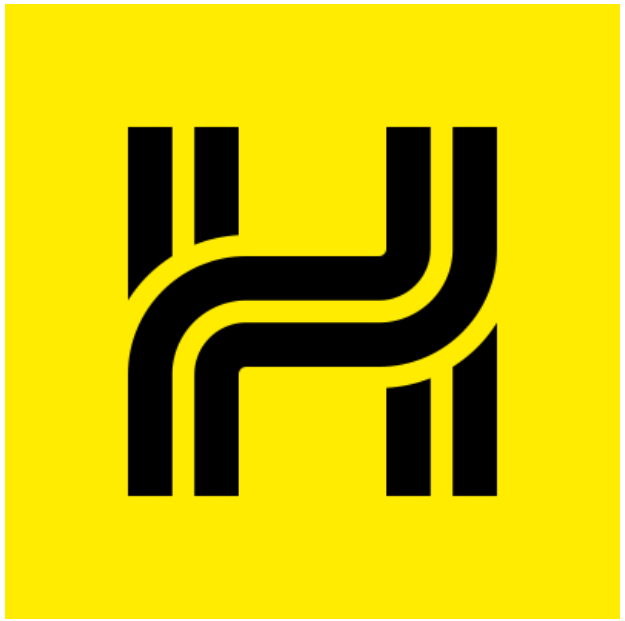
オープンサイエンスとデジタルアーカイブの二刀流で取り組むメタデータ：DIASとCODHの事例から



北本 朝展（Asanobu KITAMOTO）

ROIS-DS人文学オープンデータ共同利用センター (CODH)

国立情報学研究所

http://codh.rois.ac.jp/

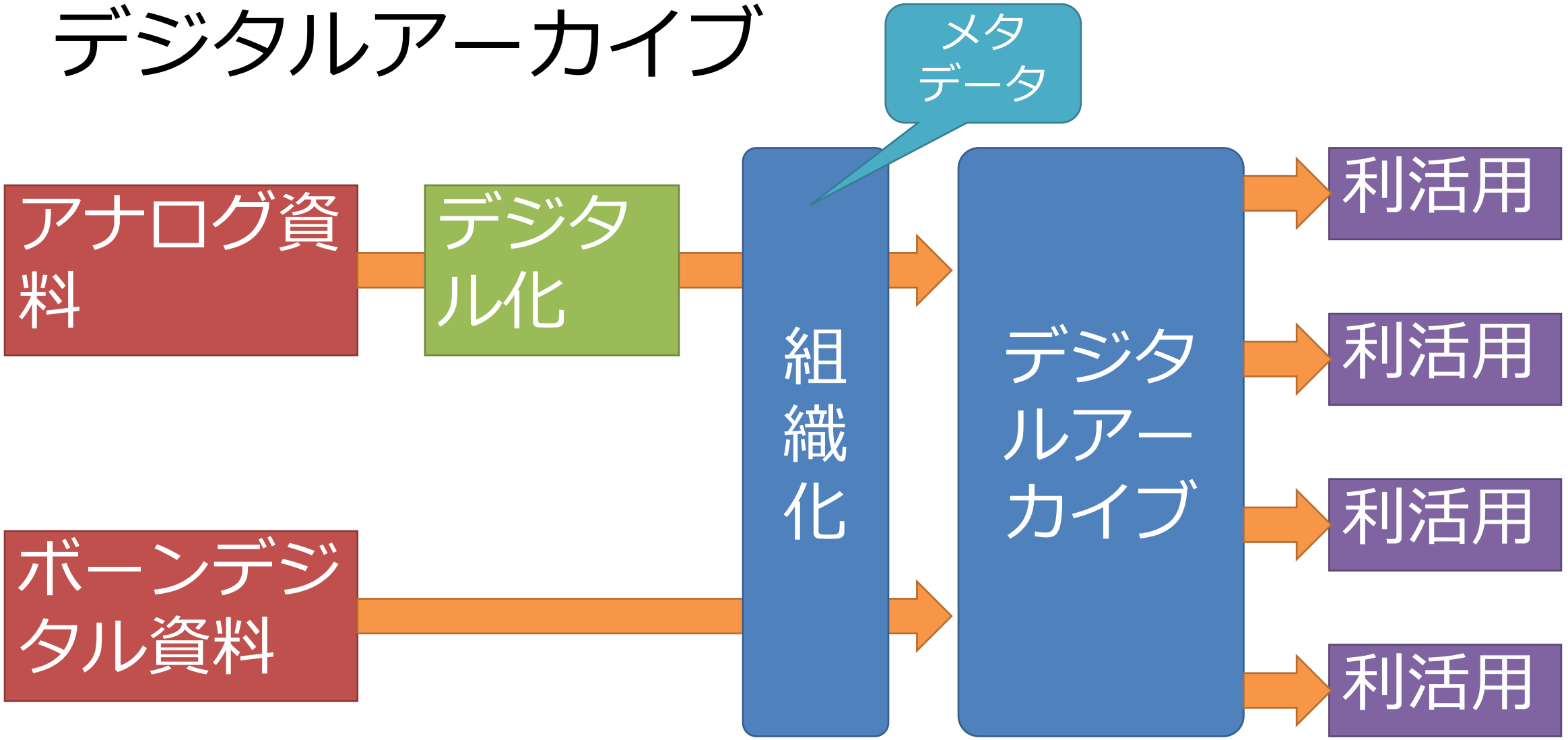
https://researchmap.jp/kitamoto/

デジタルアーカイブとオープンサイエンス

デジタルアーカイブとオープンサイエンスに関しては、情報や知識の共有という点において一見親和性が高いものの、具体的な議論や接点の模索は行われてこなかった

1. **デジタルアーカイブ（DA）**情報や知識の共有は目的
2. **オープンサイエンス（OS）**情報や知識の共有は手段
3. **手段と目的のギャップ**が、両者の距離を感じる原因？

