平成25年度

学術情報システム総合ワークショップ
グループ2 最終報告

大西 賢人（京都大学）
鳥谷 和世（神戸大学）
柴田 育子（一橋大学）

本日の発表内容

1. 選択したテーマと目標設定
2. グループ内での分担
3. 中間発表後の課題設定と目標
4. 課題(1): ID分析―最終分析結果
5. 課題(2): IDマップの作成
6. 課題(3): IDマップの活用法
7. まとめにかえて

1. 選択したテーマと目標設定

1. 選択したテーマと目標設定

テーマ

デジタル化資料のデータベース（NDLサーチ, HathiTrust等）と連携した検索環境整備

目標

CiNii BooksとNDLサーチ, HathiTrust等のデジタル化資料DBとの連携の可能性をさぐる

調査目的

各DBがもつ識別IDをキーにしたリンクの可能性を調査

書誌のIDマップの検討・作成

2. グループ内での分担

2. グループ内での分担

鳥谷
CiNii Books, NDLサーチ, NDLデジタル化資料のIDマップ作成に向けた調査／検討

大西
CiNii Books, HathiTrustのIDマップ作成に向けた調査／検討

柴田
CiNii Books, WorldCat(xISBN), HathiTrustのESTCの識別子を使用したIDマップ作成に向けた調査／検討

3. 中間発表後の課題設定と目標

3.1. WS中間発表後の課題

- **最終目標**
  グループ2ではNCIDに対応するデジタル化資料のIDマップを作成し，CiNii Booksからデジタル化資料のリンク形成のための情報を提供します。

- **具体的な課題**
  1. ID分析
  2. IDマップの作成
  3. IDマップの活用法

3.2. 課題(1): ID分析

鳥谷
JPNOを含む/含まないNC書誌データの分析（対NDL）

大西
ISSN, ISBN, LCCN, OCLC numberを含むHathiTrustメタデータとNC書誌データの分析

柴田
ESTC番号を含むNC書誌データからタイトルでWorldCatを通じてHathiTrustにリンク形成ができるか検討

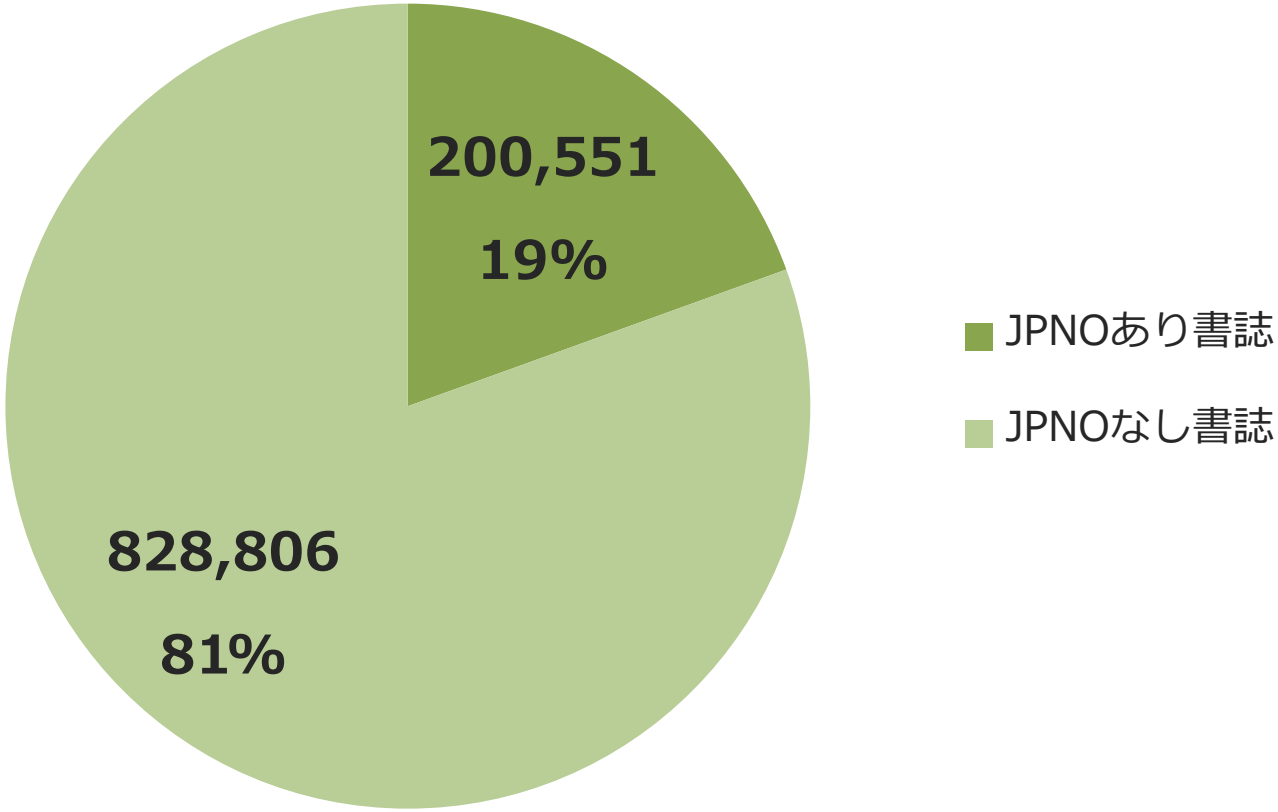
メタデータ分析 → **NCIDとの対応関係**

4. 課題(1): ID分析―最終分析結果

4. 課題(1): ID分析―最終分析結果

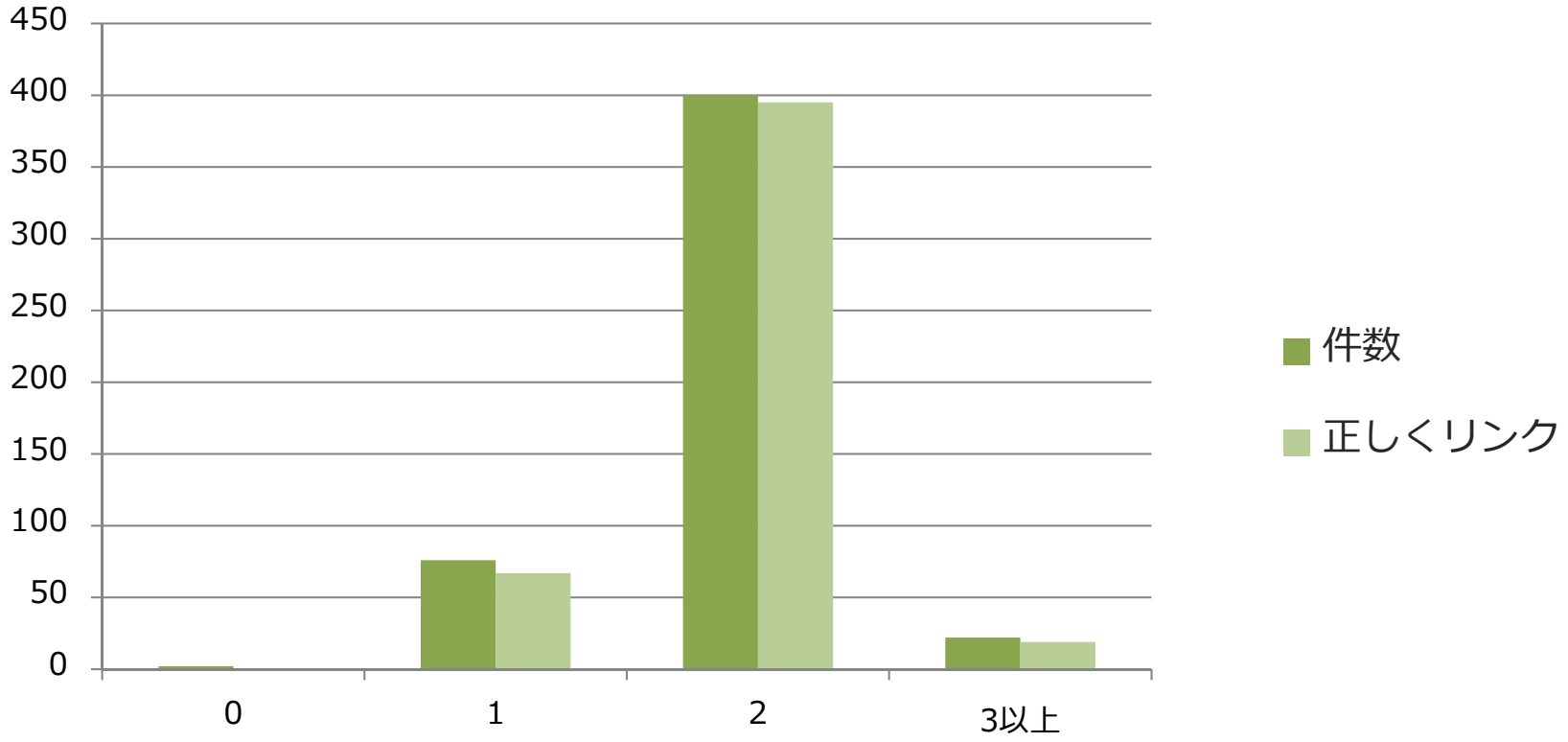
4.1. JPNOを含む/含まないNC書誌データの分析（対NDL）

4.1.1. NC書誌のJPNO保有状況



データ抽出条件：1968年以前発行，かつ，TTLLがjpn

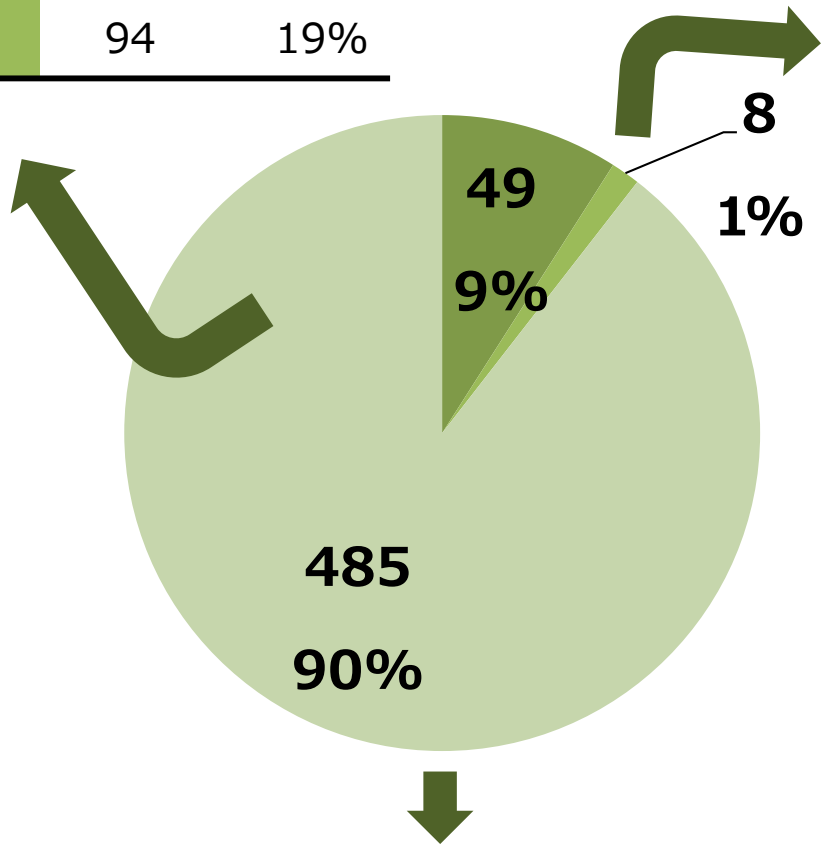
4.1.2. NC書誌保有JPNOの精度
—500件のJPNOによるNDLサーチヒット状況



1件ヒット:　　デジタル化されていない
2件ヒット:　　図書書誌とデジタル資料が1:1
3件以上ヒット:　図書書誌とデジタル資料が1:2以上

4.1.2. NC書誌保有JPNOの精度
—542件のNC書誌の対応NDL書誌との同定可能性

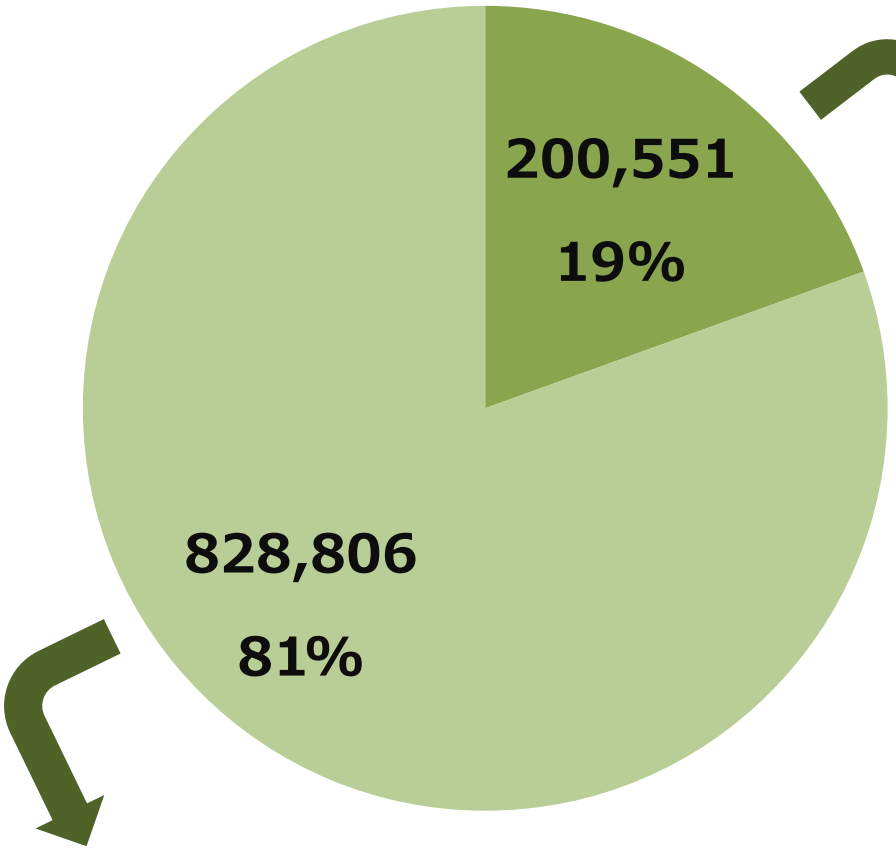
デジタル化されていない	67	14%
館内公開	324	67%
公開	94	19%



JPNO間違い	13
所蔵なし	36

一部にのみリンク可：485件のうち26件 (5%)

4.1.2. NC書誌保有JPNOの精度―総括



90%程度の精度で
NDL書誌にリンク可

ただし、対応するNDL書誌が複数の場合，その一部にのみリンクするケース有

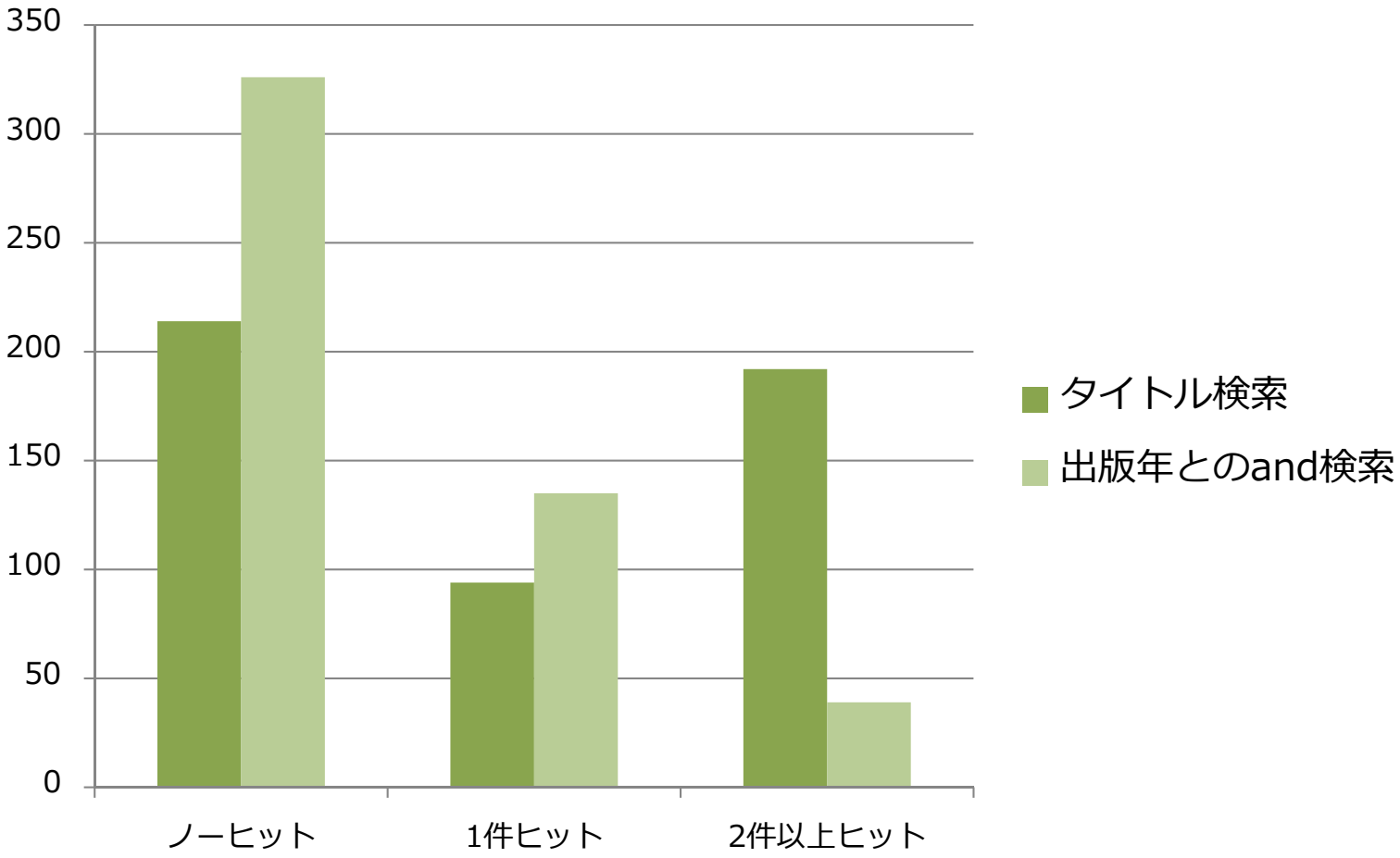
シンプルな作業で現在の保有JPNOと遜色ない精度のJPNOを把握する方法はないか？

