Study on Handling Dark Data in HPCI Shared Storage System using the WHEEL Workflow Tool

International Data Week 25/SciDataCon 25
Data and Research & Data Science and Data Analysis

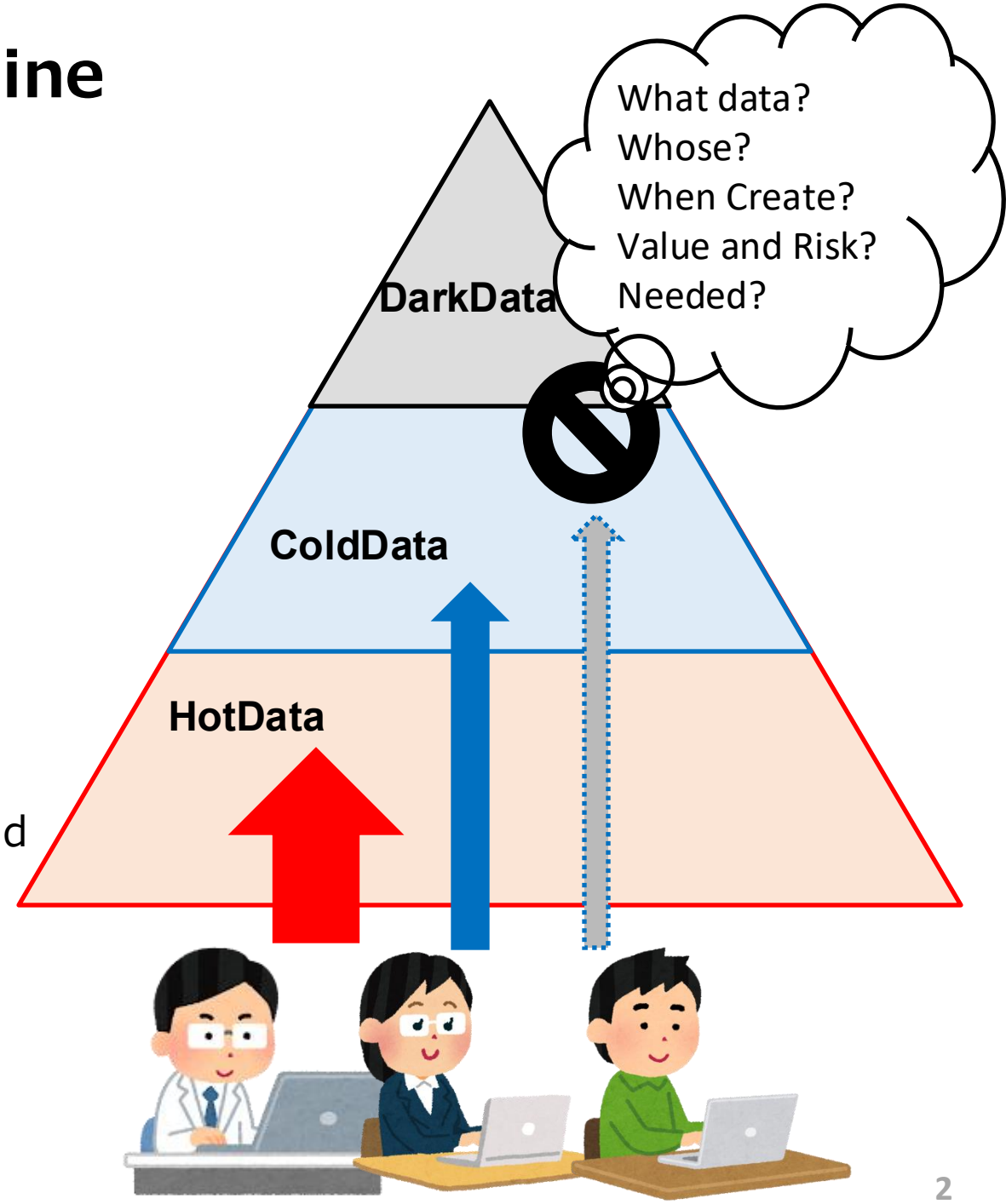
October 13, 2025

RIKEN Center for Computational Science(R-CCS)
- **Hidetomo Kaneyama**
- **Tomohiro Kawanabe**
- **Hiroshi Harada**

