

スキルギャップ解消によるメタデータ作成支援

矢代 寿寛[†] 松村 敦[‡] 宇陀 則彦[‡]

[†]総合研究大学院大学複合科学研究科情報学専攻 〒101-8430 東京都千代田区一ツ橋 2-1-2

[‡]筑波大学大学院図書館情報メディア研究科 〒305-8550 茨城県つくば市春日 1-2

E-mail: [†]yashiro@nii.ac.jp, [‡]{matsumur, uda}@slis.tsukuba.ac.jp

あらまし 汎用的なメタデータの作成には関連するスキルが要求されるが、作成者のスキルにギャップがあるため、メタデータが不均質となる。本研究では、メタデータ、オブジェクト、ドメイン、それぞれに関するスキルがギャップの要因であると仮定して、メタデータ作成支援のためのスキルギャップ解消手法について考察した。また、手法に基づくメタデータ作成支援システムについても言及する。

キーワード メタデータ作成支援, スキルギャップ

Metadata creation supported by filling the skills gap

Kazunori YASHIRO[†] Atsushi MATSUMURA[‡] and Norihiko UDA[‡]

[†]School of Multidisciplinary Science, Department of Informatics, Graduate University for Advanced Studies 2-1-2 Hitotsubashi, Chiyoda-ku, Tokyo, 101-8430 Japan

[‡]Graduate School of Library, Information and Media Studies, University of Tsukuba 1-2 Kasuga, Tsukuba-shi, Ibaraki, 305-8550 Japan

E-mail: [†]yashiro@nii.ac.jp, [‡]{matsumur, uda}@slis.tsukuba.ac.jp

Abstract Professional skills are required to design metadata schema that metadata creators can acquire common understanding. Skills gap between professionals and non-professionals decreases interoperability among metadata schemas and reliability of metadata collections. In this paper, we discuss methods filling the skills gap concerning metadata itself, object, and domain. In addition, we describe a supporting system for metadata creation.

Keyword Supporting Creation of Metadata, Skills Gap

1. はじめに

情報組織化技術の一つに、「data about data」と定義されるメタデータがある[1]。何らかの対象を記述していれば、書誌情報からレイティング、ペットボトルのラベルや名刺などもメタデータに含まれる。しかし、情報資源共有などで汎用的に利用するためには、記述の適切さが前提となるほか、記述規則や語彙を定義したスキーマに従って記述され、RDFのような機械可読形式である必要がある。また、スキーマには標準化された既存のものを利用して、相互運用性を確保することが望ましい。さらに、実装や運用を考慮して、デジタルアーカイブにおける OAIS 参照モデル (Open Archival Information System Reference Model)[2]のような指針も参考にすべきであろう。

このような狭義のメタデータを作成するためには多様なスキルが要求される。具体的には、ダブリンコア[3](Dublin Core: DC)等の国際標準規格をはじめと

したスキーマ、XML や RDF のような表現技術、記述対象である何らかのオブジェクト、オブジェクトのドメイン、それぞれに関する知識と技能である。本研究ではこれらを合わせてスキルと呼ぶ。

スキルが要求されることによって、メタデータの作成にはいくつかの問題が生じている。その一つが、メタデータの不均質さである。一般に、同じ分野の専門家同士であっても持っているスキルは様でない。スキルが様でなければ、作成されるものが不均質となることは明らかである。不均質なメタデータが流通することにより、個々のメタデータの相互運用性やコレクションとしてのメタデータの信頼性が損なわれる。

そこで本研究では、不均質さの要因は狭義のメタデータの作成に要求されるスキルの水準と作成者の持っているスキルのギャップにあると仮定し、このギャップを解消することで均質なメタデータの作成を支援する手法を提案する。

2. 関連研究

メタデータ作成支援に関連する既存研究としては、メタデータレジストリ、メタデータ・オントロジ検索エンジン、メタデータジェネレータ、メタデータエディタの4種が挙げられる。

