非標準フォーマットを含む埋め込み型メタデータの抽出と統合による メタデータ生成手法

本間 維[1]　永森 光晴[1,2]　杉本重雄[1,2,3]

筑波大学図書館情報メディア研究科[1]
筑波大学図書館情報メディア系[2]
筑波大学知的コミュニティ基盤研究センター[3]
〒 305-8550 茨城県つくば市春日 1-2
E-Mail : {tsuna，nagamori，sugimoto}@slis.tsukuba.ac.jp

概要

Web 上で公開されている Web ページ中にメタデータを記述するために，RDFa や Microformats，Microdata など複数の標準的な記述フォーマットが提案されている．しかし，標準的な記述フォーマットに従ったメタデータを持つ Web ページの数はまだ少なく，メタデータを利用した情報流通支援を行うには，より積極的なメタデータ付与が求められる．本稿では，DCMI Description Set Profile を基にした情報抽出テンプレートによる，相互利用性向上や作成コスト軽減を意識したメタデータ生成手法を提案する．

キーワード

メタデータ，メタデータスキーマ，Application Profile，情報抽出，セマンティックウェブ

Metadata Creation from Resources with Embedded Metadata - Extraction and Integration of Embedded Metadata in Standard and Non-standard Formats

Tsunagu HONMA[1]　Mitsuharu NAGAMORI[1,2]　Shigeo SUGIMOTO[1,2,3]

Graduate School of Library, Information and Media Studies, University of Tsukuba[1]
Faculty of Library, Information and Media Science, University of Tsukuba[2]
Research Center for Knowledge Communities, University of Tsukuba[3]
1-2, Kasuga, Tsukuba, Ibaraki, 305-8550, JAPAN
E-Mail : {tsuna，nagamori，sugimoto}@slis.tsukuba.ac.jp

Abstract

Standard metadata formats, such as RDFa, Microformats, and Microdata, have been recommended to embed metadata in HTML or XHTML documents. However, many web pages have no metadata written in those formats. The other side, non-standard formats describing information for layout of web pages is used widely. For increasing metadata more proactively, we regard information in non-standard formats as metadata, and integrate with metadata in standard formats. This paper proposes a method to create metadata from resources with embedded metadata in standard and non-standard formats.

Keywords

Metadata，Metadata Schema，Application Profile，Metadata Extraction，Semantic Web

1. はじめに

Web 上で公開されている Web ページ中にメタデータを記述するために，RDFa[1]や Microformats[2]，Microdata[3]など複数の標準的な記述フォーマットが提案，利用されている．しかし，そうした標準的記述フォーマットに従ったメタデータを持つ Web ページの数はまだ少ない．情報資源の発見や分類，関連付けなどをメタデータによって支援するためには，より積極的なメタデータ付与が求められる．一方，各 Web サイトがページレイアウトのために独自に定義した HTML の class 属性の値などを用いたフォーマット（以下，非標準フォーマット）で構造情報を記述することがある．この構造情報はメタデータとしての利用を意図して付与されているものではないが，構造情報と対応する Web ページ中の記述内容を併せてメタデータとして捉えることもできる．非標準フォーマットをメタデータとして扱うことで，メタデータを持つ Web ページの数を増やすことが可能になる．こうした非標準フォーマットを含むメタデータを他の情報資源のメタデータと組み合わせて情報資源の統合や横断検索などに利用するためには，メタデータ間の相互利用性向上を考慮する必要がある．本稿では，相互利用性向上などを意識したメタデータをより少ないコストで付与するため，DCMI Description Set Profile を基にした情報抽出テンプレートによる情報抽出・統合によりメタデータを生成する手法を提案する．本手法を用い， Web ページ中に様々なフォーマットで記述されている情報を一つの構造化されたメタデータとして統合し取り出すことで，元の Web ページに手を加えることなくメタデータを付与できるようにすることと，メタデータ記述フォーマットの多様化に伴うメタデータ抽出の手間を軽減することを目的とする．

