

日本図書館研究会情報組織化研究グループ月例研究会(2013年5月18日)

インターネット資料収集保存事業 (WARP)の10年とこれから



国立国会図書館 前田直俊
warp@ndl.go.jp

本日の内容

1. ウェブアーカイブの役割
2. ウェブアーカイブのしくみ
3. WARPの10年
4. メタデータと組織化
5. 課題と展望

1. ウェブアーカイブの役割

なぜウェブサイトを集めるのか？

- ウェブサイトは情報の更新が頻繁
- 消えてしまうサイト
- ボーンデジタル
- 紙の刊行物がネット版に移行
- 後世に残すべき文化遺産

誰がウェブサイトを集めるのか？

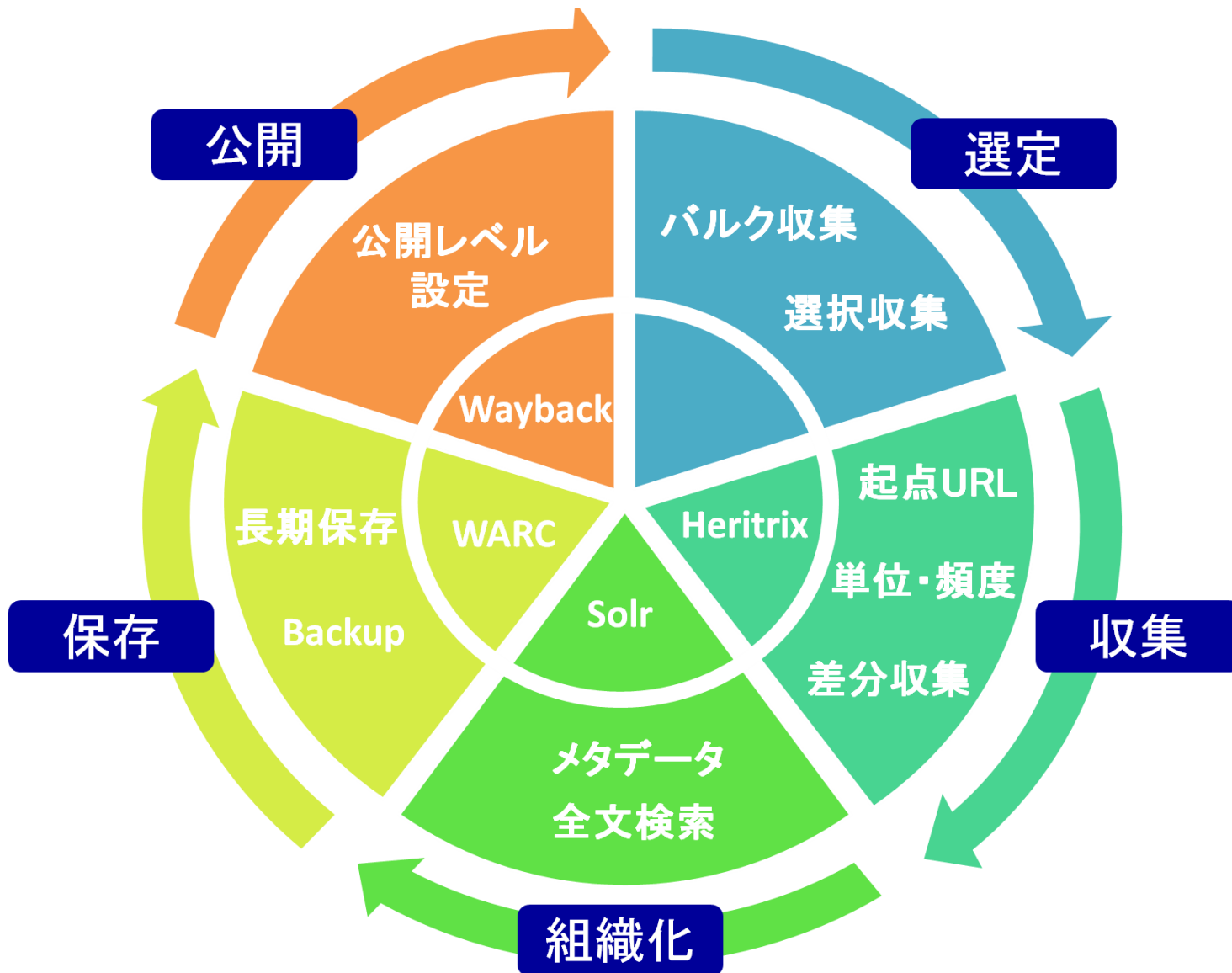
- 大規模ウェブアーカイブは公的機関が実施
(国立図書館、公文書館、大学・研究機関、etc.)
- 情報の保存と提供を担ってきた図書館
- 媒体を問わない文化遺産の保存
- 伝統的な資料群とのシームレスなアクセス保障
- 法制度に基づく安定的な運用

IIPC (International Internet Preservation Consortium)

- 世界のウェブアーカイブ機関からなるコンソーシアム
 - 第1期(2003～2006年)
 - 収集ロボットHeritrix等の基本的な技術開発
 - メンバーは12機関に限定(LC, BNF, IA, etc.)
 - 第2期(2006年～)
 - 洗練した技術の開発(WARC国際規格化)
 - ツール無償公開。改良や再配布が自由
 - メンバー拡大 ⇒ 42機関
(日本は国立国会図書館が2008年4月に加盟)

2. ウェブアーカイブのしくみ

ウェブアーカイブのライフサイクル



選定

・バルク収集と選択収集

		バルク収集 (Bulk Harvesting)	選択収集 (Selective Harvesting)
方法		ドメイン(.jp, go.jp)	ターゲット(サイト、機関、分野)
根拠		法制度、フェアユース	許諾契約、法制度
規模		大	小～中
コスト	ハード	大	小～中
	選定	小	中～大
	品質チェック	ポリシーによる	ポリシーによる
品質	収集対象	あらゆる品質	中～高(収集方針)
	データ	ポリシーによる	ポリシーによる
メタデータ		あまりない	ある程度ある