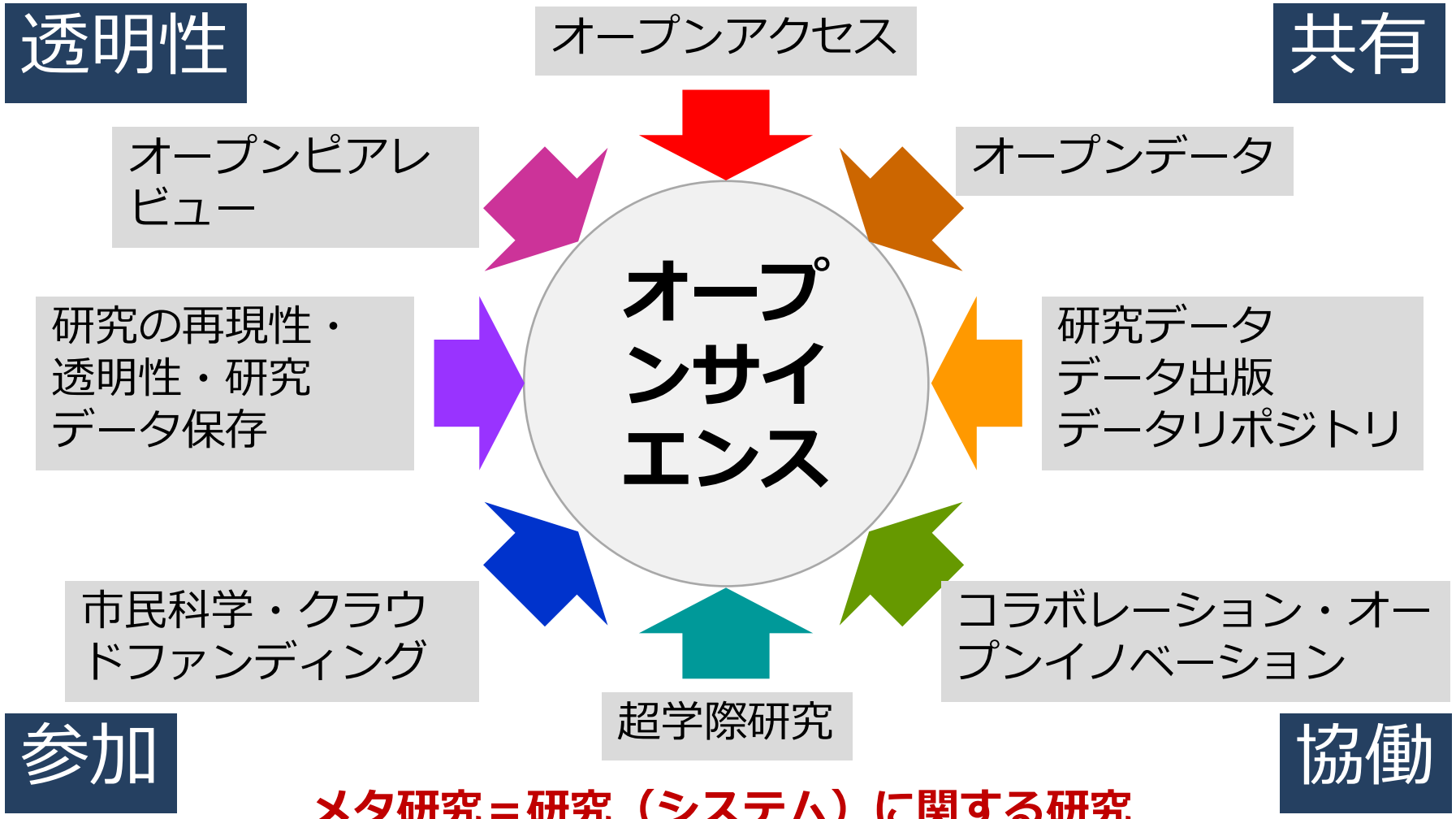
デジタルアーカイブ



デジタルアーカイブの目的

1. 伝統的なアーカイブは保存が主目的であり、共有や利用はそれに付随する目的だった
2. デジタル化の大きなメリットとして、共有や利用の促進に注目が集まった（逆に保存のメリットは微妙）
3. ボーンデジタル資料のデジタルアーカイブも最近は増えており、デジタル化は必須ではない
4. アナログ→デジタルの転換よりも、保存→利用（＝オープン化）への転換が大きい

オープンサイエンス



メタ研究＝研究（システム）に関する研究

オープンサイエンスの目的

1. 伝統的な研究はクローズドなプロセスが多く、様々な場所で不要な摩擦が生じていた
2. よりよい研究を実現する（better science）ために、「摩擦」を低減するオープン化が必要
3. 現在の研究をデジタル基盤に移行することで、アーキテクチャ的にオープン化の問題を解決する
4. OSにおける「透明性」の観点から、保存などの問題にも注目が高まる

デジタルとオープン

1. DAもOSも、共通するキーワードは「デジタル」と「オープン」
2. **DA**ではまず「デジタル」があり、その副産物として「オープン」にも価値が生まれた
3. **OS**ではまず「オープン」があり、その概念の前提として「デジタル」がある
4. **DA**はデータが中心、**OS**はプロセスが中心
5. **OS**は研究が中心、**DA**では研究は周縁？

データのライフサイクル

入力	処理	出力
研究の入力データの一つとしてDA由来のデータは重要 研究の出力を他の研究の入力に接続するのはOS領域の役割	研究の中間データをどう保存するかはOS領域の重要な課題 DAの相互運用性が高まれば、処理ワークフローとの接続が滑らかになる	研究の出力データをどう選択し保存するかはOS領域の課題 データを抽象化した知識（論文）の方が、長期保存に適する場合もある

