# Métadonnées de la numérisation et Linked Data

*Analyse des best practices*

Mars 2015

## Sommaire

**1 Introduction....2**
**2 Le workflow de numérisation....2**
**3 Les métadonnées....3**
3.1 Formats « traditionnels »....3
3.1.1 Dublin Core....3
3.1.2 METS....3
3.1.3 MODS....3
3.2 Formats basés RDF....4
**4 Un exemple : le site Chronicling America....4**
4.1 Les données RDF....4
4.2 Le système d'identifiants....7
**5 Synthèse....8**
**6 Bibliographie....9**

# 1 Introduction

Ce document donne un bref aperçu des types et formats de métadonnées générés lors de la numérisation de documents, et utilisés ensuite sur les interfaces en ligne. Le but est de faire le point sur la situation afin de connaître les enjeux d'une transformation en RDF dans ce contexte précis.

# 2 Le workflow de numérisation

La numérisation est un processus complet, qui comprend plus d'étape que la simple génération de métadonnées sur les objets numériques. Il est constitué des étapes suivantes (1) :

1. Scan d'un document
2. Renommage des images, selon une politique prédéterminée (récupération ou non de l'identifiant du catalogue, etc.)
3. Conversion des images produites (pour l'affichage web, pour l'océrisation[1])
4. Océrisation (reconnaissance de texte) des images
5. Traitement des images recueillies (rognage, etc.)
6. Lien avec un système d'informations préexistant (catalogue de bibliothèque, autres bases de données)
7. Importation des données concernées, transformation et gestion des formats souhaités : Dublin Core, METS, MODS.
8. Saisies des données structurelles propres au document (pour chaque article, pour chaque chapitre, etc. selon la précision de description souhaitée)
9. Contrôle qualité
10. Mise à disposition

L'ordre de ces étapes n'est pas fixe. Il existe des logiciels spécialisés dans la gestion du processus de numérisation, mettant à disposition de manière intégrée les divers outils nécessaires, ou les interfaces vers les outils externes souhaités. Ces outils consistent entre autres en :

- convertisseur d'images
- logiciel OCR
- éditeur XML pour les données structurelles
- interface SRU/Z39.50 pour l'import de données
- logiciel de renommage de fichiers

Le logiciel Agora, utilisé à la bibliothèque de l'ETH Zurich, en est un exemple.

1 Le mot *océrisation* est un dérivé de l'acronyme OCR, signifiant Optical Character Recognition. Il s'agit d'un processus dans lequel une image est analyse de manière automatique afin d'en extraire le texte qu'elle contient.

# 3 Les métadonnées

Quatre types de métadonnées interviennent dans les projets de numérisation (1) :

1. Métadonnées bibliographiques ou descriptives : description du document.
2. Métadonnées structurelles : structure interne du document.
3. Métadonnées administratives : droits d'utilisation, etc.
4. Métadonnées techniques : type de fichier, la taille du fichier, la résolution, etc.

## *3.1 Formats « traditionnels »*

### 3.1.1 Dublin Core

Dublin Core est un ensemble d'éléments simples. Il est indépendant de tout format d'enregistrement. Dublin Core peut donc être utilisé au sein de pages HTML, de fichiers RDF ou de documents XML.

Reconnu et utilisé intensivement dans le monde entier, Dublin Core est devenu un standard de facto dans la gestion des métadonnées sur le web.

Il est constitué d'une part des éléments de base, au nombre de quinze, qui suffisent à élaborer une description simple d'un document. Un nouveau set d'éléments a été créé en complément, les « terms », permettant des descriptions plus détaillées[2].

### 3.1.2 METS

METS (Metadata Encoding and Transmission Standard) est un format prévu pour l'échange et la gestion de documents numériques. Il se base sur un schéma XML. Souple et adaptable, METS permet d'exprimer en détails toute information liée à la structure interne des documents (2).

Un fichier METS correspond à un objet numérique (qui peut lui-même être composé de plusieurs fichiers). Dans ce document sont intégrés et liés les images de l'objet numérique ainsi que son contenu textuel. Le fichier est donc à voir comme un container de métadonnées, pointant vers divers fichiers liés.

