# Fusion Cloud

## – Open Science Platform for Fusion Research in Japan –

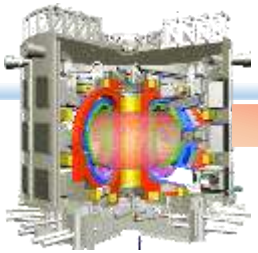
***Nakanishi H.*, Emoto M., Ohdachi S., Osakabe M., Watanabe K.Y., Nakajima N., Yamanaka K.†, Yokoyama M.***

***National Institute for Fusion Science (NIFS)***

***†National Institute of Informatics (NII)***

Note that the contents and opinions expressed in this presentation reflect the views of the presenter, and do not represent the official views of the affiliated institution.

# What is "Fusion Cloud"?



Approx. 100 PB/yr Data Transfer from ITER IO to "IFERC REC"

(Collab. with NII, QST & NIFS)

QST Rokkasho Fusion Institute
NIFS Rokkasho Research Center

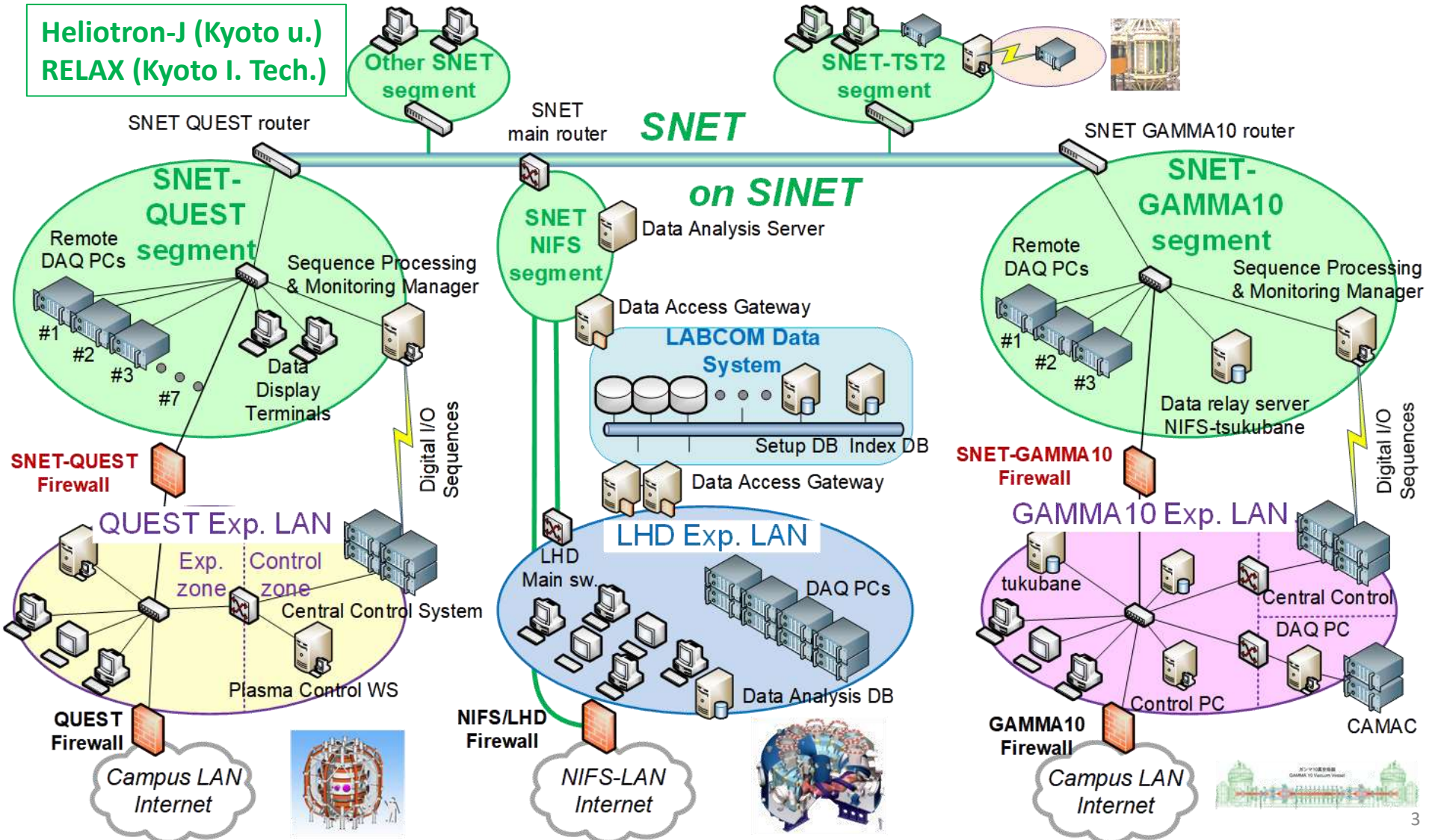

## Fusion Cloud

**Interdisciplinary Use of Fusion Data, Devices, and Computers**

- "Place of Practice" for Data-driven Sciences and "Fusion Industries"
- "Place of Education" for Expanding the Fusion Research Community
- "One-stop Services" for Collaborative R&D

