２０１６年オープンアクセス週間‘Open in Action’ワークショップ
次期デジタルリポジトリシステムに向けた COARの活動と日本の役割

国立情報学研究所　山地一禎

（於）富山大学　2016年10月28日

次期デジタルリポジトリシステム

- 国内
  - オープンサイエンス
- 海外
  - COAR Next Generation Repositories WG

オープンサイエンスと研究データの管理

- オープンサイエンス
  - 論文だけではなく研究データもオープンにして、研究の公正性や成果の再利用性を高めようとする、新しいサイエンスの進め方。
- 研究データを、
  - **公開**しなければならないのは研究者の責任。
  - **保全**する環境を整備するのは研究機関の責任。
  - **流通**させるサポートをするのは図書館の責任。

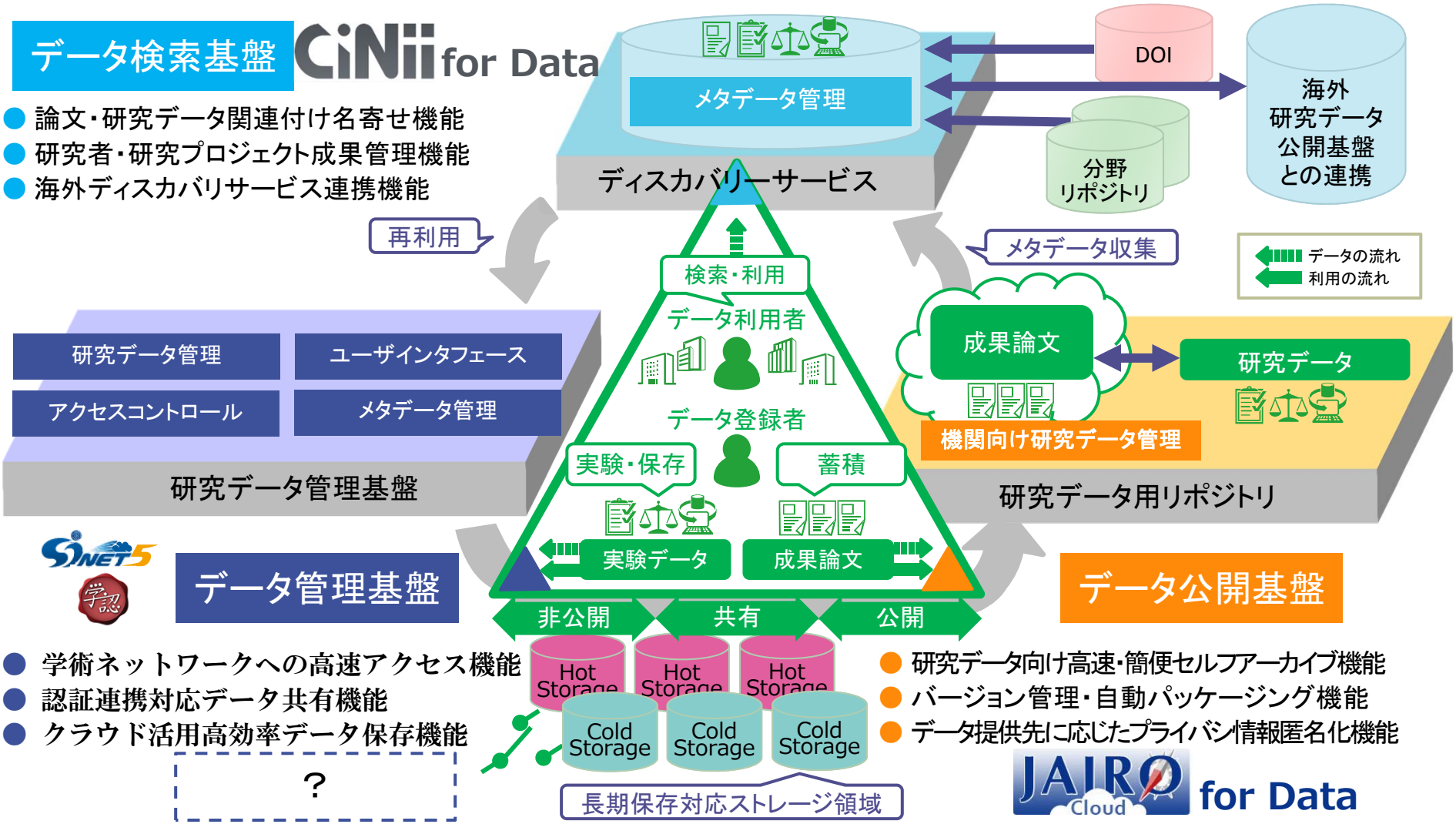
(ICSU-IAP-ISSC-TWAS working group, Open Data in Big Data World, 2015年12月　より改変)

