

# 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/>

The screenshot shows the CODH Humanities Research Data Repository interface. The header includes the logo and name 'CODH 人文科学研究データリポジトリ'. The main content area displays a search result for a presentation titled 'Center for Open Data in the Humanities (CODH): Activities and Future Plans'. The DOI 'info:doi/10.20676/00000001' is highlighted with a red box. Other details include the item type '会議発表資料 / Presentation', keywords '人文学オープンデータ活用センター, CODH, オープンデータ, CODHセミナー', and the publication date '2017-01-23'.

<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

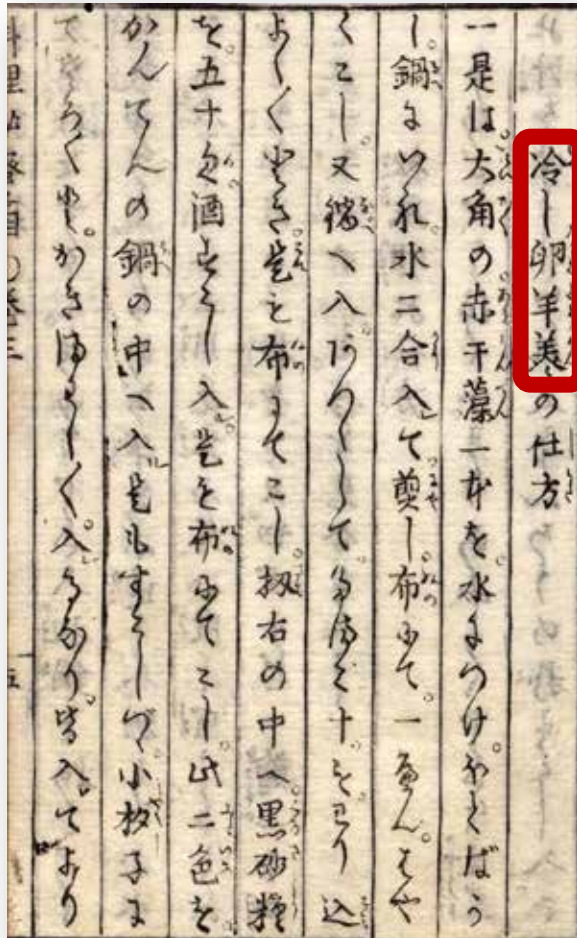
<http://codh.rois.ac.jp/pmjt/>



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/>



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

日本古典籍データセット  
(国文研所蔵)

# くずし字チャレンジ！

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

第 21 回パターン認識・メディア理解研究会  
アルゴリズムコンテスト

この文字読めますか？  
～くずし字認識にチャレンジ！～

古典籍画像の指定領域に含まれるくずし字を認識して、各字の Unicode を出力する課題です。

使用言語：C++ / Python  
開発・実行ができる  
仮想環境を配布します



詳細はアルコン HP にて

サンプル公開：2017 年 4 月末  
応募開始：2017 年 5 月 31 日  
応募締切：2017 年 8 月 31 日



@alcon2017prmu

主催：電子情報通信学会 / パターン認識・メディア理解研究会  
後援：国文学研究資料館  
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  
情報処理学会 / 人文科学とコンピュータ研究会  
データ提供：国文学研究資料館 / 人文学オープンデータ共同利用センター