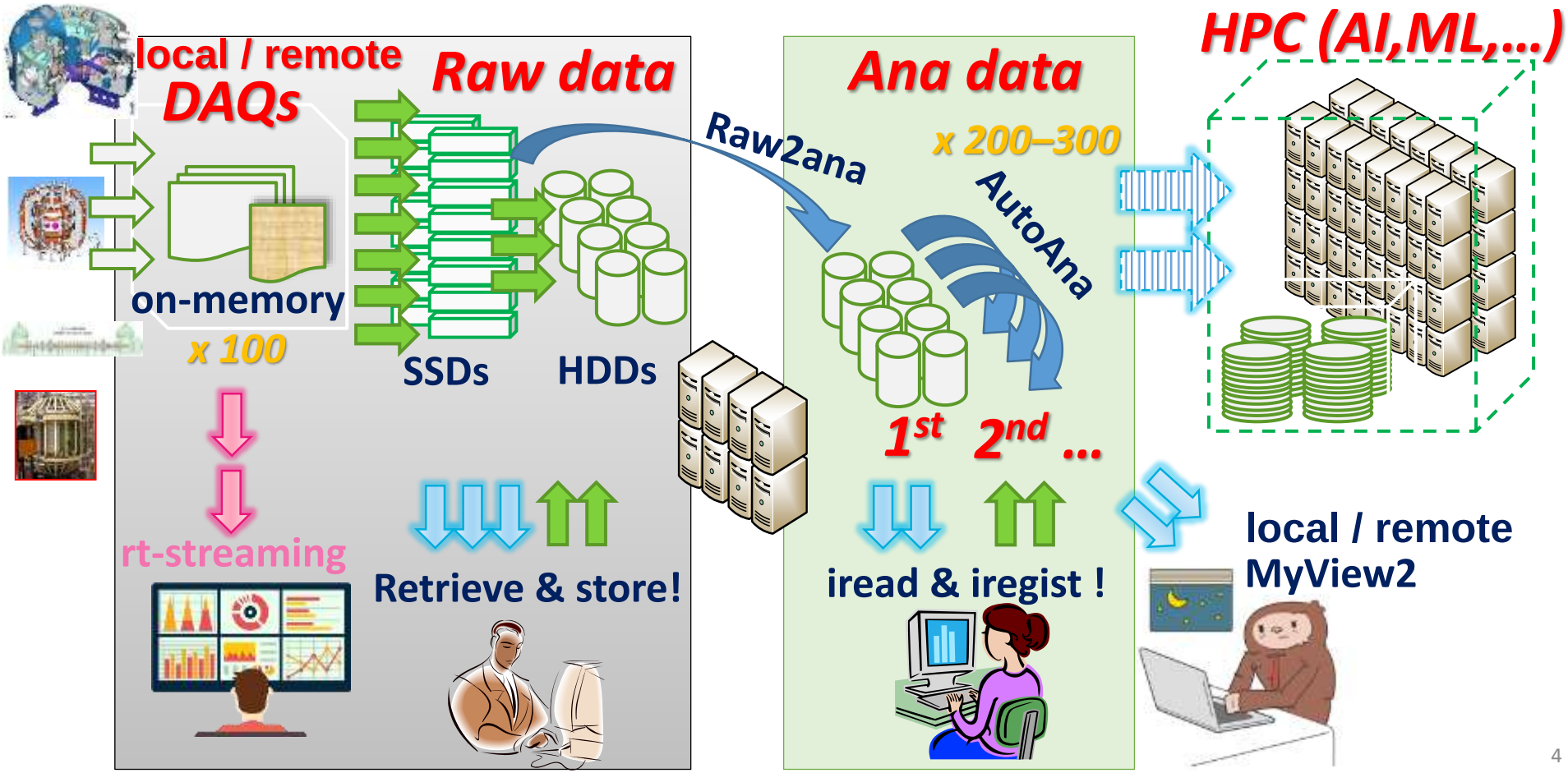
# Fusion Virtual Laboratory & SNET–LHD

**SNET and the "Storage Area Network" are all private, isolated from Internet.**



# Data chains for processing/analysis Automation

- More than 200 pre-registered processes will be woken up automatically when new input data has come to “trigger” them



# New for "Fusion Cloud"

## Trend towards DS, OS and RDM

- DS methods, such as ML, require huge datasets near HPC resources.
  ➔ Real (experimental) & virtual datasets shall be equivalently utilized.
- For OS, research data should follow the FAIR (Findable, Accessible, Interoperable and Reusable) data principles.
- Assigning global persistent identifiers (PIDs), such as DOI, can help users to directly refer to and cite data objects in their publications and presentations. ➔ Traceability