  - 研究助成団体
    - JST：OAを推奨するポリシー → 義務化＋研究データについても言及するポリシーへの変更を検討。
    - JSPS：OAなどのポリシーについて検討を開始。
    - AMED：データシェアリングポリシー（義務化）の施行。
  - 文部科学省、学術会議
    - 研究不正対策のために研究資料の10年間の保存を原則。
    - データ・バックアップ用サーバーの提供などインフラ整備は機関の責任。

研究で利用されるツールやサービス

http://dx.doi.org/10.6084/m9.figshare.1286826

研究データ基盤



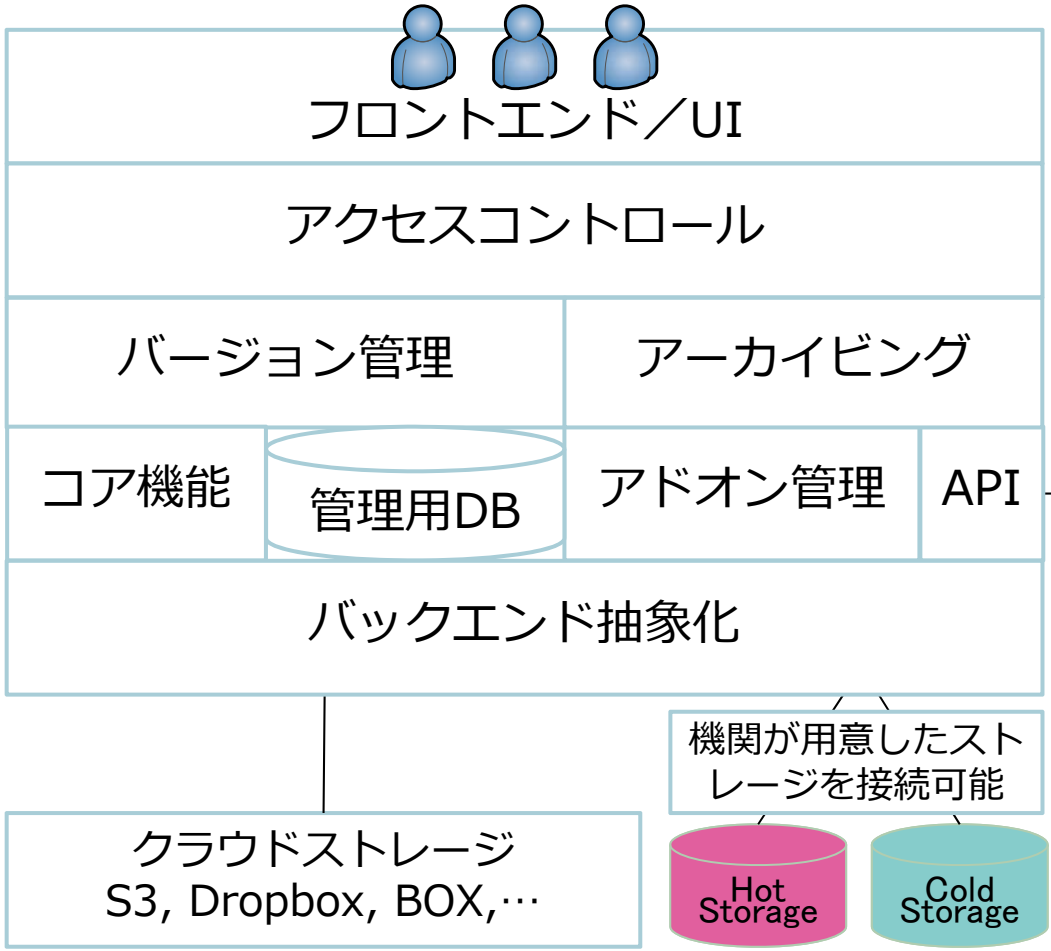
WEKO JAIRO Cloud IRP

データ管理基盤


OSF | Home
https://osf.nii.ac.jp
Open Science Framework
Browse
Support
Sign Up
Sign In
Open Science Framework
A scholarly commons to connect the entire research cycle
FREE AND OPEN SOURCE. START NOW.
Full Name
Contact Email
Confirm Email
Password (Must be 6 to 256 characters)
By clicking "Sign up free", you agree to our Terms and that you have read our Privacy Policy, including our information on Cookie Use.
Sign up free

