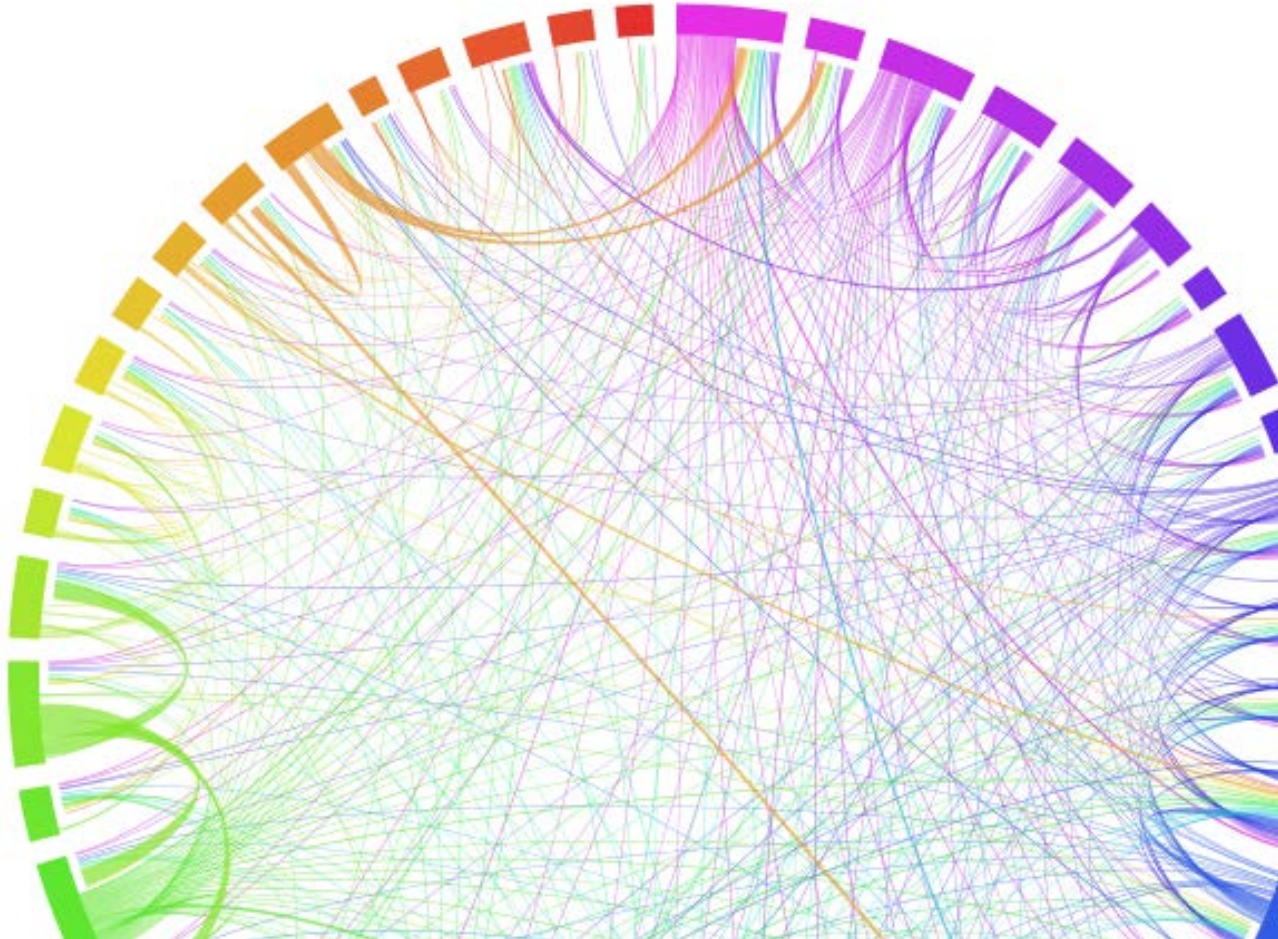
2016.7.30 NDLデータ利活用ワークショップ
～ウェブ・アーカイブの自治体サイトを可視化しよう～

