2021
July 20, 2021

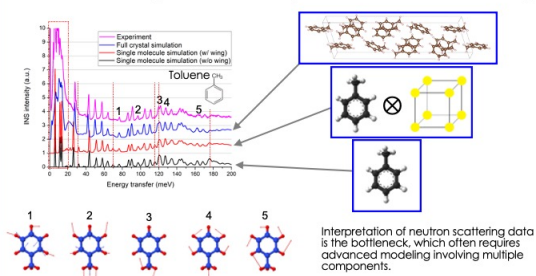
ORNL is managed by UT-Battelle LLC for the US Department of Energy

What is neutron scattering

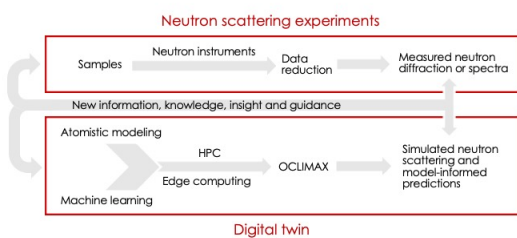
- Just like photons and electrons, neutrons can be used to bombard a material to probe its atomic-level structure and dynamics
 - Raman/infrared scattering
 - Electron energy loss spectroscopy
 - X-ray diffraction
 - Synchrotron x-ray scattering
 - Transmission electron microscope
 - X-ray photoelectron spectroscopy
- Neutron scattering can see where atoms are and how they move, but such valuable information is encoded in complex 2D, 3D or 4D data.



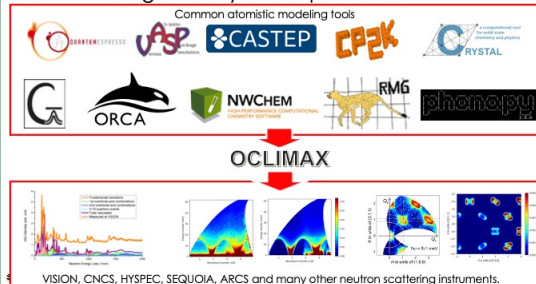
Atomistic modeling is an ideal companion for neutron scattering



The digital twin for neutron scattering



OCLIMAX bridges theory and experiments



FAIR Workflows and Data at ORNL (Neutron Sciences)

- Analysis cluster for centralized data storage and access
- Autoreduction workflow
- Data analysis workflow
- DOI for publication and citation



From Microscopic Images to Simulations via Deep Learning: A Comprehensive Workflow to Develop Physics-Based Understandings

Ayana Ghosh, PhD
Postdoctoral Research Associate
CNMS, CSED

Acknowledgments: Maxim Ziatdinov, Sergei V. Kalinin

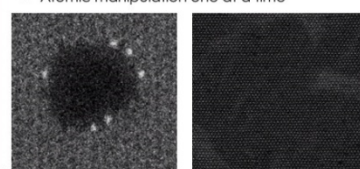
ORNL is managed by UT-Battelle, LLC for the US Department of Energy



ENERGY

ORNL: Home of the Most Powerful Microscopes in the World

- Wealth of information
 - Atomic structure
 - Quantum properties
 - Atomic manipulation one at a time



Data Collected by Ondrej Dyck

Challenges: Connecting Pathways between Experiments & Theory

- Imaging Materials:** Just imaging/characterization is not enough
- Decoding Data:**
 - Instrument Specificity: aleatoric and epistemic uncertainties
 - Implementation Complexity: formats and volumes
- Transferring Information:** experiments to theory
 - Feature finding, identify regions of interest
 - Scaling up to perform feasible simulations
 - On-the-fly analysis
- Reproducing & Extending** Workflows to Other Physical Scenarios

MIDST: Microscopic Images to Atomistic Simulations via Deep Learning

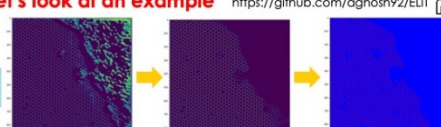
Stage 1: Microscopic Images (DCNN) → Stage 2: Atomic Types & Coordinates → Stage 3: Atomistic Simulations (Transform) → Physical Phenomena (Feature Finding & Physics-Informed Learning)

Data Machine Learning Human-in-the-Loop Learn Physics

HyperSpy Scipy Readers Universal Data Formats → AtomAI → Theory & Simulations Feed back to DL models & experiments

MIDST: Let's look at an example <https://github.com/aghosh92/EUIT>

Task 1: Feature finding



Algorithm 1 Steps to train a DCNN Network

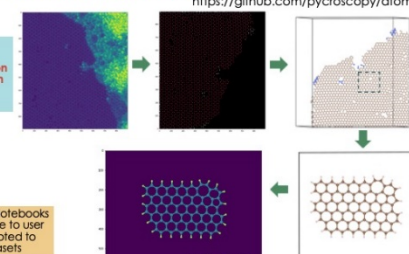
- $P_{train} \leftarrow \text{DataTrain}(P_{data})$ → Train a model on P_{train}
- $P_{test} \leftarrow \text{DataTest}(P_{data})$ → Split a model to test data
- $P_{val} \leftarrow \text{DataVal}(P_{data})$ → Choose a validation set
- $\text{for } i \in [1, \dots, N]$ **do**
- $P_i \leftarrow \text{DataTrain}(P_{train})$ → Generate training data
- $P_i \leftarrow \text{DataTest}(P_{test})$ → Test a new model on test data
- $P_i \leftarrow \text{DataVal}(P_{val})$ → Apply a model to test data
- end for**
- $P_{train} \leftarrow \text{DataTrain}(P_{train})$ → Generate a model on training set
- $P_{test} \leftarrow \text{DataTest}(P_{test})$ → Test a new model on test data

