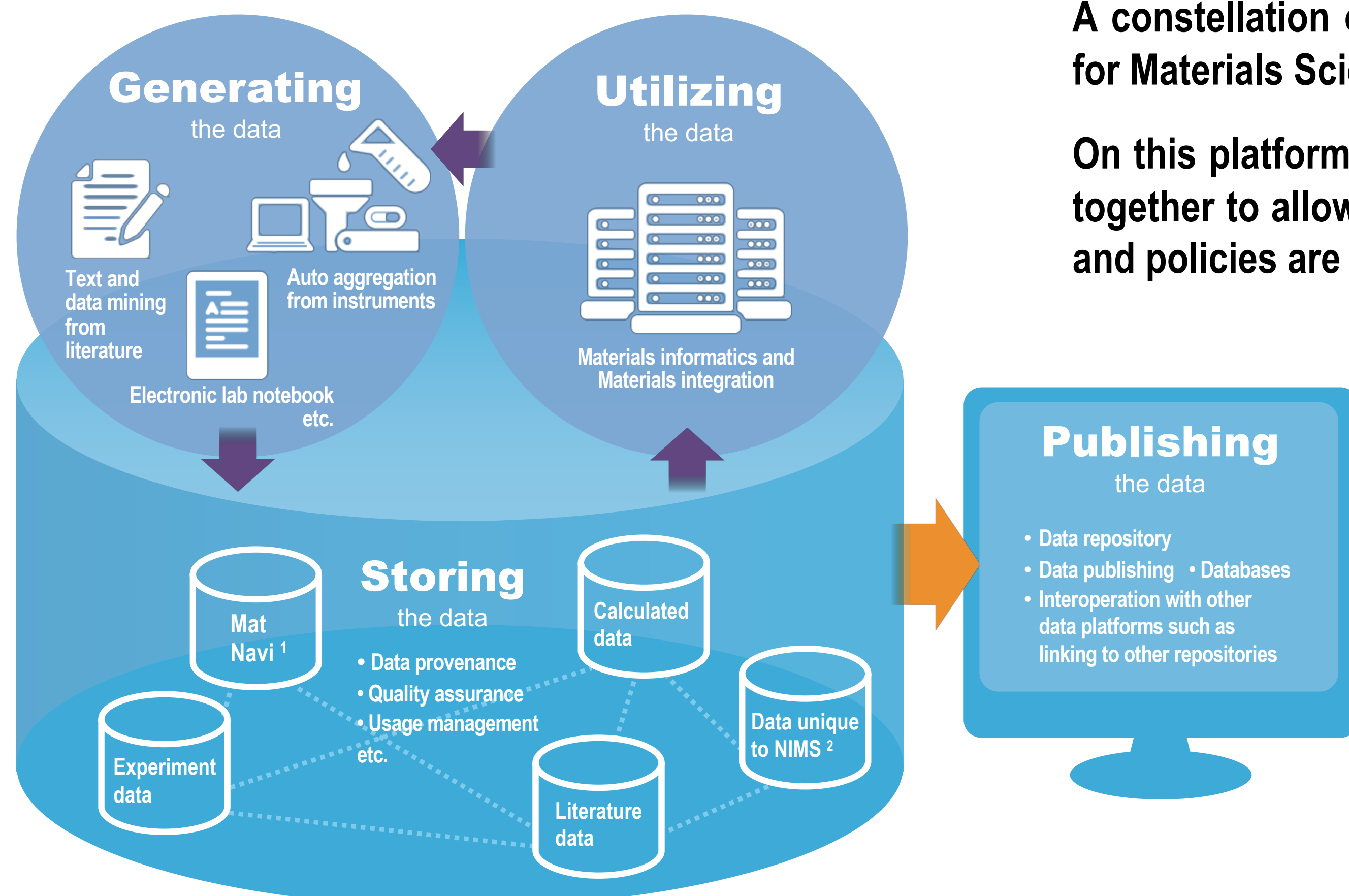# Materials Data Platform
a constellation of systems to promote materials data-oriented science

Asahiko Matsuda, Takuya Kadohira, Hideki Yoshikawa, Masashi Ishii, Masahiko Demura, Mikiko Tanifuji
NIMS, Tsukuba, Japan   matsuda.asahiko@nims.go.jp

