放送コンテンツアーカイブのためのメタデータモデル構築

萩原　和樹†　　三原　鉄也†　　永森　光晴‡　　杉本　重雄‡

筑波大学大学院図書館情報メディア研究科†

筑波大学図書館情報メディア系‡

〒305-0821 茨城県つくば市春日 1-2

概要

放送コンテンツを保存するデジタルアーカイブが増えていく一方、検索や利用の観点からアーカイブ同士の連携が求められる。しかし、保存方針の統一化が未整備のため、放送コンテンツアーカイブは同じ放送局であっても部署毎に保存や公開方法が異なっており、放送コンテンツアーカイブの利用の妨げの要因となっている。現状の放送コンテンツアーカイブの利用性を高めるという本研究では、放送番組の情報や DBpedia、NDLSH といった LOD リソースを利用し、異なるアーカイブ同士の連携を図るメタデータモデルの提案を行う。

キーワード

デジタルアーカイブ, Linked Open Data, DBpedia, NDLSH, OAI-ORE

The Construction of Metadata Model for Broadcasting Collection Archives

Kazuki Hagiwara†　Tetsuya Mihara†　Mitsuharu Nagamori‡　Shigeo Sugimoto‡

Graduate School of Library, Information and Media Studies, University of Tsukuba†

Faculty of Library, Information and Media Science, University of Tsukuba‡

1-2 Kasuga, Tsukuba, Ibaraki, 305-0821, Japan†

Abstract

Digital Archive is a system to keep digital collections safe and accessible over time. Linking digital archives is a crucial issue for users to enhance the usability of digital archives and their contents. Broadcasting stations are building their video program archives as a part of their new business. However, because of immaturity of policies and systems of video program archives the video archives at broadcasting stations are heterogeneous and it is not easy to build services across archives. This study aims to build an access service to the archived video programs across archives using Linked Open Data Technology. In this paper, we propose a metadata model to link video archives using the Web sites provided by NHK.

Keywords

Digital Archive, Linked Open Data, DBpedia, NDLSH, OAI-ORE

1. はじめに

近年、テレビ放送された番組の映像・音声を保存してアーカイブとして公開する試みが国内外でなされている。日本では日本放送協会（NHK）が 2003 年から番組やアーカイブの整備に取り組んでいる。

テレビで放送された映像や音声を公開しているデジタルアーカイブが増えていく一方、アーカイブ同士の連携が求められている。デジタルアーカイブの連携とは、複数のアーカイブから何らかの関連の持つ所蔵資料のメタデータを収集し、単一のサービスとして利用できるようにすることである。複数のアーカイブを連携させる上で課題となるのは、記述規則の異なるメタデータの統一である。既存のアーカイブの記述規則を共通化することは現実的ではなく、個別のアーカイブの要件に則した記述を無視することも適切ではない。そこで本研究では、個別のアーカイブの記述規則を保持しつつ、コンテンツ同士の関連性を持たせて、統合的に検索することを目的としたメタデータモデルの構築を行う。さらに NHK デジタルアーカイブスを対象として、Web 上の情報資源からメタデータモデルの適用と Linked Open Data（LOD）データセットの作成を進めている。

メタデータモデルの構築にあたって、ジャンルの異なる複数のアーカイブの関係を記述するために Web リソースの集合体を表現するモデルである Open Archives Initiative Object Reuse and Exchange（OAI-ORE）モデルを利用する。また、Wikipedia の情報を LOD として公開している DBpedia や国立国会図書館が作成した典拠データである国立国会図書館件名標目表（NDLSH）といった LOD リソースを利用することでコンテンツに適切な主題を付与して、意味的な検索を行えるようにする。たとえば、利用者が「日本の山」について知りたいとき、「山」というキーワードで全文検索を行うと人名に山が付いたコンテンツまで検索されてしまうことがある。複数のアーカイブの主題情報を LOD リソースで統一することで、利用者が知りたい情報と関連の深い情報を統合的に検索することができる。