2. Web ページ中に記述されるメタデータ

メタデータ作成者は，付与したメタデータが情報流通において役立つことを期待している．本章では，メタデータの持つべき特徴と，メタデータの付与される目的をメタデータの機能要件として挙げ，現在の Web ページ中のメタデータがその機能要件を十分に満たせているのかについて述べる．

2.1 メタデータの機能要件

メタデータが付与される目的として，「記述対象となる情報資源の持つ属性記述，情報資源の発見・検索，情報資源の維持管理あるいは保存，情報資源の提供や取引」があり，メタデータに求められるのは「メタデータを組み合わせて利用することとメタデータの相互運用性を高めること」であるとされている[4]．また，Dublin Core Metadata Initiative では，メタデータのあるべき姿の一つとして，資源のコンテキストを表現するものとしてのメタデータを挙げている[5].

本稿では，このような既存のメタデータへの機能要件を踏まえつつ，Web ページ利用者の視点から，メタデータが持つべき機能要件として以下の 5 つを挙げる．

１）内容を知らせる効果

メタデータは，あるデータ（もの，こと）に関するデータを記述したもの（data about data）である．その記述は，コンテンツの中身が何であるかを表現するために用いられる．Web ページ中に記述されるメタデータは，その Web ページで書かれているテーマやページ作成日，ページの作成者などがあり，これらのメタデータを利用することで Web ページに関する情報を端的に伝えられる．メタデータによってその Web ページの内容を知らせる代表的な例として，Web 検索サービスが検索結果一覧に表示する各 Web ページのスニペットが挙げられる．サービス利用者はスニペットを見ることで，それぞれの Web ページに関する情報を知ることができ，アクセスする情報の選択に役立てることができる．

２）検索・分類支援効果

プログラムによりコンテンツの内容を知らせるメタデータを収集し，情報検索のためのインデックスを構築することができる．これにより，Web ページの本文には出現しない日付や著者など構造化されたメタデータを検索対象項目として用いることで，精度の高い検索が期待できる．

また，キーワードとして利用可能な語を特定の語彙に限定するなど，メタデータとして記述する値

を統制することで，各メタデータ記述項目を Web ページの分類として利用できる．これにより，ある Web ページと同じ分類に属する Web ページの網羅的な探索などが支援される．

３）組織化支援効果

Web ページから他の Web ページへのリンクをメタデータとして記述することで，複数の Web ページ間の関連性を意識した組織化が可能になる．コンテンツ提供者が複数の Web ページを組み合わせて一つの物事に関する情報を記述する際，各々のつながりをメタデータによって明示することで，コンテンツ利用者による情報資源の結びつけと理解を支援するといったように，Web ページが独立の情報として利用されるだけでなく，Web ページ A と B を併せて読むことを推薦する，あるいは Web ページ X の前提として Web ページ Y の知識が必要であるといった関連を表現し，複数の情報資源を組み合わせた情報提示が可能になる．

４）相互利用性と長期保存性を考慮したメタデータ

メタデータの適切なデータ構造や記述形式は記述対象によって異なる．このためメタデータ記述フォーマットは多様化するが，関連する複数の Web ページを組み合わせて一つの情報資源として利用したり，メタデータを用いた複数 Web ページの横断検索を行ったりするためには，多様なフォーマット間でメタデータの相互利用性を高める必要がある．また，長期にわたって情報資源を利用するためには，いつ誰がどのメタデータ記述項目を付与したのかなど，その情報資源の管理・保存に関わる情報もメタデータとして付与されていることが望ましい．このため，メタデータ記述語彙・構文の違いを特別意識すること無く，相互に変換，マッピングが行われる必要がある．

５）メタデータ記述コストの軽減

メタデータを付与する際には，そのコストが可能な限り小さい方が望ましい．メタデータ作成のコストには，記述作業自体のコストの他，メタデータ記述フォーマットやデータモデルの学習にかかるコスト，メタデータの値として用いる語彙の調査と選択にかかるコストが大きな位置を占めている．これらの要因によるコストがメタデータにより得られる効果に見合わなければ，Web ページに対するメタデータ付与は行われないだろう．メタデータ付与作業のコストを軽減するために，簡易なメタデータ記述構文の用意，あるいはメタデータ作成を支援するツールなどの提供も行われていることが望ましい．

