#### 第1回CODHチュートリアル DOIを中心とした識別子シ ステムの全体像

北本朝展(きたもとあさのぶ)

情報・システム研究機構・データサイエンス共同利用基盤施設・ 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/

Twitter: @rois\_codh



#### 人文学オープンデータ共同 利用センター(CODH)

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

- 2017年4月1日、情報・システム研究機構 データサイエンス共同利用基盤施設にて、 正式に発足。
- 1. 情報学・統計学の技術を用いて人文学の研究を行う。
- 2. 人文学のデータを用いて情報学・統計学の研究を行う。
- CODHの冠で様々なイベントも開催。

# DOI (Digital Object Identifier) とは?

#### FAIR Data Principles

https://www.force11.org/group/fairgroup/fairprinciples

- FAIR原則:データ駆動型サイエンスにおいては、研究データをFindable, Accessible, Interoperable, Re-usableな状態にしておくことが重要である。
- Findable: <u>識別子</u>、メタデータ、リポジトリ。
- 2. Accessible: オープンなプロトコル、認証。
- 3. Interoperable: 共通フォーマット、語彙。
- 4. Re-usable: ライセンス、来歴、標準。

#### 識別子としてのDOI

10.20783/DIAS.496

10.20783 = 10.はDOI番号(Handle System由来)、 20783はIDFが付与。

オブジェクトと文字列の紐付けをグローバルに管理し、オブジェクトへの永続的アクセスを保証する仕組み。

- •単純化すれば、たったこれだけ!
- 単純に見えて実は奥深く、しかも核心的な機能、それが識別子である。

#### DOIの仕組み

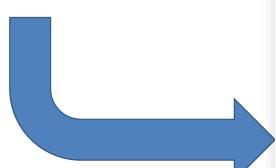
#### DOI = Prefix/Suffix

画面表示・印刷表記は DOI名の前に「doi:」を つける。

- Prefixは国際DOI財団 (IDF)が一元管理
   → グローバルに通用する識別子となる。
- 2. Suffixは独自に管理 → Unicode印刷可能 文字。この部分に意味を持たせてもよ いが、DOIはそれを意識しない。
- 3. レゾルバの運用 → actionableリンクとするには、URI表記としてDOI名の前に https://doi.org/」をつける。

### DOIレゾルバ(成功)

https://doi.org/10 .20783/DIAS.496



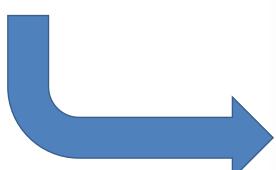
Handleシステムを利用。 http://hdl.handle.net/10.207 83/DIAS.496 でもランディ