DA的データとOS的データ

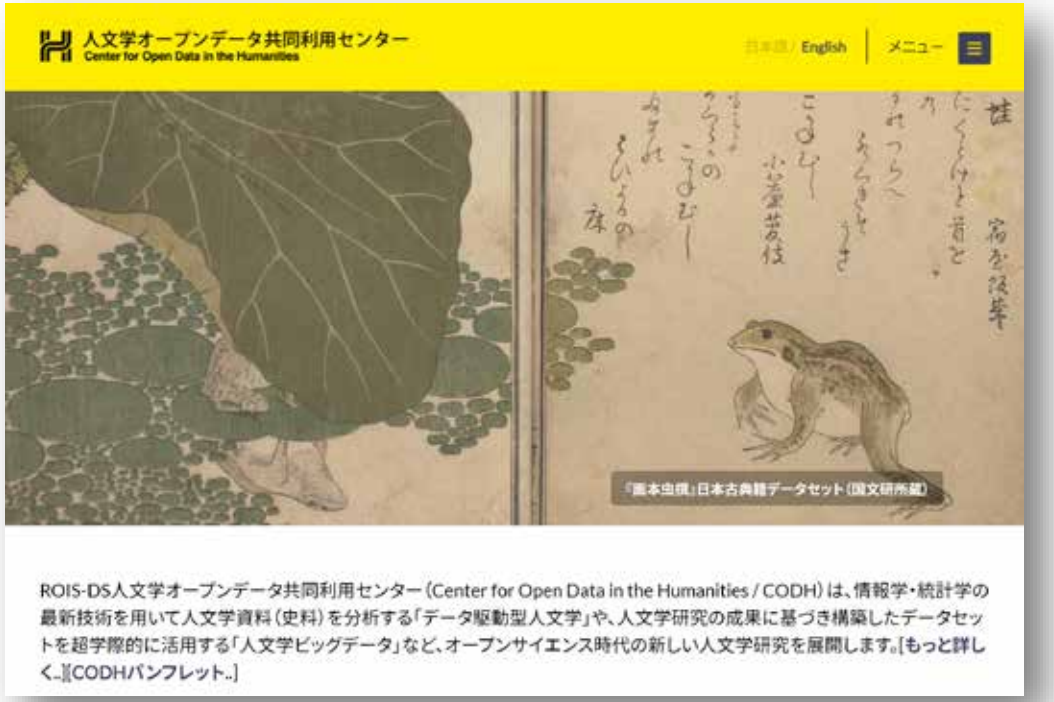
1. データに研究が紐づくデジタルアーカイブでは、メタデータはデータの記述を中心とする
2. 研究にデータが紐づくオープンサイエンスでは、メタデータはデータが生まれた文脈を中心とする
3. 研究の入力となるデータはDA的、研究の出力となるデータはOS的と言える
4. 両者の区別は、データの分野で決まるわけではなく、データの「一回性」にも依存する

メタデータ再考

1. 基本メタデータ：タイトル、作成者や連絡先など。識別子化することで、相互運用性を拡大
2. 内容メタデータ：データの内容の要約。人間が書くだけでなく、AIによる自動付与や、大規模言語モデルによる検索の高度化などが射程
3. 来歴メタデータ：データセットの品質や信頼性など。オープンサイエンス分野で知見がたまっている
4. 管理メタデータ：データを管理するための情報。主に組織内に蓄積されるため、組織外と共有する部分は少ない
5. 利用メタデータ：データの利用に関する情報。主に組織外に蓄積されるため、一望することが難しい

ROIS-DS人文学オープンデータ共同利用センター（CODH）

1. 情報学・統計学の最新技術を用いて人文学資料（史料）を分析する「データ駆動型人文学」
2. 人文学研究の成果に基づき構築したデータセットを超学際的に活用する「人文学ビッグデータ」
3. オープンサイエンス時代の新しい人文学研究を展開します



