

2025.1.16

2024年度京都大学図書館機構講演会「研究データの持続可能なライフ
サイクル構築へ向けて」 於 京都大学芝蘭会館 山内ホール

研究データの長期保存に関わる機能群を 対象としたオープンソースソフトウェアの比較

Comparison of open-source software for a set of functions related to the long-term preservation of research data

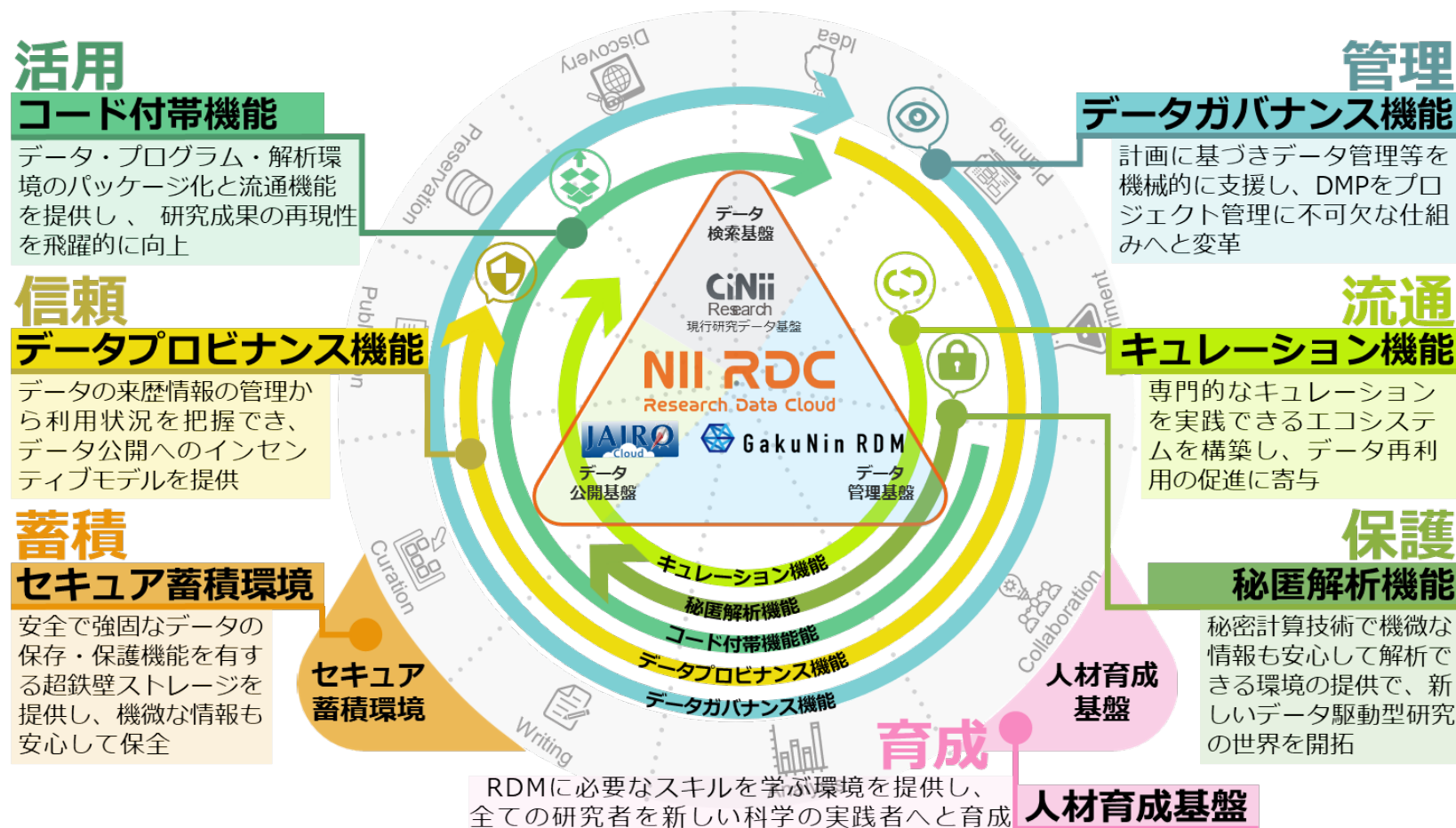
National Institute of Informatics, ROIS
Research Center for Open Science and Data Platform
Yasuyuki Minamiyama
minamiyama@nii.ac.jp
ORCID ID:0000-0002-7280-3342

自己紹介

経歴 8

2022年9月 - 現在	国立情報学研究所 オープンサイエンス基盤研究センター データキュレーション基盤担当
2019年9月 - 2022年8月	国立情報学研究所 オープンサイエンス基盤研究センター 図書館戦略担当
2018年10月 - 2019年8月	東京財団政策研究所 政策データラボ データ・ライブラリアン
2017年4月 - 2018年9月	情報・システム研究機構国立極地研究所 アーカイブ室
2014年4月 - 2018年9月	情報・システム研究機構国立極地研究所 情報図書室
2011年4月 - 2014年3月	東京大学教養学部等 図書課（駒場図書館）
2005年9月 - 2011年3月	情報・システム研究機構国立極地研究所 情報図書室
2007年7月 - 2008年3月	第49次日本南極地域観測隊 夏隊 設営一般（庶務）

