

CiNii Books とデジタル化資料データベースとの連携をめざして—ID マップを利用したデータベース間リンクの可能性

大西 賢人, 鳥谷 和世, 柴田 育子

抄録：国立国会図書館デジタルコレクションや HathiTrust などの大規模デジタル化データベースの登場によりデジタル化された様々な学術資料へのアクセスが可能になってきているが、CiNii Books からは、デジタル化され公開された資料があったとしても、たどりつけないケースが見受けられる。そこで CiNii Books からデジタル化資料へのアクセシビリティ向上の可能性を探るために、総合目録データベースと NDL Search と HathiTrust それぞれに収録された ID 間の対応をまとめた ID マップの作成を試みた。調査の結果、JPNO や LCCN がデータベース間リンクのキーとなる有力な ID であることが判明したが、より正確で網羅性の高い ID マップを実現するためには、ID 以外の書誌情報も活用しながら、ID マップのプロトタイプを作成し、その有効性について検証を重ねていくことが必要である。なお、本稿は国立情報学研究所教育研修事業「平成 25 年度学術情報システム総合ワークショップ」における成果物である。

キーワード：CiNii Books, 総合目録データベース, NDL Search, HathiTrust, ESTC, API, ID マップ

1. はじめに

国立国会図書館デジタルコレクションや HathiTrust などの大規模デジタル化データベースの登場によりデジタル化された様々な学術資料へのアクセスが可能になってきている。しかし、日本の大学図書館等に所蔵される学術資料の検索サービスとして広く利用されている CiNii Books からは、これらデジタル化資料にはアクセスできず、ユーザの探し求める紙媒体の資料がデジタル化され公開されていたとしても、たどりつけないケースが見受けられる。そこで、CiNii Books からデジタル化資料へのアクセシビリティ向上の可能性を探るために、大規模デジタル化データベースと CiNii Books のベースになっている総合目録データベース（以下 NACSIS-CAT）の ID マップ作成とその活用方法を検討した。本稿の内容は、国立情報学研究所教育研修事業「平成 25 年度学術情報システム総合ワークショップ」における成果物である。

2. ID マップ作成を目指して

2.1 ID マップとは

本稿において「ID マップ」とは、NACSIS-CAT の書誌 ID である NCID と各データベースの識別 ID との対応関係を指す。

一般的にデジタル化事業では、著作権処理などの理由から出版年の比較的古い資料がデジタル化の対象となることが多く、紙媒体の原本は劣化や破損等の理由で利用を制限されている場合も多い。所蔵する紙媒体の代替資料としてデジタル化資料を有効に活用するためには、紙媒体の資料と関連付けてユー

ザに提示することが重要である。そのベースになるのが ID マップである。

ID マップ作成を目指し、筆者らは、NACSIS-CAT と国立国会図書館サーチ（以下 NDL Search）および HathiTrust に含まれる書誌データの比較分析を行い、各データベースがもつ書誌 ID、ISBN や ISSN などの ID（以下識別 ID）をキーにしたデータベース間のリンクの可能性を調査した。

2.2 本稿の構成

以下、NACSIS-CAT と NDL Search および HathiTrust を対象にした調査結果の報告を行うがそれに先立ちまず各調査で使用したツール類の紹介を行う。いずれの調査でも NACSIS-CAT の書誌データに含まれる ID をもとに、各データベースの API を使用して書誌データを取得、分析した。また HathiTrust については ESTC 番号によるリンクとそれ以外の ID によるものと、2 パターンに分けて調査を行った。後述の通り ESTC 番号をもつ資料群は 1800 年以前に刊行されたものであり、独立して調査を行うべきと判断したためである。

前述の通り ID マップの実体は別体系の ID の対応であり、各データベースがもつ ID がデータベース間のリンクに有効かどうかを検証するのが各調査の眼目となる。調査対象としたデータベースはそれぞれ構造も異なり、また NACSIS-CAT でもつ情報も ID 毎に保有率や管理フィールド等様々に異なるため調査はそれぞれ個別のアプローチをとることとなった。得られた結果も様々であるが、調査時点での現実を切り取った結果とご理解いただきたい。

3. 使用したツールとしての Web API とソフトウェア

今回の調査にあたり、様々なデータベースのメタデータ取得が必要となり、そのツールとして各データベースの提供する API を使用することとなった。ここにそれぞれ使用した API の特徴を述べる。

また、今回 Web API を使用するにあたり、OpenRefine を多用した。そこでこちらも紹介したい。

なお以下に記述する内容は基本的に調査実行時の 2013 年時点での情報をベースに 2014 年本稿執筆時に再確認したものである。

3.1 CiNii API¹⁾

CiNii API とは CiNii が提供する API で、今回は特に CiNii Books 図書・雑誌検索 OpenSearch を使用した。API を使用するにあたって、アプリケーション ID の取得が必要なため、事前に登録、ID の取得を行った。CiNii API の主な特徴としては、クエリのパラメータの種類が豊富であるということ、様々な検索項目を指定して検索が可能である点がある。例えば、フリーワードだけでなく、タイトル、著者名、注記などもキーワードとして検索可能で、AND 検索等組み合わせも可能である。

3.2 NDL Search API²⁾

国立国会図書館が提供する API で検索用 API とハーベスト用 API がある。利用にあたって登録等の事前手続きが不要なため気軽に利用できる。NDL Search では数多くのデータプロバイダが利用できるがデータプロバイダを特定して検索することも可能である。

3.3 WorldCat Basic API

WorldCat Basic API は OCLC が提供し、商用利用でなければ誰でも使用できる API である。CiNii API と同様、アプリケーション ID の取得が必要である。クエリは 1 日 1,000 件までと制限があり、クエリ自体はフリーワード検索のみなので、OCLC 参加館が利用出来る WorldCat Search API と比べて機能がやや限定されている。ただし、2014 年 6 月現在この API サービスは終了しているようである。

3.4 xID API³⁾

xID API は OCLC が提供する API で、Worldcat Basic API と同様、アプリケーション ID の取得と使用回数制限がある。ISBN, ISSN, OCLC number などから関連する ID やメタデータを取得する

ことが可能である。冊子体とデジタル化資料の書誌データが別に作成されているような場合でも xID API を使用することで両方の ID とメタデータを一度に取得することができる。