A constellation of research data systems is being developed at the National Institute for Materials Science (Japan), to boost data-oriented approach to materials science.
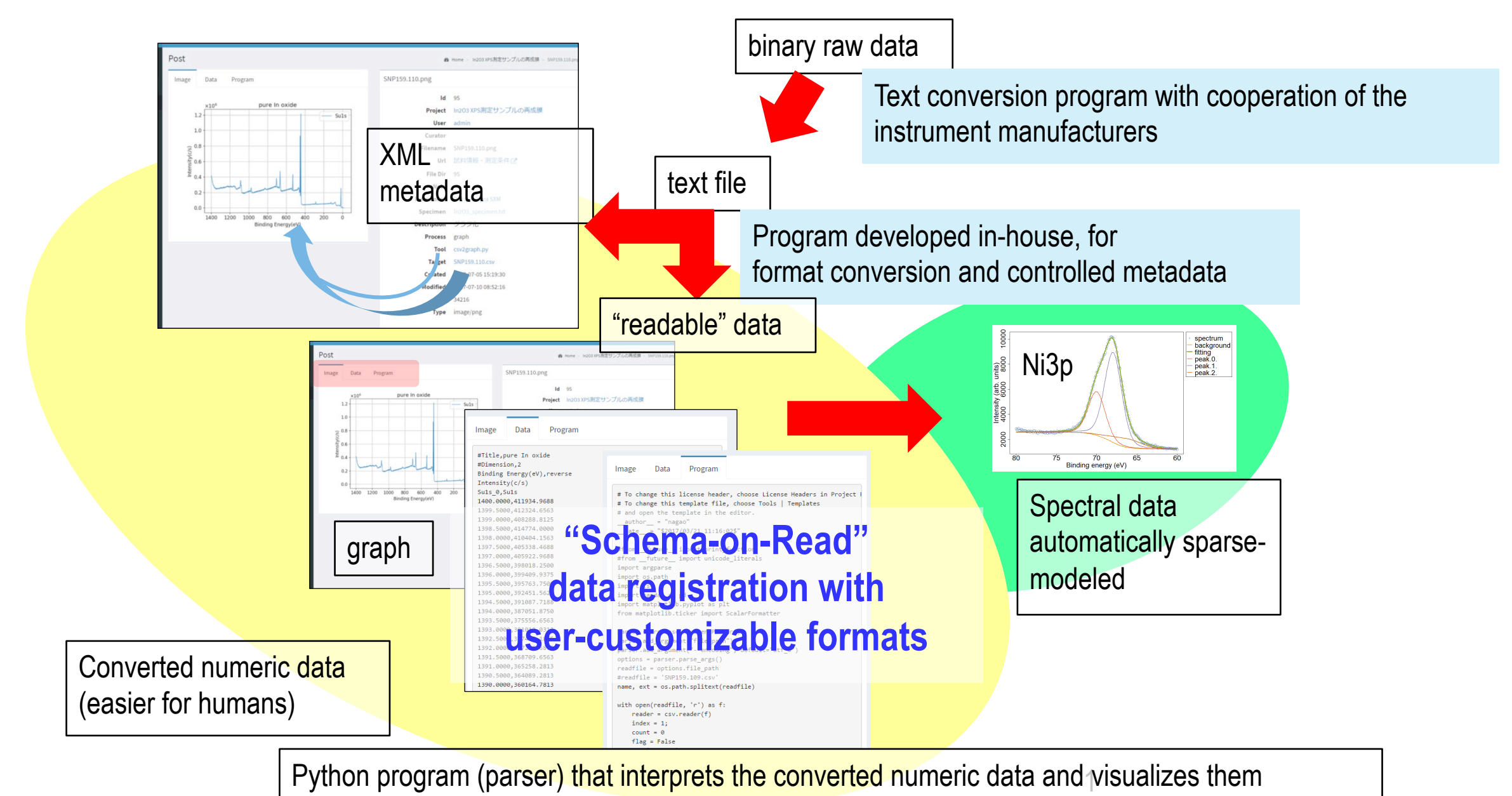
On this platform, each system provides a specific functionality, while being integrated together to allow materials data to flow. Common metadata schemas, API frameworks, and policies are being developed to allow integration between heterogeneous systems.

