平成25年度 学術情報システム
総合ワークショップ
中間報告

グループ2
京都大学　大西賢人
神戸大学　鳥谷和世
一橋大学　柴田育子

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. 途中経過と分析結果

   3-1. 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

   3-2. CiNii Books, WorldCat, HathiTrustのIDマップ作成に向けた調査／検討

   3-3. CiNii Books, WorldCat, HathiTrustのIDマップ作成に向けた調査／検討（ESTCをキーに）

4. データマッチング方法について
5. 見えてきた課題
6. 今後の予定

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, HathiTrust等）と連携した検索環境整備
- 目標
  CINiiとNDL, 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, WorldCat, HathiTrustのIDマップ作成に向けた調査／検討
- 柴田
  CiNii Books, OCLC(XISBN), 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. 途中経過と分析結果

3-1 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
3-1-1必要なデータの検討と入手方法

- NDL
  - NDLサーチ書誌データ
  - NDLデジタル化資料メタデータ
- NII
  - CiNii Books書誌データ

→全データは取得は大量と判明
（CiNii Books書誌データで1,000万件超）
→言語、年代等で絞り込みを再検討

3-1 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
3-1-1必要なデータの検討と入手方法

	NDL		NII
	NDLサーチ	NDLデジタル化資料	CiNii Books
年代			1968年以前
言語		主に日本語	本タイトルの言語コード(TTLL)が日本語
出力コード			※1参照
件数		図書のデータ89万件	約103万件
提供者		佐藤先生	NII藤井さま

※1 NCID,YEAR,CNTRY,TTLL,TXTL,ISBN,XISBN,ISSN,NBN NDLCN,GPON,OTHN,親書誌レコードID

- 年代を絞った理由→NDLのデジタル化対象が1968年受入以前分であるため

3-1 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
3-1-2書誌データ分析

- NDLのデータ（NDLサーチとNDLデジタル化資料）
  NDLのデータにはID的なものが3種類ある

IDの種類	説明	IDが含まれているサービス
・JPNO JP（日本全国書誌）番号	NDLBibID（国立国会図書館書誌ID）が付いている資料のうち、日本全国書誌の収録対象とした資料にのみ付与。	NDLサーチ
・NDLBibID （国立国会図書館書誌ID）	最も網羅性の高いID。紙の納本制度に基づく資料、デジタル化資料等の媒体変換資料なら必ず付いている。ボーンデジタルには無い。	NDLサーチ NDLデジタル化資料
・NDLJP（ 国立国会図書館で付与した永続的識別子）	NDLのデジタルアーカイブシステムに収録された資料には付与されるID。ウェブアーカイブ資料や、デジタル化資料等に付与。	NDLデジタル化資料

