WWW 上の学術文書からのメタデータ抽出

杉 木 健 二[†]　松 原 茂 樹[††]　吉 川 正 俊[†]

学術情報の電子化が進み，WWW 上の学術情報リソースの高い流通性が求められており，これらのリソースに対する索引や目録としてメタデータを付与することが重要な課題となる．本論文では，WWW 上の学術情報文書からメタデータを生成する方法について述べる．本研究では，NII 限定子を用いて Dublin Core Metadata Element Set を拡張したメタデータセットを用いた．まず，固有名詞情報やリンク元のアンカーテキストを手がかりに，大学内の Web ページをタイプ判定し，タイプ別にヒューリスティックルールを用いてメタデータの生成する．名古屋大学 Web サイト資源に対してメタデータ生成実験を行った結果，本手法の実現可能性を確認した．

Metadata Extraction from Scholarly Web Documents

Kenji Sugiki,[†] Shigeki Matsubara[††] and Masatoshi Yoshikawa[†]

By electronization of scientific information, the distributivity of the scholarly information resource on WWW is strongly required. It is becoming an important subject to give the metadata such as indexes and tables to these resources. This paper describes a method of generating the metadata from the scholarly document on WWW. This metadata is based on the metadata sets to which Dublin Core Metadata Element Set is extended by the schema of NII. First, by using information about the name entity and anchor texts of linked documents, the types of web pages are identified. Then, the metadata is generated using heuristic rules with each type. An experiment on metadata generation using Nagoya University Websites resources has shown the feasibility of our method.

1. はじめに

近年，学術情報の電子化が進み，デジタルライブラリの研究が盛んになりつつある．学術情報の流通性を高めるために，従来の図書館と同様に，デジタルライブラリにおける学術情報リソースにも目録や索引を付与することが重要となる．このような背景のもと，Dublin Core[5] をはじめ，様々なメタデータセットが提案されている．メタデータは大量のリソースを効率よく利用するうえで重要な役割を果たす．メタデータを用いることにより，WWW 上の学術情報を効率的に管理し，かつ多角的なアクセスや検索を容易に行うことができる．

Web 上のリソースに対するメタデータの付与は，これまで人手で行われることが一般的であった．しかしながら，リソースはますます増大し，更新も随時行われているため，これらの大量のデータを人手で管理することは困難になってきている．

そこで本論文では，学術情報が記述された Web ページからメタデータを生成する方法について述べる．本研究では，大学の研究者情報や広報資料など，Web 上の学術情報リソースに対してメタデータを付与する．メタデータセットとしては，NII 限定子を用いて拡張した Dublin Core Metadata Element Set[1] を用いた．まず，組織名や研究者名などの固有名詞，及びリンク元のテキストを用いてタイプを判定し，次に，タイプ別にヒューリスティックルールを適用してメタデータを生成する．Dublin Core は，機関リポジトリのメタデータ交換プロトコルとしてよく用いられる OAI-PMH[7] に対応しており，このメタデータセットに基づくメタデータを作成することによって，機関間での学術情報の流通が容易になる．

国立情報学研究所（NII）では，各大学等で登録されたデータを用いてデータベースを構築し，Web 上の学術情報のポータルサイトを提供する目的で大学 Web サイト資源検索 (junii)☆を試験的に運用している．しかし，2006 年 2 月現在の登録数は 7 万 5 千件であり，機関ごとの登録件数は多くの場合，数十〜数百件程度である．メタデータ登録の手間を考慮すると，人手による作業のみで大規模データベースを構築することは難しい．本研究では，半自動的にメタデータを生成することにより，これらの問題の解決を目指している．

以下，2 章で本研究で対象とする Web データについて述べ，その後，3 章では，本論文で用いるメタデータセット仕様について述べる．4 章では，メタデータの生

† 名古屋大学大学院情報科学研究科
Graduate School of Information Science, Nagoya University
†† 名古屋大学情報連携基盤センター
Information Technology Center, Nagoya University