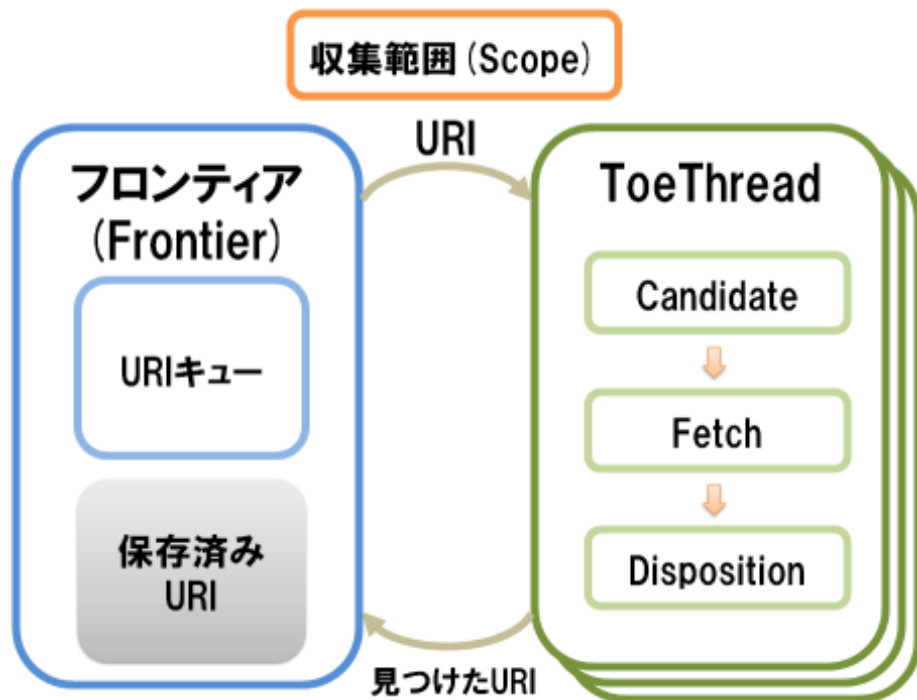
バルク、選択、法制度

Institutions	Bulk Harvesting	Selective Harvesting	Statutory Act
Internet Archive	●	●	
Austrian National Library	●	●	●
BnF (National Library of France)	●	●	●
British Library	●	●	●
Library and Archives Canada	●	●	●
Netarchive.dk	●	●	●
National and University Library of Croatia	●	●	●
National and University Library of Iceland	●	●	●
National Library of the Czech Republic	●	●	●
National Library of Finland	●	●	●
National Library of Israel	●	●	●
National Library of Norway	●	●	●
National Library of Spain	●	●	●
National Library of Sweden	●	●	●
National Library of New Zealand	● (by IA)	●	●
National Library of Australia	● (by IA)	●	
Ina (Institut National de l'Audiovisuel)		●	●
National Diet Library, Japan		●	●
National Library of Korea		●	●
National and University Library of Slovenia		●	●
California Digital Library		●	
Columbia University Libraries		●	
Harvard Library		●	
Internet Memory Foundation		●	
Library of Congress		●	
National Archives (U.K.)		●	
National Library of China		●	
National Library of The Netherlands		●	
National Library of Singapore		●	
Swiss National Library		●	

※“Member Archives, IIPC”(http://netpreserve.org/resources/member-archives)、“Report on questionnaire survey on web-archiving, 18th CDNLAO Annual Meeting”(http://www.ndl.go.jp/en/cdnlao/meetings/2010.html)、“Harvesting Practices Report”(IIPC, 2011)、各HP等より作成

Heritrix

- IIPCが開発したクローラ
- 世界のウェブアーカイブで広く使用

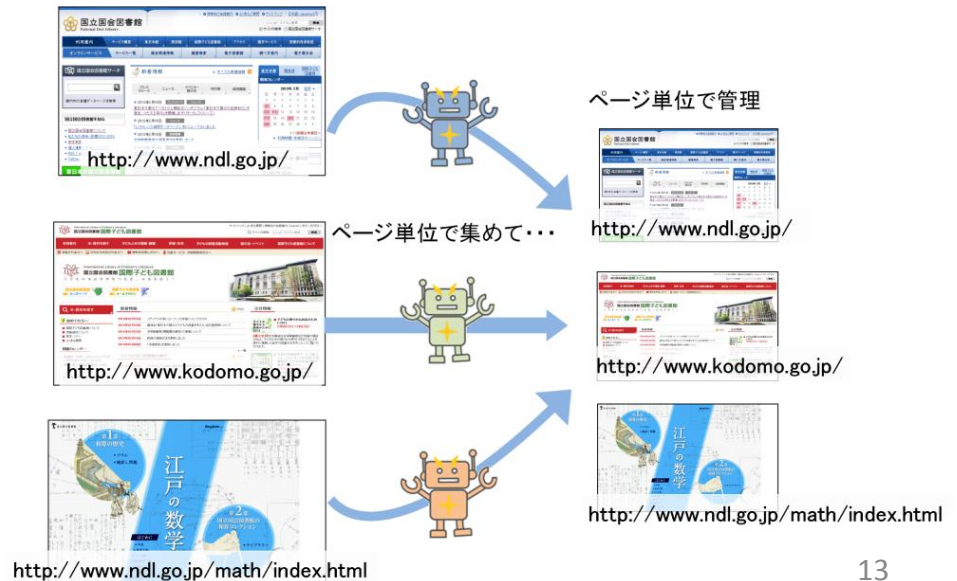


- Scope
収集範囲管理
- Frontier
URI管理
- ToeThread
リンク抽出、コンテンツ保存

収集単位

- ・ターゲット単位
- ・機関単位、ウェブサイト単位でターゲットを設定

- ・ページ単位
- ・ドメインレベルで大規模に収集し、URLのみで管理



収集頻度

・サイト更新のタイミングで収集

- ・高性能クローラが必要

(更新の自動検知、スケジュール自動設定、etc.)

・定期収集

- ・選択収集は収集ポリシー、サイト性質にあわせて設定

- ・バルク収集は1~3回/年

- ・システムリソースによる制限

(クローラの同時走行数、サーバ負担回避、etc.)

WARPの収集頻度

対象	頻度
国の機関	毎月
都道府県	年4回
政令指定都市	年4回
市町村	年4回
独立行政法人等	年4回
大学	年4回
電子雑誌	刊行頻度に合わせて

組織化

のちほど詳しく

保存

- バックアップ
- WARC (Web ARChive)
 - ウェブアーカイブの保存用ファイルフォーマット
 - 2009年に国際規格化 (ISO)
 - 「ヘッダー」と「コンテンツブロック」がセット
 - メタデータ項目とその記述方法が標準化
 - 収集日、収集方法、マイグレーション等の情報
- 長期保存
 - ファイル数の膨大さ
 - ファイル種の多様さ

公開

- Dark Archive
 - 完全非公開
- Grey Archive
 - 研究目的のみ公開
 - 特定施設内でのみ公開
- White Archive
 - ネット公開

3.WARPの10年

インターネット資料収集保存事業

- 2002年 スタート

-  Web ARchiving Project

- 「制度の10年」と「システムの10年」

制度の10年

	出来事	収集根拠
2002年	実験事業としてWARP開始	許諾
2004年	「ネットワーク系電子出版物の収集に関する制度の在り方について」(納本制度審議会答申) 「日本のWebサイトの網羅的収集、蓄積及び保存に関する調査報告書」	
2005年	「インターネット情報の収集・利用に関する制度化の考え方」に関する意見募集(パブリックコメント)	
(言論の委縮、違法情報等の懸念により、当面の制度化のターゲットを公的機関に限定)		
2006年	WARP本格事業化	
2009年7月	国立国会図書館法及び著作権法の改正	
2010年4月	改正法施行。公的機関ウェブサイトの制度収集開始	法律+許諾
2012年6月	国立国会図書館法及び著作権法の改正	
2013年7月	民間オンライン資料の制度収集開始(予定)	

制度収集

- 国立国会図書館は、公的機関のウェブサイトを許諾なく複製可能
- ネット上で一般に公開されているものが収集対象
 - LAN内のみコンテンツは対象外
- 除外対象（法的にはロボット排除設定の修正義務がないもの）
 - 事務に係る申請、届出等を受けることを目的とするもの
 - 長期アクセスを目的とし、かつ特段の事情なく消去されないと認められるもの

※民間サイトは引き続き許諾契約で収集

自動収集

・ロボット排除規約

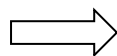
- ・ウェブサーバのルートディレクトリに「robots.txt」を置き、クローラのアクセスを排除。METAタグにも設定可能
- ・ロボット排除規約を遵守するのが基本ルール
- ・制度収集も遵守を前提とした制度設計