http://codh.rois.ac.jp/

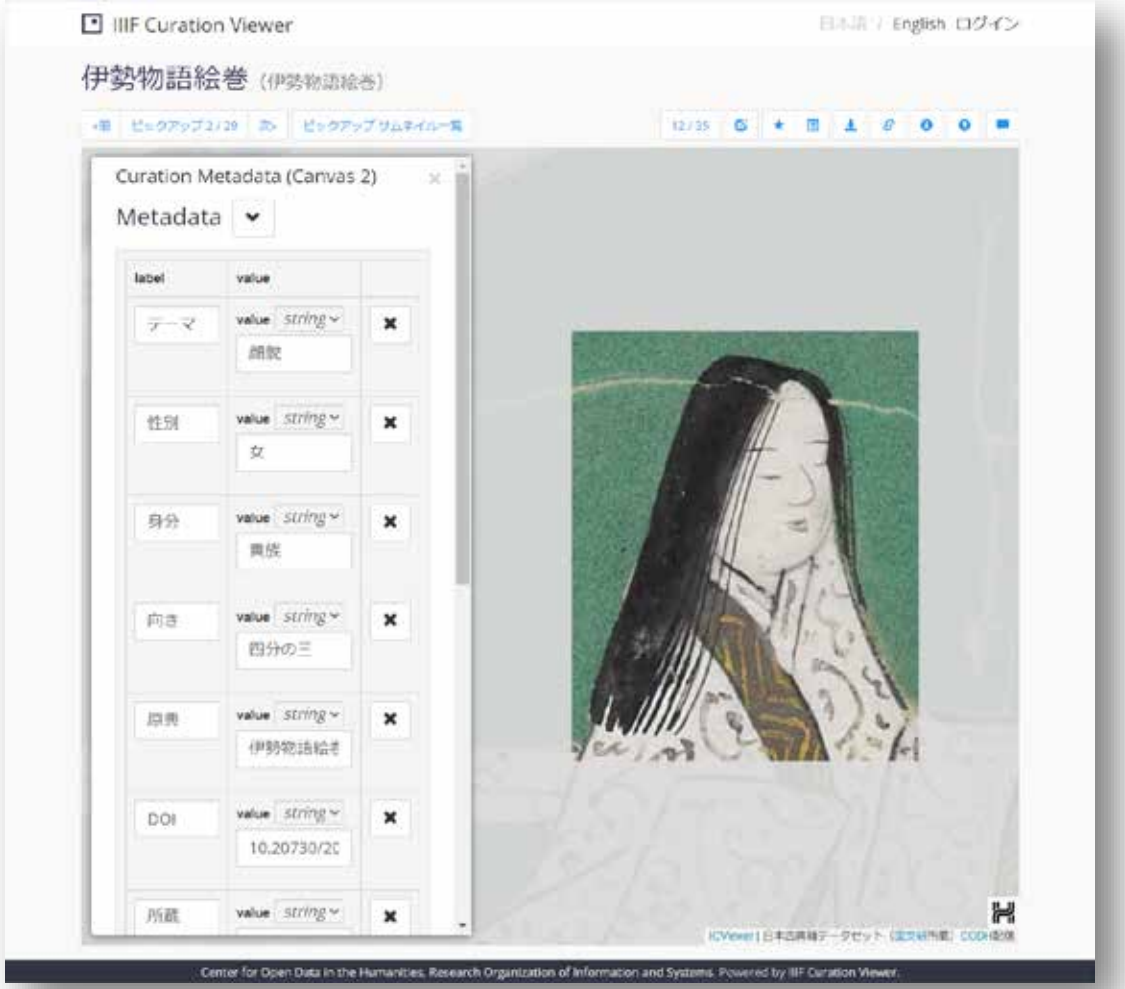
顔貌コレクション（顔コレ）

http://codh.rois.ac.jp/face/

日本美術の絵本・絵巻物などから集めた**9,683**件の顔貌を、機械学習などに活用しやすい形式でオープンデータ化

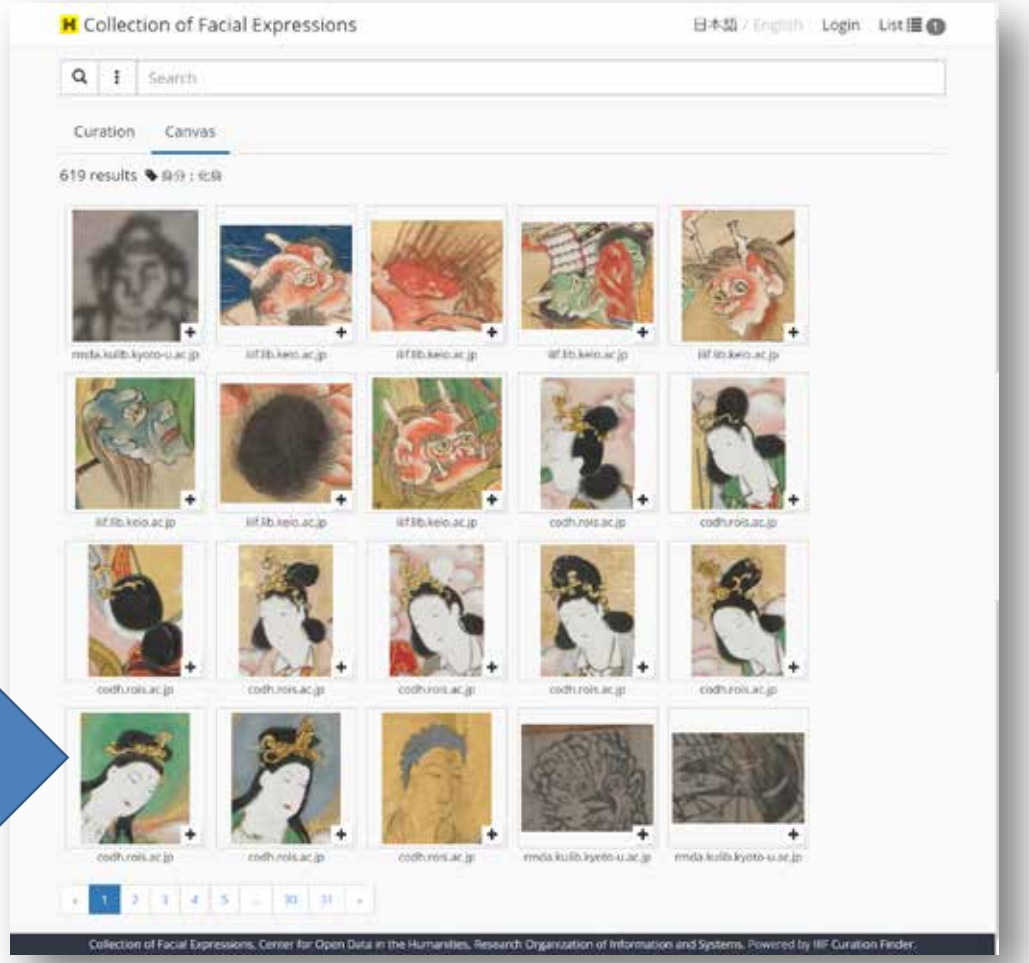
IIIF Curation Viewerでのメタデータ付与

http://codh.rois.ac.jp/software/iiif-curation-viewer/



メタデータを用いたファセット検索

http://codh.rois.ac.jp/face/iiif-curation-finder/



個々の画像にメタデータを付与すると、メタデータの値ごとに画像を検索できる

華北交通アーカイブ

http://codh.rois.ac.jp/north-china-railway/

- 原資料の翻刻
- 地理・時間情報
- 研究用データ
- AIによる自動付与タグ

写真ID	タイトル	駅	路線	撮影年月	撮影者
3803-034445-0	箭楼と機関車 東便門站	北京	京古線 京包線 大台線 通州東站線	1941年1月	安福

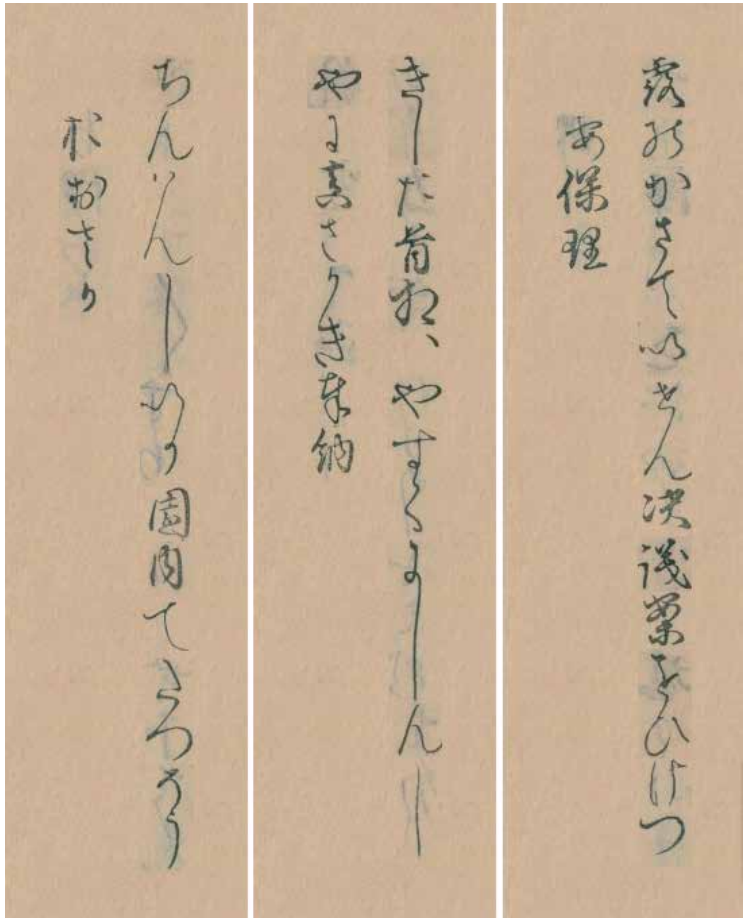
分類番号	検閲印	送付先	使用目的	備考
	軍報道課 19410313宮田			

機械タグ

steam_locomotive nematode tick thunder_snake lacewing hook quill cockroach ant common_newt isopod drumstick hatchet nail centipede banded_gecko bib missile passenger_car book_jacket