データ管理基盤アーキテクチャー

個人あるいはプロジェクト単位（含ラボでの利用）で、
研究データを管理・共有するための基盤



異なるストレージサービスや他の
外部サービスを統一UIから利用可能

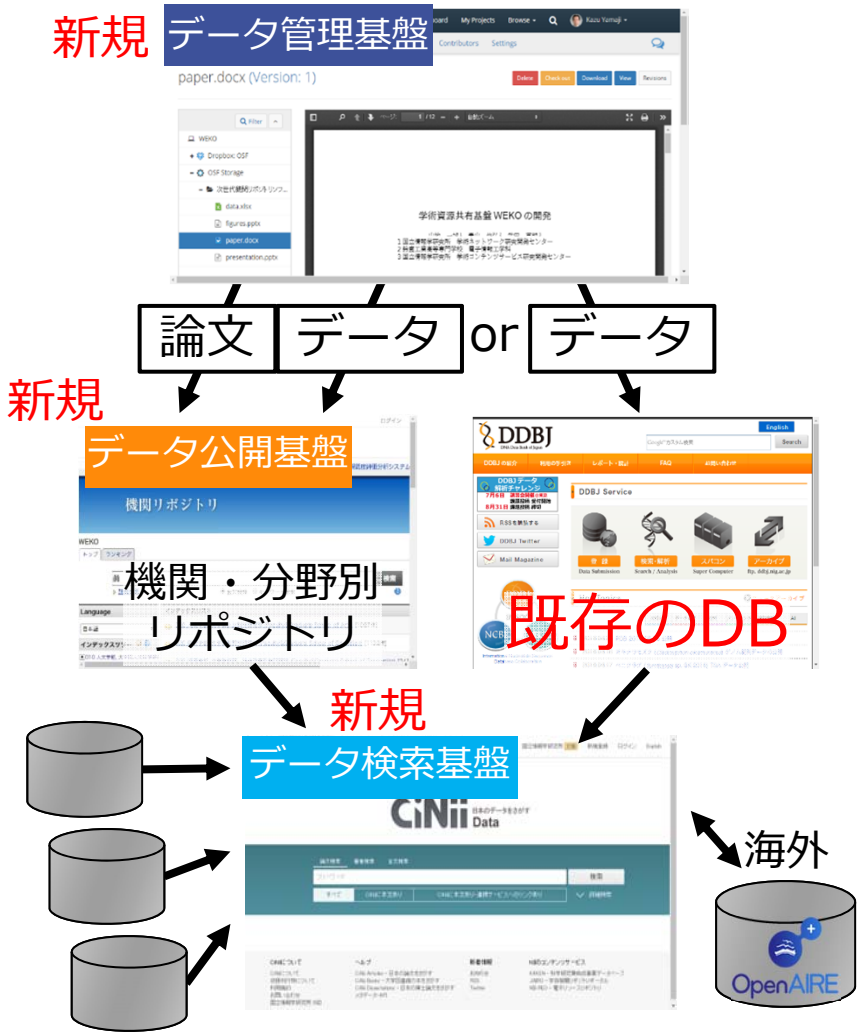
学認の属性に対応した認可機構
仮想組織（VO）を構成可能

Mendeley, figshare, GitHub,
Zotero, Dataverse, …

JAIRO Cloudにも対応

- 分野ごとに異なる機能をadd-on
として開発・組み込み可能
- メタデータ管理機能や公開基盤との
連携機能もadd-onとして機能提供

研究データ基盤の特徴



【既存】通常の研究データベース整備事業は、公開準備が整った研究データを公開するPFを提供。