Generating the data
- Text and data mining from literature
- Auto aggregation from instruments
- Electronic lab notebook etc.

Utilizing the data
- Materials informatics and Materials integration

Storing the data
- Mat Navi [1]
- Experiment data
- Calculated data
- Literature data
- Data unique to NIMS [2]
- Data provenance
- Quality assurance
- Usage management etc.

Publishing the data
- Data repository
- Data publishing • Databases
- Interoperation with other data platforms such as linking to other repositories

[1] NIMS's extensive materials database
[2] Accident investigation records, researcher profiles, etc.

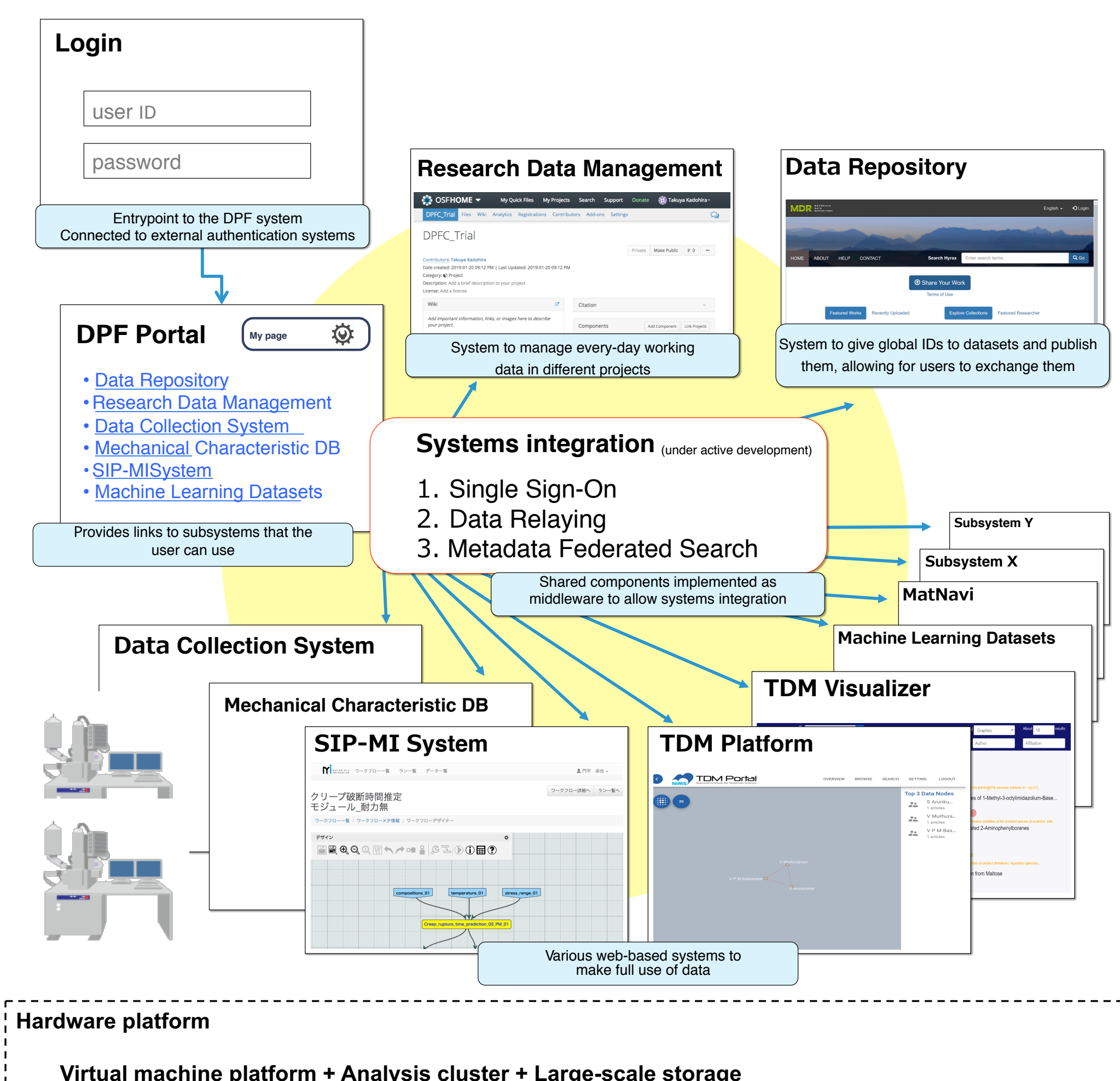| Generate	
	DCS Data Collection System
	M-DaC Data Conversion Tools
	ELN Electronic Lab Notebooks
	IoT Auto data aggregation using IoT
	TDMPF Text and Data Mining Platform
Store	**RDM** Research Data Management (based on OSF)
	VocPF Vocabulary Platform
	MatNavi Materials Database
Utilize	**SIP-MI** Materials Integration System
	DFS Data Federation System for machine learning
Publish	**MDR** Materials Data Repository (based on Hyrax)