A. Ghosh et al., *npj Comput Mater* 7, 100 (2021)

Open Access Notebooks
Freely available to user
Can be adapted to other datasets

MIDST: Let's look at an example <https://github.com/pyroscopy/atomai>

Task 2 & 3: Creating Simulation Objects & Perform Scientific Simulations




Open Access Notebooks
Freely available to user
Can be adapted to other datasets

AIMD simulation

MIDST: Summary

- Microscopic image data to atomistic simulations using Python-based infrastructure
- On-the-fly workflow in future
- Steps towards autonomous experiments



FAIR Workflows and Data at ORNL

- STEM data → atomic positions
- STEM data is acquired by CNMS facility
- we currently use AtomAI to accomplish stage 1
- AtomAI takes microscopic images as inputs
- Prepares training set, constructs NN models
- Predicts/Recognizes features (positions and type of all atoms, defects present in specific image frame)
- For training a DCNN with one typical image using Google Colab Nvidia P100/V100 GPUs facility within AtomAI.
 - Time taken → on average 0.5 hours
- Having GPUs at the edge will allow us to do it at real time of data capture
- CPU-based simulation environment (on-the-fly analysis)

Python-based Infrastructure
Automated Data Transfer
GPU/CPU-based Platform
Open Access
Reproducible
Reusable

Thank you for your attention.

CAW 2021

Experimental discovery of structure-property relationships in ferroelectric materials via active learning

Yongtao Liu
Center for Nanophase Materials Sciences
Oak Ridge National Laboratory

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

Piezoresponse Force Microscopy

Ferroelectric domains

Band-excitation piezoresponse spectroscopy (BEPS)

electric field

DKL analysis of BEPS data

Input

Output

DKL prediction

DKL-BO Autoexperiment: guided by coercive field

Input: Amplitude patches; Output: Average Coercive field

Measurement Points

Step 0: Acquisition function values

DKL Prediction

• Measurement points concentrate around the regions with dense and complicated domain walls—that is saying, the measurement points are around regions with the mixture of a/c and c/c walls.

DKL-BO AE: guided by coercive field

Input: Amplitude patches

Output: Positive Coercive field

Output: Negative Coercive field

DKL Prediction for Positive Coercive Field

DKL Prediction for Negative Coercive Field

• Here, the acquisition function is based on positive and negative coercive fields, respectively.

• The measurement points guided by positive coercive field more likely concentrate around c/c domain walls, while the measurement points guided by negative coercive field more likely concentrate around a/c walls.

FAIR Workflows and Data at ORNL

• Github

Acknowledgement


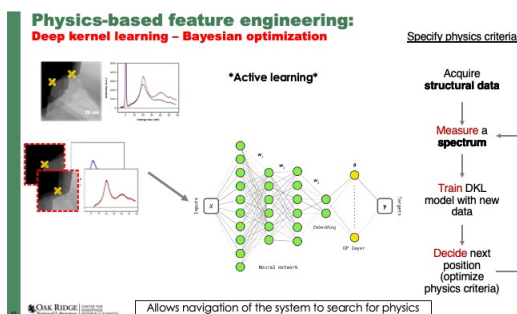
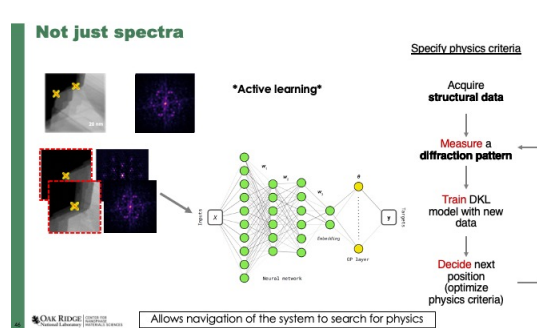
Sergei V. Kalinin
Maxim A. Ziatdinov
Stephen Jesse
Rama K. Vasudevan
Kyle Kelley

Automated Discovery of Physics in the Electron Microscope

Kevin M. Roccapriore, Maxim Ziatdinov, Sergei V. Kalinin

CAW 2021

ORNL is managed by UT-Battelle, LLC for the US Department of Energy

- ### FAIR Workflows and Data at ORNL
- As this is deployed on the instrument, how we process the data is important!
 - Must be consistent!
 - Data here is 3D hyperspectral in nature (2D space, 1D energy)
 - Metadata automatically stored during experiment, used for processing decisions
 - This workflow is compatible with common data structures in variety of instrumentation