・ロボット排除設定の変更義務

- ・制度収集の対象機関は、国立国会図書館のクローラが通れるよう設定変更の義務がある

```
User-agent: *  
Disallow: /
```

全てのクローラの
アクセスを排除



```
User-agent: *  
Disallow: /  
  
User-agent: ndl-japan  
Disallow:
```

NDLのクローラは
アクセス可能

送信・送付による収集

- ・自動収集できなかったもののうち、以下に該当するものについて、送信・送付を求めることができる

- ・送信： システムを使ってファイルをアップロード
- ・送付： 媒体に格納して送付

- | | |
|-------------------|------------------|
| 1 年鑑、要覧及び職員録 | 10 政策評価書 |
| 2 業務報告 | 11 財務諸表 |
| 3 予算書及び決算書 | 12 調査報告書 |
| 4 統計書 | 13 紀要類 |
| 5 官報、法令集、規則集及び判例集 | 14 広報資料 |
| 6 法律解説書 | 15 講演会、展示会等の関係資料 |
| 7 目録及び書誌類 | 16 審議会等の関係資料 |
| 8 議会資料 | 17 その他前各号に準ずる出版物 |
| 9 基本計画書 | |

公開

- 閲覧

- 国立国会図書館の館内で全て閲覧可能

- 複写

- 発信者から許諾を得られたものののみ、全文複写サービス提供
- 著作権法第31条第1項による複写は困難

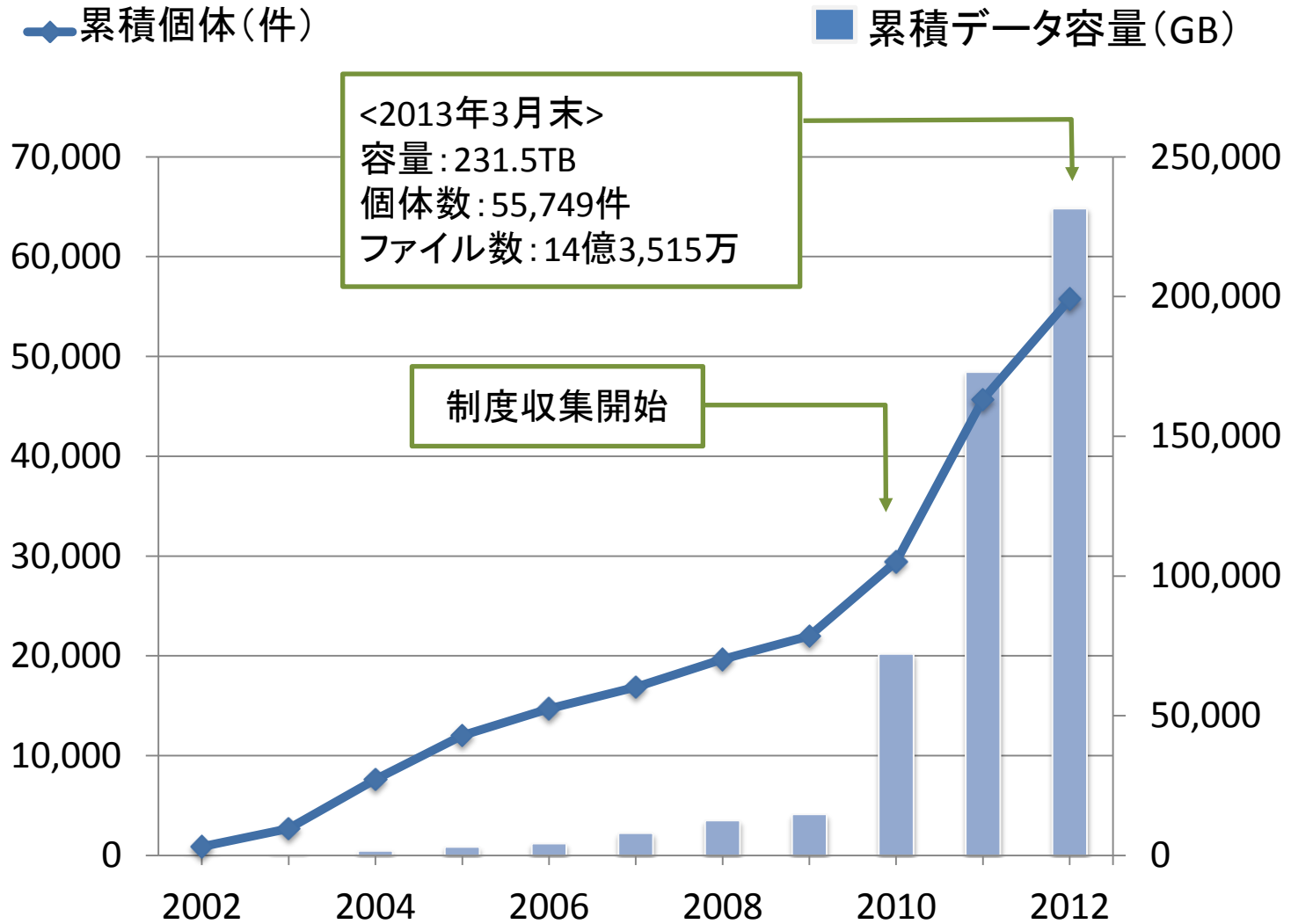
- ネット公開

- 発信者から許諾を得られたものののみ、ネット公開

※許諾率は約70%

ドメイン、ディレクトリ、ファイル単位での公開レベル設定
第三者著作物等の除外指定

収集量



システムの10年

・独自開発からオープンソースへ

	Ver.	変更点	Software			
			クローラ	収集管理	閲覧	全文検索
2002年	1.0	公開	Wget	独自	独自	—
2006年	1.x	全文検索機能				独自
2010年	2.0	全面改修	Heritrix	WCT	独自	Solr
2011年	2.1	印刷制御機能				
2013年	3.0	全面改修	Heritrix	独自	Wayback	Solr