- **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

10.20676/000000001

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。



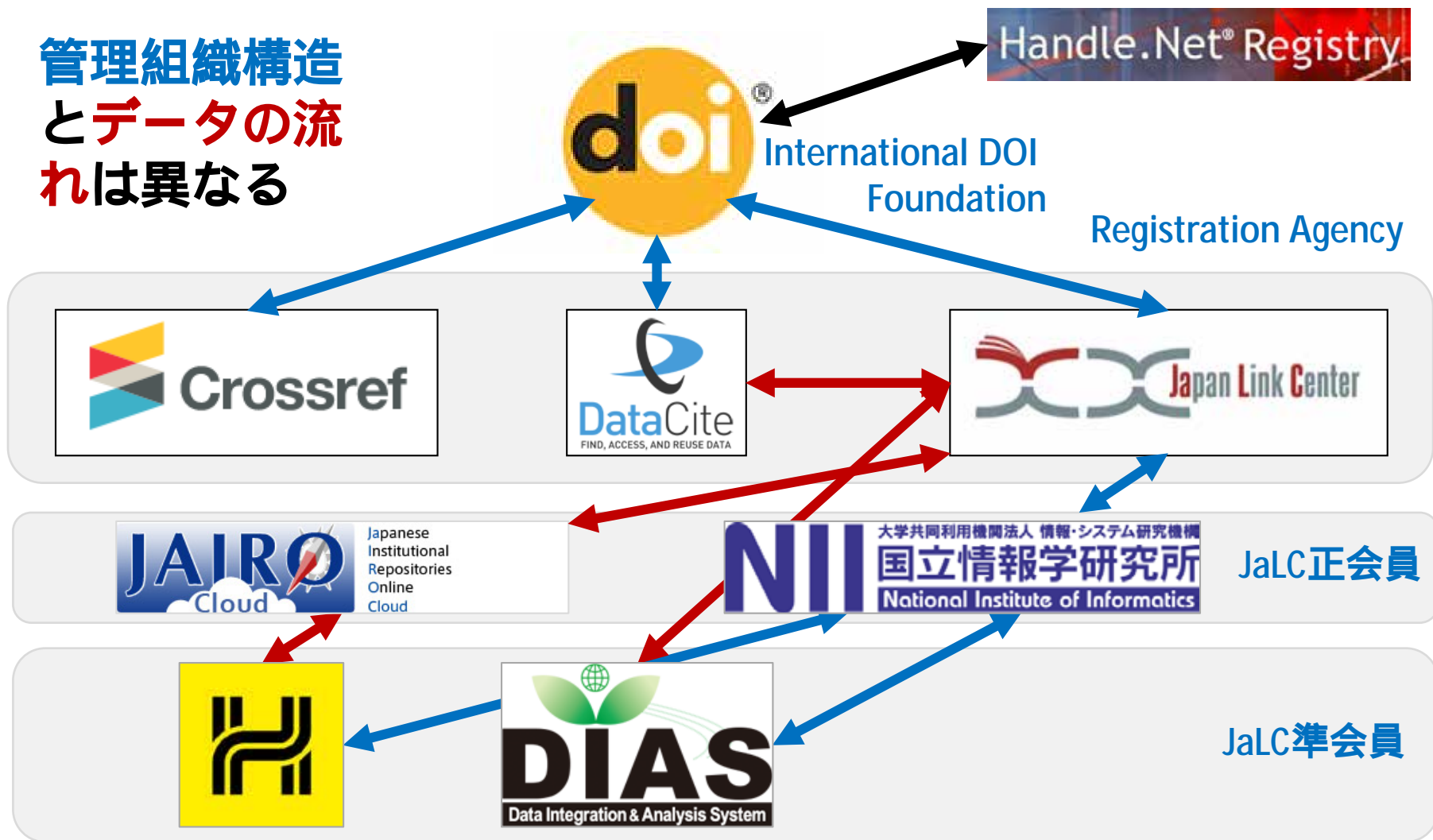
# DOIの仕組み

## 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）。

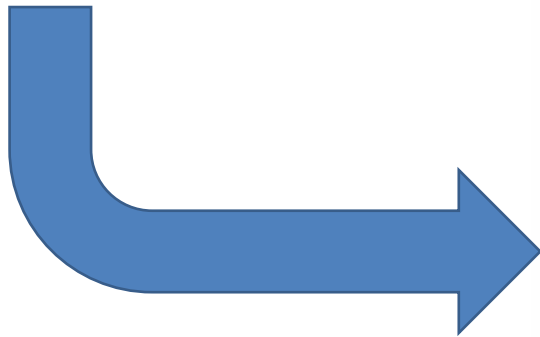
# DOIシステムの全体像

管理組織構造  
とデータの流  
れは異なる



# DOIレゾルバ

https://doi.org/10.20676/000000001



レゾルブに失敗  
してしまっただ...  
(注) 明日には成功する予定

**doi**<sup>®</sup>

HOME | HANDBOOK | FACTSHEETS | FAQs | RESOURCES | USERS | NEWS | MEMBERS AREA

### DOI Not Found

10.20676/000000001

This DOI cannot be found in the DOI System. Possible reasons are:

- The DOI is incorrect in your source. Search for the item by name, title, or other metadata using a search engine.
- The DOI was copied incorrectly. Check to see that the string includes all the characters before and after the slash and no sentence punctuation marks.
- The DOI has not been activated yet. Please try again later, and report the problem if the error continues.

You may report this error to the responsible DOI Registration Agency using the form below. Include your email address to receive confirmation and feedback.