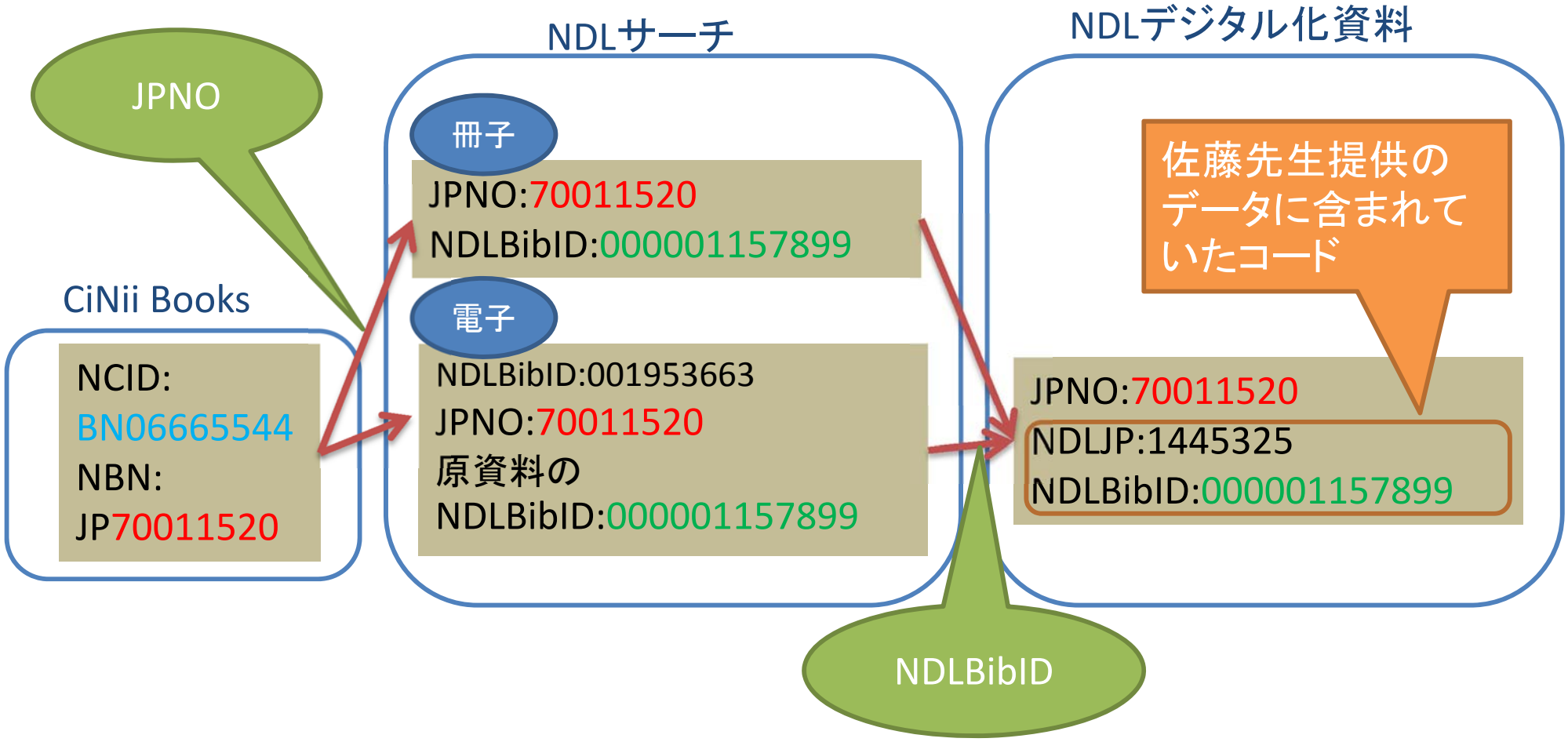
3-1 CiNii Books,NDLサーチ,NDLデジタル化資料のIDマップ作成に向けた調査／検討
3-1-2書誌データ分析

- NIIのデータ（CiNii Books）
  CiNii BooksのデータでNDLと結びつきそうなID
  →CiNii BooksでいうNBN（全国書誌番号）に入力されているID
  →JPNO JP（日本全国書誌）番号に相当する

- Webcat Plusのデータ
  NCID, JPNO, NDLBibIDを含んでいる
  →まずはCiNii Books, NDLサーチから取得するので対象外