そあん（soan）

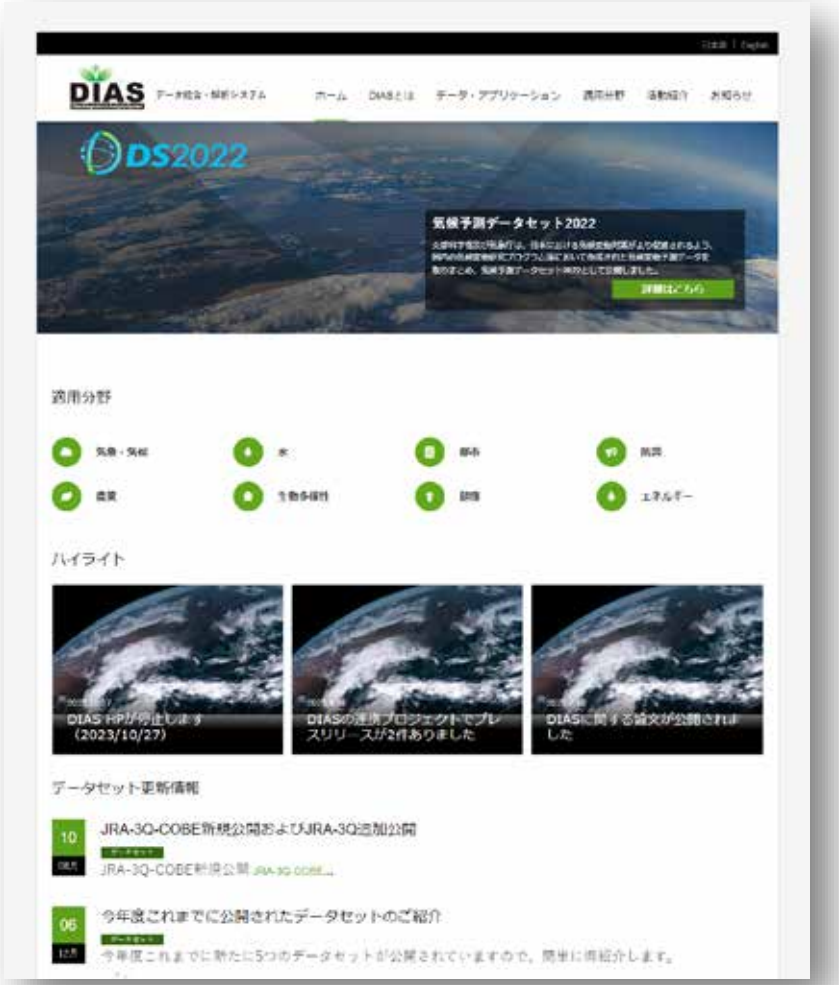
http://codh.rois.ac.jp/soan/



1. 現代日本語テキストをくずし字画像に変換し共有できるサービス
2. 現実には存在しない古活字資料の画像を生成
3. コンピュータ生成を明記するメタデータを、生成画像に埋め込み
4. 画像の来歴に関する透明性を担保し、AI学習データの「汚染」を防止

データ統合・解析システム（DIAS）

https://www.diasjp.net/



1. 地球規模の観測で得られたデータを収集、蓄積、統合、解析
2. 大規模シミュレーションから得られた気候変動予測データ等も蓄積
3. 社会課題解決に有益な情報へ変換し、アプリ・データを提供
4. 分野データリポジトリとして、大規模データを受入、公開

