DOI概論－研究基盤のオープン化に向けたIDの活用

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

CODHセミナー

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

https://codh.repo.nii.ac.jp/

- **人文情報学のトピック**に関して、専門家が様々な視点から議論。
- 運営を省力化し、**資料はDOI付きでリポジトリ公開**。
- **情報共有の場**として活用したい。

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／国文研の共同研究などが進行中。

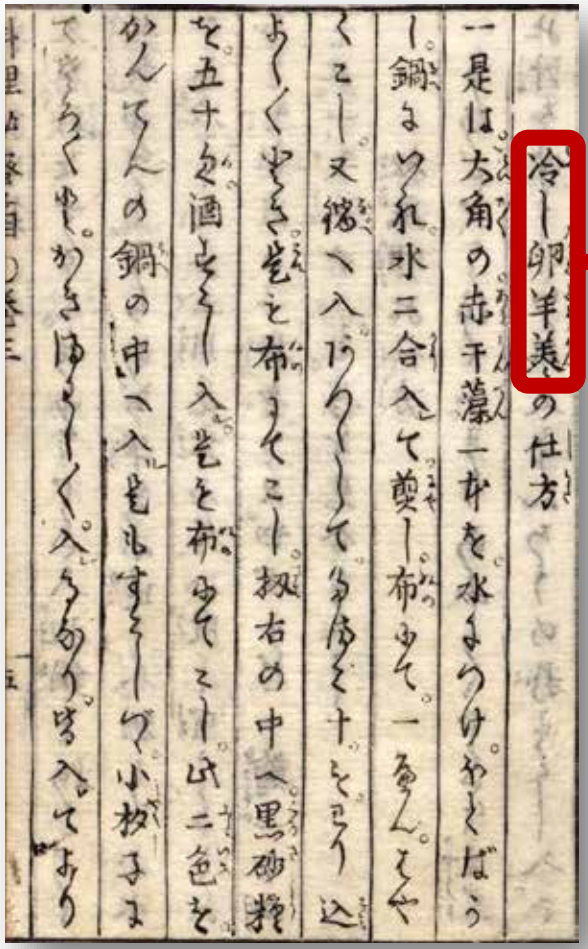
研究者のためのオープンデータ

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

日本古典籍データセット（国文研所蔵）

市民のためのオープンデータ

http://codh.rois.ac.jp/edo-cooking/



日本古典籍データセット
（国文研所蔵）

江戸料理レシピデータセット
（CODH制作）
日本古典籍データセット
（国文研所蔵）を翻案

くずし字チャレンジ！

http://codh.rois.ac.jp/char-shape/

- **人工知能**を使って、コンピュータに文字を読ませたい。
- **文字認識（OCR）**は、現代の印刷文字だと、かなりできるが…
- くずし字はまだまだ難しい。**みんなでチャレンジ**しよう！

DOI（Digital Object Identifier）とは？

識別子としてのDOI

10.20676/00000001

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。

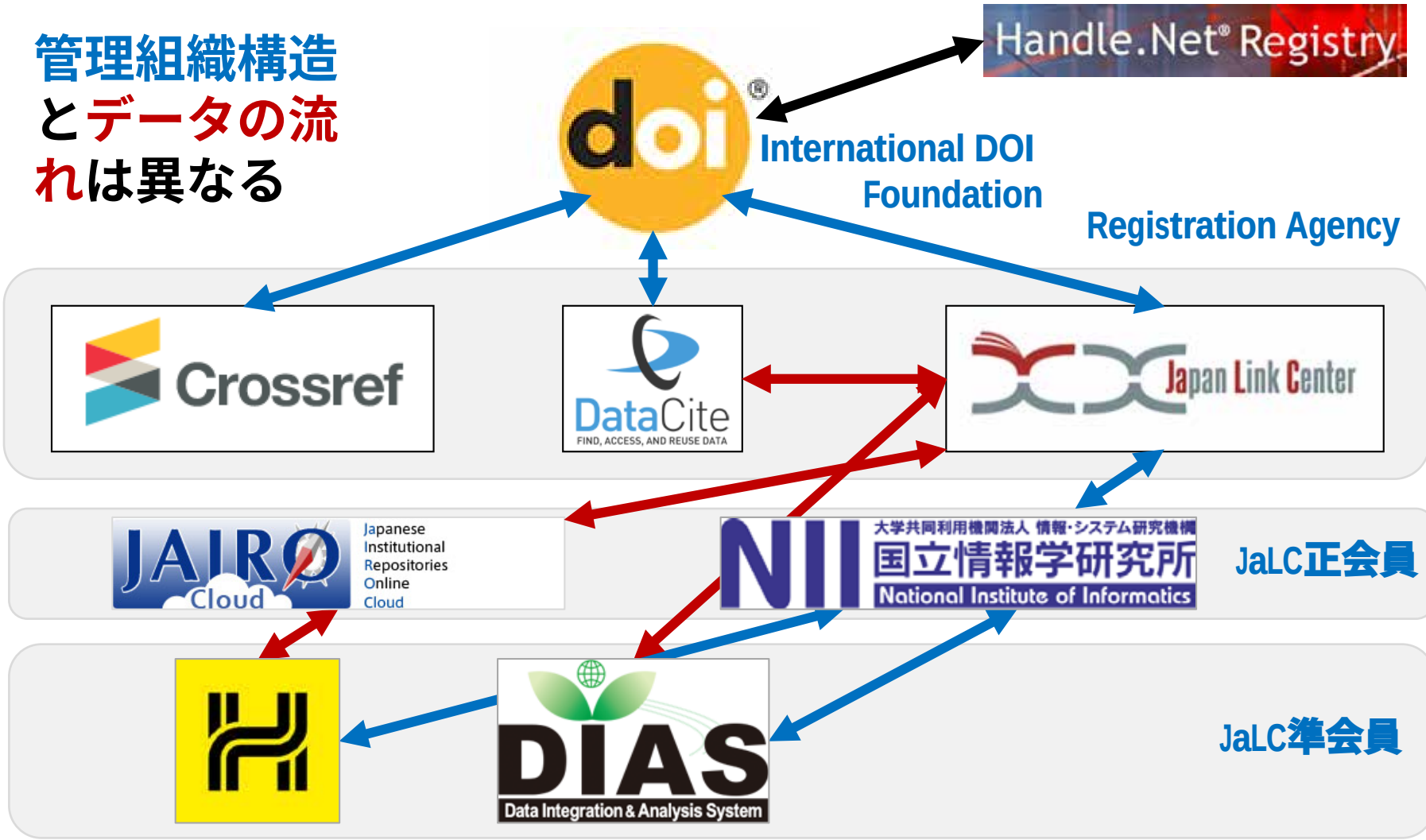
- **単純化すれば、たったこれだけ！**
- **単純に見えて実は奥深く、しかも核心的な機能、それが識別子である。**

