ResourceSync：OAI-PMHの後継規格

林 豊（九州大学附属図書館eリソースサービス室）
hayashiyutaka@gmail.com
ORCID: 0000-0001-7761-3444

情報組織化研究グループ月例研究会（2015/11/28、大阪学院大学）

元ネタ

- CA1845 - ResourceSync：OAI-PMHの後継規格 / 林 豊
  - http://current.ndl.go.jp/ca1845
- ＋2015年1月以降の動向

もくじ

1. Introduction
2. OAI-PMH：図書館界における現在のスタンダード
3. ResourceSync：概要、意義、可能性
4. まとめ

ねらい

- 対象：機関リポジトリやデジタルアーカイブ（のシステム連携）を担当したことがないひと（でも）

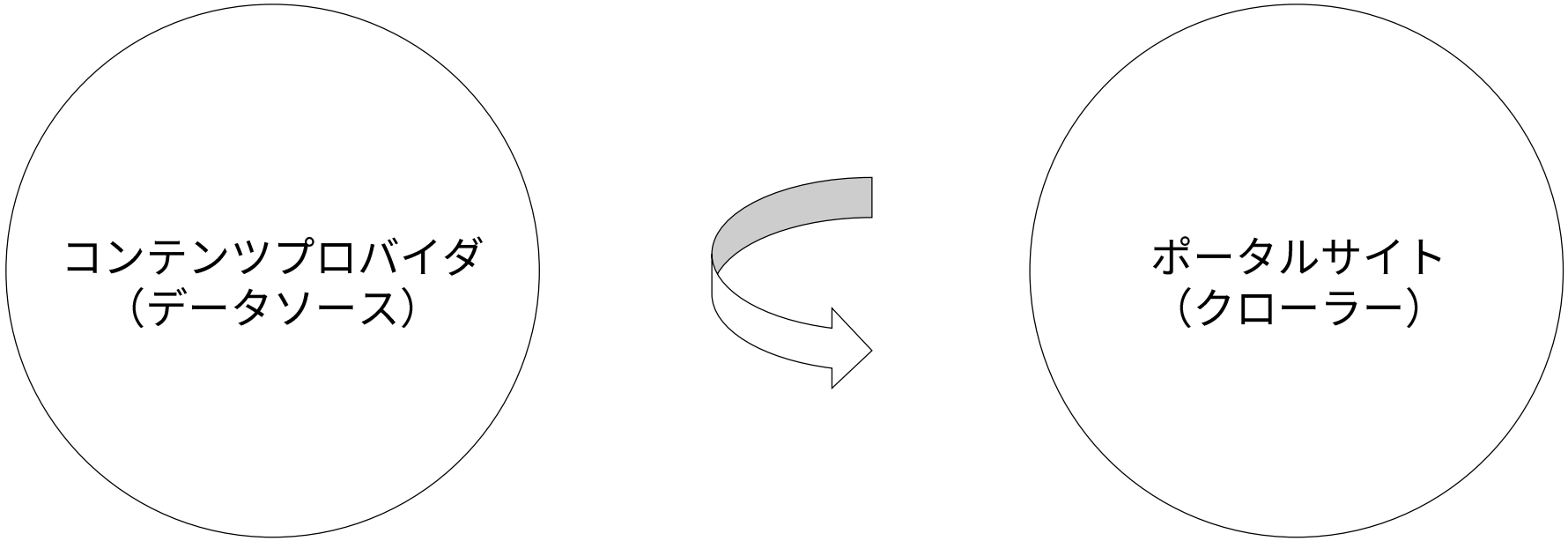
- 目的：ResourceSyncという標準規格の意義（＝あるいはOAI-PMHの問題点）を理解してもらう

サマリー

- 図書館等ではメタデータ収集（ハーベスティング）の方法（API）としてOAI-PMHがデファクトスタンダードになっている
- だが誕生から15年以上が経つOAI-PMHにはいくつかの限界がある（メタデータのみ、非リアルタイム、マイナー・旧式）
- そこでResourceSyncというコンテンツ同期のための標準規格群の策定が進められている
- ResourceSyncの実装例はまだ少ない（CiNii Dissertationsくらい？）

1. Introduction

クローリングとは



ウェブサイト
ブログ
デジタルアーカイブ
機関リポジトリ
OPAC

Google
Yahoo!
NDLサーチ
JAIRO

なぜクローリングしてもらうのか？

- コンテンツの露出度アップのため
  - 利用者の多いポータルサイトからの動線を作る
- 統合検索のため
- データをまとめる（一元管理）

クローリングに必要なもの

- ポータルサイトによる認知
  - 受動的（外部サイトからのリンクなど）
  - 主体的（Googleウェブマスターツールなどによる申請）
- コンテンツのリストアップ
  - XMLサイトマップ（＋robots.txt）
  - RSS/Atom
  - OAI-PMH
  - ResourceSync Framework Specification
- 更新タイミングの通知
  - Ping
  - PubSubHubbub（fat ping）
  - 一定間隔でのクローリング
  - ResourceSync Notification

XMLサイトマップの例

```
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9">
<url>
        <loc>http://catalog.lib.kyushu-u.ac.jp/ja/recordID/30001</loc>
        <lastmod>2015-05-02T02:00:13+09:00</lastmod>
</url>
<url>
        <loc>http://catalog.lib.kyushu-u.ac.jp/ja/recordID/30002</loc>
        <lastmod>2015-05-02T02:00:13+09:00</lastmod>
</url>
<url>
...


</urlset>
```

ウェブページのURL
最終更新日

の繰り返し

Source: http://catalog.lib.kyushu-u.ac.jp/sites/default/files/sitemap/xml/sitemap1.xml