Ver.1



Ver.2



Ver.3

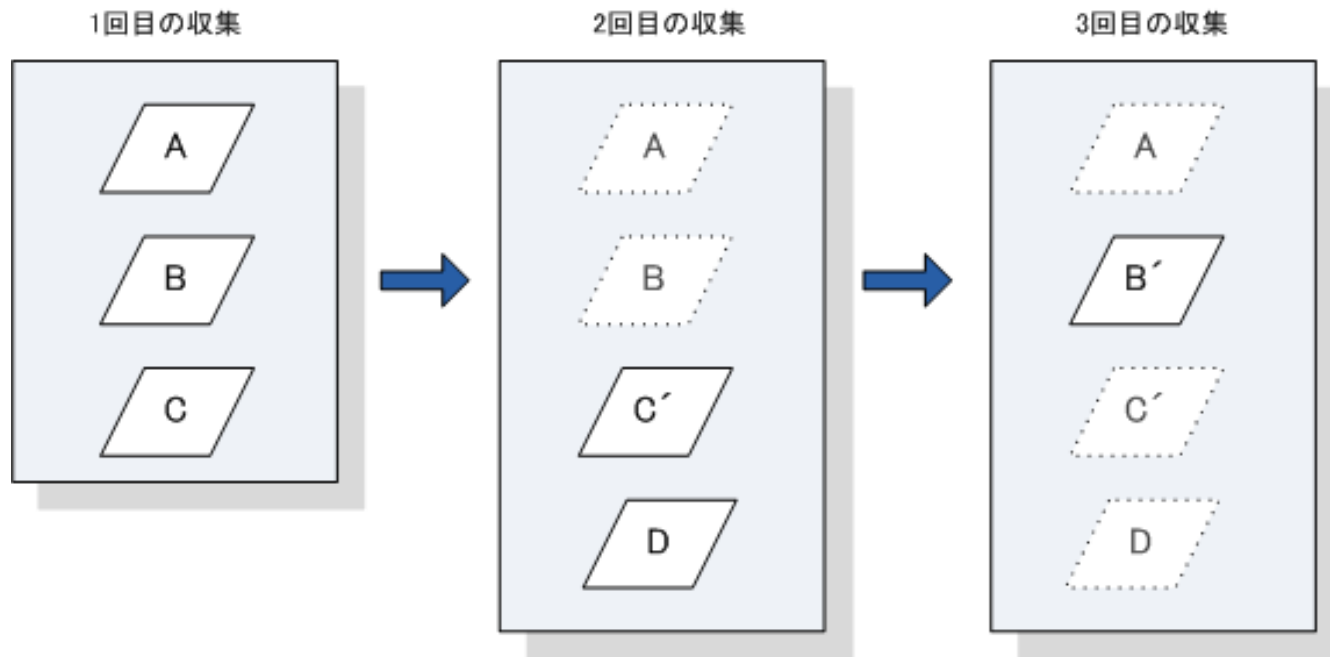


旧システム (Ver.2) の課題

- 同一ファイルの重複収集
⇒ 差分収集
- 保存用ファイルと閲覧用ファイルの2重持ち
⇒ 保存用ファイルの直接閲覧
- ハード面の強化
⇒ スケールアウト・スケールアップ
- 画面の操作性
⇒ インタフェースの改善
- 収集管理ツール (Web Curator Tool) の限界
⇒ 制度収集の運用に適した管理ツールの開発

