**Slide 1 | Trends in Open Science in the Humanities**

*[00:00-00:30]*

Hello everyone.

My name is Makoto Asaoka from the Research Center for Open Science | and Data Platform at the National Institute of Informatics (NII).

I work on the design and operation of infrastructure for sharing research data.

Today / I will talk about recent trends in sharing research data in the humanities / based on efforts in Japan | and abroad.

Please note: the views I share today are my personal opinions / not the official views of NII.

**Slide 2 | Today's Outline**

*[00:37-02:00]*

Here is today's plan.

First / / I will briefly (ブリーフリィ) explain open-science trends across academia | and when data sharing is expected in the humanities.

Next / / we will look at the research infrastructure that supports data sharing in the humanities.

We will see European | and American infrastructures / and also Japanese infrastructure.

Then / I will organize the main topic: how to share research data.

I will explain that a DMP (Data Management Plan) is very important today.

You should think about data sharing from the planning stage.

You also need to create metadata that records the context of the data.

And you / the researcher / must finally decide whether your data can be shared (シェアード).

At the end / from the viewpoint of infrastructure design / I will share a short outlook for research data sharing.

**Slide 3 | What is Open Science (OS)?**

*[02:00-03:50]*

Here / I will confirm what open science is / following the elements listed by UNESCO.

First / / open access to scientific knowledge--such as open articles | and research data.

These are part of open science.

There are three other elements.

Second / / research infrastructure.

Only when we have repositories (archives) / data management systems / discovery databases / | and computing systems / can shared knowledge actually be used.

Today's talk focuses (フォーカスィズ) on sharing research outputs | and the infrastructure that supports them.

Third / / engagement with social actors--for example / citizen participation (パーティシ**ペイ**ション) | and links with policy makers or industry.

Finally / / dialog with other knowledge systems--for example / indigenous (イン**ディ**ジェナス) or local traditional knowledge.

The humanities are closely related to this.

Open science is a practice to make research from all fields available for use / for the benefit of scientists | and society.

Based on this definition / I will now move to changes in data sharing | and infrastructure.

**Slide 4 | How the Data-Sharing Environment Has Changed**

*[03:50-05:30]*

Let us confirm the changes around data sharing.

The key point is the split "before IT" and "after IT."

Before IT: representative (レプレ**ゼン**タティブ) datasets in each field were protected as public goods of the research community.

Only some datasets for international comparison (コン**パ**リソン) were shared.

Shared datasets were mainly representative (レプレ**ゼン**タティブ) ones / | and many researchers felt that datasets belonged (ビ**ロングド**) to individual researchers or research groups.

After IT: we can share across regions and fields.

Long-term preservation of digitized (**デ**ジタイズド) cultural resources | and download distribution (ディストリ**ビュ**ーション) became practical.

Individual researchers can use datasets from other regions and fields.

At the same time / data sharing led by publishers and funders has spread.

Publishers clarified (**クラ**リファイド) peer-review | and openness requirements (リ**クワイア**メンツ) for the papers they publish.

Governments ask for openness to justify the use of public funds.

As a result / sharing research data is becoming "standard practice." Next / I will look at journal recommendations (レコメン**デー**ションズ) | and funder mandates (**マン**デイツ).

**Slide 5 | Journals Mentioning Data Sharing (France)**

*[05:30-06:30]*

This chart shows that more journals encourage data sharing.

Some journals recommend or require authors to state how to access the data that support their papers.

This statement is called a Data Availability Statement (DAS).

The French Open Science Monitor publishes (**パブ**リッシズ) the trend in the share of publications including a DAS.

As the chart shows / this share has increased year by year; in 2023 (two thousand twenty-three) / 37 (thirty-seven)% (thirty-seven percent) of journals adopted DAS.

**Slide 6| Humanities Journals with DAS Are Increasing**

*[06:30-07:00]*

This chart shows / by field in 2023 (two thousand twenty-three) / the share of publications including a DAS.

In the humanities / 22% (twenty-two percent) of journals adopt DAS.

Across disciplines (**ディシ**プリンズ) / the practice of clearly stating where the supporting data are is spreading.

Slide 7 (seven) | Data Availability Statement (DAS)

*[07:00-07:50]*

This slide shows an example of a DAS in Humanities | and Social Sciences Communications (Nature Portfolio).

A DAS is written before the references | and shows where the supporting data can be accessed (**アク**セスド).

In this example / the author uploaded (アップ**ローデ**ッド) data to their GitHub.

In other cases / authors use an institutional or a disciplinary (**ディ**シプリナリー) repository.