ングページに到達できる。



## DOIレゾルバ(失敗)

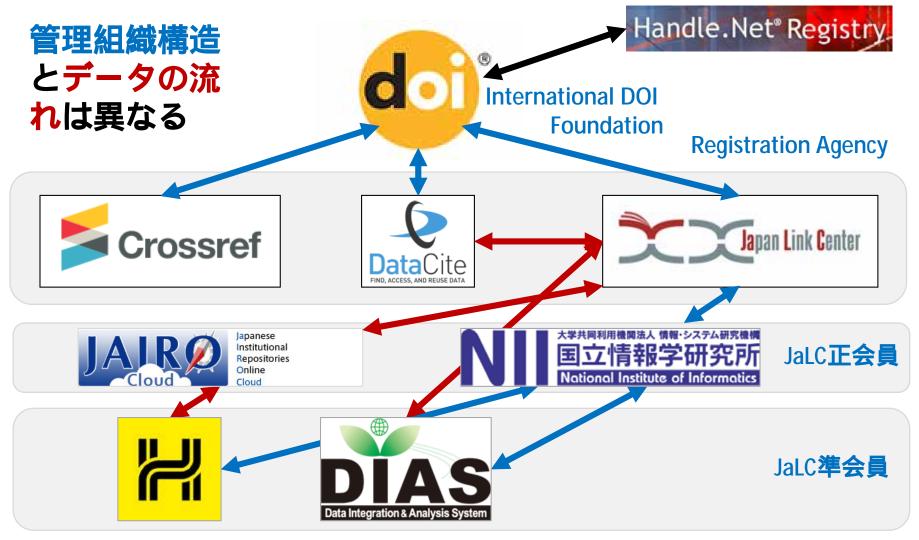
https://doi.org/10 .20676/00000001



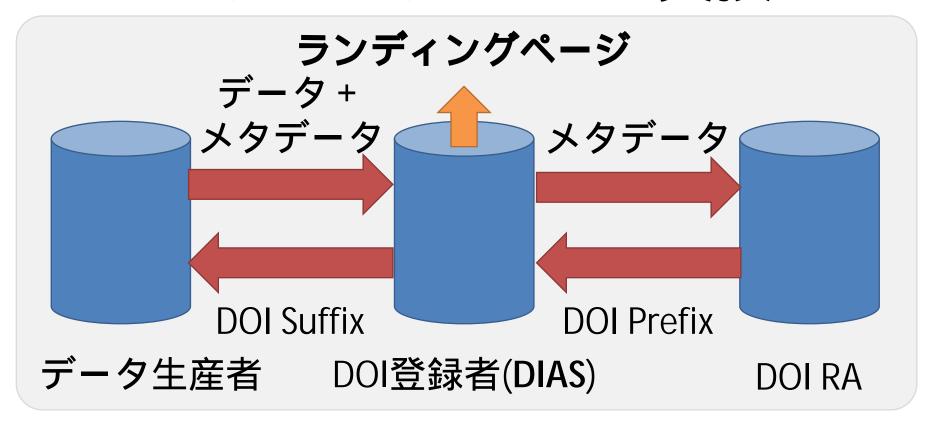
レゾルブに失敗した場合 は、メタデータが最上流 (IDF)に到達するまでし ばらく待つ。



#### DOIシステムの全体像

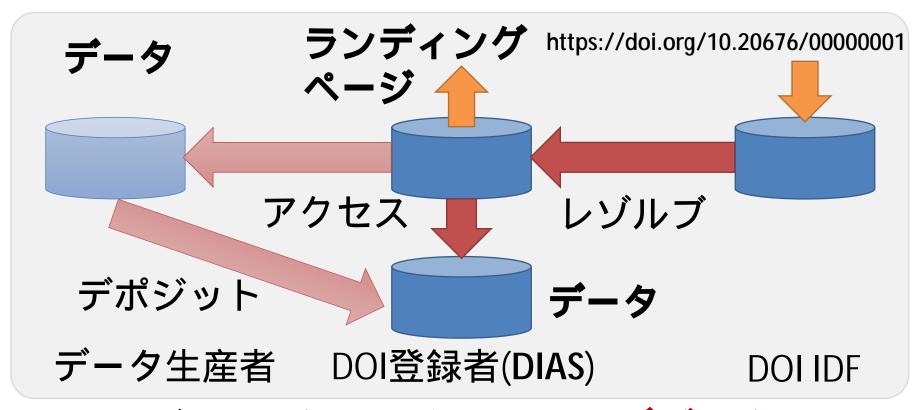


#### メタデータとDOIの交換



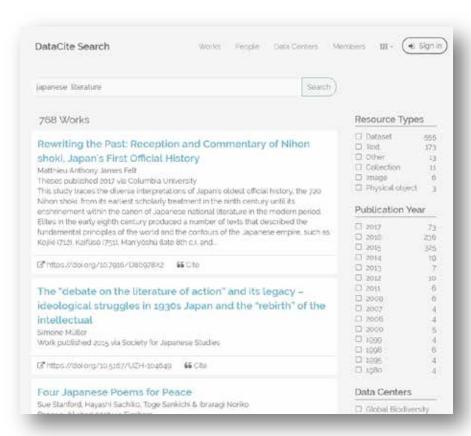
DOIをGiveする代わりにメタデータをTakeすると、 DOIチェーンの上流にメタデータが集まる。