INTERNATIONAL DATA WEEK

Outline

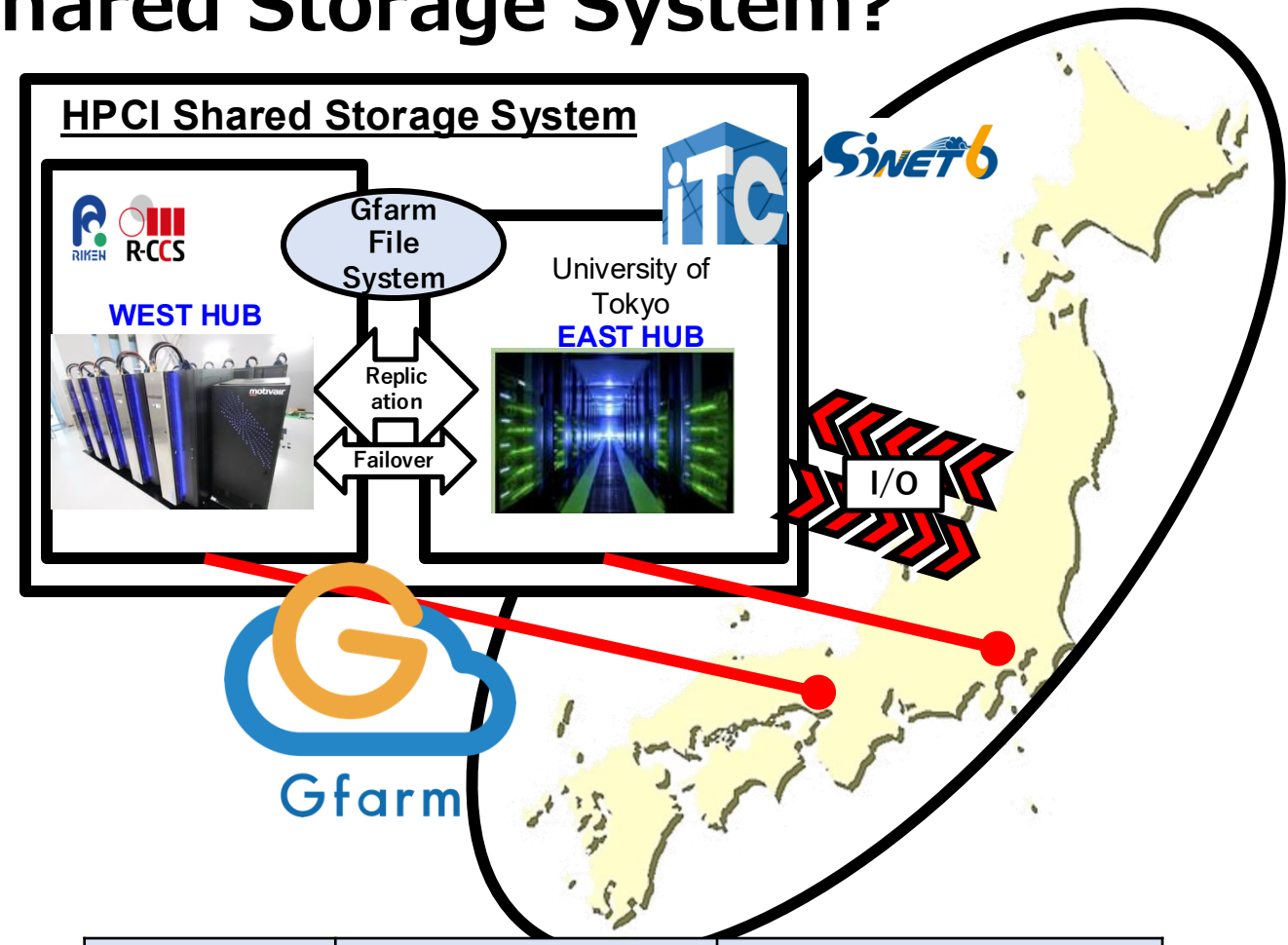
- We operates the HPCI Shared Storage Service.
- Issues:
  - Dark Data blocking “effective use” of storage, causes storage pressure.
  - Dark data:
    data unused for analysis/decision-making, with unclear value or risk.
    And, significant dark data exists in HPCI Shared Storage.
    - https://www.gartner.com/en/information-technology/glossary/dark-data
- Approach:
  - Add new functions to the workflow tool for HPC and supercomputers.
    - Automatically collect extended metadata from HPC/supercomputing systems
    - Write data to our storage with the extended metadata attached



What is HPCI Shared Storage System?

