Development of an Information Research Platform for Data-Driven Agriculture

Takahiro Kawamura, National Agriculture and Food Research Organization, Japan*

Tetsuo Katsuragi, National Agriculture and Food Research Organization, Japan

Akio Kobayashi, National Agriculture and Food Research Organization, Japan

Motoko Inatomi, National Agriculture and Food Research Organization, Japan

Masataka Oshiro, National Agriculture and Food Research Organization, Japan

Hisashi Eguchi, National Agriculture and Food Research Organization, Japan

ABSTRACT

Comprehensive research data are acknowledged as a necessity for research acceleration. Research institutes and universities are engaged in developing research data management systems. The National Agriculture and Food Research Organization of Japan (NARO) developed NARO-linked databases (Narolin DBs) in addition to a supercomputer. In the Narolin DB various research data on agriculture are cataloged using common metadata. The relationship between complicated data in natural science is described in RDF, property graph, or RDB format to facilitate the application of statistical analysis and machine learning. The system is unique in that it is connected to a data catalog, a private cloud database, a supercomputer for data analysis, and a data/service portal for business applications, such as a data pipeline. Through the development of agricultural information research platforms, NARO will accelerate data-driven agricultural research at various stages in the agricultural supply chain, ranging from genome analysis to plant breeding, cultivation, food processing, and food distribution.

KEYWORDS

Data Catalog, Research Data Management, Smart Agriculture

INTRODUCTION

In recent years, information and communication technology (ICT) use has increased, even in the field of agriculture, and research data have been digitized in large volumes. By appropriately collecting and managing these research data, sharing them to the extent necessary, and applying data science and AI technology, such as statistical analysis and machine learning, higher efficiency and profitability of agriculture and the creation of interdisciplinary agricultural research beyond the conventional field and region is expected. These processes are called "smart agriculture" or "data-driven agriculture," which are also required to strategically utilize research data that have not been previously shared by opening the data to external organizations, especially in the area of international cooperation (open data), sharing it among consortium parties, such as national projects (disclosed data), or concealing

*Corresponding Author

it for competitive utilization (closed data). Smart agriculture and data-driven agriculture have been around for a while; however, these processes are still not widely used, and the use of data is insufficient. One of the reasons for this is that many people, such as researchers, agricultural companies, farmers, do not know what data is available, what it is for, and how it can be used. So, the research issue in this paper is: How can we make better use of data in the agricultural sector? In other words, as a national research institute, what kind of comprehensive mechanism should we prepare to promote data utilization?

The National Agriculture and Food Research Organization (NARO), the largest agricultural research body that is composed of several institutes, was established in FY 2016 through the integration of multiple institutes under the Ministry of Agriculture, Forestry and Fisheries (MAFF) in Japan. NARO found it necessary to establish a unified research data strategy and infrastructure to promote data-driven agricultural research and development while also coordinating research data across diverse research fields, such as genomic information on animals, plants, and microorganisms, breeding information, cultivation and growth management information, including viruses and pest control, food processing and distribution information, robotic tractor design, and environmental information, such as climate change and soil information. According to a preliminary survey conducted in 2019, approximately 60% of the research data were stored in the researcher's PCs and Hard Disk Drives (HDDs) causing data sharing inside and outside NARO to be delayed.

Therefore, NARO developed an agricultural information research platform, including NARO linked databases on agricultural-related data (Narolin DB), a supercomputer that utilized the database for AI operations such as machine learning and a data distribution system called WAGRI. Especially, in response to the above research issue, a platform was provided as a pipeline to collect data, add unified metadata to make it easy to discover in the primary DB, prepare the data format as pre-processing for analysis in the secondary DB, and make it easy for non-ICT experts in the agricultural sector to utilize the data by data analysis environment provided by AI supercomputer, and by WAGRI to provide data via API to private companies that serve farmers as end-users. In addition to the development of such computational systems, an integrated mechanism was created that included institutional design for human systems, such as formulation of terms of use for research data and AI education. Thus, the contribution in this paper is to share the knowledge and experience acquired in the design and implementation of the comprehensive mechanism to promote data utilization in the agricultural sector.

In the rest of this paper, the systems and services related to the research data are outlined in the second section. The structure of NARO's information research platform and the common metadata of NARO are described in the third section. After the introduction of a supercomputer for AI operations in the fourth section, and WAGRI for data distribution for business in the fifth section, the authors discuss the comparison with existing systems and case studies on their platform in the sixth section, and a summary, future issues, and institutional measures for dissemination of the data-driven approach are explored in the final section.

RELATED WORK

A research data catalog or data repository, in a broad sense, is an information platform for electronic data storage and sharing. In particular, it refers to a system established by national research institutes, universities, and related organizations, which store and manage research data obtained with public funds in a form that can be reused, and additionally, the system provides data utilization services inside and outside the institutions (CabinetOffice, 2019). Apart from those individually provided as services by national institutes and universities, systems and software that can be used by third parties include CKAN, DKAN, Dataverse, Open Science Framework, figshare, data.world, Mendeley Data, Dryad, Open ICPSR, Zenodo, and others. In a July 2017 article (Dataverse, 2017, July 25), there was a comparison of the functionality of nine research data catalogs. In the following section, four systems, especially related to Narolin DB, and some other services, are introduced.

Open-source software, called CKAN (ckan, n.d.), for constructing data portals has existed since 2007, and it is widely used in the worldwide open data site Datahub.io and the data catalog site data. go.jp of the Japanese government. However, there was a problem that made the change of function and design difficult because it was not created as a content management system (CMS). The recent CMSs have functions such as providing data and displaying catalogs as part of website creation, and extending these functions makes it easier to add data and change the web interface than if you had to directly edit the code, and it provides a more accessible website. Since many services are provided using the same system, CMS support is highly convenient for customizing each service. DKAN (dkan, n.d.) is a data portal created as a distribution of Drupal, a CMS for enterprises. Since it can deal with the setting of detailed access rights, the addition of metadata items, and can freely change the design, it is used as a data platform in multiple local governments in Japan, and in the U.S. Department of Agriculture (USDA). More information on the differences between CKAN and DKAN (DKAN DOCS, n.d.) can be found at https://dkan.readthedocs.io/en/latest/introduction/dkan-ckan.html.

Dataverse (The Dataverse Project, n.d.; Trisovic et al., 2020) is an open-source web application developed by Harvard University to share, store, quote, and analyze research data. For a long time, the research community needed a way to provide appropriate academic credit and visibility to authors of research data. Dataverse supports improved discoverability of research data on the web and the exportation of metadata to document management tools to enable publication, citation, long-term access, and reuse of data. Specifically, it supports the Dublin Core and Schema.org metadata, which, in turn, supports Google Dataset Search. Schema.org is an initiative launched by Google, Microsoft, Yahoo, and Yandex to generate, maintain, and promote schemas for structured data on the Internet. Data portals from the French National Research Institute for Agriculture, Food and Environment (INRAe) and the International Crop Research Institute for the Semi-Arid Tropics (ICRISAT) are based on Dataverse.