2. OAI-PMH

図書館界における現在のスタンダード

2.1. 概要

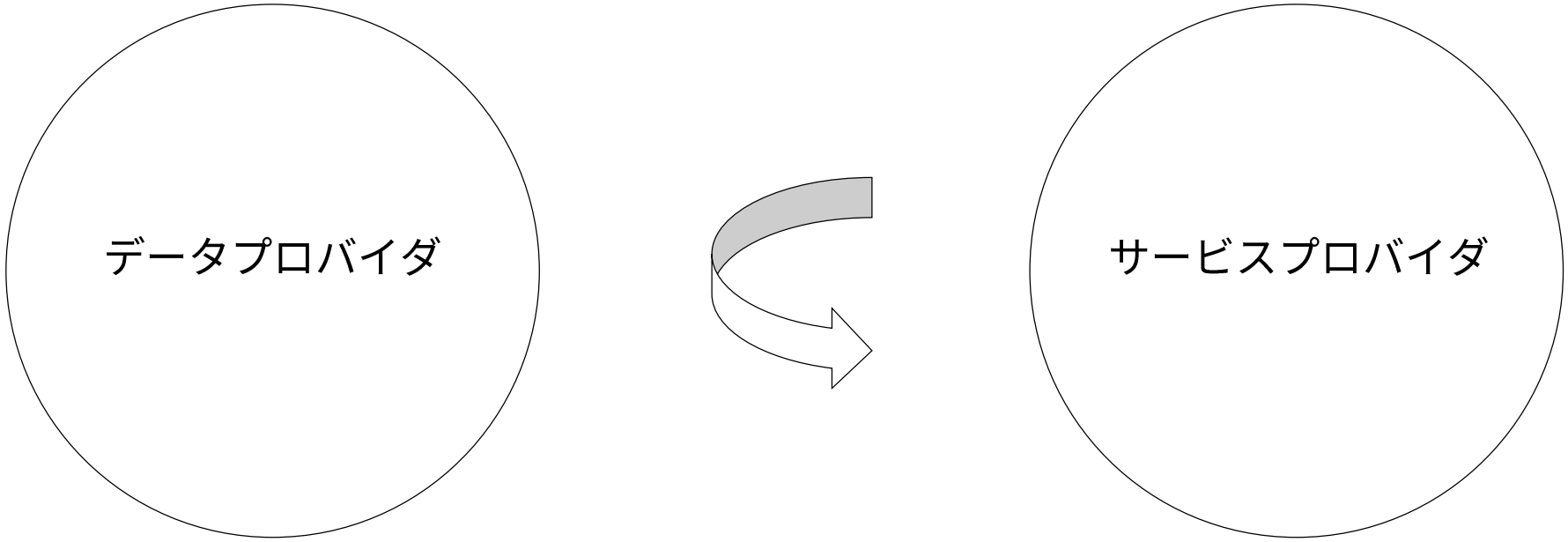
OAI-PMHとは

- Open Archives Initiative Protocol for Metadata Harvesting

- 歴史
  - 1999年10月：米国サンタフェでeプリントアーカイブの会議　→OAIの誕生
  - 2001年1月：OAI-PMHバージョン1.0リリース
  - 2002年6月：OAI-PMHバージョン2.0リリース
  - 現在でも国際標準規格にはなっていない

- 図書館界におけるメタデータ収集方法のデファクトスタンダード

①特定のURLにアクセス
http://api.lib.kyushu-u.ac.jp/opac/mmd_api/oai-pmh/?verb=ListRecords&metadataPrefix=junii2&from=2015-01-05&until=2015-01-12



②XMLでメタデータを返す

URLの構造

- 「?」以前の部分（ベースURL）
  - 例）http://api.lib.kyushu-u.ac.jp/opac/mmd_api/oai-pmh/
  - データプロバイダによって異なる
  - サービスプロバイダはデータプロバイダのベースURLを事前に知っておく必要がある

- 「?」以後の部分
  - 例）verb=ListRecords&metadataPrefix=junii2&from=2015-01-05&until=2015-01-12
  - OAI-PMHで記述方法が標準化されている
  - verb＝要求内容の指定（ListRecords、Identify、ListMetadataFormats、ListSets、ListIdentifiers、ListRecords、GetRecord）
  - metadataPrefix＝メタデータスキーマの指定
  - from、until＝メタデータの更新日の指定

デモ：XMLの例

- http://api.lib.kyushu-u.ac.jp/opac/mmd_api/oai-pmh/?verb=ListRecords&metadataPrefix=junii2&from=2015-11-19&until=2015-11-19

削除レコード

- メタデータが削除されると、status=“deleted”が出力される（任意）
  →サービスプロバイダ側でもそのレコードを削除処理する

```
<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xsi:sc
http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
  <responseDate>2015-11-23T06:57:53Z</responseDate>
  <request verb="GetRecord" metadataPrefix="junii2" identifier="oai:catalog.lib.kyushu-u.ac.jp:2324/1526112">
    http://api.lib.kyushu-u.ac.jp/opac/mmd_api/oai-pmh/
  </request>
  <GetRecord>
    <record>
      <header status="deleted">
        <identifier>oai:catalog.lib.kyushu-u.ac.jp:2324/1526112</identifier>
        <datestamp>2015-11-20T05:37:48Z</datestamp>
      </header>
    </record>
  </GetRecord>
</OAI-PMH>
```