#### ランディングページと責務



ランディングページには、**オブジェクト のメタデータとアクセス手段**を明示。

#### メタデータ検索



https://search.datacite.org/works?query= japanese++literature

- DOIの上流に集まってきたメタデータを対象に、検索システムを構築できる。
- JaLCに集まったメタ データ → RDF/XML形 式の「JaLCメタデー 対して、一括ダ ウンロード可能。

## DOIをどうつけるか?

#### DOIに関する典型的な疑問

- 1. どの粒度で付与するか?
- 2. DOIは「信頼の証」か?
- 3. DOIの重複はよいのか?
- •「書籍」や「論文」が比較的簡単だった のは、編集済み知的生産物だったから。
- データやモノ(アーカイブ資料も含む) は編集前の生の状態のため、それを整理 するには複数の区切り方がありうる。

## (1) どの粒度で付与するか?

データやモノそのものに、固有かつ 唯一の区切り方は存在しない。

- 1. システムIDの単位で考える。
- 2. ランディングページやメタデータ の単位で考える。
- 3. 引用の単位で考える。
- 4. 更新 / 再現性の単位で考える。

#### 1. システムIDの単位

- たいていのデータベースに既に存在する システムIDを流用する方式。
- •システムIDの目的、存続期間によっては、 DOIの識別子に適さない場合がある。
- •システムIDを使うなら、永続的なIDであることを運用上も保証すべきである。
- システムIDの改修はシステムの核心に関 わるため、そう簡単でないこともある。

#### 2. ランディングページの単位

- DOIがレゾルブされた後に到達するページの維持が、DOIを付与した者の責務。
- どんな単位でメタデータを付与したいか、 という問題に相当する。
- 単位が細かすぎると、適切なメタデータを付与できなくなる。
- 単位が粗すぎると、何のデータなのかが わからなくなる。

#### 3. 引用の単位

- DOIを付与する目的は、研究に用いた資料を引用で明示するため。
- 引用に使いやすい単位でDOIを付与する ことが、引用を通した評価につながる。
- 単位が細かすぎると、引用には多くの DOIを列挙する必要が生じる。
- 単位が粗すぎると、研究に関係する部分を特定することが難しくなる。

#### 4. 更新 / 再現性の単位

- DOI付与の一つの目的は、同一データを 用いて他者が研究成果を再現するため。
- 再現性重視派は、データの中身の更新があればDOIも更新すべきと主張する。
- 引用重視派は、データの対象は同じなの だからDOIの更新は必要ないと考える。
- バージョン管理、随時更新データ等の扱いは、運用コストも考えて決定すべき。

### (2) DOIは「信頼の証」か?

- DOIはリポジトリのガバナンスを保証するが、データの中身には関係ない。
- リポジトリが万が一閉鎖される際は、移 転先を確保しレゾルブ先を変更するべき。
- データ品質の審査 / 査読を、DOI付与の 条件とするリポジトリもある。
- メタデータ品質は、リポジトリのキュレーションにも依存する。

## (3) DOIの重複はよいのか?

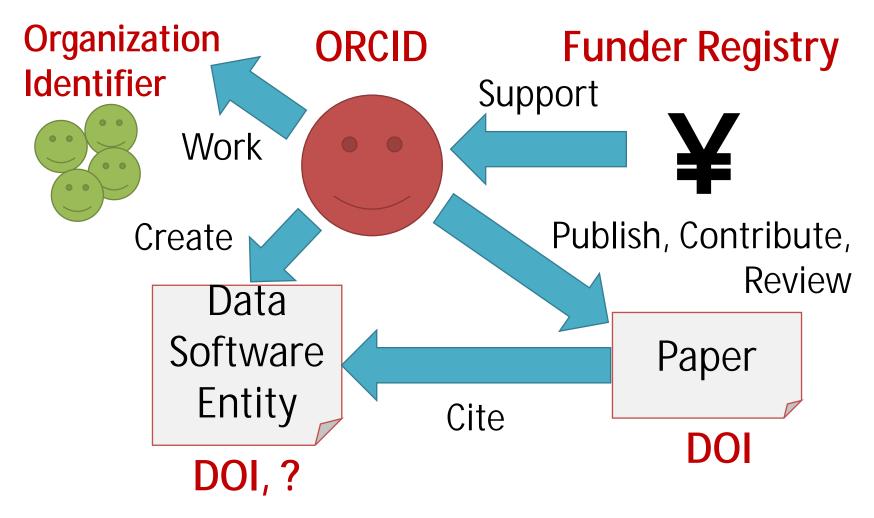
- 同一リポジトリ内で、同一オブジェクト に複数DOIを付与することは禁止すべき。
- 複数リポジトリに同一オブジェクトのコピーが存在する場合、複数DOIの付与を防ぐための技術的手段はない。
- •関係者で調整の上、最適な機関がDOIを 付与し、他者は共有することが望ましい。
- ●将来はDOIが分散しても後から統合する サービスが生まれる?