4.1.3. JPNO把握の可能性
—NC書誌から（作業手順）

1. JPNO未保有書誌**828,806**件から**500**件を抽出
2. NCIDからNC書誌のタイトル取得
3. 上記タイトルでNDLサーチに対しデータプロバイダをNDL-OPACと国立国会図書館デジタル化資料に限定してOpenSearchで検索
4. 3の結果，NDL-OPACのヒット件数が1件だったものにつき，書誌比較
5. NCIDからNC書誌の出版年を取得し，タイトルと出版年とのAND検索を3，4と同様に実施

(補) CiNiiにOpenSearchでタイトル完全一致，タイトル完全一致+出版 年で検索を行い，ヒット件数を確認

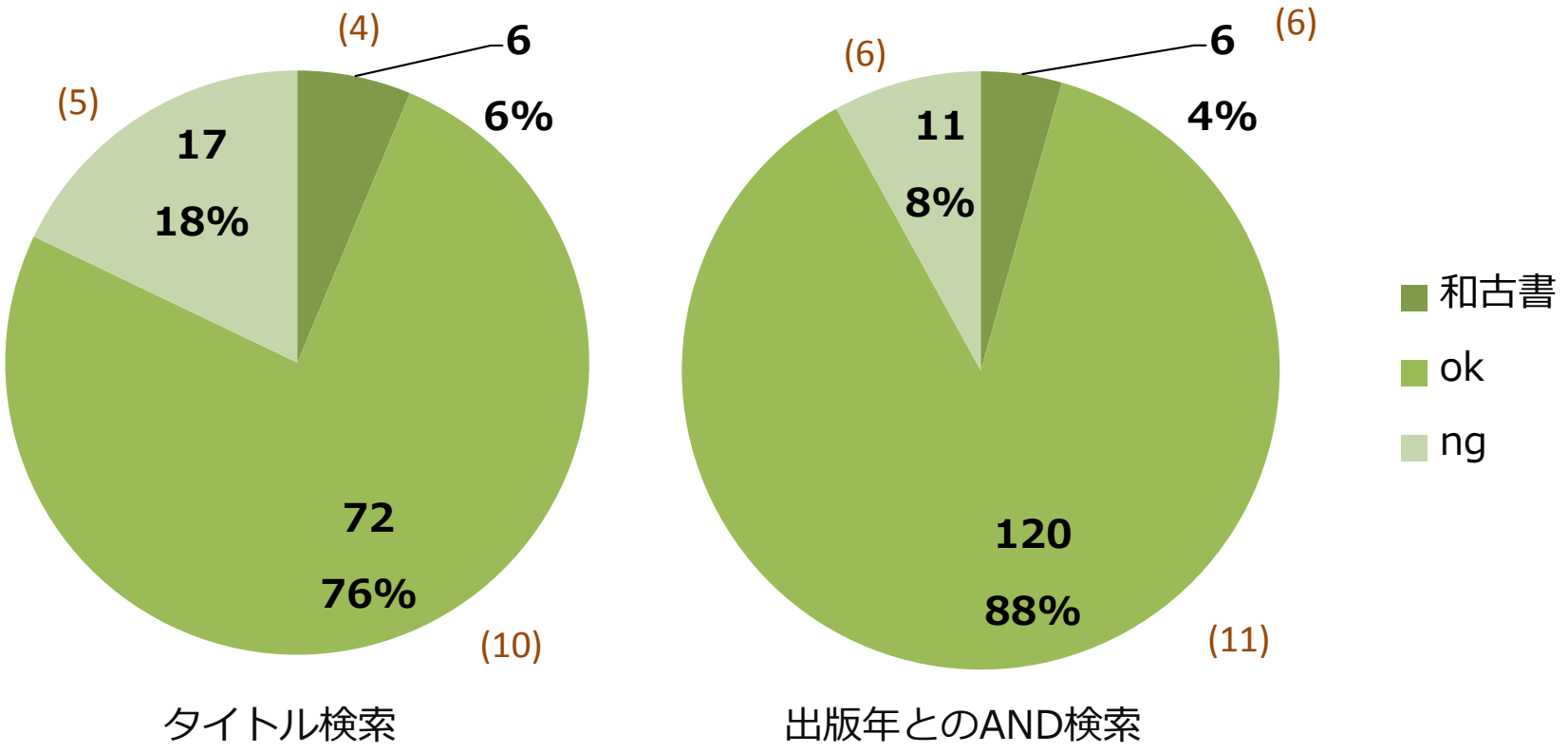
4.1.3. JPNO把握の可能性
—NC書誌から （NDL-OPACヒット状況）



ノーヒットのうち，デジタル化資料のみ1件ヒット，
1件ヒットのうちデジタル化資料ありのケースあり

検証対象
タイトル検索：95件（19%）
出版年とのand検索：137件（27%）

4.1.3. JPNO把握の可能性
―NC書誌から（NDL-OPAC1件ヒット分の同定可否）

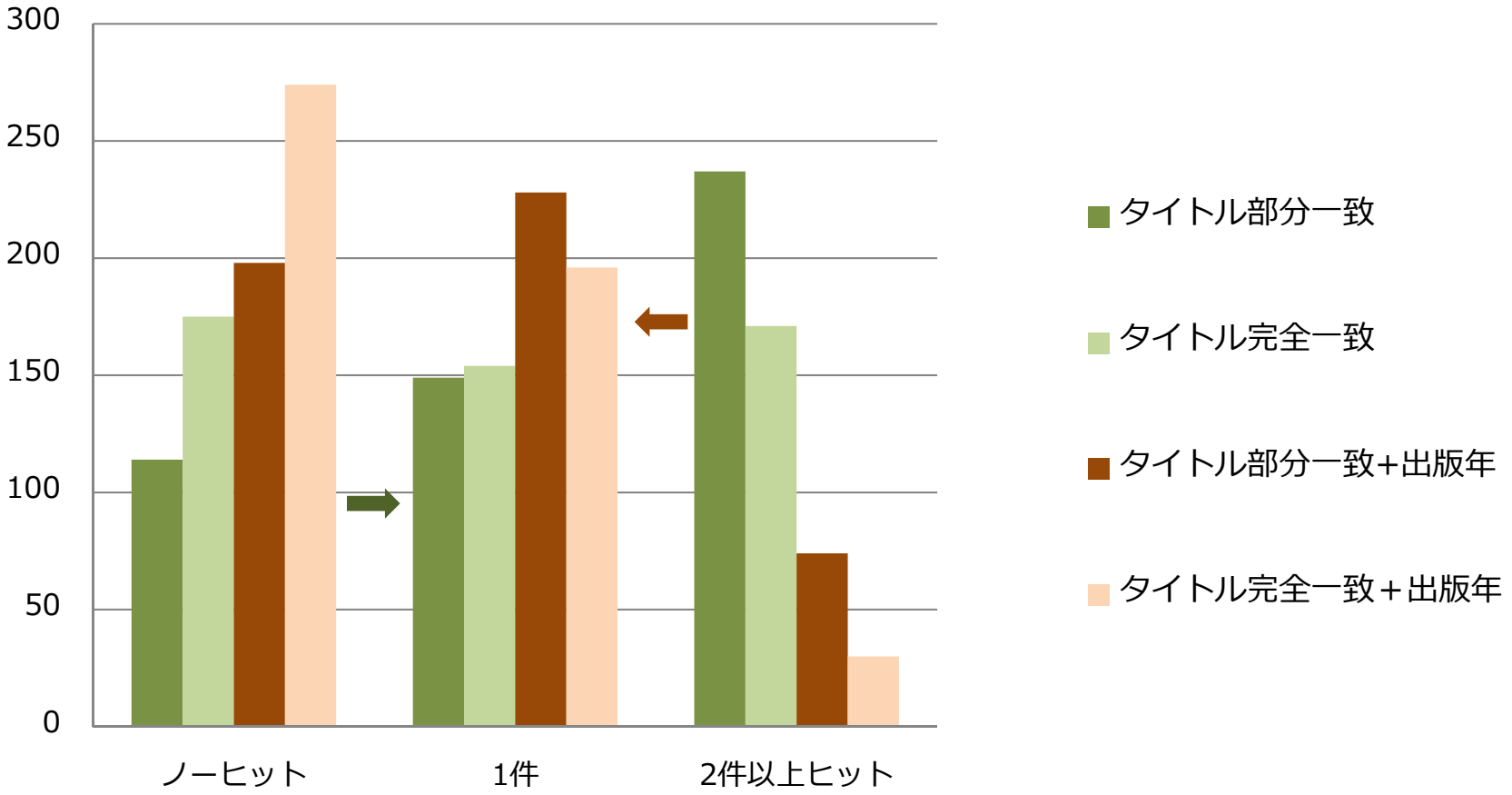


(　)はNCに同タイトルの書誌が複数ある書誌数

4.1.4. JPNO把握の可能性
—NDL書誌から（作業手順）

1. NDL図書デジタルデータ**894,275**件から**500**件を抽出
   （対象500件と同じ書誌のデジタルデータ**635**件）
2. CiNii Booksに対しOpenSearchでタイトル部分一致およびタイトル完全一致検索を実施
3. 2の結果，いずれかでヒット件数が1件だったものにつき書誌比較
4. CiNii Booksに対し出版年とのAND検索を2，3と同様に実施

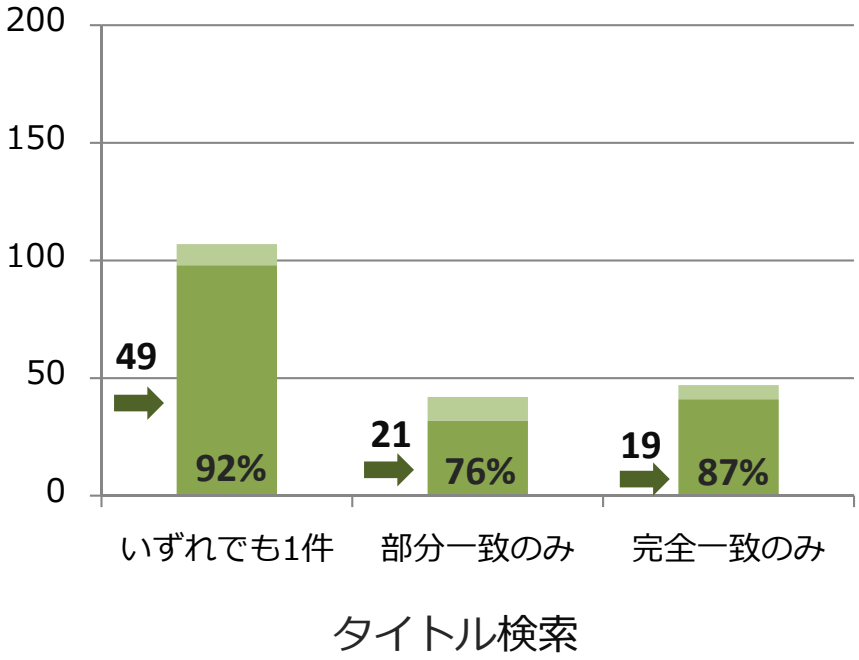
4.1.4. JPNO把握の可能性
—NDL書誌から（ヒット状況）



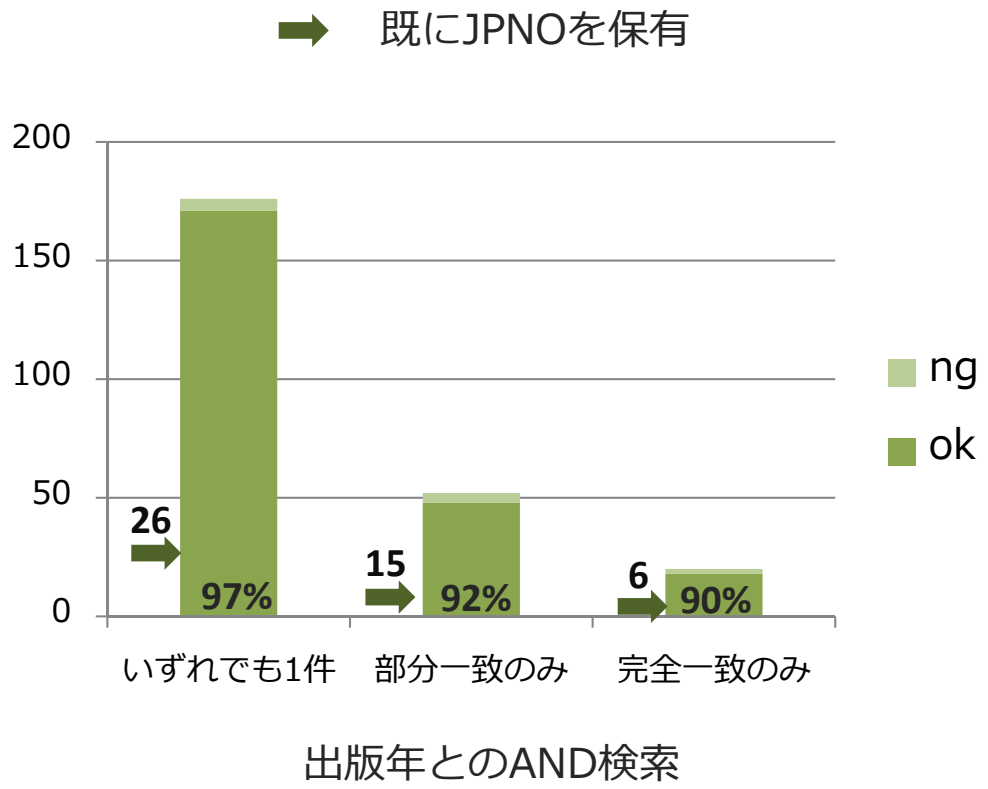
いずれも1件ヒット
タイトル検索　107件
出版年とのAND検索　176件

検証対象
タイトル検索　171件（34%）
出版年とのAND検索　237件（47%）

4.1.4. JPNO把握の可能性
—NDL書誌から（1件ヒット分の同定可否）



タイトル検索

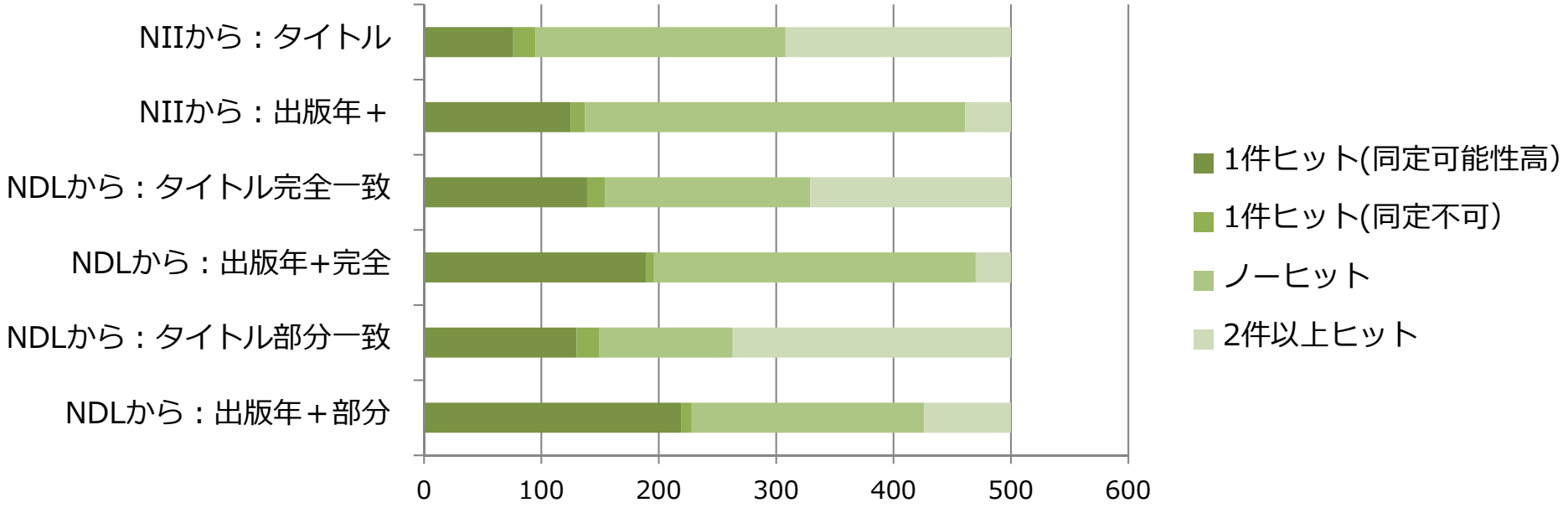


出版年とのAND検索

親書誌にヒットするケースあり

タイトル検索	15件
出版年とのAND検索の場合	3件

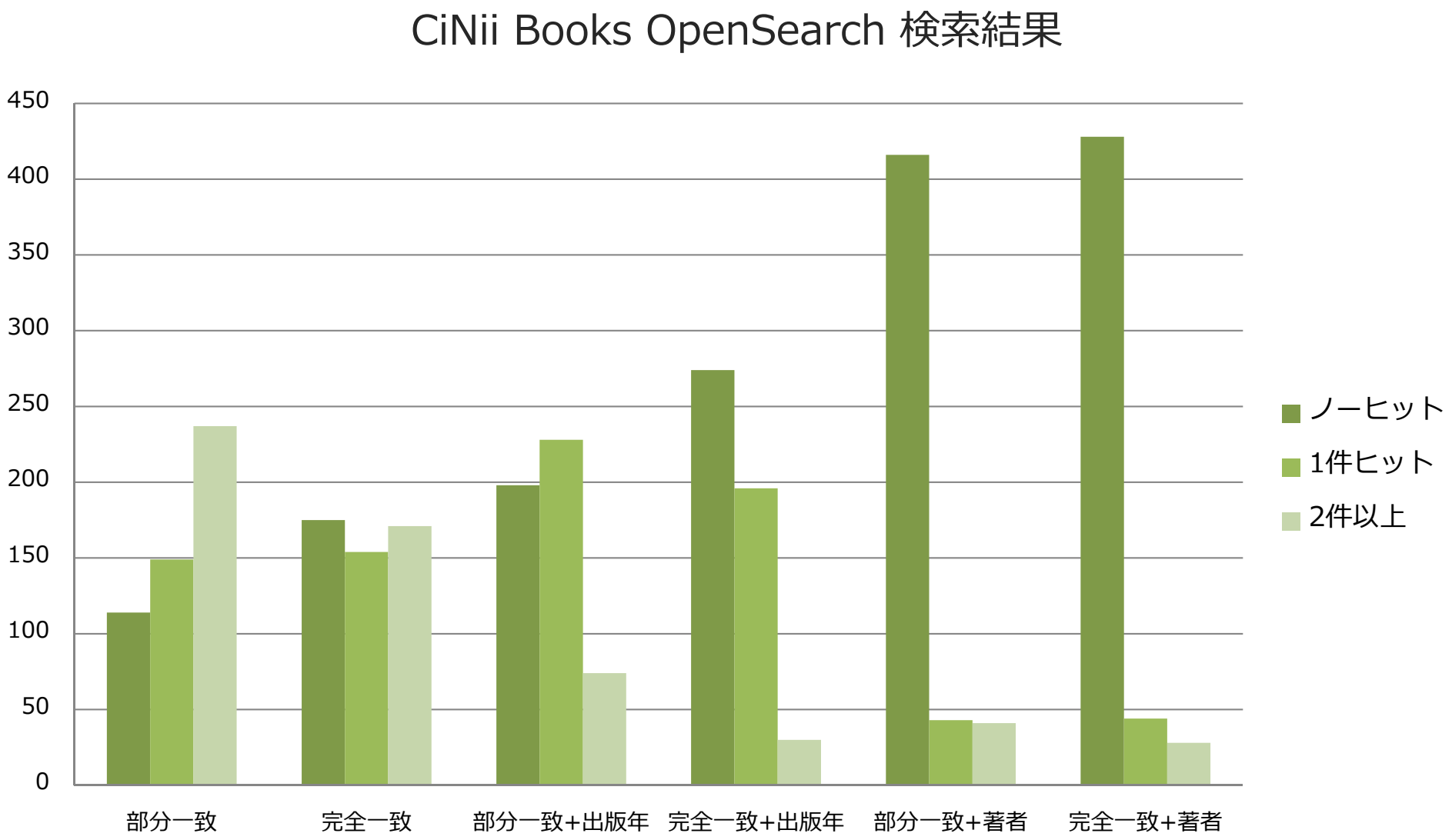
4.1.5. JPNO把握の可能性―総括



タイトルと出版年のAND検索で1件ヒットした場合に，JPNOを取得してはどうか？

- NDLが物理単位での書誌作成が基本のためか，1件ヒット率はNCからのほうが低い
- JPNOを保有しているNC書誌も，100%正確ではなくまた対応するNDL書誌を網羅してはいない