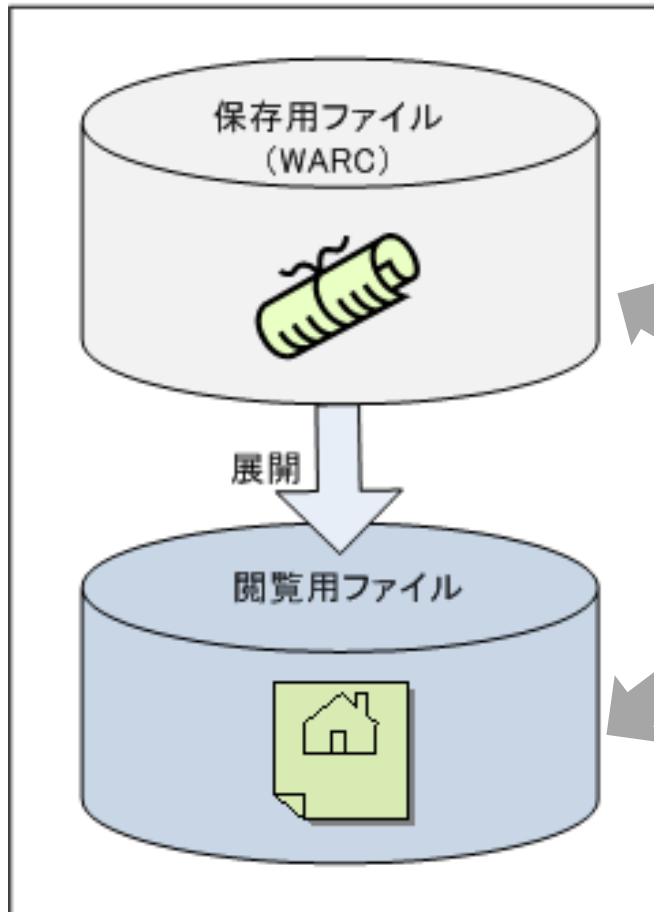
差分収集

- HeritrixのDeDuplicatorモジュールを追加
- 前回収集ファイルとハッシュ値を比較し、同値のファイルは保存しない

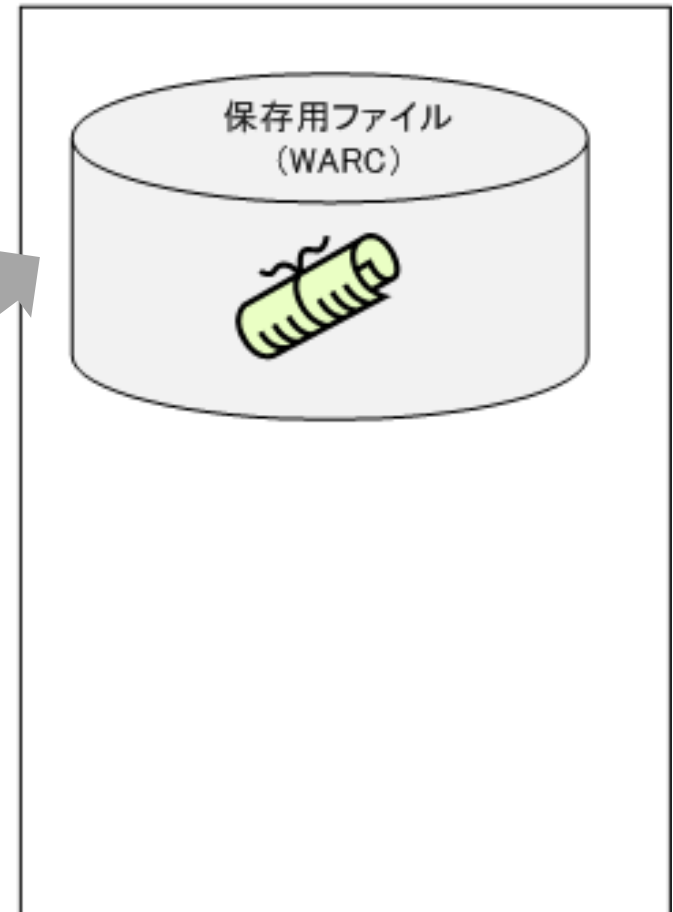


保存用ファイルの直接閲覧

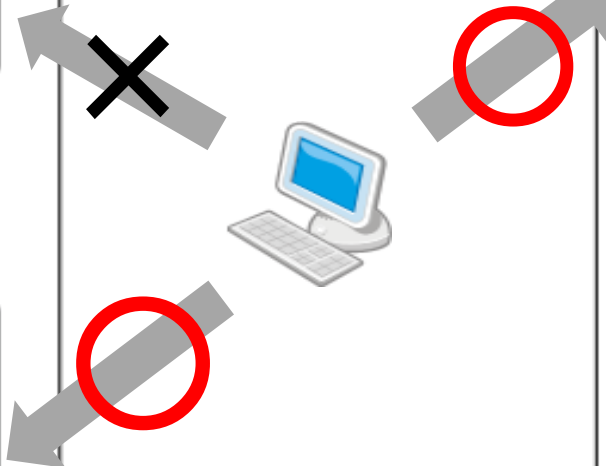
旧システム



新システム

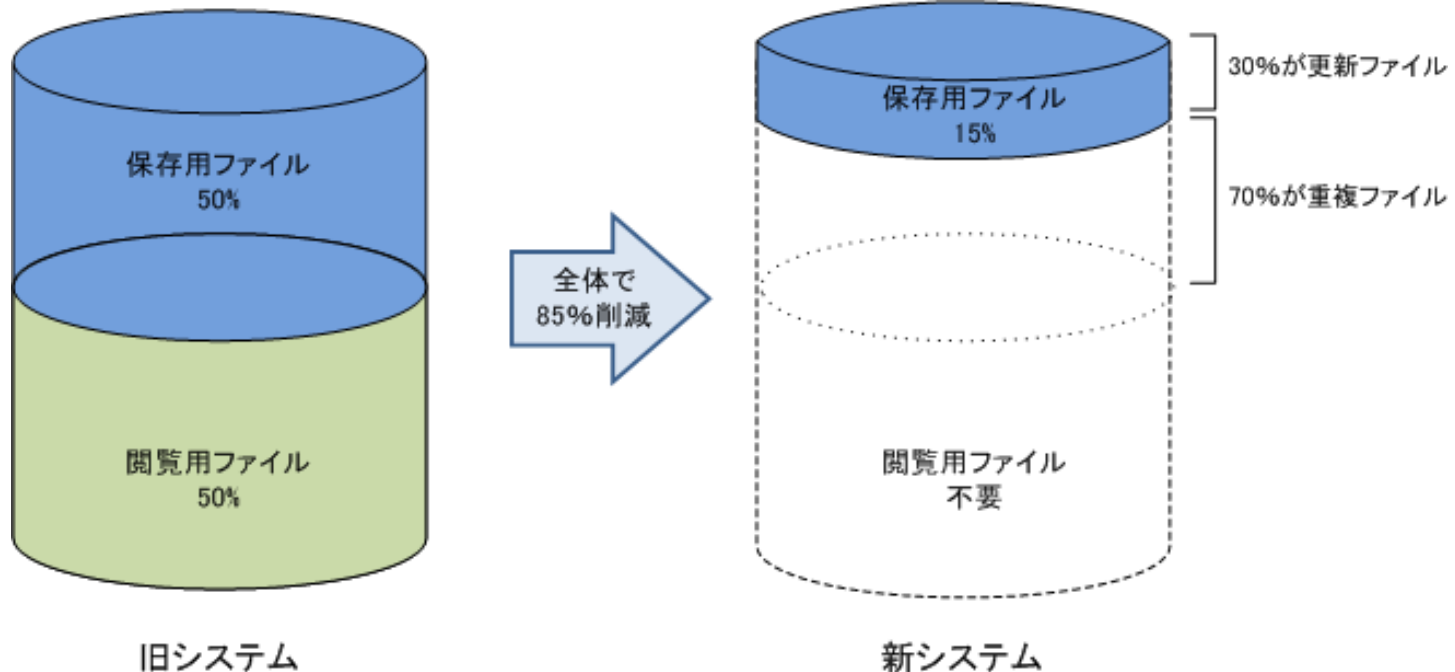


Wayback



ストレージ削減効果

- 3カ月（2013.1-3）の試行結果⇒70%が重複
（高頻度での試行のため、重複率はやや高めかも）
- 保存ファイル直接閲覧 ⇒ 50%が不要
⇒ 全体で85%の削減効果



スケールアウト・スケールアップ

・サーバ



旧: 24台

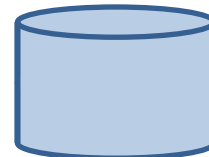
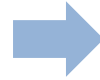


新: 50台

・ストレージ



旧: 500TB
(保存用250、閲覧用250)

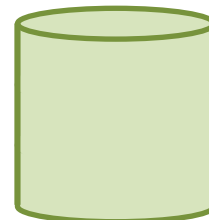


新: 500TB
(保存用のみ)

・インデクスファイル保存領域



旧: 2.8億ドキュメント



新: 8億ドキュメント

インタフェースの改善

- ・視覚的にわかりやすく
 - ・地図、組織図によるワンクリック検索
 - ・アイコンでのコレクション一覧

国立国会図書館 WebArchives
インターネット資料収集保存事業(ウェブサイト別)

▶ 本事業について ▶ 利用方法 ▶ Q&A ▶ お問い合わせ

簡易検索 | 書誌・本文検索 | ディレクトリ検索

簡易検索
(書誌情報のタイトル、編者、公開者(出版者)、起点URLと本文情報の本文、ページタイトルが検索できます。)
※ URLから検索する場合は、書誌・本文検索をご利用ください。

検索 キーワード

クリア

コレクション種別

<input checked="" type="checkbox"/> 国の機関	<input checked="" type="checkbox"/> 都道府県	<input checked="" type="checkbox"/> 政令指定都市	<input checked="" type="checkbox"/> 市町村	<input checked="" type="checkbox"/> 市町村合併
<input checked="" type="checkbox"/> 特別地方公共団体	<input checked="" type="checkbox"/> 法人・機構	<input checked="" type="checkbox"/> 大学	<input checked="" type="checkbox"/> イベント	<input checked="" type="checkbox"/> 電子雑誌
<input checked="" type="checkbox"/> その他				

収集日付範囲

年 月 - 年 月



WARP Web Archiving Project 国立国会図書館 インターネット資料収集保存事業

よくあるご質問 ヘルプ English

キーワード検索

コレクション検索

国の機関 自治体 法人・機構 大学 イベント 電子雑誌 その他

◆ 今月の特集

今月の特集は「桜咲くウェブアーカイブ」です。桜の写真や画像が使われているウェブサイトをご紹介します。

 北区(東京都) 2011年5月16日	 くら福祉保健事務組合 2011年7月31日	 学園院女子大学 2009年11月6日
---------------------------	------------------------------	---------------------------

◆ 新着情報

2013年4月3日
2013年3月の月間アクセスランキングを掲載しました。

2013年4月1日
今月の特集(2013年4月)「桜咲くウェブアーカイブ」を掲載しました。

◆ おすすめコンテンツ

ウェブアーカイブのしくみ


世界のウェブアーカイブ


特色あるコレクション


◆ 月間アクセスランキング

1位	国会事故調 (2012年10月25日)	268088
2位	財務省 (2010年6月1日)	202297
3位	国土交通省 (2009年2月17日)	183281
4位	経済産業省 (2009年7月17日)	95759
5位	総務省 (2009年1月13日)	75064

- より多くの方々に関心を持っていただきたい

-  ウェブアーカイブのしくみ

-  世界のウェブアーカイブ

-  特色あるコレクション

- アクセスランキング

- 今月の特集

4.メタデータと組織化

WARPのメタデータ

▪ DC-NDL (2011年12月版)

WARPメタデータ項目一覧

◎必須 ○あれば必須 △選択 □自動

項目	入力レベル		項目	入力レベル	
	サイト	EJ		サイト	EJ
タイトル	◎	◎	公開日	△	△
タイトル-ヨミ	◎	◎	NDC		◎
並列タイトル	○	○	ISSN		○
並列タイトル-ヨミ	○	○	ISSNL		○
編者		○	URI	□	□
編者-ヨミ		○	保存先URI	□	□
公開者(出版者)	◎	◎	NDL資源タイプ	◎	◎
公開者(出版者)-ヨミ	◎	◎	関係	△	△
巻号		○	注記	△	△
刊行頻度		○			