#### DOIに関するその他の注意点

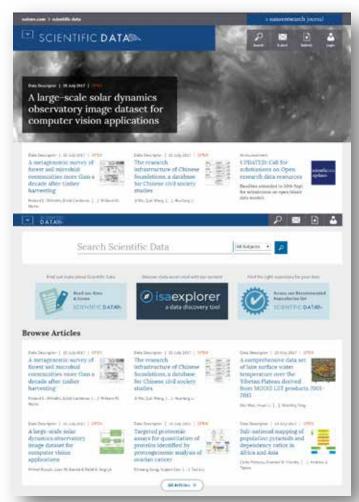
- Suffixに「意味」を与えるなら、破綻せずに永続的な維持が可能か検討すべき。
- DOIは階層構造を持てない。 DOIは独立しており、複数DOIの関係は定義できない。
- DOIは一度登録したら消去できないのが 原則。安易な登録は避けるべき。
- ランディングページはオープンアクセス。 本体へのアクセスはペイウォール等が あってもよい(例:論文データベース)。

## DOIをどう使うか?

#### グローバルな識別子の普及



#### 研究データの重視

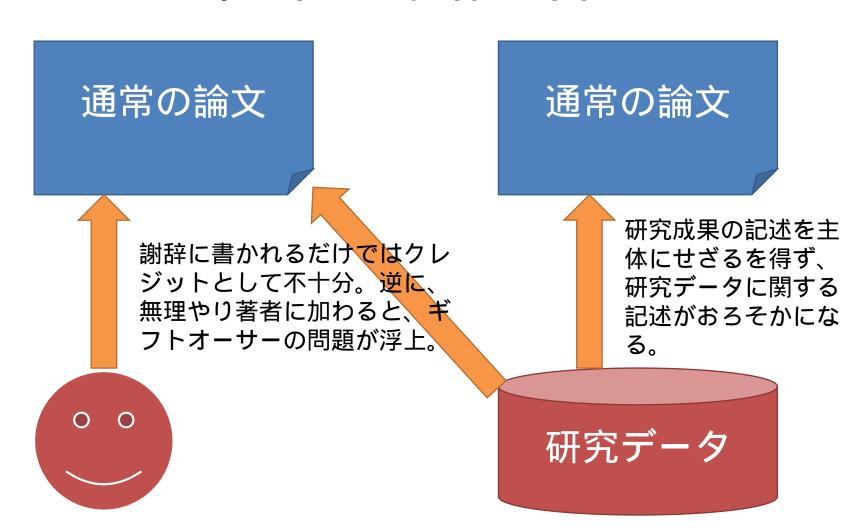


https://www.nature.com/sdata/

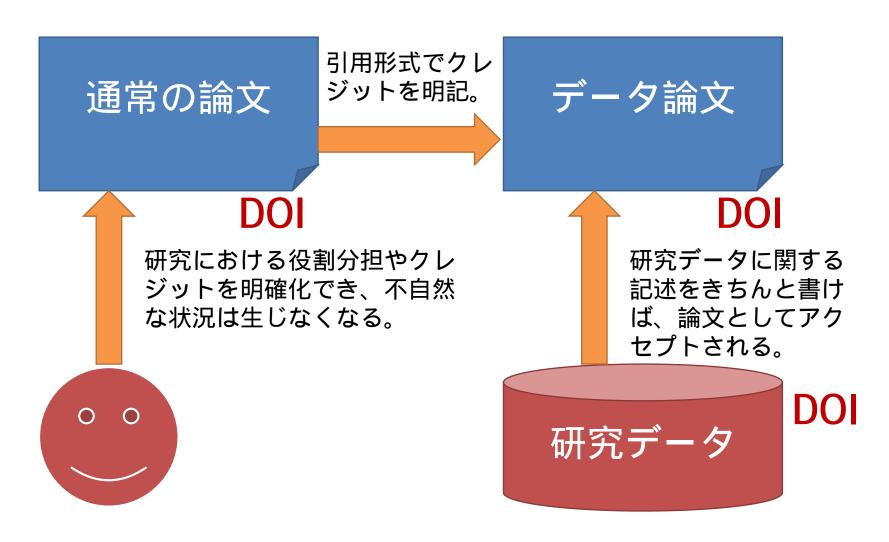
- 研究論文の根拠となるデータも、紐付けてオープン化すべき。
- 研究資源となるデータは、FAIRデータとして管理すべき。
- データ論文により、 きちんと引用して評価すべき。

25

#### 従来の学術出版



#### データ論文を含む学術出版



#### データジャーナル投稿手順

- データジャーナルごとに投稿規程が存在し、データをデポジットできるリポジトリの条件が示されている。
- 2. 有名リポジトリは条件を満たしているが、新規のリポジトリが条件を満たしているかは確認が必要。
- 3. 新規のリポジトリはジャーナル編集者 等と相談しながら、必要があれば条件 を満たすように機能を改良する。

#### Earth System Science Data Journal

https://www.earth-system-science-data.net/

- Persistent identifier: The data sets have to have a digital object identifier (DOI).
- 2. Open access: The data sets have to be available free of charge and without any barriers except a usual registration to get a login free-of-charge.