3.5 HathiTrust Bibliographic API⁴⁾

HathiTrust Bibliographic API は HathiTrust が提供する API で、誰でも使用が可能かつアプリケーション ID 取得が不要のサービスである。ただしクエリは OCLC number, LCCN, ISSN, ISBN, HTID (HathiTrust ID), record number の ID のみ使用可能で、タイトルやキーワードでの検索はできない。こちらも OCLC 同様、HathiTrust 参加館が使用できる HathiTrust Data API に比べるとサービスが限定された API である。

3.6 OpenRefine⁵⁾

OpenRefine とは以前 GoogleRefine という名で提供されていたデータクレンジングソフトである。大量のデータ処理や、データの整形等に用いられ、表記の揺れや重複の発見、削除、修正やデータの正規化などを行うことができる。また、API を通じたデータ取得も可能であり、今回の調査ではこの機能を活用した。

4. NDL Search と CiNii Books の連携可能性

4.1 NDL の ID と検証対象 ID

国立国会図書館の書誌データに含まれる主な識別 ID としては以下の 3 種類がある。

- ・JPNO (日本全国書誌番号)
- ・NDLBibID (国立国会図書館書誌 ID)
- ・NDLJP (国立国会図書館デジタルコレクションに収録された資料に付与される永続的識別子)

冊子のデジタル化資料を想定した場合、デジタル化資料の書誌にはそれ自身の識別子として NDLJP をもつと同時に原本資料の NDLBibID, JPNO をもつ。NDLBibID が最も網羅性の高い ID ではあるが、これは国立国会図書館の管理 ID であり、CiNii Books 側ではこれを把握していない。把握している可能性があるのは JPNO である。このため JPNO をキーにした連携の可能性について検証を行うこととした。

検証はリンクキーとしての量および質を確認するため、まず NACSIS-CAT 側での JPNO の保有率とその精度を確かめることから開始した。さらに実効的な ID マップ実現に向け機械的に質を改善する方法、また量を増やす方法についても検討を行った。

上記の結果を踏まえ本章の最後に ID マップ作成に向けた提案を行う。

4.2 NACSIS-CAT での JPNO 保有状況とその精度

4.2.1 検証対象

NACSIS-CAT でどの程度 JPNO を保有しているか、またその JPNO がどの程度正しいか、検証を行うため NII にデータの提供を依頼した。提供データの抽出条件は出版年が 1968 年以前、かつ本タイトルの言語が日本語の書誌のコードブロックの情報である。理由は国立国会図書館における図書のデジタル化対象が 1968 年までに受入れた資料であるからである。以下、NACSIS-CAT の書誌を NC 書誌、国立国会図書館の書誌を NDL 書誌と称す。

4.2.2 JPNO 保有状況

対象書誌数は 1,029,357 件（2013 年 8 月 6 日提供時）、うち NBN フィールドに JPNO をもつ書誌は 200,551 件であり約 19% の書誌が JPNO をもっていた。国立国会図書館と NACSIS-CAT では書誌の作成単位が異なるため、同じ JPNO を複数の書誌でもつ場合もあれば、複数の JPNO をもつ書誌もある。ユニークな JPNO 数は 193,778 件、JPNO の総数は 201,655 件であった。

4.2.3 JPNO の精度に関するサンプル調査

JPNO はコードブロックで管理されており、コードは厳密には記述されない可能性がある。NC 書誌で保有している JPNO がどの程度正確か、サンプル調査を実施した。なお下記の調査を含め、以降のサンプル調査はすべてデータ行番号をランダムに指定することで抽出を行った。

調査の実施時期は 2013 年 9 月から 10 月であり、以下の手順で行った。

1. 提供されたデータから JPNO を 500 件抽出
2. 上記 500 件の JPNO にて NDL Search を検索しヒット状況を確認
- 3.1. で対象とした JPNO をもつ NC 書誌 542 件と同 JPNO で NDL Search を検索して得た NDL 書誌を比較検証

500 件の JPNO による NDL Search の検索結果は表 1 の通りである。なお、4 章すべての調査に共通することであるが、基本的に書誌データのみの比較・参照による同定可否や重複の判断である点、ご理解をいただきたい。

表 1 JPNO500 件の NDL Search でのヒット状況

ヒット数	JPNO 数	当該 JPNO をもつ NC 書誌のいずれかに同定可能
0	2	
1	76	67
2	400	395
3 以上	22	19

一部例外はあるが、ヒット数が 1 件の場合は、デジタル化資料が存在しないケースである。同じく 2 件の場合は、図書書誌とデジタル資料が各 1 件存在するケース、3 件の場合は、図書書誌に対応するデジタル資料が複数存在するケースである。全体の約 96% について同定可能と想定される NC 書誌が存在する。

一方、NC 書誌 542 件と NDL 書誌を比較検証した結果は表 2 の通りである。

表 2 NC 書誌 542 件と NDL 書誌との同定可能性

JPNO の間違い	13
NDL に所蔵なし	36
NII 重複 (*)	8
同定可能	485

* 重複していると思われる別書誌で NDL 書誌との比較検証を行っている

重複書誌を除く 534 件のうち約 91% にあたる 485 件が NDL 書誌とおそらく同定可能である。うちデジタル化されているものは 418 件、公開されているものは 94 件である（2013 年 10 月時点）。同定可能と判断した書誌には NC 書誌に対応する NDL 書誌が複数存在し、その一部にのみリンクするケースも含んでいる。

4.3 不適当な JPNO 排除の可能性

NC 書誌がもつ JPNO が必ずしも正確ではない状況を鑑み、不正確な JPNO を機械的に排除する方法はないか検討してみた。NC 書誌と NDL 書誌では書誌の作成単位が異なるため、単純なタイトル比較では同定不可か判断できないが、NDL 書誌には本タイトル以外に別タイトル、部分タイトルなどの情報も含まれるため、作成単位の差によるタイトルの違いはある程度吸収できる可能性がある。重複の可能性のある 8 件を除いた 534 件につき双方の書誌でタイトル比較を行った結果を同定の可否別にまとめた表を以下に示す。なお、比較にはタイトル関連情報等を含まない本タイトルのみを用いた。

表3 NC 書誌と NDL 書誌のタイトル比較結果

		同定可能	同定不可
タイトル一致		362	21
タイトル不一致	字体の差	81	9
	階層の差	28	0
	タイトル採用部分の差	14	1
	JPNO でノーヒット	0	2
	タイトル情報不一致	0	16