メタデータレジストリは、既存のスキーマやマッピングを蓄積し、検索や参照ができるようにした「Authoritative Source」である。メタデータを蓄積する場合はクリアリングハウスと呼ばれる。

メタデータ・オントロジ検索エンジンは、一元的に管理することが困難な既存のスキーマやオントロジを、Web上のRDF文書から検索できるシステムである。登録された文書を中心に、レジストリ内のデータも検索することができる。代表的な例としてSwoogle[4]がある。

メタデータジェネレータは、主にWebページや動画を対象に、自然言語処理や画像認識処理、ファイルのヘッダ情報などを利用して、メタデータを自動的に生成するシステムである。既存のメタデータを別の形式に自動的に変換する場合もこれに含まれる。自動生成には技術的に限界があるため、人手で作成する場合に比べて信頼性に乏しいといわれる。しかし、主観的な評価ではそれほど差がないとするLiddyら[5]の報告もある。

メタデータエディタは、メタデータの作成、編集に特化したシステムのことである。テンプレートにメタデータを入力すると、XMLやRDFといった形式で提示または出力されるようになっている。タグ打ちの手間は省けるが、テンプレートとスキーマが1対1対応で固定されており、利用したいスキーマに応じてエディタを使い分ける必要がある。

メタデータジェネレータの多くにはメタデータエディタが組み込まれている。自動生成したメタデータの修正や、生成できない部分の入力にテンプレートを用いることで、手作業の場合に比べて作成コストが軽減でき、また自動生成だけの場合よりも信頼性の向上が見込まれる。コスト軽減の一例としては、DCを拡張したメタデータ1件当たりの作成に5分程度かかっていたものが、1分程度に短縮されたとする徳守ら[6]の報告がある。

3. メタデータ作成行動とスキルギャップ

3.1 メタデータ作成行動のモデル化

関連研究は作成それ自体のコスト軽減には有効であると考えられるが、スキルの個人差については考慮されていない。メタデータの不均質さを解決するには、スキルの個人差を考慮する必要がある。そこで、メタデータの作成にはどういったスキルが必要となるかを

明らかにするため、メタデータ作成行動のモデル化を行う。

図1にメタデータを新規作成する場合の段階的メタデータ作成行動モデルを示す。メタデータ作成行動は、オブジェクトが与えられた状態から、認識、探索、開発、適用、抽出、記述の順に推移する。第一段階ではオブジェクトの認識を行う。第二段階ではオブジェクトの記述に適切なスキーマを探査する。保存のためのメタデータを作成する場合、オブジェクト以外の周辺情報も記述可能なスキーマを探査する必要が生じる。予めスキーマが用意されている場合、または既存のスキーマを利用しない場合、この段階は省略される。この段階で適切な既存のスキーマを発見できなかった場合、第三段階としてスキーマの開発を行う。部分的に独自のスキーマを利用する場合もこの段階を経る。第二段階で適切なスキーマを発見あるいは第三段階で適切なスキーマを開発できた場合、第四段階としてスキーマをオブジェクトに適用する。ここでいう適用とは、既存のスキーマによくみられる曖昧な記述項目を、具体的なオブジェクトの情報と対応付ける事を指す。例えば『DCの「Creator」要素は図書というオブジェクトにおける著者という記述項目に対応する』といった解釈がこれにあたる。マッピングや読み替えなどもこの段階に含まれる。新たに開発したスキーマを利用する場合、このステップは省略される。この段階でスキーマがオブジェクトの記述に適切でない判断された場合、再度スキーマの探索や開発を全体的あるいは部分的に行う必要がある。第五段階として、記述するオブジェクトの情報およびその他の必要な情報の抽出を行う。予めオブジェクトの情報を抽出していた場合、このステップの一部または全部が省略される。最後に第六段階として、抽出した情報をメタデータとして機械可読形式で記述する。