2.2. 実装例

OAI-PMHに対応するとは

- サービスプロバイダ
  - データプロバイダのベースURLを収集する
  - OAI-PMHリクエストを行い、メタデータを収集・処理できるようにする

- データプロバイダ
  - OAI-PMHリクエストに応じてメタデータを出力できるようにする

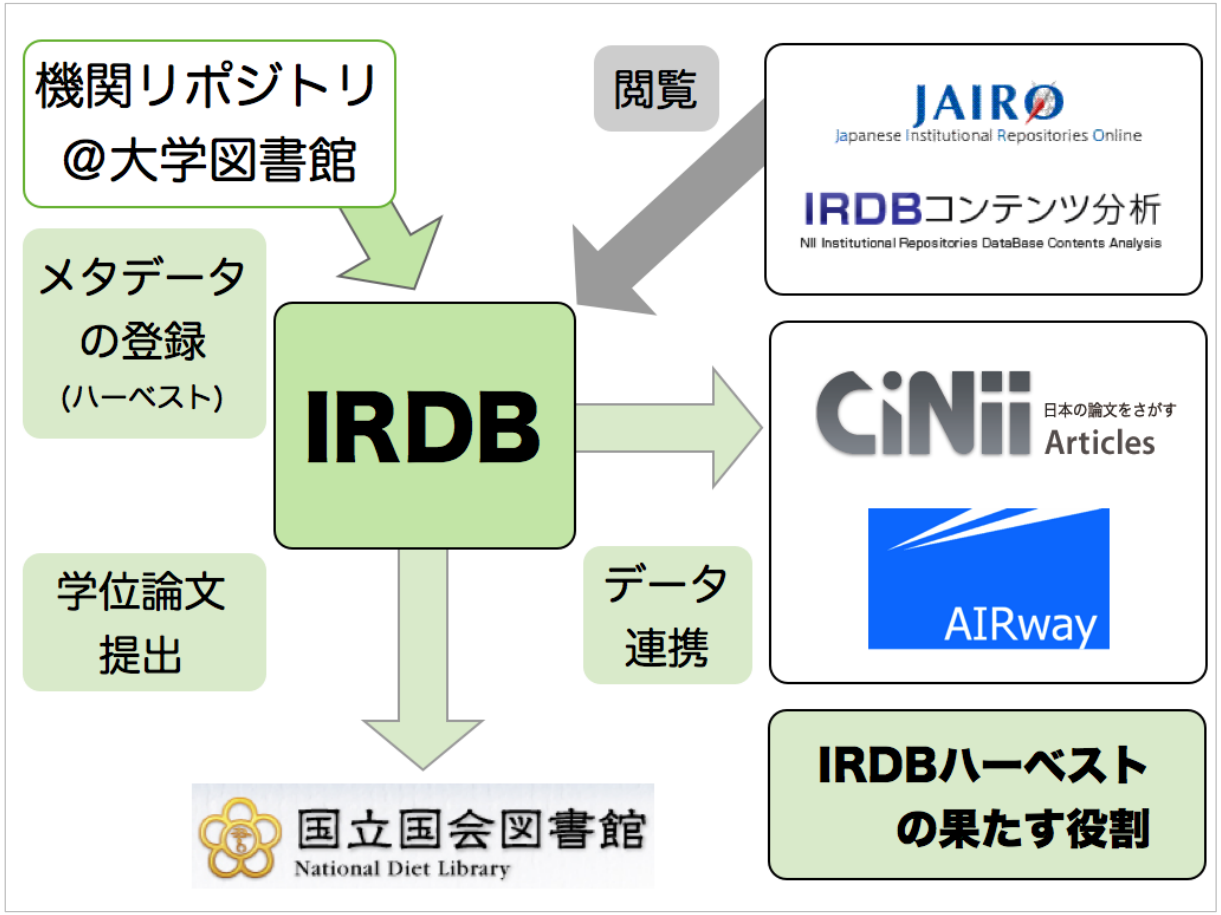
(1)デジタルリポジトリ

- データプロバイダ
  - 機関リポジトリ
  - サブジェクトリポジトリ（arXivなど）
- サービスプロバイダ
  - IRDB（JAIRO）
  - OAIster
  - RePEc

- DSpaceやEPrintsなどの主要ソフトウェアがOAI-PMHを実装

例：IRDB（JAIRO）

- JAIROでの統合検索
- IRDBコンテンツ分析
  http://irdb.nii.ac.jp/analysis/index.php
- CiNii A, Dとの連携
- NDLへの博士論文提出
- JaLC DOIの登録