参考: 各項目の検索結果の比較

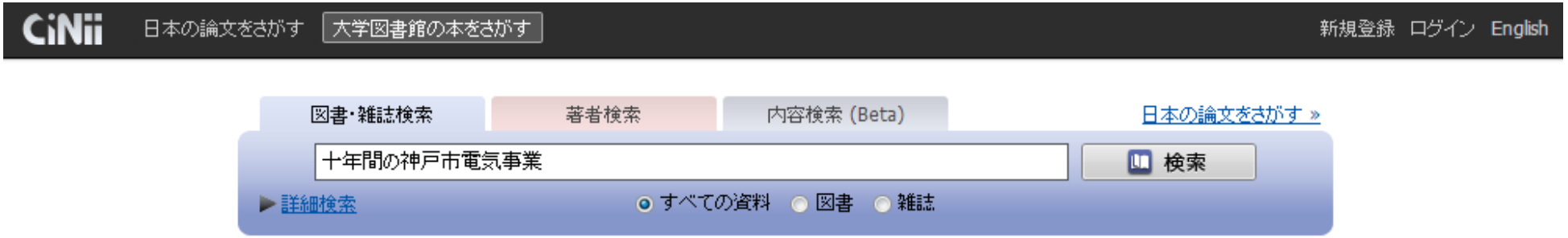


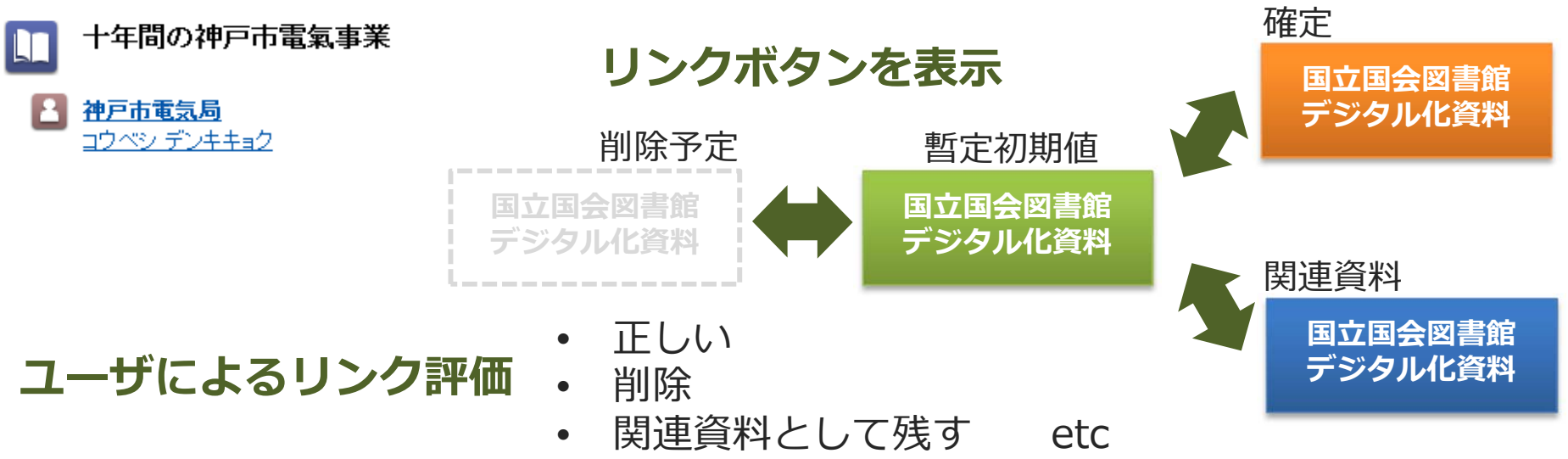
4.1.6. 運用に際して
—運用上の問題

- 現に保有しているJPNOが一定程度の誤りを含む
- 100%正確にJPNOを追加把握することは難しい

ユーザのフィードバックにより保有するJPNOの精度を評価し，それを反映できる仕組みができればユーザの理解は得られるのでは？

4.1.6. 運用に際して―イメージ





背景色を変えるなどしてリンクへの評価状況がわかるようにできないか？

4. 課題(1): ID分析—最終分析結果

4.2. ISSN, ISBN, LCCN, OCLC numberを含むHathiTrustメタデータとNC書誌データの分析

4.2. ID分析—HathiTrust

1. HathiTrust概要
2. ID分析対象データ, 使用ツール
3. ISSNによる書誌同定
4. ISBNによる書誌同定
5. LCCNによる書誌同定
6. OCLCnumberによる書誌同定
7. タイトルによる書誌同定
8. HathiTrust分析のまとめ

4.2.1. HathiTrustとは

- 米国の大学図書館などが共同で運営しているデジタル化資料のリポジトリ（2008年-）
- 80以上の図書館およびコンソーシアムが参加（2013年11月現在）
- 著作権処理，蔵書管理・構築，長期保存，メタデータ管理など多岐にわたるプロジェクト

http://www.hathitrust.org/

4.2.1. HathiTrust―コンテンツ統計

- 10,866,199 total volumes
- 5,704,683 book titles
- 284,409 serial titles
- 3,803,169,650 pages
- 487 terabytes
- 129 miles
- 8,829 tons
- **3,510,393 volumes(-32% of total) in the public domain**

http://www.hathitrust.org/statistics_info

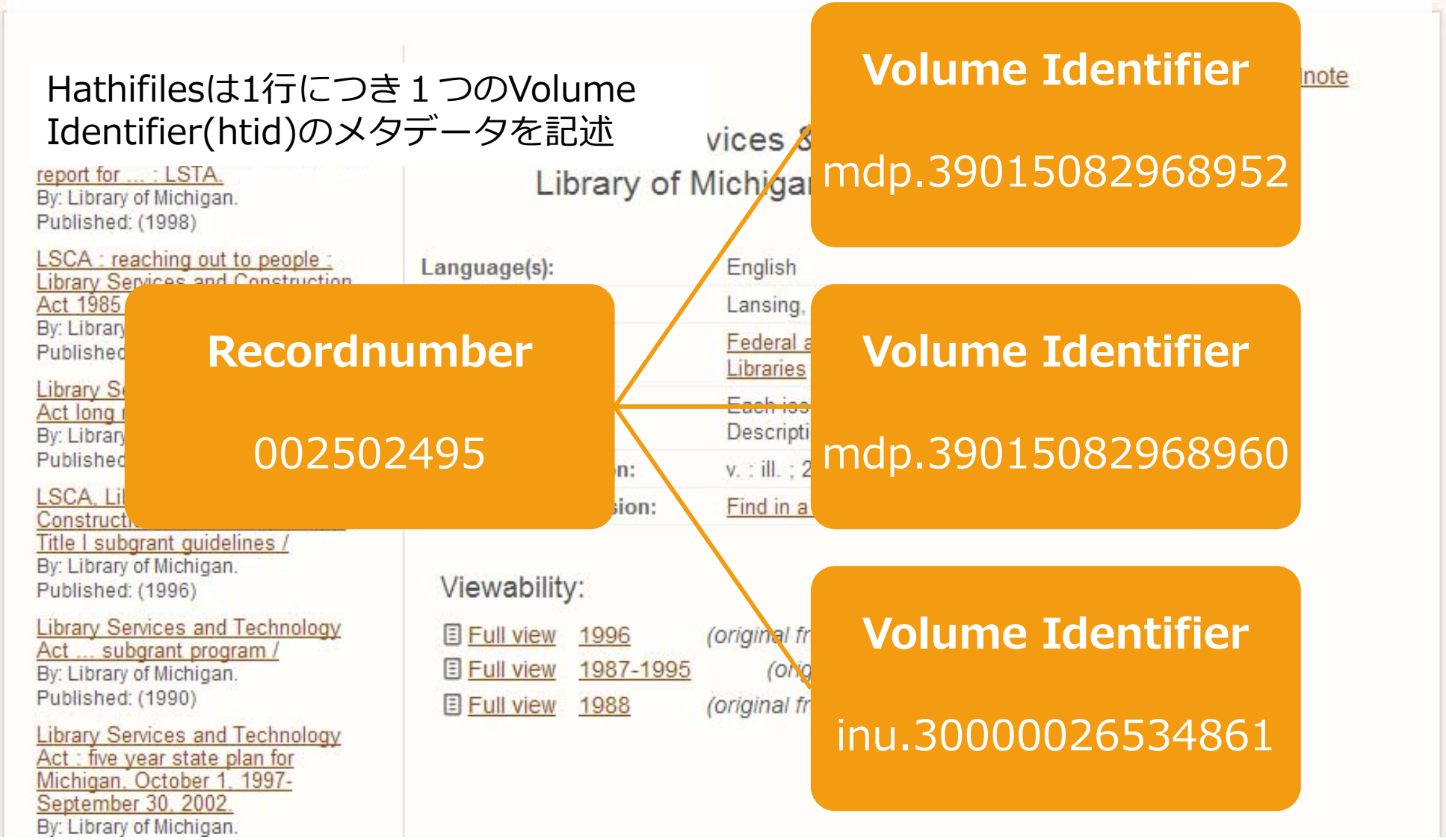
4.2.1. HathiTrust—Hathifiles

- メタデータのうち，20のエレメントを抽出したタブ区切りテキストファイル
  - **Volume Identifier (htid)**
  - Access
  - **Rights**
  - **University of Michigan record number**
  - Enumeration/Chronology
  - Source
  - Source institution record number
  - **OCLC numbers**
  - **ISBNs**
  - **ISSNs**
  - **LCCNs**
  - Title
  - Imprint
  - Rights determination reason code
  - Date of last update
  - Government Document
  - Publication Date
  - Publication Place
  - Language
  - Bibliograhic Format
- 月次の全データファイルと日次差分ファイルを公開

http://www.hathitrust.org/hathifiles

4.2.1. HathiTrust—ID構造

Hathifilesは1行につき１つのVolume Identifier(htid)のメタデータを記述



http://catalog.hathitrust.org/Record/002502495

4.2.2. ID分析―対象データ

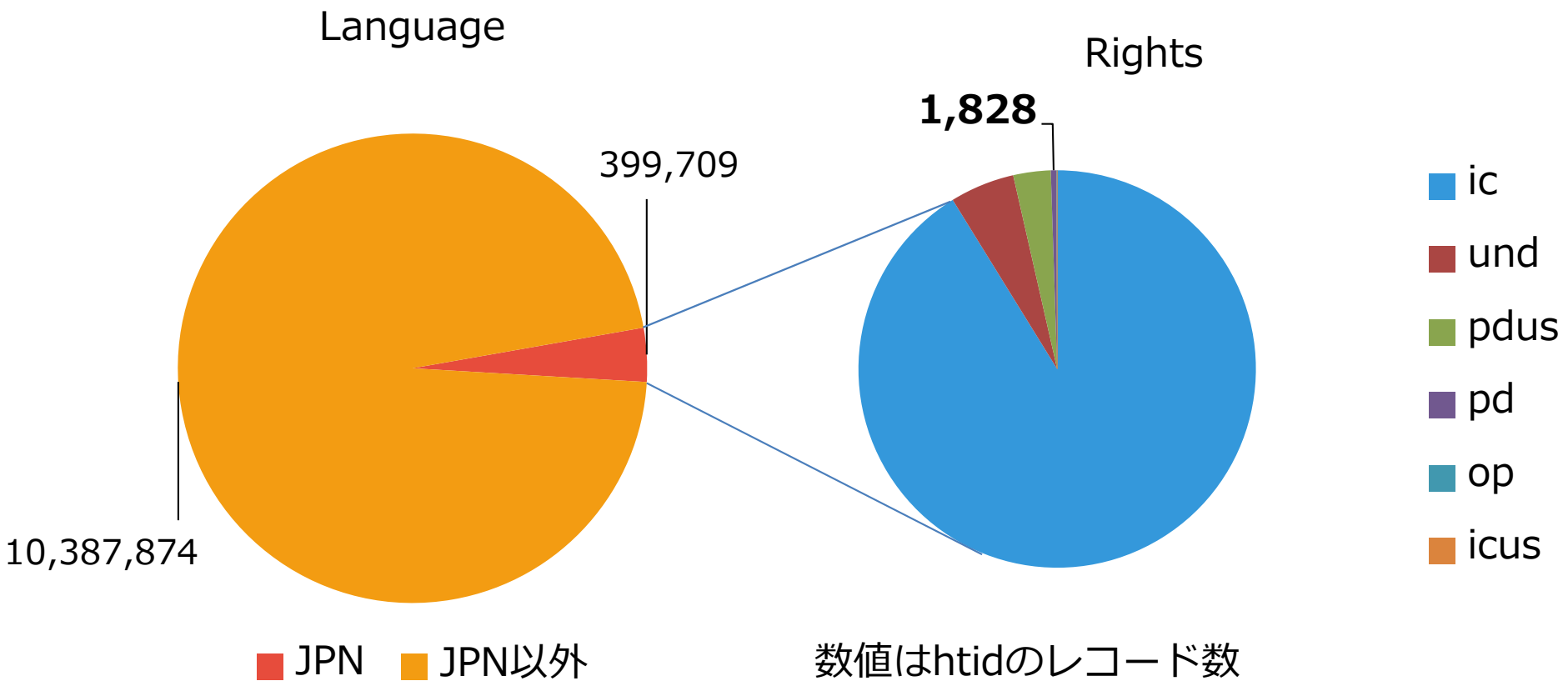
- HathiTrustメタデータ
  - Hathifiles 20130801

    http://www.hathitrust.org/sites/www.hathitrust.org/files/hathifiles/hathi_full_20130801.txt.gz
- APIで取得したメタデータ
  - CiNii Books
  - HathiTrust
  - WorldCat
  - Digital Public Library of America