同定の可能性の高い 485 件のうち、NC 書誌と NDL 書誌との相互でタイトルが一致しないものが 123 件（約 25%）あるが、字体の差、階層の差によるものが 109 件である。階層の差による場合、NDL 書誌の巻次、別タイトル、部分タイトルのいずれかに NC 書誌で採用された本タイトルと一致する情報が、字体が異なる場合もあるが、含まれている。タイトル採用部分の差は、記号の有無やその違い、タイトルに含まれる単語の有無による差である。

また同定できない 49 件も、そのうち 21 件（約 43%）がタイトルは一致する。字体のみ異なる 9 件も含めると約 61% が NC・NDL 相互で同じタイトルをもつ。タイトル等が類似する別資料の書誌からの流用入力の結果であろう。残る 18 件については、保有する JPNO で NDL 書誌にヒットしない場合も含め、タイトル情報は一致しない。

JPNO で紐づけられた NC 書誌と NDL 書誌について、字体の差を吸収する形で本タイトル以外のタイトル情報をも含め比較し一致しないものを弾くようにすれば表 3 の灰色部分を排除することになる。JPNO の誤りがある程度は排除できると思われる反面、タイトル採用部分が異なるケースなどでは一方のタイトルにのみ含まれる文言もあり正しい JPNO を除外する結果となる。ID マップに要求される精度がどのレベルか、その目標を定めた上で対応を検討すべきことではないかと考える。

4.4 JPNO 補完の可能性

4.4.1 サンプル調査による検討

4.2.2 に記した通り、国立国会図書館における図書のデジタル化対象年代に相当する日本語資料の NC 書誌のうち JPNO をもつものは 2 割弱である。国立国会図書館といえども NC 書誌が存在する資料の全てを網羅していることはないにせよ、実際に所蔵している割合はもっと高いはずである。保有率を引き上げるため、100% の正確性を追求するのは無理としても、ある程度信頼できる JPNO を機械的に把握する方法はないか、以下の手順でサンプル調

査を実施した。

なおサンプル調査に際しては NDL のデジタル化資料と NACSIS-CAT の照合を目的として、国立情報学研究所が NDL から提供を受けたデータを利用した。

4.4.2 NC 書誌をベースにしたサンプル調査

調査は以下の手順で行った。

1. JPNO を保有しない書誌 828,806 件から 500 件を抽出
2. 上記書誌の NCID から NC 書誌のタイトル取得
3. 2. で得たタイトルを用い NDL Search に対しデータプロバイダを NDL-OPAC および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、書誌を比較
5. 1. の 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
6. 3. および 5. で使用した同じ情報で CiNii Books 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

NC 書誌のタイトルおよび出版年を利用した検索結果は図 1 の通りである。出版年との AND 検索を行うことでノーヒットとなる件数が増える反面、1 件のみに絞り込まれる件数は増加する。同定の可能性も高くなることが期待される。

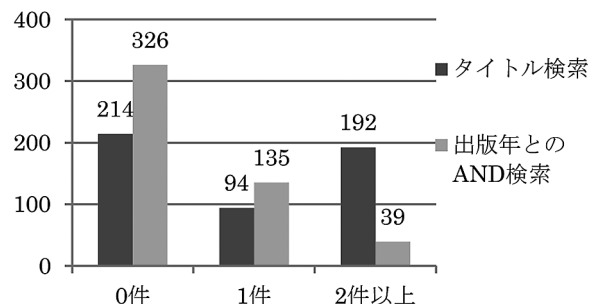


図1 NDL-OPAC ヒット状況

NDL-OPAC ではヒットせず、デジタル化資料に 1 件ヒットするケースがタイトル検索で 1 件、出版年との AND 検索で 2 件あるが、それを含め、タイトル検索で 1 件ヒットした 95 件（約 19%）、出版年との AND 検索で 1 件ヒットした 137 件（約 27%）につき、NC、NDL 双方の書誌を比較した。

同定の可能性についてそれぞれの件数、対全体比をまとめたのが表4と表5である。和古書については個別書誌作成が基本であるため NC 書誌の注記にその旨記載のある書誌について別にカウントした。なお右端は手順の6で確認した、同タイトルの NC 書誌が複数存在する書誌数である。

表4 タイトル検索での書誌比較結果

	件数	全体比	タイトル重複件数
和古書（個別書誌作成）	6	6%	4
同定可能	72	76%	10
同定不可	17	18%	5

表5 出版年との AND 検索での書誌比較結果

	件数	全体比	タイトル重複件数
和古書（個別書誌作成）	6	4%	6
同定可能	120	88%	11
同定不可	11	8%	6

4.4.3 NDL 書誌をベースにしたサンプル調査調査の手順は以下の通りである。

1. NDL 図書デジタルデータ 894,275 件（書誌数 698,556 件）から 500 件を抽出（対象の 500 件と同一書誌のデジタルデータは 635 件存在）
2. 上記 500 件のタイトルから 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 に対し出版年とタイトルとの AND 検索について 2., 3. と同様に実施

NDL 書誌のタイトルおよび出版年を利用した検索結果は図2の通りである。この場合も出版年との

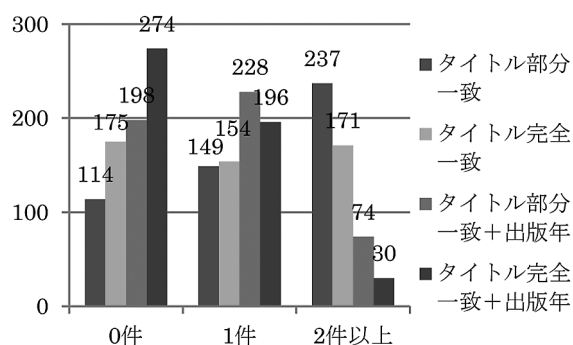


図2 CiNii Books のヒット状況

AND 検索を行うことで 1 件に絞りこまれる件数が増加する。

CiNii Books ではタイトル検索を部分一致で行うか完全一致で行うかを指定できる。部分一致検索・完全一致検索の双方で 1 件ヒットした場合、部分一致検索でのみ 1 件ヒットした場合、完全一致検索でのみ 1 件ヒットした場合、それぞれのケース毎に NC, NDL 双方の書誌を比較した結果が表6と表7である。同定の可能性について、件数および対全体比、および JPNO を既にもっている書誌数（()）はうち JPNO が間違っている書誌数）をまとめている。