Source: http://www.nii.ac.jp/irp/archive/system/irdb_harvest.html

(2)デジタルアーカイブ

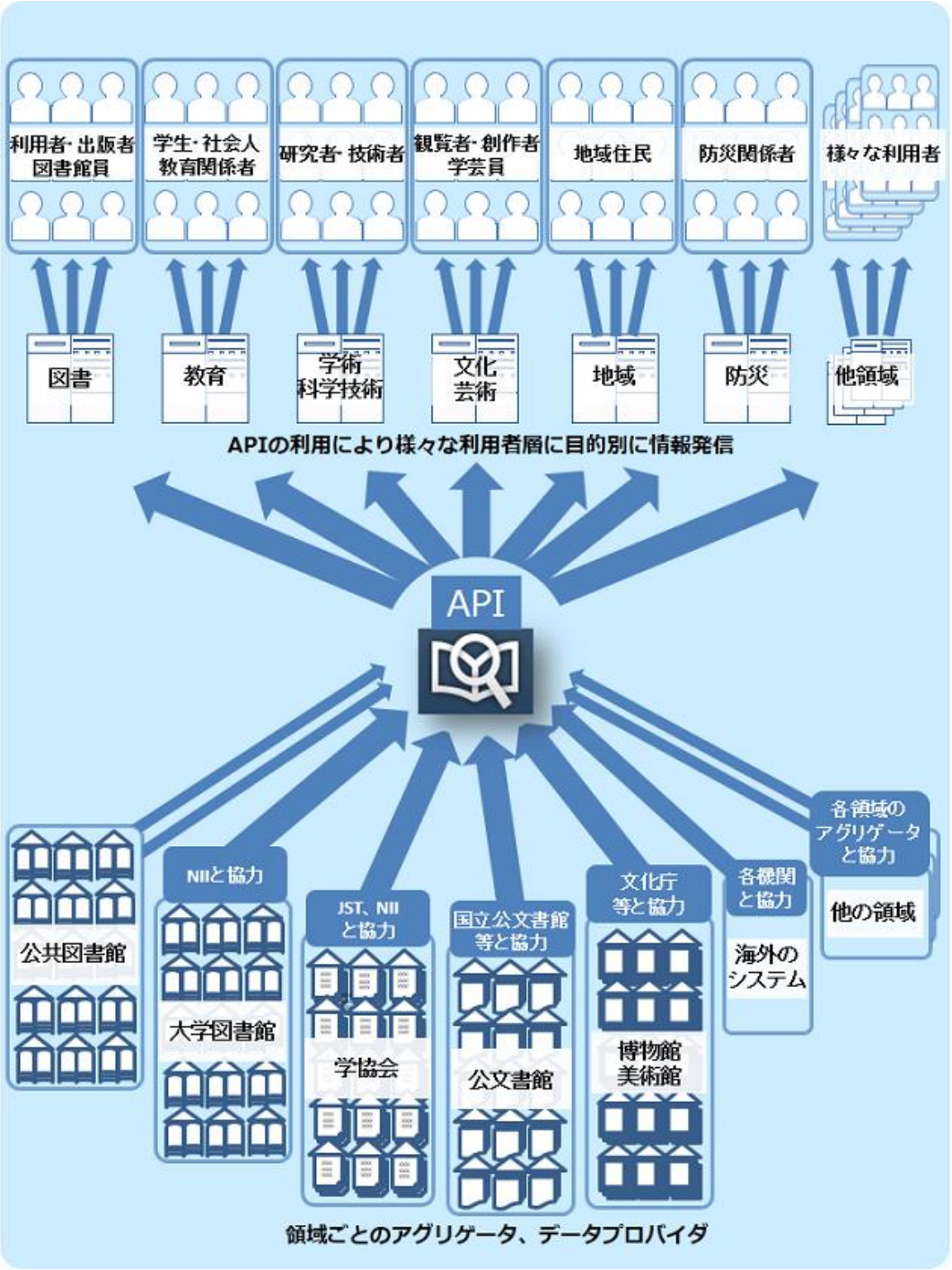
- NDL東日本大震災アーカイブ「ひなぎく」
- Europeana
- 米国デジタル公共図書館（DPLA）
- American Memory（LC）

★図書館界以外の機関・企業のサービスも収集対象になりうる
★メタデータだけでいいの？

(3)ディスカバリーサービス

- 国立国会図書館サーチ
  - 国立国会図書館サーチ連携拡張に係る実施計画（2015年4月）
    http://iss.ndl.go.jp/information/2015/04/03_announce-2/
    ★API推し
  - OAI-PMHの要点（2015年10月）
    http://iss.ndl.go.jp/information/2015/10/02_announce_1/

- ウェブスケールディスカバリーサービス
  - OPAC→
  - FTPでやっているところもある



2) 公共図書館
2-1) 総合目録

○ 総合目録については、既に都道府県及び政令指定都市からのデータ提供がほぼ実現しているため、NDL サーチにデータを提供する際の方式を OAI-PMH に切り替えることに重点を置く。これにより、書誌データの更新頻度の向上、NDL サーチから公共図書館の書誌画面へのアクセスの改善等を実現する。

○ 現在 12 館について OAI-PMH への切替えが完了しているが、今後年間約 5 館程度ずつ OAI-PMH への切替えを実施し、5 年後には、データ提供館 66 館のうち 6 割弱の37館程度について OAI-PMH でのメタデータ授受が実現することを今後 5 年間の目標とする。最終的には全てのデータ提供館との間で、OAI-PMH でのメタデータ授受が実現することを目標とする。

Source:
http://dl.ndl.go.jp/view/download/digidepo_9207570_po_iss_plan.pdf?contentNo=1

2.3. 限界

(1)メタデータしか収集できない

- メタデータの記述対象であるコンテンツ自身の収集に対応していない
  - 多様なアクセスポイントの提供　→　OK
  - 長期保存のためのバックアップ　→　×
  - 全文検索機能の提供　→　×
  - ミラーサイト設置（arXiv）　→　×

- 注）かつてSompelらがOAI-PMHによるコンテンツ収集を提案したが、その方法はあまり普及しなかったと後年振り返られている
  http://www.dlib.org/dlib/december04/vandesompel/12vandesompel.html
  http://www.dlib.org/dlib/september12/vandesompel/09vandesompel.html

(2)図書館界以外では普及していない

- 2008年、Google SitemapsがOAI-PMHのサポートを終了
  - 利用が少ないという理由でXMLサイトマップに一本化
  - http://googlewebmastercentral.blogspot.jp/2008/04/retiring-support-for-oai-pmh-in.html

- OAI-PMHは現在のウェブの標準的なスタイルではない
  - 2001年策定（Web 2.0よりも前）
  - RESTfulじゃない（Z39.50→SRW/SRU）