- 3. Liberal copyright: Anyone must be free to copy, distribute, transmit, and adapt the data sets as long as he/she gives credit to the original authors (equivalent to the Creative Commons Attribution License).
- 4. Long-term availability: The repository has to meet the highest standards to guarantee long-term availability of the data sets and permanent access.

#### リポジトリの種類

#### 汎用リポジトリ

データの種類 や由来を問わず、広ぐデポ がいた受け付しる。



#### 機関リポジトリ

機関が生み出した成果物を、 一般的な標準 を適用し、広く発信する。



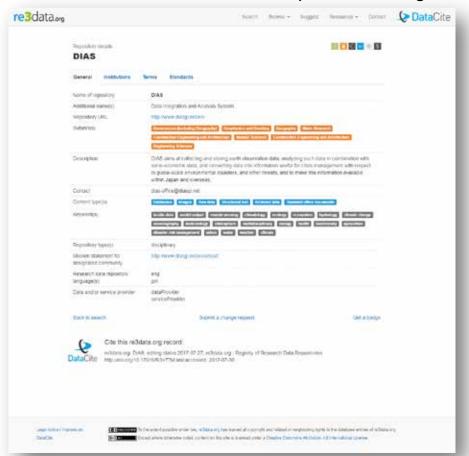
#### 分野リポジトリ

分野のコミュ ニティが、自 らの標準を適 用し、成果を 共有する。



#### レジストリへの登録

https://doi.org/10.17616/R3V77M



- Re3data: Registry of Research Data
- Repositories

- 研究データリポジト リの概要をメタデー タ化して登録。
- リポジトリの信頼性 の判断を助ける情報 を提供(認証は別)。
- データジャーナル等 への投稿には、所定 の条件を満たすリポ ジトリの利用が必須。

#### データ引用支援



http://dsr.nii.ac.jp/toyobunko/VIII-5-B2-7/

- DOIつきのデータ引用サンプルを表示。
- The Citation Style
   Languageを使って
   ジャーナル等に合
   わせた自動整形。
- ウェブサービス有 https://citation.cross cite.org/