DOI:

URL of Web Page Listing the DOI:

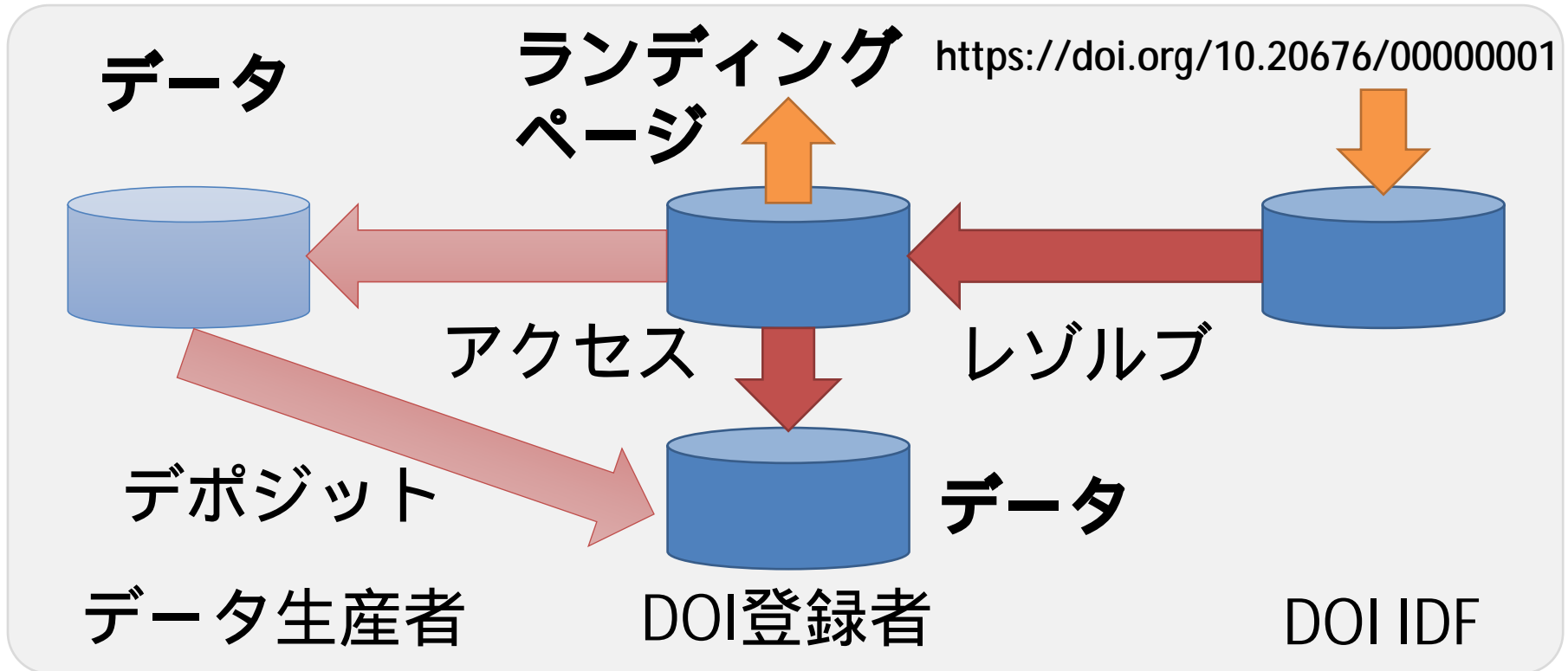
Your Email Address:

Additional Information About the Error:

[DOI System Proxy Server Documentation](#)