2. 放送コンテンツアーカイブとその連携

2.1 デジタルアーカイブと放送コンテンツ

デジタルアーカイブとは「図書・出版物、公文書、美術品・博物品、歴史資料等公共的な知的資産の総デジタル化を進め、インターネット上で電子情報として共有・利用できる仕組み」である[1]。デジタルアーカイブが増えていく一方、それらのアーカイブを連携する必要性が出てくる。デジタルアーカイブの連携とは、コンテンツの内容や管理方法、提供方法が異なるアーカイブ同士のメタデータを収集し、1 つのサービスとして利用できるようにすることである。アーカイブの連携によって、1 つのサービスから複数のアーカイブに

対して検索、利用できるようになることが求められる。2012 年に総務省が公開した「デジタルアーカイブの構築・連携のためのガイドライン[2]」によるとデジタルアーカイブを連携させることで、統合的な検索を実現し、デジタル資料の流通性を高めることで、知的資産の保存が図れるのみならず、新しい利用方法を生み出す効果があると述べている。

デジタルアーカイブにおける統合検索の先行事例として、ヨーロッパの図書館、博物館、美術館、文書館等が持つデジタル資料を連携して提供する Europeana[3]がある。Europeana では絵画や書籍、映画といった異なるメディアのコンテンツを独自のメタデータモデルである Europeana Data Model（EDM）で管理・統一している。また、EDM は複数のアーカイブが公開している異なるコンテンツの統合検索を可能としている。

日本の最近の事例として、2013 年に公開された国立国会図書館東日本大震災アーカイブ（通称:ひなぎく）[4]がある。ひなぎくは東日本大震災に関する音声・動画・写真・ウェブ情報等のデジタルデータや、関連する文献情報を統合的に検索できるポータルサイトである。2014 年 10 月 15 日時点で 39 のデータベースを対象とした横断検索が可能であり、各デジタル資料のメタデータは国立国会図書館ダブリンコアメタデータ記述[5]をもとに作成され、統一されている。

本研究では、デジタルアーカイブの中でもテレビやラジオで放送された映像や音声（放送コンテンツ）を公開しているものを放送コンテンツアーカイブとして定義する。放送コンテンツアーカイブの事例は、1936 年から英国放送協会（BBC）とフランスの国立視聴覚研究所（INA）が挙げられる。BBC は全放送番組を対象に保存しており、その量は 65 万時間分のテレビ番組、35 万時間分のラジオ番組に上る。INA では 2006 年から所有する映像資料、約 10 万番組、1 万時間に及ぶコンテンツがインターネットを通じて自由に閲覧することが可能となった。[11]

日本でも 2003 年から NHK が保存活動を始めており、NHK アーカイブスと呼ばれる組織で NHK のテレビ・ビデオ番組等の映像・音声の保存を行っている。NHK アーカイブスがインターネットを通じて公開している NHK デジタルアーカイブス[6]では約 1 万本の映像がテーマ別に配信されている。NHK では独自に放送コンテンツを収集し公開している。しかし、日本における放送コンテンツの保存に関する統一的な制度は決まっておらず、放送局の部署毎に Web サイトで放送コンテンツを公開するに留まっている。そのため、放送コンテンツの公開方式やメタデータがアーカイブ毎に異なり、これらを統合的に検索するには難しいという問題がある。

2.2 放送コンテンツアーカイブ連携のための要件分析

本節では、連携実現のために放送コンテンツのメタデータモデルの要求要件について述

べる。

放送コンテンツアーカイブの統合にあたって、各アーカイブの基本的なメタデータはそのまま保持したままの連携することが求められる。これは、他のアーカイブと連携しやすいように元のデータの変更を行うことで本来のアーカイブで付与されたメタデータの欠落や項目の意味の解釈が変わる恐れが出てくるためである。したがって、元からあるメタデータの記述項目はそのままに、他のアーカイブと連携するためのメタデータを新たに記述する必要がある。

次の要件はアーカイブが持つコンテンツのリンクを行う主題に関する情報を紐付けることである。放送コンテンツには検索や利用のために付与されている元のメタデータがあるが、その記述規則はそれぞれ異なるため、複数のアーカイブで統一された主題情報が必要となる。

最後の要件は放送コンテンツが持つ番組情報を利用することである。現在放送コンテンツアーカイブはテレビ・ラジオで放送された番組の映像・音声もしくはその一部を収録しているものが多い。この番組情報を用いることで、同じ番組や異なった番組で収録された放送コンテンツの中で「似通ったジャンルの番組の映像」や「同時期に放送された番組の映像」といった連携を可能にすることができる。