HPCI-SS(HPCI Shared Storage) System for Preserving and Sharing Data from Japanese Supercomputers.

- **Objectives**
  - Data sharing between supercomputers (computational resources)
  - Long-term preservation of research data
  - Public dissemination of research data (utilization of open/public datasets)
- **Core Software**
  - **Gfarm File System**
    https://github.com/oss-tsukuba/gfarm
- **Key Features**
  - It is provided free of charge to research projects using Japan's HPC systems.
  - Parallel data I/O transfer from HPCI resources
  - Disaster recovery through inter-site data replication
  - High availability (operational uptime >99%, downtime/unavailability <1%)
  - High-performance network storage capable of transfers exceeding 200 Gbps

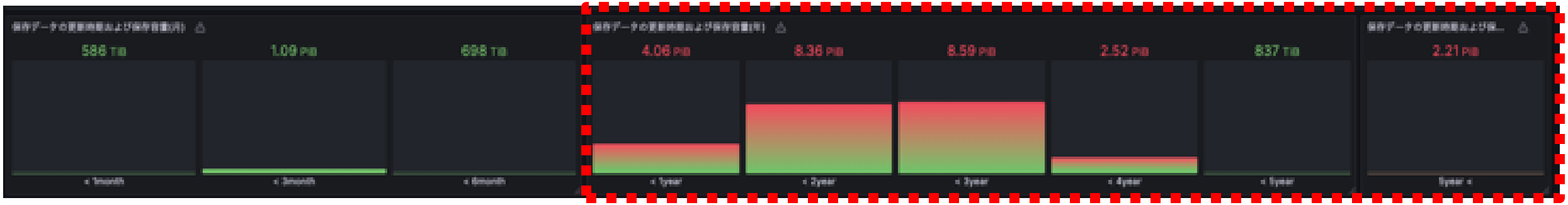


Gen	Periods	Capacity(Logical)
1st	FY2012 – FY2018	10PB → 15PB
2nd	FY2018 - FY2024	45PB → 50PB
3rd(now)	FY2025 -	95PB+

https://www.hpci-office.jp/info/pages/viewpage.action?pageId=111380786

Cold Data in HPCI-SS

Number of Files



※Cold data is 'File not Accessed in over ONE YEAR'

Total	Capacity	33PB
	Number of files	188 million
Cold data ratio (include Dark data)	Capacity	90.2%
	Number of files	88.8 %

Capacity

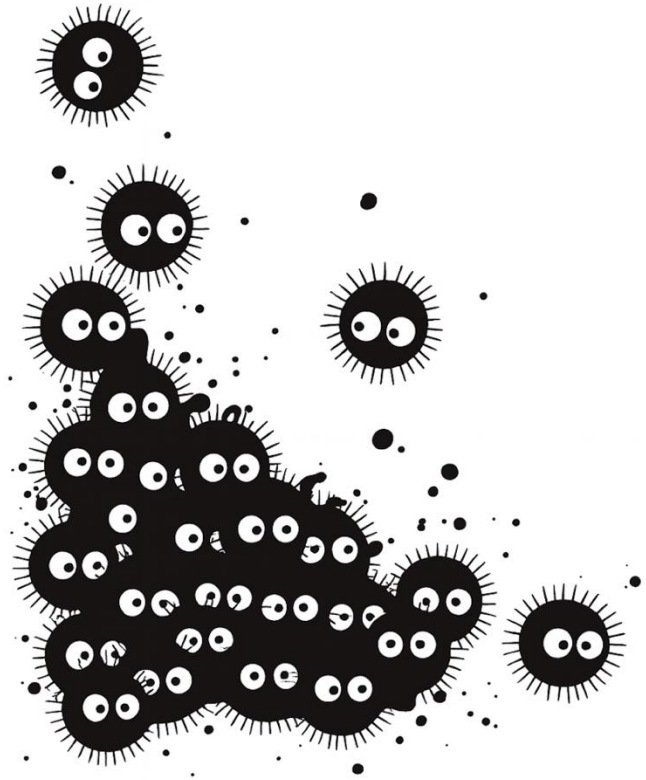
Cold data grow in HPCI Shared Storage.

