RDA から MARC21 への複雑なマッピングを読み解く
－効果的な表示法の検討とネットワーク分析の適用－

谷口 祥一（元慶應義塾大学）
s.taniguchi@keio.jp

最新の RDA エレメントから MARC21 書誌・典拠フォーマットへのマッピングは、規模が大きく複雑であり、全体の把握や理解は困難な状況にある。本研究は、このマッピングに対して、a)基本的な集計と分析、b)エレメント階層の適用と、インディケータとサブフィールドの包含関係によるマッピング先 MARC フィールドの包含関係を適用した効果的な表示法の提示、c)ネットワーク分析の適用によるグラフ分割の試行などによって、特徴把握が一定程度実行できた。

１．はじめに

RDA 運営委員会が策定した、最新の RDA エレメントから MARC21 書誌・典拠フォーマットへのマッピングが、RDA Registry 上で公開されている [1]。入念に作成されたものとはいえ、マッピング全体の規模が大きく、かつ複雑であり、全体の把握や理解は困難と言わざるをえない。RDA のエレメント数の膨大さ、包括的なマッピング設定すなわち多数のマッピング先の提示、さらには依拠する概念モデルの相違を含むスキーマの構造的相違などが複合して、マッピング全体の複雑度は極めて高い。また、マッピングの策定方針などは明示されておらず、メタデータ変換ツールへの実装を意図したものとも考えにくい。

本研究は、この複雑なマッピングに対して、適切な集計と分析、効果的な表示法（可視化）の提示、ネットワーク分析の適用などによって、マッピング全体を読み解くことを試みる。

RDA エディタである RIMMF は、3R プロジェクト以後の最新版 RDA 語彙に対応したバージョン 4 では、MARC21 レコードからの読み込みなど、マッピングに基づく変換機能は実装されていない。また、RDA のマッピングを直接取り上げた研究は殆どないが、発表者は以前に複数のマッピングを組み合わせることの妥当性検証事例において、RDA から MARC21 へのマッピングを試用した [2]。

２．基本的な集計と分布

最新の RDA 語彙（v5.1.0）において定義されたマッピングを取り上げる。それゆえ、MARC フィールドへのマッピングが定義されている RDA エレメントのみ対象とする。また、MARC21 書誌フォーマットと典拠フォーマットへのマッピングを統合して扱い、固定長フィールドと可変長フィールドの両者へのマッピングを含めたものとする。

RDA Registry では、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]」として同じ内容が公開されており、本研究では「マッピング」と総称する。先の事例は、RDA エレメント title proper (rdam:P30156)の「非構造記述」の値は、MARC21 書誌フォーマットのフィールド 245 (Title Statement)、サブフィールド a (Title) に対応づけることを表している（第 1・2 インディケータの「*」は任意の値を表す）。

RDA では、「非構造記述（unstructured description）」以外に、「構造記述（structured description）」、「識別子（identifier）」、「IRI」（URI と基本的に同義）という、合計 4 つの記録方法（recording method）があり、本研究では「エレメント＋記録方法」をマッピングの単位とし、以降はこれを便宜的に「RDA エレメント」と呼ぶ。

同様に、マッピング先 MARC 可変長フィールドは、第 1・2 インディケータとサブフィールドを組み合わせたものを、以降では便宜的に「MARC フィールド」と呼ぶ。MARC21 書誌の場合は接頭辞「B」、典拠の場合には「A」を付加して区別する。そこでは、「B245 ** $a, c, p」や「B710 ** $a, b, c, d, f, g, k, l, n, p, t, u (if subfield $t is used)」といったマッピング先も登場する。他方、「B245 ** $7(dpecou)」、「B245 ** $7 (dpermw) (data provenance)」など、data provenance サブフィールドと名づけられた値の由来の記録に該当するマッピングは、本研究の対象外としてすべて除外した。

