研究データのインパクトを計測可能に～データ引用とMahaloプロジェクト～

国立情報学研究所
ROIS-DS人文学オープンデータ共同利用センター
北本 朝展（Asanobu KITAMOTO)

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

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

研究成果の多様化

1. **研究成果が「論文・書籍だけ」という時代は終わった。**
2. 研究のデジタル変革＆第四の科学＆オープンサイエンスのトレンドが合流し、研究方法も変わりつつある。
3. （機械可読な）**データ**、（機械を動かす）**ソフトウェア**、（機械を利用者に提供する）**インフラ**などは、**学術コミュニティの持続的な成長に不可欠**となる。
4. 研究方法の変化と共に研究成果の評価も変え、**研究を長期的に積み上げるための貢献**に取り組む人を増やす。

研究成果と引用計測プラットフォーム

1. **研究成果＝「巨人の肩」。他者が自分の研究成果の上に立ち、その先を見通せるようになることが貢献。**
2. 査読を通ること＝**事前評価**（他者による活用を期待）
3. 引用が増えること＝**事後評価**（他者が実際に活用）
4. インパクトファクターは、査読主体に基づき引用を予測する仕組みだが、ばらつきが大きい。
5. 引用計測プラットフォームが発展すれば、**論文ごとに他者の研究への貢献**を直接的に評価できる。

引用計測プラットフォームの汎用化

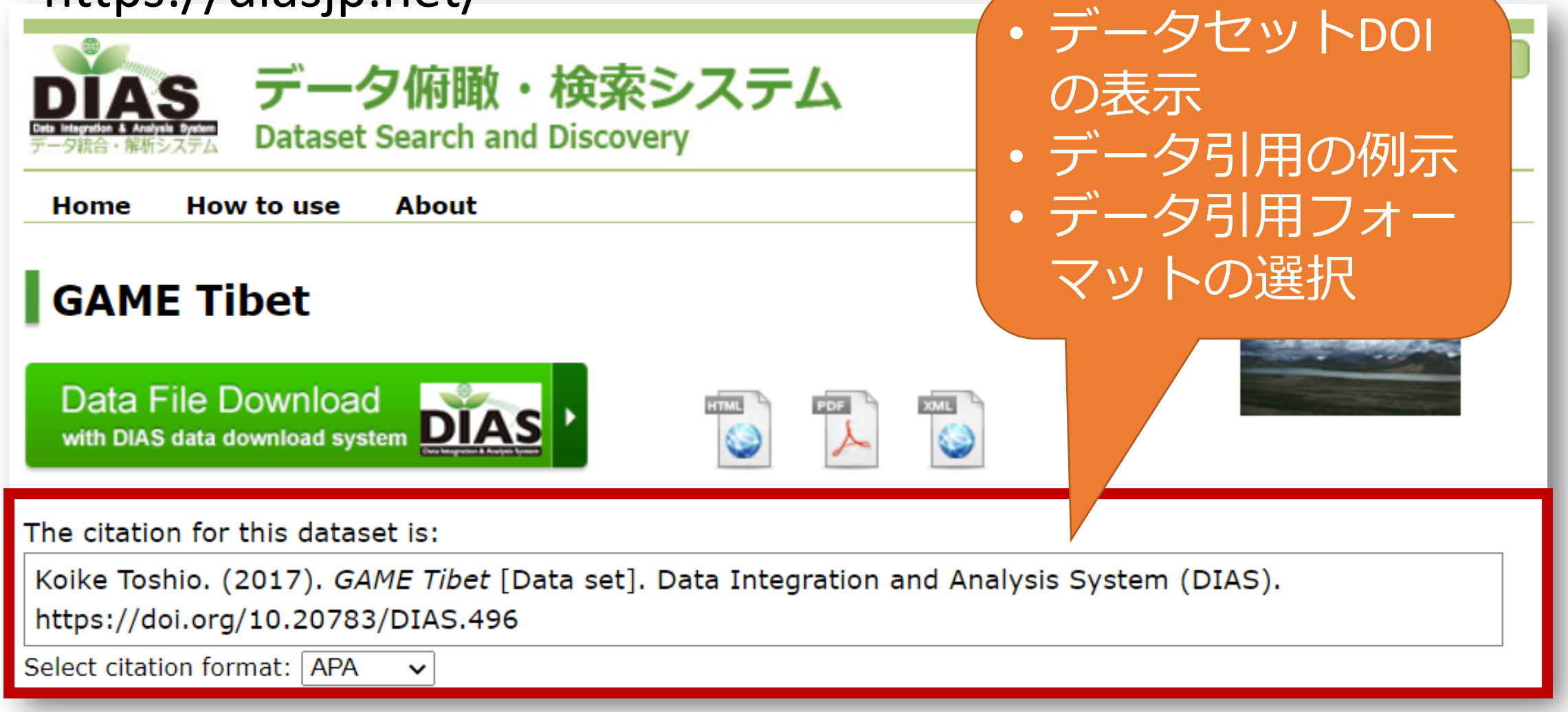
1. 論文のための引用計測プラットフォームには、他のタイプの研究成果も相乗りできるのではないか？
2. データは「**データ引用**」、ソフトウェアは「**ソフトウェア引用**」など、識別子があれば引用を計測できる。
3. 研究成果は、各種の識別子をリンクした巨大な知識ネットワーク（有向グラフ）のノードに埋め込まれる。
4. **注：謝辞はこのネットワークに含まれない**ため、徐々に役割を縮小していく。