4.2.2. ID分析―対象データ

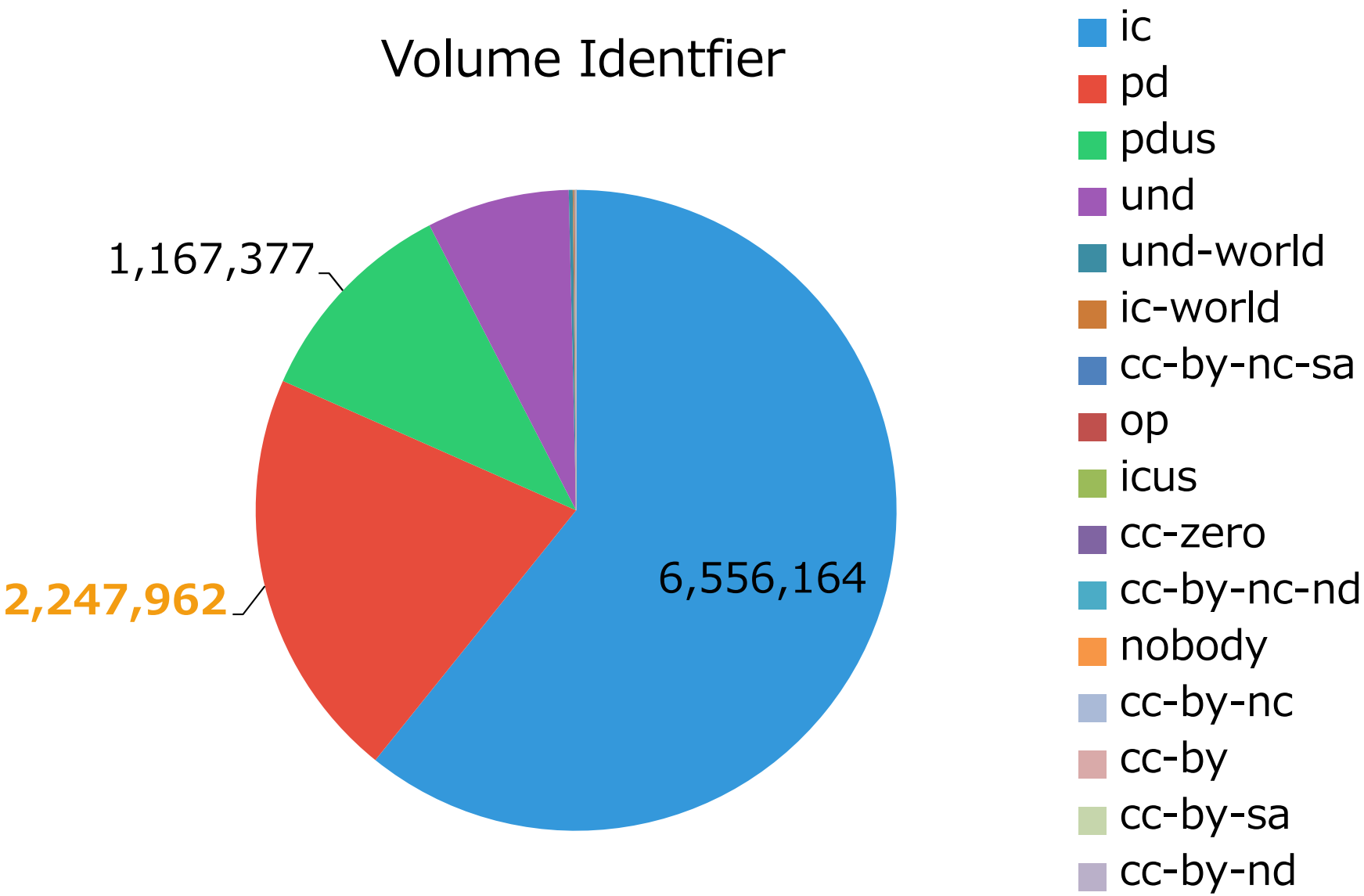
- NC書誌データ その1（NII提供）
  - 抽出条件: TTLLがjpn以外でISBN, XISBN, ISSN, NBN, LCCN, OTHN(ISSN か LCCN か NBN)のいずれかのIDを持つデータ
  - 抽出項目: ID, YEAR, CNTRY, TTLL, TXTL, ISBN, XISBN, ISSN, NBN, LCCN, GPON, OTHN, 親書誌レコードID
- NC書誌データ その2（NII提供）
  - 抽出条件: OTHNあるいはNOTEに「OCLC:」「OCLC data」「OCoLC」のいずれかが含まれているデータ
  - 抽出項目: ID, YEAR, CNTRY, TTLL, TXTL, GMD, TR, ED, PUB, PHYS, VT, NOTE, 親書誌レコードID

4.2.2. ID分析—Hathifiles JPN

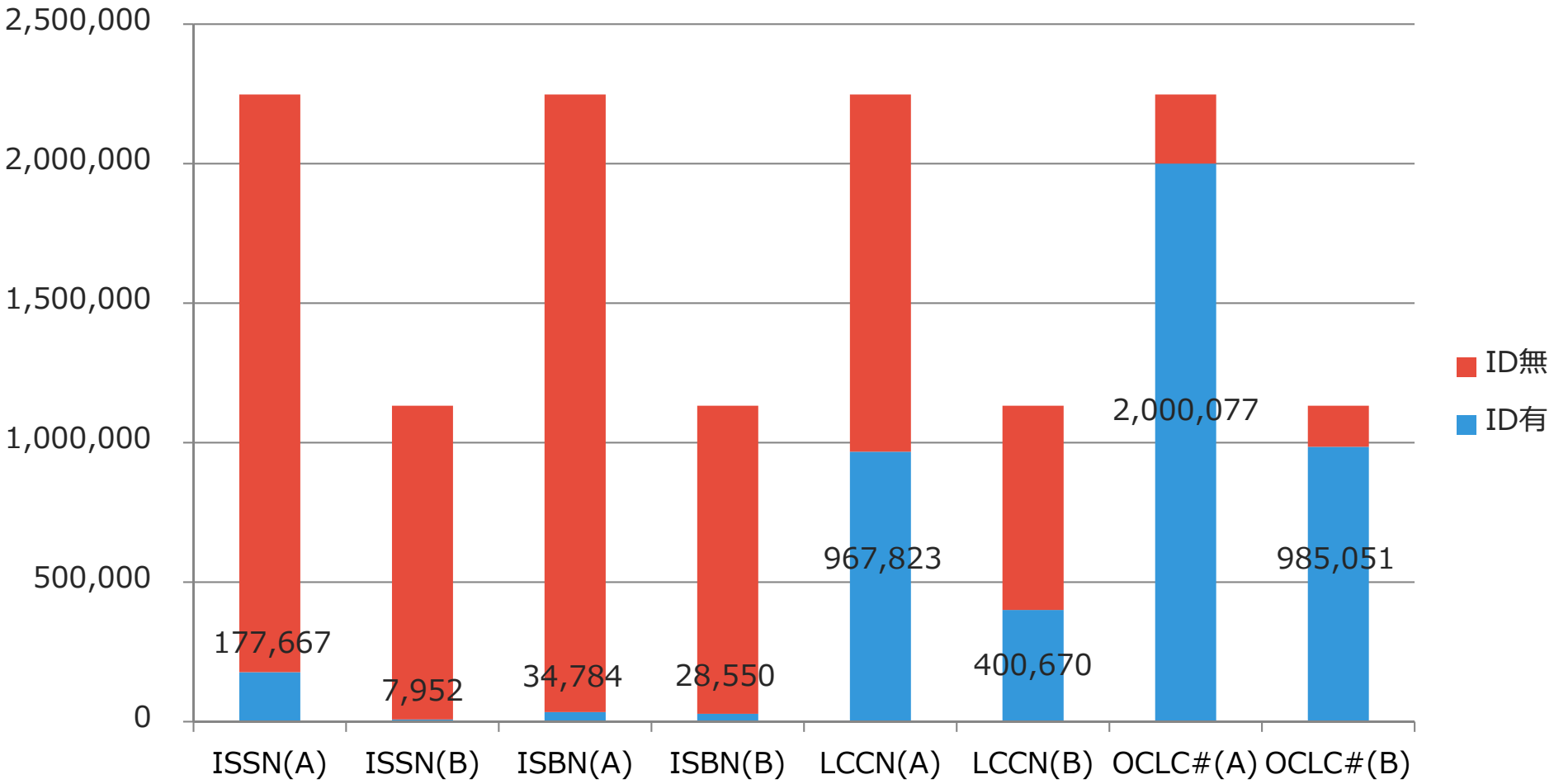


日本からアクセス可能な日本語コンテンツは少ない

4.2.2. ID分析—Hathifiles Rights



4.2.2. ID分析—Hathifiles PD



(A): Rights(pd)のhtidの中で各IDを保有するhtidの件数

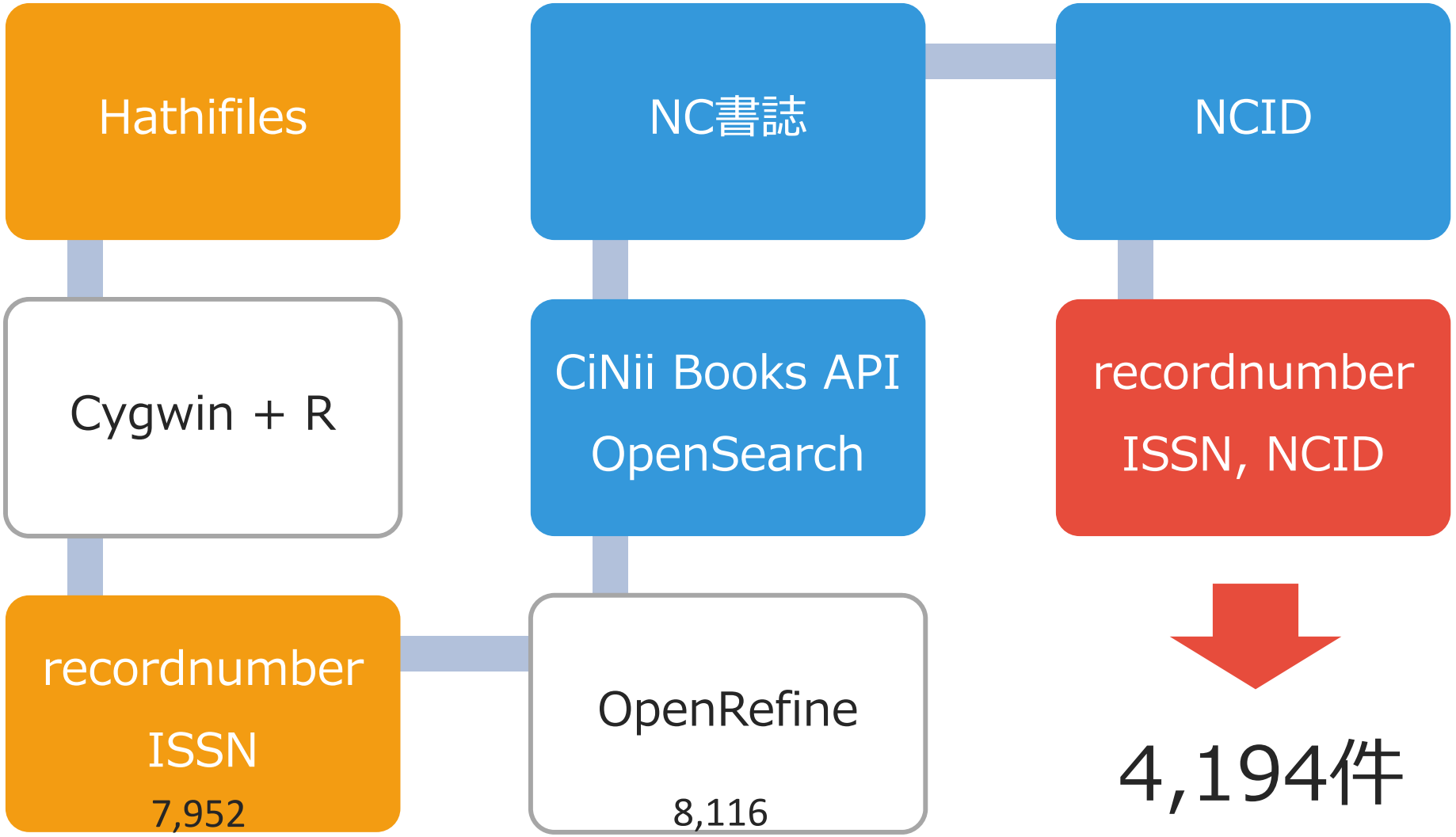
(B): (A)のうちhtid以外のIDの組み合わせが重複しているものを除去後，各IDを保有するhtid件数

4.2.2. ID分析―使用ツール

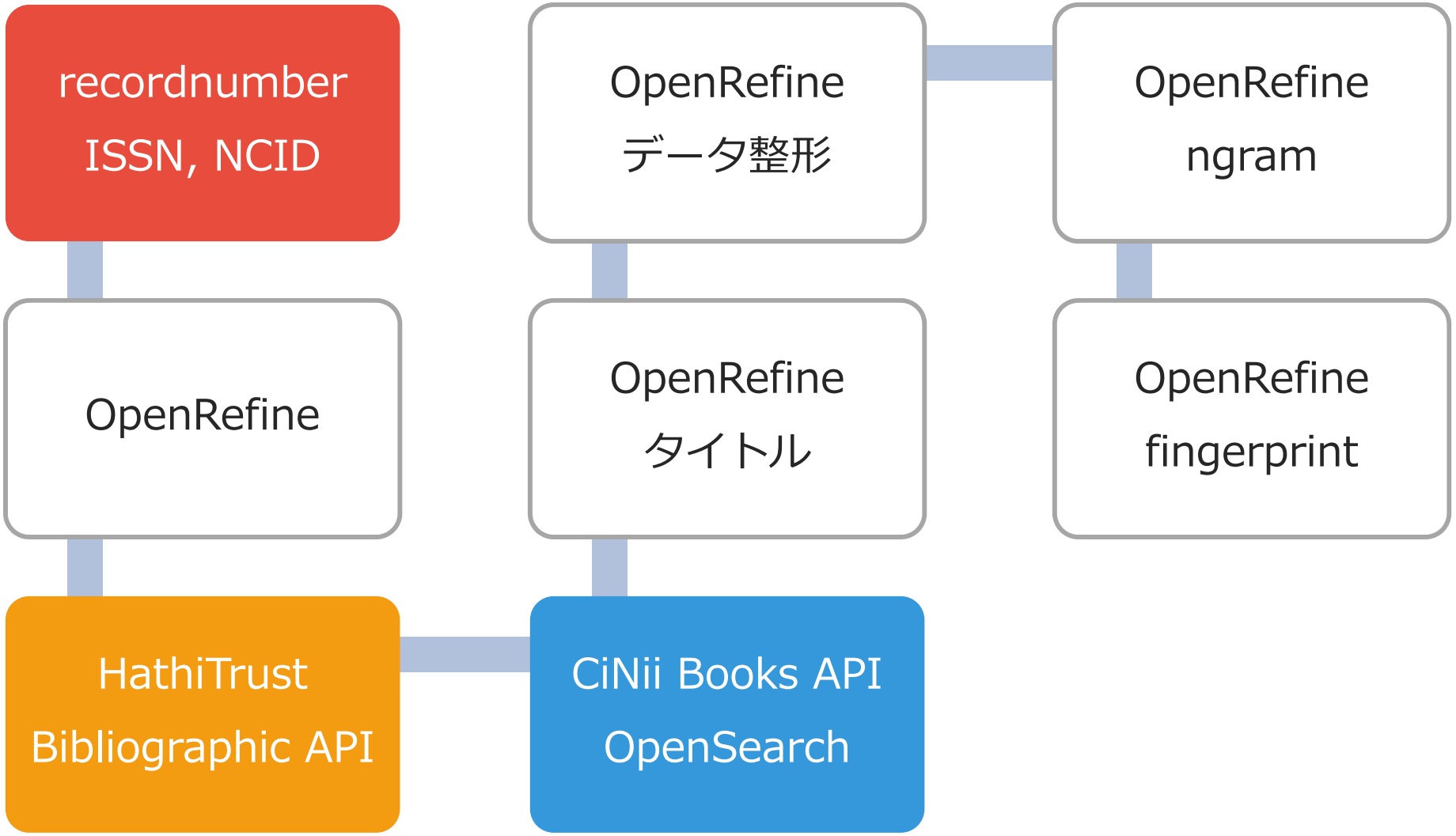
- ツール
  - **OpenRefine**
  - Cygwin
  - R
  - テキストエディタ
  - Excel
- API
  - CiNii Books API
  - HathiTrust Bibliographic API
  - WorldCat Basic API
  - Digital Public Library of America API

OpenRefineはデータの整形だけでなく，API問い合わせやデータの分析などにも使用できる非常に便利なツール！

4.2.3. ISSNによる書誌同定

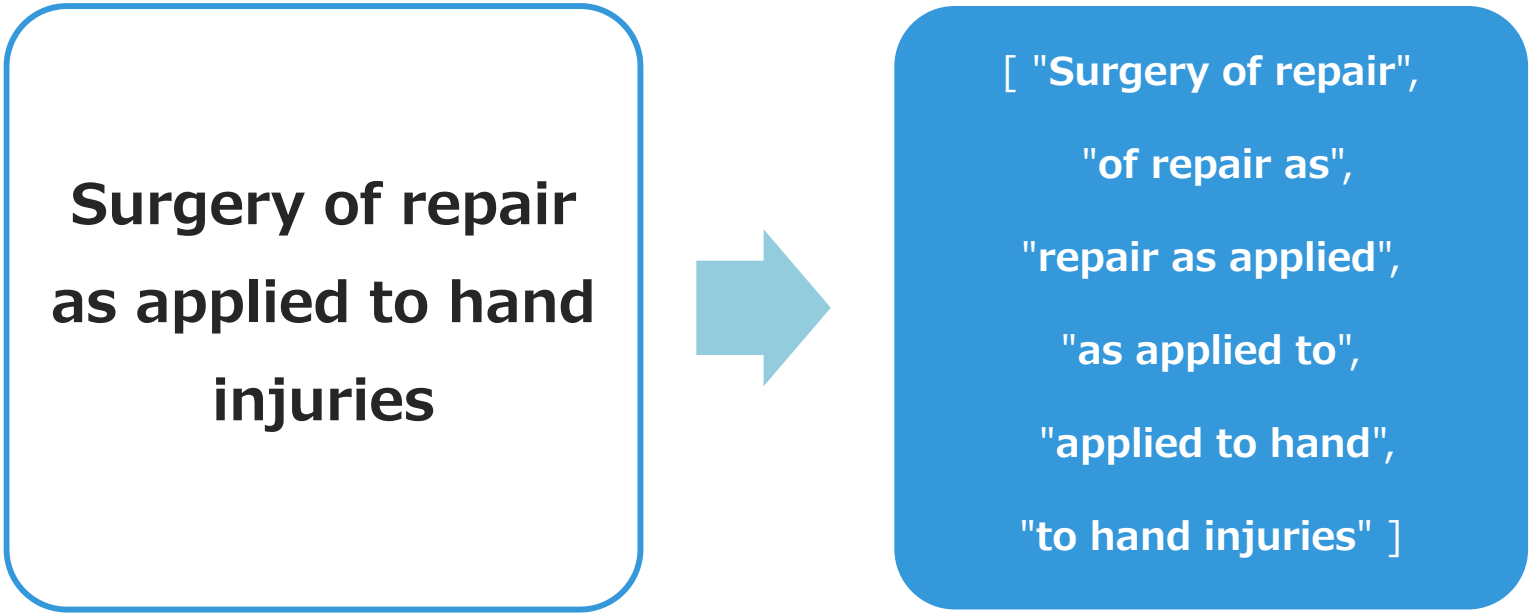


4.2.3. ISSNによる書誌同定―検証



4.2.3. OpenRefine - ngram

- ngram(string s, number n)
  - 対象文字列の単語N-gramを配列の形でかえす

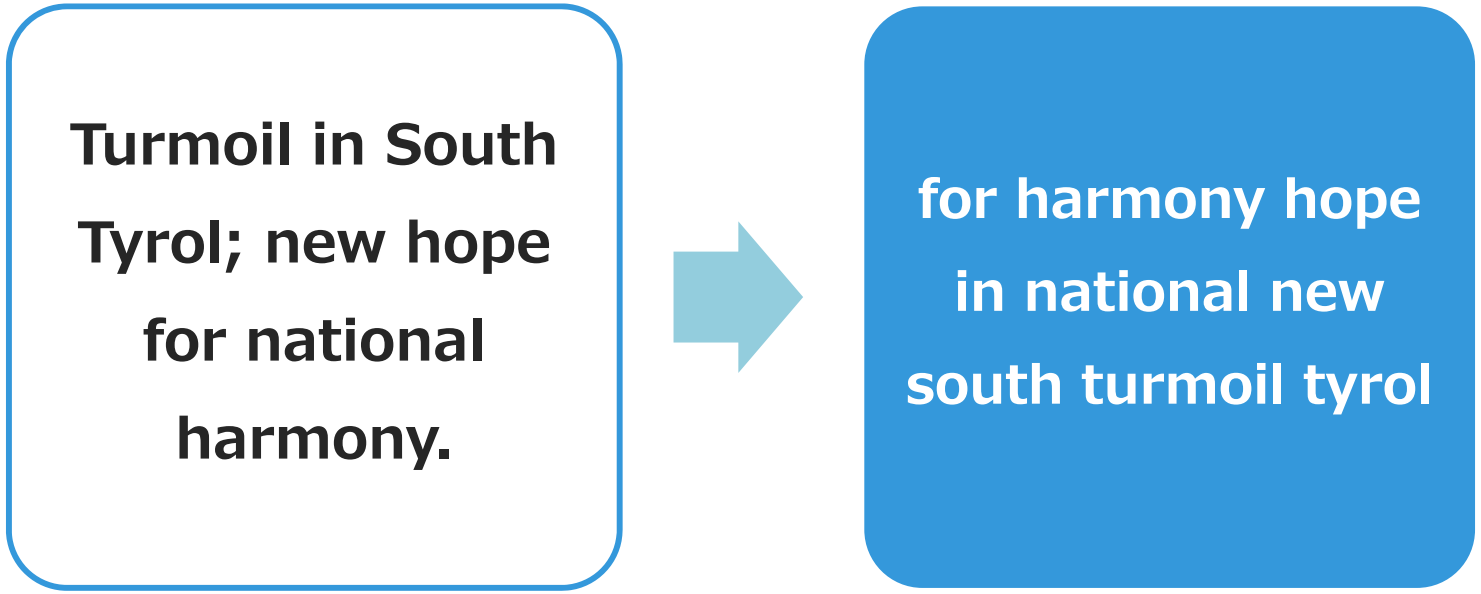


ngram(value,3)