32

# 識別子の実例と研究分野の固有性

#### DOIシステムの全体像



#### DIASEDOI

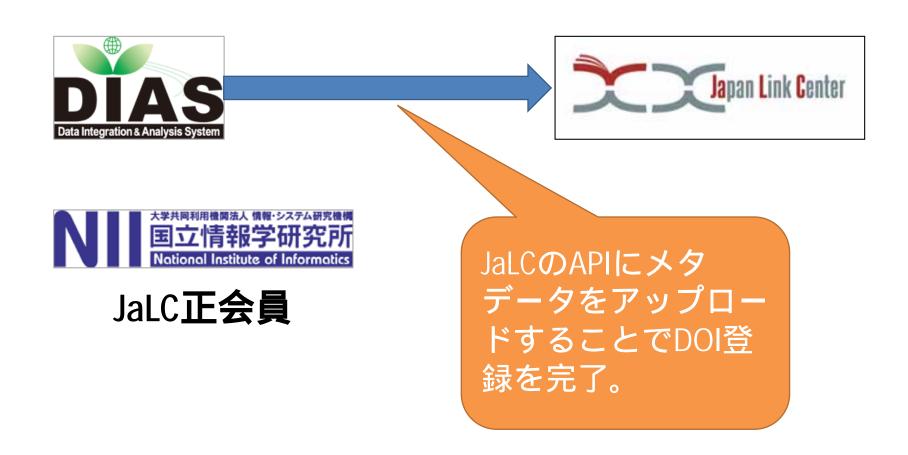
https://dias.ex.nii.ac.jp/



http://www.diasjp.net/infomation/press-release-dias-first-doi-registration/

http://search.diasjp.net/en/dataset/GAME\_ Tibet

#### JaLC正会員方式



### 東洋文庫アーカイブとDOI

http://dsr.nii.ac.jp/toyobunko/

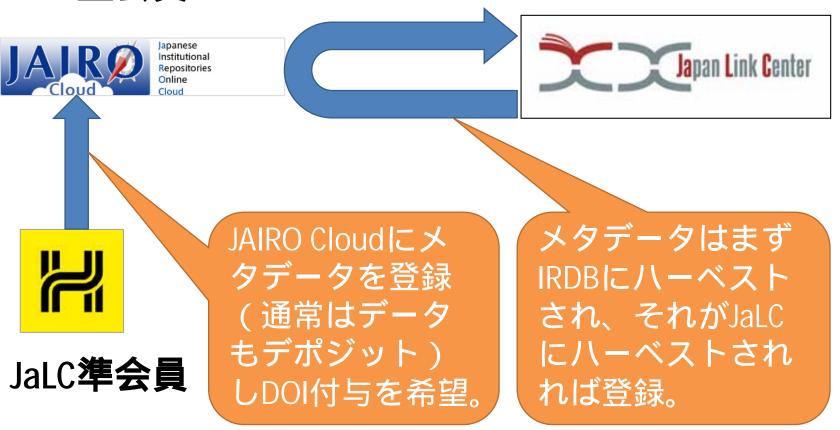


DOIについては、NIIが 提供するクラウド型機 関リポジトリ環境提供 サービスJAIRO Cloud ト に「人文学研究データ リポジトリ」を開設し、 ここにコンテンツのメ タデータ134件を入力 してDOIを取得してい ます。

http://current.ndl.go.jp/node/34507

### JaLC準会員方式





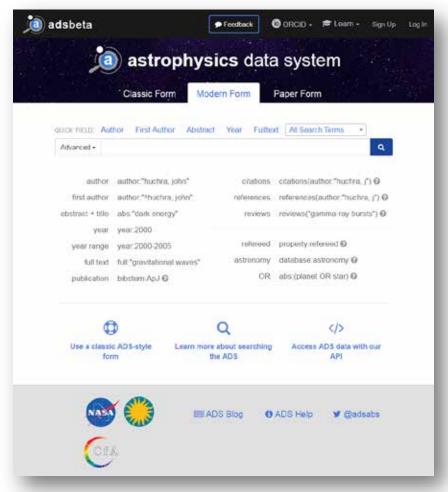
# データDOIの種類

- DIASではDataCite DOIを利用(英語)。
- CODHではJaLC DOIを利用(日本語)。
- JaLC正会員方式では、最初にどちらでも 選べる(ただし後からの変更は困難)。
- JAIRO Cloud方式ではJaLC DOI しか選べない(今のところ?)
- DOIの種類 = メタデータの種類。選んだ DOIに合わせてメタデータを作成する。