DOIの仕組み

DOI = Prefix/Suffix

1. 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 → 意味を持たせる流儀と、持たせない流儀が混在する。
3. レゾルバの運用 → https://doi.org/<DOI> はランディングページのURLに自動変換（Handleシステムの機能も利用）。

DOIシステムの全体像



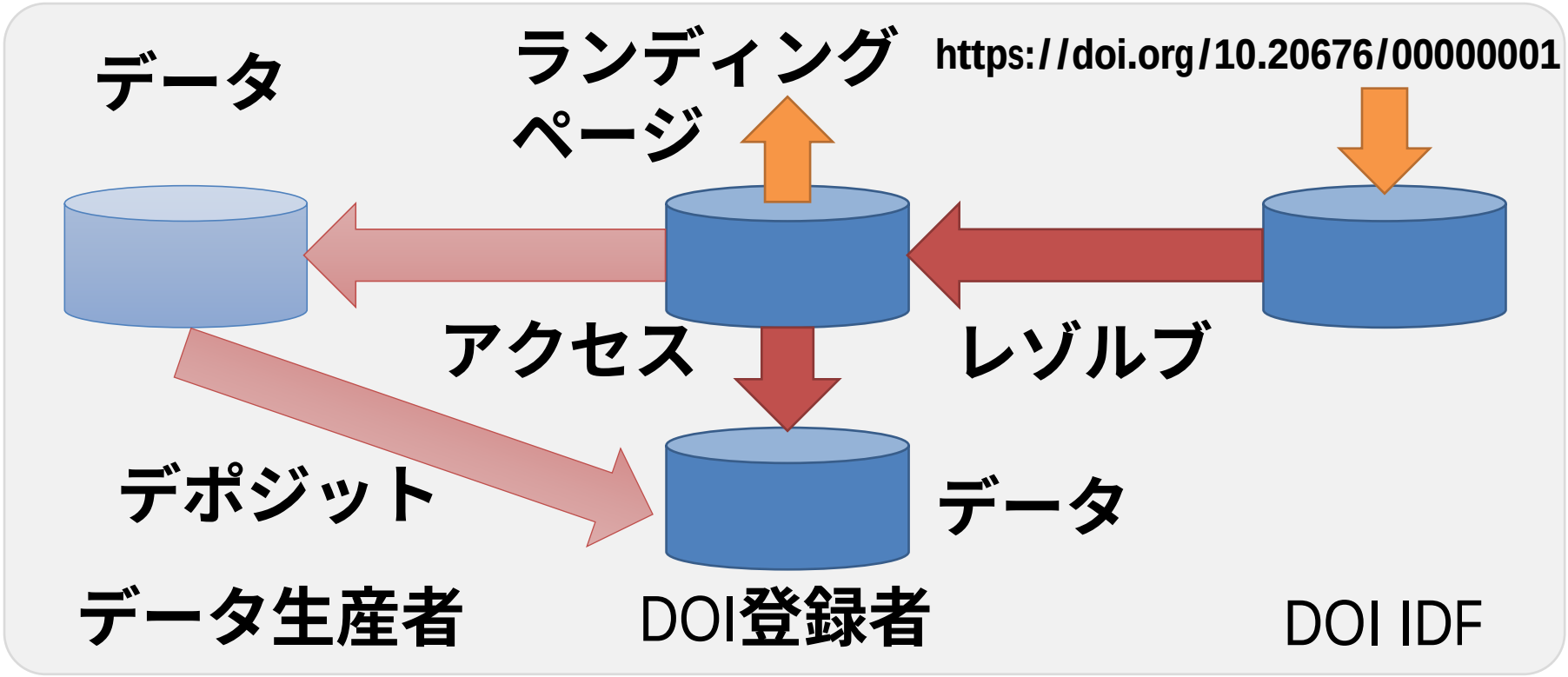
DOIレゾルバ

https://doi.org/10.20676/00000001

レゾルブに失敗してしまった…

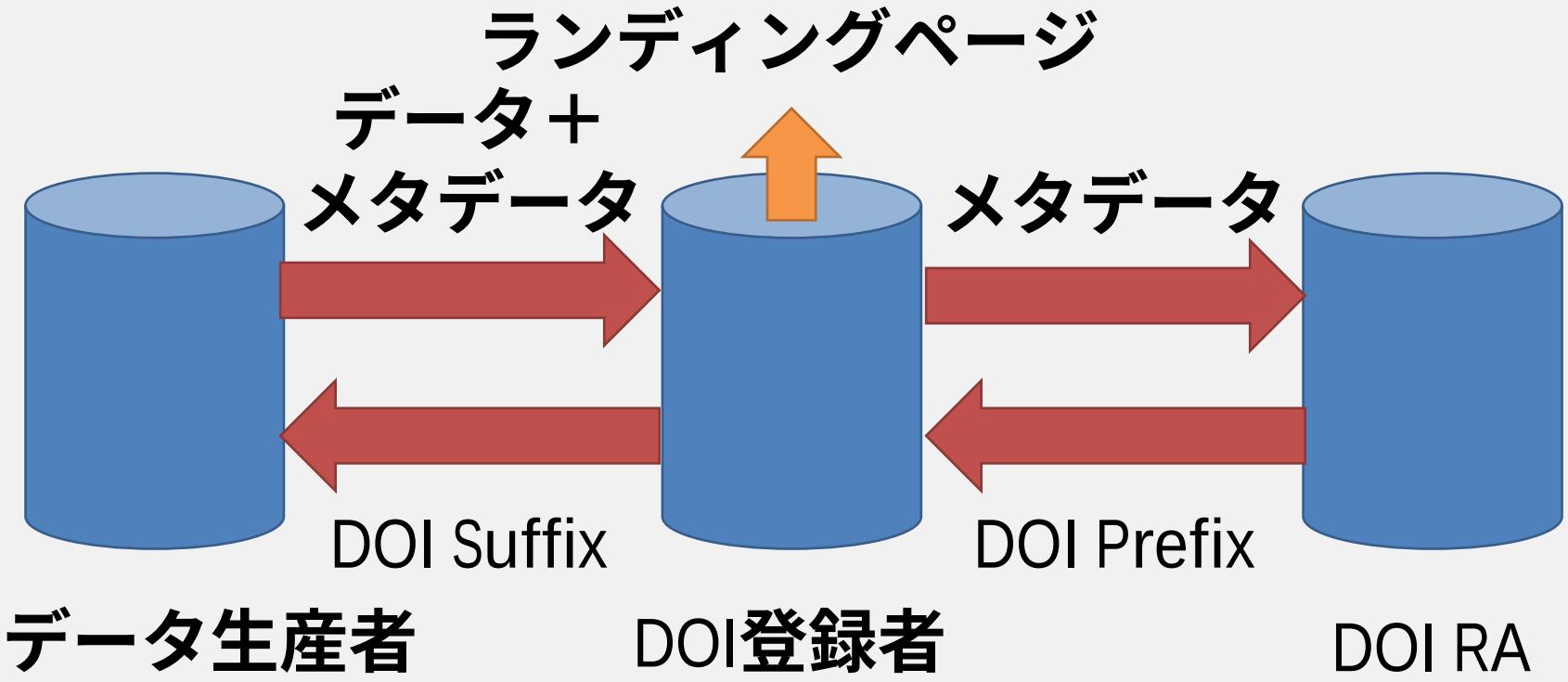
（注）明日には成功する予定

ランディングページと責務