このように， Web ページの利用者によってメタデータが有効に用いられるためには，いくつかの機能要件を満たす必要がある．メタデータに求められる上記の機能要件を踏まえて，次節では，現在 Web ページに付与されているメタデータを例に挙げ，メタデータへの要求がどのように満たされているのかを述べる．

2.2 Web ページ中のメタデータの問題点

前節で述べたメタデータに期待される効果と特徴に対して，現在多く用いられている RDFa や Microformats，Microdata といったフォーマットに従って記述されるメタデータでは十分に対応できていない点がある．

Web ページ中にメタデータを記述する場合，メタデータ作成者が任意のタグを用いて，独自の構造でデータを埋め込むといった方法が考えられる．しかし，そうして記述されたメタデータは，メタデータ作成者とそのコミュニティ以外では利用が困難なデータとなってしまう．そのため，現在 Web ページ中にメタデータを埋め込む記述フォーマットとして，RDFa や Microformats，Microdata といった，標準的なフォーマットを用いることが推奨される．記述フォーマットを標準的なものに限定することで，Google などの Web 検索サービスでは，メタデータによって検索結果一覧の情報をより充実させるサービスを実現している．

しかしながら，標準的なメタデータ記述フォーマットを用いてメタデータ記述を行っている Web ページはまだ少ない．2008 年から 2010 年の調査[6][7]によると，RDFa によってメタデータが記述されている Web ページは 5%にも満たず，多くの Web ページで標準的なメタデータ記述フォーマットが用いられていない．2011 年 6 月に，Google や Yahoo，Microsoft など大手の検索サービスが，Microdata

によってメタデータを記述する際の指針を示す Schema.org[8]という取り組みを発表したことで，今後標準的なフォーマットで Web ページ中に記述されるメタデータが増加することが予想されるが，メタデータの付与をより促進するためには，メタデータ付与コストを軽減する手法やツールの提供が必要である．メタデータ作成コスト軽減のためには，過去の情報資源にメタデータを付与する作業の支援も欠かせない．また，すべての Web ページが Schema.org の指針のみに従うことは考えにくく，他のフォーマットで記述されるメタデータと併せて利用する，あるいはすべてのメタデータを Schema.org 以外のフォーマットで記述することもありうる．その場合，メタデータの記述フォーマットが多様になるため，そのすべてに対応したサービスやアプリケーションを作ることが困難となる．

一方で，各 Web サイトがページレイアウトのために HTML の class 属性や id 属性に記述している「mainColumn」や「keywords」といった独自の値をメタデータとして捉えることもできるが，属性値として用いる値の意味が明示的でなく，プログラムによって情報資源を横断的に利用することを困難にしている．

Web ページ中に記述されるメタデータが持つ上記の問題の中心は，メタデータ付与作業のコストに関する問題と，生成されるメタデータの相互利用性の問題として捉えることができ，これを解決することが重要である．

3. 情報抽出テンプレートによるメタデータの抽出と統合

メタデータ作成コストの軽減や相互利用性向上を実現するため，本研究では DCMI Description Set Profile[9]（以下，DSP）を基にした情報抽出テンプレートにより，相互利用性向上を意識したメタデータをより低コストで作成し利用するための手法を提案する．DSP を拡張して各記述項目の抽出ルールを定義することにより，メタデータの作成と利用を支援する．

3.1 DCMI Description Set Profile

Web ページのメタデータスキーマに沿って情報抽出を行うには，各 Web ページでメタデータスキーマが定義されていることが前提となるが，既存の Web ページでスキーマを明示しているものはごく限られた少数であり，大多数の Web ページではスキーマを持たない，あるいはスキーマが機械可読な情報として用意されていない．そのため，本研究ではまず Web ページのスキーマを明確にする必要がある．既存のメタデータスキーマや語彙を再利用して新たなメタデータスキーマを作成する手法としては，Dublin Core Application Profile[10]がある．その中では，メタデータの構造制約を記述する方法として DCMI Description Set Profile（以下 DSP）を決めている．DSP は，メタデータの記述対象が持つ記述項目とその出現回数，値の制約などを表現する記述方法であり，図 1 のように記述される．