- Users face heavy burden in identifying unnecessary files
- Some data has been inherited (e.g., from former members)
- Lack of metadata makes deletion decisions difficult

Issue of Data Management in HPCI-SS

The HPCI Shared Storage Service needs to provide function of "automatic extended metadata annotation" tie-up HPC/Supercomputers.

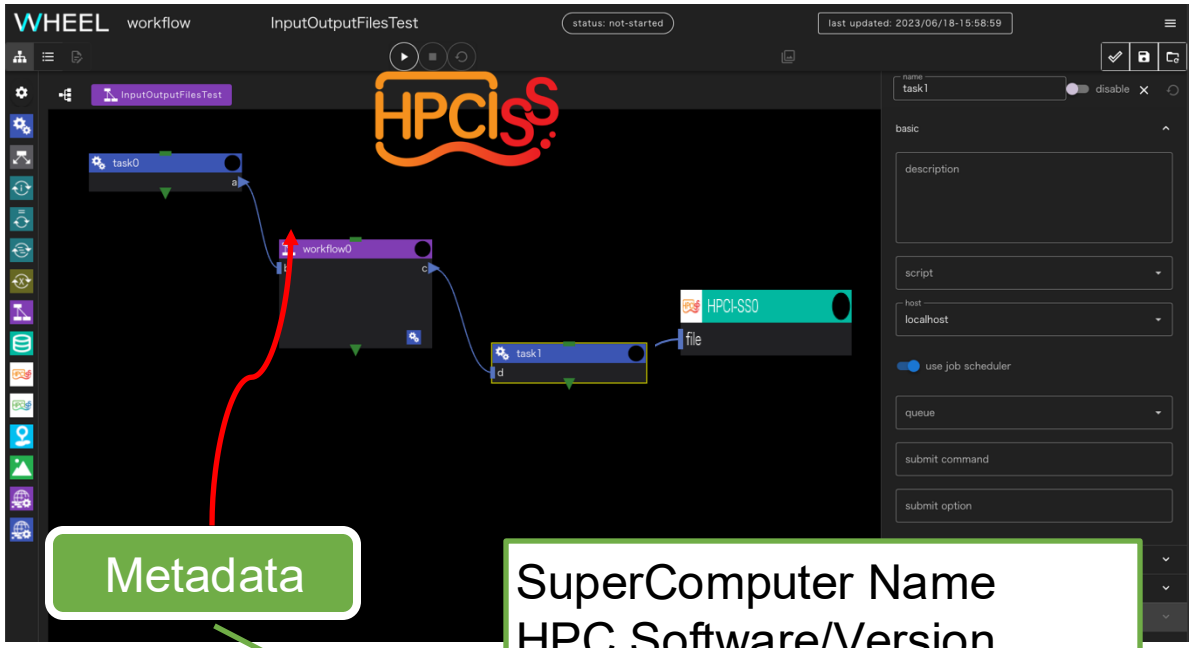
- **ISSUE[1]: How to approach to Cold/Dark Data?**
  - Ask users to remove old and unused data
  - Dark data is hard to judge due to unclear content
  - Actual stored of dark data remains unknown (only discovered when users clean up)
  - Visualization provides usage/capacity/file counts, but not data content or usefulness
- **ISSUE[2]: How to challenge open-science and AI?**
  - AI/Open Science requires labels and metadata
  - Current HPC & HPCI storage lacks tagging/labeling mechanisms
  - Object storage often missing or underutilized
  - Critical info (“who used what, and how data was obtained”) is not preserved
  - Some users keep records manually (e.g., Excel) → high management burden