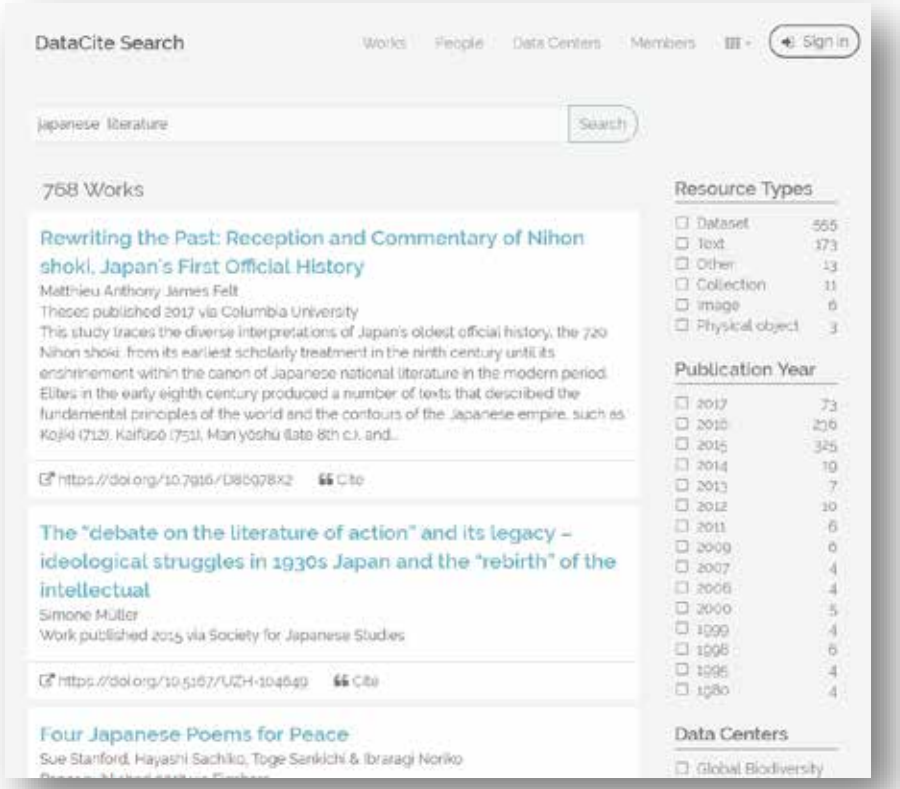
ランディングページには、**オブジェクトのメタデータとアクセス手段**を明示。

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。

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メタデータ」として、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を付与する目的は、**同一データを用いて他者が研究成果を再現する**ため。
- **データ更新は再現性に重大な影響を与える**ため、DOIを更新すべきと考える。
- **オブジェクトとは何か？**修復前後のモノ、過去と現在のモノは「同じ」なのか？
- **バージョン管理**、**随時更新データ**等の扱いは、利用と運用を考えて決定すべき。