In Japan, the National Institute of Informatics (NII) released GakuNin RDM (RCOS, n.d.) in 2020, a domestically localized version of the Open Science Framework (OSF) (COS, n.d.; Foster & Deardorff, 2017), which is open-source software for managing research data in the Center for Open Science in the United States. GakuNin RDM (Komiyama & Yamaji, 2017) is a web service for managing research data, including papers, and can share the data within laboratories and with collaborators. In addition to standard storage with size limitations, it also provides a way to connect on-premises storage for organizations and storage in the cloud. It is also possible to connect with the research data disclosure infrastructure (WEKO3, n.d.) of NII using the repository linkage function, and to disclose the data. It also provides version control, source code repositories, document management tools, diagram sharing services, data analysis, and add-ons to workflow engines.

In Japan, J-STAGE Data (J-STAGE Data1, n.d.), a research data-sharing platform provided by the Japanese funding agency, is available for open access data such as the scientific data in research papers. This makes it possible to display data serving as research evidence results on the screen of papers registered in J-STAGE (J-STAGE2, n.d.), a comprehensive system for disseminating and distributing scientific and technological information.

In the EU region, European Open Science Cloud (EOSC) (EOSC, n.d.) is to provide European researchers, innovators, companies, and citizens with a data catalog platform where they can publish, find, and re-use data, tools, and services for research, innovation, and educational purposes. EOSC is developed and mainly used under Horizon 2020 and Horizon Europe to support the reliable reuse of research data and all other digital objects produced along the research life cycle, such as methods, software, and publications. However, when accessing the EOSC Portal Catalogue and Marketplace, you will find a catalog site of catalog sites for each research field. There are seven sites registered in the agricultural field.

One example of such an individual catalog site that specializes in a particular field is OpenSILEX (OpenSILEX, n.d. ; Neveu et al., 2018). OpenSILEX is a collaborative project developed at a joint research unit of INRAe and SupAgro. OpenSILEX is an ontology-driven platform designed for life

science data, mainly for plant phenomics. Thus, a phenotyping information system is developed on OpenSILEX for high-throughput plant phenotyping in close interaction to plant phenotyping communities. Two application ontologies can describe objects and events involved in life science experiments. The international standardized breeding API (BrAPI) is also supported. There are many other field-specific data catalog services in Japan and abroad.

Research data on space and the earth environment are relatively easy to publish, and there are numerous data catalogs. DATA TERRA (DATA TERRA, n.d.), an international initiative, has developed a data and services research infrastructure for the Earth System. Data Terra aims to organize data access in an integrated way, making available data, products, and services related to Earth System observations via the thematic poles. In Japan, a satellite data platform called Tellus (Tellus, n.d.) is also in operation. Tellus is the data platform to democratize Japanese national satellite data with a user-friendly environment. Tellus is one of the few platforms the authors have found that has a data trading function. The authors are also considering collaboration with WAGRI, which will be discussed later.

On the other hand, there is an attempt to define a specification for managing research data with metadata, rather than a specific system or platform. RO-Crate (RO-Crate, n.d.; Soiland-Reyes, 2021) of Research Objects is a community effort to establish a lightweight approach to packaging research data with their metadata. It is based on schema.org annotations in JSON-LD and aims to make best-practice in a formal metadata description accessible and practical for use in a wider variety of situations, from an individual researcher working with a folder of data, to large, data-intensive, computational research environments. RO-Crate specification 1.1 was released in February 2021.

In the private sector, Google launched a dataset search for scientists and data journalists in September 2018. Like Google Scholar, it crawls publishers, digital libraries, and personal web pages to help find the necessary datasets from the vast amount of web information. The information is organized following the aforementioned Schema.org.

The reason for the newly developed NARO's research data catalog is that software has been chosen that will allow for the ordering of improvements and feature additions, which will be described in the following sections to the programmers or software companies who developed that software, since there is a reasonable overhead to localize and customize functionality to open-source software that has multiple developers, and it is written in a foreign language with comments and documents.

NARO Linked Databases

This section describes two basic approaches adopted in constructing the information research platform for agriculture or agricultural research institutes. The main components and functions of the Narolin DB are also outlined.

APPROACH

The Narolin DB has two roles. Its first role is as a research data catalog, and its other role is as an intensive database server on a private cloud in NARO. As a result of interviews with 18 research institutes and centers in NARO that took place for about three months in 2019, it was found that approximately one to two petabytes (PB) of various fields and types of data existed in NARO (see Table 1 and Figure 1). "Strategic opening" (Hayashi, 2018), considering the characteristics of each research field is required for the sharing of research data and openness, but various fields are included in the agricultural field, from the agricultural environment, which ranges from comparatively easy to open to a relatively closed variety development, and a completely closed agricultural drug discovery; thus, researchers in each field have very different ways of handling data. Additionally, even if it is a research field that is open to the public, the research is divided into subfield A and subfield B, although this trend is not limited to the agricultural field. There are not always many exchanges between subfields. The royalties to the organization of the researcher are low, and there are many cases in which the data

Table 1. Disciplines of the NARO Research Data

| Category	Data contents	#dataset (approx.)
Plant	genome, breeding, cultivation, images, etc.	230
Animals	genome, livestock diagnosis, management, etc.	20
Insects and nematodes	pests, diagnostic methods, etc.	50
Microorganisms	diseases, disease photographs, disease identification, etc.	60
Foods	functional ingredients, ingredient analysis, etc.	30
Environment	weather, facility environment, soil, etc.	60
Misc.	agricultural work, agricultural economy, agricultural land use status, water use, water quality, infrastructure development related, ground survey, experiment notebook, etc.	50

Figure 1. Formats of the NARO research data
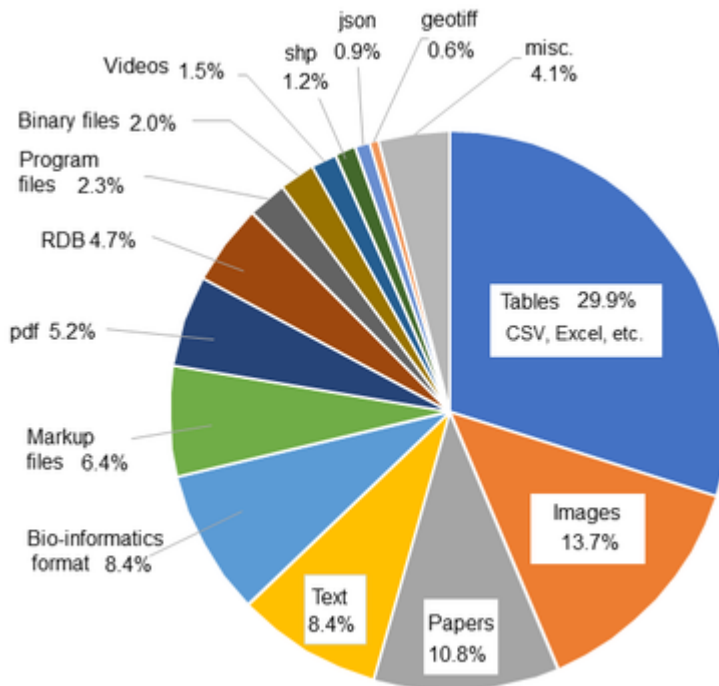
are directly submitted to individual academic societies and organizations of the subfield. In addition, there are numerous fieldwork results, and the ICT environment of such research has not always shown improvement. However, the latent possibilities of data and AI applications in agriculture is high, and the need for productivity improvement, quality improvement, and high value-added crops, such as functional agricultural products, are extensively required from both an economic and political aspect.