- 3-1-3対応図

NCID:BN06665544(『朝鮮金融組合史』)の関係



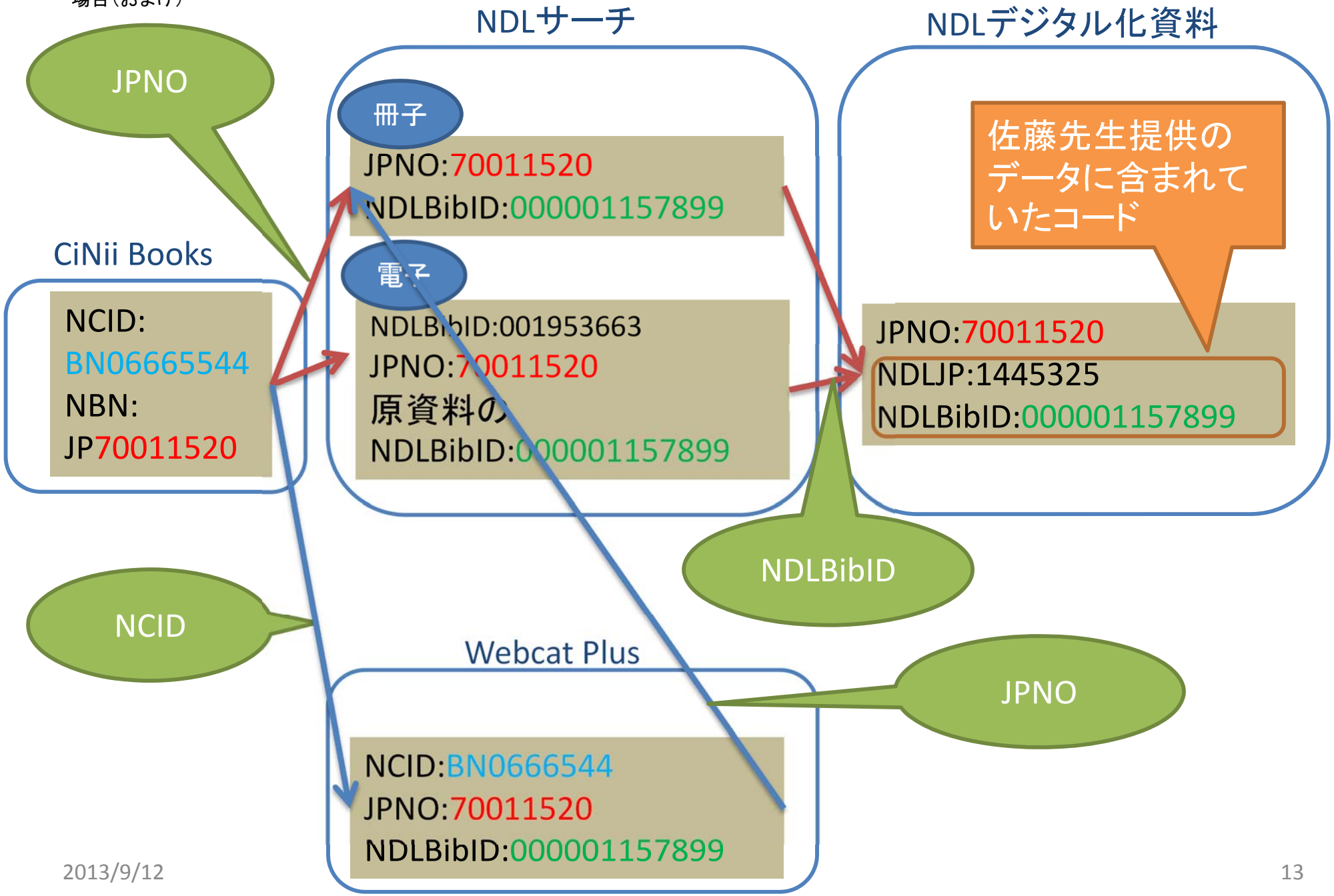
JPNO:70011520
NDLBibID:000001157899

(JPNO←→NDLBibID)この対応関係がわかるデータがあればCiNii BooksとNDLデジタル化資料が結びつく

NCID:BN06665544(『朝鮮金融組合史』)の関係

- 3-1-3対応図

もしWebcat Plusの書誌データを利用した場合(おまけ)



3-1 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
3-1-4分析結果

- CiNii Books -**NDLデジタル化資料**

CiNii Books

書誌総数	JPNOあり書誌	JPNO数	平均JPNO数	ISBNのみあり書誌	ISBN数	平均ISBN数
1,029,357	200,552	20,1656	1.01	8,067	153,255	19.00
	19.48%			0.78%		

NDLデジタル化資料

総数	NDLBibIDあり	利用可能*
894,274	889,848	284,683
	99.51%	31.83%

*利用可能：館内公開と未処理を除く

- 課題：原資料のNDLBibIDと紐づいたJPNOの入手

3-1 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
3-1-4分析結果

- リンク状況サンプル調査

サンプルJPNO数	NII書誌数	ndlサーチにヒット	正しくリンクしたJPNO数	正しくリンクしたNII書誌数
200	221	198	193	195

NDL側デジタル化状況

館内公開	127	77.44%
公開	37	22.56%
未デジタル化	29	

3-2 CiNii Books,WorldCat,HathiTrustのIDマップ作成に向けた調査／検討
3-2-1必要なデータの検討と入手方法

- HathiTrust
  - HathiTrustメタデータ
- NII
  - CiNii Books書誌データ

3-2 CiNii Books,WorldCat,HathiTrustのIDマップ作成に向けた調査／検討
3-2-1必要なデータの検討と入手方法

- HathiTrust
  - HathiTrustは「Hathifiles」という名前でタブ区切りテキスト形式でメタデータが提供されており、誰でもダウンロード可能
  - http://www.hathitrust.org/hathifiles
  - Hathifilesに含まれている要素