(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。
- リポジトリを閉鎖する前に、移転先を確保しレゾルブ先を更新する努力を求める。
- **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？

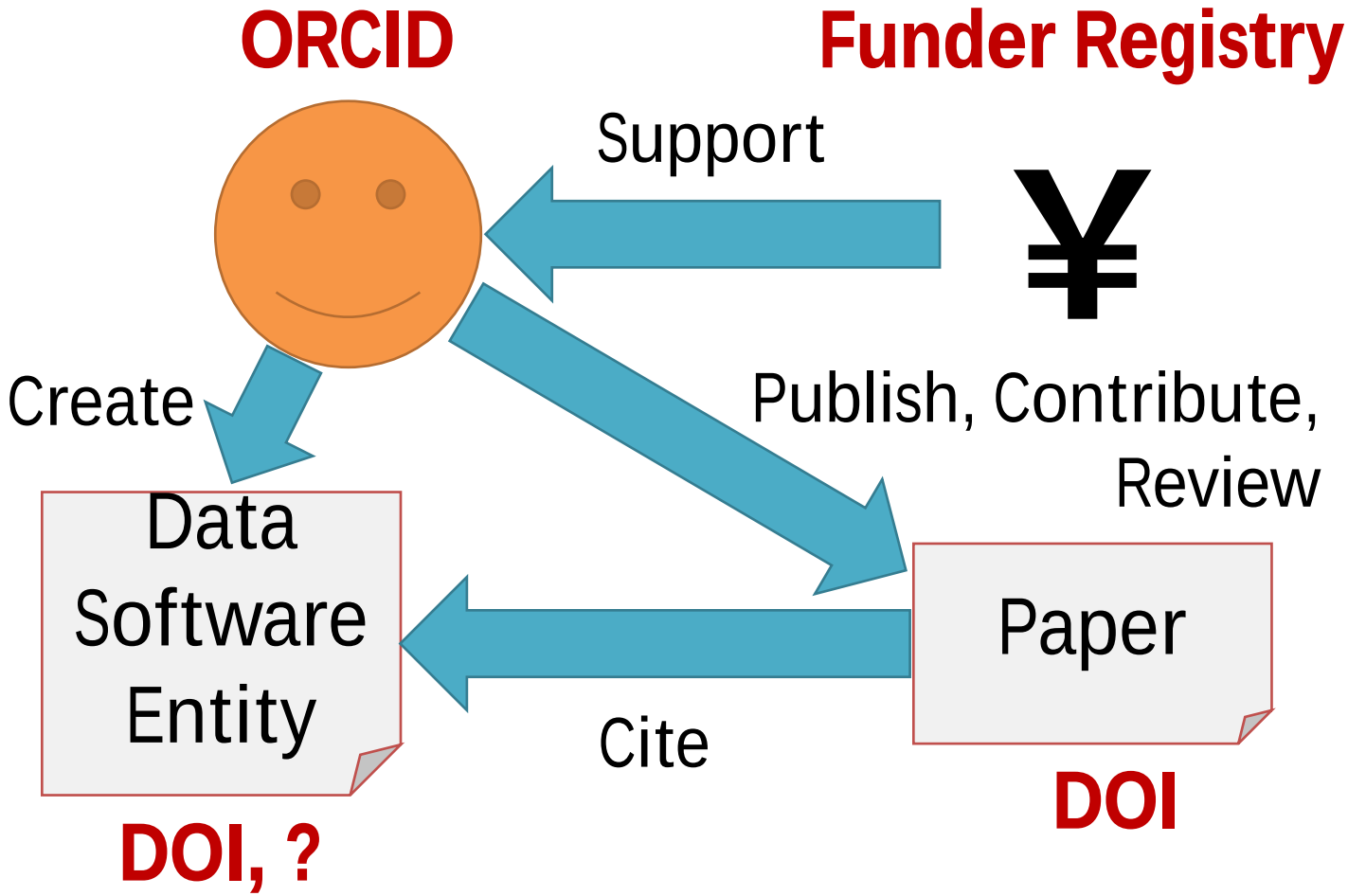
- **同一リポジトリ内**で、同一オブジェクトに複数DOIを付与することは禁止すべき。
- **複数リポジトリ**に同一オブジェクトのコピーが存在する場合、複数DOIの付与を防ぐための**技術的手段はない**。
- 関係者で調整の上、**最適な機関がDOIを付与**し、他者は共有することが望ましい。
- 将来は**DOI統合サービス**が生まれる？