メタデータの例

- ・ターゲット（機関）のもと、保存日に分けて管理・公開

メタデータ	
書誌ID	000000001607
タイトル	総務省
並列タイトル	Ministry of Internal Affairs and Communications
公開者(出版者)	総務省
起点URL	http://www.soumu.go.jp/
過去の起点URL	http://www.soumu.go.jp/ http://www.soumu.go.jp/index.html
コレクション	国の機関
NDL資源タイプ	サイト

保存したウェブサイトを見る

全88件

◀◀ 1 2 3 4 5 ▶▶

保存日 (永続的識別子)

<http://www.soumu.go.jp/>

[2013/01/23 \(info:ndljp/pid/6086243\)](#)

本文検索可

[2013/01/15 \(info:ndljp/pid/6086212\)](#)

本文検索可

[2012/12/18 \(info:ndljp/pid/6021728\)](#)

本文検索可

[2012/11/12 \(info:ndljp/pid/4002126\)](#)

本文検索可

[2012/11/01 \(info:ndljp/pid/3947914\)](#)

本文検索可

収集単位とメタデータ

- WARPはターゲット単位で収集

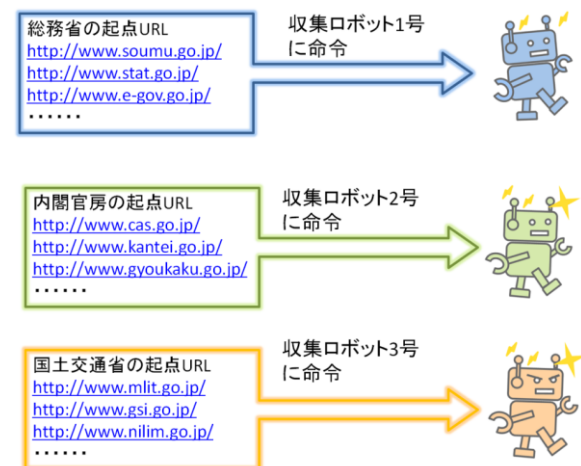
- 機関単位でメタデータを付与(収集単位のメタデータ)

- 1ターゲットに対して、複数の起点URLを設定

(例)「総務省」は、総務省、統計局、電子政府(e-Gov)など十数のURL

- 制限条件下(クローラのプロセス数、収集頻度等)での効率性

- そもそもウェブサイトの切れ目とは・・・
ドメイン? サブドメイン? ディレクトリ?
⇒いずれとも限らない



URL構造

・WARPのURL構造

http://warp.da.ndl.go.jp/info:ndljp/pid/1283840/www.ndl.go.jp/index.html

a

b

c

- a. 固定部分 : WARPで保存したウェブサイトに通で付与される部分
- b. 永続的識別子 : 特定ウェブサイトの単位で保存日ごとに付与されるID
- c. オリジナルサイトのURL

全文検索

- ・WARPの全文検索エンジンはSolr
- ・インデクスファイルの膨大化
- ・インデクス処理能力、スケールアップが課題

メタデータ 0件 本文 8809件

本文の検索結果は最大1000件までしか表示できません。

検索結果1000件中 1 ~ 10 件を表示

適合度順 降順 10 件ずつ 表示

◀ ◁ 1 2 3 4 5 6 7 8 9 10 ... 100 ▶ ▷ ▶▶

[環境省 平成22年度大気汚染状況 | PM2.5 / pm25.html](#)[HTML]
www.env.go.jp/air/osen/jokyo_h22/pm25.html [保存日:2012/03/06 - 2013/02/05]
環境省 > 大気環境・自動車対策 > 大気汚染状況・常時監視関係 > 大気環境モニタリング実施結果 > 大気汚染状況について 平成22年度大気汚染状況について (PM2.5) 微小粒子状物質 (PM2.5) 平成22
環境省

[微小粒子の年平均値の推移 / keinenn-pm25.htm](#)[HTML]
www.city.kawasaki.jp/30/30kansic/home/nennpou-index/data/data-keinenn/keinenn-pm25.htm [保存日:2012/04/10 - 2012/10/09]
微小粒子の年平均値の推移 (Trend of annual average for PM2.5) 微小粒子 (PM2.5) 詳細データ
川崎市

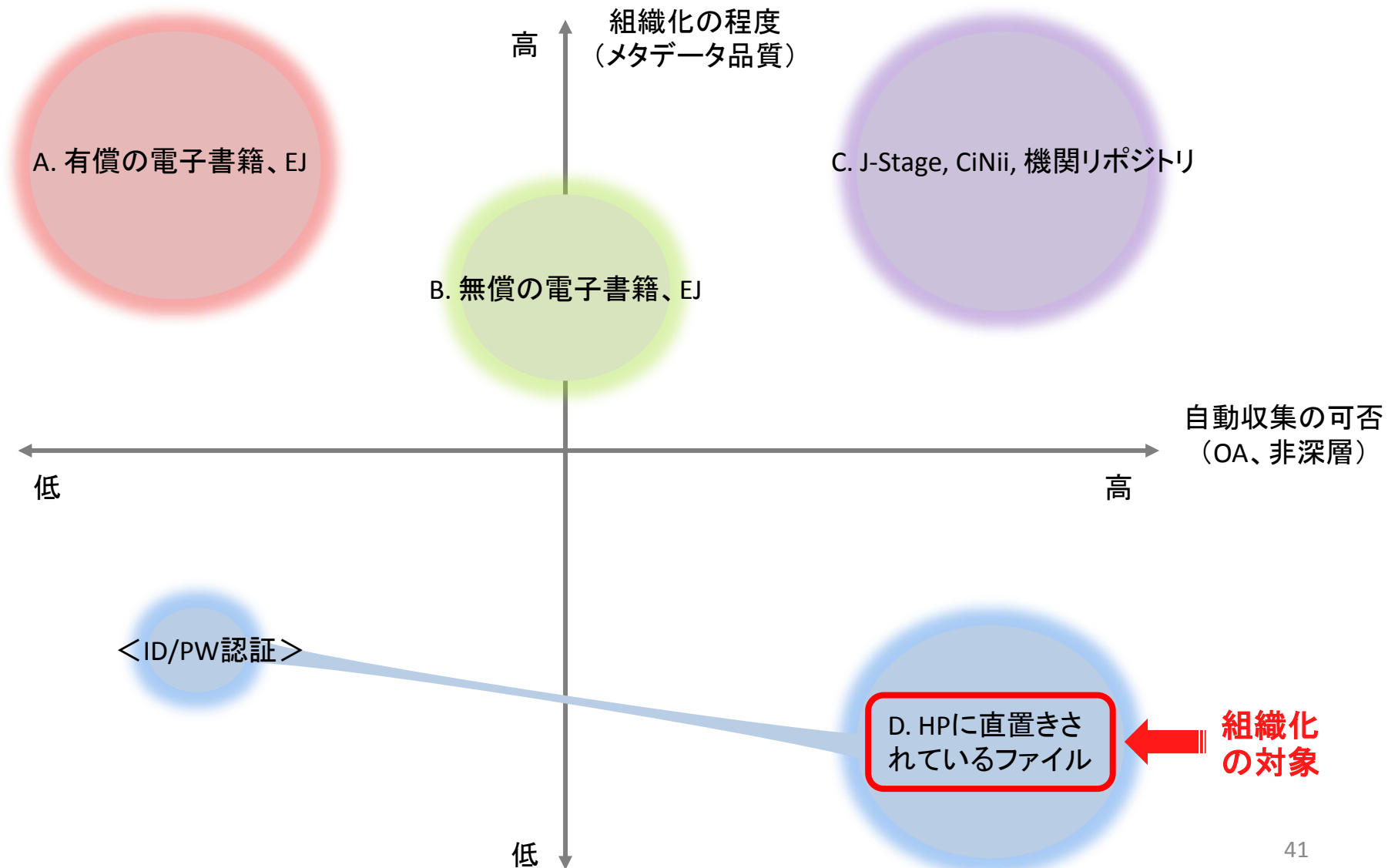
[PM2.5測定データダウンロード / taikidata-pm2.5.htm](#)[HTML]
www.city.kawasaki.jp/30/30kansic/home/html/PM2.5/taikidata-pm25.htm [保存日:2012/09/04 - 2012/10/09]
PM 2.5の測定データ 4月 5月 6月 7月 8月 9月 10月 11月 12月 1月 2月 3月 2011年度 2012年度 上記の表からダウンロードすることができるPM 2.5測定データの
川崎市