DIAS
受け入れ
データ

DIAS
プロジェクト
データ

データ登録審査・メタデータ入力・DOI粒度相談

DIASメタデータ管理システム

ランディングページ

データセット
ドキュメントメタデータ
超大規模ストレージ

DIAS
Data Integration & Analysis System

DIASメタデータ
→JaLCメタデータに
変換

DataCite

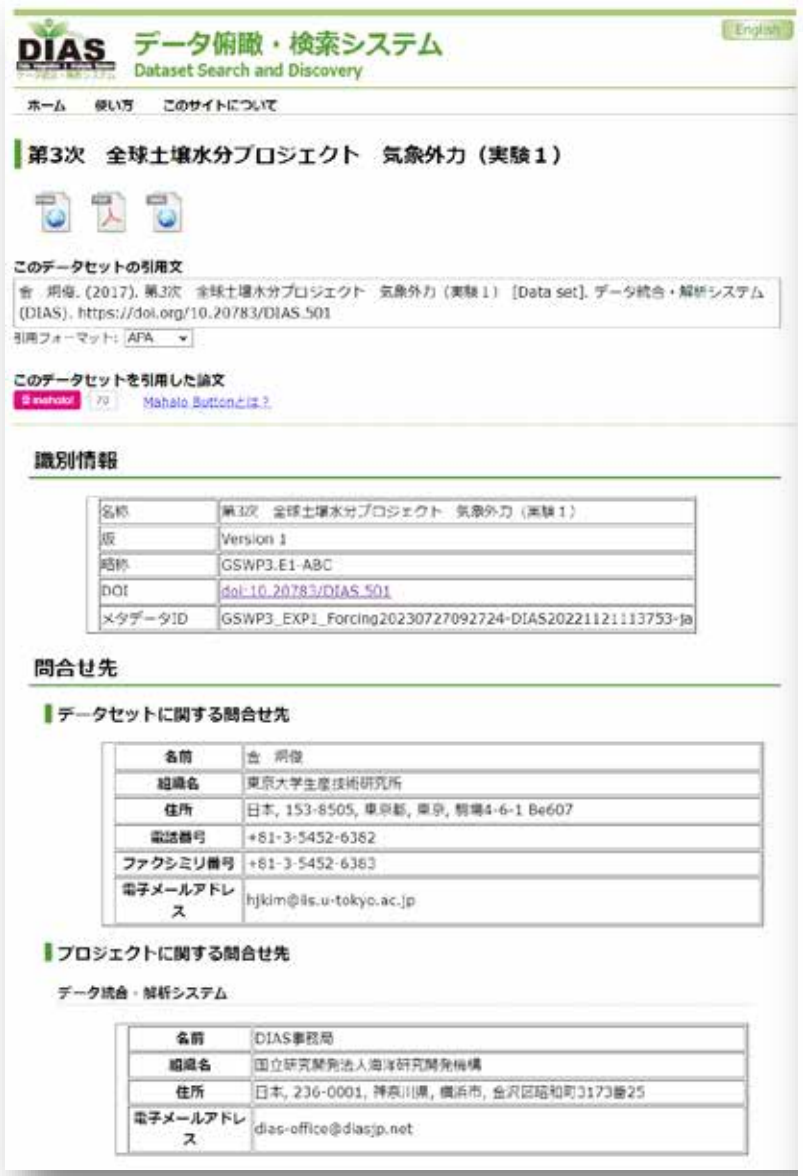
Japan Link Center

API経由で
DOI登録

DIASにおけるメタデータの状況

https://diasjp.net/apps_search/metadata-in-dias/

1. 地理的なメタデータの国際標準規格（ISO19115）をベースに、メタデータを作成する
2. ランディングページにメタデータを表示し、俯瞰検索システムで検索可能にする
3. メタデータを**DataCite**に提供して、DOIを取得する
4. メタデータを**GEOSS**に提供して、検索可能にする
5. DIAS外部（JAMSTEC等）のメタデータをハーベストして、統合検索を実現する



DIAS データ俯瞰・検索システム
Dataset Search and Discovery

ホーム　使い方　このサイトについて

第3次　全球土壌水分プロジェクト　気象外力（実験１）

このデータセットの引用文

金　炯俊. (2017). 第3次　全球土壌水分プロジェクト　気象外力（実験１）[Data set]. データ統合・解析システム (DIAS). https://doi.org/10.20783/DIAS.501

引用フォーマット: APA

このデータセットを引用した論文

Mahalo Buttonとは？

識別情報

名称	第3次　全球土壌水分プロジェクト　気象外力（実験１）
版	Version 1
略称	GSWP3.E1-ABC
DOI	doi:10.20783/DIAS.501
メタデータID	GSWP3_EXP1_Forcing20230727092724-DIAS20221121113753-ja

問合せ先

データセットに関する問合せ先

名前	金　炯俊
組織名	東京大学生産技術研究所
住所	日本, 153-8505, 東京都, 東京, 駒場4-6-1 Be607
電話番号	+81-3-5452-6382
ファクシミリ番号	+81-3-5452-6383
電子メールアドレス	hjkim@iis.u-tokyo.ac.jp

プロジェクトに関する問合せ先

データ統合・解析システム

名前	DIAS事務局
組織名	国立研究開発法人海洋研究開発機構
住所	日本, 236-0001, 神奈川県, 横浜市, 金沢区昭和町3173番25
電子メールアドレス	dias-office@diasjp.net

ドキュメント作成者

名前	金　炯俊
組織名	東京大学生産技術研究所
電子メールアドレス	hjkim@iis.u-tokyo.ac.jp

データ作成者

名前	金　炯俊
組織名	東京大学生産技術研究所
電子メールアドレス	hjkim@iis.u-tokyo.ac.jp

ドキュメント作成年月日

2023-07-27

データ作成年月日

1. creation：2017-06-01

データセット概要

序論

英文参照

トピックカテゴリ(ISO19139)

1. climatologyMeteorologyAtmosphere

時間情報

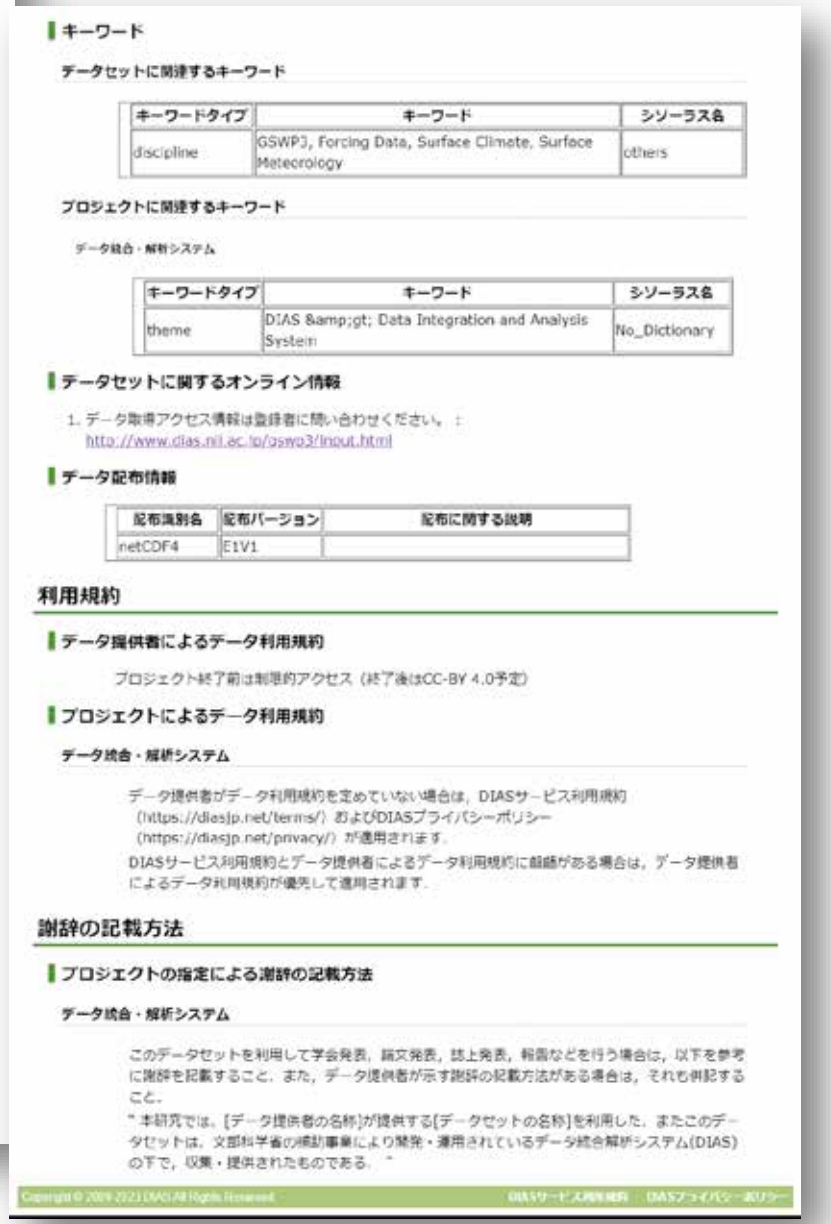
開始日	1901-01-01
終了日	2010-12-31
時間分解能	3hourly