表6 タイトル検索での書誌比較結果

	同定可能		同定不可		JPNO 有
部分一致・完全一致双方	98	92%	9	8%	49 (1)
部分一致のみ	32	76%	10	24%	21 (1)
完全一致のみ	41	87%	6	13%	19

表7 出版年との AND 検索での書誌比較結果

	同定可能		同定不可		JPNO 有
部分一致・完全一致双方	171	97%	5	3%	26
部分一致のみ	48	92%	4	8%	15 (1)
完全一致のみ	18	90%	2	10%	6

なお完全一致のみで 1 件ヒットした場合、同定可能と判断したものに親書誌にヒットしたものがタイトル検索で 15 件、出版年との AND 検索で 3 件含まれている。

4.4.4 JPNO 補完の可能性に関する検討結果

NC 書誌と NDL 書誌について、一方のタイトル・出版年の情報から他方を検索した結果を図3に整理した。

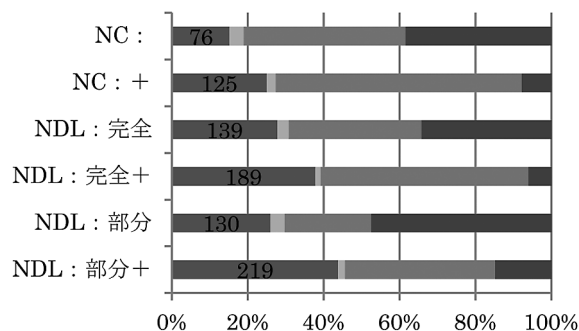


図3 NC 書誌・NDL 書誌の相互検索結果

左軸は検索に使用した情報が NC, NDL のいずれであるか、および検索方法を示している。+がないのはタイトルのみの検索であり、+があるのは出版年とタイトルの AND 検索である。グラフは左から、1 件ヒット (同定可能)、1 件ヒット (同定不可)、ノーヒット、2 件以上ヒットの割合を示しており、同定可能と思われる 1 件ヒットについては件数も表示した。

国立国会図書館では物理単位での書誌作成が基本であるためか、1 件に検索結果が絞り込まれる確率は NC 書誌からの検索の方が低い。またタイトルと出版年との AND 検索で 1 件ヒットした場合、書誌同定の可能性は完全一致、部分一致の別を問わず 9 割以上である。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 書誌を網羅していない現状を考えれば、タイトルと出版年の AND 検索で 1 件ヒットした場合、機械的に JPNO を取得することも一つの方法であると考ええる。

4.5 ID マップ作成に向けて

4.5.1 現有情報の精度

現に NC 書誌が保有している JPNO の精度が 90% 程度であること、またその保有率が 2 割弱にとどまることは前述の通りである。精度については不適切な JPNO をある程度、機械的に排除しうる可能性はあるものの、すべて排除することは不可能であり、また有効な JPNO を除外する可能性もある。保有率については現有する JPNO と遜色ない程度の精度であれば機械的に取得し保有率を引き上げることも可能だと考える。

ID マップとして許容される精度がどの程度か意見の分かれるところであろうが、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、その評価情報も合わせて提供すれば 100% 正確でなかったとしても、ユーザの理解は得られるのではないかと考える。ユーザの協力を得て ID マップの精度向上をめざす方法である。

具体的には ID マップをもとに CiNii Books の画面にクリックボタンを表示する際に、以下のような方法を検討する必要があると思われる。

- ・ユーザがリンク情報を評価できるようにする
- ・ユーザの評価状況に応じてリンクボタン自体の色や形を変化させるなど個々のリンク情報の信頼度を視覚的にも明示する

4.5.2 ID マップの具体的イメージ

これまでの調査を踏まえると、ID マップで管理するのが望ましいと思われる情報には以下のような

ものがある。なお NCID と JPNO とは必ずしも 1 対 1 の対応にはならないが、NCID 単位での情報管理を前提としている。

・NC 書誌に関する情報

- 親子の別
- 親書誌 NCID
- 複製か否か
- 和古書かどうか

・リンクの信頼性に関する情報

ユーザの評価情報

調査は JPNO の精度に焦点をあてて行ったが国立国会図書館によるデジタル化資料の公開状況はそれぞれに異なる。また平成 26 年 1 月からは国立国会図書館のデジタル化資料送信サービス⁶⁾が開始されており、入手困難な資料について国立国会図書館の承認を受けた図書館等の指定された端末で閲覧・複写が可能となる。ID マップは基本的には識別 ID との対照表であるが、デジタル化の有無およびその公開条件、権利情報を関連付けて管理することで CiNii Books の IP アドレスによる認証機能との連携により当該資料が所属機関の図書館の指定端末で利用可能かユーザに提示することも可能であろう。

CiNii Books は勿論、NDL Search も日々成長し続けているデータベースであり、ID マップは常に情報の更新を図る必要がある。OAI-PMH などを利用してメタデータ等の収集を定期的に行いデータの修正・追加を行うとともにユーザからのフィードバック情報も反映していくことが望ましい。

5. 各種 ID による HathiTrust と CiNii Books の連携の可能性

5.1 HathiTrust の概要

HathiTrust は 2008 年から開始された米国の大学図書館などが共同で運営するデジタル化資料のリポジトリである⁷⁾。2013 年 11 月時点で 80 以上の図書館およびコンソーシアムが参加しており、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をおこなっている。収録されているデジタル化資料は、図書 570 万タイトル、雑誌 28 万タイトルなど約 1,086 万冊にものぼる。収録されているデジタル化資料のメタデータについては、HathiTrust の検索プラットフォームや API⁸⁾のほか Hathifiles⁹⁾というテキストファイル形式でも公開されている。Hathifiles には、デジタル化資料のメタデータのうち、Volume Identifier (htid)、Access, Rights, University of Michigan record number (record number)、OCLC numbers, ISBNs, ISSN, LCCNs, Title, Publication

Date など約 20 のエレメントが含まれている。Hathifiles では 1 つの htid に関するレコードが 1 行で表現されている。なお、HathiTrust のメタデータには htid と record number という二つのユニーク ID が存在する。htid は一つ一つの物理単位に付与される ID で、record number は NCID に相当する書誌 ID である。