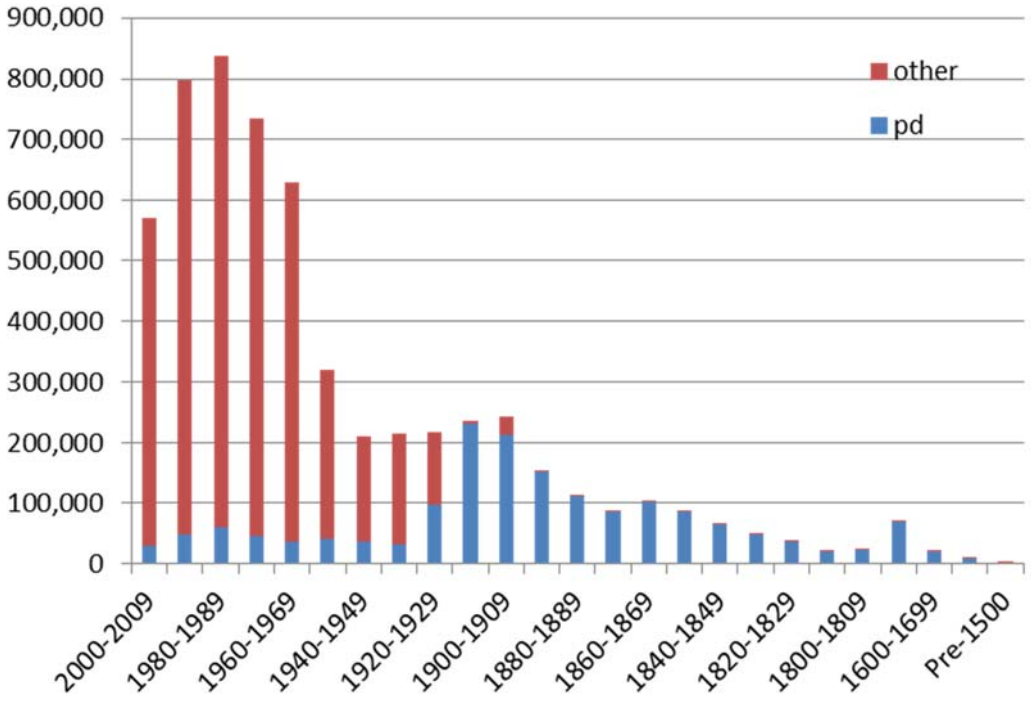
Volume Identifier	Title
Access	Imprint
Rights	Rights determination reason code
University of Michigan record number	Date of last update
Enumeration/Chronology	Government Document
Source	Publication Date
Source institution record number	Publication Place
OCLC numbers	Language
ISBNs	Bibliograhic Format
ISSNs	
LCCNs	

3-2 CiNii Books,WorldCat,HathiTrustのIDマップ作成に向けた調査／検討
3-2-1必要なデータの検討と入手方法

	HathiTrust	NII
	HathiTrust	CiNii Books
年代	すべて	すべて（図書）, 1929年以前（雑誌）
言語	すべて	本タイトルの言語コード（TTLL）がjpn以外
条件		ISBN, XISBN, ISSN, NBN, LCCN, OTHN (ISSN,LCCN,NBN,OCLC)がある（図書のみ）
出力コード	Hathifiles メタデータ	ID, YEAR, CNTRY, TTLL, TXTL, ISBN, XISBN, ISSN, NBN, LCCN, GPON, OTHN, 親書誌レコードID
件数	約1078万件（Volume）	約345万件（図書）, 約6万7千件（雑誌）
提供者	HathiTrust	NII藤井さま

3-2 CiNii Books,WorldCat,HathiTrustのIDマップ作成に向けた調査／検討
3-2-2書誌データ分析

- Hathitrust
  - 参加館の有無に拘わらず、電子コンテンツが見れるパブリックドメインのものを中心に分析



pd
ic
op
orph
und
umall, ic-world, nobody
pdus
cc-by, cc-by-nd, cc-by-nc-nd
cc-by-nc, cc-by-nc-sa, cc-by-sa
orphcand
cc-zero
und-world
icus