研究データ管理を支えるプラットフォーム Developing platforms that support RDM



研究データの保存、共有、検索等をサポートするプラットフォーム開発が進んでいる
There has been international progress in developing platforms that support the reproducibility and reusability of research data

用語の定義

Working definition

Relation of terms: Curation⊃Archiving⊃Preservation

- **キュレーション**：データが目的に適合し、発見や再利用できるようにするために、データが作成された時点から、データの利用を管理し促進する活動。動的なデータセットの場合、これは目的に合ったデータを維持するための継続的なエンリッチメントまたは更新を意味する。より高度なレベルのキュレーションには、アノテーションや他の出版物とのリンクを維持することも含まれる
- **Curation**: The activity of, managing and promoting the use of data from its point of creation, to ensure it is fit for contemporary purpose, and available for discovery and re-use. For dynamic datasets this may mean continuous enrichment or updating to keep it fit for purpose. Higher levels of curation will also involve maintaining links with annotation and with other published materials¹
- **アーカイビング**：データが適切に選択され、保管され、アクセスできるようにし、セキュリティと真正性を含め、論理のおよび物理的な整合性が長期間にわたって維持されるようにするキュレーション活動
- **Archiving**: A curation activity which ensures that data is properly selected, stored, can be accessed and that its logical and physical integrity is maintained over time, including security and authenticity
- **プリザベーション**：データの特定の項目が時間の経過とともに維持され、技術の変化によってもアクセスして理解できるようにするためのアーカイビングの活動
- **Preservation**: An activity within archiving in which specific items of data are maintained over time so that they can still be accessed and understood through changes in technology.

各分野におけるキュレーション活動

Curation activities in various fields

Fields	Earth Sciences		Life sciences	Social sciences		Humanities/ Linguistics	Multiple		
Institutions	USGS	DataONE	EMBL	UK Data Archive	ICPSR	CLARIN-NL	DPCVocab	Data Curation Network	DCC
Steps				Transfer of data					Conceptualise
		Plan	Collecting	Assigning processing standard	Proposal development and data management plan	Identification and assessment	Ingest		Create or receive
	Plan	Collect	integrating	Data processing			Representation		
	Acquire	Assure	processing	Documentation processing	Project start-up	Development of a curation plan	Provenance management	Ingest	Appraise & select
	Process	Describe	analyzing	Metadata creation	Data collection and file creation	Curation	Systems management	Appraise/Accept	Dispose
	Analyze	Preserve	storing	Additional user information	Data analysis	Validation	Data storage	Curate	Ingest
	Preserve	Discover	sharing	Publishing data	Preparing data for sharing	Archiving	Policies	Access	Preservation action
	Publish/Share	Integrate	publishing	Delivering data	Depositing data		Preservation	Preserve	Store
		Analyze	finding	Preserving data			Public access provision		Access, use & reuse
									Transform

各分野ではデータのライフサイクルを独自に定め、キュレーション活動を実施している
Each field has its own data life cycle and curation activities