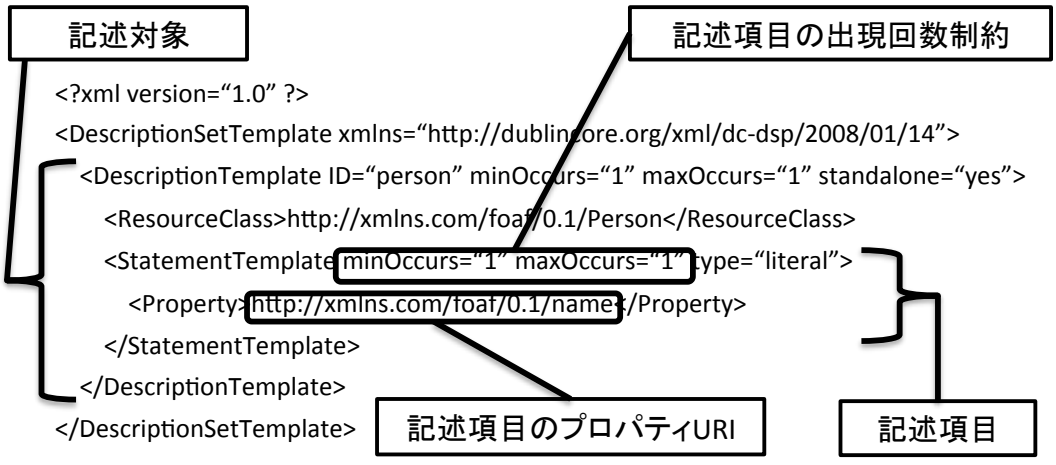

図 1 DCMI Description Set Profile の記述例（[9]より引用）

3.2 情報抽出テンプレート

本研究で提案する情報抽出テンプレートは，Web ページ中に記述されている情報をメタデータとして抽出するためのひな形である．DSP はメタデータスキーマが持つ記述項目名，プロパティ URI，値や出現回数の制約が定義可能となっている．これを拡張して，記述項目が Web ページ中のどの位置に出現するかを XPath 等で指定することにより，Web ページ中の任意の箇所をメタデータ記述項目として捉える．図 2 は DSP の記述項目情報にメタデータの出現位置を加えたものである．メタデータ出現位置の記述方法については，メタデータが非標準フォーマットで記述されている場合は XPath 等の記法でメタデータ出現位置を定義し，RDFa など標準的なフォーマットに従っている場合は，その属性名やプロパティ名，グラフパターンなどを用いることでメタデータとして抽出する値を指定する．このテンプレートにより，Web ページ中に書かれた情報を任意の記述項目のメタデータとして扱うことが可能になる．図 2 の(a)で名前空間接頭辞 dxl を用いて記述している箇所は，DSP で定義されておらず，本手法で新たに追加した記述である．この例は，Web ページ中”//title”など<dxl:value>内で記述された XPath の指し示す箇所に，dc:title として扱う情報が記述されていることを表している．

3.3 テンプレートの適用

情報抽出テンプレートを用いたメタデータの抽出・利用は以下の手順で行う．

1）メタデータ記述対象のメタデータスキーマの明確化と DSP の作成
当該 Web ページが持つ，タイトルや著者といった記述項目を列挙することで，どのようなメタデータが記述されているのかを明らかにし，Web ページの DSP を作成する．

2）情報抽出テンプレートの作成
(1)で挙げたメタデータ記述項目が Web ページ中のどの位置に出現するのかを XPath 等で表現し，DSP の記述に加えることで，情報抽出テンプレートを作成する．

