複数のメタデータスキーマ・マッピングの組み合わせは妥当なマッピングを導くか

谷口 祥一（慶應義塾大学文学部）
taniguchi@z2.keio.jp

メタデータスキーマ間の多様なマッピング（アラインメント、クロスウォーク）が現在策定されているが、その策定作業には多大な人的労力が必要になる。本研究では、複数のマッピングの機械的かつ比較的単純な組み合わせから妥当な第 3 のマッピングが導かれるのかを、実例をもって検証を試みる。図書館目録のメタデータに関わり公開されているマッピングから選択し、それぞれ規模が異なる 3 つの事例について検証を行った。その結果、マッピングとして妥当なペアとそれ以外の両者が含まれていることを確認したが、総体としての有効性には限界がある。

１．はじめに

メタデータスキーマのマッピング（mapping）とは、通常、ある語彙（実体と、その属性・関連のエレメントと値の集合）から他の語彙への対応づけである。アラインメント（alignment）、クロスウォーク（crosswalk）とも呼ばれるが、本研究ではこれらを区別しない。メタデータの機械的変換に使用するものから、意味的な関連性を確認するものまで、マッピングには多様性がある。現在、スキーマ間の多様なマッピングが策定されているが、その策定作業には多大な人的労力が必要になる。

マッピングに関わる課題はこれまでにも複数指摘されており、マッピングの評価にも、一貫性（consistency；一貫したマッピングとなっているか）、包括性（completeness；含まれる語彙を網羅しているか）、互換性（compatibility；構造、フォーマット、意味の点で適切なマッピングとなっているか）、文脈上の関連性（contextual relevance；特定のコンテクストなどによる制約を満たしたものとされているか）など、複数の観点がありうる。

本研究では、図書館目録のメタデータに関わり、それぞれ独立して策定された複数のスキーマ・マッピングの機械的かつ比較的単純な組み合わせから、妥当な第 3 のマッピングが導かれるのか、あるいは人手によるマッピング策定作業の軽減という支援となりうるのか、実例をもって検証を試みる。

本研究では、以下の手順で検証する。

①公開されているマッピング（セット）から、マッピングの意図や推移性の有無、語彙の粒度など、その特性を確認した上で採用する。語彙が依拠する概念モデルなどの、メタデータの構造的要素については本研究では捨象し、属性・関連のエレメントに該当する部分のみ対象とする。よって、エレメントの定義域や値域、値自体となる実体（クラス）や概念のマッピングは、付随的に含まれる場合以外には対象としない。また、RDF の適用を特には前提としない。

②共通した語彙を中継点にして、最小限の人手による前処理を加えた後にマッピング間の機械的な照合を実行し、新たなマッピングを機械的に生成する。

③生成されたマッピングについて、各種の集計を行うとともに、目視によりその妥当性を確認する。ただし、正解を準備した上での定量的な評価などは行わない。

２．事例１（小規模事例）

RDA Steering Committee が策定し RDA Registry[1)]において公開しているマッピングおよびアラインメント（v5.0.13）から、「RDA → IFLA LRM (LRM)」と「RDA → Dublin Core (DCT)」を採用し、方向性をもたないマッピング「LRM ←→ DCT」を生成した。

２．１　マッピング「RDA → LRM」と「RDA → DCT」

マッピング「RDA → LRM」は、「title proper (rdam:P30156) - rdfs:subPropertyOf - lrmer:R13 (has appellation)」といった意味的包含関係にある（すなわち推移性をもつ）マッピングであり、3,024 の RDA エレメントから 103 の LRM エレメントへのマッピングとしている（マッピングペア数は 3,024）。

また、マッピング「RDA → DCT」も「title proper (rdam:P30156) - rdfs:subPropertyOf - dct:title」といった包含関係にあるマッピングであり、1,145 の RDA エレメントから 33 の DCT エレメントへのマッピングとしている（マッピングペア数 1,147）。

２．２　マッピングの組み合わせ結果「LRM ←→ DCT」

RDA を起点としたマッピングの組み合わせを単純に生成すると 1,147 ペアのマッピングとなり、そこから RDA エレメントを除去し、さらに重複となるペアを削除した結果、42 の LRM エレメントと 33 の DCT エレメントの組