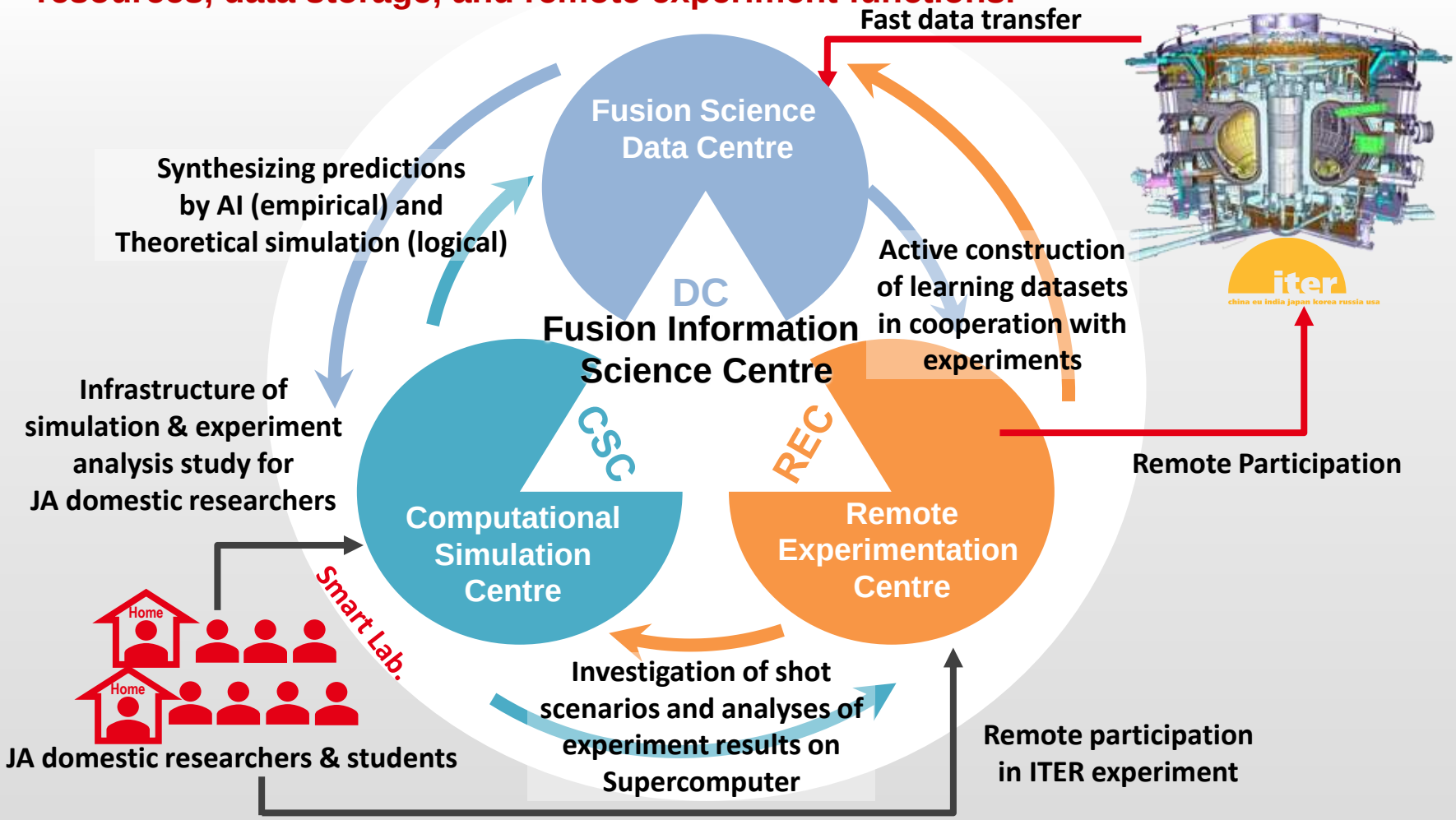
## Neighbor Activities

- Fusion Information Science Centre (FISC) is a new project of QST Rokkasho Fusion Institute in Japan, which aims to realize an infrastructure for fusion data analysis and model calculation.
- The IMAS (Integrated Modeling and Analysis Suite) project is