3）テンプレートによる情報抽出
情報抽出テンプレートで定義された各記述項目の出現位置情報を利用して，Web ページ中の情報をメタデータとして取り出す．

4）抽出したメタデータの統合
取り出したメタデータを RDF グラフとして統合することで，異なるフォーマットで記述されたメタデータを一つにまとめる．

テンプレートによってメタデータとして定義された情報を取り出して利用する際は，各記述項目の出現位置情報を基に Web ページから値を抽出し，一つの RDF グラフとして出力する．多様なフォーマットで記述されるメタデータを RDF に統合して出力することで，メタデータ利用者はフォーマット毎の対応を行わずに，その RDF グラフの処理に専念することができる．出力する RDF トリプルの主語，述語，目的語は，それぞれ以下のように定義する．

1） 主語：メタデータ抽出対象となる Web ページの URI
2） 述語：情報抽出テンプレートによって定義された記述項目のプロパティ URI
3） 目的語：Web ページ中の，情報抽出テンプレートで定義された出現位置に記述されている情報

```
<StatementTemplate minOccurs="1" maxOccurs="infinity" type="literal">
  <rdfs:label>記事タイトル</rdfs:label>
  <Property>&dc;title</Property>
  <dxl:location ex:format="general_html">
    <dxl:value>//title</ex:value>
    <dxl:value>//div[@id='HeadLine']/h1</dxl:value>
    <dxl:value>//meta[@name='mixi-check-title' or @property='og:title']/@content</dxl:value>
  </dxl:location>
</StatementTemplate>
```

(a)