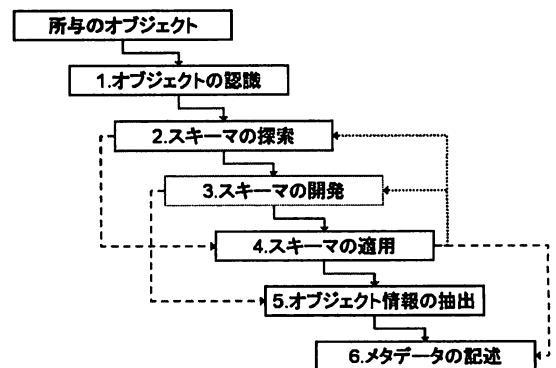


図1 段階的メタデータ作成行動モデル

3.2 スキルとギャップ

次に、モデルの各段階でどういったスキルが要求されるかということ、ギャップの存在がメタデータの不均質さにどう影響するかを整理する。

メタデータ作成行動の第一段階では、オブジェクトを適切に認識するために、オブジェクトとそのドメインに関する知識が必要となる。オブジェクトによっては特別な装置やそれを取り扱う技術も必要となる。ギャップが存在すると、適切でないスキーマが選択、開発される可能性があり、スキーマの適用段階で支障が生じる。第二段階では、ドメインに関する知識と、スキーマについての知識、情報探索技術が必要となる。ギャップが存在すると、既存のスキーマが発見できず、新たにスキーマの開発を行う必要が生じる。第三段階では、第二段階と同様に、ドメインとスキーマの知識が必要となる。また、ステークホルダーの間でスキーマについての合意が形成される必要がある。ギャップが存在すると、第一段階と同様に、スキーマの適用に支障が生じる。また、合意が得られない場合、メタデータの相互運用性が損なわれる。第四段階では、オブジェクト及びドメインについての知識と、スキーマの解釈が必要となる。例としてDCの「Coverage」要素の定義と解説を図2に示す。この情報から「Coverage」要素に適切なメタデータを記述することは可能であろうか。理解できない場合に、解説論文や書籍、ウェブサイト等を参考にしたりしても、その解説や説明について理解できる場合とできない場合がある。ギャップが存在すると、スキーマが適切に解釈されず、メタデータの記述が適切に行われない。また、不適切と解釈されてしまった場合、再度適切なスキーマを探索するか、開発しなければならない。第五段階では、オブジェクト、ドメインについての知識が必要となる。第一段階と同様に、オブジェクトによっては、特別な装置と取り扱い技術が必要となる。ギャップが存在すると、メタデータの記述が行えなくなる可能性がある。また、オブジェクトの持つ情報を損なう恐れがある。第六段階では、XMLやRDFといった表現形式の知識が必要となる。ギャップが存在すると、メタデータの相互運用性が損なわれる。

Label	Coverage [en-US]
Definition	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant. [en-US]
Description	Spatial topic may be a named place or a location specified by its geographic coordinates. Temporal period may be a named period, date, or date range. A jurisdiction may be a named administrative entity or a geographic place to which the resource applies. Recommended best practice is to use a controlled vocabulary such as the Thesaurus of Geographic Names [TGN]. Where appropriate, named places or time periods can be used in preference to numeric identifiers such as sets of coordinates or date ranges. [en-US]

図2 DCのCoverage要素

4. ギャップの解消手法

メタデータ作成行動のモデル化により、作業のどの段階でどういったスキルが必要となるかということ、ギャップの存在がメタデータ作成行動とメタデータに与える影響について整理した。これをもとに、メタデータ作成行動の段階に応じたギャップの解消を行う。ギャップの解消は、オブジェクトのプロファイル作成、言い換えによる読解支援、スキーマ候補の推薦、入力テンプレートの生成の4手法を組み合わせで行う。それぞれの手法と、メタデータの作成行動のどの段階のギャップを解消するかについて述べる。