### 3.1.3 MODS

MODS (Metadata Object Description Schema) est un format prévu pour le traitement des données bibliographiques. Basé sur un schéma XML, il permet de décrire tout type de document, mais en particulier les documents MARC21, pour lesquels il a été conçu à l'origine. Il s'agit d'un dérivé de MARC, en plus simple.

> *« Il est particulièrement intéressant dans le cadre de projets de description de documents numérisés car il contient des éléments permettant de renseigner les données relatives à la description d'une version numérisée d'un document. Il peut notamment constituer l'un des composants (volet description bibliographique) des métadonnées de ressources numérisées pour lesquelles le schéma METS est utilisé. »* Bibliothèque nationale de France (3)

MODS peut donc se voir comme une extension de METS, orientée données bibliographiques de bibliothèque.

L'équivalent de MODS pour les données d'autorité est MADS.

---

2 Les éléments et « terms » Dublin Core sont décrits sur cette page : http://dublincore.org/documents/2012/06/14/dcmi-terms/ (consultée le 18 mars 2015)

## 3.2 Formats basés RDF

Le modèle RDF peut s'exprimer au moyen de divers format de sérialisation, compatibles les unes avec les autres.

Les données doivent néanmoins remplir certaines conditions pour pouvoir être représentées selon le modèle RDF (qu'importe la sérialisation). Elles doivent être hautement structurées et les ressources décrites (documents, personnes, etc.) doivent posséder des identifiants unique appelés IRI[3]. Ces IRI doivent être pérennes et doivent pouvoir être déréférencées.

# 4 Un exemple : le site Chronicling America

Pour illustrer concrètement la mise en application du modèle RDF pour des données de documents numériques, l'exemple du site Chronicling America[4] a été choisi.

Chronicling America propose une interface donnant accès à toute la presse américaine publiée depuis l'an 1690 à nos jours. En outre, il met à disposition des éditions digitalisées de titres parus entre 1836-1922. Le site est développé et maintenu par la LOC, en partenariat avec le *National Endowment for the Humanities*. Les données sont mises à disposition en divers formats, dont le format RDF/XML. Il s'agit d'un exemple d'implémentation Linked Open Data en bibliothèque digitale qui est concret, cohérent et qui prend en compte la complexité de la structure informationnelle des publications en série.

## 4.1 Les données RDF

Pour chaque revue, chaque numéro et chaque image (scan), des représentations RDF sont disponibles. Elles sont saisies au moyen des ontologies *Dublin Core (DC)*, *Bibliographic Ontology (BIBO)* et *Object Reuse and Exchange (ORE)*[5].

Un exemple concret permet d'illustrer ce système et de le comprendre en détail : la revue *Durant weekly news*. Ce journal se décline, sur le site, en une notice générale, en plusieurs notices pour chacun de ses numéros parus, ainsi qu'en une notice pour chaque page des numéros.

La notice de périodique (notice générale) se présente ainsi (illustration 1, la notice a été abrégée pour des raisons de lisibilité).

3 IRI est l'abréviation de *Internationalized Resource Identifier*.
4 http://chroniclingamerica.loc.gov/ (consulté le 3 septembre 2014)
5 Une ontologie développée et maintenue par l'Open Archives Initiative