⇒図書館界で収集するだけなら問題がなくても、一般のサービスを対象する場合には？（特に震災アーカイブ）

(3)収集がリアルタイムではない

- OAI-PMHはプル型
  - コンテンツが生成されたことをデータプロバイダからサービスプロバイダに通知（プッシュ）する手段がない

⇒コンテンツの生成から収集までにタイムラグが生じがち
  - メタデータだけでなくコンテンツも収集するなら、リアルタイム性は重要

3. ResourceSync

概要、意義、可能性

3.1. 概要

“ResourceSync”の規格群

- **ResourceSync Framework Specification**
  - コアとなる規格

- ResourceSync Archives（ベータ版）
  - コアを拡張
  - Memento（ウェブアーカイビングプロジェクト）との関連

- ResourceSync Notification（ベータ版）
  - コアを拡張
  - プッシュ通知

特徴

- URIを持つあらゆるウェブリソースが対象　⇔　OAI-PMHの限界(1)

- Sitemapプロトコルベース　⇔　限界(2)
  - SEOのためにSitemapファイルを公開していれば多少の変更で対応可能
  - Sitemapファイルとしても使える

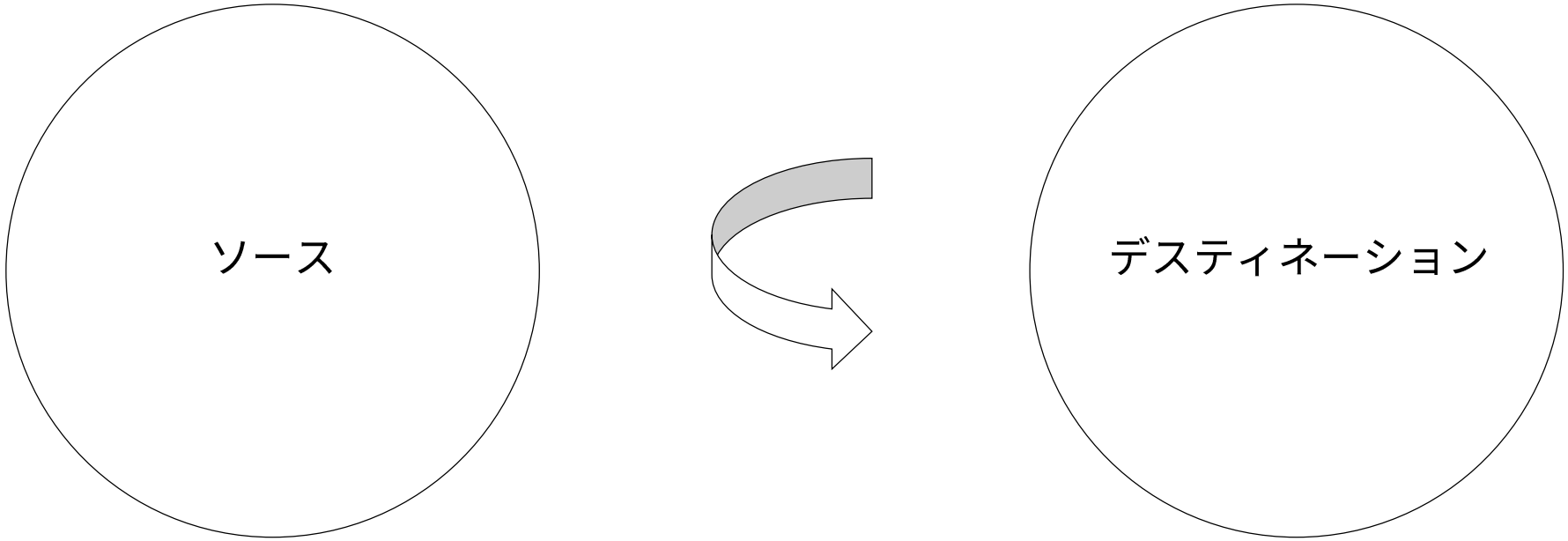
- プッシュ通知に対応　⇔　限界(3)
  - 同期のタイムラグを減らすことが可能になる

- スケーラビリティへの配慮
  - コンテンツ数が数百万件以上の大規模環境
  - 秒単位で頻繁に更新されるようなコンテンツ

OAI-PMH 2.0	ResourceSync
メタデータの収集	URIを持つあらゆるリソースの同期
標準規格ではない（デファクトスタンダード）	ANSI/NISO Z39.99-2014
2002年6月策定	2014年4月策定（コア）
独自プロトコル	Sitemapプロトコルベース
コンテンツプロバイダ、サービスプロバイダ	ソース、デスティネーション
Identify、ListMetadataFormats、ListSets、ListIdentifiers、ListRecords、GetRecord	Resource List、Change List、Resource Dump、Change Dump、Resource Dump Archive、Change List Archive、Change Dump Archive、など
ベースURL	Source Description、Capability List、Resource List
プル型	プル型（コア）・プッシュ型（ResourceSync Notification）