As an approach to the entire NARO project, it was decided to first visualize where and what kind of research data exists in NARO, to arouse interest in similar researchers and other fields, then to show as a story that the use of AI technology would lead to novel research. At least it would make the conventional work easier. The Research Center for Agricultural Information Technology of NARO would support the linkage of related data and the design of databases. It is necessary that the researcher is conscious of the incentive of the data registration. Therefore, the NARO agricultural information research platform was constructed as a pipeline, where the primary DB, as a research data catalog, is connected to the secondary DB, which consists of RDBs and graph DBs as a preparatory phase for AI use. The secondary DB is connected to an AI supercomputer for data analysis and a data distribution system for business, WAGRI. Figure 2 illustrates the agricultural information research platform. In addition, it supports commercialization and interdisciplinary research through institutional measures.

In addition, as individual approaches in each research field, the authors conducted the construction of an open data catalog mainly for environmental data, participation in the Cabinet Office program for sharing the data related to breeding domestically, and a project for the business use of image data on disease and insect damage through WAGRI.

Additionally, the research data catalog is generally made to share and open the research result, and the registration of the research result is often carried out only at the research's end and fiscal year-end. However, in research in which research periods, such as plant variety development and environmental problems are extensive, it takes a long time until results are obtained and shared with others; thus, there is a concern that the data utilization cycle will be delayed. In addition, data generated in the course of a study are often useful to others for other purposes. In the primary DB, the features of the shared folder that are used daily in the research process and the catalog of the research result are combined. By giving the operability of PCs, such as the concept of folder and commands like move, copy, rename, and a symbolic link of the file to the research data catalog, the publication of the static data as a research result and sharing of the data under study were accomplished in the same system. The utilization of the data, which crosses research fields, occurred in faster cycles.

Figure 2. Overview of the agricultural information research platform
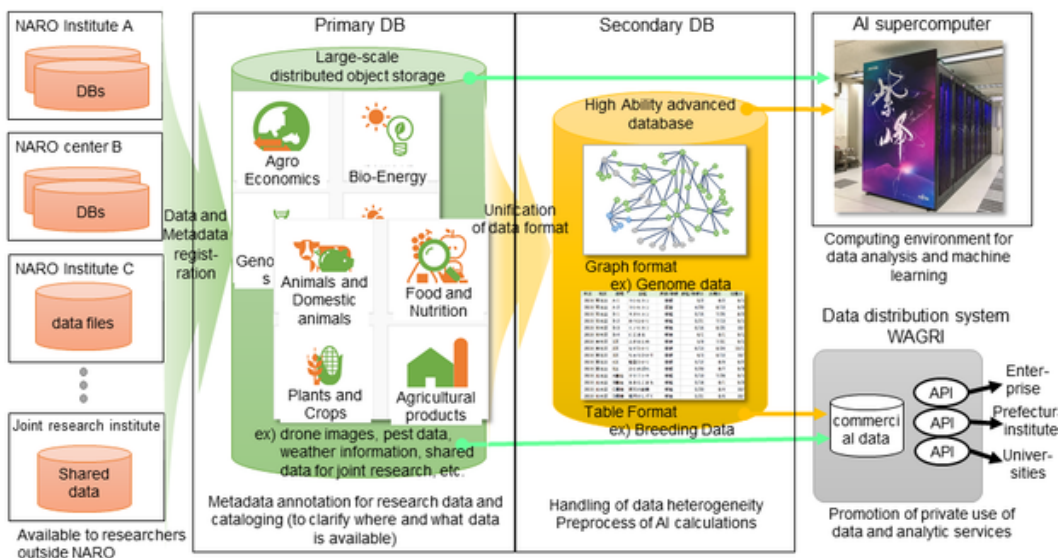
PRIMARY DB

The primary DB first catalogs and visualizes the existence of useful research data by sharing data within the range of access rights set by registrants and aims to promote co-creation among researchers by finding novel research themes and developing existing themes. Therefore, the annotation of NARO Commons metadata, which is specialized for the agricultural field, is required for data registration, and the following functions are the main ones provided (see Figure 3):

- Move, copy, rename, and symbolic link of folders and files.
- Facet search capability for NARO Commons Metadata.
- Full-text retrieval function mainly for document data such as research records, experimental notes, reports, papers, and so forth. Document structure analysis is performed, and document retrieval is conducted based on field information, except for encrypted ones.
- Versioning management function.
- Comment and bulletin board functions.
- Dashboard functions for data utilization, such as the number of downloads and comments.
- Cooperation with research result management systems for personnel management.

In addition, access rights setting can be provided in proportion to the organizational structure and the cross-sectional projects. Some of the systems described in the second section are for open data, and detailed access rights cannot be set; however, strict access rights setting is the most consequential function that encourages researchers to register their data. In particular, since the Narolin DB has a role as a catalog of research results and a role as a shared folder of data under study, the access right setting is a delicate issue.

Specifically, the following four types of access rights to institutes/departments/research groups or arbitrary groups/individuals, can be granted for datasets, which are equivalent to folders with metadata, and individual data files under hierarchically segmented research subjects. The research subjects are classified into the major, medium, and small subjects in NARO and cross-institutional projects. If permissions are duplicated, the higher permissions take precedence.

Figure 3. Screenshot of the research data list and search in the primary DB

Facet Search:
Data sets can be searched by metadata items

Dataset list

- **Owner:** The user who created the project, dataset, or file. All operations can be performed on objects at lower levels.
- **Editor:** An organization or user with edit permission granted by the owner. All operations can be performed on the appropriate dataset/file.
- **User:** Organization and user authorized by the owner. They cannot modify or overwrite information but can perform operations such as a download.
- **Viewers:** Organizations and users who have been granted viewing rights by their owners. Information cannot be modified, overwritten, or downloaded but can be viewed.

Newly created datasets and files inherit the permissions of the project and the upper-level datasets, and access permissions are automatically set. In addition, a user having the editor right can change the setting of the access right, and when the access right of the dataset is changed, it is automatically applied to all lower datasets and files. The metadata is viewable by all users, but the metadata itself can be made invisible for projects that need to hide the existence of the data itself. In addition, it has functions such as authority management of internal IDs and external IDs from outside NARO, and a function to automatically change IDs and access rights in response to organizational changes in cooperation with user management systems such as Lightweight Directory Access Protocol (LDAP).

SECONDARY DB

Among the data files stored in the primary DB, which is a data catalog, the authors selected data files mainly for projects that promote collaboration between datasets and those that are planned for collaboration in the research plans, according to the progress of the project, data quantity, budget, and so forth. Then, after the improvement of each data format and item, and the conversion to RDB or graph format, the transfer, and new establishment of the secondary DB are advanced. The choice between RDB and Graph DB depends on the nature of the data; Graph DB can represent information in multiple dimensions, but currently does not have high search performance. On the other hand, RDB can express information in a limited way but has high search performance. In addition, the selection is also based on the state of development of peripheral programs. For the conversion to graph DB or RDB, the authors often consult with researchers in the field, design the data schema based on how the data is to be used, such as what kind of analysis you want to do with the data, and then develop individual conversion programs to convert the data from Excel files, and so forth. This is where NAROs data scientists specialize. In addition, PDFs and images will have their URLs attached. The conversion program described above is used to add data to the secondary DB as an Extract/Transform/Load (ETL) process. The following WAGRI also converts them through the ETL process to distribute them in a JSON format.