4.1 オブジェクトのプロファイル作成

既存のメタデータを参考にして、予めオブジェクトとそのドメインごとに適切とされるスキーマを全て収集してDBに格納した。これをオブジェクトプロファイルと呼ぶ。

オブジェクトプロファイルを作成する理由は、メタデータ作成行動において、スキーマの探索結果がその後の作業に大きな影響を与えるからである。例として、DCはネットワーク上の情報資源の発見を目的とし、IEEE-LOM[7]は教育用コンテンツの記述を目的としている。このように、既存の標準規格化されたスキーマには目的や記述の対象となるオブジェクトが予め想定されている。そのため、オブジェクトが決定すれば必然的に適切な既存スキーマは絞られてくる。しかし、現在メタデータの規格は多数存在するため、画像などでは、Z39.87、Dublin Core、IEEE-LOM、METSなどの中から、状況や目的に応じて適切なものを適宜選択する必要がある。これがメタデータ作成行動でスキーマの探索が必要な理由であり、ギャップの原因でもある。この行動を、オブジェクトプロファイルを作成することで不要とする。行動がなければスキルも必要とならないため、ギャップは発生しない。

4.2 言い換えによるスキーマ読解支援

オブジェクトプロフィールに対して、想定されるメタデータ作成者のスキルのレベルに合わせた言い換え(Paraphrase)を行なった。これにより、オブジェクトプロフィールは、収集した本来の記述要素の表現に加え、レベルごとの表現も持つことになる。言い換えとは、意味がほぼ同じ別の語句への変更であり、読解支援に用いられる。メタデータ関連の研究においては、多様な記述が可能な解説文などに対して行われていた[8]が、本研究では記述要素に行く。その理由はメタデータ作成行動の推移に読解が大きく影響しているからである。言い換えの基準とするのは、適切さである。この場合の適切さには、オブジェクトとそのドメインにおけるコンテキストと、メタデータ作成者のコンテキストの二つの観点がある。例えばDCの「Creator」という記述項目はオブジェクトが写真であるというコンテキストでは、その写真を撮影または編集した人間を指し、「撮影者」などと表現される。そして、メタデータ作成者が小学生であるというコンテキストでは、「撮影者」というよりも「写真をとった人」と表現した方がより適切と思われる。これら二つの観点からの適切さを保証するための規格として、オブジェクトとドメインのコンテキストには文化資源情報共有研究でのマッピング事例[9]を、メタデータ作成者のコンテキストには国語辞典と日本語能力試験の出題基準を用いる。適切な言い換えにより、スキーマの解釈に必要なスキルの要求水準がメタデータ作成者のスキルに近づくため、メタデータ作成行動のスキーマの適用におけるギャップが解消される。

4.3 スキーマ候補の推薦

情報推薦の一種である内容ベース方式[10]を参考に、DBに格納されたオブジェクトプロフィールをスキーマ候補として推薦する。推薦は、言い換えで用いた二つのコンテキストに合わせてオブジェクトプロフィールを重みづけし、メタデータ作成者のスキルに応じたオブジェクトプロフィールだけを取り出して行う。推薦するのは、メタデータとして記述すべき項目と、その繰り返しの可否や推奨度、項目ごとの使用可能語彙、規定値である。推薦したスキーマ候補は、メタデータ作成者が必要に応じて修正できる。

スキーマ候補の推薦により、メタデータ作成行動のスキーマ開発におけるスキルの要求水準が下がり、ギャップが解消される。

4.4 入力テンプレートの生成

推薦したスキーマを元に、メタデータの入力テンプレートを生成する。テンプレートの入力項目の表現は、

オブジェクトとドメイン及び作成者のコンテキストに適切な形に言い換えられたものとなる。入力項目の数や語彙は、作成者の決定したものとなる。つまり、先の述べた3手法を組み合わせたものがテンプレートとなる。テンプレート例を図3に示す。このテンプレートに従って入力したメタデータは、実際には図4に示すような、本来の記述項目で保存されるため、相互運用性が保たれる。これにより、メタデータ作成行動のメタデータの記述において、XML形式のタグを打ち込むといったメタデータの記述作業の一部が不要となることで、関連するスキルが不要となり、ギャップが発生しなくなる。また、オブジェクトとドメインに適切な記述をするためのスキルの要求水準が、言い換えと推薦によって作成者のスキルに近づくため、その間のギャップが解消される。スキルが異なる複数のメタデータ作成者に、一つのスキーマを各々のコンテキストに適切な表現で利用してもらうことで、作成者のスキルが一律でないにもかかわらず、均質なメタデータを作成することができる。