3. 放送コンテンツアーカイブ連携のためのメタデータモデル

3.1 OAI-ORE モデル

本研究のモデルの構築にあたってはEDMで採用されているOAI-OREモデル[9]を利用する。OAI-OREモデルはWeb上のリソースの集まりを集合体と呼ばれるオブジェクトで表現するモデルで、図1はOAI-OREモデルの構造を示したものである。デジタルアーカイブは複数のコンテンツを集積した集合体であり、そのコンテンツにはアーカイブ独自のメタデータが付与されている。本研究では、独自のメタデータと連携を行うためのメタデータを別々に記述するために集合体とプロキシの概念を用いている。

集合体（Aggregation）とはore:Aggregation型のリソースであり、他のリソースの集合である。主語にAggregation型のリソースを持ち、述語となるore:aggregatesの目的語となるリソースは集合リソース（Aggregated Resource）と表現される。OREモデルでは集合体を1つのリソースとみなしてメタデータを記述することができる。OREモデルはプロキシ（Proxy）と呼ばれるore:Proxy型のリソースを用いて集合体毎に異なるメタデータを記述する。例えば、論文1はそれぞれ集合体であるWebサイトAとWebサイトBで公開されており、それぞれのサイトで異なるメタデータが記述されている。WebサイトAでは論文1に対するメタデータをプロキシAに記述し、WebサイトBはプロキシBに記述す

る。これにより、同一リソースに対するメタデータを集合体毎に記述することができる。

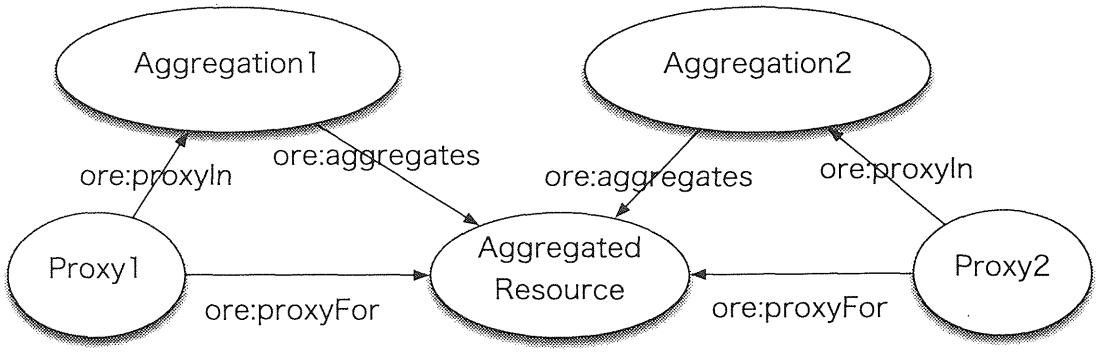

図 1. ORE モデルの基本的な構造

3.2 プロキシの概念を用いたアーカイブ間の連携手法

提案するメタデータモデルでは放送コンテンツアーカイブを集合体、アーカイブに収蔵されている放送コンテンツを集合リソースとして記述する。本研究では、連携のための概念として、複数のアーカイブのコンテンツを全て収集したものを統合アーカイブスとして集合体と定義する。2.2 節で述べた要件を満たすために ORE モデルの概念であるプロキシを用いる。放送コンテンツに対して、各アーカイブがプロキシを通じてメタデータを記述する一方、統合アーカイブス側でも別のプロキシを通じてメタデータを記述する。

図 2 は統合アーカイブスとそのうちの 1 つであるアーカイブ A とそのコンテンツの関係を記述したものである。集合体であるアーカイブ A が所蔵している 1 つの放送コンテンツは統合アーカイブスにも属していると表現している。A におけるメタデータは ProxyA リソースを主語として記述する。統合アーカイブスの ProxyT リソースにはアーカイブの連携のためのメタデータである番組情報や主題情報について記述する。番組情報は映像の収録元であるテレビ番組の情報を使用する。異なるアーカイブであっても放送コンテンツに共通する、ジャンルや放送期間といった番組情報を用いることで関連するコンテンツのリンクを作成する。主題情報はコンテンツのタイトルや説明文から抽出した主題情報をアーカイブに共通する語彙として、LOD リソースでもある DBpedia[7]と国立国会図書館件名標目表（NDLSH）[8]とリンクしたものを使用する。DBpedia は Wikipedia の情報を LOD 化したものであり、NDLSH と同様に汎用性のある辞書として多くの単語データを含んでいる。放送コンテンツに対して LOD リソースと関連のある単語を主題として付与することで、放送コンテンツ間の主題によるリンクを可能にする。