み合わせからなる 87 ペアが残った。表 1 にその一部を示す。

「LRM → DCT」のマッピングとして見た場合、エレメントの意味的包含関係にあるかを人手により確認したところ、「lrmer:E1A1 (has category of res) → dct:type」や「lrmer:E2A1 (has category of work) → dct:type」など、41 ペアは妥当と判断された。また、意味的な関連性の観点から見たときには、「lrmer:E1A2 (has note) → dct:description」など、10 ペアをさらに追加して認めることができる。ここには、マッピング先の複数個を足し合わせたときに、マッピング元の意味を復元可能なものを含む。他方、「lrmer:E1A2 (has note) → dct:abstract」など 36 ペアは、直接的な意味上の関連を認めがたい。

逆に「DCT → LRM」のマッピングとして見たときには、「dct:abstract → lrmer:E1A2 (has note)」など、26 ペアが意味的包含関係にあると考えられる。

表 1　事例 1 により生成されたマッピング「LRM ←→ DCT」（一部）

LRM	DCT
lrmer:E1A1 (has category of res)	dct:type
lrmer:E3A1 (has category of expression)	dct:format
lrmer:E3A1 (has category of expression)	dct:type
lrmer:E3A2 (has extent of expression)	dct:extent
lrmer:E3A2 (has extent of expression)	dct:format
lrmer:E3A3 (has intended audience of expression)	dct:audience
lrmer:E3A6 (has language of expression)	dct:language

３．事例２（大規模事例）

RDA Steering Committee が公開しているマッピング「RDA → MARC21 Bibliographic」と、LC が公開している「MARC21 Bibliographic → BIBFRAME」[2]を組み合わせ、「RDA → BIBFRAME」という方向性をもつマッピングを生成した。

３．１　マッピング「RDA → MARC21 Bibliographic」

アラインメント「title proper / unstructured description / aligns with / 245 ** $a」とマッピング「rdam:P30156 - rdakit:hasM21 - 245 ** $a [unstructured description] . 」の両形式で同じ内容が公開されている。「unstructured description」は、RDA による 4 つの記録方法のうち、「非構造記述」であることを表す。また、意味的包含関係に該当しない対応づけもあるため、独自のプロパティ「rdakit:hasM21」が用いられている。よって、上記では title proper (rdam:P30156)の非構造記述の値は、MARC21 フィールド 245 (Title Statement)、サブフィールド a (Title) に対応づけることを指示している(第 1･2 インディケータの値「*」は任意の値を表す)。

なお、示されたマッピングは意味的な関連をもつものを幅広く網羅したものであり、例えば title of work (rdaw:P10088) は、非構造記述の場合、「100 ** $t」・「110 ** $t」・「111 ** $t」・「130 ** $a」・「245 ** $a, c, p」など、多数のマッピング先が示されている。さらには、記録の方法が識別子の場合と IRI (URI)の場合が加わり、それぞれについても複数のマッピング先が示され、結果的に多数のマッピングペアが形成されている。

すべての RDA エレメントに対する「rdaw:P10088 - rdakit:hasM21 - 245 ** $a, c, p [unstructured description] . 」といった単位によるマッピングでは、合計 17,799 行となった(サブフィールド単位に分割したときには、51,367 行)。ここには MARC 固定長フィールドへのマッピングも含まれている。

RDA エレメント数（異なり数）は 1,606、記録の方法と組み合わせたときには 5,638 であった。一方、出現した MARC エレメント数（前記の単位による異なり数）は 1,918 である。

３．２　マッピング「MARC21 Bibliographic → BIBFRAME」

MARC 固定長フィールド (00X) およびレコードリーダからのマッピング、可変長フィールドからのマッピングがそれぞれ Excel による表形式といくつかの補足説明文書として、これらに基づき実装した変換ツールとともに公開されている。フィールドごと、インディケータごと、サブフィールドごとにそれぞれ変換先が指示されており、複数の変換処理（最大 3 つ）が示されている場合もある。また、変換に際しての各種の条件や複雑な処理内容が付記されている場合もある。

Excel データから機械的にマッピングを生成できるよう最小限の手作業による前処理を加えた後、プログラムによりマッピングペアを生成した。指示がある変換の付帯的な条件などは捨象した場合も多い。その結果、可変長フィールドについては、フィールド数 181、「フィールド＋インディケータ＋サブフィールド」数