☆ 大学 Web サイト資源検索− junii 大学情報メタデータポータル
http://ju.nii.ac.jp/

表 1 本研究で使用するメタデータセット

要素	修飾子	限定子
Title		
	Transcription	
	Alternative	
Subject		NDC
		NII
Creator		
	Transcription	
	Alternative	
Description		
Identifier		URL
Coverage	Spatial	NII
Date		
Type		NII
		DCMI
Publisher		
	Transcription	
	Alternative	
Format		IMT
Relation	hasVersion	URL
	isReferencedBy	URL
Source		URL
Language		ISO639-2

成手法について述べ，5 章でメタデータ生成実験の結果と考察について述べる．

2. 名古屋大学 Web サイト資源

名古屋大学では，2004 年度より学内の学術情報のアーカイビングプロジェクトを推進している．plum では，大学内の大量の Web ページを定期的に収集し，蓄積した Web ページを利用して，キーワード検索，カテゴリ検索[2),3)]，アーカイブ検索，研究者検索など多角的な検索環境を提供している[8)]．2005 年度に実施した収集では，名古屋大学内のドメインの総数は 541，ファイルサイズの合計は約 106GB，ファイルの総数は約 100 万であった．そのうち Web ページ（text/html）の総数は約 30 万ページであった．

本研究では，収集した Web ページに対してメタデータの生成を試みる．作成したメタデータデータベースを学術機関リポジトリ[4)] などとともに公開し，また，より高度で多角的な検索環境の提供，及び，学術情報資源の管理の実現を目指している．

3. メタデータのタグセット

Open Archive Initiative[6)](OAI) は，電子図書館リポジトリ間の相互運用のためにメタデータ収集プロトコル OAI-PMH[7)](The Open Archives Initiative Protocol for Metadata Harvesting) を策定した．OAI では，リポジトリの収集（ハーベスティング）のために Dublin Core に基づいたメタデータを提供している．本研究では，この Dublin Core を基本とし，さらに NII の限定子を用いて拡張したメタデータセットを用いた．本研究で用いるメタデータセットを**表 1** に示す．Contributer, Rights 項目を除く 13 項目の要素と，その要素に対する修飾子，限定子を定義した．以下，要素，修飾子，限定子の組を [要素].[修飾子].[限定子] と表すことにする．ただし，修飾子，あるいは限定子が空白の場合は，それらを省略して表す．例えば，Title.Alternative，Type..NII，Relation.isReplacedBy.URL のように表す．