地理的範囲

北限緯度	90
西限経度	-180
東限経度	180
南限緯度	-90

グリッド

次元の名称	次元の分割数	次元の解像度
time	321416	180 (minute)
row	360	0.5 (deg)
column	720	0.5 (deg)



キーワード

データセットに関連するキーワード

キーワードタイプ	キーワード	シソーラス名
discipline	GSWP3, Forcing Data, Surface Climate, Surface Meteorology	others

プロジェクトに関連するキーワード

データ統合・解析システム

キーワードタイプ	キーワード	シソーラス名
theme	DIAS > Data Integration and Analysis System	No_Dictionary

データセットに関するオンライン情報

1. データ取得アクセス情報は登録者に問い合わせください。：
http://www.dias.nii.ac.jp/gswp3/input.html

データ配布情報

配布識別名	配布バージョン	配布に関する説明
netCDF4	E1V1	

利用規約

データ提供者によるデータ利用規約

プロジェクト終了前は制限的アクセス（終了後はCC-BY 4.0予定）

プロジェクトによるデータ利用規約

データ統合・解析システム

データ提供者がデータ利用規約を定めていない場合は，DIASサービス利用規約（https://diasjp.net/terms/）およびDIASプライバシーポリシー（https://diasjp.net/privacy/）が適用されます．
DIASサービス利用規約とデータ提供者によるデータ利用規約に齟齬がある場合は，データ提供者によるデータ利用規約が優先して適用されます．

謝辞の記載方法

プロジェクトの指定による謝辞の記載方法

データ統合・解析システム

このデータセットを利用して学会発表，論文発表，誌上発表，報道などを行う場合は，以下を参考に謝辞を記載すること．また，データ提供者が示す謝辞の記載方法がある場合は，それも併記すること．
" 本研究では，[データ提供者の名称]が提供する[データセットの名称]を利用した．またこのデータセットは，文部科学省の補助事業により開発・運用されているデータ統合解析システム(DIAS)の下で，収集・提供されたものである． "

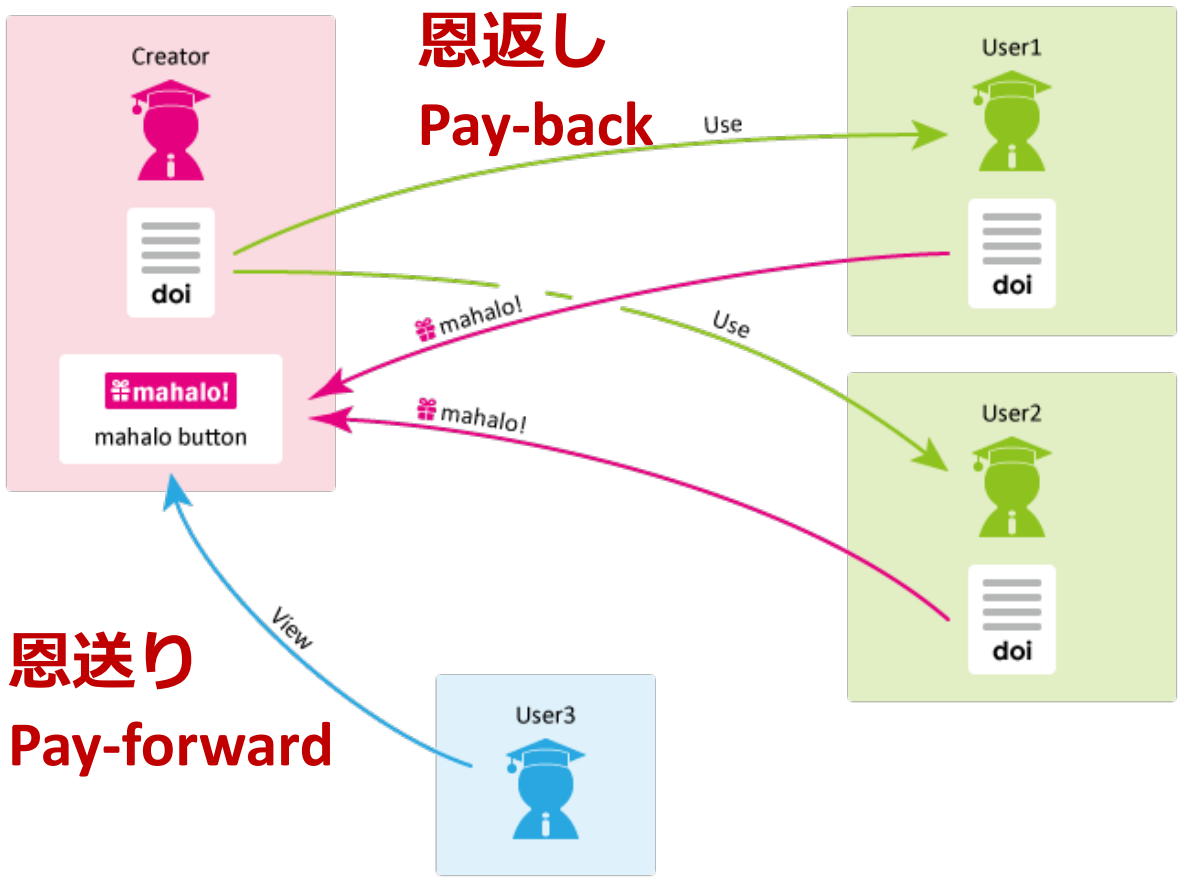
Copyright © 2009-2023 DIAS All Rights Reserved.　DIASサービス利用規約　DIASプライバシーポリシー