Note that some authors share data on request / or explain that they cannot provide data due to rights issues.

**Slide 8 | Immediate Open Access for Publicly Funded Research**

*[07:50-09:50]*

Next / / I will look at research outputs funded by public money.

In short / / mandates (**マン**デイツ) are growing not only in Europe and the U.S.

but also in Japan.

The EU announced the principles (**プリン**シプルズ) of Plan S in September 2018 (two thousand eighteen).

It requires (リ**クワイ**アーズ) that publicly funded research articles be open immediately (イ**ミ**ディアットリー) after publication.

It has been implemented (**イン**プリメンテッド) since 2021 (two thousand twenty-one) for research funded by 11 (eleven) national agencies in Europe.

In the U.S. / the OSTP announced in August 2022 (two thousand twenty-two) that immediate open access is required.

This includes not only articles / but also the supporting data.

Implementation is from fiscal (フィスカル) year 2025 (two thousand twenty-five) .

In Japan / / the Cabinet Office announced a similar (シミラー) basic policy in February 2024 (two thousand twenty-four).

From fiscal (フィスカル) year 2025 (two thousand twenty-five) / for new competitive research funds / peer-reviewed articles | and their supporting data must be deposited (ディポジテッド) immediately (イ**ミ**ディアットリー) after publication in institutional repositories / etc (/ エトセトラ).

Here / "supporting data" means the research data that the target journal's author guidelines require to be made public.

If the guidelines do not require "supporting data / " there is no mandate to open that data.

So / far / I have shown that data sharing is progressing / | and humanities researchers cannot ignore it.

Next / / I will introduce the infrastructures that support research activities / including data sharing.

**Slide 9 | Research Infrastructure for the Humanities in Europe/U.S.**

*[09:50-11:00]*

First / / I will explain the humanities infrastructure in Europe/U.S..

Two pillars (ピラーズ) support data sharing: a format for contextual (コンテクスチュアル) information (metadata) | and a platform to host data and metadata ("repositories").

Humanities data have complex context: how they were created / how to interpret them / | and how they relate to other data.

To compare and search across projects / we need standardized (/ スタンダーダイズド) description.

For this reason: TEI is used for text data; IIIF for image distribution (ディストリ**ビュ**ーション); CIDOC-CRMnships in cultural heritage; | and schema.org to improve web discoverability (ディスカバラ**ビリ**ティ).

For hosting | and long-term preservation / data plus metadata are shared (シェアード) in repositories.

Examples include Europeana / Open Context / and DPLA.

They are designed for cross-institution (**クロス**・インスティ**テュ**ーション) use via APIs | and platform interoperation (インタロパ**レー**ション).

There are three challenges: (1) rights clearance / (2) multilingual (マルチ**リン**ガル) support / | and (3) sustainable operations.

These become heavier (**ヘヴ**ィアー) at larger scale.

Keeping this in mind / I will next explain Japan's infrastructure.

**Slide 10 | Japan's Social Science Research Data Infrastructure**

*[10:10-11:20]*

In Japan / / data sharing is led by infrastructures run by each institution.

But standardization (スタンダーダイ**ゼー**ション) of metadata | and connections across institutions are behind / so it is hard to find and reuse data.

As a result / opportunities (オポ**チュ**ニティーズ) for collaboration are sometimes lost / | and there has been concern (コン**サーン**) about delays (ディ**レイズ**) in humanities and social sciences infrastructure.

Therefore / / JSPS implemented (インプリメンテッド) a five-year project (FY2018 (fiscal (フィスカル) year two thousand eighteen) (fiscal year two thousand eighteen)-FY2023 (fiscal year two thousand twenty-three)) to build Humanities | and Social Sciences Data Infrastructure.

It standardized (スタンダーダイズド) datasets from each institution with common metadata.

From October 2023 (two thousand twenty-three) / a follow-up enhancement project started to strengthen (ストレングセン) the infrastructure.

Next / / I will explain the infrastructure built in this project.

**Slide 11 | What Infrastructure Do HSS Researchers Need?**

*[11:20-12:30]*

NII built the infrastructure for this project with three concepts.

First / / a cross-search platform.

We built a platform that can connect with repositories in each field using standard HSS metadata.

Second / / web-based execution (エグゼ**キュー**ション) environments (エン**ヴァイ**ロンメンツ).

In classes | and joint research / sharing the same notebook and steps is important.

We prepared environments (エン**ヴァイ**ロンメンツ) with sample data and notebooks.

Third / / a repository for HSS.

NII had developed WEKO3 / a general data repository system.

We added (**アデ**ッド) support for DDI (Data Documentation Initiative) for social sciences | and for access control on research data.