WARPとデータセット



国立国会図書館

WARP

WARPとは

・ウェブサイトのアーカイブ

・2002年に始めて15年目

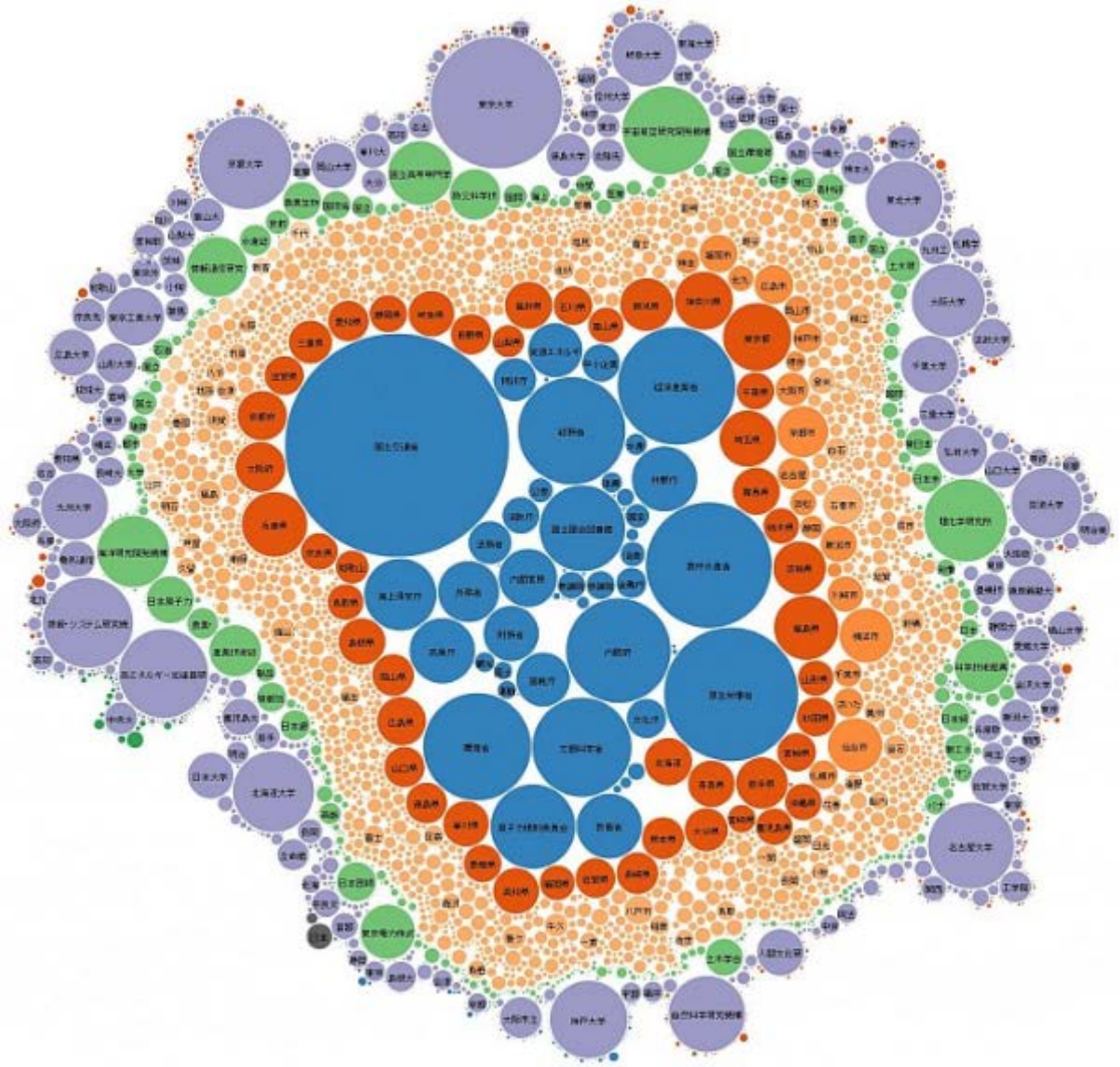
・2010年から公的機関サイトを大規模に収集

http://warp.da.ndl.go.jp/

集めているウェブサイト

区分	根拠	対象	サイト数	収集頻度
公的機関	法律	国の機関	5,400	月1回
		地方自治体		年4回
		独立行政法人		
		国公立大学		
民間	契約	公益法人、私立大学、政党、イベント、震災、電子雑誌	4,600	年1～4回