は 3,067 となった。なお、インディケータについては、値が任意を表す「*」の場合も追加してマッピングを生成した。

例えば、「245 ** $a」は、BIBFRAME の「I - title - Title - mainTitle - literal (remove trailing =, : or / punctuation)」と「W - title - Title - mainTitle - literal (remove trailing : or / punctuation)」の 2 つの記述にマッピングされている。ここで「I」・「W」と「Title」はそれぞれクラス Instance・Work と Title を指し、title と mainTitle はプロパティを表す。Turtle に沿った記載とすれば、「URI リソース bf:title [a bf:Title ; bf:mainTitle "リテラル"] . 」（URI リソースは bf:Instance または bf:Work のインスタンス）となる。

同様に、固定長フィールドについては、例えば「007/03 electronic resource」（フィールド 007 のポジション 00 によって electronic resource が指定されている場合の、ポジション 03）は、「W - colorContent - ColorContent」にマッピングされる。これらはフィールド数 5 とそれにレコードリーダ 1 が加わり、上記のようなマッピングペアが総計 156 となった。

３．３　マッピングの組み合わせ結果「RDA → BIBFRAME」

前記の 2 セットのマッピングを組み合わせ、新たなマッピング「RDA → BIBFRAME」を生成した。RDA エレメント数は 1,576 であり、記録の方法と組み合わせたときには 3,895、出現した BIBFRAME 記述（前記した単位によるもの）数は 540、これらの間のマッピングペアの合計は 23,202 となった。表 2 にその一部を示した。

title proper の非構造記述は、前記した 2 つの BIBFRAME 記述へとそのままのマッピングとなったが、例えば title of work は非構造記述が 10 の BIBFRAME 記述、識別子による記述が 5 つの BIBFRAME 記述へとマッピングされる結果となった。最大のマッピングペアは、related work of work の構造記述の場合（128 ペア）であり、続いて related work of expression、related work of manifestation、related work of item がいずれも 114 ペアを形成した。マッピング先の BIBFRAME 記述の側では、「## - contribution - Contribution – [agent - Agent - [rdfs:label - literal ; identifiedBy - Identifier]; role - Role] 」が 2,761 回、「I - note - Note - rdfs:label」が 2,012 回それぞれ出現している。

４．事例３（中規模事例）

RDA Steering Committee が公開しているマッピング「RDA → MARC21 Authority」と、LC-BIBFRAME-Wikidata-Project[3] による「MARC21 Authority → Wikidata」を組み合わせ、「RDA → Wikidata」を生成した。

４．１　マッピング「RDA → MARC21 Authority」と「MARC21 Authority → Wikidata」

マッピング「RDA → MARC21 Authority」は、MARC21 Bibliographic と同様、意味的関連性を有するものを幅広く網羅したマッピングであり、アラインメント「access point for agent / structured description / aligns with / 100 ** $a, b, c, d, g, q」などとしている（マッピングも同内容）。これにより、MARC フィールド 100 (Heading--Personal Name)のサブフィールド a (Personal name)、b (Numeration) などにマッピングされることが示されている。

一部の MARC21 Bibliographic へのマッピングも含まれており、「RDA エレメント＋記録の方法」単位のマッピング行数 16,298、サブフィールド単位に分割したときにはマッピング行数 56,445 となった。

一方、「MARC21 Authority → Wikidata」のマッピングは、プロジェクトの参考用に作成されたものであり、網羅的でも規範的でもないと記されている。可変長フィールド 0XX、1XX、3XX、4XX、6XX のみ取り上げており、「フィールド、第 1 インディケータ、サブフィールドとラベル」から「Wikidata プロパティとラベル」へのマッピングとしている。例えば、「100 0/1 $a personal name → P742 (pseudonym)」、「100 0 $a personal name (name in direct order) → P735 (given name)」、「100 1 $a personal name (surname) → P734 (family name)」などである。

マッピング行数 250 であり、出現した MARC フィールド数 33、「MARC フィールド＋インディケータ＋サブフィールド」数 124、また出現した Wikidata プロパティ数 110 であった。