図 2 情報抽出テンプレート（一部）

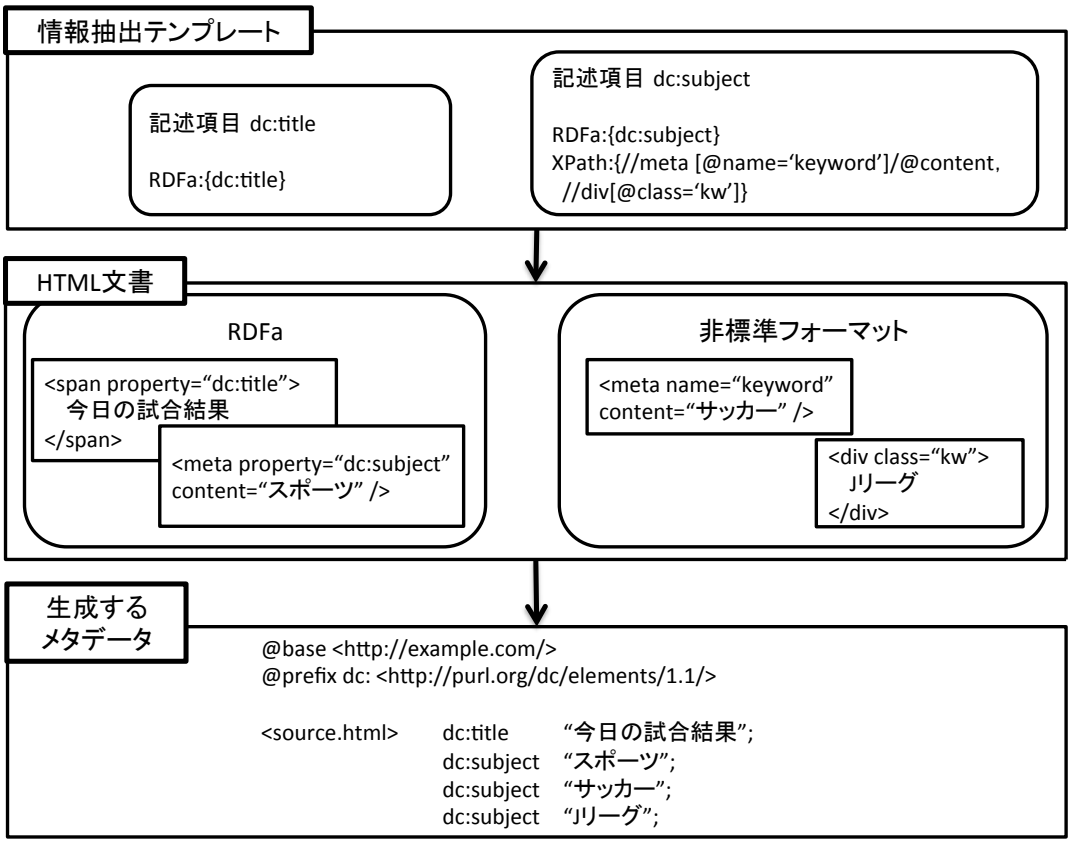

図 3　情報抽出テンプレートによるメタデータ生成

図 3 は，上記の方法で Web ページからメタデータを生成する流れを図示したものである．タイトルと主題の出現位置を RDFa のプロパティ名や XPath で指定すると，HTML 文書から該当箇所を抽出し，一つの RDF グラフとしてメタデータを統合・出力する．

4. システムの実現

本手法の情報抽出テンプレートを用いたメタデータ生成システムを構築した．本システムは登録されたテンプレートを利用し，Web ページ閲覧時にメタデータを生成・蓄積する．本システムは，Web プロキシ, メタデータインテグレータ, メタデータ抽出サーバ, RDF リポジトリという 4 つのモジュールから構成される．それぞれのモジュールは図 4 に示したように他のモジュールとデータの送受信を行い，メタデータを生成・蓄積する．

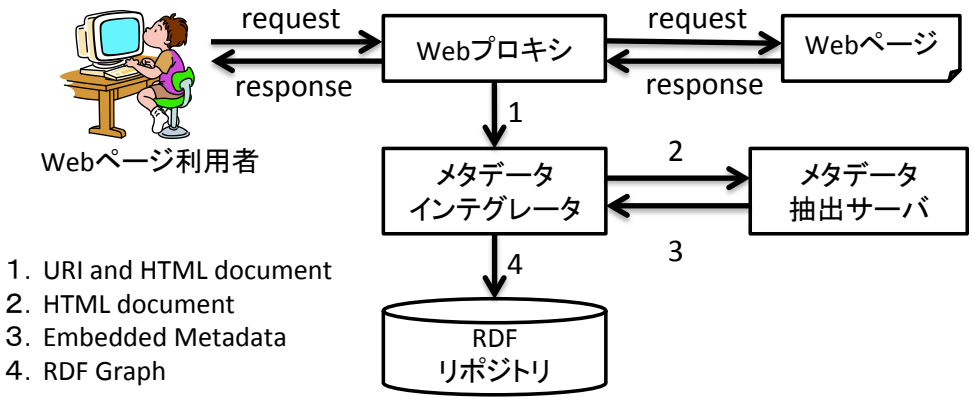

図 4　メタデータ生成システムの構成図

1）Web プロキシ:Web ページ閲覧時に, 当該 Web ページの URI と HTML document をメタデータインテグレータに送信する.
2）メタデータインテグレータ : Web プロキシから送信された Web ページ中のメタデータを RDF グラフとして統合して RDF リポジトリに送信する.
3）メタデータ抽出サーバ : メタデータインテグレータから HTML document を受け取った後，情報抽出テンプレートを利用して抽出したメタデータをメタデータインテグレータに送信する.
4）RDF リポジトリ：メタデータインテグレータから送信された RDF グラフを保存する.

5. 考察

本システムでは抽出対象とする標準的なフォーマットとして RDFa，Microformats，Microdata を扱っているが，Microformats は RDF グラフへの統合が困難であった．最終的な出力となるメタデータを RDF グラフとしたとき，RDFa など RDF をデータモデルとして採用しているフォーマットは特別な操作を必要としないが，Microformats のように key-value 型のデータモデルで記述されるメタデータを RDF グラフに統合する場合，メタデータの記述対象が明確でない，あるいは記述対象が URI を持たない文字列であることが問題となる．本研究では仮の記述対象を空ノードとして設けることで RDF グラフへの統合を行っているが，これによりメタデータ作成者の意図と異なるメタデータが生成される可能性がある.