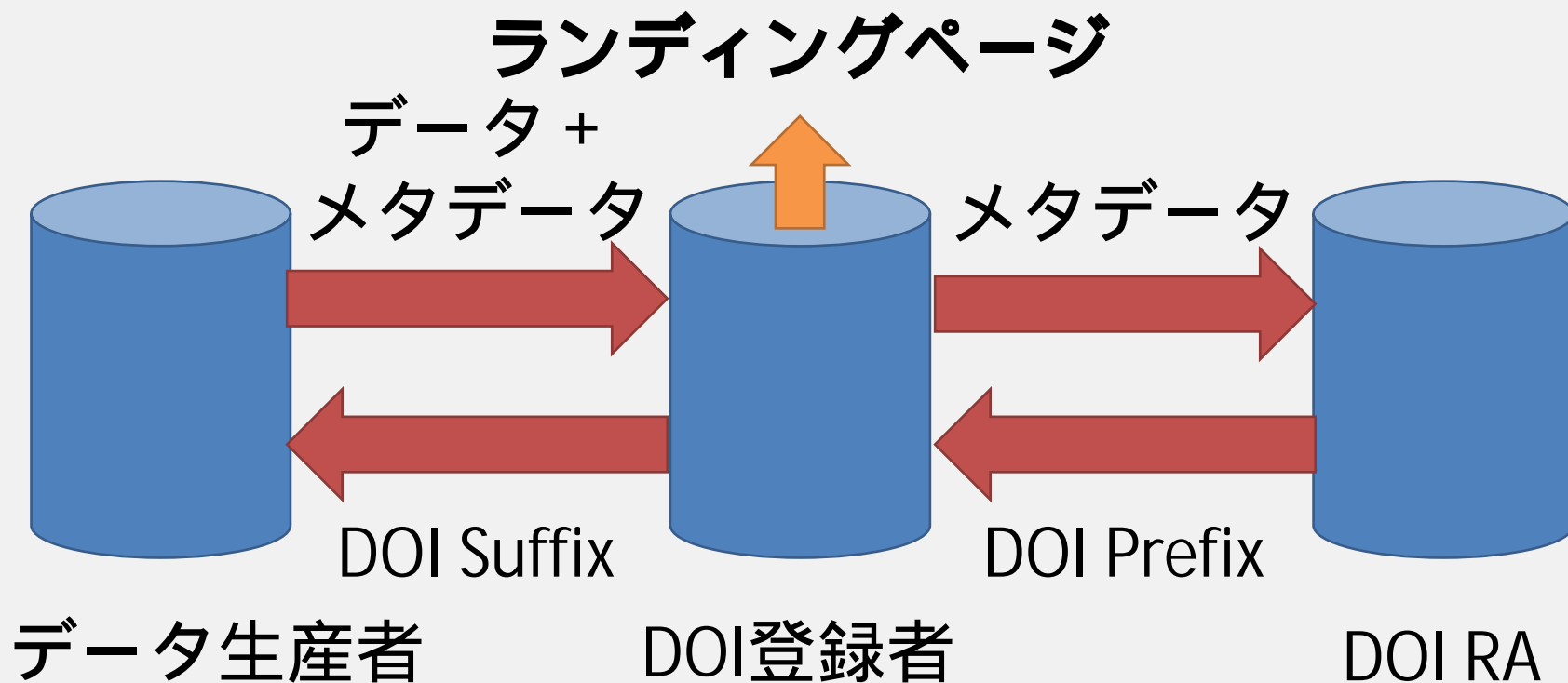
doi<sup>®</sup>, DOI<sup>®</sup>, DOI.ORG<sup>®</sup>, and shortDOI<sup>®</sup> are trademarks of the International DOI Foundation.

# 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

The screenshot shows the DataCite Search interface. The search query is 'japanese literature', resulting in 768 works. The first result is 'Rewriting the Past: Reception and Commentary of Nihon shoki, Japan's First Official History' by Matthieu Anthony James Felt, published in 2017. The second result is 'The "debate on the literature of action" and its legacy - ideological struggles in 1930s Japan and the "rebirth" of the intellectual' by Simone Müller, published in 2015. The third result is 'Four Japanese Poems for Peace' by Sue Stanford, Hayashi Sachiko, Toge Senkichi & Ibaragi Noriko, published in 2015. On the right side, there are filters for 'Resource Types' and 'Publication Year'. The 'Resource Types' filter shows: Dataset (565), Text (173), Other (13), Collection (11), Image (6), and Physical object (3). The 'Publication Year' filter shows a range from 2017 (73) down to 1980 (4). At the bottom, there is a 'Data Centers' filter with 'Global Biodiversity' selected.