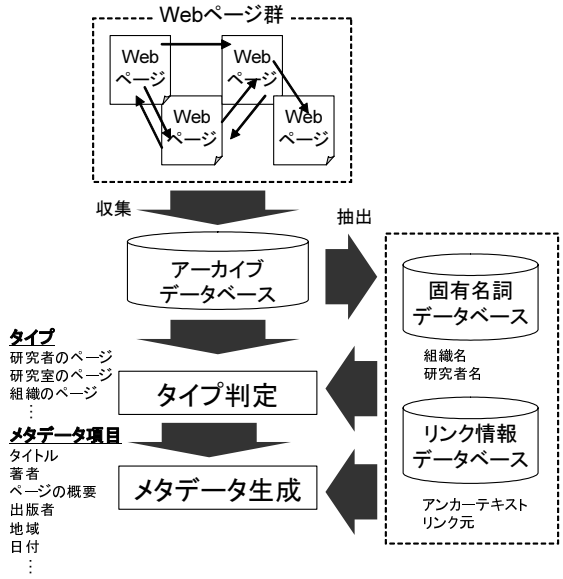


図 1 メタデータ生成方法の概要

4. メタデータの生成方法

4.1 生成方法の概要

メタデータ生成方法の概要を**図 1** に示す．まず，大学内の Web サイトをクローリングし，Web ページを収集する．次に，収集した Web ページから研究者名や組織名などの固有名詞データベースを作成し，アンカーテキストとリンク先 URL を保持したリンク情報データベースを作成する．続いて，大学内の Web ページから固有名詞データベースとリンク情報データベースを利用して，「組織のトップページ」「研究者の個人ページ」「機関広報資料」といった Web ページのタイプを特定する．特定するタイプ項目については 4.2 節で詳しく述べる．タイプ判定後，タイプ別に Title，Publisher，Creator などの要素を付与し，それ以外の要素については各タイプ同様の処理により値を付与し，メタデータを生成する．

以下では，(1) 前処理，(2) Web ページのタイプ判定，(3) メタデータの生成，のそれぞれのフェーズについて詳しく述べる．

4.2 前 処 理

前処理として，Web ページの収集，固有名詞データベース，リンク情報データベースを作成する．

- **大学内の Web ページの収集**
  2 章で述べたような学内の Web ページを収集する．収集にはクローラ（ソフトウェアロボット）を用いる．

表 2 NII の資源タイプ語彙集（の一部）

（第一階層）	（第二階層）
研究者情報	個人のページ
	研究室トップページ
図書館情報	図書館・室トップページ
広報資料	機関トップページ
	下部組織トップページ
	機関広報資料

- **固有名詞データベースの作成**
  組織，研究科，研究者の固有名詞データベース (日/英) を作成する．組織，研究科の名称は人手で作成し，研究者の名前は，学内で公開されている研究者データベースから研究者の名前，ヨミ，英語名をそれぞれ抽出する．
- **ドメインと NII 主題語彙集（Subject..NII）との対応付け**
  ドメインに対して，NII のメタデータ語彙集☆の１つである主題語彙集（Subject..NII）（比較的粒度の粗い主題語彙）との対応付けを人手で行う．この主題語彙集の，第一階層 9 項目，第二階層 78 項目に分類を行う．
- **リンク情報データベースの作成**
  収集した全ての Web ページから，アンカータグのリンク先の URL とテキストを取得し，それら２つの組を保持したリンク情報データベースを作成する．Web ページのリンクには相対リンクになっているものや，同一の URL でも表記の異なるものが存在するので，同一の URL となるように統一する．

4.3 Web ページのタイプ判定

本研究では，取得する Web ページのタイプとして NII の定める資源タイプ語彙集（Type..NII）を用いる．本研究で対象としたタイプ（語彙集の一部）を**表 2** に示す．メタデータ上では，タイプ（Type..NII）を（第一階層）-（第二階層）と表記するが，以下では，Web ページのタイプを NII 語彙集の第二階層のみで示すことにする．これらのタイプに基づいて，タイプ別に Web ページを取得し，分類を行う．「個人のページ」,「機関トップページ」など，固有名詞データベースが利用できる場合は，検索エンジンを用いて固有名詞の含まれるページを取得し，さらにアンカーテキストに固有名詞の含まれているリンク先 URL を取得する．リンク先の URL の中で上位ものを取得し，さらに，リンク先に固有名詞が含まれるものを取得する．「下部組織トップページ」や「研究室トップページ」など，固有名詞データベース中に名称が含まれないページを取得する場合は，アンカーテキストに「専攻」,「講座」,「研究室」などのキーワードが文末に含まれるリンク先の URL を取得し，さらに，リンク先のページのタイトルかヘッダタグ，もしくは先頭 300 字以内にこのキーワードが含まれるものを取得する．それ以外のページについては，人手で取得し，分類する．