引用から評価に至る手順

1. データに識別子を付与しておくことで、識別子からデータを逆引きできるようにする。
2. 引用情報に識別子を含めることで、名寄せすることなしに、引用関係を機械的に処理可能とする。
3. 引用関係を網羅的に収集し分析するサービスを活用することで、データの引用数を追跡できる。
4. データの引用数がデータ作成者に還元されることで、データ作成者の業績として計測可能となる。

DIASにおけるDOIの付与

https://diasjp.net/



- データセットDOIの表示
- データ引用の例示
- データ引用フォーマットの選択

疑問

1. データセットと識別子の関係は常に明確なのか？データセットDOIは常に引用されるのか？
2. データの引用数は、データ作成者にきちんと還元されるのか？
3. 引用関係の網羅的な分析は独力で行うのが難しいため、大規模事業者のサービスに囲い込まれてしまうのではないか？

浮世絵顔データセットの例

http://codh.rois.ac.jp/ukiyo-e/face-dataset/

アノテーションデータを利用する場合

『ARC浮世絵顔データセット』（Yingtao Tian、 ROIS-DS CODH作成、 ARCから収集）, **doi:10.20676/00000394**

メタデータや画像を利用する場合

立命館大学アート・リサーチセンター (2020): ARC所蔵浮世絵データベース. 国立情報学研究所情報学研究データリポジトリ. (データセット). **doi:10.32130/rdata.2.1**

データセットに関する研究内容を参照する場合

Yingtao Tian, Tarin Clanuwat, Chikahiko Suzuki, Asanobu Kitamoto, "Ukiyo-e Analysis and Creativity with Attribute and Geometry Annotation", **arXiv:2106.02267**, 2021.

3個の識別子をすべて引用すべき？機械学習分野では、最後の1個だけになる可能性が高い。

顔コレデータセットの例

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

データセット

『顔コレデータセット』（CODHが**複数の機関**から収集）, **doi:10.20676/00000353**.

原典画像公開者一覧

- 日本古典籍データセット（国文学研究資料館・ROIS-DS人文学オープンデータ共同利用センター） **http://codh.rois.ac.jp/pmjt/**
- 慶應義塾大学メディアセンターデジタルコレクション（慶應義塾大学） **http://dcollections.lib.keio.ac.jp/**
- 京都大学貴重資料デジタルアーカイブ（京都大学附属図書館） **https://rmda.kulib.kyoto-u.ac.jp/**

データセットに関する論文

Yingtao Tian, Chikahiko Suzuki, Tarin Clanuwat, Mikel Bober-Irizar, Alex Lamb, Asanobu Kitamoto, "KaoKore: A Pre-modern Japanese Art Facial Expression Dataset", **arXiv:2002.08595**.