3.2. 同期のプロセス

①robots.txt　→　Resource Listなどにアクセス



②SitemapベースのXMLファイルを返す

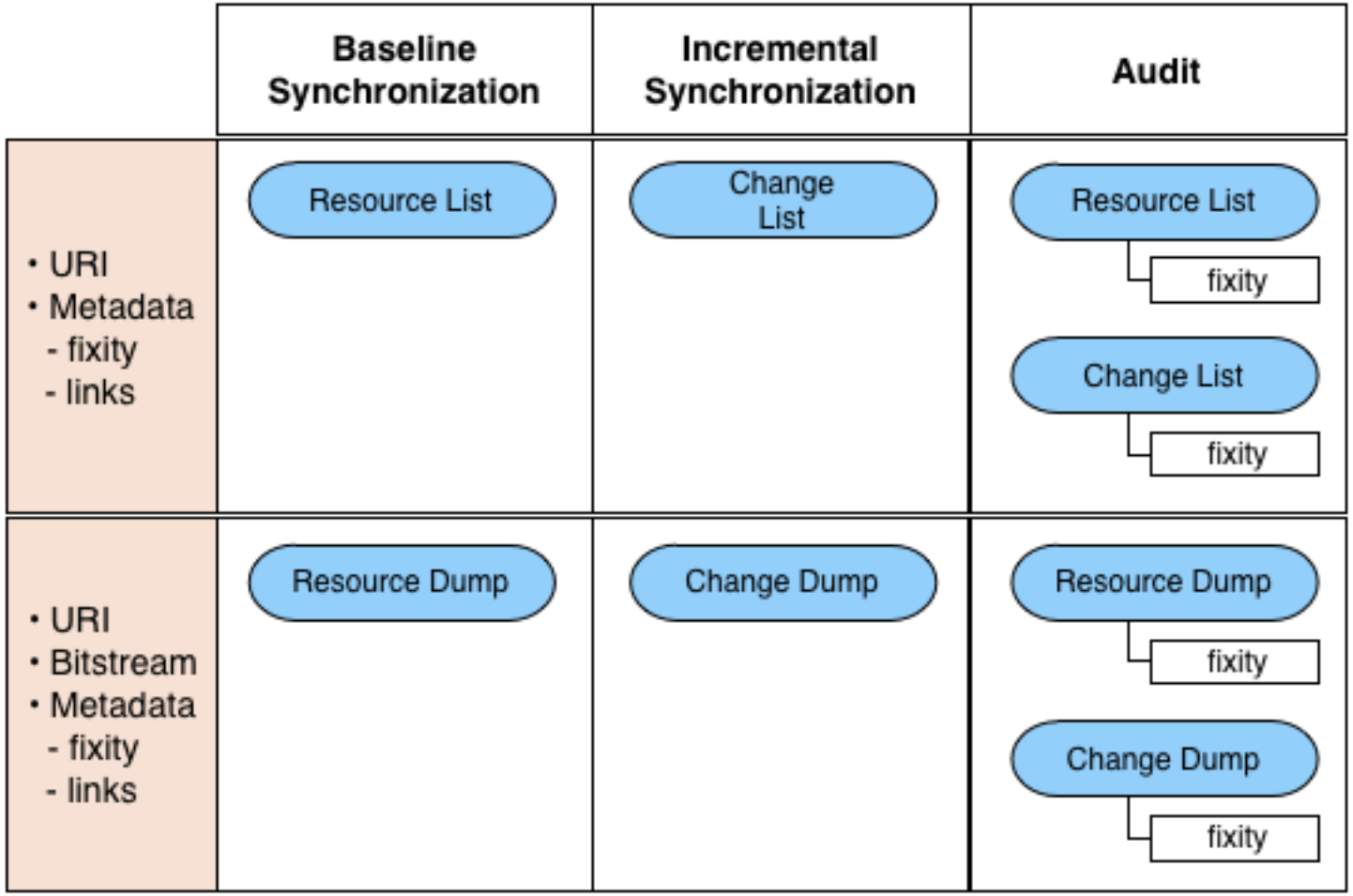
デスティネーション側のプロセス

- Baseline Synchronization（ベースライン同期）
  - 初回同期
  - 全件同期やり直し
- Incremental Synchronization（増分同期）
  - 差分更新
- Audit（監査）
  - 同期にミスがないか

ソース側のcapability（機能）

- Resource List
  - ある時点におけるすべてのリソースのURIの一覧
- Resource Dump
  - ある時点におけるすべてのリソースをパッケージングしたZIPファイルのURI
- Change List
  - ある期間に変化（追加・更新・削除）したリソースのURIの一覧
- Change Dump
  - ある期間に変化したリソースをパッケージングしたZIPファイルのURI

★それぞれSitemapベースのXMLファイル（を出力する機能）の名称



Source: Figure 3, http://www.openarchives.org/rs/1.0/resourcesync

capabilityの基本構造

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns="http://www.sitemaps.org/schemas/sitemap/0.9"
        xmlns:rs="http://www.openarchives.org/rs/terms/">
<rs:ln rel="up" href="http://ex.com/capabilitylist.xml"/>

    <rs:md capability="..." ... />          # capabilityで機能を指定する
    <url>
        <loc>http://...</loc>               # リソースのURI
        <lastmod>...</lastmod>              # リソースの最終更新日
        ...
    </url>
    <url>
        ......
    </url>
        ......

</urlset>
```

Resource List

```
<rs:md capability="resourcelist" at="2015-01-03T09:00:00Z"/>

<url>
        <loc>http://ex.com/content.pdf</loc>
        <lastmod>2015-01-02T13:00:00Z</lastmod>
        <rs:md hash="md5:1584abdf8ebdc9802ac0c6a7402c03b6"
                        type="application/pdf"/>
</url>

<url>
        ......
</url>
```

Resource Dump

```
<rs:md capability="resourcedump" at="2015-01-03T09:00:00Z"/>

<url>
        <loc>http://ex.com/package-1.zip</loc>
        <rs:md type="application/zip" length="4765"
                                at="2015-01-03T09:00:00Z"/>
        <rs:ln rel="contents" href="http://ex.com/manifest-1.xml" type="application/xml"/>
                                        # Resource Dump ManifestのURI
</url>