https://github.com/OpenRefine/OpenRefine/wiki/GREL-String-Functions

4.2.3. OpenRefine - fingerprint

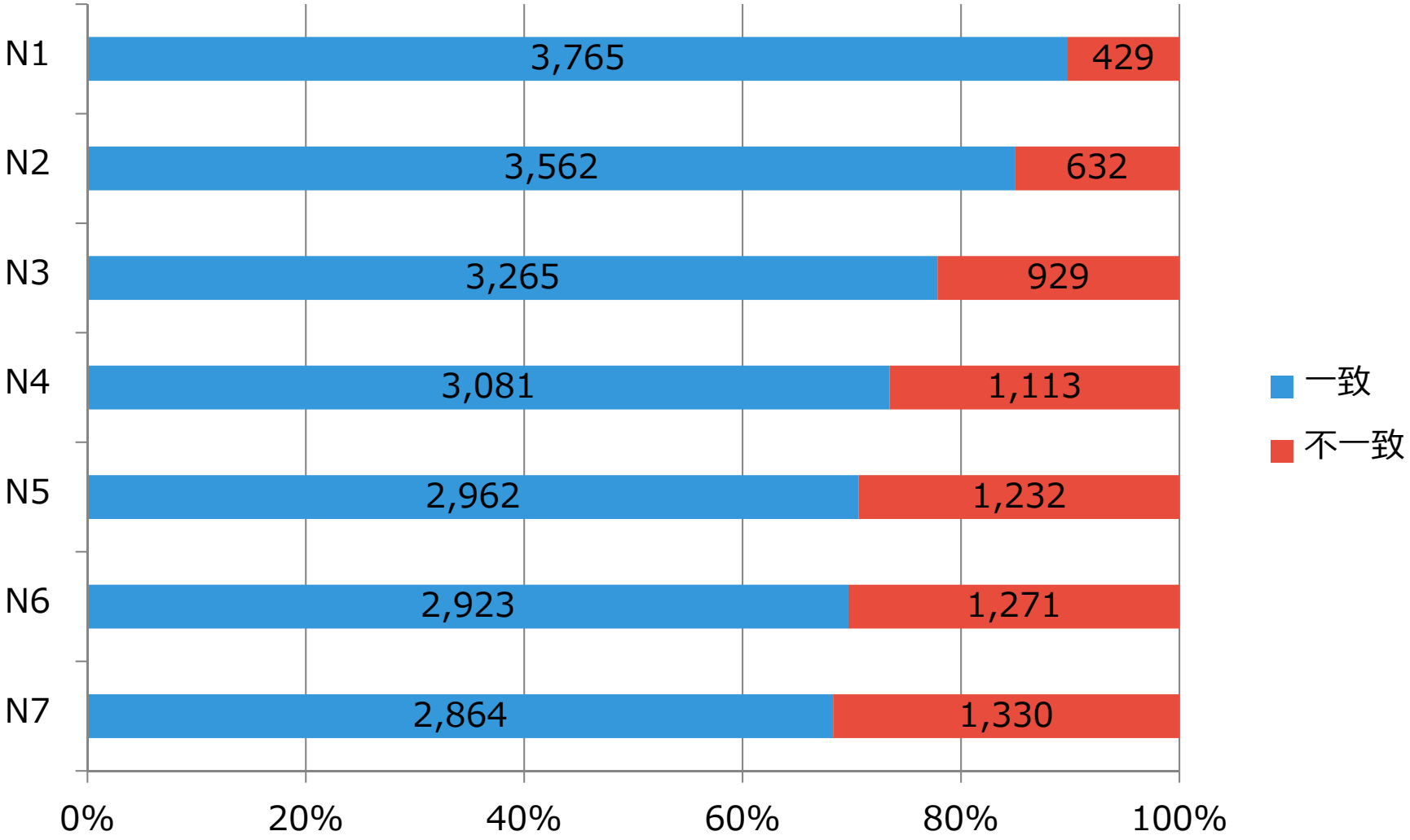
- fingerprint(string s)
  - 冒頭末尾空白・制御文字など削除，小文字化，ソート



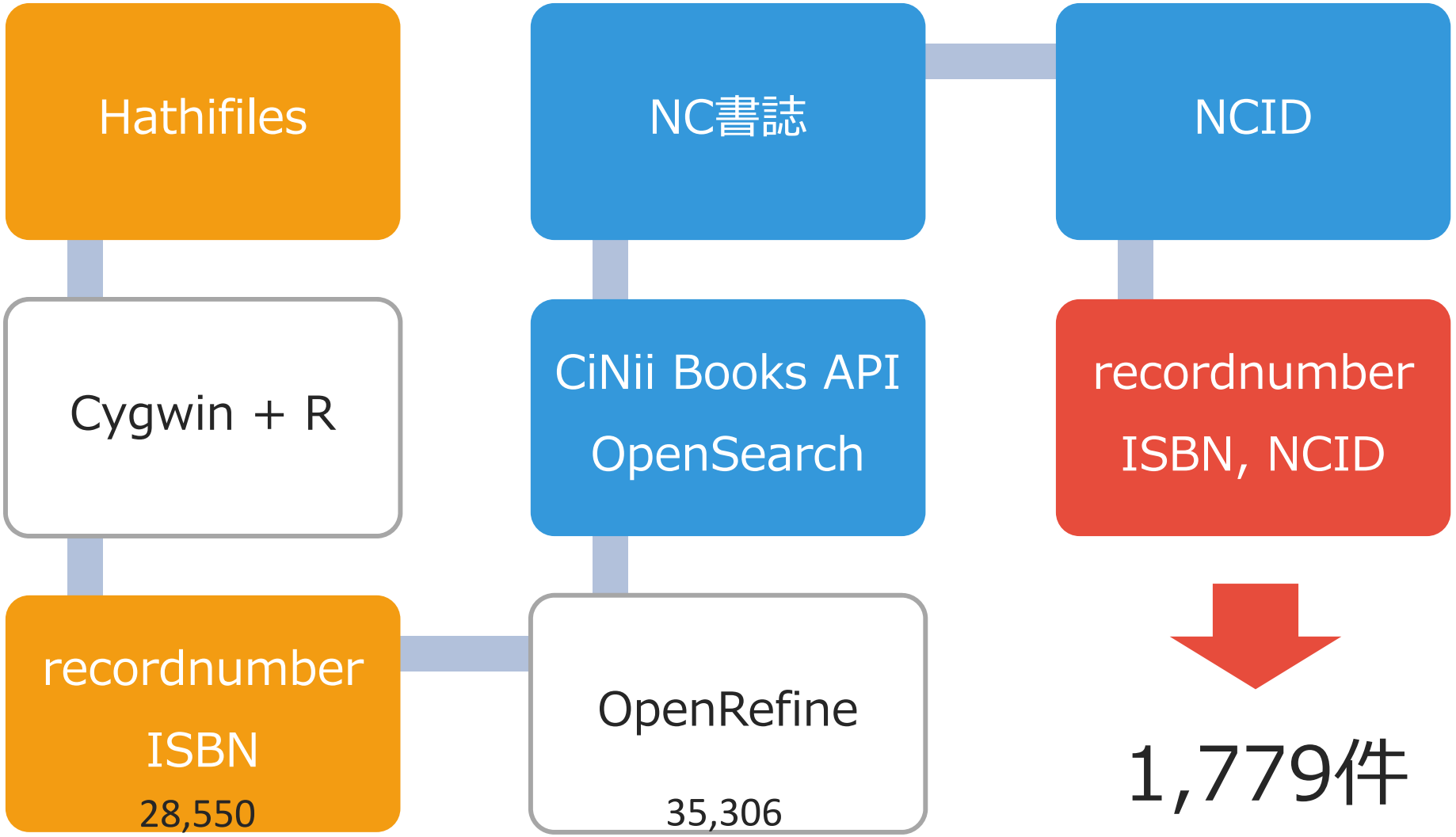
fingerprint(value)

