メタデータスキーマと XPath を用いた HTML 文書からのメタデータ生成モデル*

本間維(学籍番号 201021754)
研究指導教員:杉本重雄
副研究指導教員:永森光晴

1. はじめに

WWW 上で公開されるリソースに他のリソースへの参照を表すメタデータを付与し,異なるリソースを組み合わせて利用する基盤を構築するために,Linked Data[1]と呼ばれる試みが注目されている.Linked Data による情報共有の基盤をより有用なものとするためには,多くの情報が公開され共有されることが重要である.既に多様な情報を公開する手段として用いられている HTML 文書から Linked Data に適した形式のメタデータを生成する手法を確立することが求められる.しかし,メタデータを用いた処理を行うために有用な,HTML 文書中に記述されたメタデータの解釈に必要なメタデータスキーマや,メタデータの抽出に必要なメタデータ記述項目の出現位置情報は十分に与えられていない.

本研究では,メタデータの記述項目を定義する Statement Template と,HTML 文書中でメタデータの値が出現する位置を指定する XPath 式を組み合わせることで,HTML 文書からのメタデータ抽出と解釈を支援する情報抽出テンプレートを提案した.

2. HTML 文書中のメタデータとメタデータスキーマ

メタデータは,メタデータの記述対象,メタデータのプロパティ,メタデータの値という3つの要素で構成される.メタデータはメタデータの記述に用いる語彙や規則を定めたメタデータスキーマに基づき作成される.メタデータスキーマは,記述されたメタデータを計算機が解釈する際に,どの値をどのようなプロパティと対応させて解釈するのかといったルールとして利用できる情報である.

計算機が HTML 文書中に表現されているメタデータを抽出し解釈するためには,1)値の出現位置:HTML 文書中のどこに記述されている値を抽出するのか,2)値の解釈に用いるルール:抽出した値がどのような規則に基づいて記述されているのか,といった定義が必要である.しかし,HTML 文書中のメタデータの抽出に必要な定義と解釈に必要な定義は関連づけられておらず,計算機がどの値にどのルールを適用するべきか明確でない.そこで本研究では,XPath とメタデータスキーマにより値の出現位置と値の解釈に用いるルールを定義し,それらを関連づける情報抽出テンプレートを提案した.

3. HTML 文書からのメタデータ生成モデル

3.1. 値の出現位置定義とメタデータ記述項目定義の作成

メタデータを HTML 文書中に記述するフォーマットとしては RDFa や Microdata などが標準規格として推奨されるが,2010 年の Mika らの調査[2]から,標準的なフォーマットの利用が進んでいないことが分かる.このため本研究では,HTML 文書中の要素を柔軟に指し示すことができる XPath 式を利用して,メタデータの値の出現位置を定義した.

メタデータの解釈に用いる定義には,Description Set Profile(DSP)[3]を利用した.DSP は,メタデータ記述規則の作成モデルである DCMI Application Profile において,メタデータの構造的制約などを計算機に解釈可能な形式で記述した定義である.DSP はメタデータの記

* "A Metadata Creation Model for HTML Document using Metadata Schema and XPath" by Tsunagu HONMA

述対象を定義する Description Template と，各記述対象が持つそれぞれの記述項目を定義する Statement Template で構成されている．DSP を用いて，当該 DSP に基づくメタデータが何を対象に記述されており，そしてどのような記述項目を持ちうるのかといった情報を計算機に与えることができる．

3.2. XPath 式と記述項目の対応付け

本研究では，計算機による HTML 文書からのメタデータ抽出とその解釈の支援を目的として，メタデータスキーマの定義と HTML 文書中のメタデータ記述位置の定義を対応付ける情報抽出テンプレートを提案した．図１はメタデータスキーマと HTML 文書中の値，そしてその間を結ぶ情報抽出テンプレートとの関係を表した図である．

HTML 文書中の値の指定は XPath を用い，記述項目の定義は Statement Template を用いている．本研究では，Statement Template に XPath を割り当てるための繰り返し可能なプロパティを設けることで，値と記述項目の定義を対応付けた．計算機が HTML 文書中のメタデータを抽出して解釈する際は，1)対象 HTML 文書に適用する情報抽出テンプレートを指定，2)情報抽出テンプレートで定義された XPath 式を利用し，汎用の XML 処理器で HTML 文書から値を抽出，3)抽出した値を，それぞれの値の抽出に利用した XPath と対応付けられた Statement Template に基づきメタデータ記述項目として解釈するといった手順を踏む．

情報抽出テンプレートの利用例として，Web ブラウザで閲覧した HTML 文書中のメタデータを

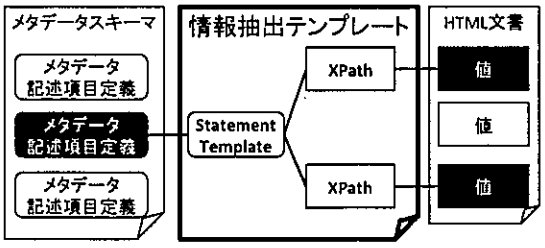

図 1 メタデータ記述項目定義と HTML 文書中の値の関連付け

Linked Data のための標準形式で蓄積するシステムの実装を行った．

4. 課題

本研究で提案したメタデータ生成手法には，1)情報抽出テンプレートの基となるメタデータスキーマの作成コストが大きい，2)HTML 文書から抽出される値が文字列であり外部へのリンクとなっていないといった問題がある．メタデータスキーマの生成コストについては，類似する既存のメタデータスキーマを再利用し，要求に合わせてカスタマイズすることで新たなメタデータスキーマを作成する手法が考えられる．HTML 文書から抽出される値を外部へのリンクとし，より Linked Data に適したメタデータを生成するためには，文字列として抽出される値を，URI で識別可能なリソースと対応付ける仕組みを構築することで，共通のリソースを参照するメタデータ間でのリンクを発生させることができる．

5. おわりに

本研究では，計算機が HTML 文書からメタデータを抽出し解釈するために必要な情報として，値の出現位置を指す XPath 式とメタデータ記述項目を定義する Statement Template を対応付けた情報抽出テンプレートを提案した．

情報抽出テンプレート作成支援環境を構築し，HTML 文書からのメタデータ生成をより容易にすることが必要である．また，HTML 文書から抽出される値をリソースとして扱うことで，より多くのメタデータを関連付ける仕組みが求められる．

文献

[1] Berners-L,T. Linked Data Design Issues. http://www.w3.org/DesignIssues/LinkedData.html

[2] Mika, P. The role of Linked Data in Search and Online Media. FIA Ghent 2010.

[3] Nisson,M. DCMI Description Set Profile. http://dublincore.org/architecturewiki/DescriptionSetProfile