４．２　マッピングの組み合わせ結果「RDA → Wikidata」

前記 2 セットのマッピングを組み合わせ、新たなマッピング「RDA → Wikidata」を機械的に生成した。その結果、マッピング行数 264 となり、出現した RDA エレメント数 138、「RDA エレメント＋記録の方法」の数 184 であった。表 3 にその一部を示した。RDA エレメントと単純に 1 対 1 対応で Wikidata プロパティが存

在する場合には適切なマッピングであるが、1対多である場合には容易には判断がつかない結果となった。

注・参照文献

1) RDA Registry. http://www.rdaregistry.info/
2) LC. MARC21 to BIBFRAME 2.0 Conversion Specifications. https://www.loc.gov/bibframe/mtbf/
3) LC-BIBFRAME-Wikidata-Project. https://www.wikidata.org/wiki/Wikidata:WikiProject_PCC_Wikidata_Pilot/LC-BIBFRAME-Wikidata-Project

表 2　事例 2 により生成されたマッピング「RDA → BIBFRAME」（一部）

RDA	RDA 記録の方法	フィールド＋サブフィールド ラベル	BIBFRAME 記述
title proper	unstructured description	TITLE STATEMENT -- Title	I - title - Title - mainTitle - literal (remove trailing =, : or / punctuation)
title proper	unstructured description	TITLE STATEMENT -- Title	W - title - Title - mainTitle - literal (remove trailing : or / punctuation)
title of work	unstructured description	COLLECTIVE UNIFORM TITLE -- Uniform title	W - title - CollectiveTitle - mainTitle - literal
title of work	unstructured description	TITLE STATEMENT -- Name of part/section of a work	I - title - Title - partName - literal
title of work	unstructured description	TITLE STATEMENT -- Name of part/section of a work	W - title - Title - partName - literal
title of work	unstructured description	TITLE STATEMENT -- Statement of responsibility, etc.	I - title - Title . I - responsibilityStatement-literal
title of work	unstructured description	TITLE STATEMENT -- Title	I - title - Title - mainTitle - literal (remove trailing =, : or / punctuation)
title of work	unstructured description	TITLE STATEMENT -- Title	W - title - Title - mainTitle - literal (remove trailing : or / punctuation)
title of work	unstructured description	Uniform titles -- Uniform title	Work - expressionOf - Hub - [contribution - Contribution - agent - Agent ; title - Title - mainTitle - combine all subfields (concatenated)]
title of work	unstructured description	Uniform titles -- Uniform title	Work - expressionOf - Hub - title - Title - mainTitle - combine all subfields (concatenated)
title of work	unstructured description	VARYING FORM OF TITLE -- Title proper/short title	I - title - VariantTitle - mainTitle - literal
title of work	unstructured description	VARYING FORM OF TITLE -- Title proper/short title	W - title - VariantTitle - mainTitle - literal

表 3　事例 3 により生成されたマッピング「RDA → Wikidata」（一部）

RDA	RDA 記録の方法	フィールドラベル	Wikidata プロパティ
access point for agent (rdaa:P50373)	structured description	Corporate Name	P1813 (short name) ; P276 (location) ; P585 (point in time)
access point for agent (rdaa:P50373)	structured description	Meeting Name	P1545 (series ordinal) ; P276 (location) ; P478 (volume) ; P585 (point in time) ; P958 (section, verse, paragraph, or clause)
access point for agent (rdaa:P50373)	structured description	Personal Name	P1035 (honorific suffix) ; P112 (founded by) ; P1317 (floruit) ; P1635 (religious name) ; P2031 (work period (start)) ; P2032 (work period (end)) ; P31 (instance of) ; P410 (military rank) ; P5056 (patronym or matronym for this person) ; P511 (honorific prefix) ; P512 (academic degree) ; P569 (date of birth) ; P570 (date of death) ; P7338 (regnal ordinal) ; P734 (family name) ; P735 (given name) ; P742 (pseudonym) ; P97 (noble title) ; P1449 (nickname) ; P1477 (birth name) ; P1559 (name in native language) ; P1813 (short name)
address of agent (rdaa:P50418)	unstructured description	Address	P131 (located in the administrative territorial entity) ; P17 (country) ; P281 (postal code) ; P6375 (street address) ; P968 (e-mail address)