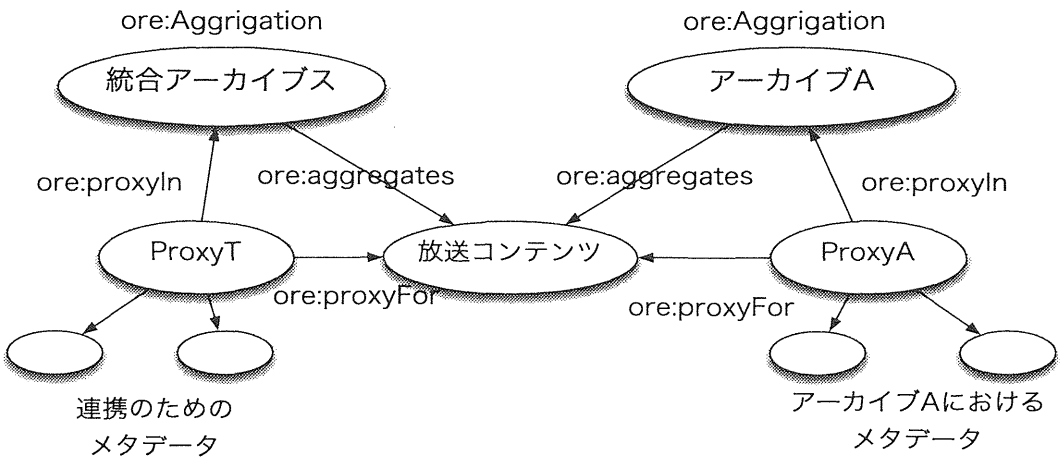

図 2. 統合アーカイブスとプロキシの関係

4. NHK デジタルアーカイブスへのメタデータモデルの適用事例

4.1 NHK デジタルアーカイブス

本研究で提案するメタデータモデルを、NHK デジタルアーカイブスを構成するアーカイブに適用する。NHK デジタルアーカイブスは NHK が保存している映像を各ジャンルに沿って収録したデジタルアーカイブ 6 サイトを公開している。

表 1 は NHK デジタルアーカイブスを構成するアーカイブ及びそのコンテンツとメタデータ項目を示している。アーカイブ毎に独自のカテゴリーが割り振られている場合や地理情報の有無など形式が異なるので、メタデータの 1 つの標準に整理することは難しい。しかし、収録番組や放送年月日、キーワードの項目については共通するものが多い点が見受けられる。

表 1. NHK デジタルアーカイブスのコンテンツとメタデータ項目

アーカイブ名	コンテンツの種別	メタデータ項目
戦争証言アーカイブス	番組	番組名,戦地,放送日,番組内容,チャプター名,タイムコード
	証言	タイトル,氏名,戦地・収録年月日,再生テキスト,チャプター名,タイムコード,出来事の背景,証言者プロフィール,キーワード
	日本ニュース	タイトル,公開日,再生テキスト,チャプター名,タイムコード
	戦時録音	タイトル,年月日,再生テキスト,背景,関連資料
東日本大震災アーカイブス	証言	タイトル,氏名,年齢,場所,番組名,取材時期,チャプター名,再生テキスト
	復興映像	タイトル,場所,番組名,チャプター名,再生テキスト
エコチャンネル	動画	タイトル・収録番組・放送年月日・動画説明文・カテゴリー・キーワード
NHK映像マップ みちしる	動画	タイトル・サブタイトル・詳細情報・都道府県・主な撮影地・関連テーマ
NHKクリエイティブ・ライブラリー	動画	タイトル・動画説明文・番組名・副題・放送・カテゴリー・キーワード
みのがし なつかし	番組	タイトル・放送期間・コンテンツの番組放送年・主な出演者・番組詳細