As examples, the authors try to support the discovery of correlation and causality through the following data linkages:

- Data on functional ingredients of agricultural products are converted into RDBs, and then into graph DBs to add various relationships, such as nutritional components, physical characteristics, and taste, to them. In addition, the connection between genomic information and phenotypic information is also made into a graph DB for plant breeding and development from the upstream of a supply chain.
- For accuracy improvement of the forecast of pest outbreak, meteorological information, soil information, and pest outbreak situations in a field unit are linked in the graph DBs.
- To make the data interoperable, database linkage of genome information and phenotype information across the other research institutes and varieties was also constructed.

Since it is said that the man-hours for preprocessing occupies 90% of the time for data analysis, the knowledge discovery by statistical analysis and machine learning needs to be supported by resolving

the heterogeneity of data semantics by arranging the items in advance using predefined schemas and ontologies and unifying them into the data format suitable for standard or mutual use in the research field in the secondary DB. However, the schema and the ontology used in the DBs are different for each dataset. Designing them is an activity such as standardization or specification, which is time-consuming work. One such activity the authors are conducting is the Open API for Agricultural Machinery that will be described in the "Discussion" section. In this project, a specialized JSON format and the terms of individual items were designed by comparing the existing data formats of manufacturers. The existing terms are also used, for example, the names of pesticides and fertilizers provided by the Food and Agricultural Materials Inspection Center (FAMIC, n.d.).

When a database is installed in a secondary DB, a dataset of only metadata and an access destination (web server) is added in the primary DB. Therefore, the database in the secondary DB can also be referred to as one of the catalogs in the primary DB, and the perspicuity of the related dataset and database is ensured on the primary DB through the metadata. It seems that many data will move from the primary DB to the secondary DB in the future, but the primary DB is positioned as a portal site to the user.

SYSTEM ARCHITECTURE

Distributed object storage compatible with Amazon S3 was installed as the primary DB, and thus, the scalability and dispersibility of the business continuity plan were ensured. The authors physically located storage devices to multiple locations in Japan in preparation for an earthquake, and so forth, but those broadly coordinate and logically perform as a single storage unit. As a result of the abovementioned interviews on research data, the storage size was established as 5 PB, effectively 3 PB, except for backup.

Figure 4 shows the system architecture of the NARO agricultural information research platform, including the primary and secondary DBs. In the secondary DB, an external ID authentication system and Web Application Firewall (WAF) are placed in the cloud, and the DB

Figure 4. The architecture of the agricultural information research platform
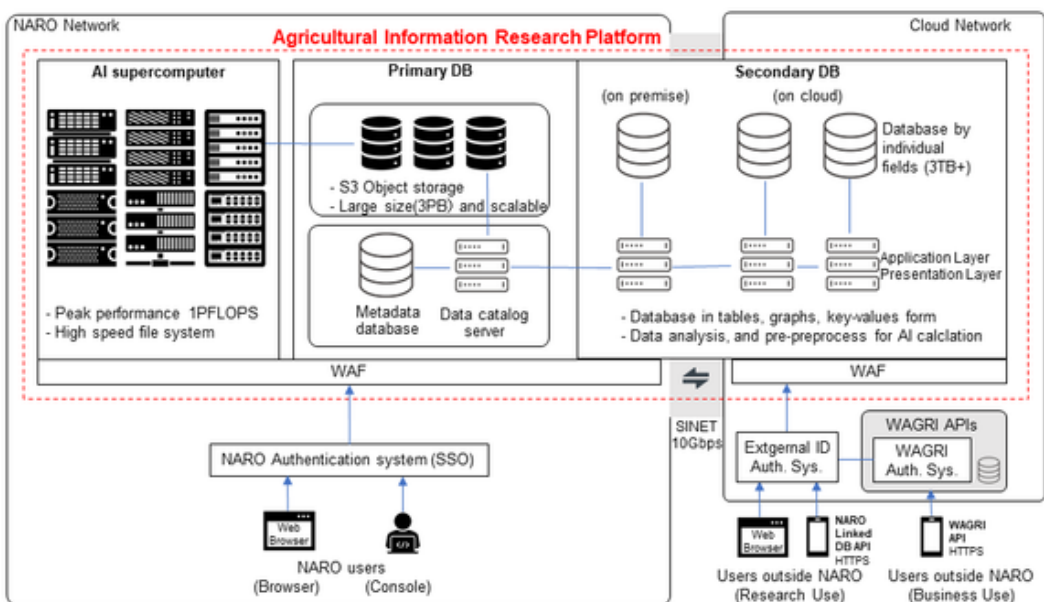
and the storage are placed in the cloud and on-premises in a hybrid configuration. In addition, multiple servers are arranged on the platform of the secondary DB as the application layer and presentation (web) layer. At present, each research institute of NARO individually manages and operates a large number of servers including web servers, OS, and databases; however, by managing them collectively, the cost and burden of maintenance is reduced and a uniform policy on security can be applied.

The Narolin DB links to other institutional DBs, but the way of linking varies. In Japan, it is a common practice in the field of genome data to publish them in the DNA Data Bank of Japan (DDBJ) (DDBJ, n.d.). Therefore, even in the Narolin DB, genome data is managed with the same metadata as DDBJ by using the custom metadata function, and when it is published, the metadata is downloaded in TSV format and manually uploaded to the DDBJ site. The same applies to the data. On the other hand, metadata of Japanese national projects are obligated to be integrated into GakuNin RDM, and metadata of public data is transferred to GakuNin RDM via API linkage on the primary DB. In the former, the correctness is not checked and left to the researcher's sense of justice. In the latter case, the correctness can be checked because the data has been assigned a Digital Object Identifier (DOI) obtained from a certification authority.

SECURITY

For the use of Narolin DB and an AI supercomputer, an ID was also issued to researchers outside the NARO to use in joint projects with NARO. In principle all functions in the Narolin DB can be utilized, even by users outside the NARO; however, in many cases, the accessible range is limited to the data of the project and consortium. As of April 2021, the external ID was approximately 100. In addition, at least a single internal ID was issued for all employees within NARO, and the total number was approximately 3000.

Although the details are omitted because of security concerns, the security measures for outside use are composed of the following three stages:

Stage 1: MAFF Research Network (MAFFIN) firewall:
- ◦ Monitoring and preventing unauthorized external access to the entire MAFFIN including the NARO.

Stage 2: The NARO firewall:
- ◦ Allow access only from the research base where the user is located.
- ◦ Monitoring and preventing unauthorized access not only from outside, but also from inside.

Stage 3: Security measures in the Narolin DB/supercomputer:
- ◦ Access is allowed only from the research base where the user is located (double setting with 2nd stage).
- ◦ Allow only specific PCs (public key authentication).
- ◦ Introduction of intrusion detection tools.
- ◦ Capture of an audit log that records user operations.
- ◦ Introduction of a function for detecting server firmware tampering and switching to normal firmware, and so forth.

The above three stages are positioned as filters, and security becomes stricter as the stage increases. Note that the access permission setting function can set the data in the Narolin DB to one of the publication restriction levels 1–3 (private, restricted sharing, restricted publishing); however, for level 4 (unrestricted publishing), the authors established a separate server for data publishing with an automatic DOI assignment function, which is placed at the second stage of the security measures.