5.2 検証方法

5 章では HathiTrust のメタデータに含まれる ISSN, ISBN, LCCN, OCLC number をキーにして CiNii Books とリンク形成ができるかどうか検証した。まず、HathiTrust のデジタル化資料のうち、日本国内からアクセス可能な資料のメタデータを Hathifiles から抽出して、それぞれの ID の保有率を調査した。次に、Hathifiles で同じ record number と ID をもつ重複レコードを削除した上で、ID をキーにして NC 書誌と HathiTrust のメタデータの同定作業を行った。

今回の分析対象として使用したデータは、HathiTrust については、Hathifiles20130801 のほか HathiTrust, WorldCat, Digital Public Library of America の API¹⁰⁾ を通じて取得したメタデータ、NACSIS-CAT については NII 提供の NC 書誌データ¹¹⁾ と CiNii Books の API を通じて取得した書誌データである。取得したデータに含まれる NCID, record number, ISBN, ISSN, LCCN, OCLC number の同定作業については、OpenRefine, Cygwin¹²⁾, R¹³⁾, テキストエディタ, Microsoft Excel などのツールを用いた。なお、各種 API によるデータの取得は 2013 年 8 月から 12 月にかけておこなった。

5.3 HathiTrust のパブリックドメインコンテンツの ID 保有率

Hathifiles20130801 に含まれる htid の数は 10,787,584 件で、そのうち Rights が pd (public domain) のもの¹⁴⁾、つまり日本からデジタル画像へのアクセスが可能なコンテンツ数は 2,247,962 件 (20.8%) であった¹⁵⁾。パブリックドメインのコンテンツの htid ごとのメタデータにおける各種識別 ID の保有率は、ISSN が 177,667 件 (7.9%)、ISBN が 34,784 件 (1.5%)、LCCN が 967,823 件 (43%)、OCLC number が 2,000,077 件 (88.9%) であった。以下、これらの ID を有する HathiTrust のメタデータを対象として NC 書誌データとの同定を試みた。

5.4 ISSN による書誌同定

Hathifiles から ISSN を持つレコードのみを抽出し、重複を排除すると、record number と ISSN の組み合わせの数は 8,116 件となった。この ISSN を用いて CiNii Books の API で NC 書誌データを取得して、OpenRefine で ISSN をキーにした record number と NCID の対応関係を調査した。その結果、NCID と NCID に対応する record number, ISSN の組み合わせで 4,194 件の書誌がリンク可能であることがわかった。ただし書誌データに誤って異なる資料の ISSN が入力されている可能性もあるため、次の方法で同定結果の検証を行った。まず、同定結果に含まれる record number と NCID を使用して、HathiTrust および CiNii Books の API を用いてそれぞれの書誌データに含まれるタイトルを抽出する。次に OpenRefine の ngram と fingerprint function¹⁶⁾ を使用してタイトルの冒頭の 1 語から 7 語までを抽出して、すべての単語を小文字に正規化した後 ABC 順にソートして比較することで、タイトルの類似度を判定した。タイトルの冒頭から 1 単語のみ使用して比較した場合を N1 と表記すると、N1 で 3,765 件、N2 で 3,562 件、N3 で 3,265 件、N4 で 3,081 件、N5 で 2,962 件、N6 で 2,923 件、N7 で 2,864 件が一致した。準拠する目録規則によりタイトルの採録方法が異なる場合もあり不一致の場合も同定に誤りがあるとは言い切れないが、少なくとも 4,194 件のうち、タイトルの冒頭 7 語 (N7) まで一致した 2,864 件、つまり全体の約 7 割は ISSN をキーにしたリンクが可能である。

5.5 ISBN による書誌同定

ISBN を持つレコードについても、ISSN の場合と同様の方法で同定作業をおこなった結果、Hathifiles の record number と ISBN の組み合わせの数は 35,306 件、NCID と record number, ISBN の組み合わせで 1,779 件の書誌がリンク可能であった。同定作業の検証を OpenRefine でおこなったところ、N1 で 1,592 件、N2 で 1,566 件、N3 で 1,534 件、N4 で 1,493 件、N5 で 1,461 件、N6 で 1,437 件、N7 で 1,421 件が一致した。少なくとも 1,779 件のうち、タイトルの冒頭 7 語 (N7) まで一致した 1,421 件、つまり全体の約 8 割は ISBN をキーにしたリンクが可能である。

5.6 LCCN による書誌同定

LCCN については CiNii Books の API では LCCN を使用した書誌データの取得がおこなえないため、NII から提供を受けた NC 書誌データを用いて同定

をおこなった¹⁷⁾。Hathifiles の record number と LCCN の組み合わせの数は 400,670 件で、NCID と record number、LCCN の組み合わせで 31,757 件の書誌がリンク可能であった。LCCN については NC 書誌を流用入力にて作成する際に対象資料と異なる LCCN が削除されずに残ってしまう可能性もあるため、タイトルだけでなく出版年も使用して同定の検証をおこなった。タイトルと出版年がともに一致したのは、N1 で 23,271 件、N2 で 22,642 件、N3 で 21,317 件、N4 で 20,015 件、N5 で 18,846 件、N6 で 17,987 件、N7 で 17,436 件であった。一方タイトルと出版年がともに不一致だったのは、N1 で 3,456 件、N2 で 3,626 件、N3 で 3,900 件、N4 で 4,125 件、N5 で 4,355 件、N6 で 4,477 件、N7 で 4,600 件となり、全体の 10% から 15% 程度であった。これらの不一致レコードについては双方で同じ LCCN を有していても異なる資料である可能性が高い。

5.7 OCLC number による書誌同定

OCLC number も CiNii Books の API を使用して書誌データを取得できないため、NII から提供を受けた NC 書誌データを用いて同定をおこなった。なお、WorldCat では同一資料に対して異なる OCLC number を保有している場合があり、NC 書誌データに対応する資料が HathiTrust でデジタル公開されていても同じ OCLC number をもつとは限らない。NC 書誌データに含まれる OCLC number をキーにして NC 書誌データと HathiTrust の同定をおこなうには、WorldCat の xID (xOCLCNUM) API を使用して同一資料の OCLC number グループを取得してから、HathiTrust の OCLC number とマッチングすることが必要となる。NC 書誌データから抽出した OCLC number は 5,826 件。この OCLC number を WorldCat の xID API を使用して HathiTrust の OCLC number を取得すると 239 件がヒットした。レスポンスデータには複数の OCLC number をもつものがあるので整形すると、NC 書誌データの OTHN に含まれている OCLC number と関連する OCLC number をもつ HathiTrust のデータ件数は 279 件となった。このうちパブリックドメインのコンテンツは 244 件であった。