https://github.com/OpenRefine/OpenRefine/wiki/GREL-String-Functions

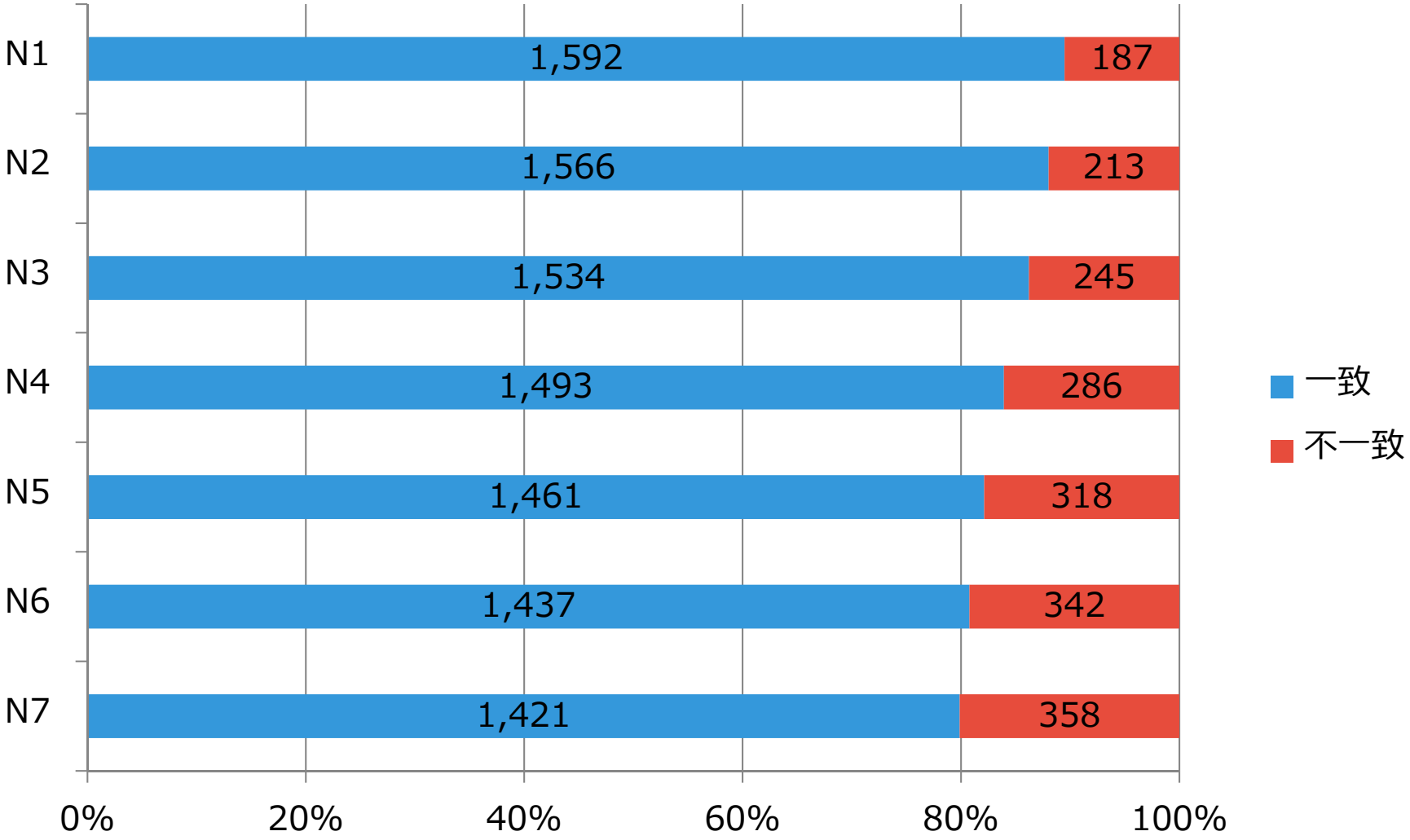
4.2.3. ISSNによる書誌同定―検証



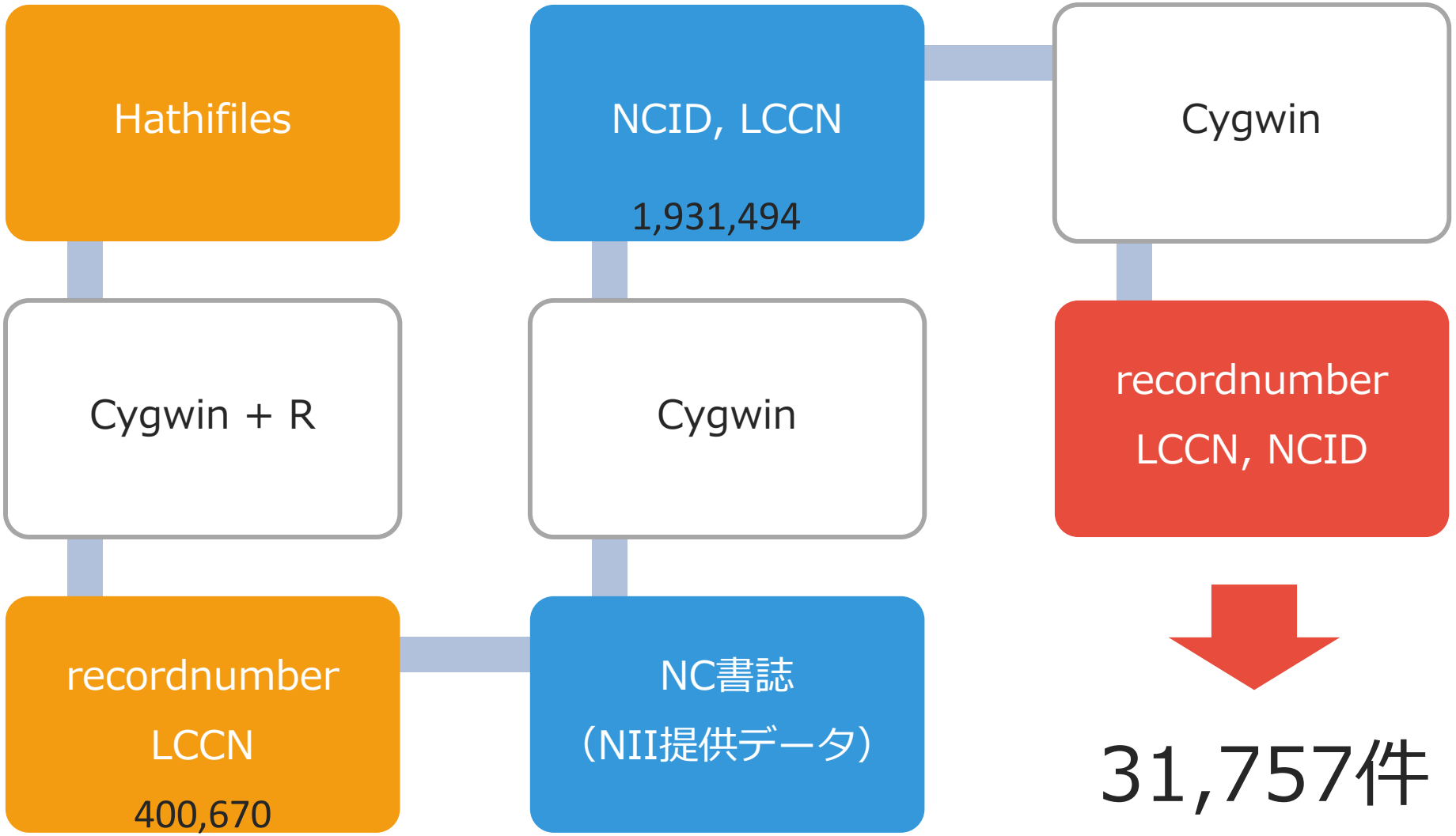
4.2.4. ISBNによる書誌同定



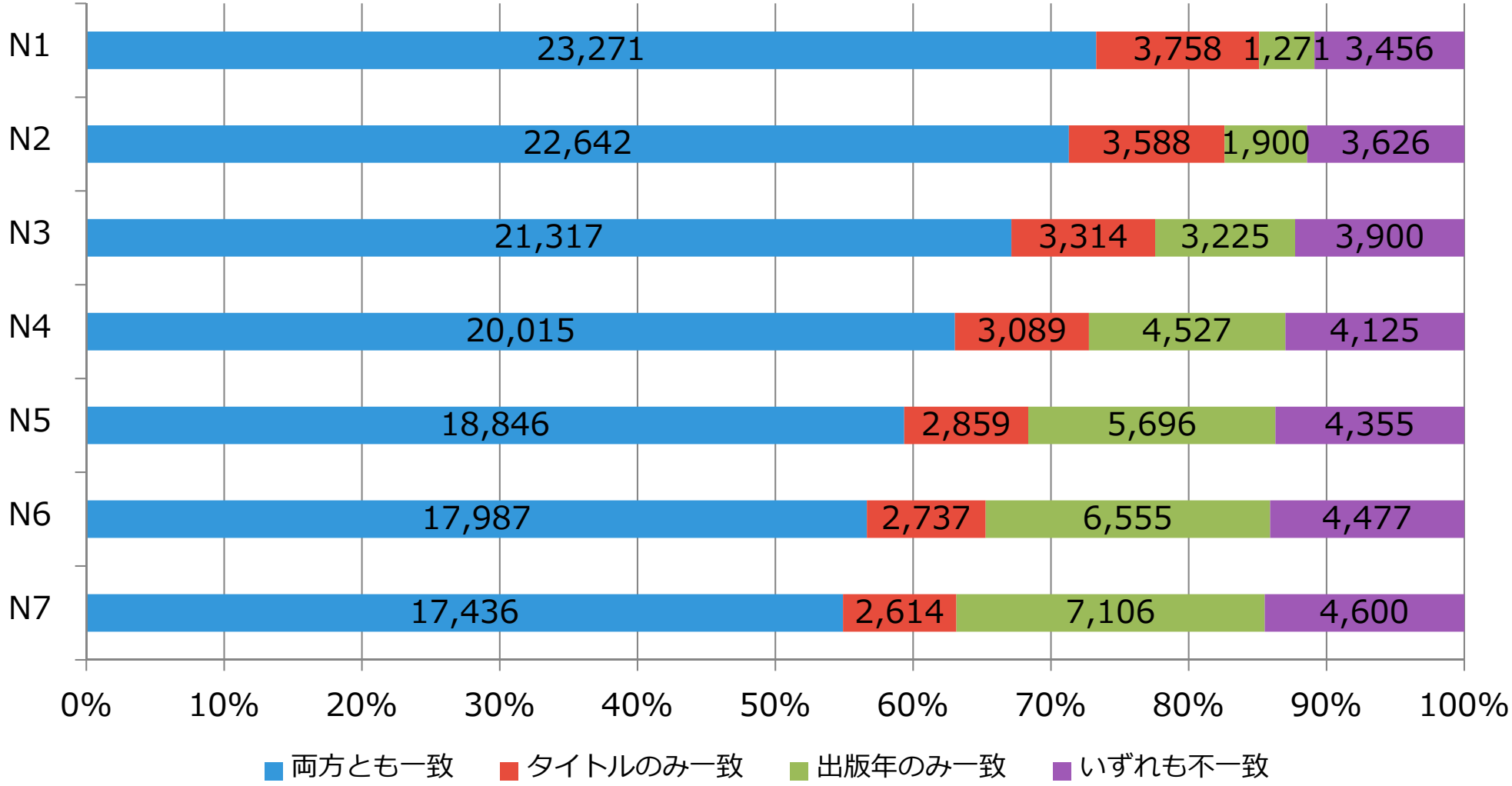
4.2.4. ISBNによる書誌同定ー検証



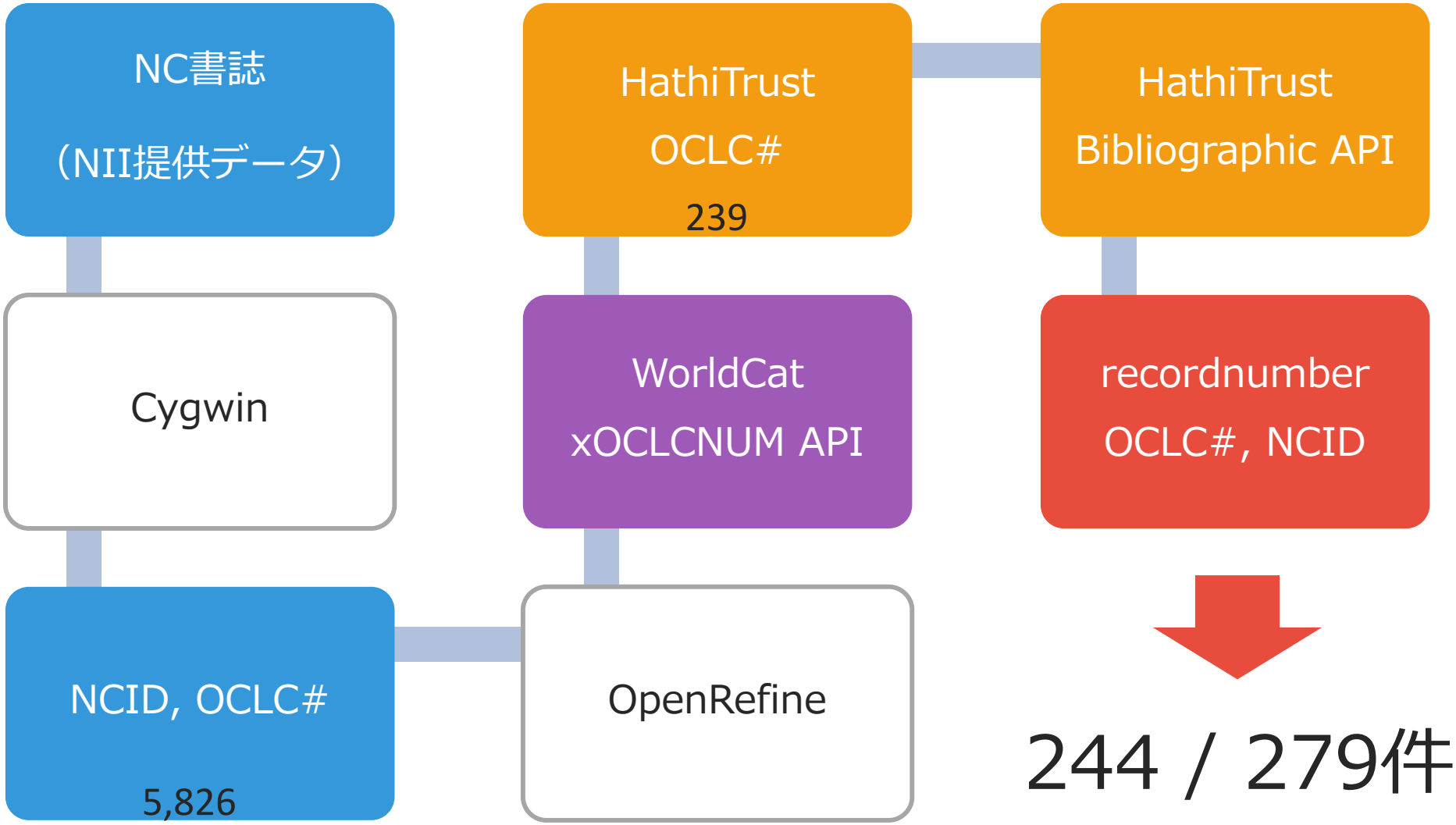
4.2.5. LCCNによる書誌同定



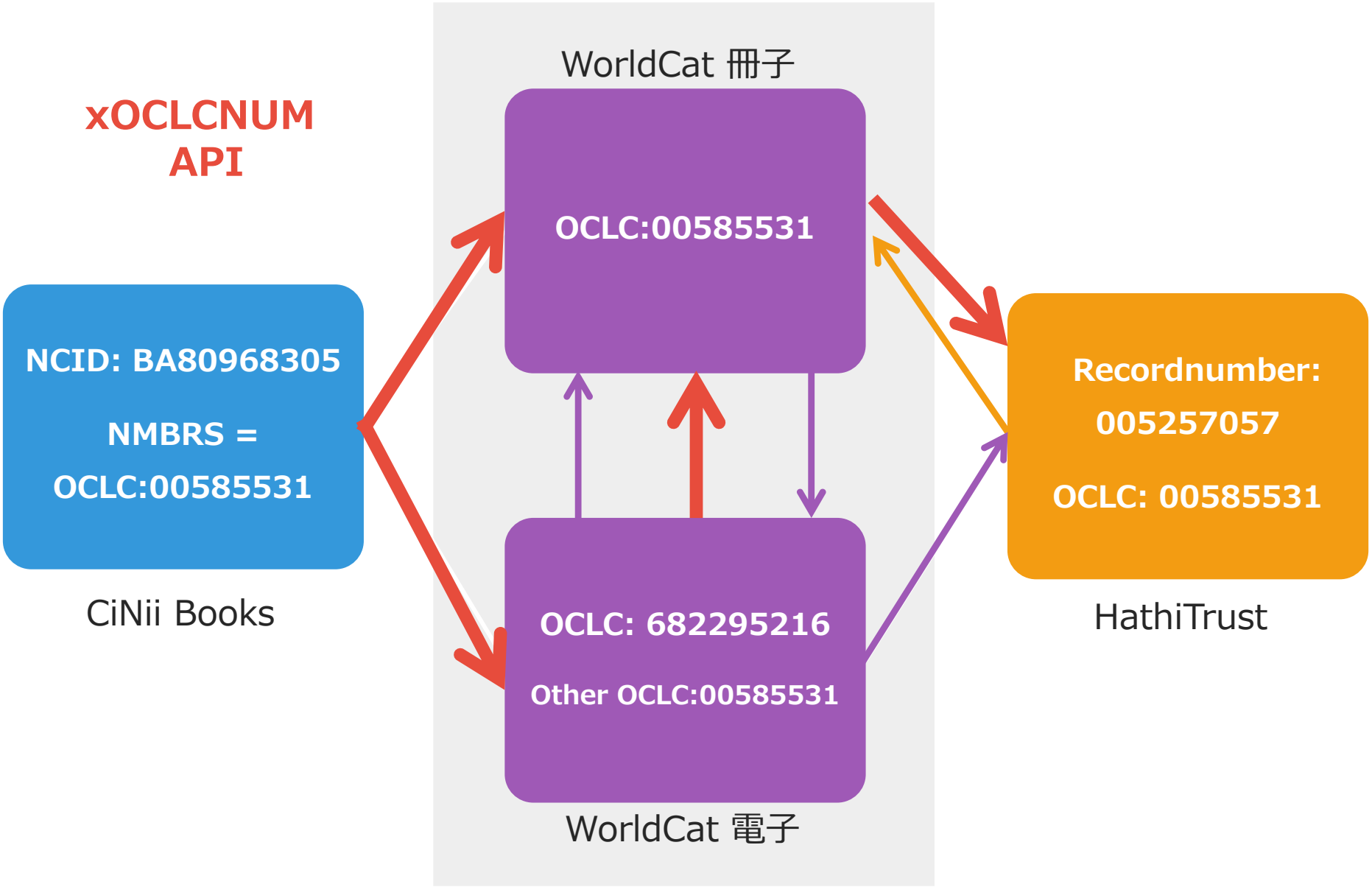
4.2.5. LCCNによる書誌同定ー検証



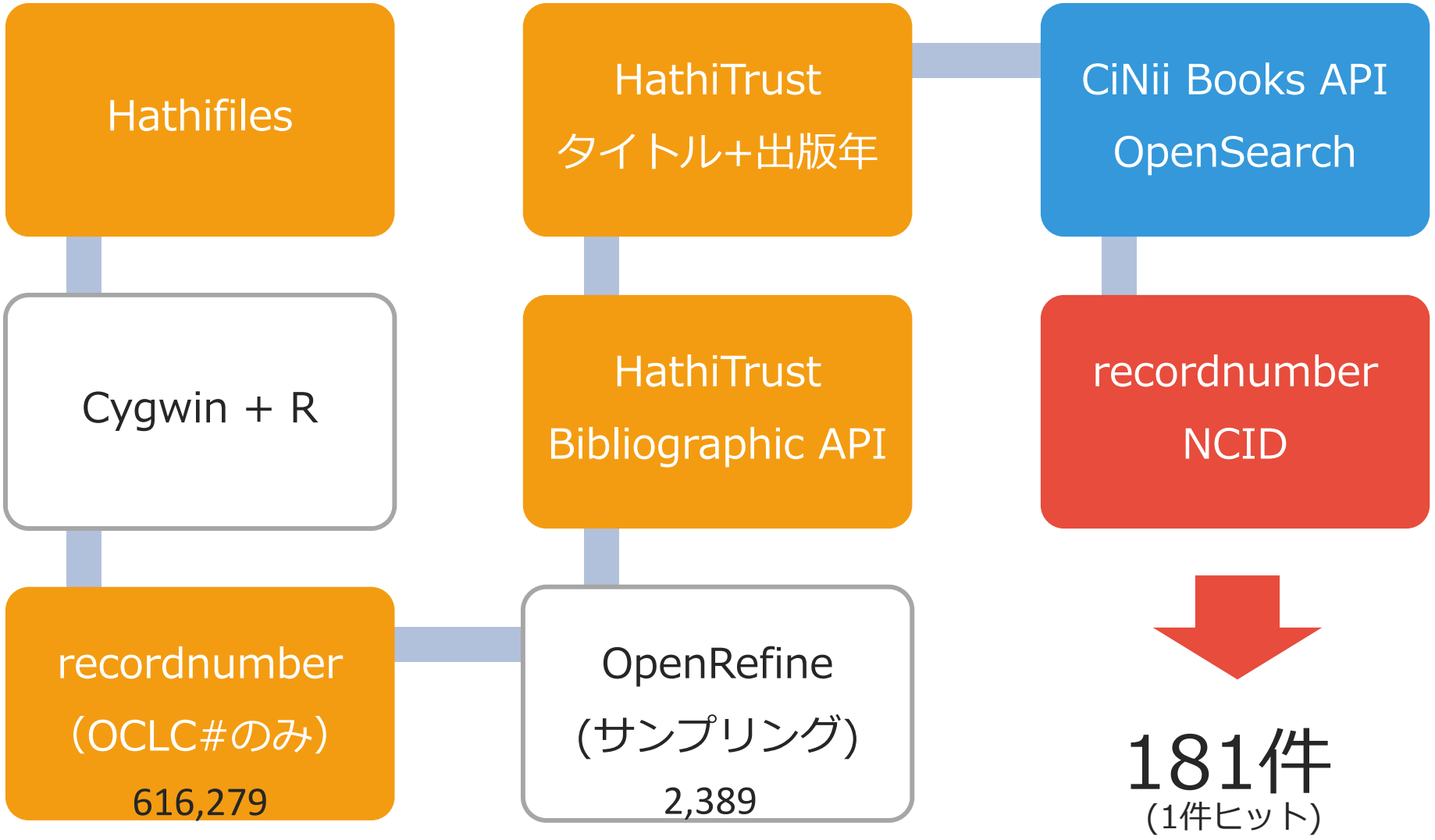
4.2.6. OCLCnumberによる書誌同定



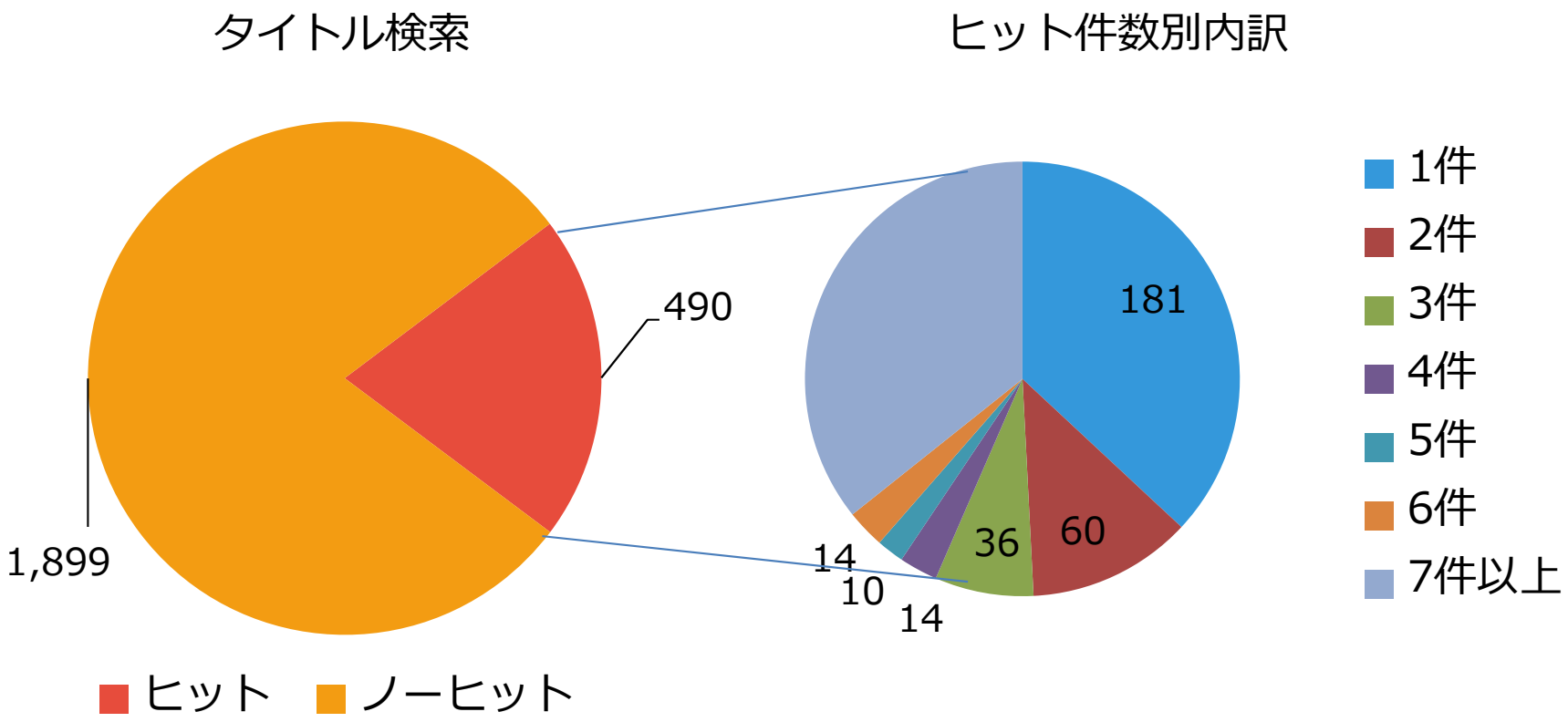
4.2.6. OCLCnumberによる書誌同定



4.2.7. タイトルによる書誌同定

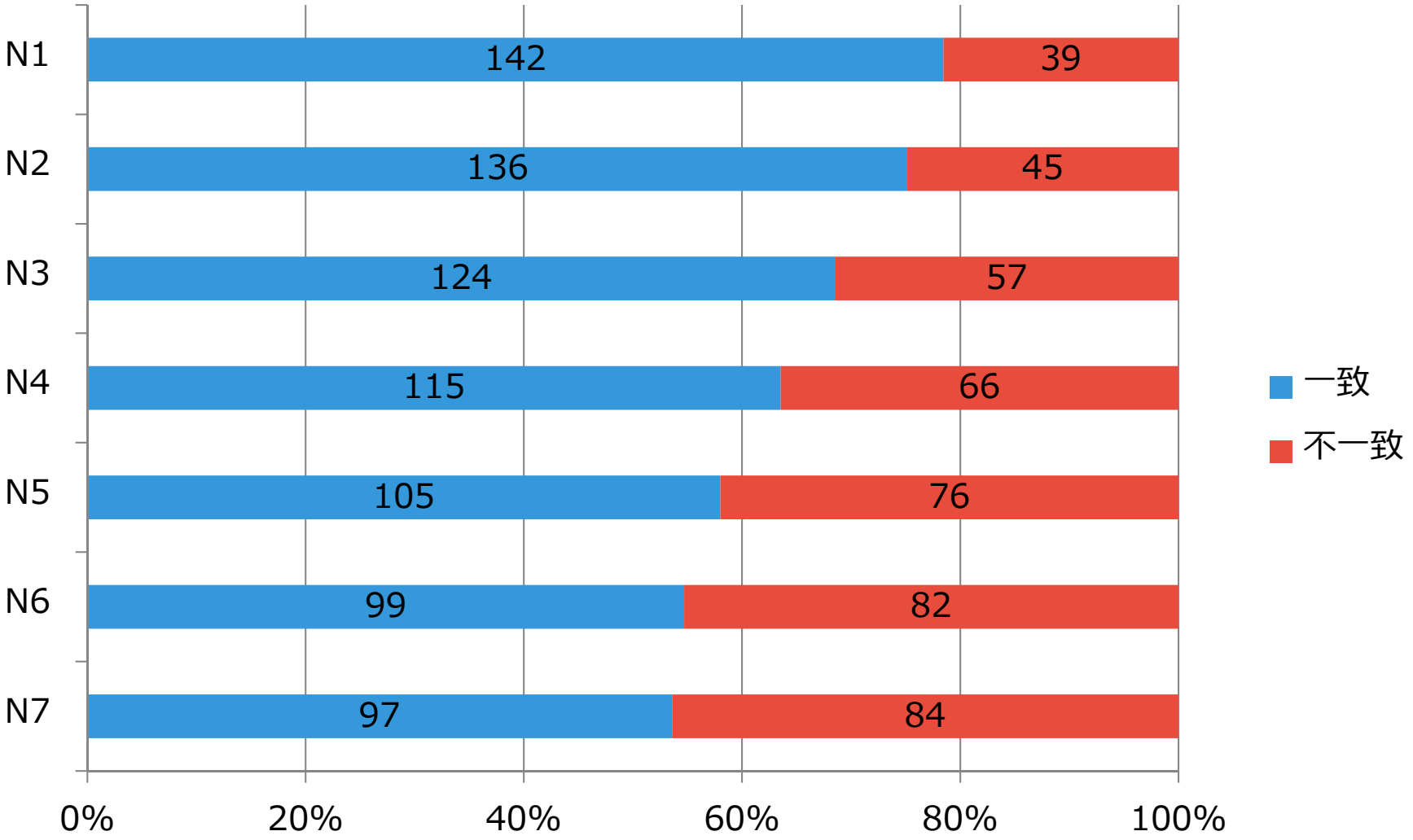


4.2.7. タイトルによる書誌同定―検証



ヒットした場合，3件までのヒット数が5割を超える

4.2.7. タイトルによる書誌同定―検証



4.2.8. HathiTrust分析のまとめ

- IDによる同定可能な書誌数はLCCNが最も多い
- 検証結果からISBN，ISSNの同定はLCCNやOCLCnumberと比較して精度が高いと推測
- 流用入力時に入力された明らかに異なるIDはタイトルや出版年を組み合せて検証することによって推測は可能
- タイトルと出版年のみによる同定はIDを用いた場合と比べて同定率は減少するが同定可能な書誌もある
- 日本語コンテンツは少ないが，慶應義塾大学データの投入以後再度分析する必要がある

4. 課題(1): ID分析―最終分析結果

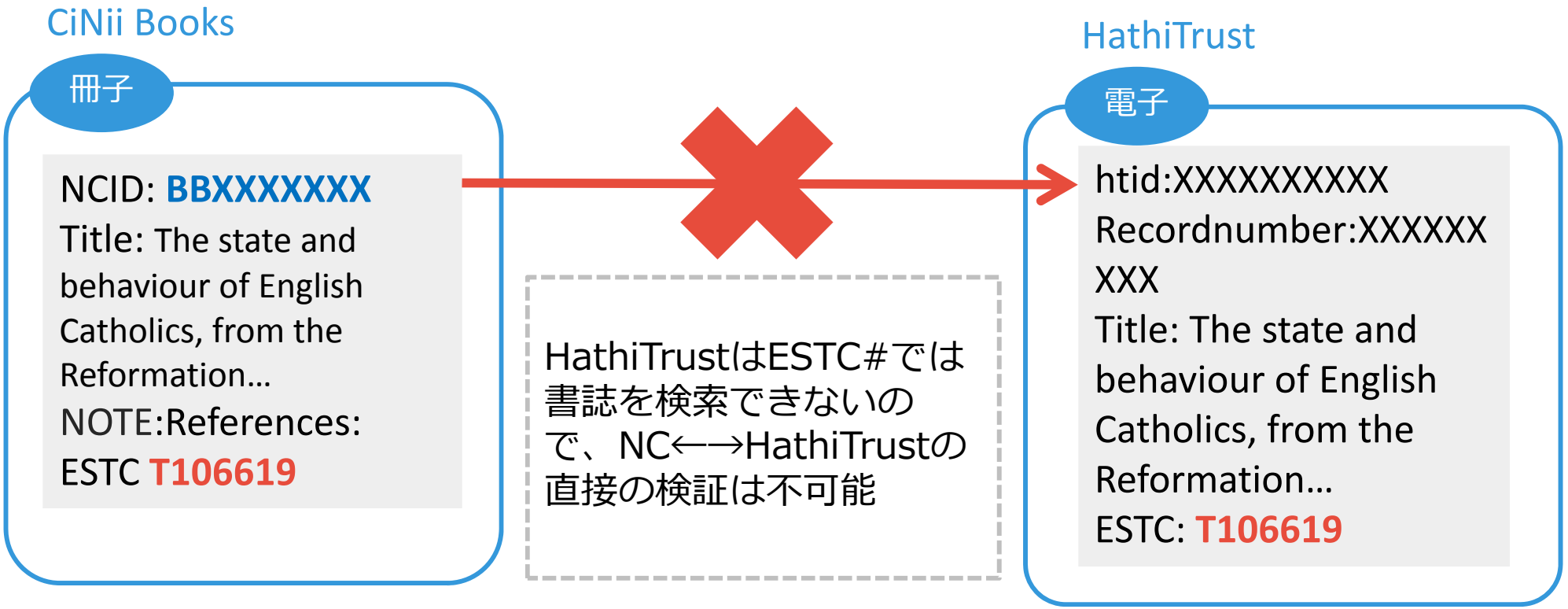
4.3. ESTC番号を含むNC書誌からタイトルでWorldCatを通じてHathiTrustにリンク形成の検討

4.3. ESTC番号を含むNC書誌からタイトルでWorldCatを通じてHathiTrustにリンク形成の検討

1. 検討内容と前提
2. 検証方法
3. 具体的な検証方法