# 識別子とドメイン固有性

- DOIのRAが定めたメタデータ形式に不満が生じた場合はどうするか?
- ドメインごとの独自メタデータ形式と固 有IDをグローバル展開する道もある。
- DataCiteの公認ID: bibcode(天文学)、IGSN(地質学)、LSID(生命科学)など。
- ・識別子は単なる文字列ではなく、永続的な運営を考慮した社会的システムである。

# 天文学の例



SIMBAD Astronomical Database Join the LISA VIII (Library and Information Services in Astronomy) conference in Strasbourg June 2017 What is SIMEAD ? Documentation Information Queries User's guide basic scerch by identifier by coordinates Image thumbnals by triteria Query by urb. reference quecy Nomenciabure Dictionary Object types scripts TAP queries List of journals SerWatch D Measurement description options Spectral type coding SIMBADA 1.5.11 - Feb-2017 User annotations documentation Release history Display all user annotations **Acknowledgment** Content Basic search The SIMBAD astronomical database provides basic data, cross-identifications, bibliography and measurements for astronomical objects outside the solar system. SIMBAD can be queried by object name, coordinates and identifier, coordinates (radiusetti accron), or bibcode various criteria. Lists of objects and scripts can be SIMBAD search Links to some other on line services are also provided. Install the Smbad basic search in your tool bar Acknowledgment Simbad contains on 2017.05.29 If the Simbad database was helpful for your research work, the following acknowledgment would be appreciated: 9,209,417 objects 24,790,606 Identifiers This research has made use of the SIMBAD database, 331,128 bibliographic references operated at CDS, Strasbourg, France 15,791,761 citations of objects in papers 2000,A&A5,143,9, "The SIMBAD astronomical database". Wenger et al.

Purtal Simbad VicieR Alade X-Match Other- Help

https://ui.adsabs.harvard.edu/

2017/08/26

http://simbad.u-strasbg.fr/simbad/

41



# International Geo Sample Number (IGSN)の例

- SESAR (System for Earth Sample Registration) が管理する地質標本番号。
- IGSN:HRV003M16 は以下のURLでレゾルブ可能。http://igsn.org/HRV003M16
  https://doi.org/10273/HRV003M16
  http://hdl.handle.net/10273/HRV003M16
- •標本が現実空間を移動しても同一識別子。
- ドメインが定めるメタデータ形式を利用。

# 永続的識別子の選択

# 識別子 = エコシステム

- どの識別子を使うか? = どのエコシ ステムを使うか?
- DOIはエコシステムが確立しているので、その上に乗ることができるという利点がある。
- 識別子システム = 運営組織。国際 DOI財団 (IDF) に相当する組織を、 独自に立ち上げられるか?

### 識別子の選択

- システムに閉じた識別子なら、システム 内でのエコシステムの充実度で選択。
- グローバルに開かれた識別子なら、それを支えるエコシステムの充実度で選択。
- ・識別子の粒度の選択は識別子の選択と (ほぼ)直交するため、独立に検討可能。
- 識別子の複合も可能。例えばSuffixをARK とするDOIも作れる。

### ARKとDOIの比較

- ARK (Archival Resource Key) も永続的識別子 としての機能を有する。
- N2t.net (Resolver Name->Things)やその他のNMAHがレゾルバとして使える。
- Ark:/NAAN/Name[Qualifier]という構造は、NAAN=Prefix・Name=Suffixに対応する。
- QualifierはDOIに対応するものがないので、 興味深い(後述)。

# DOIIは階層を表現できない

- DOIの識別子はすべて独立しており、階層的な表現はできない。
- •書籍のDOIが登録してあっても、その中の特定ページを指すことはできない。
- 階層を導入するにはDOIの根本的な作り 直しが必要であり、ほぼ不可能。
- 細かい粒度の識別子が欲しいことがよく あるが、今のところ解決策なし。