EOSCによる実践例 Practical examples by EOSC

European Open Science Cloud architecture

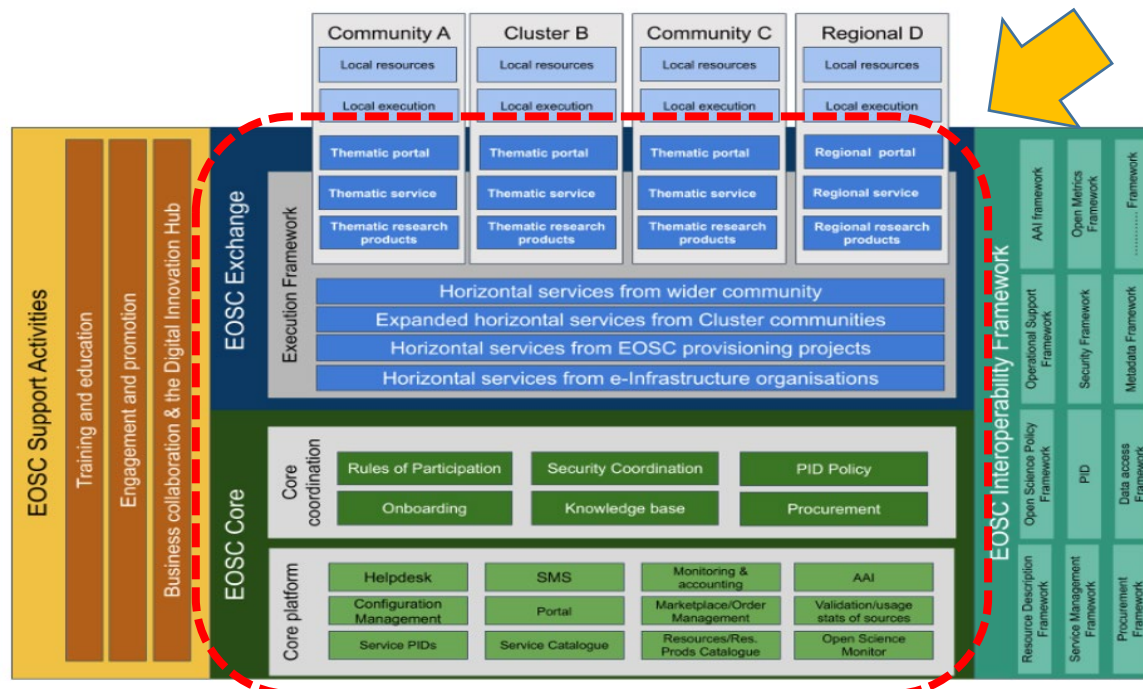


Figure 1. High-level architecture for EOSC [credits: EOSC Future Consortium [not yet published]].

機能群を“EOSC Core”と“EOSC Exchange”に分け、複数の情報システムを統合する
アーキテクチャを採用している

Service architecture that integrates multiple information systems by dividing functions into
“EOSC Core” and “EOSC Exchange”

開発のアプローチ

Development approach

1. 各分野に共通する活動を特定、実装したい
We aim to identify common activities and implement a set of related functions
2. 分野独自のキュレーション活動と連携したい
We also aim to integrate with the field-specific curation activities



共通する活動として、キュレーション活動の基礎となる
長期保存に関する活動に着目

Focus on common activities related to long-term
preservation as the basis for curation activities

NII RDCが提供するサービスの分析

Analysis of services provided by NII RDC

データの長期保存活動に対応する機能要件とNII RDCサービスのマッピング

Mapping of functional requirements and NII RDC services for long-term data preservation activities

Activities	Corresponding functional requirements	GakuNin RDM	JAIRO Cloud
データキュレーションの中止	公開期間が終了したデータに対して、墓標となるランディングページやメタデータレコードを提供できること	×	△
エミュレーション	公開対象となるデータを再現可能なレガシーシステムの構成を提供できること	×	×
ファイル監査	デジタル浸食（ビットロット等）やハードウェア障害を監視できること	○	○
コンタクト情報の管理	公開されたデータに対する問い合わせ先情報を管理できること	○	○
マイグレーション	陳腐化したファイルフォーマットを変換できること	×	×
ソフトウェアの登録	データファイルに対応するソフトウェアのレジストリを作成できること	△	×
ストレージの確保	定期的にバックアップされ、物理的に保護されたストレージ環境を提供できること	○	○
サクセッションプランニング	データを外部に預託する機能を備えること	×	×
技術動向のモニタリング・更新	リポジトリのハードウェア・ソフトウェアを最新の状態に保つことができること	×	×
利用状況の追跡	データへのアクセス件数等を追跡できること	○	○
バージョン管理	データファイルのバージョン履歴、及び各バージョンで行われた変更を記録できること	○	○

既存のNII RDCサービスではカバーされない機能群を特定
Identify functional groups not covered by existing NII RDC services

参考：データキュレーションプロセスオントロジー ref: Data Curation Process Ontology

複数分野におけるデータキュレーション活動の実態を調査し、活動に含まれるプロセスを形式化
Survey data curation activities in multiple fields and formalize the processes involved in the activities

形式化までのステップ Steps for formalization

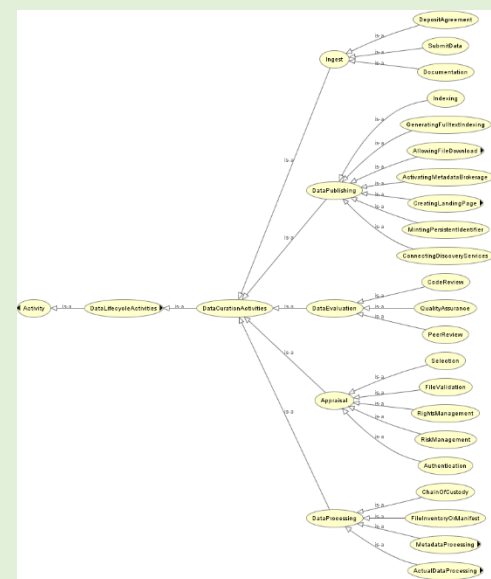
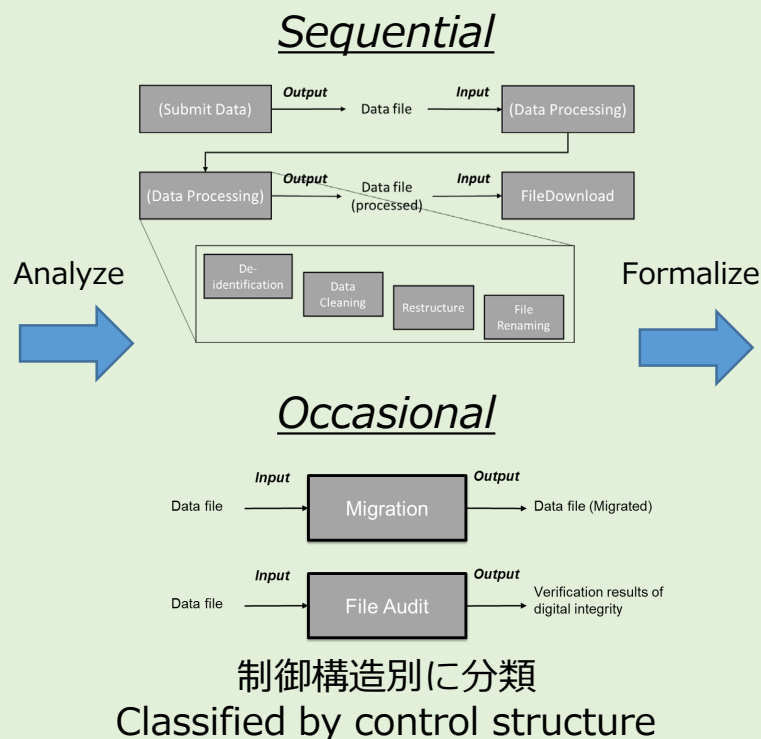
Data Curation Activities