4. 使用したWeb APIの種類と特徴
5. 検証結果
6. ESTC#分析のまとめ

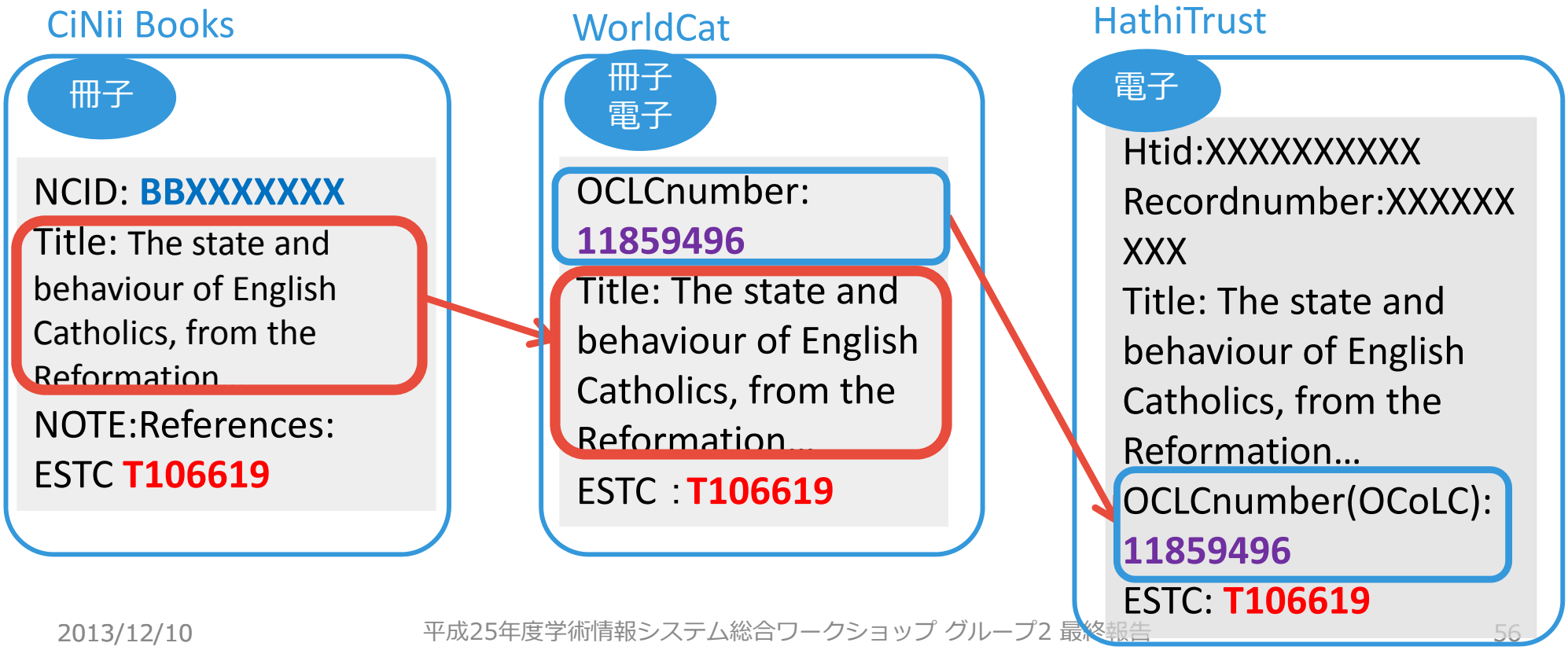
4.3.1. 検討内容と前提

NC書誌に記述されている「ESTC番号」をキーにしてHathiTrustとリンクができるか検証する

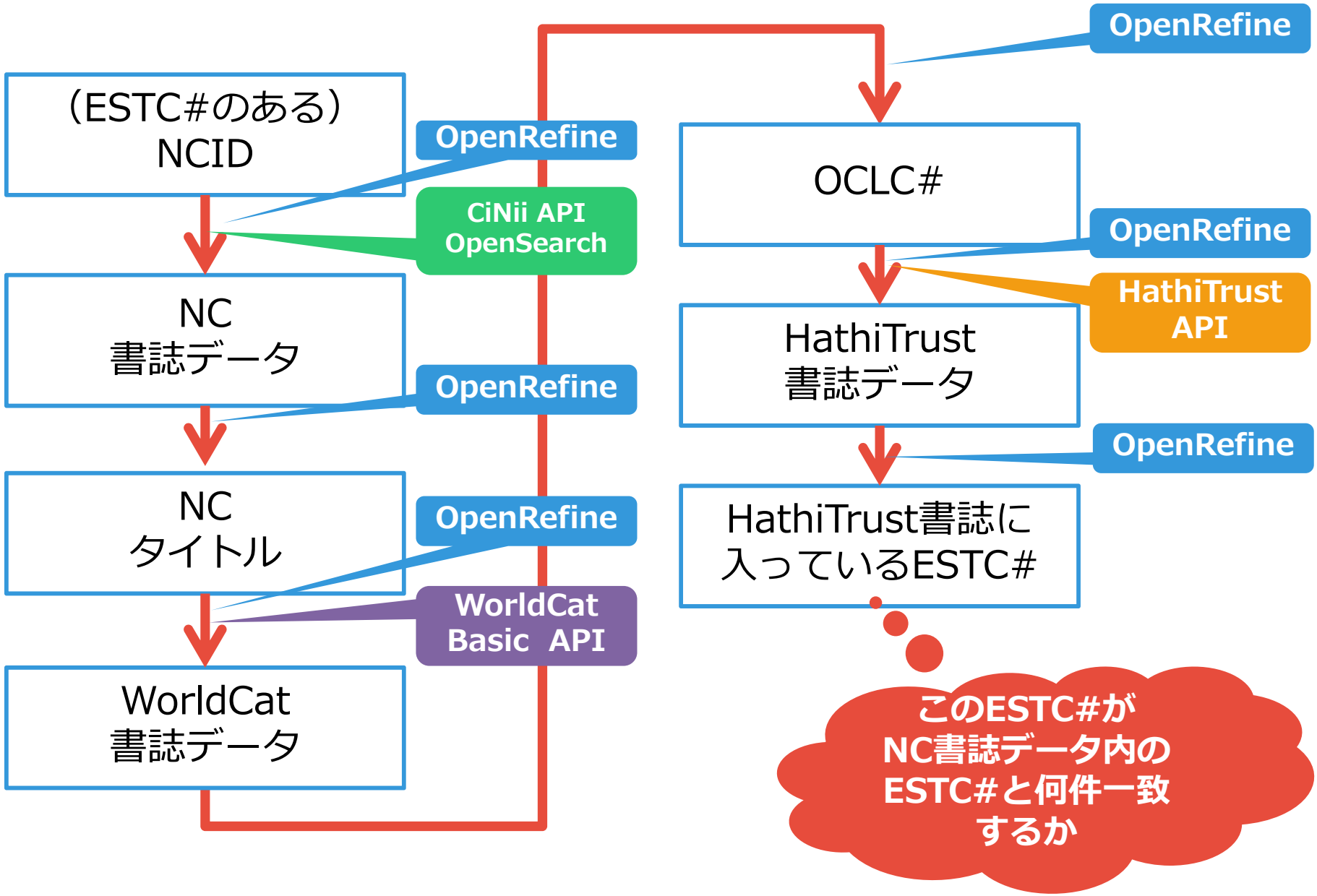


4.3.2. 検証方法

- NCとHathiTrustとの間にWorldCat APIを利用しOCLC#を取得
- 取得したOCLC numberをHathiTrustで検索
- HathiTrustの書誌データはOCLC number, LCCNを持つ割合が高い（WS中間発表・大西さん）



4.3.3. 具体的な検証方法



4.3.4. 使用したWeb APIの種類と特徴

名称	形式	特徴
CiNii API *CiNii Books 図書・雑誌検索OpenSearch	OpenSearch	• クエリのパラメータ種類が豊富 • NCIDやTitleなど検索項目を指定しての検索が可能
WorldCat Basic API	OpenSearch	• フリーワード検索しかない • 1日1,000件の問い合わせが限度
HathiTrust Bibliographic API	--	• 以下のIDでのみ問い合わせ可能 oclc, lccn, issn, isbn, htid, recordnumber • タイトル，キーワードでの検索はできない

使用したツール：OpenRefine, Excel

4.3.5. 検証結果(1)

- 提供データの確認（重複の削除）

	総数	備考
ESTC番号整理データ（鳥谷さん提供）	2,653	
重複削除後	2,539	*114件削除

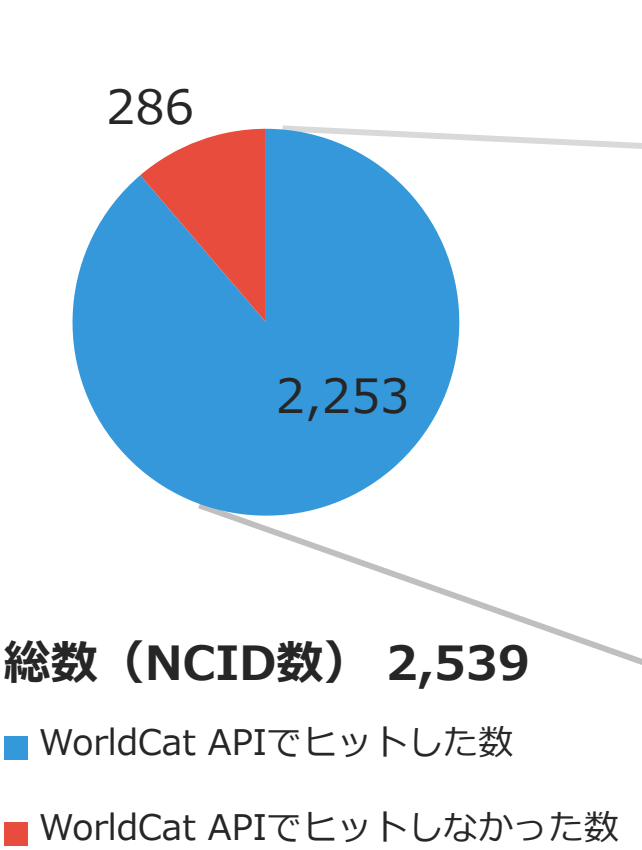
- 検証結果

項目	件数
ESTCをNOTEに含むNCIDの総数	2,539
OCLC APIでタイトル検索してヒットした数	2,253
取得できたOCLC番号の総数	13,171
HathiTrust APIでOCLC番号をキーに検索し書誌データがヒットした結果(=A)	629
Aのうち，ESTC番号を含んでいた件数(=B)	55
ESTC番号をもっている割合(=B/A)	9%
ヒットした(=A)うちESTC番号がNCのESTC番号と同一(*完全一致のみ)	**18**
ヒットした(=A)うちタイトルがNCと同一(*完全一致のみ)	81
ヒットした(=A)うちESTC番号とタイトルがNCと同一	3