making steady progress in the development of a suite of related software to realize the computation platform for ITER.
- FAIR4Fusion – European activity toward OS/OD in fusion research

# Fusion Information Science Center (FISC)

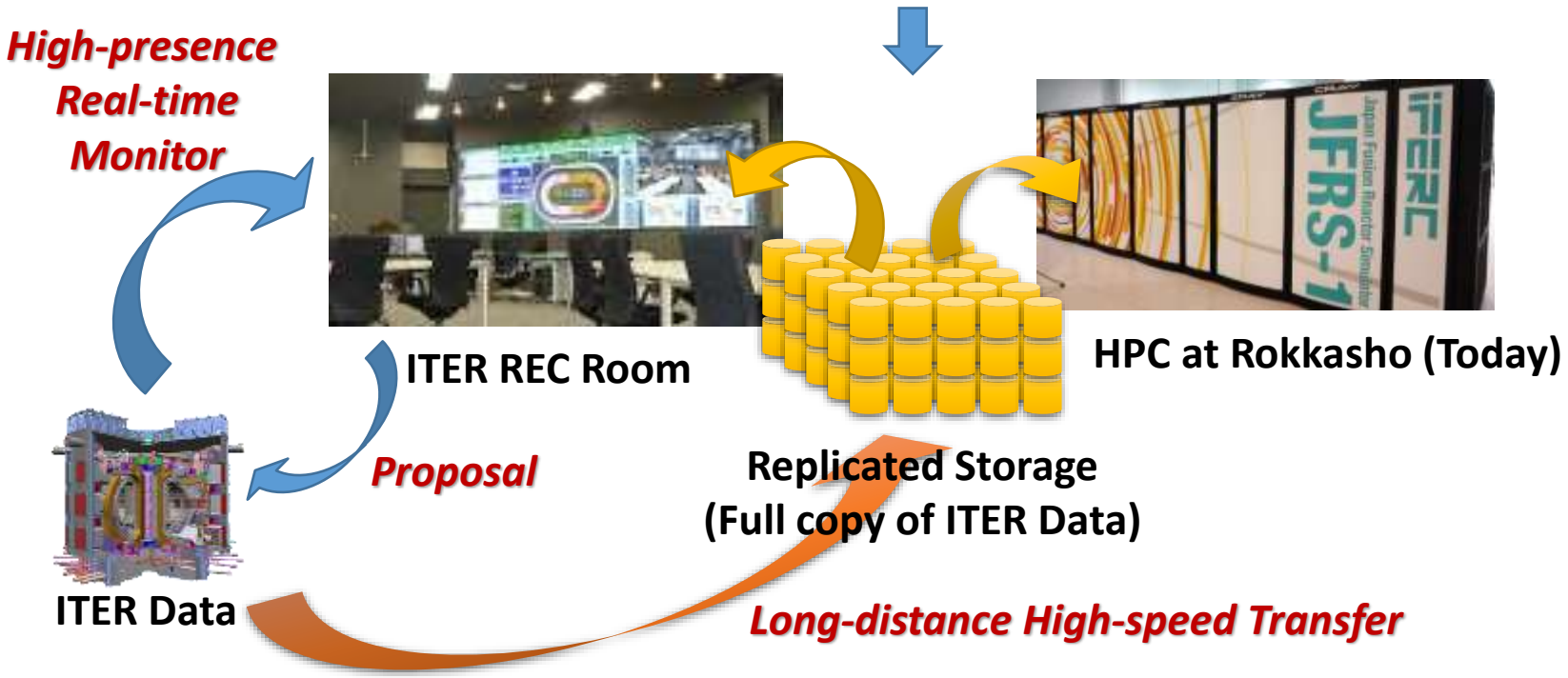
*cf. S. Tokunaga, et al., ITC30 19Ea2 (2021).*