4.2 LOD リソースと番組情報を利用した NHK デジタルアーカイブスの連携

本節では、NHK デジタルアーカイブスを対象としてメタデータモデルの適用事例を述べる。図 3 は NHK デジタルアーカイブスの 1 つである「エコチャンネル」と統合アーカイブス及びその放送コンテンツの関係を記述したものである。図の放送コンテンツはエコチャンネルが公開しているものの 1 つであり、統合アーカイブスにも属していると表現されている。図の ProxyE は当該放送コンテンツに対するメタデータを記述するためのプロキシである。ProxyT には連携のためのメタデータを記述することでコンテンツ間のリンクを行う。このとき、メタデータには DBpedia や NDLSH の LOD リソースに存在する主題情報と番組情報を記述する。

LOD リソースを用いて主題情報を付与する場合、アーカイブに付与されているキーワードに利用する方法がある。コンテンツに付与されているキーワードを用いて、LOD リソースから関連のある単語を判別し、URI を付与する。また、コンテンツと関連の深い主題情報を記述するために、タイトルや動画説明文からそのコンテンツの主題に適したキーワードを抽出し、メタデータを新たに記述する。その場合、タイトルや説明文に対して形態素解析プログラムを使用して DBpedia や NDLSH の単語を抽出する。この 2 つの方法で作成したメタデータを統合アーカイブスにおけるメタデータとして付与することで放送コンテンツ同士の主題による関連を作成する。

また、NHK デジタルアーカイブスのコンテンツには放送されたテレビ番組名をメタデータとして持つものが多い。テレビ番組はジャンル、放送期間などそれ自体に多くの情報を持つ。また、テレビ番組は毎週放送されるものもあり、放送された番組毎に出演者や放送日、撮影地等のメタデータを持つ。番組の基本的な情報と放送された番組ではそれぞれメタデータが異なる。本研究では、アーカイブのコンテンツと番組情報をより結びつけるために番組情報の構造化を行う。タイトルや放送期間、ジャンル等のテレビ番組の基本的な情報を「番組情報」として扱い、実際にテレビで放送された番組を「放送番組」として放送コンテンツと結びつけた。その構造を図 3 に示す。

以上のように本来のメタデータに加えて新たに付与される LOD リソースや番組情報は異なるアーカイブ間を結び付ける、アーカイブ間にまたがった共通の情報である。こうした新たな情報を付与することで異なるアーカイブのコンテンツをより意味的に結びつけ、統合検索に加えて新たな利用方法を生み出すことが出来ると考察する。

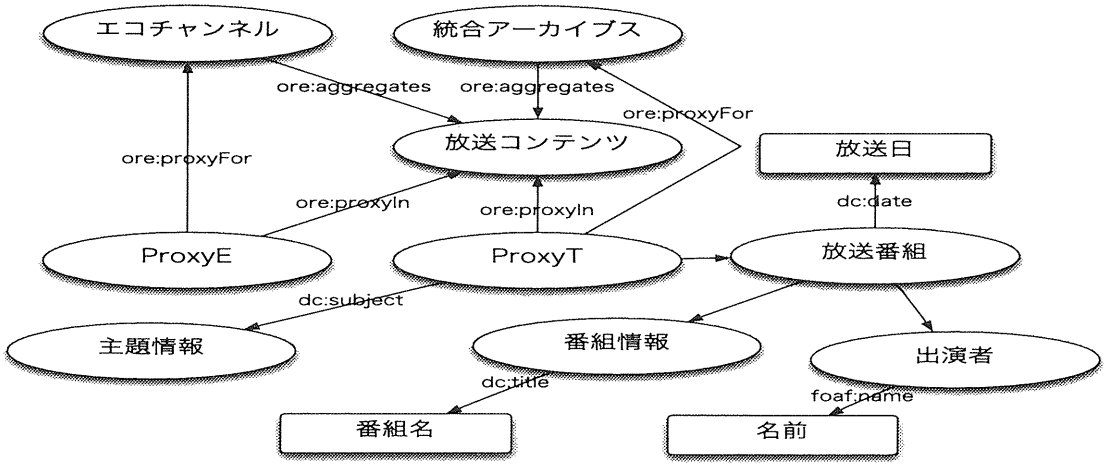

図 3. NHK デジタルアーカイブス連携のためのメタデータ記述

5. アーカイブコンテンツの LOD 化事例 -みちしる LOD-

3 章ではアーカイブ間の結びつきを作成するための手法について述べたが、連携のためのメタデータを作成する際にはアーカイブ毎にメタデータモデルを設計する必要がある。本章では NHK デジタルアーカイブスの 1 つである「NHK 映像マップ みちしる」(みちしる)を例にあげて、その LOD 化について説明する。みちしるは NHK が保存している歴史的映像から厳選した日本の映像を視聴できる Web サービスである。