[微小粒子測定方法 / houhou-pm25.htm](#)[HTML]
www.city.kawasaki.jp/30/30kansic/home/nennpou-index/jyoukan/houhou/houhou-pm25.htm [保存日:2012/04/10 - 2012/10/09]
微小粒子状物質測定方法 (フィルター振動法) (Measuring method for PM2.5) 測定方法 (Measuring method) フィルター振動法 (Tapered

著作単位の組織化

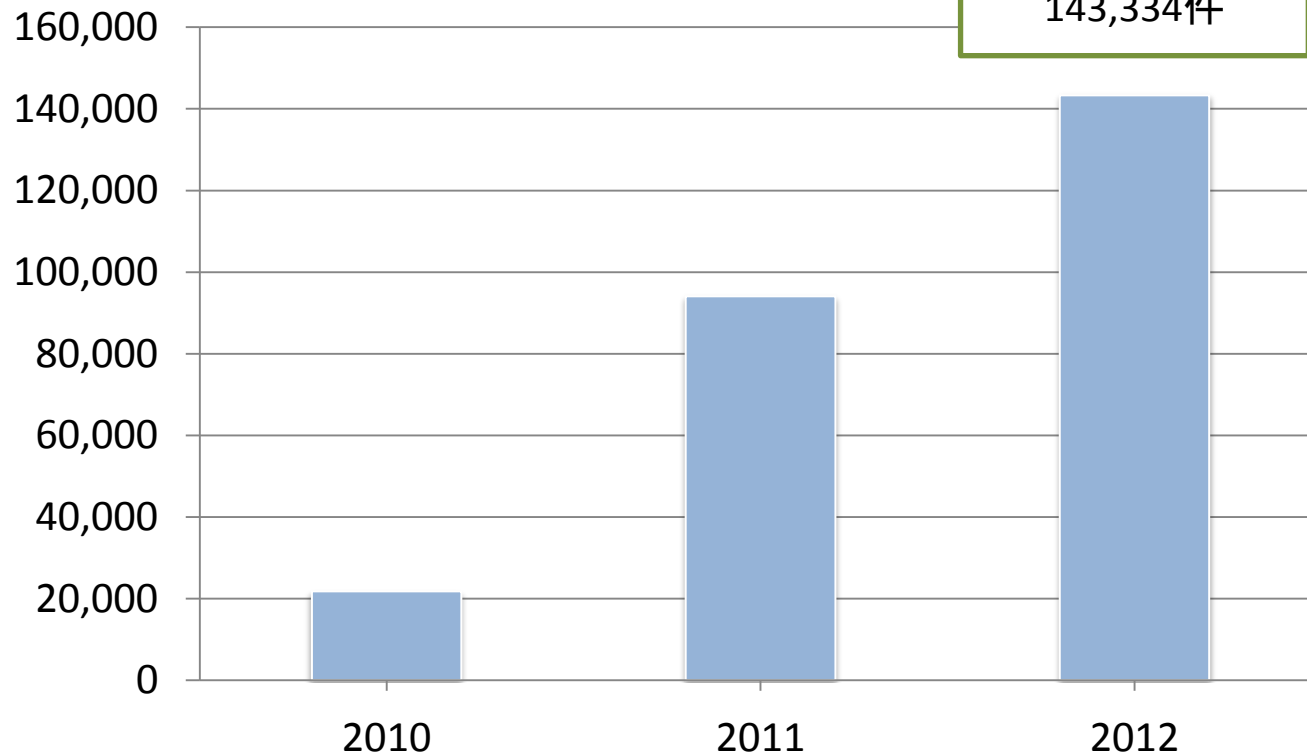
- 機関単位のメタだけでは粒度が粗い
- 全文検索はノイズが多い
- 「以後電子」との連続性
- WARPから著作単位で切り出してメタデータを付与
 - 国の刊行物、電子雑誌など重要コンテンツが主なターゲット
- タイトル単位、巻号単位、記事単位のメタデータ
- 人的コストがかかる。
 - 選別コスト、メタ付与コスト ⇒ 自動化の可能性は？

著作物ファイルの分布(概念図)



メタデータ量

累積メタデータ(件)



著作単位のメタデータ

・DC-NDL(2011年12月版)

著作単位メタデータ項目一覧

◎必須 ○あれば必須 △選択 □自動

項目	入力レベル				項目	入力レベル			
	単行	雑誌				単行	雑誌		
		タイトル	巻号	記事			タイトル	巻号	記事
タイトル	◎	◎	◎	◎	主題	△			△
並列タイトル	○	○	○	○	URL	□	□	□	□
シリーズ名	○	○	○		ISBN	○			
並列シリーズ名	○	○	○		ISSN	○	○		
シリーズ巻次	○				ISSNL	○	○		
部編名	○	○	○		永続的識別子	□	□	□	□
巻	○	○	○		Relation		○		
著者	○	○	○	○	HasVersion	○	○	○	
版表示	○	○	○		IsPartOf(ISSN,ISSNL)	○	○	○	○
出版者	◎	◎	◎	◎	IsFormatOf	△	△		
内容記述	△	△	△	△	HasFormat	○	○		
目次	○		○	○	Source	○	○	○	○
保存日	○	○	○	○	号			○	
発行日	○		○	○	通号			○	
言語	◎	◎	◎	◎	掲載雑誌タイトル				○
フォーマット(IMT)	◎		◎	◎	掲載雑誌巻号				○
階層レベル	◎	◎	◎	◎	アクセス制限	◎	◎	◎	◎
コレクション情報	◎	◎	◎	◎					

メタデータの例

書誌情報

簡易レコード表示にする

タイトル (title)

イギリスの2011年議会任期固定法

著者 (creator)

河島太郎

掲載雑誌名 (publicationName)

外国の立法：立法情報・翻訳・解説

掲載巻号 (publicationVolume)

(254)

出版者 (publisher)

国立国会図書館

出版年月日 (W3CDTF形式) (issued:W3CDTF)

2012-12

フォーマット (IMT形式) (format:IMT)

application/pdf

前の巻 (永続的識別子) (previous:NDLJP)

info:ndljp/pid/4023706

次の巻 (永続的識別子) (next:NDLJP)

info:ndljp/pid/4023708

上位資料 (永続的識別子) (relation:isPartof-ndljp)

info:ndljp/pid/4023705

上位資料 (ISSN) (isPartOf:ISSN)