DOIに関するその他の注意点

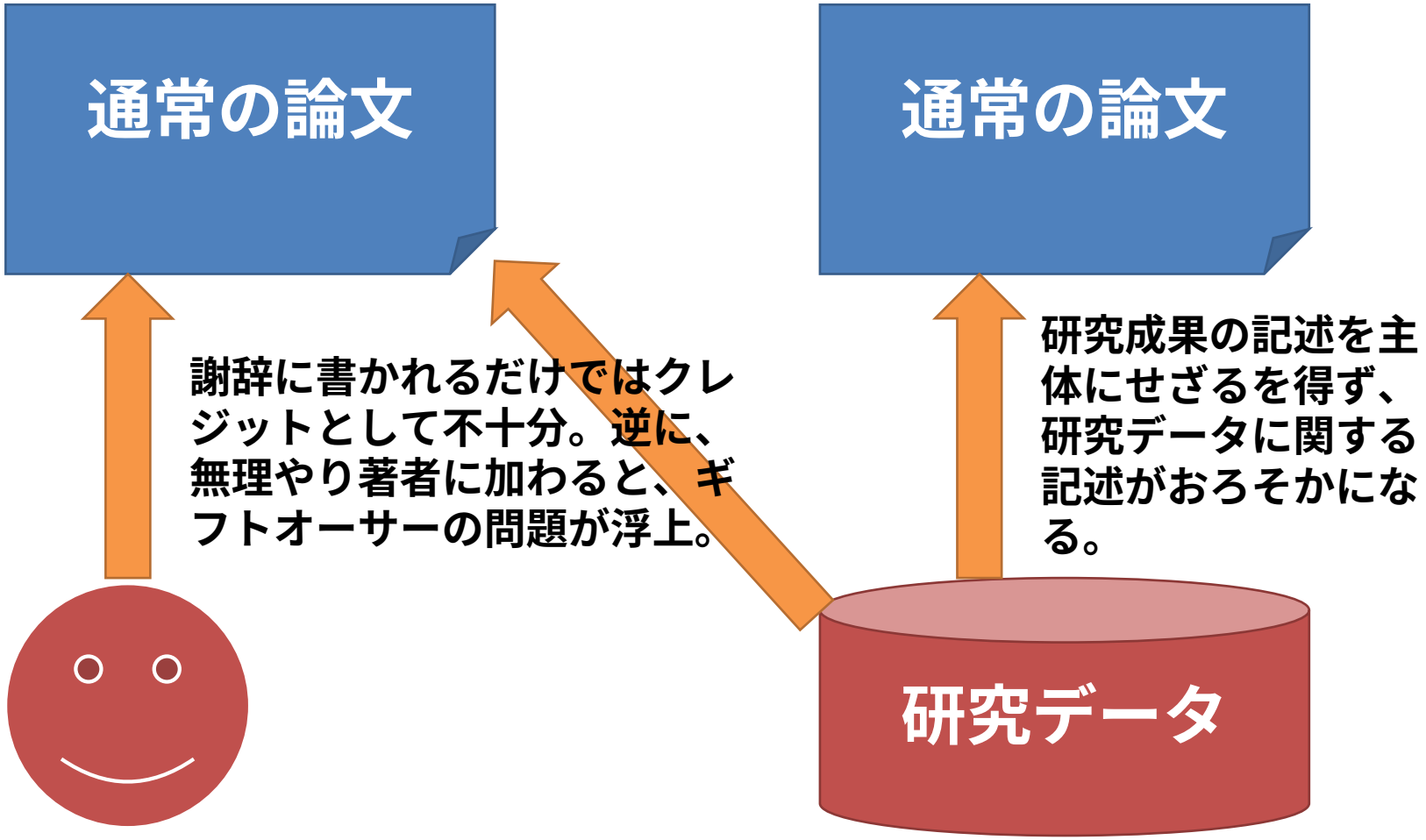
- Suffixに「意味」を与えるなら、破綻せずに永続的な維持が可能か検討すべき。
- DOIは階層構造を持てない。DOIは独立しており、複数DOIの関係は定義できない。
- DOIは一度登録したら消去できないのが原則。安易な登録は避けるべき。
- ランディングページはオープンアクセス。本体はオープンでない場合も許容する。