<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、引用を通じた評価につながる。
- 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？

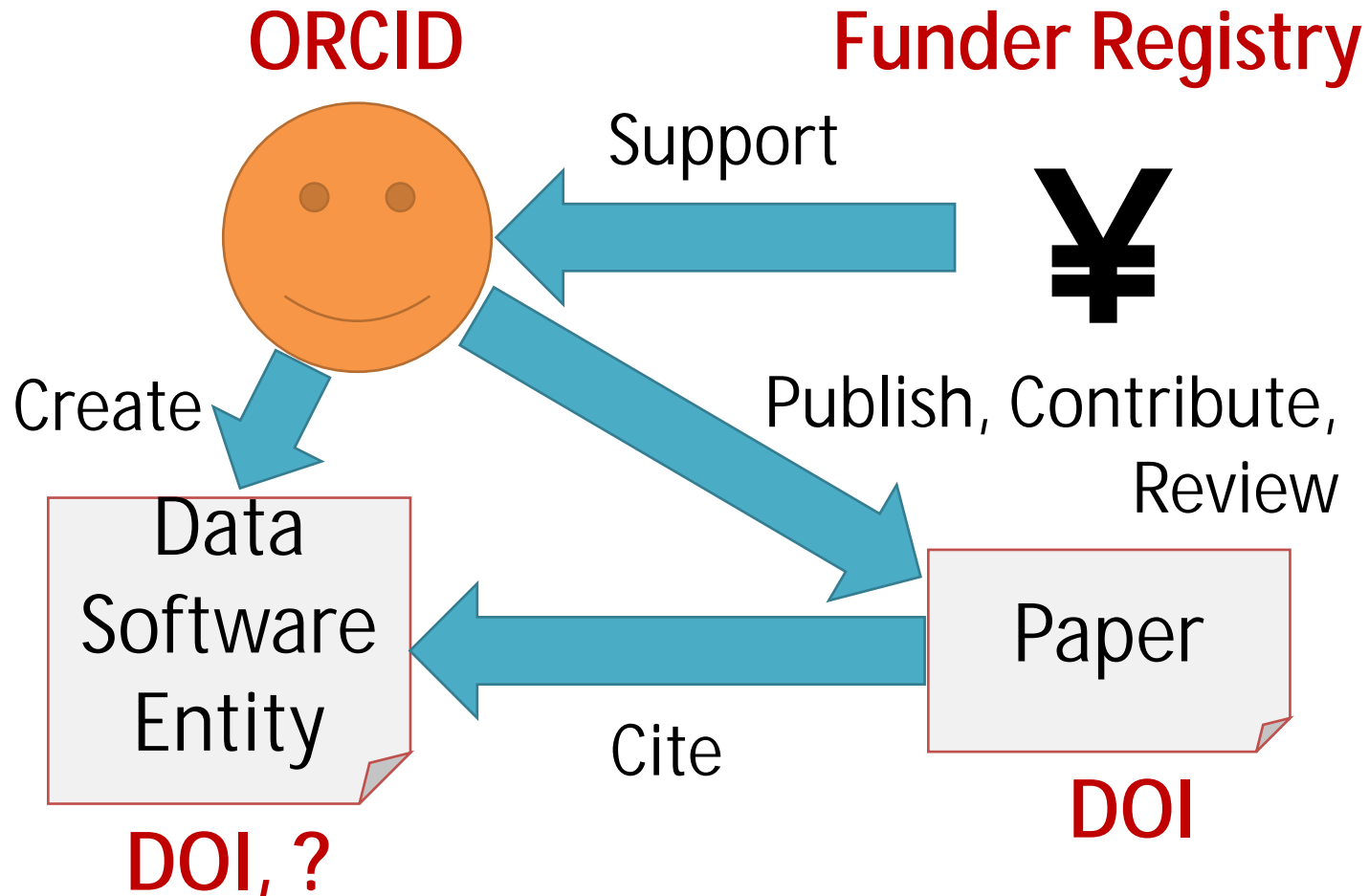
# DOIに関するその他の注意点

- 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は階層構造を持ってない。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。