METADATA SCHEMA

Metadata are indispensable to explain the contents of data. According to the Cabinet Office's Guideline for the Development and Operation of Research Data Repositories (CabinetOffice, 2019), data management systems are required to "Ensuring interoperability of research data and sharing including disclosure of research data by preparing metadata as ancillary information to promote data reuse" in accordance with the FAIR Data Principles (Findable, Accessible, Interoperable, Reusable), which is an international standard for the sharing of research data.

In the Narolin DB, based on metadata for general research data, such as Data Catalog Vocabulary (DCAT, 2020) and DataCite Metadata Schema (DataCite, n.d.), the original NARO Commons metadata schema was prepared by incorporating metadata items for agriculture in the USDA Ag Data Commons, INRAe in France, and ICRISAT. Table 2 lists the metadata items.

In Japan, the Open Access Repository Promotion Association (JPCOAR) (JPCOAR, n.d.) has established a metadata item called the JPCOAR schema to improve the interoperability of metadata and to distribute academic results. The NARO Commons 37 items contain 30 items required by the JPCOAR schema. In addition, information on species and fields necessary for agricultural research and NARO, subject numbers, and so forth was added. Figure 5 shows the relationship between NARO Commons metadata and other schemas.

Additionally, to improve the searchability by facets, the following controlled words are defined based on the results of the interviews in NARO described above. The 10 major categories of data classification include genomics and genetics, plants and crops, agricultural products, food and

Table 2. NARO Commons metadata schema

Metadata items	Description	How to input	Required
Title	data title	free writing	yes
Alternative name	data alias	free writing	
Author Name	creator's name	obtained automatically	yes
Author ID	NARO ID, Google Scholar, ORCID, Researchmap	obtained automatically	
Affiliation	creator's affliation	obtained automatically	
Project	research subject of creator	select from menu	yes
Author Contact address	creator's e-mail	obtained automatically	
Contact person	contact person's name	free writing	yes
Contact person ID	NARO ID, Google Scholar, ORCID, Researchmap	obtained automatically	
Affiliation	affiliation of contact person	obtained automatically	
Project	research subject of contact person	select from menu	yes
Contact address	contact person's e-mail	obtained automatically	
License holder	license holder's name	free writing	yes
License holder ID	NARO ID, Google Scholar, ORCID, Researchmap	obtained automatically in NARO, otherwise free writing	
Affiliation	affiliation of license holder	obtained automatically in NARO, otherwise free writing	
Project	research subject of license holder	select from menu	yes
Contact address	license holder's e-mail	obtained automatically in NARO, otherwise free writing	
Coworker ID in NARO	NARO ID	separate commas in the case of multiple persons	yes
Coworker ID in other insititute	Researchmap ID	separate commas in the case of multiple persons	
Coworker name in other insititute	coworker's name	separate commas in the case of multiple persons	
Subject	data categoriy	select from menu	yes
Keyword	keywords	Controlled keywords	yes
Taxonomy Name	taxonomy name of target species	select one from species, Taxonomy ID or NA	yes
Description	data description	free writing	yes

Figure 5. Relationship of NARO Commons and other metadata schemas
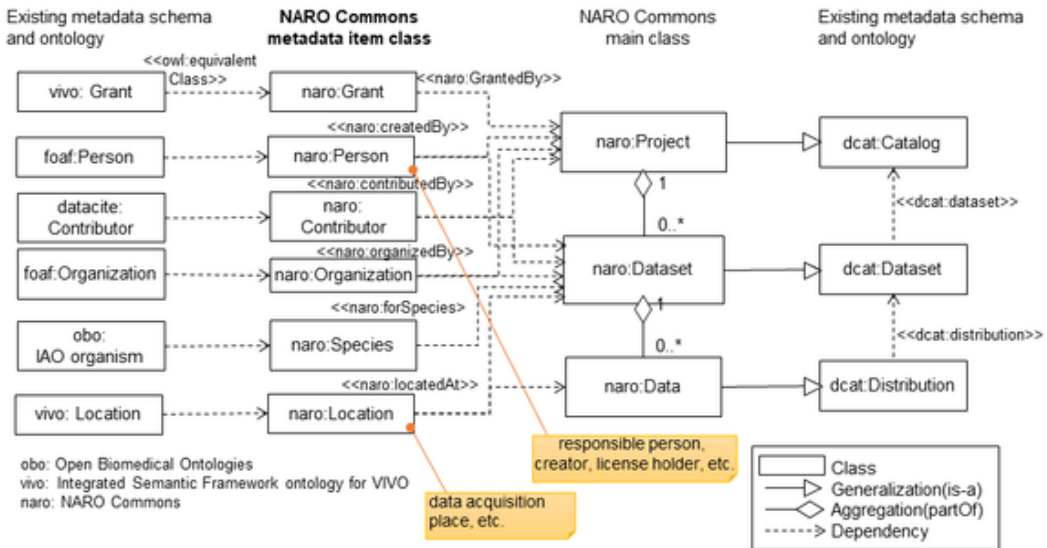
nutrition, agroecosystem and environment, bioenergy, animals and livestock, agroeconomics, maps and multimedia, and online databases:

- Data classification for research fields of major categories 10, middle categories 60, and small categories 106. The data classification was developed for NARO's research field, referring to the Japanese funding agency classification for agriculture and the USDA's field classification.
- Controlled words of 14,604 species. The Japanese version of the NCBI taxonomy is used as a list of biological species.
- Keywords count of 21,016 in Japanese and English. The keywords are based on the Ag Data Commons of USDA.

The metadata entry has a completion function. When a user is halfway through the input, existing names will be listed and the user can select one of them. After entering the name, existing names will be tagged. The entry will be surrounded by a square. This allows the user to know that the name does not exist. Also, to save the effort of metadata input, the metadata of the upper folder/dataset is automatically inherited to the lower folder/dataset and data file. The authors also provide the ability to register metadata by writing it in a Microsoft Excel file and uploading it with the data. This allows metadata to be reused. However, since the NARO Commons was created as a common set of diverse data in NARO, it is possible to add an extension set for each research field. Note that the authors do not edit metadata in a collaborative environment. Metadata is assigned by the data creator or the data registrant.

SUPERCOMPUTER FOR AI APPLICATION

In the introduction of this supercomputer, the required calculation resources were assumed based on the situation of computer resources in NARO, while a supercomputer called AI Bridging Cloud Infrastructure of the National Institute of Advanced Industrial Science and Technology was referred to as an advanced case in Japan. The authors' supercomputer adopts an NVIDIA Tesla V 100, which

excels in matrix operation performance often used in AI calculations for GPUs, and it is equipped with eight GPUs connected by NVLink in a single compute node. A total of 16 compute nodes are connected at InfiniBand 100 Gbps, and a theoretical peak performance of one PFLOPS is achieved with a total of 128 GPUs. This is the first time that a PFLOPS class computer has been introduced into agricultural research in Japan. However, the authors' supercomputer not only aims to boost computing resources, but also aims to offer a data processing environment that an agricultural researcher can easily use. Thus, in addition to the conventional method mainly of command and batch operations, usability, such as PC operations, are offered through the user interface (Jupyter Notebook, R studio, etc.). These can be operated from the web browser and applications can be remotely operated from the user's PC. It also offers virtual machines with the latest machine learning packages preinstalled so that users can take advantage of them immediately.

The supercomputer has a high-performance file system installed to input and output data to be analyzed, and the data can be copied to the primary DB by selecting "Export" from the menu, and vice versa. In the future, each research institute in NARO that reaches over 20 bases in Japan will be connected in order through the Science Information NETwork (SINET) at 10 Gbps, which will increase the convenience of the Narolin DB and the supercomputer.