**Slide 12| Outcomes of the Data Infrastructure Project**

*[12:30-13:40]*

This diagram shows the ecosystem we built based on the three concepts.

Researchers can (1) search -> (2) access -> (3) analyze data in a seamless flow.

Humanities researchers have used CiNii to search articles.

Now they can use JDCat / an integrated (**イン**テグレイテッド) cross-search system made in this project / to search across repositories.

Metadata from institutional repositories is aggregated (**アグ**リゲイテッド) / | so you can search by title / creator / period / region / etc ( エトセトラ).

Data found by the search can be accessed from each repository.

It is linked with the online analysis system built in the project.

Using Jupyter Hub / you can reuse data together with code directly online.

We also prepared guidelines for providing data to disciplinary (**ディ**シプリナリー) repositories / | so that researchers who want to share data have clear steps.

**Slide 13 | How Do We Share Research Data?**

*[13:40-16:10]*

So / far I explained (エクス**プレ**インド) infrastructure.

From here / I will introduce how to share data in the humanities based on the project guidelines.

**Slide 14 | How Do We Share Research Data? (Guide)**

*[about 1 minute]*

I will introduce data sharing in the humanities based on the HSS data sharing guide.

The ALLEA ( アレア) report also provides a humanities-oriented guide.

ALLEA is the European Federation of Academies of Sciences and Humanities.

You can access these documents from these QR code.

Both documents emphasize the importance of DMP-based data sharing.

So / / I will briefly (**ブリーフ**リィ) introduce DMP next.

**Slide 15 | Data Management Plan (DMP)**

*[21:50-23:10]*

A DMP is a plan that describes (ディ**スクラ**イブズ) how you will collect / manage / preserve / | and share data.

It is not a final report; it should be updated as the project progresses.

Some funders require a DMP at the start.

In Japan / / from FY2024 (fiscal (フィスカル) year two thousand twenty-four) (fiscal year two thousand twenty-four) / JSPS requires (リ**クワイ**アーズ) PIs (ピー・アイズ) of adopted projects to prepare a DMP.

A DMP includes: formats / creators | and managers / whether the data include sensitive information / and the sharing policy.

As the project goes on / you update the DMP and share data according to it.

**Slide 16 | Steps for Data Sharing (Plan -> Collect/Create)**

*[23:10-24:50]*

Now I will explain the concrete steps to publish research data, in line with the progress of a research project.

First is the planning stage.

Before the research starts, plan **what data you will produce**.

Decide whether it will be **primary sources** or **derived data** made from primary sources.

Decide whether it will be **text or images**, and **which file formats** you will use for storage.

Next, check the **legal and ethical constraints (コンストレインツ)** on the data.

Review copyright / portrait rights / cultural property law (**カル**チュラル・**プロ**パティ・ロー) / etc (/ エトセトラ).

In the humanities / cultural heritage and personal data often become bottlenecks.

Write these points in your DMP.

If the planned outputs change/ as the project goes on, update the DMP each time.

If plans change / update the DMP.

Next is the collection/creation stage.Collect/Create stage: decide storage formats for data.

As the project progresses, you collect and create data.

Decide in advance the formats for storing the data.

Even if you use other formats during processing and analysis, for stored files prefer sustainable open formats such as CSV, TIFF, and WAV.

Also decide the metadata schema that will describe the context of your data, and create metadata as you collect and produce the data.

Library-style bibliographic schemas such as Dublin Core and JPCOAR exist, but there are other schemas that describe humanities data in more detail.

As I will explain shortly: TEI is common for text, and IIIF for high-resolution images, and so on.

I will explain metadata in the next slides.

**Slide 17 | Metadata Required by Public Funders**

*[24:50-26:30]*

From FY2025 (fiscal ( フィスカル) year two thousand twenty-five) (fiscal year two thousand twenty-five) / immediate open access starts.

In principle / supporting data for publicly funded research must be open.

Even when data cannot be shared (シェアード) / metadata should be open.

This table shows the actual items to publish: project/funding information / dataset title / description / field / size / where it is shared (シェアード) / creators / managers / etc (エトセトラ). These map to items you can fill from the DMP.

But note: such metadata only describe the dataset; they do not capture humanities-specific context.

Next / / I will introduce metadata used in the humanities.

Slide 18 (eighteen) | Metadata in the Humanities

*[26:30-27:50]*

This table lists major metadata used in the humanities.

Roughly two groups: Dublin Core / schema.org / JPCOAR are general-purpose schemas.

The other five are specialized for the humanities.

Dublin Core : a minimal set of elements.