複合的なデータセットの問題

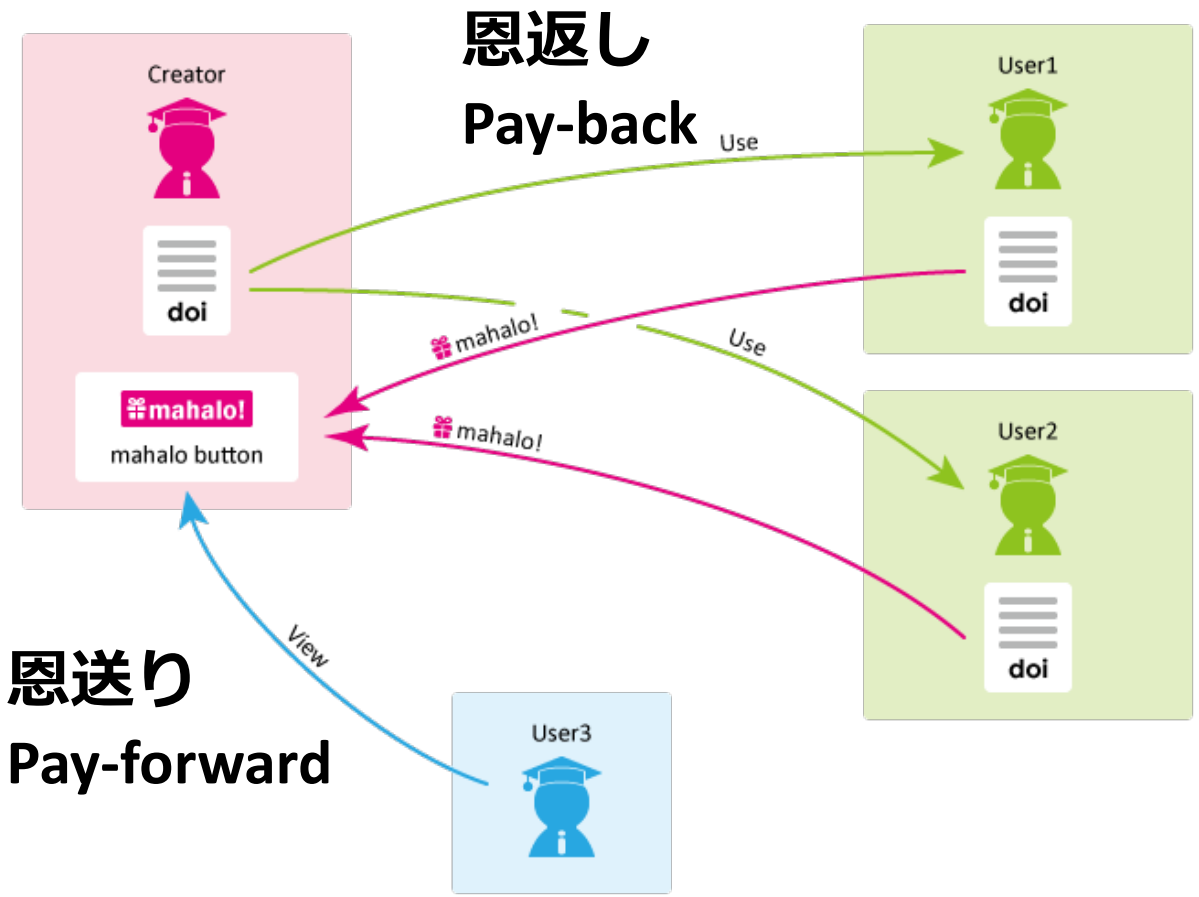
1. データセットに関する識別子が複数存在する場合、現実的には引用されるものとされないものが生じる。
2. 複数の元データセットをまとめた集約データセットが引用された場合、元データセットは不可視となる。
3. 多数の元データセットをまとめた新データセットに加え、多数の元データセットをすべて引用するように要求するのも現実的ではない。
4. 結局のところ、**ユーザに一番近いところが目立つ**？

データの利活用状況の把握

1. データのアクセス実績（ダウンロード数など）は、**ウェブサーバのアクセス解析**から把握できる。
2. データの利用実績は、契約に基づくデータ共有であれば、**定期的な報告**等で事後的に（一部）把握できる。
3. データの引用実績を網羅的に計測するには、**大規模事業者のサービス**に頼らざるを得ない。
4. **Mahaloプロジェクト：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。

mahalo project

'Mahalo' is a Hawaiian word for ‘thank you,’ but it has a broader meaning such as admiration, praise, esteem, regards and respects (Mary Kawena Pukui et al., Hawaiian Dictionary, 1986).



1. **恩返し**：データ利用者は**データ作成者に対する感謝**を、Likeボタンのように**Mahaloボタン**で伝えることができる。
2. **恩送り**：データ利用者が感謝と共に**自分の利用例を紹介**することで、ボタンは**将来の利用者のための情報ハブ**となる。

Mahaloボタン

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

1. **データ公開者**はMahaloにログインし、ボタンに紐づく一意な識別子（UUID）を生成する。
2. **データ公開者**はボタンのSnippetを貼り付けることで、データのランディングページにボタンを設置する。
3. **データ利用者**は自らの研究成果のDOIをボタンに登録することで、データの利用実績を自発的に報告する。
4. **データ利用検討者**は、ボタンに集まるデータの過去の利用実績を一覧し、その経験を自分の研究に活かす。

まとめ

1. 研究成果の多様化からデータ引用およびその評価まで、一連の経緯と最終的な目的を説明した。
2. 研究データの引用に関するいくつかの例を紹介し、実例における課題や問題点などを議論した。
3. 研究データの貢献をボランタリーな仕組みで把握するプラットフォーム「Mahalo」を紹介した。
4. **Mahaloは6月中に公開予定**（本当は今日公開したかったのですが…）

参考資料

1. **Mahaloプロジェクトウェブサイト**
   - https://mahalo.ex.nii.ac.jp/
2. Asanobu KITAMOTO, "Mahalo project: a lightweight solution for connecting data creators and users with pay-back and pay-forward incentives", Japan Open Science Summit 2018, June 2018.
   - http://agora.ex.nii.ac.jp/~kitamoto/research/publications/joss18b.html.en
3. 北本 朝展, "X-インフォマティクス：第四パラダイムに基づく科学研究の変化とデータ中心科学の発展", 情報の科学と技術, Vol. 71, No. 6, pp. 240-246, 2021
   - https://doi.org/10.18919/jkg.71.6_240