<url>
        ...
        # パッケージは分割可
</url>

...
```

Change List

```
<rs:md capability="changelist" from="2015-01-03T00:00:00Z"/>
<url>
        <loc>http://ex.com/content1.html</loc>
        <lastmod>2015-01-03T11:00:00Z</lastmod>
        <rs:md change="created"/>          # 新規作成
</url>
<url>
        <loc>http://ex.com/content2.pdf</loc>
        <lastmod>2015-01-03T13:00:00Z</lastmod>
        <rs:md change="updated"/>          # 更新
</url>
<url>
        <loc>http://ex.com/content3.tiff</loc>
        <lastmod>2015-01-03T18:00:00Z</lastmod>
        <rs:md change="deleted"/>          # 削除
</url>
...
```

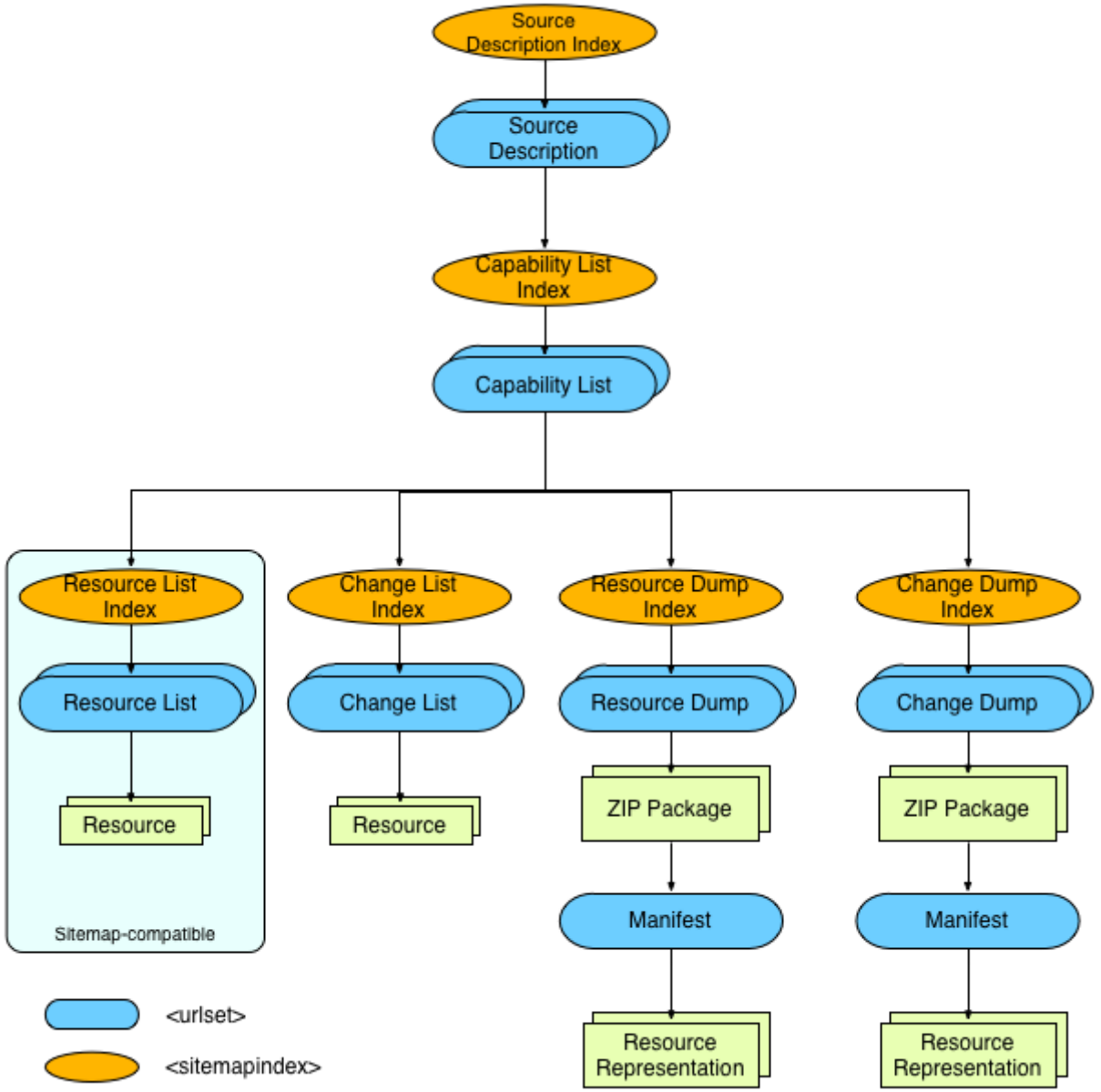
Change Dump

```
<rs:md capability="changedump” from="2015-01-01T00:00:00Z"/>
<url>
        <loc>http://ex.com/20150101-changedump.zip</loc>
        <lastmod>2015-01-01T23:59:59Z</lastmod>
        <rs:md type="application/zip” length="3109"
                        from="2015-01-01T00:00:00Z”
                        until="2015-01-02T00:00:00Z"/>
        <rs:ln rel="contents"
                        href="http://ex.com/20150101-changedump-manifest.xml"
                        type="application/xml"/>
                                # Change Dump Manifest の URI
</url>
<url>
        …
        # パッケージは分割可
</url>
…
```

ファイルの分割

- Sitemapプロトコルの制限
  - 1つのSitemapファイルのサイズは5万URLまたは10MBまで
  - 超える場合はファイルを分割し、Sitemap indexファイルからリンクする

- ResourceSyncもこの流儀に従う
  - Resource List　→　Resource List Index
  - Change List　→　Change List Index



Source: Figure 4, http://www.openarchives.org/rs/1.0/resourcesync

メタデータハーベスティング

- ResourceSyncはメタデータの収集にも対応できる
  - 一般のコンテンツと同じ方法
  - メタデータもURIを持ったひとつのリソースでなければならない

- また、コンテンツとメタデータの関係を記述する方法も用意されている（rel = describedby, describes）