*Illustration 1: Métadonnées RDF d'un périodique*

```
<rdf:Description rdf:about="/lccn/sn96093062#title">
    <rdfs:seeAlso
rdf:resource="http://www.loc.gov/chroniclingamerica/lccn/sn96093062/issues"/>
    <rdfs:seeAlso rdf:resource="http://lccn.loc.gov/sn96093062"/>
    <rdfs:seeAlso rdf:resource="http://www.worldcat.org/oclc/35709904"/>
    <dc:publisher>Lewis Paullin</dc:publisher>
    <dcterms:date
rdf:datatype="http://www.loc.gov/standards/datetime#edt">190?/1944</dcterms:date>
    <frbr:successor rdf:resource="/lccn/sn96093065#title"/>
    <frbr:successor rdf:resource="/lccn/sn96093061#title"/>
     <rda:placeOfPublication>Durant, Choctaw Nation, Indian Territory
[Okla.]</rda:placeOfPublication>
    <dcterms:title>Durant weekly news.</dcterms:title>
    <ore:isDescribedBy rdf:resource="/lccn/sn96093062.rdf"/>
    <dcterms:description rdf:resource="/essays/604/"/>
    <dcterms:relation rdf:resource="/lccn/sn96093064#title"/>
    <dcterms:coverage rdf:resource="http://dbpedia.org/resource/Durant%2C_Oklahoma"/>
    <dc:subject>Durant (Okla.)--Newspapers.</dc:subject>
    <dc:subject>Choctaw Nation of Oklahoma--Newspapers.</dc:subject>
    <dc:subject>Bryan County (Okla.)--Newspapers.</dc:subject>
    <dcterms:hasFormat rdf:resource="/lccn/sn96093062/marc.xml"/>
    <owl:sameAs rdf:resource="info:oclcnum/35709904"/>
    <owl:sameAs rdf:resource="info:lccn/sn96093062"/>
    <owl:sameAs rdf:resource="urn:issn:2163-8012"/>
    <dcterms:language rdf:resource="http://www.lingvoj.org/lang/en"/>
    <rdf:type rdf:resource="http://purl.org/ontology/bibo/Newspaper"/>
    <ore:aggregates rdf:resource="/lccn/sn96093062/1918-02-08/ed-1#issue"/>
    <ore:aggregates rdf:resource="/lccn/sn96093062/1918-04-26/ed-1#issue"/>
</rdf:Description>
```

Les particularités de cette dernière sont les suivantes :

- classe RDF : bibo:Newspaper
- liens avec une propriété FRBR vers les titres précédents et suivant du périodique
- liens avec la propriété *ore:aggregates* vers les notices des numéros
- lien avec la propriété *rdfs:seeAlso* vers la notice du périodique sur la catalogue de la LOC.

Une notice de numéro se présente ainsi (illustration 2).

*Illustration 2: Métadonnées d'un numéro de périodique*

```
<rdf:Description rdf:about="/lccn/sn96093062/1905-05-19/ed-1#issue">
    <dcterms:title>Durant weekly news. - 1905-05-19</dcterms:title>
    <dcterms:issued rdf:datatype="http://www.w3.org/2001/XMLSchema#date">1905-05-
19</dcterms:issued>
    <ore:aggregates rdf:resource="/lccn/sn96093062/1905-05-19/ed-1/seq-2#page"/>
    <ore:aggregates rdf:resource="/lccn/sn96093062/1905-05-19/ed-1/seq-7#page"/>
    <ore:aggregates rdf:resource="/lccn/sn96093062/1905-05-19/ed-1/seq-3#page"/>
    <ore:aggregates rdf:resource="/lccn/sn96093062/1905-05-19/ed-1/seq-6#page"/>
    <ore:aggregates rdf:resource="/lccn/sn96093062/1905-05-19/ed-1/seq-5#page"/>
    <ore:aggregates rdf:resource="/lccn/sn96093062/1905-05-19/ed-1/seq-1#page"/>
    <ore:aggregates rdf:resource="/lccn/sn96093062/1905-05-19/ed-1/seq-8#page"/>
    <ore:aggregates rdf:resource="/lccn/sn96093062/1905-05-19/ed-1/seq-4#page"/>
    <ore:isDescribedBy rdf:resource="/lccn/sn96093062/1905-05-19/ed-1.rdf"/>
    <rdf:type rdf:resource="http://purl.org/ontology/bibo/Issue"/>
    <ore:isAggregatedBy rdf:resource="/lccn/sn96093062#title"/>
    <ore:isAggregatedBy rdf:resource="/batches/batch_okhi_greer_ver01#batch"/>
</rdf:Description>
```

Les particularités de cette dernière sont les suivantes :

- classe RDF : bibo:Issue
- la date du numéro apparaît dans la propriété du titre
- lien avec la propriété *ore:isAggregatedBy* vers la notice du périodique
- liens avec la propriété *ore:aggregates* vers les pages constituant le numéro

Une notice de « page » se présente ainsi (illustration 3).