4.4 メタデータの生成

本研究では，タイプ別に Web ページを分類し，それぞれのメタデータ項目を生成する．

4.4.1 Title，Creator，Publisher の生成

Web ページのタイプ別ごとに Title，Creator，Publisher を生成する．

- **組織トップページ**
  固有名詞データベースを用いて，Title，Transcription.Alternative に組織の名前，英語名をそれぞれ付与する．Title.Transcription に付与する組織名のヨミは，茶筌[10]（形態素解析器）を用いて付与する．茶筌には，あらかじめ人名や組織名の辞書を追加しておく．また，Creator，Creator.Transcription，Creator.Alternative，Publisher，Publisher.Transcription, Publisher.Alternative も同様にして，それぞれ組織名の付与を行う．Web ページのタイトルか H1 タグのテキストは Title.Alternative に記述する．ただし，Title と同じ場合は記述しない．
- **個人のページ**
  「組織のトップページ」とほぼ同様の処理を行うが，「個人のページ」の場合は，Publisher が組織であるか研究者であるかの２通りの場合がある．そこで，あらかじめ組織の研究者ページの URL を登録しておき，マッチすれば組織名が，それ以外の場合は研究者の名前が Publisher となるようにする．
- **下部組織トップページ**
  Web ページ中から「研究科」,「専攻」などのキーワードが含まれるテキストを抽出，結合し，で正式な組織名へと補完し，Title，Publisher，Creator を設定する．Transcription 修飾子は，それぞれ茶筌を用いてヨミを付与する．「図書館・室トップページ」の場合も同様の方法を用いる．
- **研究室トップページ**
  「下部組織のトップページ」の場合とほぼ同様である．固有名詞データベースから研究科の英語名と研究者の名字の英語名を取得し，さらに文字列「Lab.」を付け加え，Publisher.Alternative，Creator.Alternative を設定する．ただし，研究室名が研究者の名字でない場合は，表記しない．
- **機関広報資料**
  取得した機関や下部組織，図書館の Publisher,Creator を利用して，Web ページのドメインが同じ場合，同様の Publisher，Creator を付与し，Relation.isPartOf.URL をその組織のトップページの URL に設定する．

4.4.2 Description の生成

Description 項目を記述するために，要約文を生成す

☆ 詳しくは「NII メタデータ・データベース入力マニュアル」[9] を参照されたい

る．まず，Web ページを各タグとそのタグに含まれるテキストとの組に分割する．サイトのトップページなどでは，メニュー，見出しなどが画像で記述されている場合が多いので，IMG タグの ALT 属性の値もテキストとして追加する．以下のヒューリスティックな削除ルールを各タグに適用することにより要約文を生成する．以下に列挙する削除対象にマッチするタグやそのテキストをその順に削除していき，最終的に 300 字以下となるようにする．

(1) 「english」「here」「日本語」「ジャンプ」「戻る」など本文の内容と関係のないと思われるテキスト
(2) 「・」「：」「■」などの記号
(3) バイト数が 3 バイト以下のテキスト
(4) テキストが文であるものを削除するために，助動詞，代名詞，句点を含むテキスト
(5) 先頭が「の」で始まるテキスト，最後が「へ」,「が」,「の」で終わるテキスト
(6) 「について」「のページ」などの文字列
(7) 括弧と括弧内のテキスト
(8) 「TEL」,「FAX」,「電子メール」など連絡先を表すテキスト
(9) 文末のコピーライト（©,（C）など）が出現するテキスト
(10) リストタグ内に出現するリストタグ
(11) アンカータグの次に出現するリストタグ
(12) Web ページのドメイン外へリンクを張っているアンカータグ
(13) 他のタグと重複しているテキスト
(14) 上記のルールを用いても 300 字以上の場合，後方のタグを削除し 300 字以下にする