項目	メタデータ
タイトル	学校周辺の秋の草花 その5 オニアザミ
撮影者	阿部真男
キーワード	オニアザミ
キーワード	学校周辺の草花
キーワード	秋の草花
キーワード	総合的な学習の時間
写っているもの	オニアザミ
公開者	総研町立総研小学校
協力者	総研写真館
撮影日	2007年09月02日
ファイルの種類	Image
ファイルのフォーマット	Image/jpeg
写真のID番号	25
撮影場所	総研鎮龍研町
著作権者	総研町立総研小学校

出力形式を選んでください

図3 画像メタデータ入力テンプレート例



図 4 画像メタデータ XML 形式出力例

5. メタデータ作成支援システム

5.1 システムの全体像

考察したギャップ解消手法を組み合わせ、メタデータ作成支援システムとして構築した。システムは、メタデータ作成行動モデルに沿った流れで、対話的な操作により既存のスキーマを核とする独自スキーマの開発と、それに基づくメタデータの作成が行えるように意図されている。

図 5 に本システムの構成を示す。図中央の実線で囲んだ部分がメタデータ作成支援とスキーマ開発支援のためのシステム全体像である。Web ブラウザから利用する Web アプリケーション形式をとっている。関連研究で述べたメタデータ自動生成、メタデータ・オントロジ検索、メタデータレジストリとの連携によって、作成コストの軽減も試みる。今回はギャップ解消手法の有効性を検証するため、システム全体像の左側に位置する破線部分を構築した。

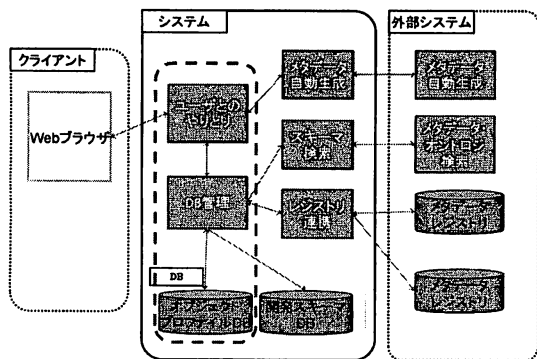


図 5 システムの全体像と完成部分

5.2 検証用システム

検証用システムは、オブジェクトプロファイルを格納した DB とスキーマを元にテンプレートの生成を行うサーバ部の 2 つから成っている。オブジェクトプロファイルを最新の状態に保つためのレジストリとの連携部分は省いている。検証用システムでは、図 1 に示したメタデータ作成行動モデルに沿って対話的に操作を行うことで、スキーマの簡易的な開発と、開発したスキーマを利用したメタデータ作成が行えるようになっている。

具体的な利用の流れは、まずオブジェクトと要求する記述の水準、メタデータの利用目的を選択する。それにより、スキーマの候補が提示される。現時点では、公式レジストリの存在や関連ドキュメントの豊富さからスキーマを DC のみとし、電子的なファイルに限定してオブジェクトプロファイルを作成し、DB に格納している。各記述項目には、ギャップ解消手法である作成者のメタデータに対する知識レベルに応じて言い換えた表現と、スキーマ候補の推薦のためのエレメントのクラスや推奨度、既存のスキーマとの関係なども併せて記録している。これらを判断材料としてスキーマ候補として提示している。具体例としては、簡潔な記述を望むメタデータ作成者に対しては SimpleDC の Element のみを候補として提示する。このときの記述項目は、最も平易な表現となる。提示された表現が不適切と感じられた場合、作成者は自身のスキーマのレベルを変更することで、より適切な表現に修正することができる。スキーマ候補は、記述項目名、記述項目の繰り返し、推奨度、利用可能語彙、デフォルト値を自由に修正することができる。このスキーマ候補をもとに、メタデータの入力テンプレートを生成し、表示する。生成したテンプレートを利用して、XML、RDF、CSV の各形式でメタデータを出力することができる。メタデータを入力する際の記述項目名は、作成者に応じた表現または自由に修正されたものであるが、出力される際には DC 本来の記述項目となる。