Data curation activities can be arranged around five steps of data curation life-cycle: Ingest, Appraise (Accept), Curate, Access, and Preserve.

Ingest

- Authentication: The process of confirming the identity of a person, generally the depositor, who is contributing data to the data repository. (e.g., password authentication or authorization via digital signature). Used for tracking provenance of the data files.
- Chain of custody: Intentional recording of provenance metadata of the files (e.g., metadata about who created the file, when it was last edited, etc.) in order to preserve file authenticity when data are transferred to third-parties.
- Deposit agreement: The certification by the data author (or depositor) that the data conform to all policies and conditions (e.g., do not violate any legal restrictions placed on the data) and are fit for deposit into the repository. A deposit agreement may also include rights transfer to the repository for ongoing stewardship.
- Documentation: Information describing any necessary information to use and understand the data. Documentation may be structured (e.g., a code book) or unstructured (e.g., a plain text "Readme" file).
- File validation: A computational process to ensure that the intended data transfer to a repository was perfect and complete using means such as generating and validating file

活動を表現する語彙
Existing vocabularies



分野横断的なデータキュレーションプロセスを抽出
Extract inter-disciplinary data curation process

長期保存に関わる機能群を対象としたOSS比較

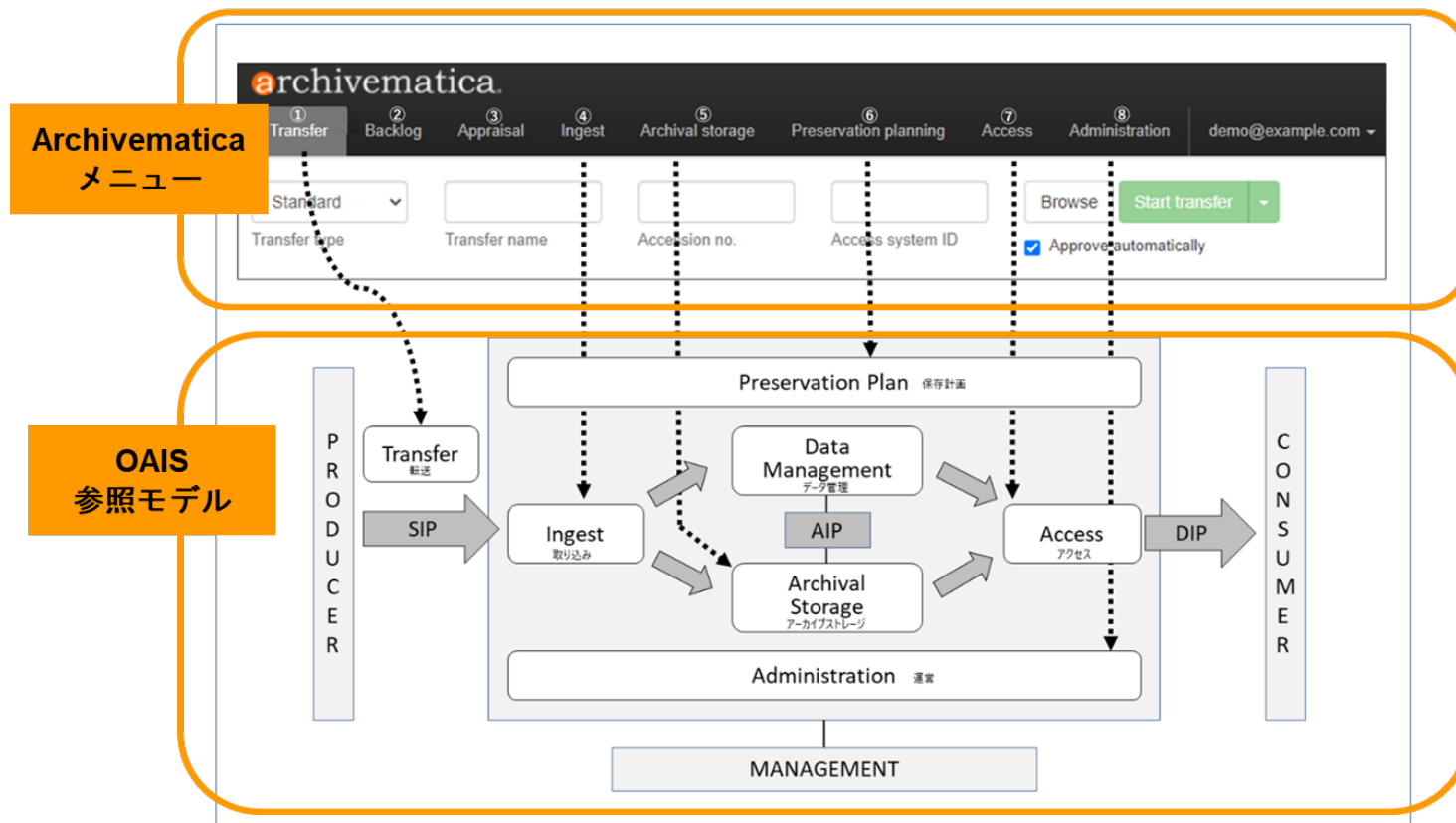
OSS comparison for functional groups related to long-term preservation