Description の生成例を**図 2** に示す．この例の場合，要約により 1294 字から 222 字（要約率 17%）になっている．

4.4.3 その他の項目の設定

その他の項目の設定について以下に示す．

Identifier..URL Web ページの URL を記述する．

Subject..NDC 「377」（大学, 高等・専門教育, 学術行政）とする．

Coverage.Spatial.NII Web ページ中に県名が含まれていればそれを付与し，それ以外は「日本」とする．

Source..URL アーカイブサーバの URL を記述する．

Language..ISO639-2 文字コードによって自動判別し，記述する．

Date サーバーから Web ページを取得するとき，HTTP ヘッダのメタ要素に Last-Modified があればその日付を記述する．

Type..DCMI 「text」とする．

Format..IMT 本論文では HTML のみを対象としたので「text/html」とする．

Relation.hasVersion.URL Web ページ上のアンカーテキストに文字列「English」「英語」などがあれば，そのリンク先を設定する．

松原 茂樹, 名古屋大学, 情報連携基盤センター, 学術情報開発研究部門, 助教授, 工学部電気電子情報工学科情報工学コース 担当, 大学院情報科学研究科社会システム情報学専攻 担当, 附属図書館研究開発室 担当, 新着, 教育, 研究, 共同研究, プロフィール, 研究成果, 自然言語処理, 音声言語処理, 情報検索デジタル図書館, 松原グループのページ, 離散数学及び演習, 知識社会システム論セミナー, 社会システム情報学特論, 音声情報処理, 言語情報処理

図 2 Description の生成例
(http://www.el.itc.nagoya-u.ac.jp/ matubara/)

Relation.isReferencedBy.URL 収集した Web ページ全体を解析し，Web ページのリンク元ページがメタデータが生成された URL であれば，その URL を記述する．

5. メタデータの生成実験

5.1 生成実験の結果

名古屋大学 Web サイト資源を用いてメタデータ生成実験を行った．タイプごとのメタデータ数を**表 3** に示す．Web ページのタイプ判定を行い，メタデータが生成されたページ数は 932 であった．そのうち，半自動でタイプ判定を行ったものは 893 ページ，人手で行ったものは 39 ページである．実際に作成したメタデータの例を**図 3** に示す．この例の場合，タイプは「機関のトップページ」であるので，Title, Creator, Publisher 項目は組織の名前とし，また，修飾子 Transcription, Alternative には，それぞれヨミ，英語名を付与した．Subject..NDC はドメインで設定しており「情報学」とした．Description 項目は，Web ページを要約することにより 1225 字から 207 字（要約率 17%）となり，それを記述した．これは，Web ページ中の重要な箇所を抽出できた例である．

5.2 考　　察

5.2.1 Web ページのタイプ判定

本手法では，Web ページのタイプ判定として，固有名詞や特定のキーワードがアンカーテキストに含まれているかどうかによって判定した．研究者や研究室の数を考慮すると，タイプ判定を行うことができた Web ページの数が少ないように思われる．この方法を用いることに

表 3 タイプ別メタデータ数

（第一階層）	（第二階層）	生成数
研究者情報	個人のページ	513
	研究室トップページ	275
図書館情報	図書館・室トップページ	16
広報資料	機関トップページ	51
	下部組織トップページ	54
	機関広報資料	23
合　計		932

よって精度は高くなるが，固有名詞やキーワードがアンカーテキストに出現しなければ目的とする Web ページは取得できないので，再現性が低くなる傾向があり，本手法のみでは必ずしもすべて取得することができるとは限らない．また,「個人のページ」に見られるように，研究者の名前のリンク先が「研究室トップページ」となる場合がある.「個人ページ」が「研究者のページ」である場合もあるが，リンク先として,「個人のページ」ではなく「研究室トップページ」を指している場合も存在した．よって，アンカーテキストのみを考慮するのではなく，前後のテキストからも固有名詞やキーワードを探す，もしくは，Web ページの特徴（構造的特徴など）を捉えることによって，タイプ判定を行う必要がある．

5.2.2 Creator,Publisher

メタデータを生成するうえで，Creator，Publisher を誰にするかということが問題となる．例えば，研究室の Web ページが研究科のサーバに存在する場合，Publisher が誰になるのか，研究科の公式ページの研究者情報のページの Creator は誰になるのかということである．本研究では，特定の URL であれば Creator，Publisher を変更するなどの設定を行った．人手で記述する場合も Creator，Publisher を判断することが困難である場合も存在すると思われる．