5.8 タイトルによる書誌同定

パブリックドメインのコンテンツ全体の約 9 割近くのメタデータに OCLC number が存在するが、5.7 で述べた通り、NC 書誌データに OCLC number をもつレコードが少ないため OCLC number をキーにした同定は困難である。そこでタイトルによ

る書誌同定を試みた。まず Hathifiles で ISSN, ISBN, LCCN がなく OCLC number のみ ID をもつレコードを抽出し、同じ record number をもつ行を統合すると 616,279 件になった。この中から 2,389 件をランダムに抽出して、record number を用いて API で HathiTrust からタイトルと出版年を取得し、さらに取得したタイトルと出版年で CiNii Books の API で問い合わせをおこなった。その結果 2,389 件のうち 1,899 件がノーヒット、490 件は 1 レコード以上がヒットした。さらに 490 件のうち 1 レコードのみがヒットしたのは 181 件であった。この 181 件について ISSN の場合と同様にタイトルを用いて OpenRefine による検証をおこなうと、N1 で 142 件、N2 で 136 件、N3 で 124 件、N4 で 115 件、N5 で 105 件、N6 で 99 件、N7 で 97 件が一致した。タイトルを用いた同定はほかの ID をキーにして同定をした場合に比べて精度はかなり低い、ID が使用できない場合に補完的に使用できる可能性がありそうである。

5.9 検証結果のまとめ

それぞれの ID をキーにして CiNiiBooks から HathiTrust へリンク可能な件数は、ISSN (4,194 件)、ISBN (1,779 件)、LCCN (31,757 件)、OCLC number (244 件) であった。このうち、ISBN, ISSN をキーにした同定は検証作業から LCCN や OCLC number の場合と比較して精度が高いことがわかった。しかしデジタル化資料の原本の出版年が古い場合 ISBN がないメタデータも多い。一方、ID による同定可能な数が最も多い LCCN は NC 書誌データにも一定程度 ID が保有されているものの、タイトルや出版年が異なる資料について同じ LCCN が入力されている場合もあることから、LCCN のみだけでなくタイトル等の書誌事項を組み合わせて検証する必要がある。また ID を使用しない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精度がかなり低くなることがわかった。

5.10 ID マップ作成に向けて

5.10.1 ID マップの構成

ISBN や ISSN など識別 ID がある場合は NCID, 識別 ID, HathiTrust record number の組み合わせ、識別 ID を持たない場合は NCID と HathiTrust record number の組み合わせを ID マップとして作成する。CiNii Books 書誌詳細画面に表示される NCID を用いて ID マップを参照することで HathiTrust へのリンクが可能になる。今回は前提

条件としてパブリックドメインのものに限定して分析をおこなったが、現在未公開のデジタル化資料で今後パブリックドメインになるものも多数含まれている。このため HathiTrust に含まれるすべての書誌データについて分析をおこない権利情報についてもあらかじめ ID マップに含めておくか、HathiTrust の API を用いて権利情報を動的に取得してパブリックドメインの資料のみリンクを生成するといった運用方法が考えられる。

5.10.2 検討すべき課題

ID マップ作成には単純に NACSIS-CAT 対 HathiTrust = 1 対 1 の関係で結びつけることが可能だと思うが、いくつかの検討すべき項目が挙げられるのではと思う。NACSIS-CAT 側ではユニーク ID として NCID が挙げられるが、HathiTrust 側では二つのユニーク ID が存在しており、そのどちらに NCID を紐づけたら良いのか迷う部分である。

5.1 で記した通り、HathiTrust には htid と record number という ID がある。NACSIS-CAT と HathiTrust の ID 関係は NCID と htid を対にした場合、NC 書誌から直接 HathiTrust の電子化コンテンツに結びつけることができる利点がある。一方で版違いに気づきにくい面もある。一方 NCID と record number の場合は同じ書誌データの平行なリンク関係で、版違いや関連書誌等発見しやすいのではないと思われる。その際、NACSIS-CAT の VOL 積み書誌はどう対応すべきか等、他にもいくつか課題が挙げられる。

6. ESTC 番号による HathiTrust と CiNii Books の連携の可能性

6.1 ESTC とは

ESTC とは English Short-Title Catalog の略称で 1473 年～1800 年に主にイギリスや北米で刊行された資料を収録対象としている目録である。この目録には ESTC 番号と呼ばれるユニーク ID があり、先頭にアルファベット 1 文字、それに続いて数字が 4 桁～6 桁程度続く。例えば「R38489」というような ID である。

NC 書誌には主に NOTE フィールドに「References: ESTC R38489」といった記述がされている。

6.2 検討内容と前提

6 章では NC 書誌の NOTE フィールドに記述されている「ESTC 番号」をキーにして HathiTrust とリンク形成ができるかどうかを検証した。NC 書誌と HathiTrust には ESTC 番号という共有の ID

があるので、直接マッチングすれば良いと考えられるが、1 点問題があった。それは HathiTrust のメタデータに ESTC 番号の記述があったとしても、HathiTrust の検索画面では ESTC 番号をキーワードとして書誌を検索することができないのである。よって別のキーワードで HathiTrust の検索を行う必要がある。そのためには、別の書誌データを間に介在させる必要があると考えた。

6.3 検証方法

HathiTrust の検索画面では OCLC number での検索が可能であるため、検証方法として、NACSIS-CAT と HathiTrust との間に WorldCat サービスを利用することを検討した。検証には、前述の OpenRefine と各サービスが提供する API を使用した。

始めに NII より ESTC 番号が記述されている NCID のリスト提供をしてもらった。次に OpenRefine を利用し、CiNii API を使って、対象 NCID から NC 書誌のタイトルを取得、その後 OpenRefine でデータをエンコードし、WorldCat Basic API を使って、タイトルをキーワードにして検索した書誌データを取得した。その書誌データに含まれる OCLC number を抽出し、さらに OCLC number をキーワードにして HathiTrust API を使用し、HathiTrust メタデータを取得した。その後エクセル等で ESTC 番号が入っているメタデータを抽出し、元の NC 書誌に含まれていた ESTC 番号と一致するかを比較検証した。