みちしるの LOD 化の手順を大きく分けると以下の 3 つとなる。

1. Web サイトからのスクレイピングを用いたデータの収集
2. メタデータモデルの設計
3. 抽出したデータから RDF 形式のファイルへの変換

1 では、Web サイトの HTML ファイルの木構造を利用したスクレイピングプログラムを実行し、要素を抽出することで放送コンテンツのメタデータを収集した。その際に、URL・タイトル・都道府県・テーマ・画像タイトル・画像 URL・動画 URL・サブタイトル・緯度・経度・撮影地・説明文を取得した。

2 では、手順 1 で収集したメタデータを元にメタデータスキーマを定義した。取得したメタデータをグラフで表現したものを図 4 に示す。メタデータを設計する際はダブリンコア等の LOD のための標準を使用した。また、みちしるのコンテンツに付与されているキーワードに対して、DBpedia の項目や NDLSH の典拠データへのリンクを作成することで他の LOD リソースとの連携を図った。

3 では、収集したデータを定義したメタデータスキーマを元にデータの変換を行った。コンテンツの数は約 2500 件、RDF トリプルの数は約 12000 件である。RDF/Turtle 形式で作成したメタデータは Web サイト上へと公開し、利用できるようになっている[10]。

6. 終わりに

本論文は、異なる放送コンテンツアーカイブを連携するためにEuropeanaを参考にして放送アーカイブのWebサイト向けに設計したメタデータモデル、ならびにLODデータセットについて述べた。異なるアーカイブは所蔵するコンテンツのメタデータの定義が異なるため連携して統合検索することが難しい。そのため、放送コンテンツアーカイブに共通する番組情報とLODリソースを用いてコンテンツ同士の関連付けについての手法の提案を示した。アーカイブとコンテンツの関係をOAI-OREモデルを利用したデータモデルで記述することで元のメタデータを保持しつつ、関連付けを行うためのメタデータの付与することができることを明らかにした。今後は、番組情報の構造化データと典拠データやDBpediaといった主題情報を追加したLODデータセットの作成と実際にメタデータを利用した統合検索システムの開発と実験を進めていく予定である。

参考文献

[1] 知のデジタルアーカイブに関する研究会開催要綱,
http://www.soumu.go.jp/main_content/000101009.pdf（アクセス：2014.10.17）

[2] デジタルアーカイブの構築・連携のためのガイドライン,
http://www.soumu.go.jp/main_content/000153595.pdf（アクセス：2014.10.17）

[3] Europeana, http://www.europeana.eu/（アクセス：2014.10.18）

[4] 国立国会図書館東日本大震災アーカイブ, http://kn.ndl.go.jp/（アクセス：2014.10.17）

[5] 国立国会図書館ダブリンコアメタデータ記述,
http://dl.ndl.go.jp/view/download/digidepo_8295098_po_dcndl201112.pdf?contentNo=1（アクセス：2014.10.18）

[6] NHKデジタルアーカイブス, http://www.nhk.or.jp/archives/digital/(アクセス:2014.10.15)

[7] DBpedia, http://dbpedia.org/（アクセス：2014.10.19）

[8] 国立国会図書館件名標目表, http://id.ndl.go.jp/auth/ndla（アクセス：2014.10.19）

[9] ORE Specification - Abstract Data Model,
http://www.openarchives.org/ore/1.0/datamodel（アクセス：2014.10.15）

[10] NHK映像マップ みちしるLOD, http://mdlab.slis.tsukuba.ac.jp/lodc2013/michishiru/(アクセス：2014.10.15)

[11] 文化審議会 著作権分科会 過去の著作物等の保護と利用に関する小委員会（第6回）議事録・配付資料 ［資料3］－文部科学省,
http://www.mext.go.jp/b_menu/shingi/bunka/gijiroku/021/07073007/003.htm（アクセス：2014.10.19）