https://sca25.sc-asia.org/wp-content/uploads/2025/04/pos121s1.pdf

Approach

- **Our Approach**
  - User Side
  - Seek improvements without direct system modifications
  - Collaborate with WHEEL workflow tool
  - Automatically collect extended metadata during workflow execution
    - HPC Name
    - Project Name
    - JobID
    - Pre/PostName
    - UserName
    - Software Name
  - Store enriched metadata + datasets into HPCI-SS
  - Long-term goal: integrate with DOI assignment
- **Progress**
  - FY2024: Enabled WHEEL I/O to HPCI-SS via Gfarm API
  - FY2025: Developing automated metadata collection during HPC compute & pre/post-processing



https://sca25.sc-asia.org/wp-content/uploads/2025/04/pos121s1.pdf

Discussion to Dark Data in HPCI-SS

■ Why use Workflow Tools?

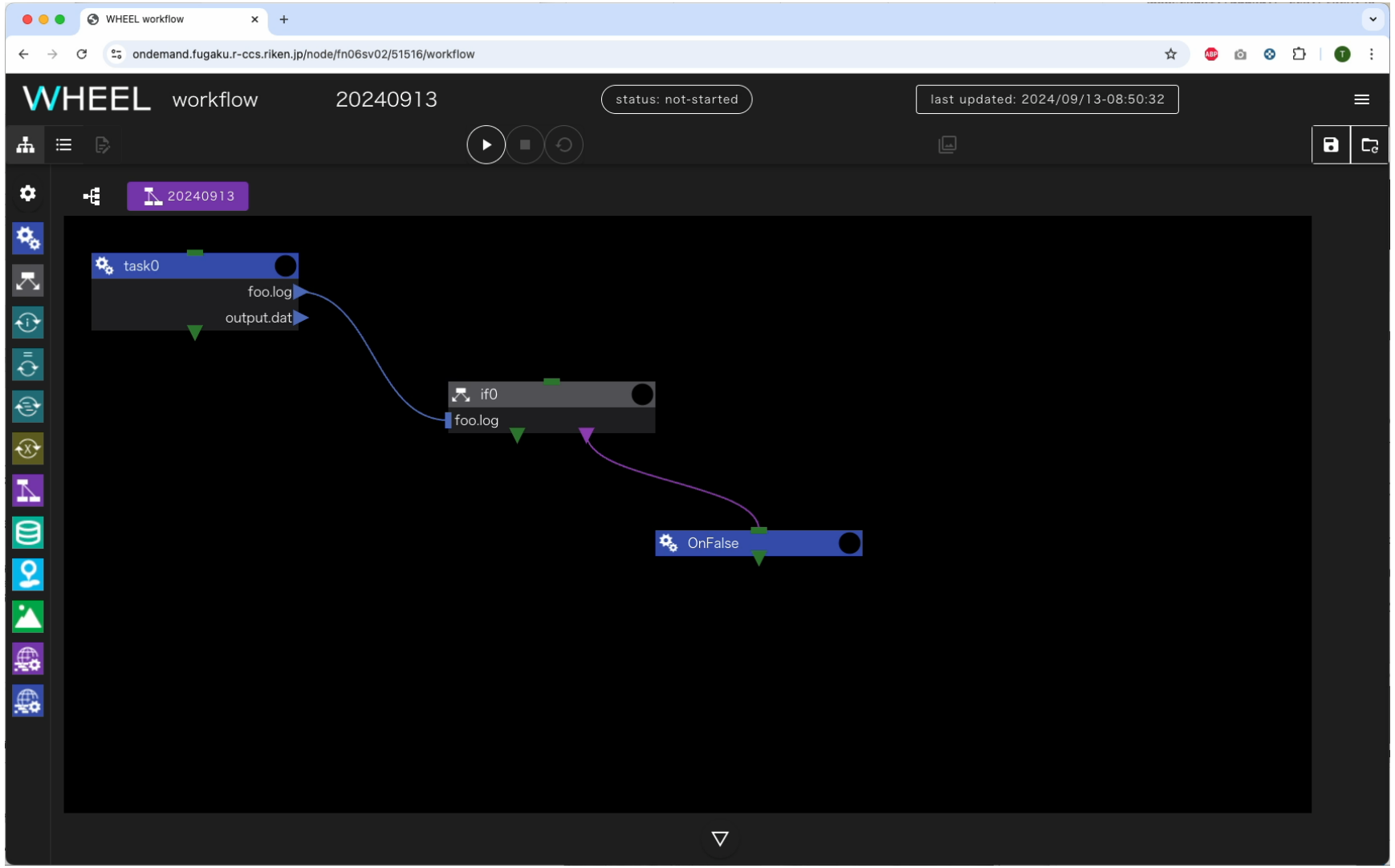
- Low Cost:
  - Replacing entire AI or Metadata controls storage(e.g. VAST) is too expensive.(We need more capacity)
  - Commercial metadata products (e.g. Starfish, IBM AFM System)
    - Management authority at each supercomputer sites.
    - High license and operational costs.
- External Factor:
  - Data management need to starts at the point of data generation(Computation to used HPC/Supercomputers).
  - HPC storage (e.g., Lustre, BeeGFS) is fast, but nothing expanded metadata functions.
- **Implement metadata management on “User side = Workflow side”.**