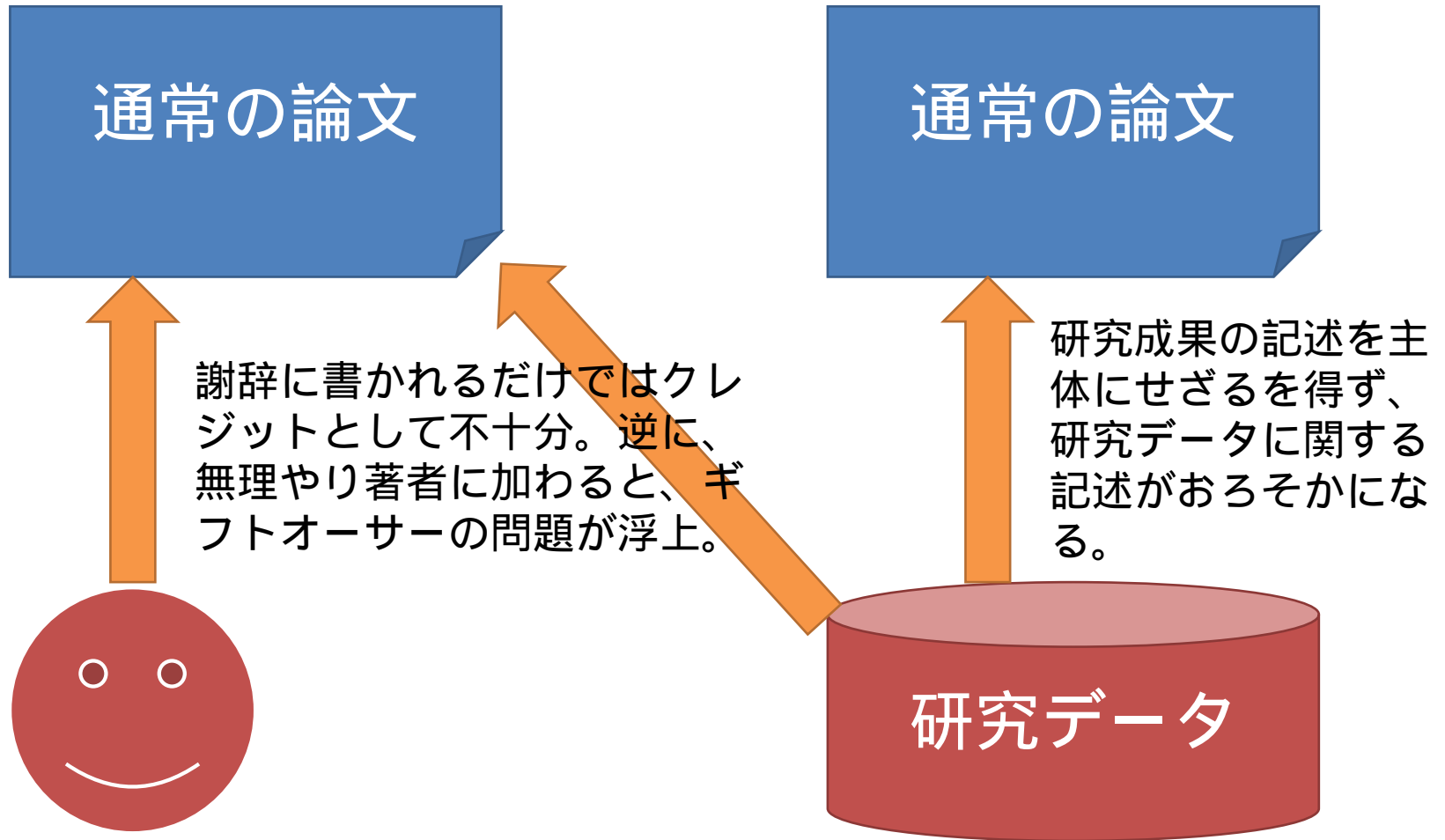


# なぜDOIが必要な のか？

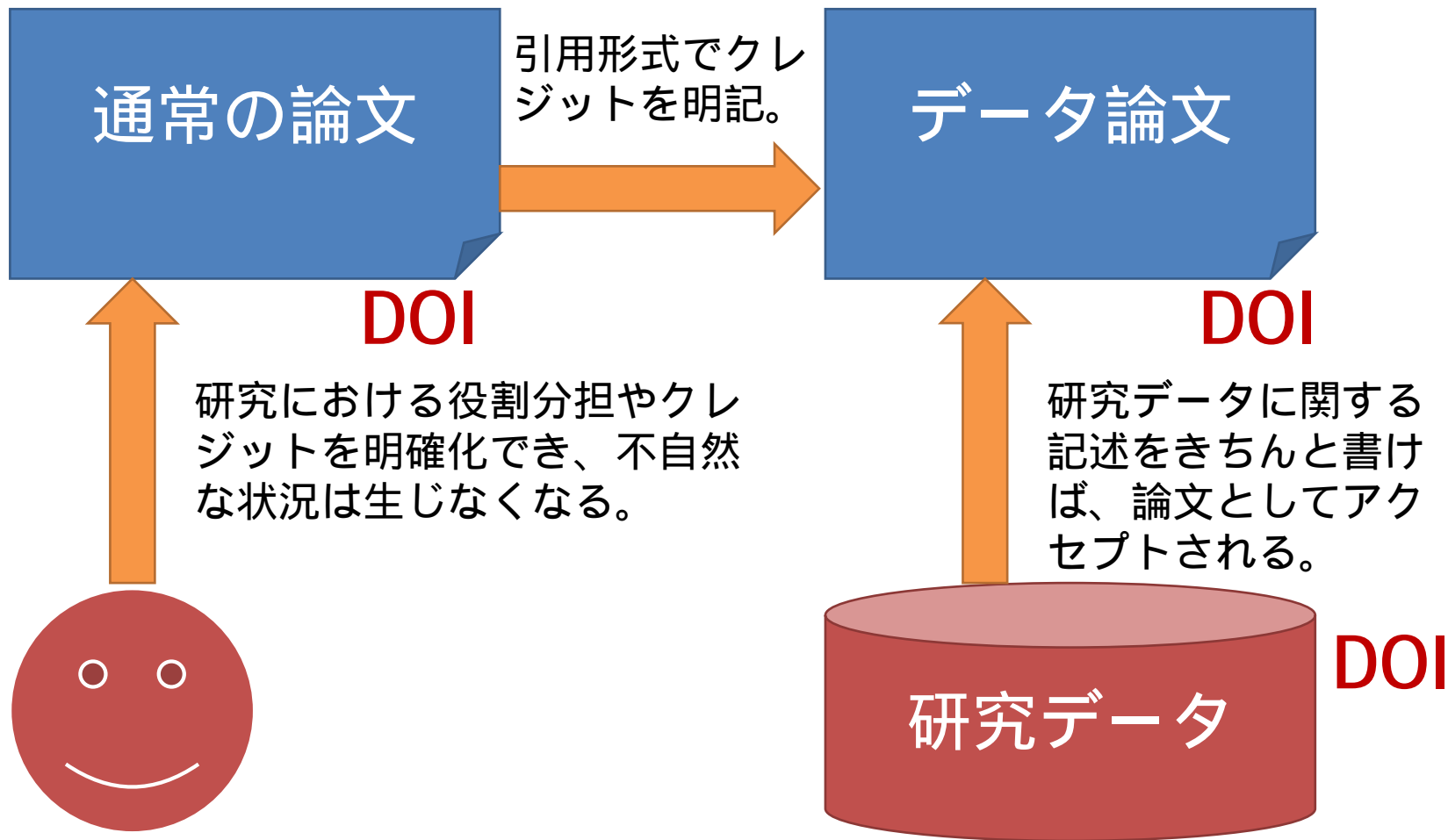
# オープンサイエンスと グローバルな識別子



# 従来の学術出版

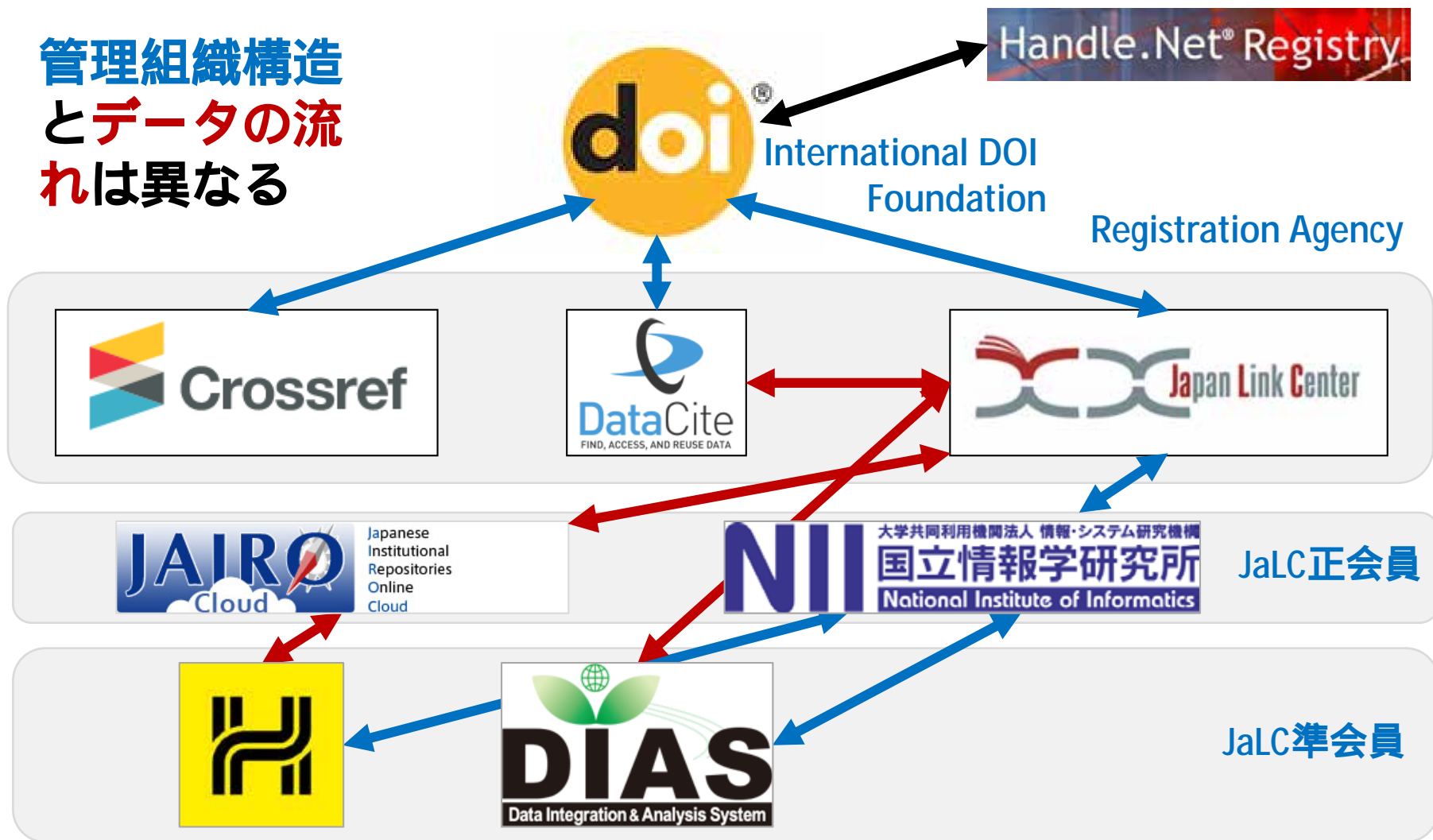


# データ論文を含む学術出版



# DOIシステムの全体像

管理組織構造  
とデータの流  
れは異なる



# 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

adsbeta Feedback ORCID Learn Sign Up Log In

astrophysics data system

Classic Form Modern Form Paper Form

QUICK FIELD: Author First Author Abstract Year Fulltext All Search Terms

Advanced +

author author:"huchra, john" citations citations(author:"huchra, j")

first author author:"^huchra, john" references references(author:"huchra, j")

abstract + title abs:"dark energy" reviews reviews("gamma ray bursts")

year year:2000

year range year:2000-2005 refereed properly refereed

full text full "gravitational waves" astronomy database:astronomy