容量で可視化



本日のターゲット

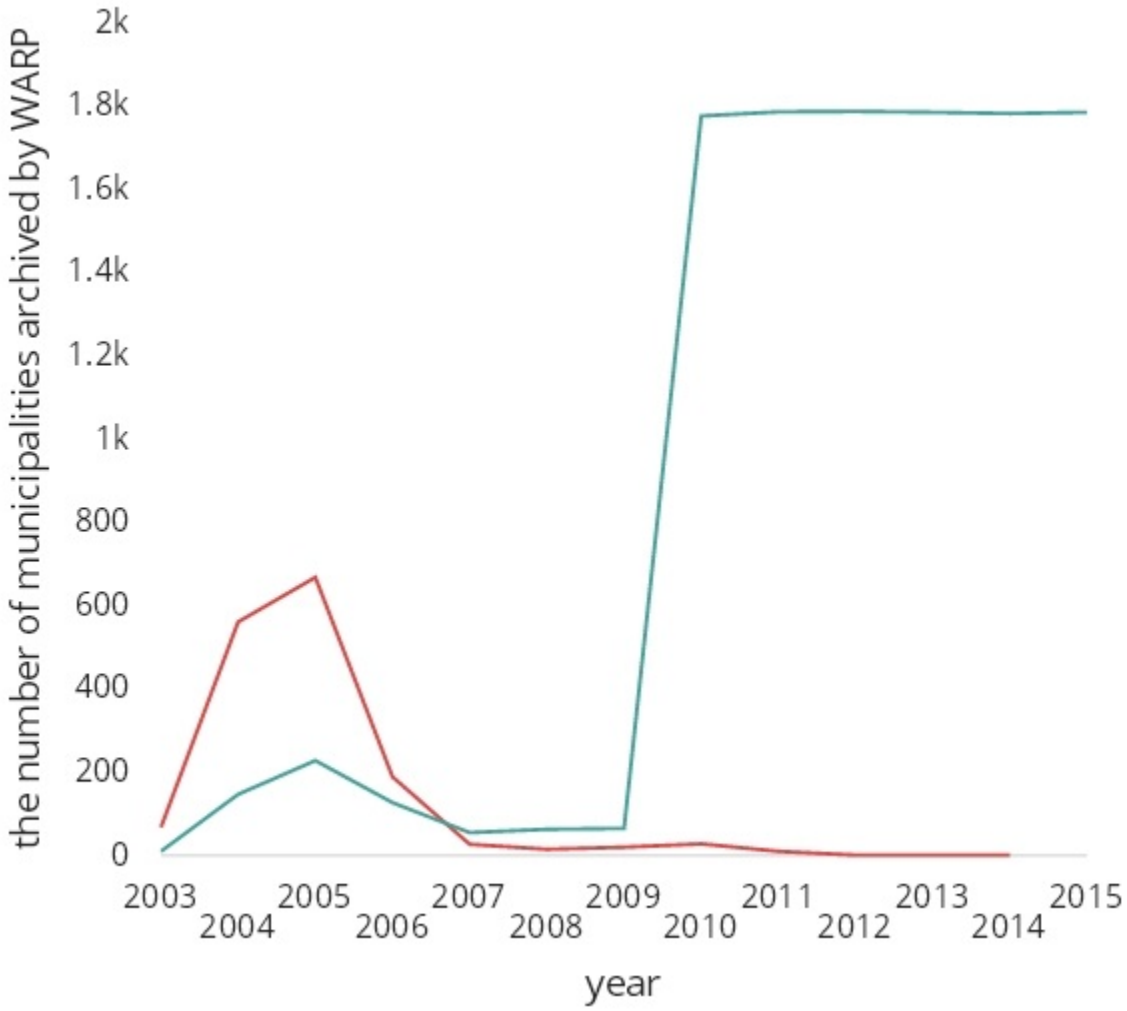
- 都道府県
- 政令指定都市
- 市町村
- 特別地方公共団体（東京23区を含む）