WAGRI FOR DATA DISTRIBUTION

The Narolin DB has also connected with WAGRI (WAGRI, 2018) for agricultural data distribution mainly for the business domain, which was the result of the Strategic Innovation Program of the Cabinet Office of Japan. WAGRI is a membership-based application programming interface (API) portal site for distributing (selling) agriculture-related data and services through APIs. The architecture of WAGRI is shown in Figure 6. WAGRI is based on the Business-to-Business-to-Customer (B2B2C) model and thus sits between data/service providers and WAGRI users receiving those data/services. WAGRI users, such as agricultural ICT vendors, can develop data-driven agricultural services, for example, for smartphones of farmers, by combining WAGRI APIs. WAGRI provides several data/ services ranging from free-of-charge weather and farmland data to commercial services, such as plant growth prediction, and is currently providing approximately 100 APIs. Table 3 shows the main
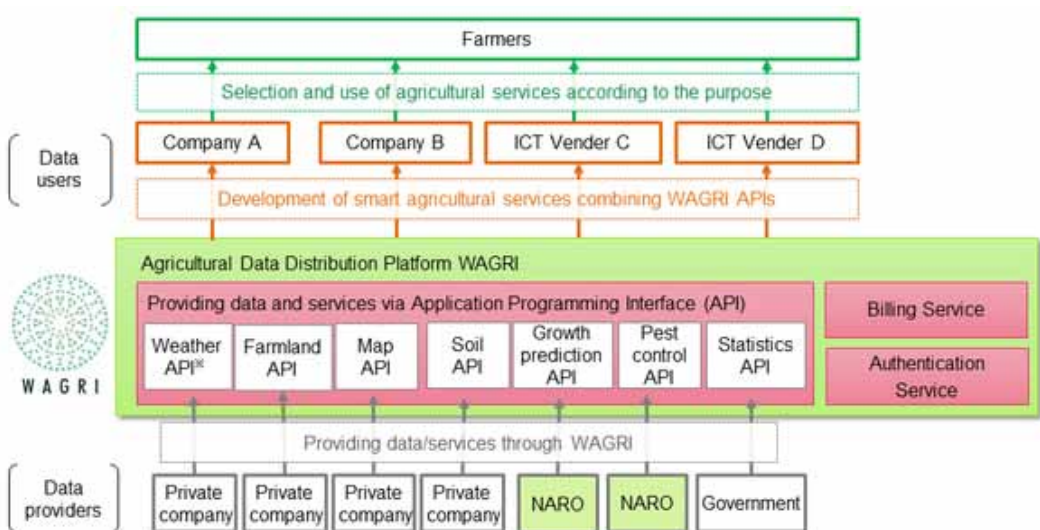
Figure 6. The architecture of WAGRI

Table 3. Parts of the data and services provided by WAGRI

Category	Data/Service	Description
Weather	Hourly weather information up to 3 days ahead (1 km square grid)	Weather information announced by the Japan Meteorological Agency with mesh generation and original correction processing
"	Daily weather information up to 26 days ahead (1 km square grid)	1 km mesh long-term forecast information in which 14 types of fixed, predicted, and average values are seamlessly connected
Map	Map data and aerial photo image data	Spatial information content that boasts the highest level of accuracy and completeness in Japan
Soil	Digital soil map showing soil type and distribution	Information on the type and distribution of soil nationwide
Farmland	Farmland polygon information	Information on about 31 million farms nationwide
"	Latitude-Longitude information for farmland (farmland pin data)	Location, parcel number, land category, area, and area classification of 1,595 of the 1,724 municipalities in Japan
"	Integrated farmland data	Integrated data from farmland plots, farmland pins, and soil data across the country
Fertilizer	Fertilizer registration information	Information on about 20,000 fertilizer brands registered by the Minister of Agriculture
Pesticide	Pesticide registration information	Information on registration of about 7,400 pesticides
Growth prediction	Services provision of growth prediction programs	Rice, wheat, soybeans, open field vegetables, gardening (Tomatoes, paprika, cucumbers, and strawberries)
Pest control	Service provision of disease judgment programs	Program for determining disease from photos taken by smartphone and aggregating the photos for future improvement

data and services provided, such as hourly/daily weather information, map data and aerial photo images, digital soil maps, farmland polygon information, latitude-longitude information, fertilizer information, pesticide information, growth prediction programs, and disease judgment programs. WAGRI currently has more than 50 paying members.

It is built on the Microsoft Azure cloud services and includes databases but often the data are provided through access to the data provider's database. WAGRI is a kind of faucet of data, and Narolin DB acts as one of the dams of data. At present, the open data that are available on the website of the MAFF, and the data of prefecture and public research institute are stored in the Narolin DB, and they are offered by API from WAGRI. When accessing the Narolin DB through WAGRI, the client can double the authentication token and access it through Representational State Transfer (REST), and after passing through WAGRI authentication, it is transferred from the WAGRI API. After passing through external ID authentication of the Narolin DB, the API of the Narolin DB can be executed. The advantage of WAGRI is that both data providers and users can enjoy "one-stop shopping" for agricultural data and related services using unified user authentication and billing services.

DISCUSSION

Table 4 provides a comparison of the five systems introduced in the Related Work: CKAN, DKAN, Dataverse, OSF, and the authors' platform (AIRP). The list of functions is based on the existing comparison table (Comp, 2017) and is based on the list of functions that are considered necessary for an agricultural research dataset management system, as far as the authors know. This list is not necessarily exhaustive, as there are many extension modules available. The AIRP refers not only to the data catalog (the primary DB), but also to the entire platform proposed here. Also, some of the individual services use the systems listed here and are not included in the comparison (e.g.,