なぜDOIが必要なのか？

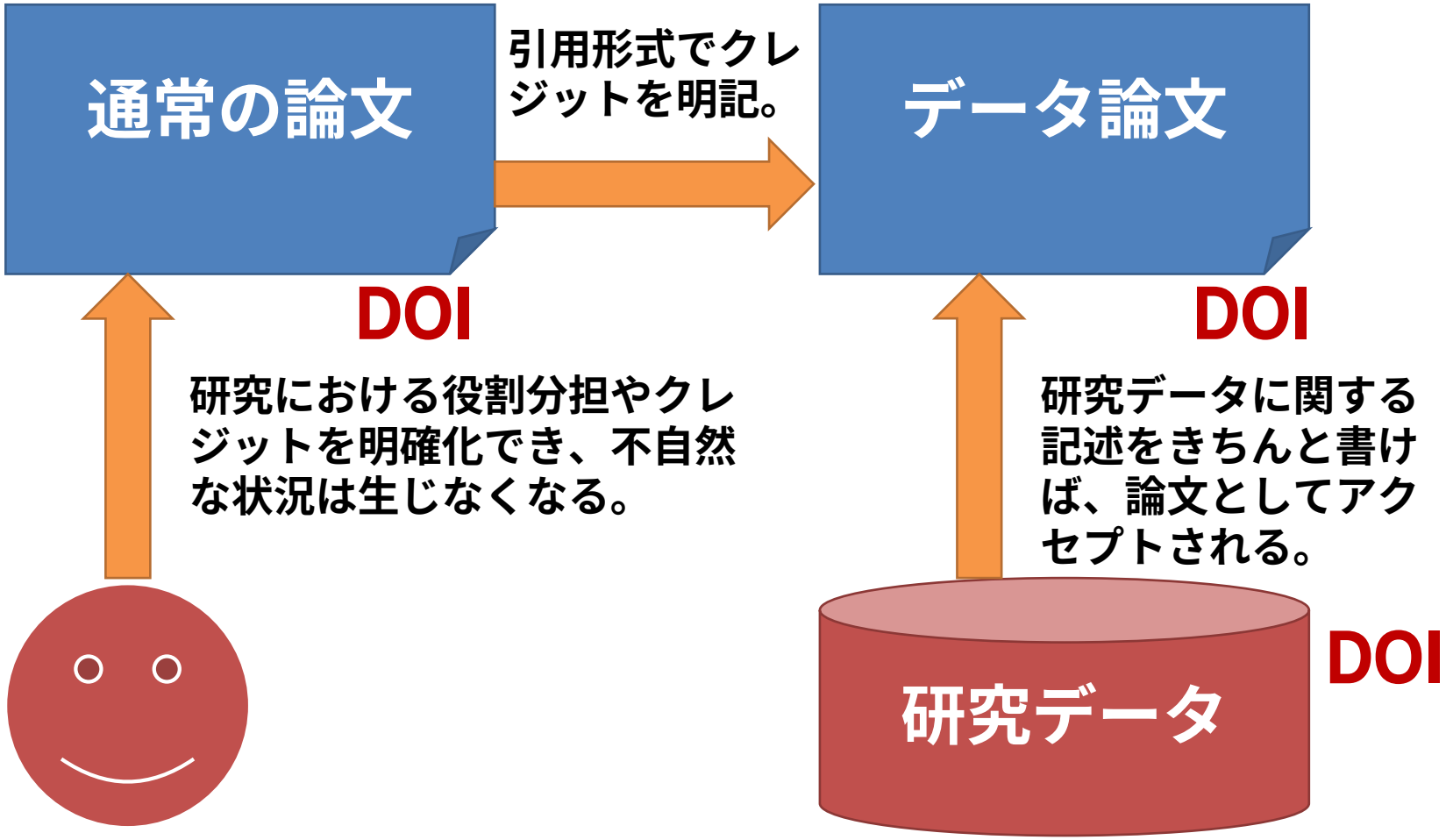
オープンサイエンスとグローバルな識別子



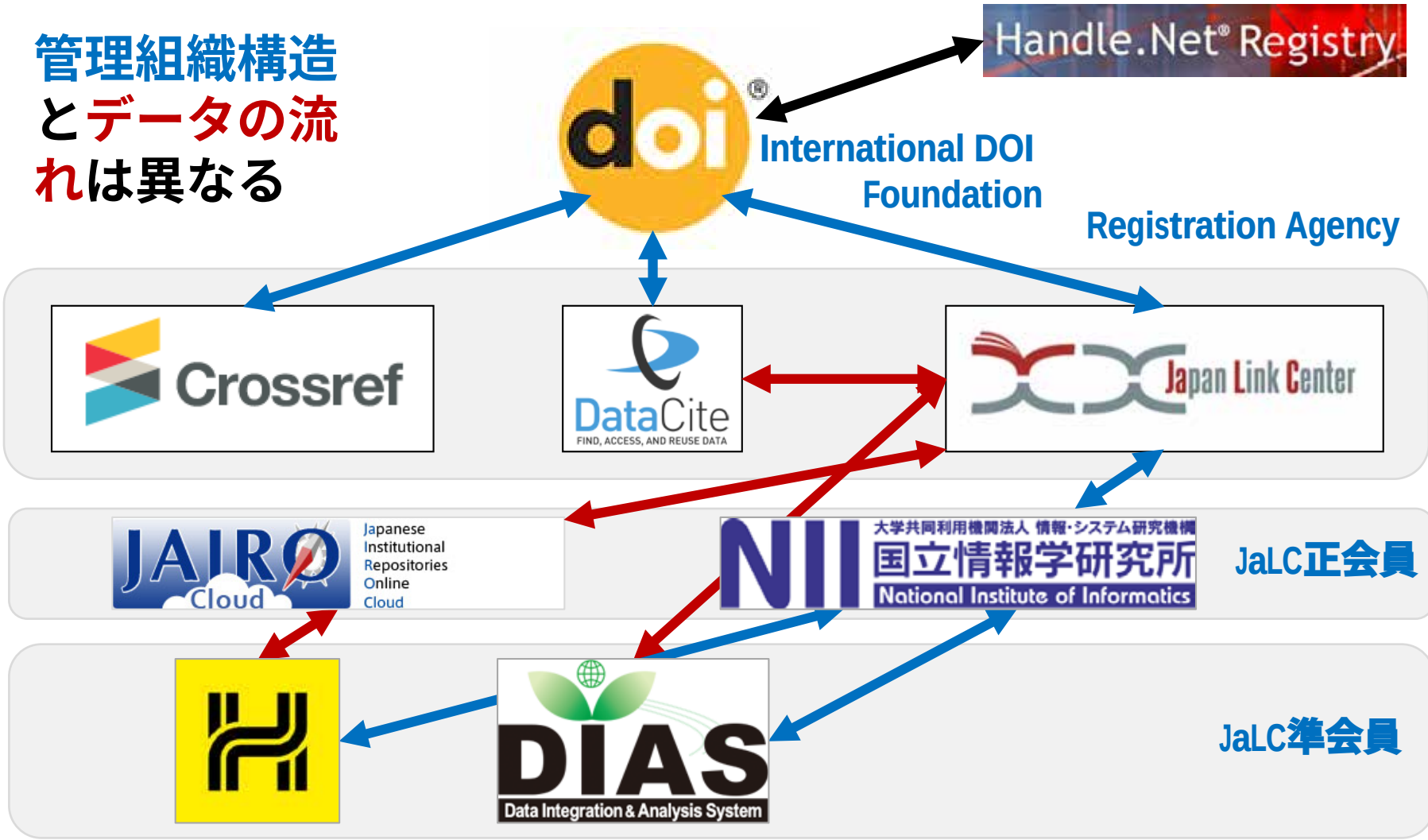
従来の学術出版


通常の論文
通常の論文
謝辞に書かれるだけではクレジットとして不十分。逆に、無理やり著者に加わると、ギフトオーサーの問題が浮上。
研究成果の記述を主体にせざるを得ず、研究データに関する記述がおろそかになる。
研究データ