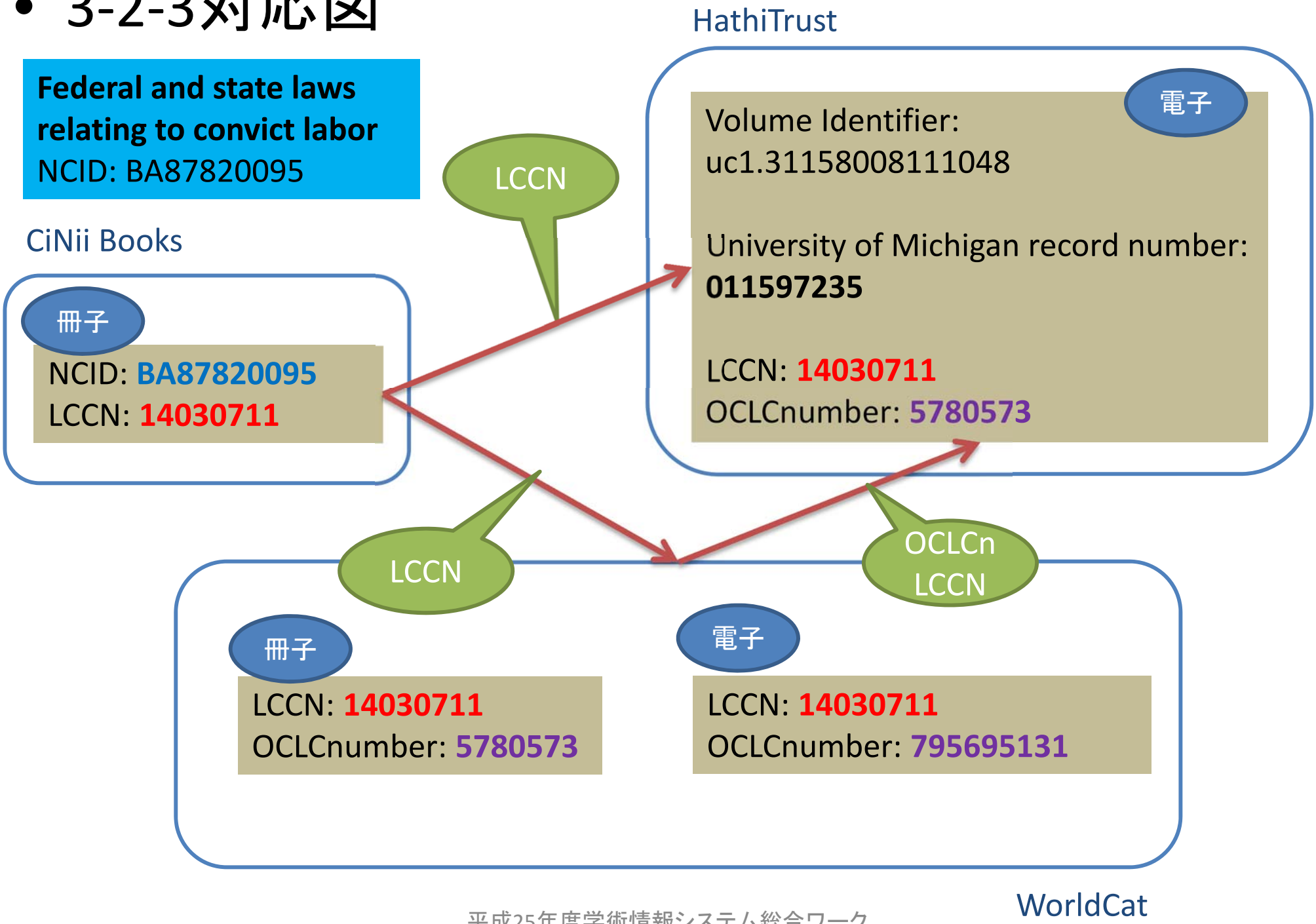
3-2 CiNii Books,WorldCat,HathiTrustのIDマップ作成に向けた調査／検討
3-2-2書誌データ分析

- NIIのデータ（CiNii Books）
  - ISBN, XISBN, ISSN, NBN, LCCN, OTHN (ISSN,LCCN,NBN,OCLC)
- HathiTrustのデータ
  - OCLC numbers, ISBNs, ISSNs, LCCNs

ID	IDあり	IDなし	ID保有率(pd)
OCLC numbers	2,000,073 (978,294)	247,889	88.97%
ISBNs	34,780 (27,390)	2,213,182	1.55%
ISSNs	177,663 (6,739)	2,070,299	7.90%
LCCNs	967,483 (380,556)	1,280,479	43.04%

- OCLC Number, LCCNを持つ割合が高い
- ISSNは書誌単位でカウントするとかなり限定される
- 言語が日本語のデータにはISBN, ISSN, LCCNがほとんどない