また，情報抽出テンプレートの作成についても課題が残っている．本システム構築時に複数の Web サイトのメタデータスキーマを調査し情報抽出テンプレートを作成した際，各ニュース配信サイトに共通して表れるメタデータ記述項目が多いことが分かった．情報抽出テンプレートはそれぞれの Web サイトごとに作成しているが，各ニュース配信サイトに共通する記述項目を基にニュース用テンプレートを作成し，記述項目の出現位置情報を各 Web サイトにあわせて変更するといった，記述項目の共有によるテンプレート作成の仕組みが必要である.

6. 関連研究

非標準的なフォーマットで記述された情報を抽出する方法として，RDF グラフを出力するための XSLT スタイルシートを XHTML 中で指定する GRDDL[11]や，Ruby や Python など各種プログラミング言語で書かれたパーサを Web 上で共有する ScraperWiki[12]などが挙げられる．これらはメタデータスキーマの作成を行わずに Web ページ中からメタデータを抽出する方法であり，相互利用性向上のためにメタデータスキーマを情報抽出に利用する本研究のアプローチとは異なる.

7. おわりに

本稿では， DSP を基にした情報抽出テンプレートを用いることにより Web ページ中の非標準フォーマットを含む情報を抽出・統合してメタデータを生成する手法を提案した．これにより，Web ページへのメタデータ付与を促し，他の Web ページとの関連付けやメタデータによる横断検索を支援する.

今後の課題として，RDF 形式ではないメタデータを RDF 形式に変換・統合する手法の検討や，情報抽出テンプレートの共有基盤を作成する必要がある．また，本研究では，メタデータ相互利用性向上のために，異なる語彙・フォーマットで記述されたメタデータを情報抽出テンプレートにより一つの RDF グラフに統合した．その一方，各 Web ページで生成される RDF 間での相互利用性向上も重要である. メタデータスキーマの共有や語彙の関連付けを行うメタデータスキーマレジストリ[13]を用いて，情報抽出テンプレート間の記述項目間マッピングによる相互利用性向上なども今後の課題として挙げられる.

謝辞

本研究の一部は平成 23 年度日本学術振興会科学研究費補助金（課題番号:23500295）による.

参考文献

[1] RDFa in XHTML: Syntax and Processing. http://www.w3.org/TR/2008/REC-rdfa-syntax-20081014/（参照：2011-10-08）

[2] Microformats Wiki. http://microformats.org/wiki/Main_Page（参照：2011-10-08）

[3] HTML Microdata. http://dev.w3.org/html5/md/（参照：2011-10-08）

[4] 日本図書館情報学会研究委員会編. 図書館目録とメタデータ. 勉誠出版. 2004.

[5] DCMI Metadata Basics. http://dublincore.org/metadata-basics/（参照：2011-10-08）

[6] Mika, Peter. et al. Investigating the Semantic Gap through Query Log Analysis. Lecture Notes in Computer Science. 2009. vol. 5823. p. 441-455.

[7] Mika, Peter. "Microformts and RDFa deployment across the Web". Tripletalk. 2011-01-25. http://tripletalk.wordpress.com/2011/01/25/rdfa-deployment-across-the-web/（参照：2011-10-08）.

[8] Schema.org. http://schema.org（参照：2011-10-08）

[9] DCMI Description Set Profile. http://dublincore.org/architecturewiki/DescriptionSetProfile（参照：2011-10-08）

[10] The Singapore Framework for Dublin Core Application Profile. http://dublincore.org/documents/singapore-framework/（参照：2011-10-08）

[11] Gleaning Resource Descriptions from Dialects of Languages (GRDDL). http://www.w3.org/TR/grddl/（参照：2011-10-08）

[12] ScraperWiki. https://scraperwiki.com/（参照：2011-10-08）

[13] Mitsuharu Nagamori. et al. Meta-Bridge: A Development of Metadata Information Infrastructure in Japan. DC-2011. 2011.