■ What does this solve?

- Automatically collects expanded metadata during HPC/Supercomputer computations.
- Independent of supercomputer site storage systems.
- Manage of expanded metadata per data or dataset (supports compressed and archived data).
- Metadata-based search available in HPCI-SS.
- **Dark Data can be identified by automatically adding and saving metadata extracted from the workflow information.**

WHEEL: A Web-based Workflow tool

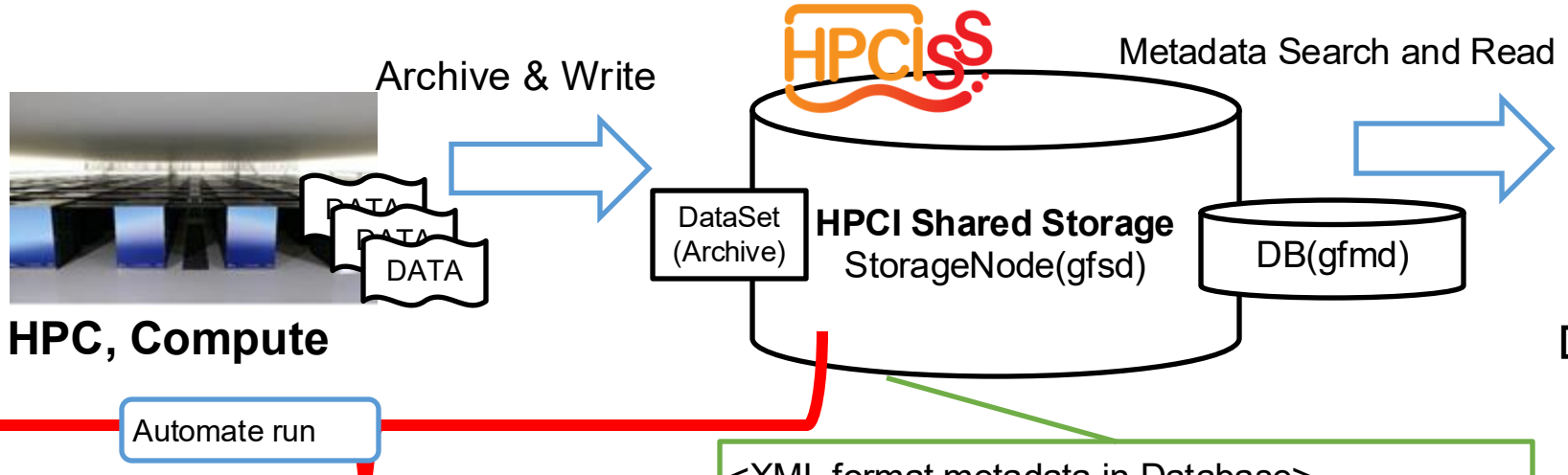
- A web-based GUI workflow build and execution tool
- JavaScript application using Node.js
- Open-source software distributed under the BSD-2 license