Five key concepts

1. Quality of the data
- Identify who/what/when/how in the metadata
- Integrity of the data (hashes)

2. Accessibility
- URI / DOI / PID - based management
- Lab → RDM with DMP → MDR → analysis
- MDR ⇔ other repos and DBs

3. Usability of the data
- Licensing (CC, MIT, etc.)
- Machine-readability of datasets and its metadata

4. Safe environment
- CAS authentication and permissions
- 10-year preservation of data
- User policies (for depositors, downloaders)

5. Research aiding functionalities
- Vocabulary for TDM and materials informatics
- Data analysis environments
- API to connect platform services

Data Collection System for efficient measurement data collection and automatic conversion

- binary raw data
- Text conversion program with cooperation of the instrument manufacturers
- XML metadata
- text file
- Program developed in-house, for format conversion and controlled metadata
- "readable" data
- Ni3p
- graph
- Spectral data automatically sparse-modeled
- **"Schema-on-Read" data registration with user-customizable formats**
- Converted numeric data (easier for humans)
- Python program (parser) that interprets the converted numeric data and visualizes them

Interconnected systems

Login
- user ID
- password
- Entrypoint to the DPF system Connected to external authentication systems

DPF Portal — My page
- Data Repository
- Research Data Management
- Data Collection System
- Mechanical Characteristic DB
- SIP-MISystem
- Machine Learning Datasets
- Provides links to subsystems that the user can use

Research Data Management
DPFC_Trial
- System to manage every-day working data in different projects

Data Repository
- System to give global IDs to datasets and publish them, allowing for users to exchange them

Systems integration (under active development)
1. Single Sign-On
2. Data Relaying
3. Metadata Federated Search
- Shared components implemented as middleware to allow systems integration

- Subsystem Y
- Subsystem X
- MatNavi
- Machine Learning Datasets

Data Collection System
- Mechanical Characteristic DB
- SIP-MI System
- TDM Visualizer
- TDM Platform
- Various web-based systems to make full use of data

Hardware platform
Virtual machine platform + Analysis cluster + Large-scale storage

Automatic data conversion programs, made freely available

https://www.nims.go.jp/MaDIS/en/about/en_Top_of_M-DaC.html

Meta information extraction tool of measurement data to enhance machine readability

M-DaC Materials Data Conversion Tools

Three concepts of M-DaC
Make creating, storing, sharing measurement data easy to use for AI and machine learning more efficient and convenient.

Convert
- From machine-specific binary data to human-readable and interoperable / reusable data
- Visualization of measured data by spectrum etc.

Extract
- Minimum measurement conditions that are indispensable for reproduction measurement etc. is extracted as "main parameters"
- Redefining main parameters that commonly use measurement device unique terms in common terms

XML Output
- XML format output of main parameters with meta information added

Provide solutions for these challenges
- Even with devices of the same maker, mutual comparison is difficult if the data format differs if the devices are different
- It is difficult to search for target data because measurement conditions etc. are not recorded in the file
- I would like to use measurement data for machine learning and AI, but there is no metadata for mashup

ULVAC-PHI, INC. — XPS (x-ray spectroscopy) Quantera XSM series
Rigaku — XRD (x-ray diffraction) SmartLab series

National Institute for Materials Science (Tsukuba, Japan)
Materials Data and Integrated System, Materials Data Platform Center

MaDIS — NIMS MATERIALS DATA and INTEGRATED SYSTEM

NIMS