13492071

上位資料 (Linking ISSN) (isPartOf:ISSNL)

0433096X

永続的識別子 (identifier:NDLJP)

info:ndljp/pid/4023707

URL (identifier:URI)

<http://dl.ndl.go.jp/info:ndljp/pid/4023707>

主題タグ (分野) (subject:genre)

議会

選挙

憲法

主題タグ (国・地域) (subject:area)

イギリス

言語 (ISO639-2形式) (language:ISO639-2)

jpn

コレクション情報 (type:collection)

国の機関-国会-国立国会図書館-国立国会図書館調査及び立法考査局

受理日 (W3CDTF形式) (dateAccepted:W3CDTF)

2012-12-12T00:42:01Z

提供者 (provider)

国立国会図書館調査及び立法考査局連携協力課_001

提供制限 (accessRights)

インターネット公開

階層レベル (type:biblevel)

3

Web入手区分 (type:Web-get)

1

URL

<http://dl.ndl.go.jp/info:ndljp/pid/4023707>

著作単位の公開

- ・デジタル化資料と一緒に公開
(デジタル化資料との親和性)



国立国会図書館デジタル化資料


検索 館内限定公開資料を含める

国立国会図書館で収集・集積されているさまざまなデジタル化資料を検索・閲覧できるサービスです。

インターネット資料

当館が収集したインターネット上の刊行物をご覧いただけます。国の機関や地方公共団体、独立行政法人、大学などがウェブサイトに掲載した白書、年鑑、報告書、広報誌、雑誌論文などを収録しています。

※収集・保存したウェブサイトはインターネット資料収集保存事業(WARP)をご覧ください。



詳細検索へ [→](#)

インターネット資料のタイトル一覧 [↑](#)

平成 19 年度

農林水産省年報

農林水産省

農林水産省年報
農林水産省大臣官房情報評価課

農林水産省の年報。農林水産行政の各分野において講じた施策等を記している。

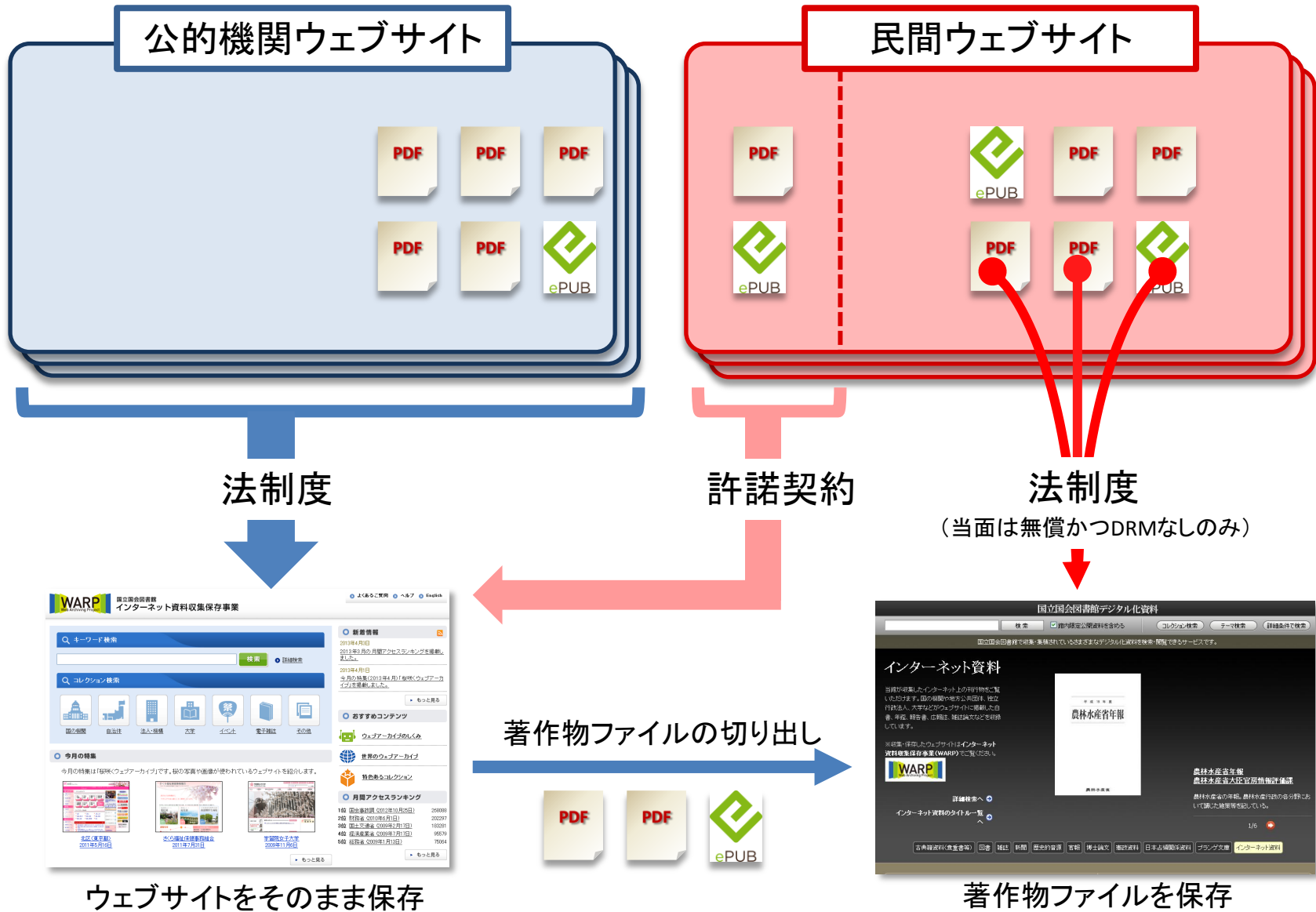
1/6 [→](#)

古典籍資料(貴重書等) 図書 雑誌 新聞 歴史的音源 官報 博士論文 憲政資料 日本占領関係資料 プランゲ文庫 **インターネット資料**

民間オンライン資料の制度収集

- インターネットで公開されているうち、図書、雑誌に相当するものがオンライン資料
- 2013年7月より、私人のオンライン資料は納入義務
- 当面は無償かつDRMのないもの
- 特定コード（ISBN、ISSN、DOI）があるもの、もしくは特定フォーマット（PDF、EPUB、DAISY）が対象
- 自動収集、送信、送付の何れかの方法

ウェブコンテンツの収集・組織化モデル



5.課題と展望

収集が難しいもの

- JavaScriptで呼び出されるファイル
- ストリーミングファイル(動画)
- データベース内のファイル(深層ウェブ)
- SNSは技術的だけでなく制度的な課題も(robots.txtの修正義務が及ばない)
- 新技術への常なる対応

⇒世界各国(IIPC)と共同して課題解決

いかに利用するか

- 「いかに集めるか」だけでなく「いかに利用するか」

- データマイニング

 - “Web archiving use cases”

 - <http://netpreserve.org/resources/web-archiving-use-cases-0>

 - (Text mining、Link analysis、技術変遷分析、etc.)

- 過去データのデポジット機能

 - オリジナルサイトから過去データを消去してWARPに誘導
(総務省、国土交通省、文部科学省、etc.)

- 切り出し自動化、検索機能の高度化

 - 対象発見(セマンティック)

 - メタデータ付与(正解集合に基づくパターン認識)

ありがとうございました！