5 章で触れているように、HathiTrust の書誌データは OCLC number、LCCN をもつ割合が高いことが分かっていたので、OCLC number を取得できればある程度ヒット率はあるのではないかと予想された。

6.4 検証結果

API 等を使用して、NII 提供のデータを起点として検証を行った。提供データのうち、重複データや、ESTC 番号の有無の確認後 2,539 件の ESTC 番号を含む NCID のデータを使用した。この NCID から CiNii API を使用しタイトルを取得し、WorldCat Basic API を使用したところ、2,539 件のうち 2,253 件が何かしらヒットした。

WorldCat Basic API はデフォルトで 1 つのクエリに対し 10 件（最大 100 件）検索結果を返すが、のべ 13,171 件ヒットした。この 13,171 件からそれぞれ OCLC number を抽出し、HathiTrust Bibliographic API で OCLC を指定して検索をかけたところ

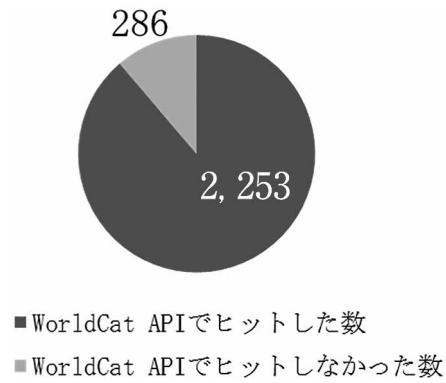


図4 WorldCat Basic API 問い合わせ結果

ろ、ヒットしたのはわずか629件だった。13,171件のうち、629件だったので、この時点でのヒット率は4.7%である。

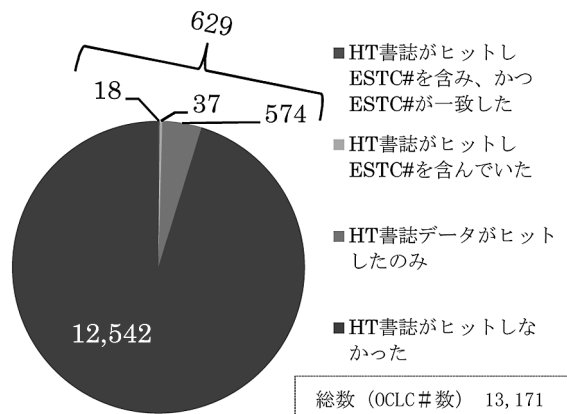


図5 HathiTrust Bibliographic API 問い合わせ結果

この629件のうち、ESTC番号を含んでいたメタデータは55件で629件全体の8.7%である。さらにその中55件のうち、NC書誌データ中のESTC番号と一致するのは18件だった。

表8 検証結果まとめ

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 (= 8/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 書誌と同一	3

6.5 検証結果から考察できること

この検証で明らかになったことは主に3つある。一つ目は、HathiTrustのメタデータがそもそもESTC番号を保有していないということ、二つ目は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%と半数にも満たないこと、三つめはESTC番号よりタイトルをキーにするほうがヒット率は若干だが高いということである。

タイトルは一般的な名詞を多く含んでいるタイトルなら、タイトルは同一でも中身が異なる書誌同士でもヒットする可能性はあるが、それでもESTC番号よりヒット率が高いことはESTC番号での照合が困難であることを示しているように思われる。ちなみに、今回ESTC番号が一致した18件のうち、ESTC番号とタイトル両方が一致したものはわずか3件になった。

この結果からESTC番号をキーにしたWorldCat経由でのHathiTrustのリンクの可能性はほとんど無いと結論づけた。

さらに検証中、ESTC番号が一致していたHathiTrustのメタデータを再び取得したところ、ESTC番号が無くなった状態でメタデータが更新されていることが分かった。このような点からもESTC番号を利用したリンク関係の形成はメタデータが流動的なこともあり、ますます困難と考える。検証結果からESTC番号をキーとしたIDマップの作成は困難かつ実用性に乏しいと判断されるため具体的な考察を行わないこととする。

7. 調査を終えて

7.1 IDマップの作成について

本稿では、異なるデータベース間のリンク、特にNACSIS-CATを基点としたリンクの可能性を調査するために、各データベースに蓄積されたIDに着目して分析をおこなった。NDL SearchはJPNOをキーとしたリンクが、NACSIS-CATの保有情報だけでもある程度可能であり、また賛否はあろうが欠落しているJPNOを補完しうる可能性を示すことができた。HathiTrustは多様なIDの情報をもつがNACSIS-CATとの連携という点で考えればLCCNが有力候補になろう。ただし正確性の面で考慮が必要という結果となった。またESTC番号は残念ながら少なくともNACSIS-CATとHathiTrustとの書誌データリンクという点では現状、有効に機能しえない可能性が高いことも確認できた。

これまで記述してきたように各データベースがもつIDは保有状況も精度も様々であり、単純にデー

タを突き合わせるだけでは必ずしも有効かつ正確な ID マップをつくりあげることにはできない。より正確で網羅性の高い ID マップを実現するためには、ID 以外の書誌情報も活用しながら、ID マップのプロトタイプを作成し、その有効性について検証を重ねていくことが必要である。

7.2 ID マップの活用方法

ID マップは、実現すればユーザの学術資料へのアクセシビリティ向上という目的以外にも様々な活用方法が期待できる。各機関の所蔵資料のデジタル化状況の確認が容易となり、書誌データの充実や相互運用性の向上にも活用可能である。例えば、各データベースの書誌データに含まれる基本件名標目表 (BSH)、国立国会図書館件名標目表 (NDLSH) や米国議会図書館件名標目表 (LCSH) について相互に補完することができる。また、書誌データを介した著者名などの典拠データの相互リンクも可能となる。NDL 書誌に含まれる典拠データは既に国立国会図書館典拠データ検索・提供サービス (Web NDL Authorities)¹⁸⁾ を通じて OCLC (Online Computer Library Center, Inc.) が提供するバーチャル国際典拠ファイル (Virtual International Authority File, 略して VIAF)¹⁹⁾ とリンクされており、ID マップで対応関係がとれた NC の典拠データもリンクさせることが可能になる。現在、国内の大学図書館を中心に導入がすすんでいるディスカバリーサービスでは国内外の様々なデータベースプロバイダから提供されたメタデータをセントラルインデックスに統合して検索可能にしているが、ID マップはこのようなセントラルインデックスにおける書誌統合のための情報源としても活用が期待できる。