**QST Rokkasho Institute plans to integrate the computing resources, data storage, and remote experiment functions.**



# ITER Data Analyses with HPC at *FISC/REC*

- **FISC is preparing to perform ITER remote experiments from Rokkasho, Japan. Goals are:**
  1. **REC room will provide high presence as if people were on-site control room**
  2. **All the ITER data will be replicated to the REC storage almost in real-time for enabling high-performance data analyses off-site using HPC.**



# The FAIR Guiding Principles

*As open as possible, as closed as necessary*

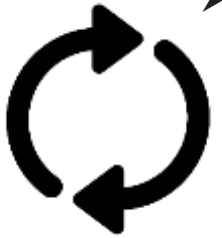
already ok
somehow
not yet

*cf. FORCE11, https://www.force11.org/fairprinciples*

- **To be Findable:**
  - ✓ F1. (meta)data are assigned a globally unique and persistent identifier
  - ✓ F2. data are described with rich metadata (defined by R1 below)
  - ✓ F3. metadata clearly and explicitly include the identifier of the data it describes
  - ✓ F4. (meta)data are registered or indexed in a searchable resource

- **To be Accessible:**
  - ✓ A1. (meta)data are retrievable by their identifier using a standardized communication protocol
  - ✓ A1.1 the protocol is open, free, and universally implementable
  - ✓ A1.2 the protocol allows for an authentication and authorization procedure, where necessary
  - ✓ A2. metadata are accessible, even when the data are no longer available

- **To be Interoperable:**
  - ✓ I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
  - ✓ I2. (meta)data use vocabularies that follow FAIR principles
  - ✓ I3. (meta)data include qualified references to other (meta)data



- **To be Reusable:**
  - ✓ R1. (meta)data are richly described with a plurality of accurate and relevant attributes
  - ✓ R1.1. (meta)data are released with a clear and accessible data usage license
  - ✓ R1.2. (meta)data are associated with detailed provenance
  - ✓ R1.3. (meta)data meet domain-relevant community standards

# Data License and Persistent ID (PID)

## Data Licensing

- In general, popular licenses are often adopted for data usage agreement.
  - ✓ GNU General Public License (GPL) and the derivatives (v2, v3, LGPL, …)
    - – strong license inheritance for the secondary products
  - ✓ BSD license – weaker inheritance
  - ✓ Creative Commons (CC) License
    - – preserves the copyright with "some rights reserved", neither "all rights reserved" nor "public domain".
  - ✓ "CC BY NC ND" is often adopted for scientific data open, which requires the impose of credit Attribution, Non-Commercial, and No-Derivatives conditions.