【新規】**データ管理基盤**は、研究データの日々の管理や公開が容易にできる汎用的な機能を提供。個人やラボ、共同研究者間でのクローズドな研究データ管理が主目的。既存のDBが存在する分野の研究者にとっても有用な基盤。

【既存】通常の研究データベース整備事業は、特定の分野の研究コミュニティを対象に設計。

【新規】**データ公開基盤**は、研究データ公開のための汎用的なリポジトリとして整備。大学や研究所機関や、新しい分野における利用を対象。既存のDBからの移行も可能な柔軟性を実現。オープンサイエンスの裾野を広げるために必須な基盤。

【既存】これまでにも既存のDBを横断的に検索できるサービスは存在したが、特定の分野を対象。

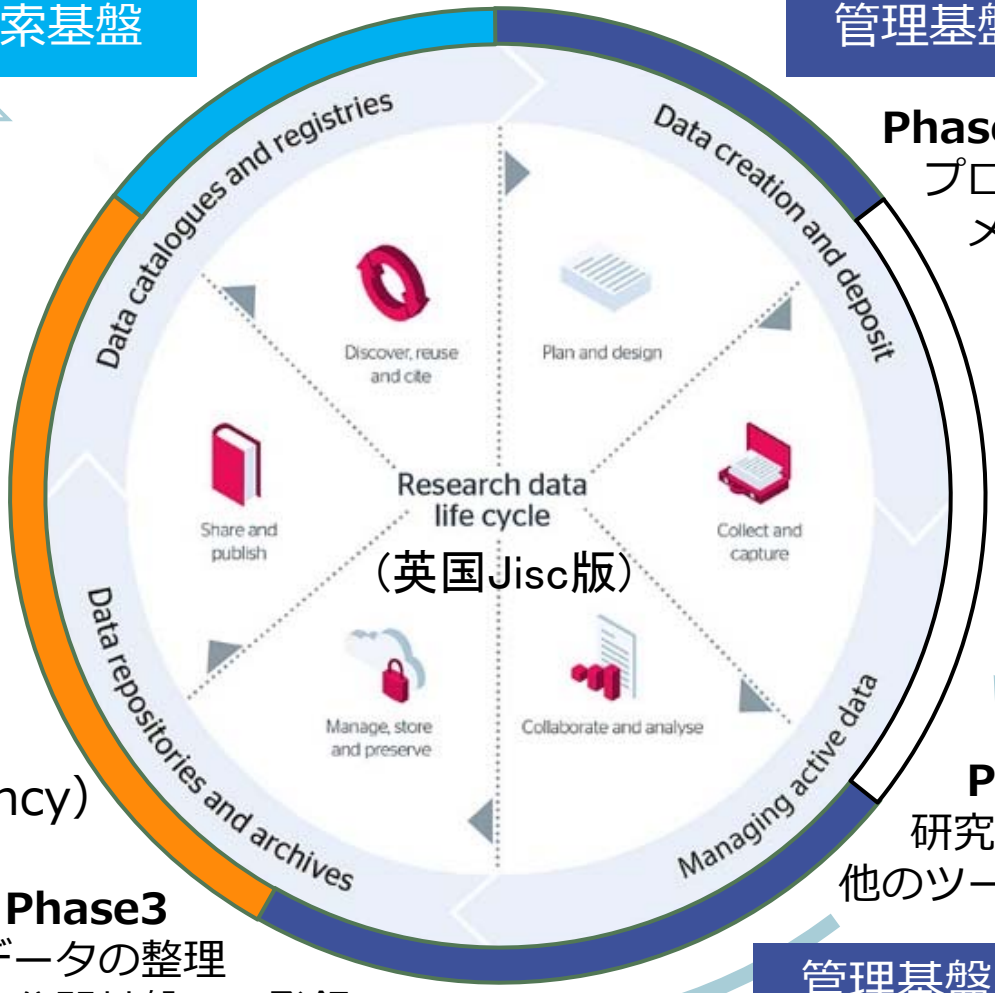
【新規】**データ検索基盤**は、分野を超越した研究データの検索サービス。論文や研究者、研究プロジェクトとの関連情報も対象。国際的な研究データ検索基盤とも連携し、国や分野を問わない研究データの発見とアクセスを実現。