今回検証対象とした NDL Search や HathiTrust 以外にも Internet Archive, Digital Public Library of America, Europeana などデジタル化資料の大規模データベースはほかにも存在する。無論個別のデータベースごとにその実現性につき検証は必要だが、図書館はこれまで書誌データという形で蓄積してきた情報資産を活用して ID マップを作成・維持していくことで、これらのデジタル化資料の効率的な利用を促すことができるであろうし、またそういった貢献も求められているのではないだろうか。

本稿は平成 25 年 7 月～12 月にかけて国立情報学研究所において開催された平成 25 年度学術情報システム総合ワークショップで実施した調査をベースとしている。調査にあたってはデータ提供の面のみならず地理的に離れた参加者間でも必要な意思疎通が図れるよう種々配慮をいただくなど、国立情報学

研究所の皆様には様々な点でご協力をいただいた。講師の皆様からも講義等を通じて貴重な示唆を頂いたが、とりわけ佐藤義則教授 (東北学院大学)、大向一輝准教授 (国立情報学研究所) のお二方には 5 か月に及ぶワークショップ全期間を通じて適切な助言や提言を頂いた。心より感謝の意を表して本稿の結びの言葉としたい。

注

- 1) “CiNii 全般 - メタデータ・API”. CiNii. http://ci.nii.ac.jp/info/ja/api/api_outline.html (参照 2014-06-16).
- 2) “外部提供インターフェース (API)”. 国立国会図書館サーチ. <http://iss.ndl.go.jp/information/api/>, (参照 2014-06-12).
- 3) “xID API”. OCLC. <https://oclc.org/developer/develop/web-services/xid-api.en.html>, (参照 2014-05-03).
- 4) “HathiTrust Bibliographic API”. HathiTrust Digital Library. http://www.hathitrust.org/bib_api, (参照 2014-06-16).
- 5) OpenRefine. <http://openrefine.org/>, (参照 2014-06-16).
- 6) “図書館向けデジタル化資料送信サービス”. 国立国会図書館. http://www.ndl.go.jp/jp/library/service_digi/, (参照 2014-05-03).
- 7) HathiTrust Digital Library. <http://www.hathitrust.org/>, (参照 2014-05-03).
- 8) 前掲脚注 4) 参照.
- 9) “Hathifiles”. HathiTrust Digital Library. <http://www.hathitrust.org/hathifiles>, (参照 2014-05-03).
- 10) “API Codex”. Digital Public Library of America. <http://dp.la/info/developers/codex/>, (参照 2014-05-03).
- 11) 2013 年 7 月に「TTLL が jpn 以外で ISBN, XISBN, ISSN, NBN, LCCN, OTHN (ISSN か LCCN か NBN) のいずれかの ID をもつデータ」という抽出条件で「ID, YEAR, CNTRY, TTLL, TXTL, ISBN, XISBN, ISSN, NBN, LCCN, GPON, OTHN, 親書誌レコード ID」を抽出したデータと「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」という条件で「ID, YEAR, CNTRY, TTLL, TXTL, GMD, TR, ED, PUB, PHYS, VT, NOTE, 親書誌レコード ID」を抽出したデータの 2 種類。
- 12) Cygwin. <http://www.cygwin.com/>, (参照 2014-05-03).
- 13) “The R Project for Statistical Computing”. <http://www.r-project.org/>, (参照 2014-05-03).
- 14) HathiTrust では pdus (public domain in united states) も含めてパブリックドメインのコンテンツ

- と広報する場合があるが pdus の資料は日本からはアクセスできない。
- 15) 2014 年 1 月には HathiTrust に慶應義塾大学のデジタルコレクション 8 万点が収録された。ただし、2014 年 6 月現在、8 万点すべてがパブリックドメインにはなっていない。”Update on January 2014 Activities”. HathiTrust Digital Library. http://www.hathitrust.org/updates_january2014, (参照 2014-03-17).
 - 16) “GREL String Functions · OpenRefine/OpenRefine Wiki”. GitHub. <https://github.com/OpenRefine/OpenRefine/wiki/GREL-String-Functions>, (参照 2014-05-03).
 - 17) 2014 年 5 月 15 日に CiNii Books 図書・雑誌検索の API で LCCN (フリーワード) を使用した書誌データの取得が可能になった。”CiNii Books における、内容検索の対象追加、図書・雑誌詳細表示画面への LCCN (米国議会図書館管理番号) 表示追加等について (2014 年 5 月 16 日)”. CiNii. http://ci.nii.ac.jp/info/ja/index_2014.html#20140516, (参照 2014-06-24).
 - 18) “Web NDL Authorities (国立国会図書館典拠データ検索・提供サービス)”. 国立国会図書館. <http://id.ndl.go.jp/auth/ndla>, (参照 2014-05-03).
 - 19) VIAF. <http://viaf.org/>, (参照 2014-05-03).
-
- ＜ 2016.6.13 受理 おおにし まさと 京都大学 附属図書館情報サービス課相互利用掛, とや かずよ 神戸大学附属図書館情報管理課雑誌情報係, しばた いくこ 一橋大学学術・図書部学術情報課＞

Masato ONISHI, Kazuyo TOYA, Yasuko SHIBATA

Towards Linkage between CiNii Books and Databases of Digitalized Materials: The Possibility of Linkage among Databases Using ID Maps

Abstract : Thanks to the emergence of large-scale digitalized databases, such as the National Diet Library Digital Collections and HathiTrust, we can now access a variety of digitalized academic materials. In some cases, however, we cannot access such materials on CiNii Books even if digitalized materials are publicly available. Therefore, to explore the possibility of improving accessibility to digitalized materials on CiNii Books, the authors created ID maps for NACSIS-CAT, NDL Search, and HathiTrust respectively and examined how they could be utilized. This paper is a deliverable of the Education and Training Program of the National Institute of Informatics called “2013 Workshop on Scientific Information Systems.”

Keywords : CiNii Books / NACSIS-CAT / NDL Search / HathiTrust / ESTC / API / ID map