4.3.5. 検証結果(2)

WorldCat Basic API

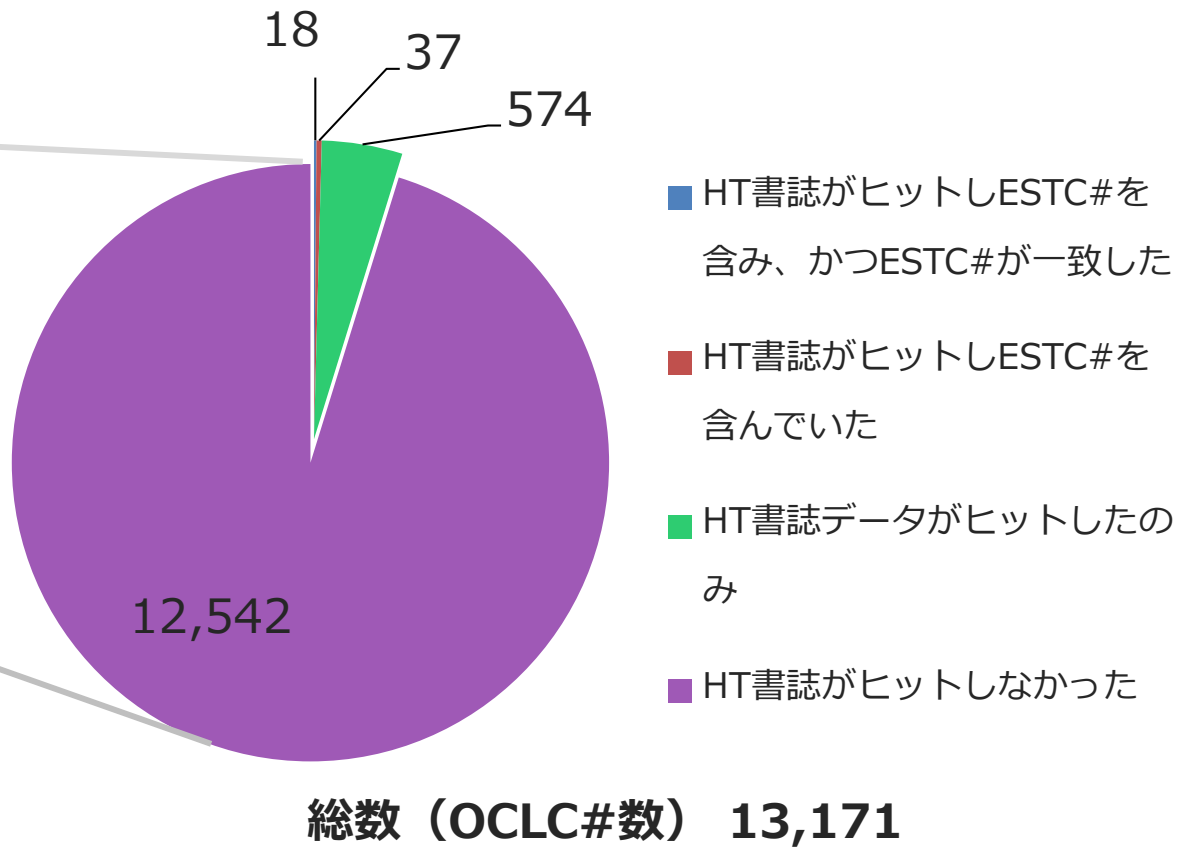
問い合わせ結果



総数（NCID数） 2,539

HathiTrust Bibliographic API

問い合わせ結果



総数（OCLC#数） 13,171

4.3.6. ESTC#分析のまとめ

- HathiTrustのメタデータがそもそもESTC#を保有していない（今回の検証で約9%）
- 保有していたとしても番号が完全一致するのは約32%，ほぼ一致するのは約43%
- ESTC#よりタイトルをキーにするほうが，ヒット率は若干だがあがる（約12%）

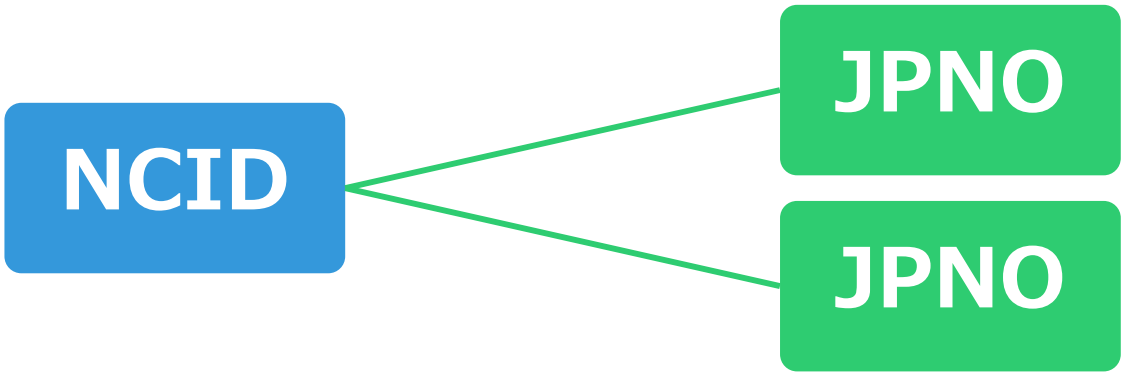
ESTC#をキーにしたWorldCat経由でのHathiTrustのリンクの可能性は困難

5. 課題(2): IDマップの作成

5. 課題(2): IDマップの作成

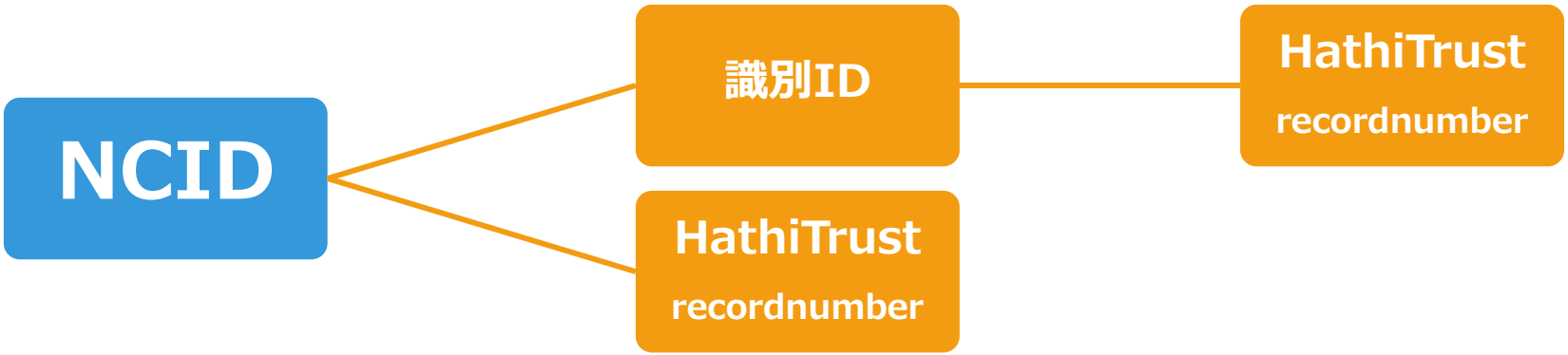
1. NCID—JPNO
2. NCID—HathiTrust recordnumber
3. ESTC#の場合
4. IDマップのメンテナンス

5.1. NCID—JPNO



あったほうがよいと思われる情報	リンクの信頼性（クリック数カウント）	デジタル化状況
• 親子の別 • 親書誌NCID • 複製か否か • 和古書かどうか	• 承認 • 未承認 • 関連（表示可）	• デジタル化の有無 • 公開状況 • **図書館送信対象**

5.2. NCID—HathiTrust recordnum



あったほうがよいと思われる情報	リンクの信頼性（クリック数カウント）	公開状況
• 親子の別 • 親書誌NCID • 複製か否か • 識別ID	• 承認 • 未承認 • 関連（表示可）	• public domain • copyright

5.3. ESTC#の場合

NCID	HathiTrust recordnumber
BA84441556	007652501
BA84512235	007662635
BA84771732	007702558
BA84833217	007669744
BA85090774	001746912
BA85731711	001919548
BA8577523X	001734887
BA90064988	001381481
BB04184562	000107180
BB0430135X	000108262
BB04583717	002203409
BB04599896	001139777
BB04821382	001748292
BB0490715X	001396889
BB05219990	001734887
BB05286998	001164253
BB09905865	001959904
BB10908135	001603508

IDマップ作成において，NCIDとHathiTrustのどのIDを対応させるか。

- htid=volume ID
  =University of Michigan record number
- Recordnumber=書誌ID

①NCID←→htid の場合

○ 書誌から直接電子化コンテンツに結びつけることができる。

△ 版違いに気が付かない。

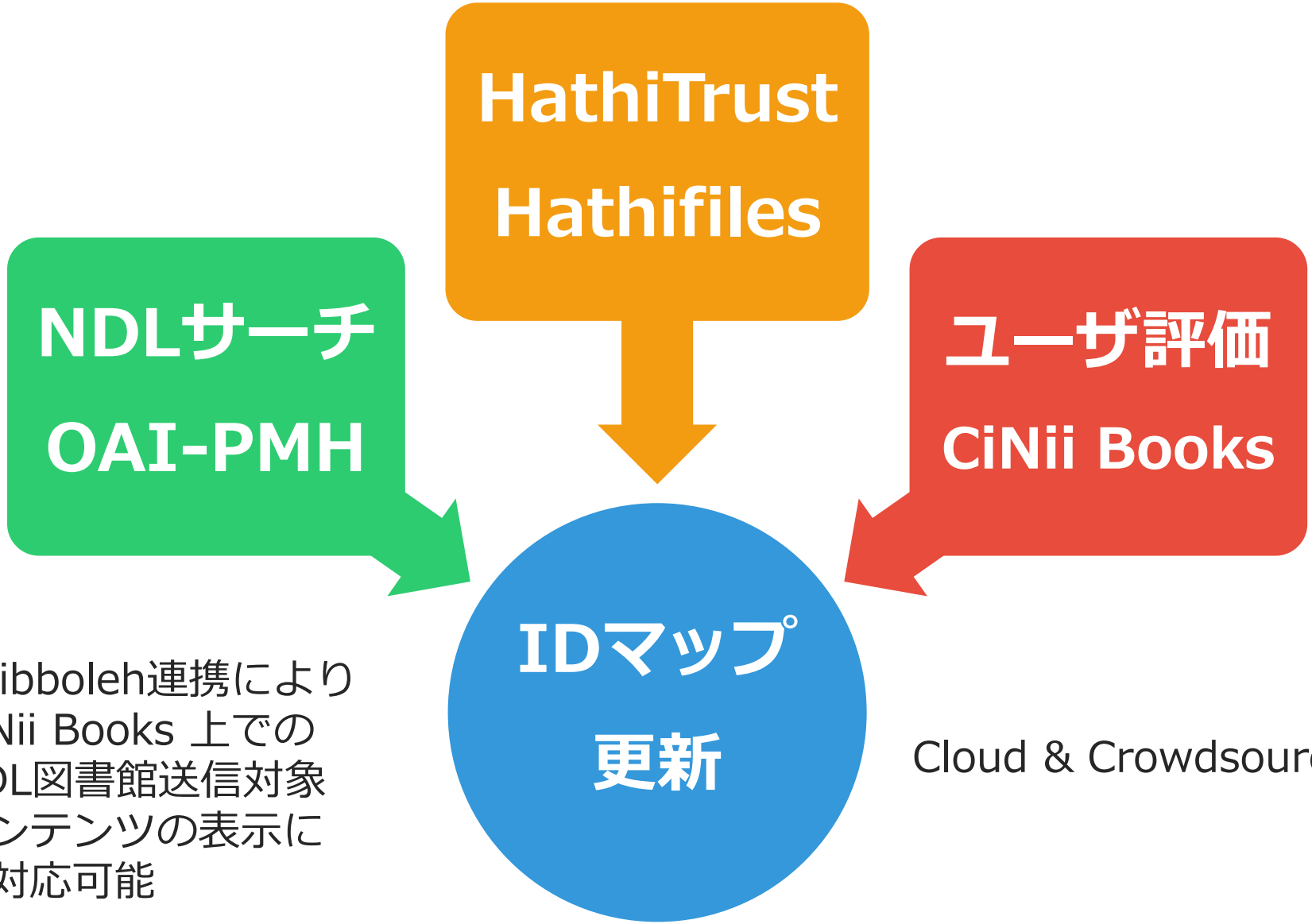
△ コンテンツが重複している場合は？

②NCID←→Recordnumberの場合

○ 書誌データ←→書誌データのパラレルなリンク関係で，版違い，関連書誌等発見しやすいか。

△ NCのVol積みはどう対応？

5.4. IDマップのメンテナンス



Shibbolеh連携により
CiNii Books 上での
NDL図書館送信対象
コンテンツの表示に
も対応可能

Cloud & Crowdsourcing

6. 課題(3): IDマップの活用法

6.1. デジタル化資料へのリンク

- CiNii Booksでデジタル化資料へのリンク表示
- 資料保存の効率化
  - 所蔵資料のNDLでのデジタル化状況の確認が容易に
  - 所蔵資料の中でデジタル化すべきものが把握できる
    →コスト削減
- 目録作成時，同定判断の一助に
- ILL業務の効率化
  - ユーザのデジタル化資料に対する可視性が向上

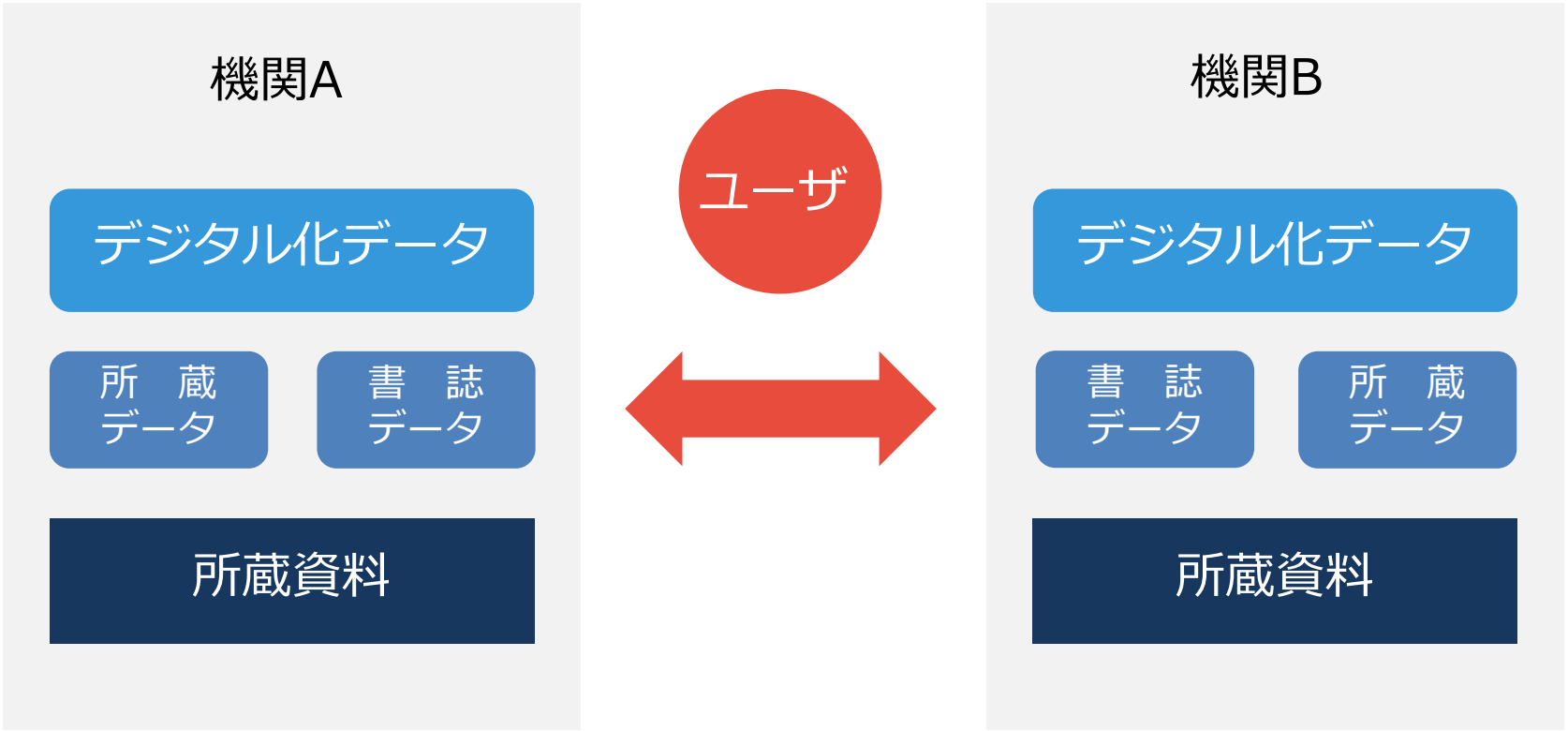
6.2. 他サービスとの連携

- 書誌単位で件名の相互補完
  - BSH ⇔ NDLSH ⇔ LCSH
- 書誌レコードを通じた典拠DB間の連携
  - NC典拠 ⇔ NDL典拠 ⇔ NDL Authorities ⇔ VIAF ⇔ ORCID
- ディスカバリサービスのセントラルインデクスにおける書誌統合のための情報提供

7. まとめにかえて
―デジタル化資料の海の中で

7.1. 図書館が担うべき機能

各図書館/組織	書誌・所蔵データ作成 所蔵資料のデジタル化
図書館/組織間	「つなぐ」ための仕掛け ユーザへの提供



7.2. 図書館員に求められるもの

デジタル空間のナビゲータとして

新たな技術動向のキャッチアップ

技術を有効活用していくための仕組みを知る，仕掛けを作る

利用を最大化するために

未来のユーザ，他機関のユーザを想定する

他機関の利用を想定する

目録＋機関相互での識別のための書誌データ作成を