先ず、マッピングデータの整合性をプログラムで機械的に点検し、明らかな誤りや不整合と見られる箇所については手作業で修正した。

基本的な集計結果は、以下の通りである。

a）合計 34,228 ペア（MARC 書誌 19,104、典拠 15,124）のマッピングデータとなった。出

現した RDA エレメントの異なり数 8,258（記録方法を組み合わせない、エレメントのみの場合は 2,481）、MARC フィールド異なり数 1,259 であった。

b) RDA エレメントごとに平均 4.1（SD 2.8）ペア（すなわちマッピング先 MARC フィールド数）をもち、最大は rdae:P20298 [identifier] (related work of expression)の 30、続いて rdaw:P10198 [identifier] (related work of work)、rdaw:P10219 [structured description] (date of work)、rdaw:P10256 [identifier] (subject)の 29 ペアとなった。分布の偏りを示す歪度が 1.83、ジニ係数が 0.34 であった。

c) MARC フィールドごとに平均 27.2（SD 111.7）ペア（すなわち RDA エレメント数）が出現し、最大は「B500 ** $a」の 1,498、続いて「A510 ** $0」と「A510 ** $1」の 1,139 ペアであった。歪度 6.75、ジニ係数 0.89 という、偏りの大きな分布である。

d) マッピングのカーディナリティ（基数）は、それぞれ排他的に分けると、「1 対 1」が 58、「1 対多」は RDA エレメント数 50 で MARC フィールド数 184、「多対 1」は RDA エレメント数 86 で MARC フィールド数 18 となり、これらはそれぞれ 58、50、18 個の部分グラフ（連結成分）を構成している。残りはすべて「多対多」のマッピングとなり、RDA エレメント数 8,064、MARC フィールド数 995、そして部分グラフ数 39 であった。よって、部分グラフの合計数は 165 となる。

1 対多や多対多に属する、単一 RDA エレメントから複数 MARC フィールドへのマッピングは、その該当事例を見ると、マッピング先が「論理和」の場合と「排他的論理和」の場合とが混在している。前者は単一の値をマッピング先の複数フィールドに同時に記録（変換）することを許容し、後者は提示されたフィールドのいずれか 1 つに記録する扱いに該当する。いずれにしても、可能性のある MARC フィールドを幅広く網羅した結果であり、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おり、複雑と言えよう。

３．効果的な表示法の検討

RDA エレメントは階層構造をもつ。以前は「エレメント・サブタイプ」と名付けられていた関係であり、rdfs:subPropertyOf によって定義される。なお、以前にあった「サブエレメント」の関係は最新 RDA では採用されていない。

この階層関係に依拠した表示法が有効と考える。ただし、単一エレメントが複数の上位エレメントをもつ場合もあり、単純な構成ではない。これは定義域とするクラスの階層（agent – collective agent – corporate body など）に基づくエレメントの階層と、例えば行為主体の役割に基づく階層（creator – author – screenwriter など）とが組み合わされるからである。

３．１　マッピングの階層表示

RDA エレメントの階層関係を適用すると、階層を構成するエレメント数 7,904（ペア数 32,962）、階層に位置づけられず孤立したエレメント数 354（ペア数 843）となった。階層の最上位となるエレメント 228 と孤立エレメント 354 を足した 582 が、相互に独立した階層木の数を示す。階層木ごとの平均出現行数 28.6（SD 109.1）、最大 849、平均階層レベル数 2.0（SD 1.7）、最大 9 であった。

単一エレメントが複数の上位エレメントをもちうることから、場合によっては階層木内での同一エレメントの複数回出現が数多く発生し、その結果、階層木を巨大化させている。同一エレメントが階層内で最大 28 回出現するエレメントが 42 個も見られた。

階層表示において、2 回目以降の出現かつ下位エレメント群をもつ場合には、これら下位エレメント群の表示を省略し、表示行数を減らし全体的な見通しをよくする方式も併せて開発した。ただし、副作用もあり、いずれが適切な表示法であるのか、にわかには判断できない。