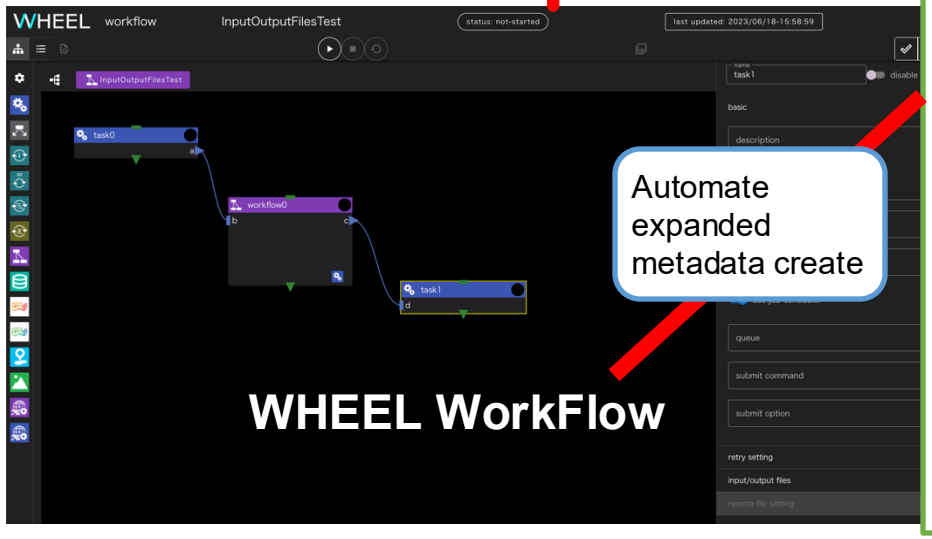


https://github.com/RIKEN-RCCS/OPEN-WHEEL

Method And Future Challenges

Future challenges





Automate expanded metadata create

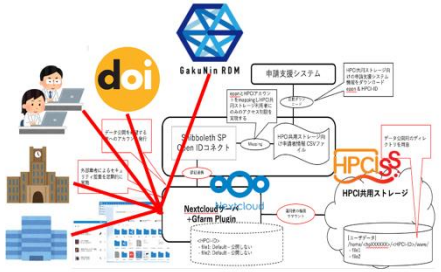
<XML format metadata in Database>
PreComputeNodeName
PreComputeSoftwareName/Version
HPCName
HPCSoftware/Version
HPCJobID
PostComputeNodeName
PostComputeSoftwareName/Version
Purpose
UserName(GroupName)
Create Time(HPC Using)
Tags
etc…

[1] Develop Access/Search Platform

[2] Visualization of Expanded Metadata

[3]Collaboration to DOI

[4] Connect to/for RDM or Open Science Service



https://sca25.sc-asia.org/wp-content/uploads/2025/04/pos121s1.pdf

Keyword & Reference

■ Keywords

- **Dark Data**
  - https://www.gartner.com/en/information-technology/glossary/dark-data
- **HPCI Shared Storage**
  - https://www.hpci-office.jp/info/pages/viewpage.action?pageId=111380786
- **Gfarm File System**
  - https://github.com/oss-tsukuba/gfarm
- **WHEEL Workflow Tools**
  - https://github.com/RIKEN-RCCS/OPEN-WHEEL

Gfarm

WHEEL

■ References

- [1] Tatebe, O., et al. (2010). Gfarm grid file system. New Generation Computing, 28(3), 257–275. https://dl.acm.org/doi/10.1007/s00354-009-0089-5
- [2] Bauer, D., et al. (2022). Revisiting data lakes: The metadata lake. In Proceedings of the 23rd International Middleware Conference Industrial Track (pp. 8–14). ACM. https://doi.org/10.1145/3564695.3564773
- [3] Kawanabe, T., et al. (2024). Introduction of WHEEL: An analysis workflow tool for industrial users and its use case on supercomputer Fugaku. In Proceedings of the 2024 IEEE International Conference on Cluster Computing Workshops (pp. 180–181). IEEE. https://www.computer.org/csdl/proceedings-article/cluster-workshops/2024/834500a180/21EtRZFluLu
- [4] Jain, A., et al. (2015). FireWorks: A dynamic workflow system designed for high-throughput applications. Computational Materials Science, 96, 118–124. https://doi.org/10.1016/j.commatsci.2014.10.037
- [5] Chiapparino, G., et al. (2024). From ontology to metadata: A crawler for script-based workflows. INGGRid. https://www.inggrid.org/article/3983/galley/3912/download/


Thank you