publication bibstem A<sub>p</sub>J OR abs:(planet OR star)

Use a classic ADS-style form Learn more about searching the ADS Access ADS data with our API

NASA ADS Blog ADS Help @adsats CFA

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

CDS Portal Simbad VizieR Aladin X-Match Other Help

SIMBAD Astronomical Database

Join the LISA VIII (Library and Information Services in Astronomy) conference in Strasbourg June 2017

What is SIMBAD ?

Queries	Documentation	Information
basic search	User's guide	Presentation
by identifier		
by coordinates		Image thumbnails
by criteria	Query by url	
reference query	Nomenclature Dictionary	
scripts	Object types	SimWatch
TAP queries	List of journals	
	Measurement description	
options	Spectral type coding	Release: SIMBAD 1.5.11 - Feb 2017
	User annotations documentation	Release history
Display all user annotations	Acknowledgment	

**Content**

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 various criteria. Lists of objects and scripts can be submitted.

Links to some other on-line services are also provided.

**Basic search**

Identifier, coordinates (radius in arcmin), or bibcode

SIMBAD search clear help

Install the Simbad basic search in your tool bar

**Acknowledgment**

If the Simbad database was helpful for your research work, the following acknowledgment would be appreciated:

*This research has made use of the SIMBAD database, operated at CDS, Strasbourg, France*

2000.A&AS,143,9, "The SIMBAD astronomical database", Wenger et al.

Statistics
Simbad contains on 2017.05.29
9,209,417 objects
24,790,606 identifiers
331,128 bibliographic references
15,791,761 citations of objects in papers

<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。