Table 4. Comparison of the existing platforms

Categories	Features	CKAN	DKAN	Dataverse	OSF	AIRP
Data Handling	Ability to embargo files (Toggle public / private)	No	Yes	No	Yes	Yes
Data Handling	Handling sensitive data	No	No	No	Yes	Yes
Data Handling	Users are able to control dataset file hierarchy + directory structure	No	No	Yes	Yes	Yes
Data Handling	Users are able to create a symbric link to the other directory	No	No	No	No	Yes
Data Handling	file-level versioning	No	No	Yes	Yes	Yes
Data Handling	Preview of pptx, docx, xlsx, pdf, and images (figures)	Yes	Yes	Yes	Yes	Yes
Data Handling	Preview of zip files	No	Yes	No	Yes	No
Data Handling	Analyzing tabular data with predefined statistical function	No	Yes	Yes	Yes	Yes
Data Handling	Published data is automatically attached with Permanent ID	No	No	No	No	Yes
Data Handling	Handling large data (more than 1 GB)	Yes	Yes	Yes	Yes	Yes
Data Handling	Automatic anti-virus scan of uploaded files	No	No	No	No	No
Data Handling	Source code management	No	No	No	Yes	No
Data Handling	Ability to sell data, i.e., Billing function	No	No	No	No	Yes
Search	Easily accessible Help page (not pdf document)	No	No	No	Yes	Yes
Search	Faceted search based on metadata	Yes	Yes	Yes	Yes	Yes
Search	Full-text search for uploaded text data	No	No	No	No	Yes
Metadata	Annotaion of metadata from external file (import)	No	No	Yes	Yes	Yes
Metadata	Exporting metadata to external files	No	No	Yes	Yes	Yes
Metadata	Metadata customization for domain-specific needs	Yes	Yes	Yes	Yes	Yes
Metadata	Support for controlled vocabulary terms in metadata	No	Yes	Yes	No	Yes
Metadata	Mapping of Geospatial files in metadata	Yes	Yes	Yes	No	Yes
User Account Management	Access control of each data file for each user	No	Yes	Yes	Yes	Yes (Fine-grained control is possible according to organizational and subject structures)
User Account Management	Do users have to actively acknowledge any conditions of use?	No	No	Yes	No	Yes
User Account Management	ID federation service (Shibboleth, Google, Open ID, Active Directory, etc.)	Yes	No	Yes	Yes	Yes
External Linkage	Are there APIs?	Yes	Yes	Yes	Yes	Yes
External Linkage	Migration of data files and metadata to external DBs	No	No	No	Yes	Yes
External Linkage	Migration of data files to computational resourse with id federation	No	No	No	No	Yes
External Linkage	External storage mount (Dropbox, etc.)	Yes	No	Yes	Yes	Yes
External Linkage	Embed data to web pages and other application	Yes	Yes	Yes	Yes	No
External Linkage	Data citation automatically generated	No	No	Yes	Yes	No
External Linkage	Link to workflow mgt. system	No	No	No	Yes	No
Metrics + Reports	Citations tracking	No	No	No	Yes	No
Metrics + Reports	Show reports of download#, access#, etc. (Dashboard)	No	No	No	No	Yes
Metrics + Reports	Are users emailed about changes in their datasets?	No	No	Yes	Yes	Yes
Customization	Ability to customize the interface (My Page, etc.)	No	Yes	Yes	No	Yes
Customization	User communinity support via Group, BBS, e-mail, etc.	No	Yes	No	Yes	Yes
	Total Yes	9	14	20	25	29

GakuNin RDM uses OSF). In addition, the authors did not describe any features in this paper where a dedicated explanation was not necessary, such as features that are obvious from their names. It can be seen that the authors' AIRP provides the most complete set of features. However, they have also identified features that need to be added in the future, such as embedding data to web pages and other applications, data citation automatically generated, and citation tracking.

In addition, the followings are a few examples of data-driven agricultural services that have been deployed on the authors' data platform. One is an image-based AI service. In the disease and

pest determination service (NARO n.d.), farmers can take photos of disease and pest marks with their smartphones and send them to the authors. The service tells farmers what the cause is and what pesticides to prescribe. The WAGRI APIs are used as the front-end of the service, and a huge amount of image data is stored in the primary DB, while a disease assessment engine trained on the AI supercomputer runs on a web server in the secondary DB. In the growth management service using drones, Normalized Difference Vegetation Index (NDVI) values are calculated from images taken from above by a drone, and the service grasps the growth status of each area of the field to apply fertilizers and pesticides appropriately. Here, a large number of images are also collected in the Narolin DB, shared, and analyzed by the AI supercomputer, and the service is provided as APIs through WAGRI. There is also a calving prediction service based on cow images. This way, the authors' platform has made it relatively easy to develop and apply the same type of service, which in turn has promoted the use of data. Presently, they are all being developed nationwide in Japan.

The disease and pest identification service used here as an example is performed locally by the user using a smartphone and requires a fast response time (latency). On the other hand, drone-based growth management services require large-volume transmission, as a single drone image can be several hundred MB in size. These are the two main performance indicators that the authors are working to improve. For latency, the API processing function has been improved, and the number of data processed has been increased from 10 cases/minute to 300 cases/minute. For latency, the API processing function has been improved, and the number of data processed has been increased from 10 to 300 per minute. For transfer speed, writing to storage has been improved so that 300 images at a time, totaling 5 GB of aerial photography data, can be stored within one hour. These are all needs that have been identified in the process of using the service on site. Qualitatively, budgeting for computing resources (CPUs and GPUs) to increase speed will improve latency because API processing speed will increase. On the other hand, using more expensive storage devices (SSDs, HDDs, tapes) as storage resources will enable large-volume transfers, so the budget ratio should be considered according to performance indicators. The authors will continue to make improvements as the use of the service expands.

As mentioned in the secondary DB section, concerning the linkage between functional components and genomes, the creation of a lactobacillus database can be specifically mentioned. Among microorganisms, lactic acid bacteria are biological materials that are expected to be applied to healthy foods and pharmaceuticals in recent years. Therefore, the authors constructed a graph DB on the secondary DB that interconnects different types of information stored on the primary DB, such as functional components, genomic information, and phenotypic information (Genebank Project, n.d.). In response to this, we are providing data to companies following the open-close strategy via WAGRI, and promoting pharmaceutical development using the AI supercomputer.

On the other hand, the Agricultural Machinery Open API (API, 2021) is not an AI service, but a data exchange service. It unifies the formats of data obtained from agricultural machinery of various manufacturers, stores them in a DB via WAGRI, and provides them from the DB via WAGRI as needed. This unification of data from agricultural machinery manufacturers will make it easier for agricultural ICT vendors' cultivation management support systems to cope with the data. As a result, farmers with multiple farm machines from different manufacturers can now link their farm machines to the cultivation management support system they are using, which has facilitated data-driven farming. Although it is difficult to quantitatively evaluate the contribution of the platform in each case, it is clear that the data platform has helped in the rapid deployment of services by eliminating the need to develop systems for each service.

Regarding Agricultural Machinery Open API, four domestic and foreign manufacturers (Iseki, Kubota, Mitsubishi Mahindra, and Yanmar) have already announced in a press release that they will maintain mutual compatibility. In the future, the Japanese government plans to take measures such as subsidies for proposals that do not use machinery compliant with this standard. While the ISO 11783 standard already exists for large tractors, this standard is for small and medium-sized tractors, combine

harvesters, rice trans-planters, dryers, and horticultural equipment, and is mainly intended for data exchange with domestic and foreign cultivation management support systems (Kubota KSAS, BASF Xarvio). This will allow farmers to easily consolidate and transfer data at any time, regardless of the manufacturer of the equipment they own, or the cultivation management support system they use.

Additionally, standardization of agricultural data itself is currently underway with the IEEE Standard Association P2992 Working Group—Data Expression, Exchange, and Processing in Smart Agriculture starting in the Fall of 2021, and the Strategic Advisory Group on Smart Farming will be formed in ISO in the Spring of 2022, with Subgroup 8 discussing Data and Subgroup 6 Terminology and Semantics. These results will also be implemented in the AIRP.

CONCLUSION AND FUTURE WORK

This paper describes the agricultural information research platform, in which the NARO started operations in FY 2020 to properly share research data, along with the open and close strategies and the promotion of data-driven agriculture. By setting the backup destination of the data file to the Narolin DB, the researcher is released from the trouble and duty (10-year storage of the research data) of data management, and the data can be stored safely and permanently.

In the future, functions to be installed in the Narolin DB include a single sign-on (SSO) mechanism for domestic researchers by a unified identity provider and a metadata linkage function with external research data catalogs. Additionally, since external access without user registration is not permitted, a metadata crawl by an external retrieval service is not permitted; however, the server for open data permits retrieval by an external service such as Google.