一般的な研究データ
ライフサイクルと研究データ基盤との関係

検索基盤

管理基盤

公開基盤群の
情報を集約
検索可能



Phase1
プロジェクト開始（申請）
メンバー管理
基盤の初期設定

各実験装置など

Phase2
研究データの管理
他のツールとのデータ連携

管理基盤

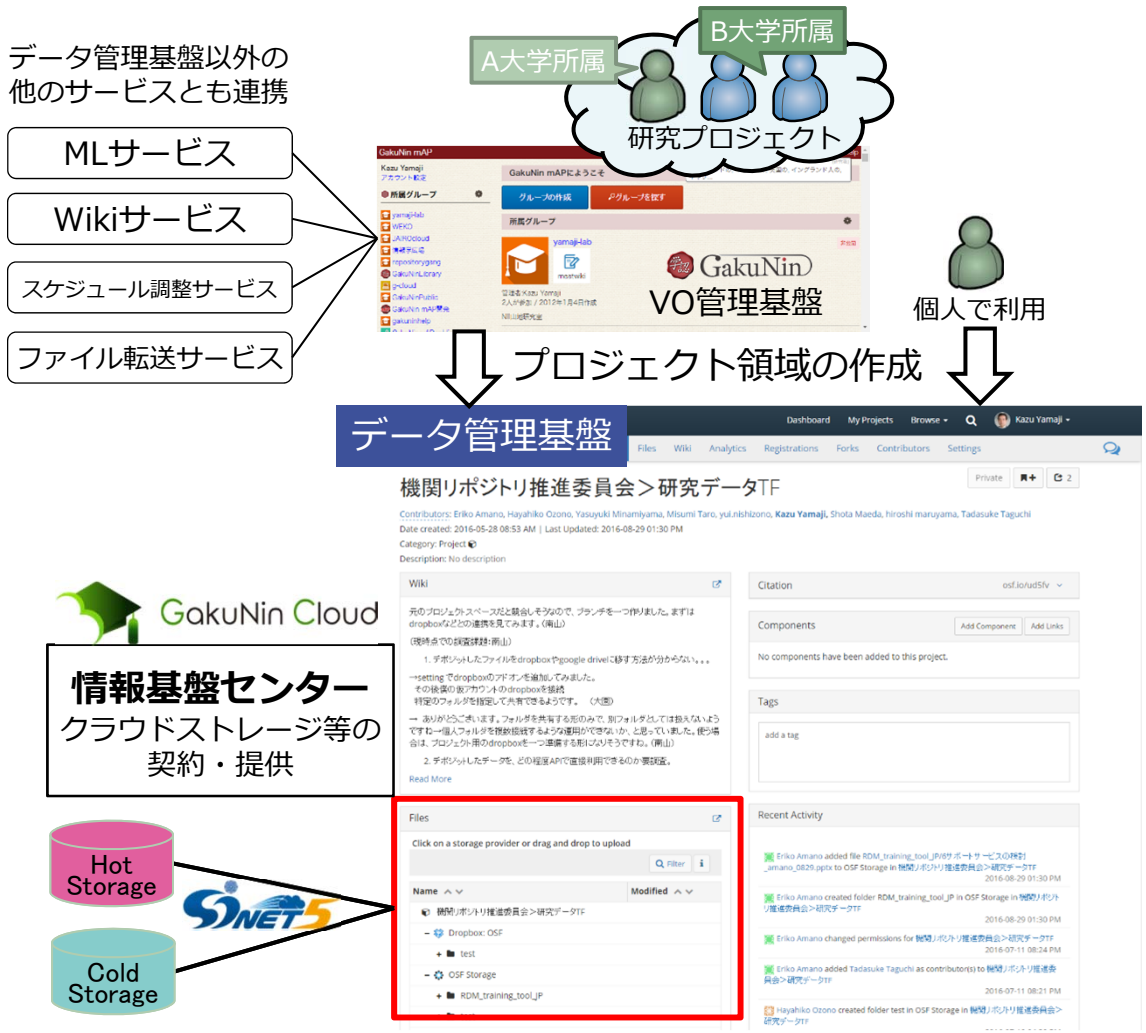
Phase3
論文執筆・付随データの整理
公開基盤への登録

公開基盤

機関/分野/
FA（Funding Agency）
リポジトリ

Phase1 管理基盤によるプロジェクト開始

― 研究過程におけるデータ管理基盤の使い方の例 ―



1. プロジェクトの発案
2. 初期メンバーでVOを作成
3. ML等で初期的な議論
4. メンバーの拡充

管理基盤上で
プロジェクト領域を作成

5. 申請書類等の共有
6. DMPの作成
7. 申請書類の完成

応募

8. ヒアリング資料等の共有

採択

9. プロジェクト開始

※外部資金プロジェクトだけではなく、ラボの管理にも活用可能

Phase2 管理基盤によるデータの管理
― 研究過程におけるデータ管理基盤の使い方の例 ―

各種クラウドタイプのストレージ接続だけでなく、外部の文献管理ツール（Mendeley）やソースコードリポジトリ（GitHub）、データ解析環境（R）やそのノートツール（Jupyter Notebook）などとも連携し研究者の利便性を向上

- 共同研究者間でオリジナルデータからの派生データや差分データを体系的に管理
- データや処理方法に関するアノテーションやメタデータを管理
- ホットストレージとコールドストレージを使い分ける機能により肥大化するストレージ領域を効率的に管理

Phase3 論文執筆・付随データの公開

― 研究過程におけるデータ管理基盤の使い方の例 ―

論文　データ　and / or　データ

機関リポジトリ

分野/FAリポジトリ

公開・非公開例

- 条件に応じたエンバーゴを設定
- 非公開データはメタデータのみを登録
- 非公開実データは管理基盤上でフリーズ

1. 論文原稿のバージョン管理
2. Mendeley等の外部ツール（add-on）を利用し引用情報等の管理
3. 論文と紐づく根拠データの管理・整理
4. 図表等の管理