```
<url>
        <loc>http://ex.com/content.pdf</loc>
        <rs:ln rel="describedby" href="http://ex.com/metadata.xml">
</url>
```

3.3. 実装例

実装実験

- Simeon Warner（コーネル大学図書館）
  - arXivのデータを用いてResource ListやChange Listを公開
  - http://resync.library.cornell.edu/

- Martin Klein et al.（ロスアラモス国立研究所）
  - 更新の激しいDBpedia Liveを題材にプッシュ通知の実験
  - http://arxiv.org/abs/1402.3305

- Herbert Van de Sompel（ロスアラモス国立研究所）
  - プッシュ通知を使ってリソース同期を行うデモ動画
  - https://www.youtube.com/watch?v=H2Le9_Bbkdw

ツール

- Cottage Labs
  - メタデータハーベスティング用のJavaライブラリとDSpaceモジュール
  - Resource DumpやChange Dumpによるコンテンツの同期はまだ
  - http://cottagelabs.com/news/resourcesync-module-for-dspace
- resync & ResourceSync Simulator
  - Pythonクライアント＆シミュレータ
  - https://pypi.python.org/pypi/resync/1.0.0
  - https://github.com/resync/resync-simulator
- ResourceSync PuSH
  - ResourceSync NotificationのPython実装
  - https://github.com/resync/resourcesync_push

CiNii Dissertations !!!!! (2015.6)

```
Disallow: /feedback
Disallow: /*/refworks
Disallow: /*/endnote
Disallow: /*.bix$
Disallow: /*.ris$
Disallow: /*.bib$
Disallow: /*.tsv$
Disallow: /*.txt$
Disallow: /d/search
Disallow: /d/openurl
Disallow: /d/link
Disallow: /d/search/export
Sitemap: http://ci.nii.ac.jp/sitemaps/sitemapindex.xml
Sitemap: http://ci.nii.ac.jp/sitemaps/sitemap_authorindex.xml
Sitemap: http://ci.nii.ac.jp/books/sitemap/sitemap_index.xml
Sitemap: http://ci.nii.ac.jp/d/sitemaps/resourcelist.xml
```

Source: http://ci.nii.ac.jp/robots.txt

4. まとめ

あるいは余談

サマリー

- 図書館等ではメタデータ収集（ハーベスティング）の方法（API）としてOAI-PMHがデファクトスタンダードになっている
- だが誕生から15年以上が経つOAI-PMHにはいくつかの限界がある（メタデータのみ、非リアルタイム、マイナー・旧式）
- そこでResourceSyncというコンテンツ同期のための標準規格群の策定が進められている
- ResourceSyncの実装例はまだ少ない（CiNii Dissertationsくらい？）

ResourceSyncの可能性

- OAI-PMH（メタデータハーベスティング）の置き換え
  - NDLはしばらくOAI-PMH推し……？
- デジタルアーカイブにおけるコンテンツ収集
  - DPLAの求人
    http://dp.la/info/2015/10/22/job-opportunity-developer-ingestion-and-operations/
- 実運用サービスでの採用
  - Project Next-L Enju？
- NACSIS-CAT 2020？

Herbert Van de Sompel

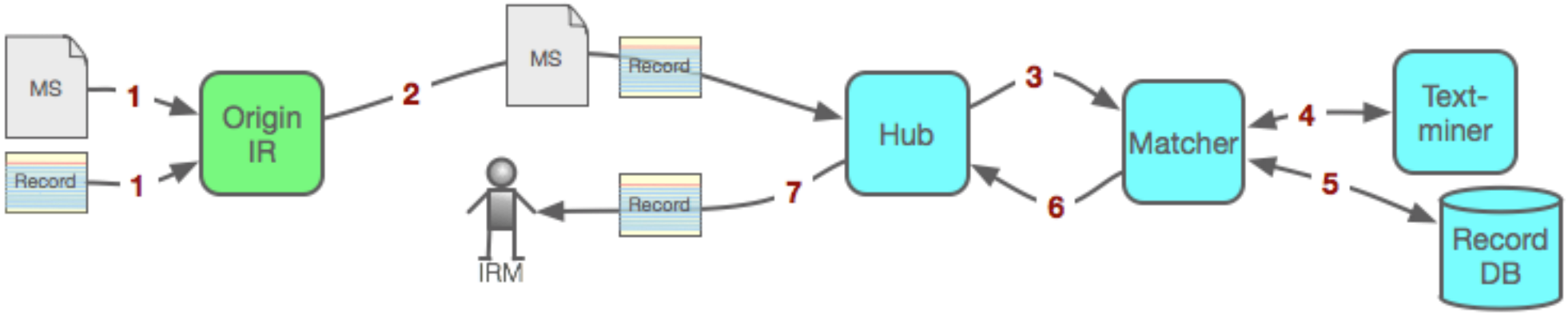
- 米国ロスアラモス国立研究所（LANL）

- リンクリゾルバの生みの親

- Reminiscing About 15 Years of Interoperability Efforts (D-Lib Magazine)
  - OAI-PMH (1999)　→　ResourceSync (2014)
  - OAI-ORE (2006)
  - Memento (2009)
  - http://www.dlib.org/dlib/november15/vandesompel/11vandesompel.html

Cooperative Open Access Exchange

- Paul Walk (EDINA) の提案
  - http://www.paulwalk.net/2015/10/19/the-active-repository-pattern/
  - http://www.paulwalk.net/2015/10/20/cooperative-open-access-exchange-coax/
- 共著論文を複数のリポジトリで登録する（router / broker）



References

- OAI-PMHについて
  - http://www.openarchives.org/OAI/openarchivesprotocol.html
  - https://www.nii.ac.jp/irp/archive/translation/oai-pmh2.0/
  - http://current.ndl.go.jp/ca1513
  - http://iss.ndl.go.jp/information/2015/10/02_announce_1/

- ResourceSyncについて
  - http://www.openarchives.org/rs/toc
  - http://current.ndl.go.jp/ca1845
  - http://b.hatena.ne.jp/kitone/ResourceSync