Name of Product	URL	Description	Service or software	OAIS compliant	Open Source or Commercial	GUI	File formats Identification	Formats-related function	Third party systems integration	Storage s integra
Archivematica	https://github.com/archivematica/archivematica/tree/ga/1.x	Free and open-source digital preservation system designed to maintain standards-based, long-term access to collections of digital objects.	Software	Yes	Open source	Yes	Siegfried (MIME/PRONOM/FDD)	Identification Characterization Extraction Normalization Transcription Validation Verification	Integration with - dSpace - CONTENTdm - Islandora - LOCKSS - AtoM - DuraCloud - OpenStack - Archivists' Toolkit - Arkivum - ArchivesSpace	Yes (S3 Amazon File System/O
DAITSS [Dark Archive in the Sunshine State]	https://github.com/daits	A digital preservation software application designed as a dark archive to service consortial and institutional preservation repositories a multi-user environment type. DAITSS is considered to be a first-party system.	Software	Yes	Open source	Yes	DROID/JHOVE	Identification Transformation (Normalization/Migration)	Unclear	Limited
LOCKSS	https://www.lockss.org/	LOCKSS applies, contributes to, and supports relevant standards for digital preservation, scholarly communications, and web archiving. These efforts help to benchmark fitness, facilitate interoperability, promote data interchange, and align with larger community initiatives.	Software (LOCKSS 1.x) / Service (LOCKSS 2.0)	Yes	Open source	Yes	FileMetadataExtractor (XML only)	Identification Extraction	Discovery system integration with - 360Link - SFX - WebBridge	Unclear
Roda [Repository of Authentic Digital Records]	https://github.com/keeps/roda	RODA is a complete digital repository that delivers functionality for all the main units of the OAIS reference model. RODA is capable of ingesting, managing and providing access to the various types of digital objects produced by large corporations or public bodies. RODA is based on open-source technologies and is supported by existing standards such as the OAIS, METS, EAD and PREMIS.	Software	Yes	Open source	Yes	MIME/PRONOM/Extensions	Identification Normalization	Unclear	Unclear
ArchivesSpace	https://archivesspace.org/	ArchivesSpace is an open-source web application to manage descriptive information from archives.	Software	Yes	Open source	Yes	Siegfried (MIME/PRONOM/FDD)	Identification Characterization	Digital preservation integration with	Yes (via API)

分散型プラットフォームへの対応×研究データへの拡張可能性を重視
Focus on support for distributed platforms x scalability to research data

ソフトウェア候補の選定 : Archivematica

Candidates: Archivematica



OAIS参照モデルに準拠し、情報パッケージを作成可能な統合ソフトウェアツール群
A set of integrated software tools that enables the creation of information packages compliant with the OAIS reference model

入力ファイルに対してMETSメタデータを自動生成
Automatic generation of METS metadata for
input files

```
graph TD; A["<metaHdr>  
meta Header"] --> B["<dmdSec>  
descriptive metadata Section"]; B --> C["<amdSec>  
administrative metadata Section"]; C --> D["<fileSec>  
file Section"]; D --> E["<structMap>  
structural Map section"]; E --> F["<structLink>  
structural Link section"]; F --> G["<behaviorSec>  
behavior Section"];
```

The diagram illustrates the structure of an XML Schema, showing a sequence of sections. Each section is represented by a colored box with its XML tag and a descriptive name. The sections are: metaHdr (meta Header), dmdSec (descriptive metadata Section), amdSec (administrative metadata Section), fileSec (file Section), structMap (structural Map section), structLink (structural Link section), and behaviorSec (behavior Section).