- 3-2-3対応図


Federal and state laws relating to convict labor
NCID: BA87820095
CiNii Books
冊子
NCID: BA87820095
LCCN: 14030711
LCCN
HathiTrust
電子
Volume Identifier:
uc1.31158008111048
University of Michigan record number:
011597235
LCCN: 14030711
OCLCnumber: 5780573
LCCN
OCLCn
LCCN
冊子
LCCN: 14030711
OCLCnumber: 5780573
電子
LCCN: 14030711
OCLCnumber: 795695131
WorldCat

3-2 CiNii Books,WorldCat,HathiTrustのIDマップ作成に向けた調査／検討
3-2-4分析結果

- CiNii Books-HathiTrust

ID	HathiTrust	CiNii Books	マッチング
LCCN	967,483 (380,556)	1,917,179	87,316 (30,674)
ISSN	177,663 (6,739)	47,888	61,664 (1,917)

- HathiTrust のメタデータに含まれるLCCNにはNormalizationされていないIDが存在
- NCのデータは桁数の違いはあるがprefixはない

3-3 CiNii Books,WorldCat,HathiTrustのIDマップ作成に向けた調査／検討（ESTCをキーに）
3-3-1必要なデータの検討と入手方法

- HathiTrust
  - HathiTrustメタデータ
    →大西さんのほうでHathifileを入手
- NII
  - CiNii Books書誌データ
- Britsh Library(BL)
  - BL書誌データ
    →入手は検討しなかった

3-3 CiNii Books,WorldCat,HathiTrustのIDマップ作成に向けた調査／検討（ESTCをキーに）
3-3-1必要なデータの検討と入手方法

	NII
	CiNii Books
年代	すべて
言語	すべて
条件	NOTEフィールドに「ESTC」と記載があるもの
出力コード	ID,YEAR, CNTRY, TTLL, TXTL, ISBN, XISBN, ISSN, NBN, LCCN,NDLCN, GPON,OTHN, NOTE, 親書誌レコードID
件数	約2,653件
提供者	NII藤井さま