It is widely used as a common vocabulary in institutional repositories | and OAI-PMH.

schema.org is for web search.

If you use JSON-LD / Google | and other engines can discover your records more easily.

It can describe more details than Dublin Core.

JPCOAR is used in institutional repositories at Japanese universities.

It is compatible with Dublin Core | and DataCite / and metadata flows to CiNii Research.

Next / / the humanities-specific ones: TEI is for text.

CIDOC-CRM (サイドック・シー・アール・エム) is a model for museums and cultural heritage.

IIIF is an image distribution (ディストリ**ビュ**ーション) API rather than a metadata schema.

I will explain these three in more detail.

EAD is for archival (アー**カイ**ヴァル) finding aids / keeping hierarchical context (fonds -> series -> file -> item).

MODS/METS are often used together in libraries: MODS for rich bibliographic description; METS for packaging | and structural maps (which file is which page).

They help organize and preserve complex digital resources.

Next / / TEI in detail.

Slide 19 (nineteen) | TEI: Standard for Text Data

*[27:50-29:10]*

TEI (Text Encoding Initiative) makes the structure / annotations / | and variant readings of texts machine-readable.

It adds background information to text data; it is not only bibliographic.

You can mark logical structure / annotations / editing notes / person | and place names with IDs / translations / and so on.

TEI uses XML and is adopted by an international humanities community.

Next / / I will explain CIDOC-CRM (サイドック・シー・アール・エム).

**Slide 20| CIDOC-CRM (SEE-dock C-R-M / サイドック・シー・アール・エム（CIDOC-CRM）) (Cultural Heritage)**

*[29:10-30:50]*

CIDOC-CRM (サイドック・シー・アール・エム) is a model to describe cultural heritage.

Strictly speaking / it is not a metadata schema.

It aims to clarify semantic connections across museums and disciplines.

It describes (ディ**スクライブ**ズ) events around an object: production / acquisition (アクイ**ジシ**ョン) / exhibition (エグジ**ビ**ション) / documentation / etc (/ エトセトラ).

Each event links who (Actor) / when (Time-Span) / where (Place) / | and what.

By linking persons / places / | and objects with IDs / museums can publish data on the web and connect cultural heritage across institutions.

Finally / / I will explain IIIF.

**Slide 21 | IIIF (Images)**

*[30:50-32:20]*

IIIF is a set of APIs | so that high-resolution images can be delivered / viewed / and shared (シェアード) in the same way anywhere.

It is also not a pure metadata schema.

When using images for research / sometimes you want to zoom only a part.

With IIIF-compatible viewers / you can zoom / crop / | and compare images side by side.

IIIF uses JSON-LD as a machine-readable format.

In Japan / / the National Diet Library / the National Museum of Japanese History / | and universities like Kyoto University use it.

That is all for metadata.

Next / / I will explain how to decide whether to share data after research ends.

**Slide 22 | After Research: Deciding on Sharing**

*[32:20-33:40]*

This diagram is a flowchart from the Guideline for Research Data Publication | and Use Condition Display by the Research Data Utilization Forum, RDUF.

Please use this flowchart when you consider data sharing.

When you plan to share data / first check whether you need to share it.

Then / / check constraints (コン**ストレイ**ンツ) based on how you obtained the data.

In the next slide / I will look at Q1 Identify the data to publish in detail.

**Slide 23 | Q1 Identify target data**

*[32:20-33:40]*

Here I will briefly explain Q1: identify target data.

The target for data publising is digital data: numbers, text, images, audio, and video.

First, sort your data into three groups.

1) Supporting data.

These data used as evidence for the article, or data that is the result itself.

From the viewpoint of verification, this group has a high priority for publish.

2) Primary data.

These data are original data from observation or collection.

Make clear whether you hold it or a third party holds it.

If a third party holds it, check the location and permission.

3) Derived data.

These Data made by processing source data—for example, cleaning, transcription, annotation, or aggregation.

Always record the citation method, rights ownership, and version information.

The following objects are not publishing data.

metadata itself, research notes or diaries, physical specimens/objects,

copyrighted works (articles, books, secondary works), or the execution environment (databases, analysis software, etc.).

There are times when data must be published.

A funder, publisher, or your institution may require release of supporting data.

If you are unsure, consult legal/IP/ethics experts.

That is all for Q1.

In the next step, we check the constraints that might block release.

**Slide 24 | Check constraints**

*[33:40-34:50]*

In the next step, Q2, we decide whether the data can be made public.

We judge this from the nature of the data.

For humanities data, publishing can be difficult.

Reasons include cultural property protection and copyright.