https://search.diasjp.net/ja/dataset/GSWP3_EXP1_Forcing

Mahalo Button

https://mahalo.ex.nii.ac.jp/

データ公開者に対する自発的な利用報告＝感謝（Mahalo）を集める仕組み

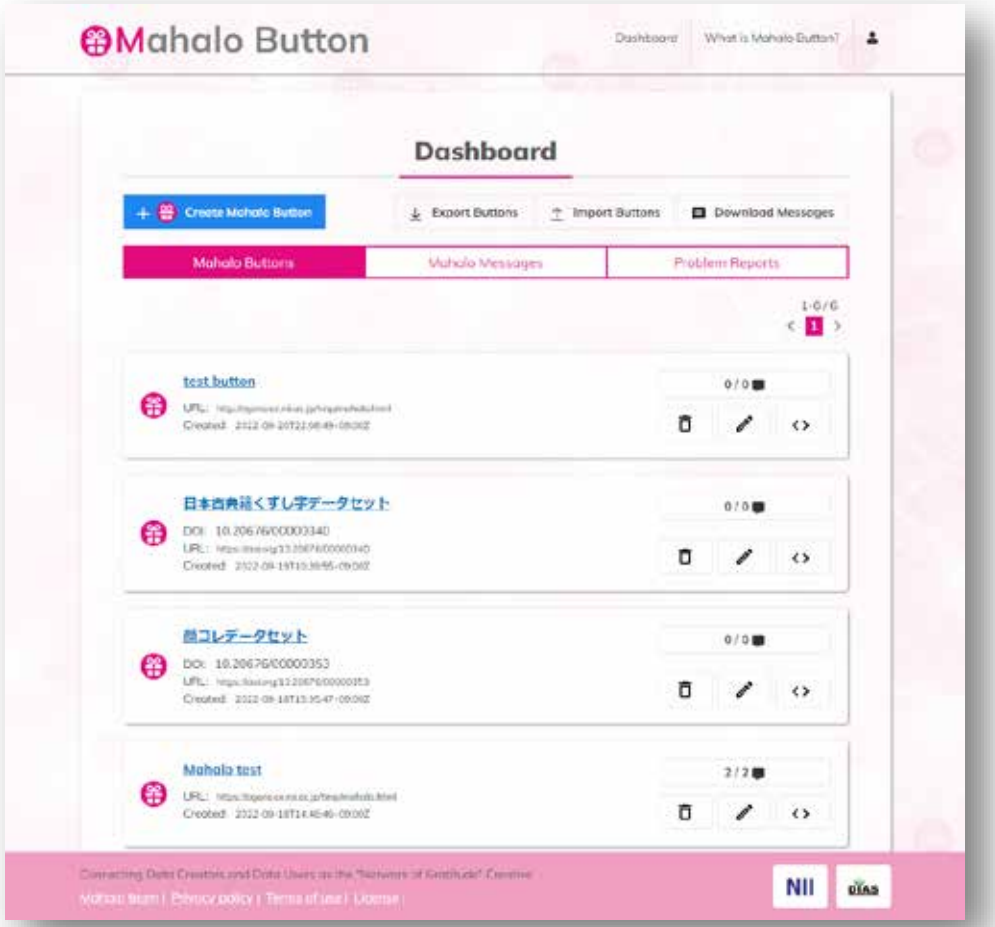


1. **恩返し**：**データ利用者から**
   データ作成者に向けて、ボタンを押して書くことで、**研究成果と共に感謝**を伝える
2. **恩送り**：**ボタンに集まった研究成果**を読むことで、**潜在的利用者**は新たな着想を得て**データの利用を進める**


DIAS
Data Integration & Analysis System
データ統合・解析システム
データ俯瞰・検索シス
Dataset Search and Discovery
ホーム
使い方
このサイトについて
第3次　全球土壌水分プロジェクト
HTML
PDF
XML
このデータセットの引用文
金　炯俊. (2017). 第3次　全球土壌水分プロジェクト
テム(DIAS). https://doi.org/10.20783/DIAS.5
引用フォーマット: APA
このデータセットを引用した論文
mahalo!
55
Mahalo Buttonとは?
Mahalo Button
Dashboard
What is Mahalo Button?
Show Mahalo Messages
55
mahalo!
Global Soil Wetness Project Phase 3 Atmospheric Boundary Conditions (Experiment 1)
DOI: 10.20783/DIAS.501
URL:
Mahalo Messages
Give Mahalo Message
or learn more about the Mahalo Message
Latest
Like
DIAS Office dias-office@diasjp.net 3 months ago
0
Decadal fates and impacts of nitrogen additions on temperate forest carbon storage: a data-model comparison
A paper using the DIAS dataset has been published. We thank the authors of the paper and the dat...
Given DOI: 10.5194/bg-16-2771-2019
DIAS Office dias-office@diasjp.net 4 months ago
0
CMIP6 Simulations With the CMCC Earth System Model (CMCC-ESM2)
Connecting Data Creators and Data Users as the "Network of Gratitude"
About | Contact | License | Privacy policy | Terms of use
NII
DIAS

Mahalo Button

https://mahalo.ex.nii.ac.jp/



DIASの外側に増えるメタデータを、DIASデータセットと紐づけるハブとしての役割

1. データの引用論文の情報を収集（これまで）
2. 論文中でデータがどのように活用されたかを収集
3. データセットを活用したコードを収集
4. 各種資料の収集を支援

メタデータへの意識

デジタルアーカイブ

- メタデータをできるだけ詳しく記述したい
- データ単体で価値があり、研究はそこに紐づく
- 文化などへの依存性もあり共通化に限界もある
- 流通への意識は低い

オープンサイエンス

- メタデータをできるだけ広く流通させたい
- 本体は研究（論文）、データはそこに紐づく
- 科学は概念が主であり、言語はその翻訳
- 記述への意識は低い

時間軸の違い

デジタルアーカイブ	オープンサイエンス
• 保存されてきたアナログデータをデジタル化	• 現在の研究を支援するためにデータを保存
• 将来に残すべきボーンデジタルデータを保存	• 現在の研究の信頼性を高めるためにデータを保存
• 利用者としても現在だけでなく将来を意識	• 利用者も現在の研究者中心（データがやや短命）
• 時間軸は現在だけでなく、過去から未来にも拡大する	• 時間軸は現在が主であり、過去や未来は優先度が低い

共通する方向性

1. メタデータは、データを発見し、処理し、成果を得るプロセスを支援する
2. メタデータは他者のための情報であり、他者が活用しやすい形式で提供する必要がある
3. 最近は、他者に「機械」が入ったため、機械可読形式で提供することの重要性が高まる
4. 今後は、デジタル基盤上で自動付与されるメタデータや、AIによる自動解析メタデータも増える

オープンサイエンスとデジタルアーカイブの二刀流

1. OSとDAでは手段と目的が異なるため、メタデータの方向性も異なる
2. DA分野では、メタデータの流通性を高めることで、OSのデジタル基盤への接続が太くなる
3. DAを用いた研究（OSやその一つとしてのデジタル・ヒューマニティーズ）の成果をDAに還流させることで、DAを豊かにしていくことが課題となる
4. OS分野は、時間軸を長く考えることで、DAの知見を取り入れやすくなり、両者の親和性が高まるのではないか