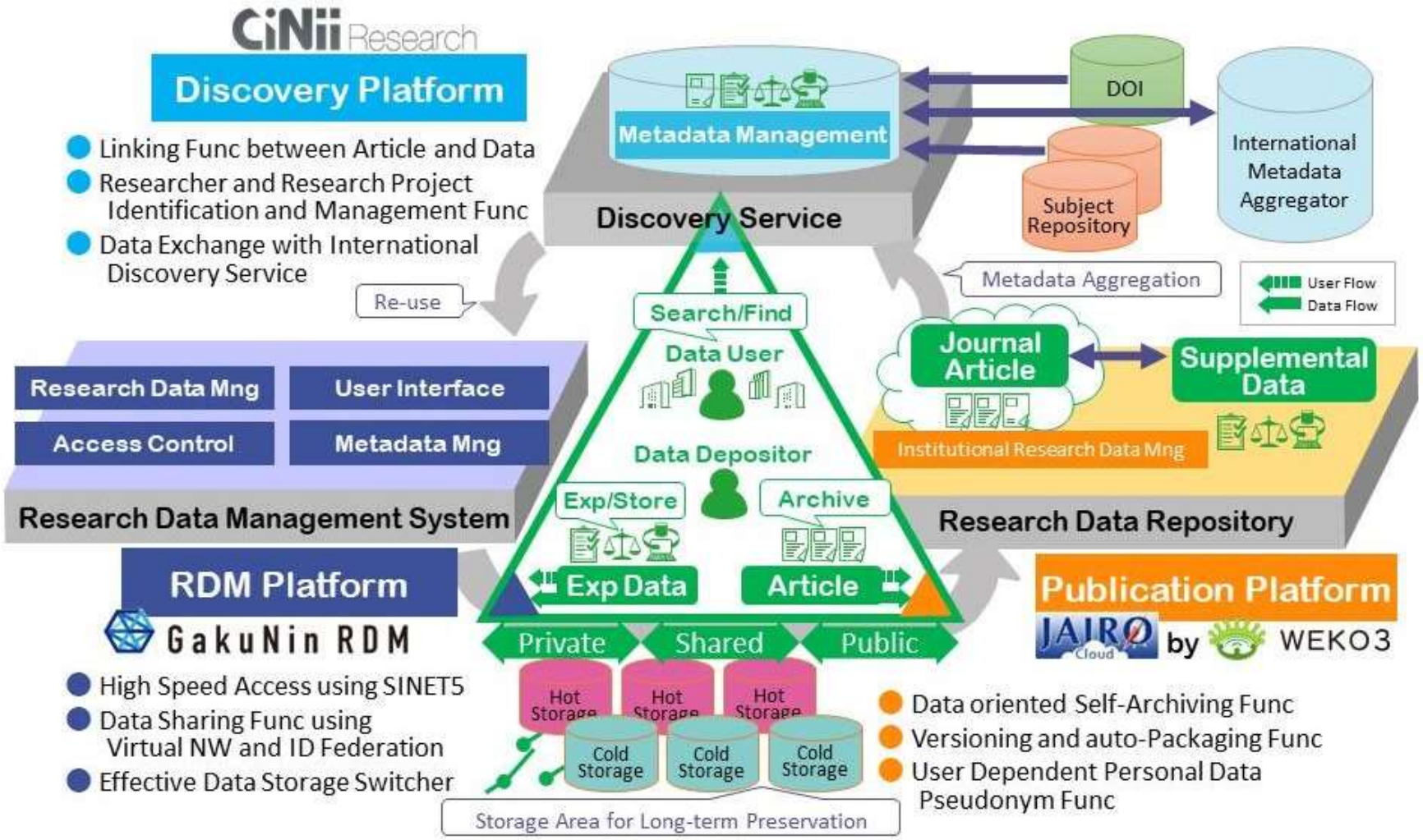

## Global Persistent ID on data

- FAIR F1. principle require PID on data.
  - ✓ LHD-SNET data have tens of million entries by "data-name × shot-no." basis.
    ➔ Granularity, registration fee, application/registration interval, …
  - ✓ Data revisions can be dealt with a unique ID ?
- Discussions to make a consensus/common rule for Open Data is just started in Japanese fusion research society.
  - ✓ International projects, ITER and the BA, may have various stakeholders …

# NII Research Data Cloud (RDC) since 2021

*https://rcos.nii.ac.jp/en/service/*

- **'FC' will utilize the NII RDC as the basis of data literacy & lifecycle infra.**
- **Interconnection between RDC and FC is needed for fusion specific use-cases**



# Summary and Issues TBD

**"Fusion Cloud" is proposed to realize the interdisciplinary data platform for both experiments and model calculations across research institutions in Japan. Cooperation with RDC, FISC, and academic community are very essential for OS.**

1. **Analyzed data**
   - ✓ Might have 'revisions' – How to manage the revision/history by time?
   - ✓ Data uploading destination/storage – Where is appropriate?
2. **Issue PIDs (DOI etc.) for each data object**
   - ✓ LHD-SNET have tens of million data entries by "data-name × shot-no." basis
   - ✓ Granularity, Registration fee, Application/Registration interval, …
   - ✓ Unique ID vs. Data revisions
3. **Data Licensing – How could the data be "as 'closed' as necessary"?**
   - ✓ Implementation for authentication, authorization, and restricted disclosure
   - ✓ Agreements are needed across projects and institutions – or common license?
4. **Standardization in fusion community**
   - ✓ Standardization also for application, publication, and reference/usage procedures
   - ✓ Cross-platform? or compatibility with *e.g.* IMAS, FAIR4Fusion, ePIC?

➔ **Need discussions in the community**