6. おわりに

メタデータの不均質さを解決するため、スキーマのギャップ解消によるメタデータ作成支援を提案した。メタデータ作成行動モデルをモデル化し、メタデータの作成に必要なスキルと、ギャップがメタデータに及ぼす影響を整理した。これを元にギャップの解消手法を考察した。ギャップの解消手法は、オブジェクトのプロファイル作成、言い換えによる読解支援、スキーマ候補の推薦、入力テンプレートの生成の 4 手法である。ギャップ解消手法によって、メタデータの作成に要求されるスキルの一部を不要とすることができ、また一

部のスキルの要求水準を低下することができる。ギャップの解消手法を組み合わせ、段階的メタデータ作成行動モデルに対応したメタデータ作成支援システムを構築した。

今後の課題は、システムの評価とシステムを使用して作成したメタデータの評価である。現時点では、ギャップ解消に対するギャップ解消手法の有効性と、メタデータの不均質さを解決する手法としてのギャップ解消の有効性が明らかでないため、システムユーザや専門家のレビューによる評価が必要である。

文 献

- [1] 杉本重雄, “XML とメタデータ-メタデータの基本概念”, 情報知識学会誌, vol.13, no. 4, pp. 16-23, Oct. 2003.
- [2] 栗山正光, “OAIS参照モデルと保存メタデータ”, 情報の科学と技術, vol.54, no. 9, pp. 461-466, Sept. 2004.
- [3] Dublin Core, “Dublin Core Metadata Initiative (DCMI)”, <http://dublincore.org/>
- [4] Swoogle, “Swoogle Semantic Web Search Engine”, <http://swoogle.umbc.edu/>
- [5] E. D. Liddy, E. Allen, S. Harwell, S. Corieri, O. Yilmazel, N. E. Ozgencil, A. Diekema, N. McCracken, J. Silverstein, and S. Sutton, “Automatic metadata generation & evaluation”. Proc. 25th Annual International ACM SIGIR Conf. on Research and Development in Information Retrieval., pp.401-402, New York, U.S., Aug. 2002.
- [6] 徳守淳也, 小野智司, 木場隆司, 北山信一, 中山茂, “コンパイラ・コンパイラを用いたメタデータ作成支援システムの開発”, 情報知識学会誌, vol.17, no.1, pp. 41-50, Feb. 2007.
- [7] IEEE Learning Object Metadata, “IEEE LTSC | WG12”, <http://ltsc.ieee.org/wg12/>
- [8] K. Aihara, T. Yamada, N. Kando, S. Fujisawa, Y. Uehara, T. Baba, S. Nagata, T. Tojo, T. Awaji, and J. Adachi, “Owlery: A Flexible Content Management System for “Growing Metadata” of Cultural Heritage Objects and Its Educational Use in the CEAX Project”, the 9th International Conference on Asian Digital Libraries (ICADL 2006), vol.4312, pp.22-31, Kyoto, Japan, Nov. 2006.
- [9] 山本泰則, 中川隆, “データベース横断検索のための民族学標本資料情報の Dublin Core による記述”, 情処研報 (CH), vol. 2005, no. 76, pp. 47-54, June. 2005.
- [10] 石川徹也, 宇田隆幸, “情報フィルタリングの利用システム:情報推薦システム”, 情報の科学と技術, vol. 56, no. 10, pp. 458-463, Oct. 2006.