5.2.3 Description

Description の生成において単純なヒューリスティックルールを用いることにより，**図 2**，**図 3** の例のように Web ページの内容の要約を表しているのは，ある程度 Web ページが構造化されているからだと思われる．本手法では，フレーズを抽出するという手法で要約を行ったが，要約手法としては文を生成するという手法も考えられる．しかし，文としてではなくタグの階層に基づいたキーワード（フレーズ）の羅列として内容を表現する Web ページが多く存在し，特にサイトのトップページは，サイト全体を手短に表現するためにその傾向が強い．また，文が主体の Web ページの場合においても文ではなくタイトルや見出しなどのフレーズが Web ページの内容を端的に表現している場合が多い．よって，Web ページの要約においては重要なキーワードを抽出するという手法が適していると思われる．

一方，Description が全く記述されないページも存在した．それは，画像や flash をクリックすると，コンテンツに移動するというサイトである場合である．また，タグを用いて階層化されていないページなどは，要約が極端に短くなる，もしくは，ページの重要でない部分が抽出されてしまう場合がある．

6. おわりに

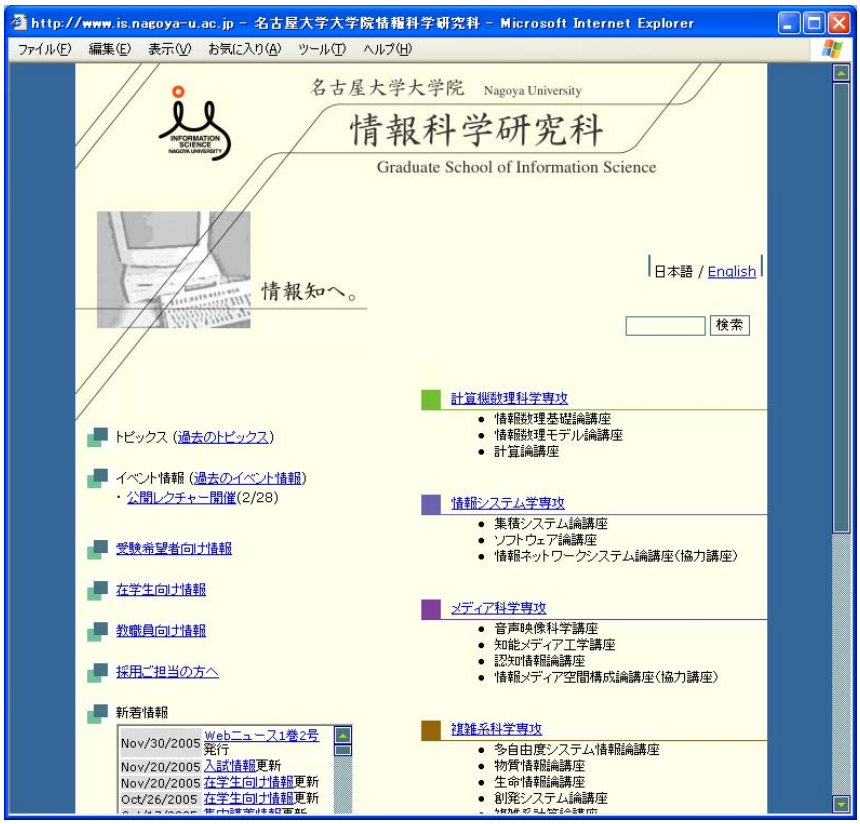
本論文では，学術情報が記述された Web ページにメタデータを生成する方法について述べた．Web ページに対して Dublin Core 形式のメタデータを生成することができた．本手法では，固有名詞データベース，リンク情報データベースとヒューリスティックルールを用いることにより，Web ページのタイプ判定，Title，Publisher，Creator，Description などの項目を生成した．