データ論文を含む学術出版



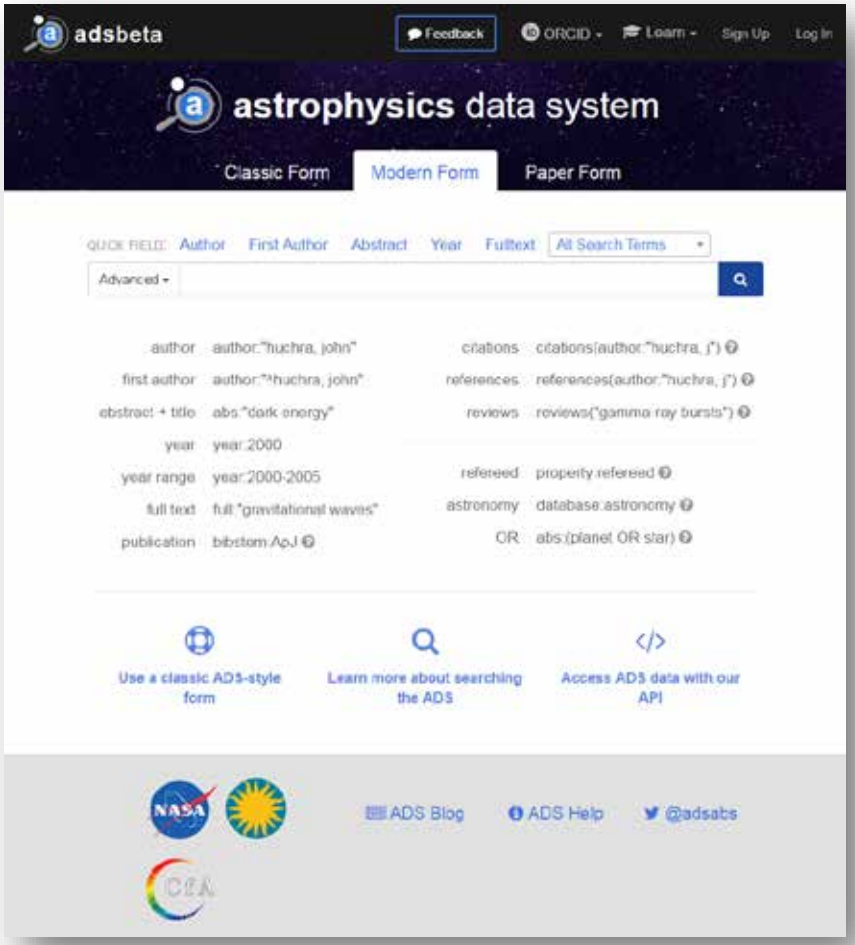
DOIシステムの全体像



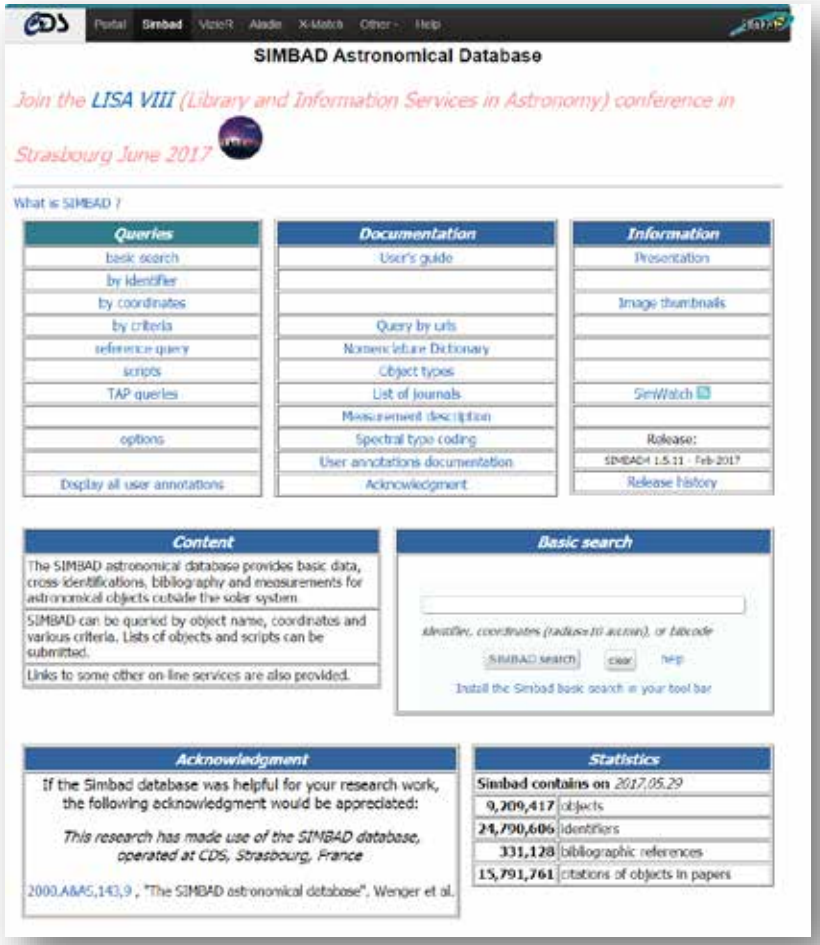
DOIとドメインの固有性

- DOI RAが定めたメタデータ形式に合わせて、メタデータを登録する必要がある。
- 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。
- DOIは単なるDBではなく、永続的な運営体制を考慮した社会的システムである。

天文学の例



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



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

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。**

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. なぜDOIが必要なのかについて、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。