However, it is still a large problem to get researchers to take part in the registration and sharing research data and to increase their active use. It is well known that data utilization does not advance easily by just preparing the infrastructure of the research DB. In the NARO, thus, as a policy of systematic and institutional aspects, the NARO version of the research DB operation guideline was formulated in FY 2019, in accordance with the Cabinet Office's Research Data Repository Development and Operation Guideline described earlier. In this guideline, it is specified that research data, as research results, should be shared in principle within NARO. Even if access is restricted on the DB, it can be obtained with the prescribed procedure. In the meantime, the standard terms of use for research data were also determined, the ownership of researchers who created the data was clarified, and the rights in the paper and patent proposals were specified. In addition, the research result management guideline determined the registration of research records, including research data, to the Narolin DB, and also the periodic confirmation by the responsible person from FY 2021.

In parallel with these measures, a hands-on AI education course on using the supercomputer and Narolin DB began in NARO in FY 2020. Classes on data science and AI, especially machine learning, equivalent to 15 semi-annual classes in universities, will be conducted through preparatory, beginner, and intermediate courses based on agricultural research themes. These courses will be offered outside NARO, within public research institutes in the prefecture from FY 2021, to spread the approach of open science and data-driven agriculture. In addition, the authors examined the mechanism and business model of agricultural data utilization in industries, universities, and governments, and cooperated with other domestic and overseas organizations, such as the Research Data Alliance (RDA), on research data.

To promote data utilization in the agricultural sector, it is important to simultaneously implement not only the preparation of data platform, but also systematic and institutional design. As a result, these will expand the agricultural researches through data and AI, and lead to the dissemination of data-driven agriculture.

FUNDING AGENCY

REFERENCES

API. (2021). To Realize agricultural data linkage beyond the system wall. Agricultural Machinery API Standardization Consortium. https://www.naro.affrc.go.jp/org/brain/iam/API/index.html

CabinetOffice. (2019). Guidelines for the Development and Operation of Research Data Repositories. Working Group on the Promotion of Open Science in Light of International Trends. Cabinet Office. https://www8.cao.go.jp/cstp/tyousakai/kokusaiopen/guideline. pdf

ckan. (n.d.). Comprehensive Knowledge Archive Network. ckan. https://ckan.org/

COS. (n.d.). Manage and share your research with OSF – an easy, integrated platform. COS Center for Open Science. https://www.cos.io/products/osf

Data Terra. (n.d.). The Data Terra Research Infrastructure. https://www.data-terra.org/ en/homepage-english/

DataCite. (n.d.). Welcome to DataCite. DataCite. https://datacite.org/

Dataverse. (2017, July 25). A Comparative Review of Various Data Repositories. Dataverse Project. https://dataverse.org/blog/comparative-review-various-data-repositories

DCAT. (2020). Data Catalog Vocabulary (DCAT) - Version 2. W3C Recommendation. https://www.w3.org/TR/vocab-dcat-2/

DDBJ. (n.d.). About DDBJ Center. DDBJ: DNA Data Bank of Japan. https://www.ddbj.nig.ac.jp/ index-e.html

DKAN DOCS. (n.d.). Comparing DKAN and CKAN. DKAN DOCS. https://dkan.readthedocs.io/ en/latest/introduction/dkan-ckan.html

dkan. (n.d.). DKAN Open Data Platform. https://getdkan.org/

EOSC. (n.d.). Welcome to the EOSC Portal Catalogue and Marketplace. European Open Science Cloud. https://marketplace.eosc-portal.eu/services

FAMIC. (n.d.). Home. Food and Agricultural Materials Inspection Center. http://www.famic.go.jp/english

Foster, E. D., & Deardorff, A. (2017). Open Science Framework (OSF). Journal of the Medical Library Association: JMLA, 105(2) https://doi.org/10.5195/jmla.2017.88

Genebank Project. (n.d.). About Genebank Project, NARO. https://www.gene.affrc.go.jp/about-micro_en.php

Hayashi, K. (2018). Open Science in Integrated Innovation Strategies. STI Horizon, 4(3), 42–47. http://www.nistep.go.jp/wp/wp-content/uploads/NISTEP- STIH4-3-00145.pdf

J-STAGE. (n.d.b). J-STAGE networks worldwide. J-STAGE. https://www.jstage.jst.go.jp/browse/ -char/en

J-STAGE Data. (n.d.a). Discover Research from J-Stage Data. J-STAGE Data. https://jstagedata.jst.go.jp/

JPCOAR. (n.d.). The Japan Consortium for Open Access Repository. JPCOAR. https://jpcoar.repo. nii.ac.jp/?page=1&size=20&sort=controlnumber&search_type=0&q=0

Komiyama, Y., & Yamaji, K. (2017, July). Nationwide research data management service of Japan in the open science era. In 2017 6th IIAI International Congress on Advanced Applied *Informatics IIAI-AAI)* (pp. 129–133). IEEE. doi:10.1109/IIAI-AAI.2017.144

NARO. (n.d.). *Launch of AI-based pest image diagnostic system via WAGRI*. NARO National Agriculture and Food Research Organization. https://www.naro.go.jp/english/laboratory/ rcait/press/aipestdiagnosis/index.html

Neveu, P., Tireau, A., Hilgert, N., Nègre, V., Mineau-Cesari, J., Brichet, N., Chapuis, R., Sanchez, I., Pommier, C., Charnomordic, B., Tardieu, F., & Cabrera-Bosquet, L. (2019). Dealing with multi-source and multi-scale information in plant phenomics: The ontology-driven Phenotyping Hybrid Information System. *The New Phytologist*, *221*(1), 588–601. doi:10.1111/nph.15385 PMID:30152011

OpenSILEX. (n.d.). Welcome to OpenSILEX, an Open Source information system. http://www.opensilex.org/

RCOS. (n.d.). GakuNin RDM (Research Data Management Platform). Research Center for Open Science and Data Platform. https://rcos.nii.ac.jp/en/service/ rdm/

RO-Crate. (n.d.). Research Object Crate. RO-Crate. https://www.researchobject.org/ro-crate/</eref>

Soiland-Reyes, Stian & Sefton, Peter & Crosas, Merce & Castro, Leyla & Coppens, Frederik & Fernández, José & Garijo, Daniel & Grüning, Björn & Rosa, Marco & Leo, Simone & Carragáin, Eoghan & Portier, Marc & Trisovic, Ana & Community, RO-Crate & Groth, Paul & Goble, Carole. (2022). Packaging research artefacts with RO-Crate. Data Science. 1-42. 10.3233/DS-210053.

Tellus. (n.d.). Satellite data platform in Japan. Tellus. https://www.tellusxdp.com/en-us/

The Dataverse Project. (n.d.). Open source research data repository software. The Dataverse Project. https://dataverse.org/

Trisovic, A., Durbin, P., Schlatter, T., Durand, G., Barbosa, S., Brooke, D., & Crosas, M. (2020). Advancing Computational Reproducibility in the Dataverse Data Repository Platform. *Proceedings of the 3rd International Workshop on Practical Reproducible Evaluation of Computer Systems*, 15–20. doi:10.1145/3391800.3398173

WAGRI. (2018, April). About WAGRI: Agricultural Data Collaboration Platform. WAGRI. https://wagri.net/en-us/

WEKO3. (n.d.). WEKO3 (Publication Platform). Research Center for Open Science and Data Platform (RCOS): https://rcos.nii.ac.jp/en/service/weko3/