3-3 CiNii Books,WorldCat,HathiTrustのIDマップ作成に向けた調査／検討（ESTCをキーに）
3-3-2書誌データ分析

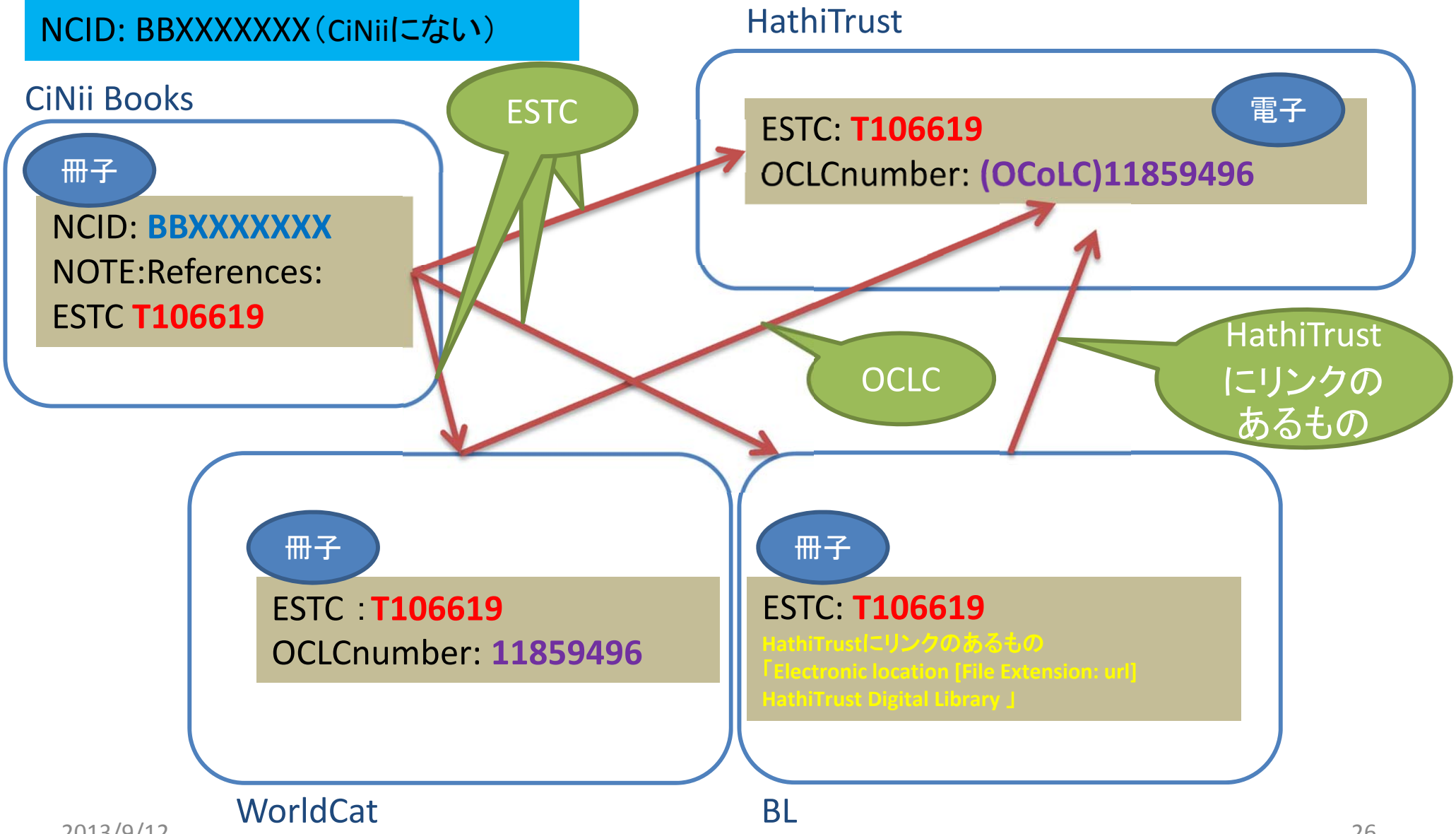
- NIIのデータ（CiNii Books）

	ESTC#の記述あり	LC#記載あり
ある	2,653	153
ない	44	4
合計	2,697	157

- HathiTrust
  - Hathifilesをみると、HathifilesにESTC#は含まれてない模様（項目なし）
  - HathiTrustのメタデータを見るとESTC#が記述されているものもある→しかし機械的にESTC#を抽出するのは難しそう…
  - HathiTrustのメタデータに高い割合で含まれているOCLC#をキーにしたマッチングが一番現実的か

- 3-3-3対応図

The state and behaviour of English Catholics, from the Reformation…
NCID: BBXXXXXXX（CiNiiにない）



4. データマッチング方法について

4.データマッチング方法について

- 書誌データを取得時に突き当たった壁
  - データが大きい
  - マッチング方法がエクセル以外わからない
  - パソコンが思うように動かない
  - APIの概念は分かっていても、具体的に何をすればわからない

4.データマッチング方法について

- Cygwin
  - Windows上でも簡易的なUNIX環境を構築
  - カウント，ソート，文字列操作
- OpenRefine
  - ローカルで動くWebアプリケーション
  - データクレンジング
  - APIを利用して外部サービスからデータ取得
    - Add column by fetching URLs

5-1 見えてきた課題(NDL)

- NDLとの書誌作成単位の差
  1対1対応にならない
  IDマップの構成に配慮

- NDLはシリーズでの管理はされていない

- NII側でのJPNO保有率の低さ

5-2 見えてきた課題(HathiTrust)

- HathiTrustにあるOCLC# をどう活用するか
- 正確なマッチングにはIDでも整形が必要
- マッチング結果が妥当かどうか検証が必要
- ESTC番号はどの程度有効なのか？

6.今後の予定

スケジュール（当初の計画）

1. 各DBのメタデータを入手する
（入手方法調査も含む）
2. 各DBのメタデータの構成を知る
3. 対象のDBのメタデータを比べ，相互にリンク関係ができそうな識別子の有無を調査する
4. あるならば，それがCiNii（NCID）全体の中で何%であるか算出する
5. IDマップが実現可能か検討する
6. IDマップを作成する
7. IDマップの活用方法を検討する

中間発表

最終報告

スケジュール（今後）

1. 各DBのメタデータを入手する
（入手方法調査も含む）
2. 各DBのメタデータの構成を知る
3. 対象のDBのメタデータを比べ，相互にリンク関係ができそうな識別子の有無を調査する
4. あるならば，それがCiNii（NCID）全体の中で何%であるか算出する
5. IDマップが実現可能か検討する
6. IDマップを作成する
7. IDマップの活用方法を検討する

最終報告