投稿

5. 査読の返事を編集・共有

採録

6. FAのポリシーに応じて公開基盤に登録
7. 出版社のOAポリシー、DMP等に従い論文・付随データを公開基盤に登録
8. 図書館員やキュレータによる統制語の付与、メタデータ記法の標準化
9. DOIの付与

公開

※OAの手段や査読方法により手順は異なる

データ管理基盤 V.S. データ公開基盤

研究者

- メタデータ管理機能
- データ管理機能
- 公開基盤連携機能
- 非公開データ長期保存機能
- …

図書館員・URA

- メタデータ管理機能
- データ公開機能
- 管理基盤連携機能
- DOI機能
- …

どちらがどのような機能を提供するのが実ワークフローに最適か？

海外の状況と日本の強み

- 欧米の現状
  - OSに関するポリシーの制定、DMPの普及
  - 必要な基盤の先行開発
  - 主要な分野でのケーススタディ

従来の個々のインフラ系のプロジェクトが乱立しており連携はこれから
日本より先行しているが広範な分野・研究者への普及はこれから

- 日本の強みは共通基盤の整備・普及力
  - 日本の機関リポジトリの普及は世界でも注目
    - 各国で日本のような基盤整備への展開を模索

欧米で開発された基盤の活用　＋　JAIRO Cloudの成功パターン
研究データ基盤の共通化により研究分野間の連携を一気に促進

次期デジタルリポジトリシステム

- 国内
  - オープンサイエンス
- 海外
  - COAR Next Generation Repositories WG

機関リポジトリに関する最近の議論

IRに関する議論が活発

- OAI-PMHのサービスプロバイダーが機能していない
- グリーンOAもセルフアーカイブも成功していない
- ゴールドOAを促進するシステムとなっているだけ

これに対して

- 日本の状況はどうか？
- 何を考えるか？

機関リポジトリとは？

- Clifford Lynch (2003)
  - a set of services that a university offers to the members of its community for the management and dissemination of digital materials created by the institution and its community members
- JISC (2016)
  - A repository is a set of services[1] that a research organisation[2] offers[3] to the members of its community[4] for the management and dissemination[5] of digital materials[6] created by its community members
    - 1～6についてより詳細な定義（引用）を提供
- Herbert Van de Sompel (2016)
  - もともとのリポジトリは、“…all kinds of digital materials created by an institution‘s staff…” というものであったけど、時代の流れで“…formally published materials created by an institution’s staff…”という認識が強くなってきている。さらに、…one of the major problems of many current IRs: they don‘t provide a service to their local community… というズレもはらんでいる。

なぜ、いまNGRか？

- 現在のRepositoryの唯一の統一的機能：OAI-PMH
  - 背景
    - Pre-Printを世の中に流通させたい → arXiv
    - 90年代のWebの技術をベースに考案
  - 動作
    - メタデータの機械的な流通
    - Pull動作によるアグリゲータからの収集

- 目的
  - 超分散型のリポジトリネットワークを、世界規模でネットワーク化された学術情報流通基盤とするために、次世代型リポジトリの仕様を考える。
- 目標
  - 学術情報流通基盤のポテンシャルを引き出すために、オープン、分散管理、統一的機能、リアルタイム伝達、集約によるイノベーション創出などをキーワードに機能的特徴を表現
  - 寄与者、機関、助成機関、プロジェクトなどの識別
  - 検索、アクセス性、品質保証、コンテンツ流通、解析、発生源トレースなどのサポート

NGRのためのユーザーストーリ

1. **Discovering Metadata that Describes a Scholarly Resource**
   リポジトリのランディングページは文献情報（メタデータ）を提供するが、文献管理ツールやクローラがランディングページからメタデータを自動的に取得する共通の方法はなく、ランディングページからヒューリスティックにスクレイピングするしか方法がない。こうしたツールやクローラに共通のフォーマットでメタデータを提供できる共通の方法を定める。
2. **Discovering the Identifier of a Scholarly Resource**
   リポジトリにはコンテンツの永久識別子をHTTP URI形式で提供する。しかしながら、そのURIはアプリケーションのランディングページにリダイレクトされるために、ツールやクローラは最終的なURIのみを認識してURIとして利用する。ツールやクローラにも永久識別子を理解させるために、相互の関係を記述できる共通の方法を定める。
3. **Recognizing the User**
   通常のブログのサイトには、著者にコメントを残す機能が備わっており、著者と読者が繋がるコミュニケーションツールとして機能している。この機能がリポジトリに備われば、論文へのコメントやアノテーションを残したり、ピアレビューの機能としても活用できる。読者のユーザIDとしてはORCIDや、もしくは、GoogleやTwitter, FaceBookアカウントを活用することも考えられる。これにより、学術的な相互作用がさらに加速する。
4. **Discovering Usage Rights**
   リポジトリから提供されるコンテンツには、できるだけ制約の少ないライセンスが適用されるべきであるが実情は異なる。人間にはロゴ等で容易に識別できるように提供されるべきであるし、機械にはライセンスが記述されたURIへのリンクが提供されるべきである。ライセンスとしては、CCが適用されるのが望ましい。