#### **ARK Qualifier**

https://tools.ietf.org/html/draft-kunze-ark-18

- http://example.org/ark:/12025/654xz321/s3/f8.05v.tiff というarkでは、"/s3/f8"はComponentPath、そして".05v.tiff"はVariantPathとなる。
- アイデアは理解できるが、この仕組みは どのように実現するのか?NMAとNAAで 多段レゾルバを実現するのか?
- ●識別子は永続性が重要なため、複雑な仕組みは維持コストを上昇させる。

## 面倒そうな識別子仕様

- http://example.org/ark:/12025/654xz321/s3/f8.05v.tiff というURL構造は、サーバ実装に制約を課す仕様。
- http://bnf.fr/ark:/13030/tf5p30086k の移転 先を以下の形式とするのは http://portico.org/ark:/13030/tf5p30086k ローカルなレゾルバを置くことが前提?
- レゾルバの維持は面倒なので、グローバルなレゾルバにお任せしたいのが本音。

# レゾルバの不審な挙動

http://gallica.bnf.fr/ark:/12148/bpt6k107371
 t

http://n2t.net/ark:/12148/bpt6k107371t の レゾルブ結果

http://visualiseur.bnf.fr/ark:/12148/bpt6k107371tが異なる。

- これで正しい動作なのか?DOIではこういう不整合は起こらないが。。
- Actionableな識別子として問題ありそう。

# レゾルバのリダイレクト

- DOIは「ランディングページ」にリダイレクトし、そこでメタデータを表示し、 そこにデータのアクセス方法を明示する。
- ARKはレゾルバがメタデータを表示する 機能をもち(inflection)、リダイレクト 先は特に決まっていない(?)
- 実利用のパターンからすれば、DOIの方が納得感がある。ARKは微妙。

### 理想主義が複雑化を招いた?

- ARKの基本思想は、NMAがつぶれても大 丈夫な分散システムを作ること。
- DOIはIDFが潰れたら終わりだが、ARKなら代役を立てられるから大丈夫!?
- DOIが潰れる兆候はなく、むしろ分散システムの複雑さが、ARKには逆効果?
- DOIは国際標準だが、ARKは国際標準ではない。むしろARKの方が心配?

### 永続的識別子の選択

http://www.ncdd.nl/en/pid/

- PID = Persistent Identifier: すべての永続的 識別子を包含する概念。
- PID Guide:自分のプロジェクトに適した PIDを選択式で選べるツール。
- Various methods of implementing Persistent Identifiers exist. We recommend using one of the three international standards applied in the Netherlands: URN:NBN, DOI or Handle.

# おわりに

### 人文学におけるDOIの活用

- 1. 研究の出力へのDOI付与:論文・書籍などを特定可能とする。
- 2. 研究の入力へのDOI付与:データ・資料 などを特定可能とする。
- 3. 実体への識別子付与:世界に存在するオブジェクトを特定可能とする。
- 4. DOI以外の可能性もあるが、グローバル に通用する識別子の立ち上げは大仕事。

### まとめ

- 1. DOIとは何かについて、DOIの仕組みと ランディングページの重要性を述べた。
- 2. DOIをどうつけるかについて、典型的な 疑問に答える形で目安を示した。
- 3. DOIをどう使うかについて、研究基盤のオープン化の観点から説明した。
- 4. 研究の現代化に識別子は必須であり、 ドメイン全体で取り組む必要がある。

# 関連情報

- ・研究データへのDOI付与
  - <a href="https://dias.ex.nii.ac.jp/doi/">https://dias.ex.nii.ac.jp/doi/</a>
- オープンサイエンス
  - http://agora.ex.nii.ac.jp/~kitamoto/research/openscience/
- ・第3回CODHセミナー人文学でのDOI活用 ~研究 データや所蔵品など研究資源へのDOI付与 ~
  - http://codh.rois.ac.jp/seminar/doi-for-humanities-data-20170530/
- 本発表のスライドも公開します。