*Illustration 3: Métadonnées d'une page de périodique*

```
<rdf:Description rdf:about="/lccn/sn96093062/1905-05-19/ed-1/seq-7#page">
    <dcterms:title>Durant weekly news. - 1905-05-19 - 7</dcterms:title>
    <foaf:depiction rdf:resource="/lccn/sn96093062/1905-05-19/ed-1/seq-
7/thumbnail.jpg"/>
    <ndnp:sequence
rdf:datatype="http://www.w3.org/2001/XMLSchema#integer">7</ndnp:sequence>
    <ore:aggregates rdf:resource="/lccn/sn96093062/1905-05-19/ed-1/seq-7.pdf"/>
    <ore:aggregates rdf:resource="/lccn/sn96093062/1905-05-19/ed-1/seq-7/ocr.xml"/>
    <ore:aggregates rdf:resource="/lccn/sn96093062/1905-05-19/ed-1/seq-7.jp2"/>
    <ore:aggregates rdf:resource="/lccn/sn96093062/1905-05-19/ed-1/seq-
7/thumbnail.jpg"/>
    <ore:aggregates rdf:resource="/lccn/sn96093062/1905-05-19/ed-1/seq-7/ocr.txt"/>
    <ore:isAggregatedBy rdf:resource="/lccn/sn96093062/1905-05-19/ed-1#issue"/>
    <dcterms:issued rdf:datatype="http://www.w3.org/2001/XMLSchema#date">1905-05-
19</dcterms:issued>
    <ore:isDescribedBy rdf:resource="/lccn/sn96093062/1905-05-19/ed-1/seq-7.rdf"/>
    <rdf:type rdf:resource="http://chroniclingamerica.loc.gov/terms#Page"/>
</rdf:Description>
```

Les particularités de cette dernière sont les suivantes :

- classe RDF issue d'un vocabulaire développé spécifiquement pour ce site web : http://chroniclingamerica.loc.gov/terms#Page
- lien avec la propriété *ore:isAggregatedBy* vers la notice du numéro
- lien avec la propriété *ore:aggregates* vers le texte brut (ocr.txt), une vignette pour l'affichage en ligne (thumbnail.jpg), le ficher PDF (.pdf), vers le fichier image en format JPEG-2000 (.jp2)

- numéro de la page informé au moyen de la propriété *ndnp:sequence*[6].
- la date du numéro et le numéro de la page apparaissent dans la propriété du titre

## 4.2 Le système d'identifiants

Comme mentionné dans le chapitre 3.2, la particularité du modèle RDF impose aux ressources de posséder des identifiants pérennes et déréférençables. Le choix de tels identifiants et surtout de leur structure nécessite une certaine réflexion, en particulier dans le cas des publications en série, qui sont des ressources relativement hierarchisées (revue / numéro / page).

Les identifiants pour les revues, numéros et pages sont étroitement liés. L'IRI d'une page se base sur l'IRI du numéro, lui-même construit à partir de l'IRI de la revue (voir tableau 1).

Des suffixes sont ajoutés aux IRIs pour en préciser le format. L'IRI de l'objet du monde réel porte ainsi le suffixe *#title* (pour une revue), *#issue* (pour un numéro) et *#page* (pour une page). Les IRIs des divers formats sont :

- / pour la page web en HTML
- .rdf pour la représentation RDF
- .json pour la représentation JSON
- /marc.xml pour une représentation en MARC/XML
- /ocr.txt pour le texte brut
- .jp2 pour l'image
- .pdf pour le document PDF
- /thumbnail.jpg pour la vignette
- /ocr.xml pour le document ALTO (contient le texte brut avec indications de mise en page)