今後は，メタデータの生成数を増加し，Web ページの主題を示す Subject 項目など，各メタデータ項目の内容を充実させるとともに，より完成度を高める必要がある．デジタルライブラリにおけるメタデータとは，その機関がどのような学術情報資源を保持しているかを示す重要なものなので，メタデータは正確に記述されるべきである．しかし，本手法で生成したメタデータについての評価を行っておらず，今後，これらのメタデータの信頼性を確認する必要がある．

謝辞　本研究の一部は，名古屋大学附属図書館への NII 委託事業「学術ナレッジ・ファクトリー（AKF）の開発及び構築」により実施した．

参 考 文 献

1) Dublin Core Metadata Element Set, Version 1.1: Reference Description, http://dublincore.org/documents/dces/
2) 松原 茂樹, 鈴木 祐介: インターネットアーカイブに基づく Web ディレクトリの設計と構築, 名古屋大学附属図書館研究年報, No.3, pp. 33-37（2004）.
3) 鈴木 祐介, 松原 茂樹, 吉川 正俊: ハイパーリンクを用いた Web 文書の自動分類, 言語処理学会第 11 回年次大会発表論文集, pp. 61-64（2005）.
4) 郡司 久: 名古屋大学における学術機関リポジトリ構築への取り組み, 情報の科学と技術, Vol.55, No.10, p.439-446（2005）.
5) Dublin Core Metadata Initiative(DCMI), http://dublincore.org/
6) Open Archives Initiative(OAI), http://www.openarchives.org/
7) OAI-PMH(The Open Archives Initiative Protocol for Metadata Harvesting), http://www.openarchives.org/OAI/openarchivesprotocol.html
8) plum(名古屋大学 Web サイト資源検索), http://plum.itc.nagoya-u.ac.jp/
9) NII メタデータ・データベース入力マニュアル, http://www.nii.ac.jp/metadata/manual/
10) 形態素解析システム『茶筌』2.3.3 使用説明書, http://chasen.aist-nara.ac.jp/chasen/doc/chasen-2.3.3-j.pdf



メタデータ項目	値
Title	名古屋大学大学院情報科学研究科
Title.Transcription	ナゴヤダイガクダイガクインジョウホウカガクケンキュウカ
Title.Alternative	Graduate School of Information Science, Nagoya University
Subject..NDC	377
Subject..NII	情報学
Creator	名古屋大学大学院情報科学研究科
Creator.Transcription	ナゴヤダイガクダイガクインジョウホウカガクケンキュウカ
Craetor.Alternative	Graduate School of Information Science, Nagoya University
Description	トピックス, イベント情報, 受験希望者向け情報, 在学生向け情報, 教職員向け情報, 新着情報, ネットワーク情報, 研究科紹介, 教員一覧, サイトマップ, 計算機数理科学専攻, 情報システム学専攻, メディア科学専攻, 複雑系化学専攻, 社会システム情報学専攻
Identifier..URL	http://www.is.nagoya-u.ac.jp/index.html
Coverage.Spatial.NII	日本
Date	
Type..NII	広報資料-機関トップページ
Type..DCMI	text
Publisher	名古屋大学大学院情報科学研究科
Publisher.Transcription	ナゴヤダイガクダイガクインジョウホウカガクケンキュウカ
Publisher.Alternative	Graduate School of Information Science, Nagoya University
Format..IMT	text/html
Relation.hasVersion.URL	http://www.is.nagoya-u.ac.jp/index.html.en
Relation.isReferencedBy.URL	http://www.cm.is.nagoya-u.ac.jp/index.html http://www.cs.is.nagoya-u.ac.jp/index.html http://www.ss.is.nagoya-u.ac.jp/index.html
Source..URL	http://plum.itc.nagoya-u.ac.jp/
Language..ISO639-2	jpn

図 3 Web ページとそのメタデータの生成例 (http://www.is.nagoya-u.ac.jp/)