Such data cannot be released, even if a funder or a publisher asks for publishing.

There are also cases where release is hard because of personal data protection.

In the next step, Q3, we consider how to handle constraints.

Even when there are constraints, sharing may still be possible by:

setting an embargo period before release,

doing anonymization or offering condition access.

If data publishing is impossible, we should:

document the decision process, publish metadata only, and

store the original data properly.

If you are not sure how to deal with a constraint, please consult legal/IP/ethics expert

**Slide 25 | Select a repository**

*[34:50-36:10]*

After checking constraints (コン**ストレインツ**) / select the repository.

There are three types: disciplinary (**ディシ**プリナリー) / institutional / | and general repository.

Disciplinary (**ディシ**プリナリー) repositories often provide rich functions for reuse.

They can support TEI or IIIF / and make citation and reuse easier.

But they may accept only specific formats / | and some services are experimental / so long-term stability may be uncertain.

Institutional repositories are run by university libraries.

If you belong to the institution / you can usually share many kinds of data.

Basic metadata (e.g. / JPCOAR) are added (**アデ**ッド) / and persistence is expected.

But metadata are limited to general schemas / | and only a few institutional repositories support restricted access on request.

General-purpose repositories (e.g. / Figshare / Zenodo / OSF; also GitHub for code) let researchers publish data quickly | and set access controls themselves.

They are convenient / | but metadata are limited to general fields; sometimes they are seen as personal file storage / and their brand as data repositories may be weaker.

They are usually free now / but may change to paid or stop services.

**Slide 26 | Specify terms of use**

*[36:10-37:20]*

Finally / / licenses (/ ライセンスズ).

Three points:

First / / confirm that you have the right to license the data.

For images of manuscripts or paintings / you may need to confirm with the holder.

Do not add new licenses (/ ライセンスズ) to public-domain material.

If third-party content is mixed / separate it under different conditions or exclude it.

For datasets mixing text | and images / you may need different conditions for images.

Second / / choose a license with as much freedom as possible.

For research data / CC BY 4.0 (four point zero) is recommended.

CC BY-NC can block use because "commercial" is vague.

CC BY-ND can block translation / correction / | and excerpting / and may hinder applying TEI.

Third / / record the license URI in metadata so machines can read it.

Many repository systems handle URLs / | but be careful if you publish on your own site.

**Slide 27 | Data Sharing FAQ (Summary)**

*[37:20-38:50]*

Let me summarize three points.

Where to share: disciplinary/ institutional / general-purpose.

A disciplinary repository is ideal for reuse / but may not accept your format.

For persistence / an institutional repository is good; if you need speed or restricted access / consider a general repository.

What to do: plan research from the start with the DMP in mind.

Even if plans change / if you document what data you will collect/create | and what rights apply / you can prepare for sharing.

How far to open: currently / only supporting data for peer-reviewed articles are mandated.

Do not open data that should not be open.

Use an open-and-closed strategy.

In the humanities / some data--such as indigenous (イン**ディ**ジェナス) or religious materials--are not personal | but are culturally sensitive.

For such data / build consensus with stakeholders before sharing.

Slide 28 (twenty-seven) | Outlook for Data Sharing

*[38:50-40:40]*

Finally / / the outlook.

Top-down decisions are building the environment.

Funders mandate openness for papers | and supporting data; publishers make DAS and data sharing standard.

Infrastructure is being built rapidly.

Going forward / institutional repositories will provide a baseline (supporting data | and basic metadata) / while disciplinary (**ディシ**プリナリー) repositories will lead reuse-oriented openness.

Standards like IIIF and TEI will help.

But it is also true that some researchers feel left behind.

So / / we also need bottom-up promotion.

If each field prepares guidelines for what to share at minimum | and how far to open / and if incentives and tracking of reuse are designed / more researchers will share data voluntarily.

**Slide 29 | Summary**

*[44:05-44:45]*

To review:

With the spread of open science / sharing research outputs is becoming standard / | and opportunities (オポ**チュニ**ティーズ) for international and interdisciplinary work are growing.

Journals clarify peer-review | and openness requirements (リ**クワイ**アメンツ); funders place duties on researchers to share supporting data.

Research infrastructure is changing / | and DMP-based management is becoming normal.

The key to progress is to use the DMP both as minimum metadata | and as a decision record at the time of sharing.

Finally / / to build a better environment / we need incentives and clearer rules.

**Slide 30 | Contact & Closing**

*[44:45-45:00]*

This is the end of my talk.

Thank you very much.

If you have questions about research data sharing / please feel free to contact NII RCOS.