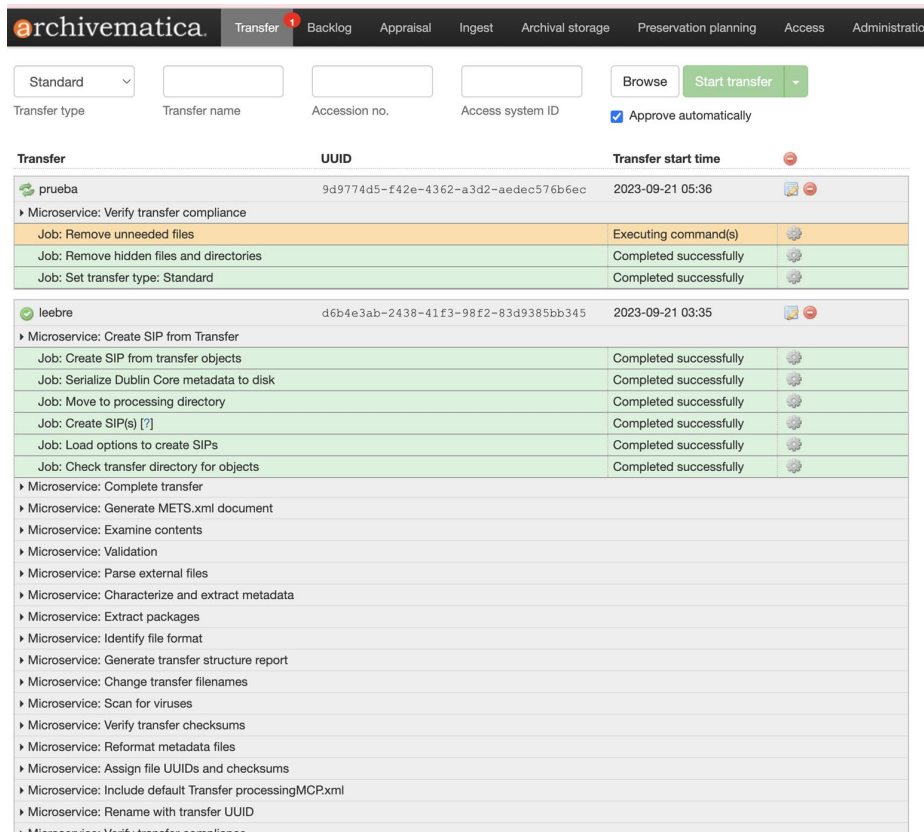
Section
<metaHdr> meta Header
<dmdSec> descriptive metadata Section
<amdSec> administrative metadata Section
<fileSec> file Section
<structMap> structural Map section
<structLink> structural Link section
<behaviorSec> behavior Section

管理に適した情報パッケージ生成・管理
Generate and manage information
packages suitable for management

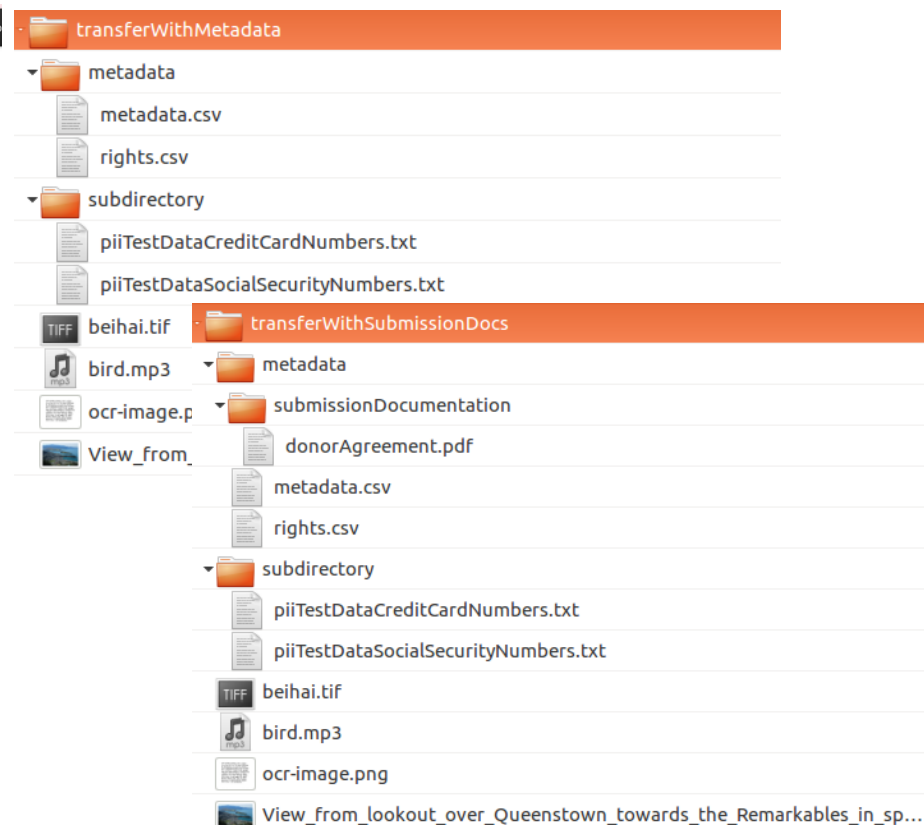
<https://www.loc.gov/standards/mets/METSPrimer.pdf>
<https://www.archivematica.org/en/docs/archivematica-1.15/>