218TB/633TB

内訳

収集対象	現存する自治体	消えた自治体	合計
都道府県	47	–	47
政令指定都市	20	–	20
市町村	1,698	989	2,687
東京23区	23	–	23
合計	1,788	989	2,777

保存状況



現存する自治体

2010年～
全自治体を保存

消えた自治体

～2009年
消えた自治体が多い

具体例をみてみましょう

・消えた町 – 佐賀県大和町

http://warp.da.ndl.go.jp/info:ndljp/pid/246720/www.saganet.ne.jp/yamato/

・2003年の香川県

http://warp.da.ndl.go.jp/info:ndljp/pid/236640/www.pref.kagawa.jp/

・2012年の・・・（うどん県）

http://warp.da.ndl.go.jp/info:ndljp/pid/6019057/www.my-kagawa.jp/udon-ken/top.html

・2015年の香川県

http://warp.da.ndl.go.jp/info:ndljp/pid/9498887/www.pref.kagawa.jp/

データセット

本日、使えるデータ

1. メタデータ

2. 検索API

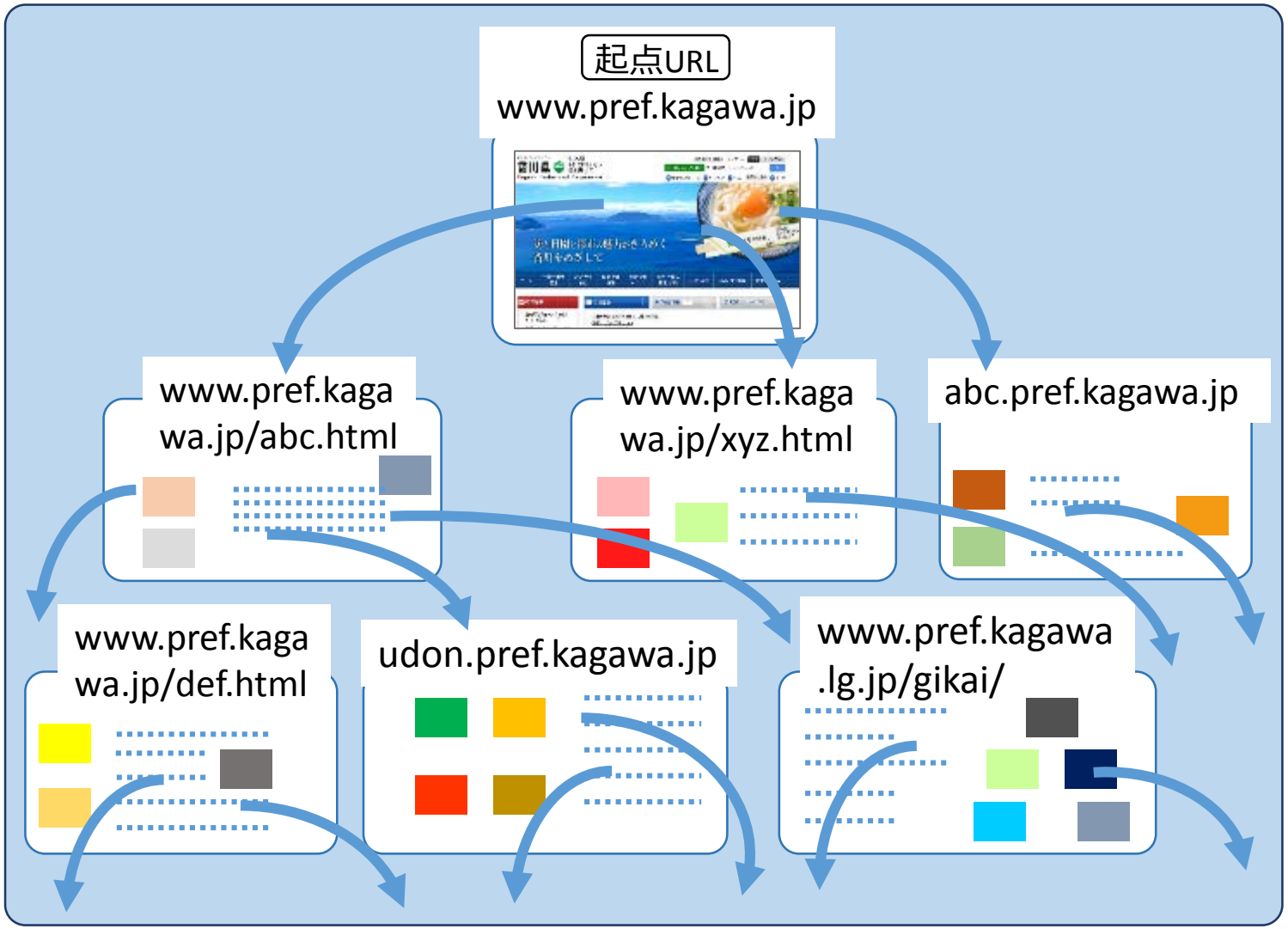
1. メタデータ

・2003年から2015年の自治体サイトの全件メタデータ

・2つのメタデータ

収集対象	（自治体）	2,777 件
収集個体	（収集回ごとのまとまり）	47,318 件

ここで収集のイメージを



クローラによる収集

① 起点URLにあるファイルを複製

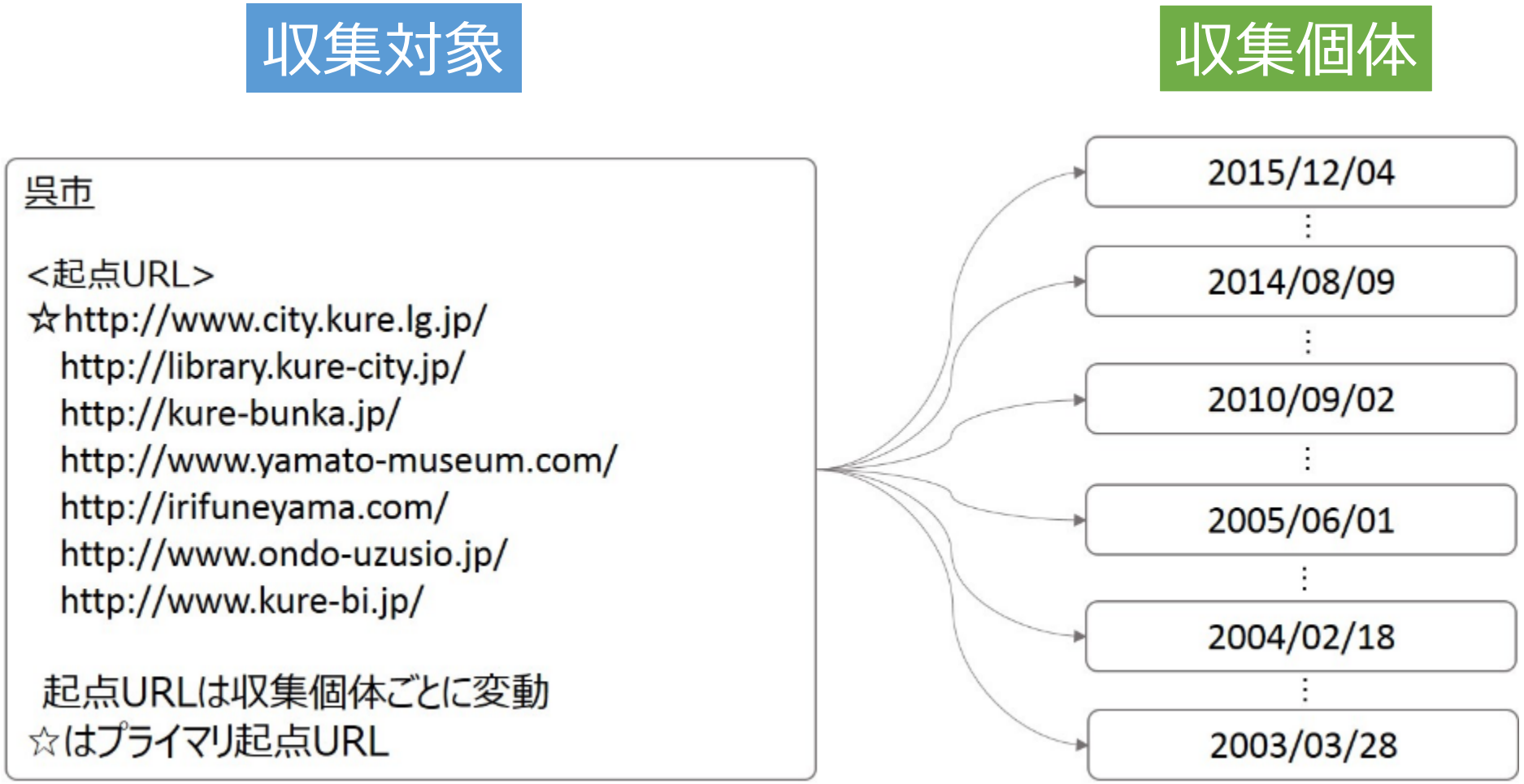
② リンクを解析してページ遷移

③ リンク先のページでも同じように
ファイル複製、リンク解析

④ 指定した範囲の全てのファイルを
複製するまで繰り返す

⑤ ファイルをひとまとめにして保存

2つの関係


収集対象
収集個体
呉市
<起点URL>
☆http://www.city.kure.lg.jp/
http://library.kure-city.jp/
http://kure-bunka.jp/
http://www.yamato-museum.com/
http://irifuneyama.com/
http://www.ondo-uzusio.jp/
http://www.kure-bi.jp/
起点URLは収集個体ごとに変動
☆はプライマリ起点URL
2015/12/04
2014/08/09
2010/09/02
2005/06/01
2004/02/18
2003/03/28

WARP画面


メタデータ
書誌ID
000000001223
タイトル
呉市
公開者(出版者)
呉市 (広島県)
公開日
1997/04/01
起点URL
http://www.city.kure.lg.jp/
過去の起点URL
http://www.city.kure.lg.jp/
http://www.city.kure.hiroshima.jp/index.html
コレクション
市町村
NDL資源タイプ
サイト
保存したウェブサイトを見る
全33件
保存日
(永続的識別子)
http://www.city.kure.lg.jp/
2016/03/04 (info:ndljp/pid/9907548)
本文検索可
2015/12/04 (info:ndljp/pid/9551856)
本文検索可
2015/09/04 (info:ndljp/pid/9493519)
本文検索可
2015/06/04 (info:ndljp/pid/9377386)
本文検索可
2015/03/06 (info:ndljp/pid/9102376)
本文検索可
2014/12/02 (info:ndljp/pid/8829203)
本文検索可
収集対象
収集個体