*Tableau 1: Identifiants utilisés pour le site Chronicling America*

| Contenu | IRI (identifiant) |
|---|---|
| Revue | http://chroniclingamerica.loc.gov/lccn/sn96093062#title |
| Revue (page HTML) | http://chroniclingamerica.loc.gov/lccn/sn96093062/ |
| Revue (représentation RDF) | http://chroniclingamerica.loc.gov/lccn/sn96093062.rdf |
| Revue (représentation JSON) | http://chroniclingamerica.loc.gov/lccn/sn96093062.json |
| Revue (représentation MARC/XML) | http://chroniclingamerica.loc.gov/lccn/sn96093062/marc.xml |
| Numéro | http://chroniclingamerica.loc.gov/lccn/sn96093062/1905-05-19/ed-1#issue |
| Numéro (page HTML) | http://chroniclingamerica.loc.gov/lccn/sn96093062/1905-05-19/ed-1/ |
| Numéro (représentation RDF) | http://chroniclingamerica.loc.gov/lccn/sn96093062/1905-05-19/ed-1.rdf |
| Numéro (représentation JSON) | http://chroniclingamerica.loc.gov/lccn/sn96093062/1905-05-19/ed-1.json |
| Page | http://chroniclingamerica.loc.gov/lccn/sn96093062/1905-05-19/ed-1/seq-7#page |
| Page (page HTML) | http://chroniclingamerica.loc.gov/lccn/sn96093062/1905-05-19/ed-1/seq-7 |
| Page (représentation RDF) | http://chroniclingamerica.loc.gov/lccn/sn96093062/1905-05-19/ed-1/seq-7.rdf |

6 Il s'agit d'une propriété venant d'une ontologie développée à l'interne par Chronicling America. Label de la propriété : « Page Sequence ». Définition : « A sequence number uniquely identifying a newspaper page on a microfilm reel. »

| | |
|---|---|
| Page (représentation JSON) | http://chroniclingamerica.loc.gov/lccn/sn96093062/1905-05-19/ed-1/seq-7.json |
| Page (texte brut) | http://chroniclingamerica.loc.gov/lccn/sn96093062/1905-05-19/ed-1/seq-7/ocr.txt |
| Page (image JPEG-2000) | http://chroniclingamerica.loc.gov/lccn/sn96093062/1905-05-19/ed-1/seq-7.jp2 |
| Page (fichier PDF) | http://chroniclingamerica.loc.gov/lccn/sn96093062/1905-05-19/ed-1/seq-7.pdf |
| Page (vignette) | http://chroniclingamerica.loc.gov/lccn/sn96093062/1905-05-19/ed-1/seq-7/thumbnail.jpg |
| Page (document ALTO) | http://chroniclingamerica.loc.gov/lccn/sn96093062/1905-05-19/ed-1/seq-7/ocr.xml |

# 5 Synthèse

Lors de la conversion de données issues de la numérisation en RDF, plusieurs points nécessitent une attention particulière en comparaison avec les données bibliographiques traditionnelles des bibliothèques (les métadonnées des catalogues en ligne). Il faut notamment concevoir un modèle suffisamment précis qui permet de distinguer une revue, un numéro et une page (si l'on souhaite fournir des données sur chaque page individuellement). Les métadonnées doivent être clairement distinguées des ressources numériques à proprement parler (les fichiers PDF consultables), au moyen des classes RDF adéquates. En outre, pour assurer les liens entre les diverses ressources, des IRI cohérentes doivent être choisies.

# 6 Bibliographie

1. ETH ZÜRICH. BIBLIOTHEK. Meilleures pratiques en matière de numérisation. *ETH Zürich* [en ligne]. 2014. [Consulté le 3 septembre 2014]. Disponible à l'adresse : http://www.digitalisierung.ethz.ch/index_f.html

2. BNF. METS : Metadata Encoding and Transmission Standard. *Bibliothèque nationale de France* [en ligne]. 21 novembre 2013. [Consulté le 3 septembre 2014]. Disponible à l'adresse : http://www.bnf.fr/fr/professionnels/formats_catalogage/a.f_mets.html

3. BNF. MODS : Metadata Object Description Schema. *Bibliothèque nationale de France* [en ligne]. 17 décembre 2013. [Consulté le 3 septembre 2014]. Disponible à l'adresse : http://www.bnf.fr/fr/professionnels/f_mods/s.mods_presentation.html