実装上の課題 Issues in implementation

UIが専門家向け
UI is for expert data curators/archivists



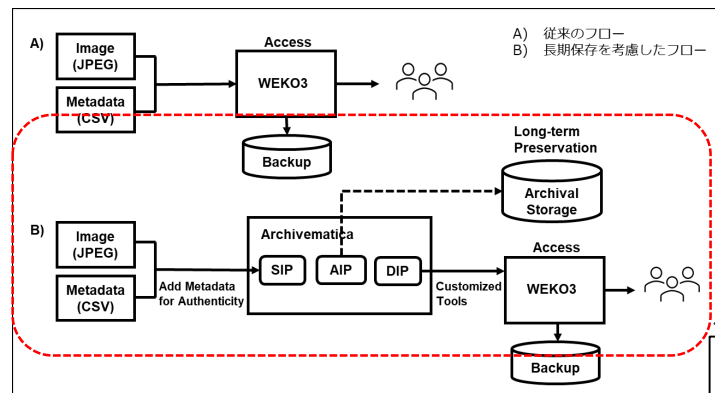
データ転送の準備が複雑
Complicated data transfer preparation



一般の研究者向けカスタマイズが必要
Customization for general researchers is required

研究者向け登録UIの試作

Trial development of registration UI for researchers



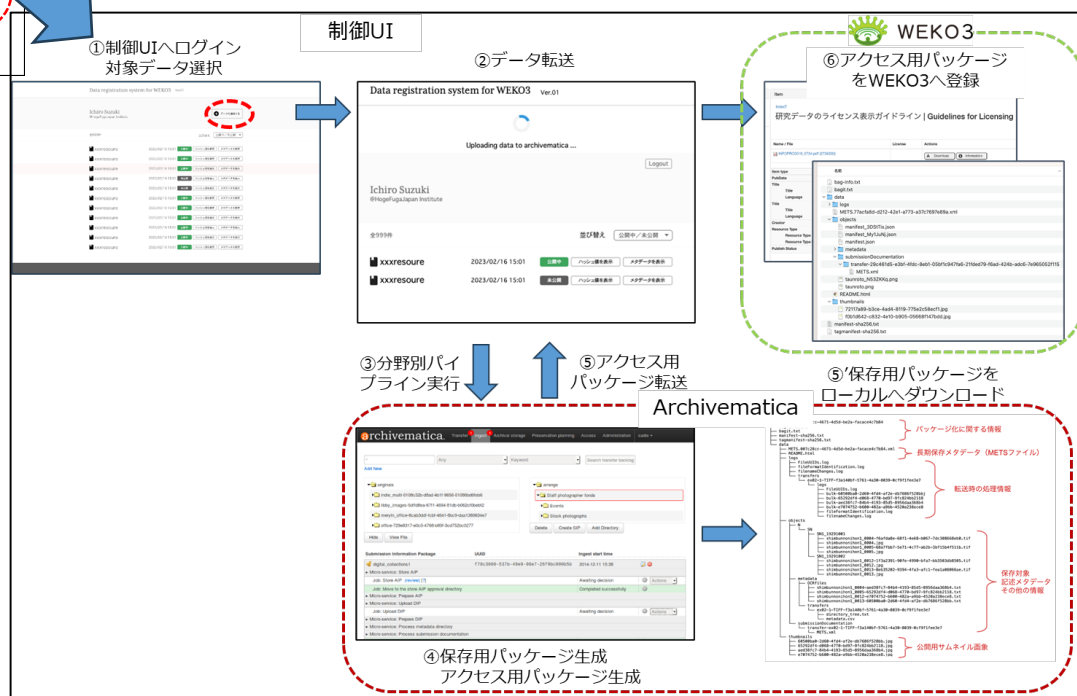
WEKO3への登録フロー
Registration flow for WEKO3

機能概要:

Feature Overview:

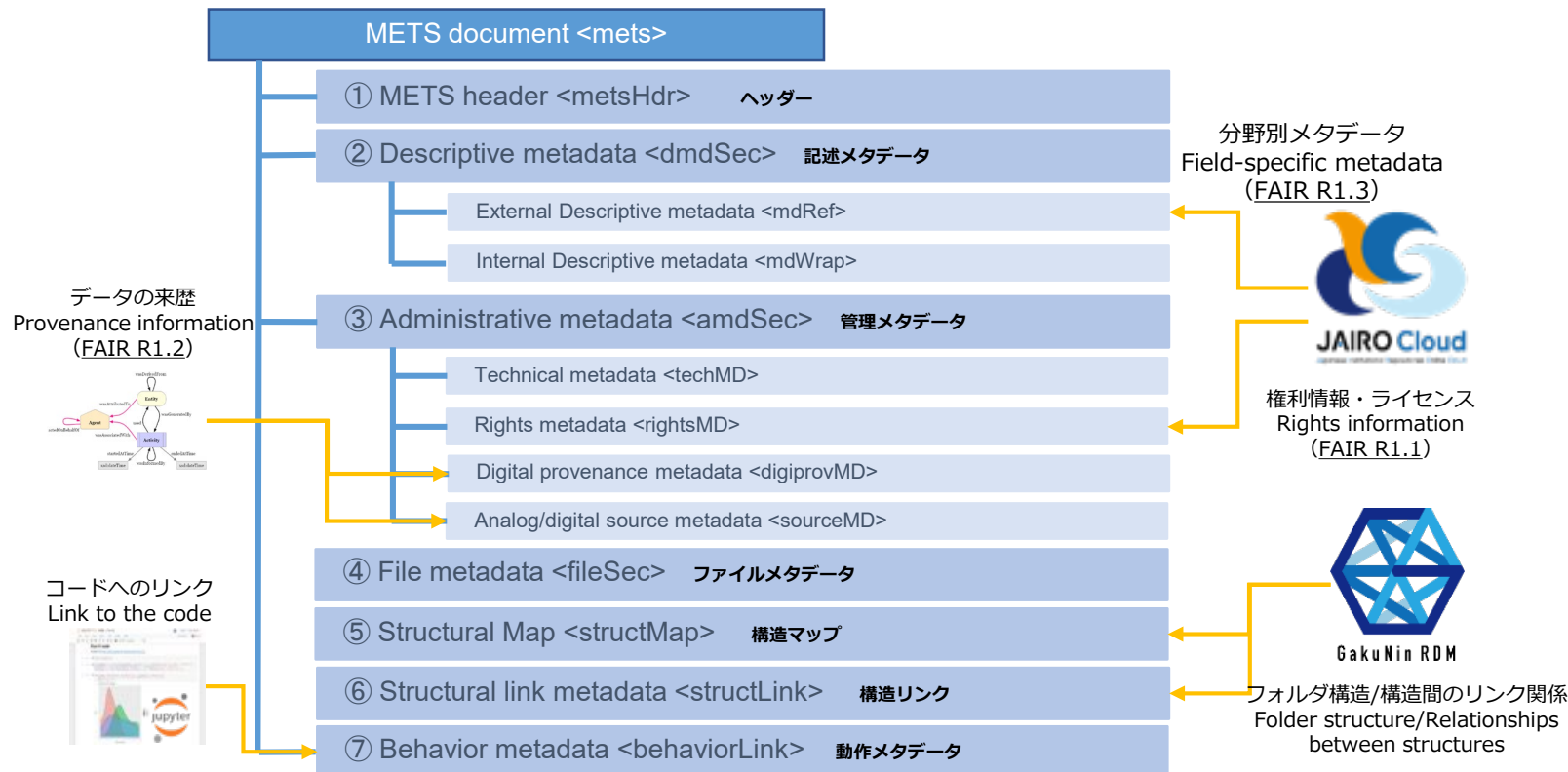
- # ユーザは制御UIを用いて各種リポジトリ (GitHub等) から対象データを取得できる
Users can retrieve target data from various repositories (GitHub, etc.) using the control UI
- # ユーザごとに個別のパイプラインを設定可能
Individual pipelines can be configured for each user
- # ユーザ管理はWEKO3上で実施
User authentication is performed on WEKO3

データの流れ Data flow



分野独自のキュレーション活動との連携に向けて Toward integrating with the field-specific curation activities

METSの構造とNII RDCで生成されるメタデータの関係 Relationship between METS structure and metadata generated by the NII RDC



FAIR原則が定める関連情報を記述でき、分野別のメタデータに対する拡張性を備えた規格として
米国国会図書館が採用するMETSを試用

Trial use of METS adopted by the U.S. Library of Congress as a standard that can describe
relevant information defined by the FAIR Principles and is extensible to field-specific metadata

まとめ

Summary

- 研究データのライフサイクルをカバーするための共通プラットフォーム開発が進んでいる
- Developing common platforms to cover the lifecycle of research data is underway
- 各分野における多様なキュレーション活動を支援するために、分野に共通する活動と分野固有の活動を整理し、統合するアプローチを採用している
- To support diverse curatorial activities in each field, we have adopted an approach that organizes and integrates common and field-specific activities
- 分野共通の活動として長期保存に着目し、NII RDCの関連基盤でカバーされない活動を特定した
- We assessed NII Research Data Cloud functionality and identified the missing set of functions related to long-term preservation as common activities
- さらに、該当する機能群に関わるオープンソースソフトウェアを比較し、Archivematicaを対象に研究者向けのカスタマイズを試行している
- We also discuss the open-source software corresponding to the set of functions and customizing Archivematica for researchers