必ずしも100%ではありません

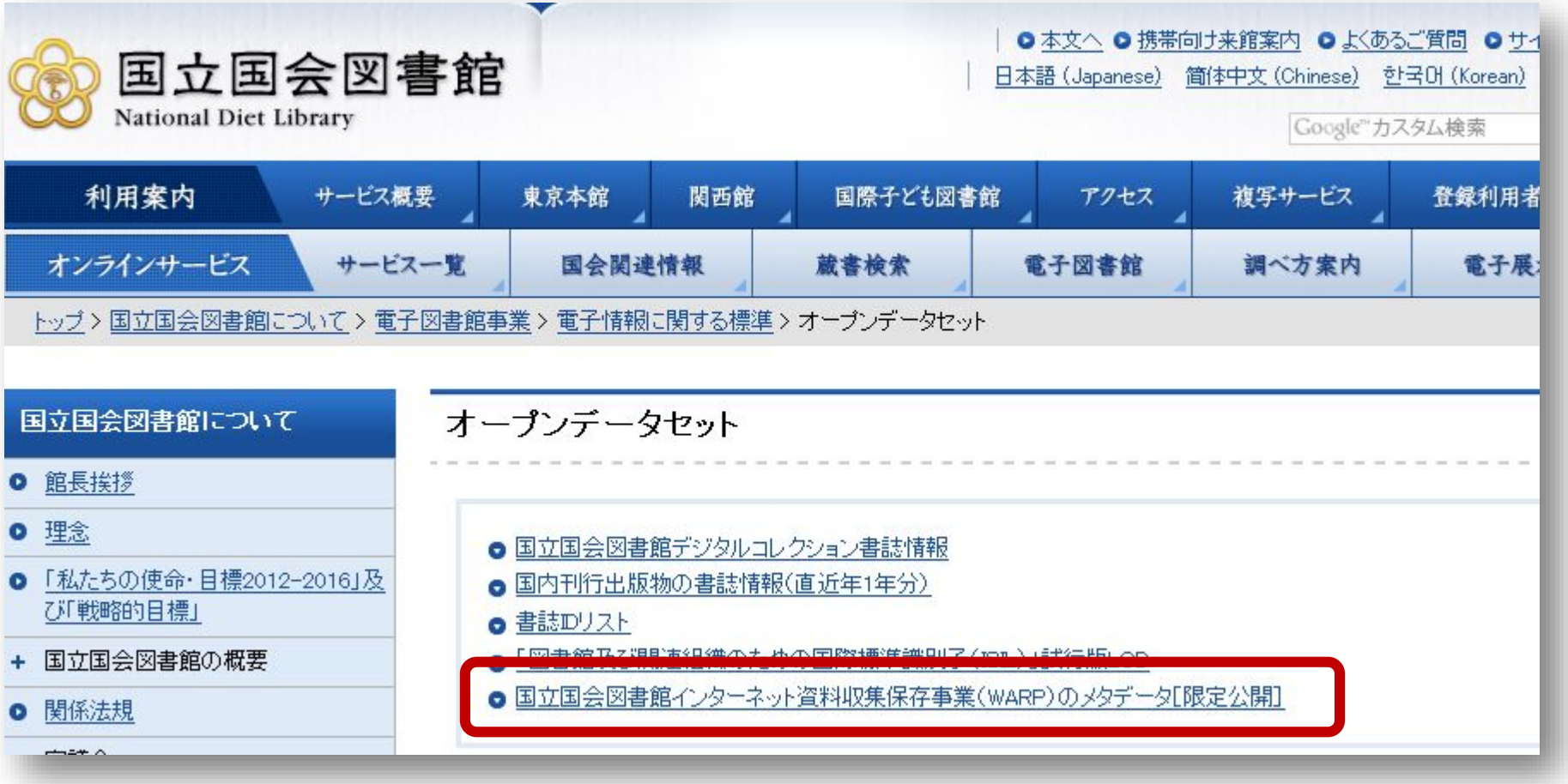
・技術的にとれないもの

・収集回ごとの時間制限（オーバーしたら停止）

自治体	2015年3月まで	2015年４月以降
都道府県 政令指定都市	5日	20日
市町村 東京23区	1日	

詳細とダウンロードはこちら

http://www.ndl.go.jp/jp/aboutus/standards/opendataset.html



2. 検索API

・自治体サイトをページ単位で検索できる

ページ数	62,286,266 ページ
自治体数	1,788（47都道府県、20政令指定都市、1,698市町村、東京23区）
対象年	2010年、2013年、2015年

・キーワード、外部リンクなど様々な情報を取得

詳細はこちら


WARP自治体サイト検索API
データ
API仕様
簡易検索
「WARP自治体サイト検索API」は、国立国会図書館のイベントのために試行的に作成し、イベント限定で公開するものです。このAPIでは、「国立
(WARP)」で保存・インターネット公開されている自治体(都道府県及び市区町村)のウェブサイトについて、データ可視化用のデータを検索するこ
NDLデータ利活用ワークショップ～ウェブ・アーカイブの治体サイトを可視化しよう～
2016年7月30日(土)10:00 ～17:00
国立国会図書館東京本館　新館3階大会議室(東京都千代田区永田町1-10-1)
イベントのページへ
API及びメタデータの利用にあたってのご注意
上記イベント及びそのプレイベントでの利用に当たっては、手続きなしでご利用いただけます。イベント外での利用及び営利目的での利用
までご相談ください。
上記イベントの中で本API、メタデータを利用して開発した作品は、「アーバンデータチャレンジ2016(外部サイト)」に応募することができま
る際には、「国立国会図書館インターネット資料収集保存事業(WARP)」のデータを利用したことを明記してください。
サーバへの過負荷により、応答速度が低下したり、応答が無くなる場合があります。予めご了承ください。
本APIは、試行版です。予告なく仕様を変更したり公開を終了したりする場合があります。本APIまたはメタデータを利用したことに起因また
図書館は責任を負いません。

補足

・館内でのみ見られるものが19%

⇒ 各グループのPCで見られます