NGRのためのユーザーストーリ

5. **Data mining**
   データ解析者のニーズとしては、リポジトリ横断側の解析ができることが望まれる。解析自体は、アグリゲータなどの第三者の基盤で実行されることが多いが、必要に応じて効率的にデータ（メタデータとコンテンツ）が集約でき、かつ、差分管理も含めて同期できるような機能が必要とされる。

6. **Supporting the Researchers' Workflows**
   研究者にとっては、オーサリングツールなどから論文投稿システムにワンクリックで投稿できる仕組みが望まれる。機関リポジトリや助成機関のリポジトリに登録する場合も同様である。複数のサイトに登録された場合にも、相互に自動的にリンクが張られるような仕組みが必要となる。

7. **Preservation**
   単に保存のための保存ではなく、研究成果がどのような派生的な成果に繋がっていったかを知るために、現状を適切に保存していく。研究成果間の複雑な相関関係をグローバルにグラフ化していく上でも重要である。

8. **Commenting, Annotating, Peer-Review**
   リポジトリを単なる一方通行の情報発信サイトに留めることなく、学術コミュニケーションの基盤として価値を向上させるためには、こうした研究者間のコミュニケーション機能を提供する必要がある。ただし、ピアレビューのような複雑な機能については、それをリポジトリ内で実現するのか、あるいは別の第三者のサービスとして独立させるかについては検討の余地がある。

NGRのためのユーザーストーリ

9. **Metadata syncing/automated updating of records**
現在のリポジトリはOAI-PMHによって、データプロバイダーとサービスプロバイダーの役割が分かれている。これに対して、あるリポジトリが保持するコンテンツに関連ある他のリポジトリのコンテンツの登録や更新の情報を受け取れるようにする（あるいは提供できるようにする）。こうしたリポジトリ間の情報交換を、リアルタイムに実現できる仕様を定める。

10. **Comparing usage of content in repositories**
著者（コンテンツ登録者）の視点からは、自分の論文やデータのダウンロード数、アクセス数、引用数が、他の研究者のものと比較し、自分の成果のインパクトを評価できる指標が欲しい。複数のリポジトリにコンテンツを登録する場合には、それらを統合する仕組みも必要になる。リポジトリの運用者には、リポジトリそのもののインパクトを評価できる指標が望まれている。

11. **Content Recommendation**
ユーザとしては、リポジトリを横断して、興味のある研究成果や研究者が検索できる仕組みを望む。こうしたユーザをナビゲートできる、リコメンデーション機能は、機関リポジトリで十分に開発されていない。これを実現するためにはリポジトリ間でログを共有する必要がある。

12. **Social Layer for Repositories**
関連する論文やコメントが登録されたりといった、変化がリポジトリ内に生じた場合に、リポジトリ間でイベントを共有し、ユーザに情報提供することによって、リポジトリを軸としたソーシャルなサービスを実現する。これを実現するために、リアルタイムにイベント情報をプッシュする仕組みを導入する。逆にそうしたイベントのアグリゲータ（あるいはハブ的なサービス）から、リアルタイムに情報を収集する機構を実現し、サービスに活用する。

考えるべきこと

- 自分たちが「これならできる」ではなく、機関リポジトリとして「本来こうあるべきだ」に立ち戻って改めて深く議論する必要がある。
  - これならできる、だけでは現状から抜け出せない。→ 紀要リポジトリ
  - しかし、IRによる紀要の収集は、日本の誇れるIRモデルであることも忘れてはいけない。

日本のアドバンテージ

- 日本には既に700近くのIRがある。
- 日本にはjunii2とIRDBがある。
- 日本にはJAIRO Cloudがある。
  - 他国がNext Generation Repositoriesと言っても、なかなか対応できない。
- 新しいデータ管理基盤というのも期待してもよい。
  - 本家のOSFは、SocArXiv, engRxiv, PsyArXivなど分野別プレプリントプラットフォームの利用が推進されている。

再構成できる材料はそろっている