３．２　マッピングの階層・包含表示

RDA エレメントの階層表示に加えて、マッピング先の MARC フィールドにおける包含関係を適用した表示法が有効と考える。

a) インディケータの包含関係：値「1」や「#」は、値「*」に合致し包含されるものと見なす。これにより、「B264 *1 $a」は「B264 ** $a」に包含される。この処理に加えて、単一エレメントからのマッピング先のインディケータ値が複数ある場合には、統合化した表示とする（例：「B264 *(1|2|3) $a」）。なかには、単一エレメントからのマッピング先自体にも包含関係が見られる場合があり（例：「B600 (*|1)* $1」）、意図されたものか不明であるが、そのまま残し採用した。

b) サブフィールドの包含関係：サブフィールドが複数指定されているものとその部分集合は包含関係にあると見なす。例えば、「A510 **

$a, b, c, d, g」に「A510 ** $a, e」は包含されると見なす。なお、単一エレメントからのマッピング先を個別サブフィールドごとに独立させている場合（「A111 1* $a」と「A111 1* $c」）と、統合して単一マッピング先としている場合とがあり、フィールドごとの使い分けがあるものと推測されるが、本研究では便宜的に一律に統合化して扱うことにした。

以上のRDAエレメント階層関係とマッピング先MARCフィールド包含関係とを組み合わせた表示法を開発した。表示全体は巨大となり、図1と2にその部分を示した。左側から階層レベル、RDAエレメント、直上位エレメントとのマッピング先包含関係、マッピング先MARCフィールド群の順に示す構成とした。

MARCフィールド群の表示は、最上位エレメントのマッピング先に列を揃えた配置としたが、図1はスペースの制約でそれに準じた表示としてある。また、直上位エレメントのマッピング先と照合し、それに含まれないMARCフィールドには、記号「+」を前置した表示としてある。

表示中の「直上位エレメントとのマッピング先包含関係」欄は、当該エレメントのマッピング先（通常、複数）が直上位エレメントのマッピング先にすべて包含されているかを判定した結果を表している。

・値「0」: 最上位レベルのエレメント。582回出現
・値「1」: 直上位エレメントのマッピング先にすべて一致（インディケータとサブフィールドによる包含を含む）。20,460
・値「2」: 直上位エレメントのマッピング先にすべて包含。9,682
・値「3」: 直上位エレメントのマッピング先に部分的に包含、すなわち一部は逸脱。2,502
・値「4」: 直上位エレメントのマッピング先からすべて逸脱。90

判定結果「3」と「4」の場合が少なからず発生することから、単純に直上位エレメントのマッピングに包含されるわけではない、つまりエレメント階層に従ったマッピングとは言えないことが分かる。

４．ネットワーク分析の適用

MARCへのマッピングを介してRDAの構造上の特徴を把握する、あるいは逆にRDAからのマッピングを介してMARCの構造上の特徴を把握する目的で、マッピングの全体を2部グラフと捉え、それぞれRDAへのプロジェクションとMARCへのプロジェクションを取り、単一モードのRDAグラフとMARCグラフに変換した。ここでは、MARCフィールドの統合化処理前のマッピングを使用している。

得られたRDAグラフはノード数8,258、エッジ数3,750,312、一方MARCグラフはノード数1,259、エッジ数6,961であり、いずれも165の部分グラフを構成している。RDAグラフの密度は0.110、平均クラスタ係数0.891、推移性（transitivity）0.908であり、一方MARCグラフは密度0.009、平均クラスタ係数0.704、推移性0.683となった。

４．１　RDAグラフの分割

RDAグラフを適切に分割してその特徴を探るために、複数の手法でグラフを分割し、分割結果の類似度を求めた。いずれの分割が妥当か正解がない状況での次善の策である。

上記のRDAグラフ（①）は、部分グラフ数165、平均ノード数50.1（SD 366.5）、歪度8.21、ジニ係数0.96であり、部分グラフのノード数3,723、2,108、2,029が突出しており、これら以外は17以下であった。

このグラフに対して、ネットワーク分析におけるコミュニティ検出手法のうち、②Louvainアルゴリズム（代表的なモジュラリティ最適化手法）、③ラベル伝播（label propagation）アルゴリズム、④非同期ラベル伝播（asynchronous label propagation）アルゴリズムをそれぞれ適用して分割を行った。形成された部分グラフ数は、それぞれ②は181、③197、④201となった。元のグラフにさらなる分割を加えた結果、部分グラフのノード数の偏りは多少とも緩和されている。表1に個々のグラフ分割結果の特徴量を整理した。

併せて、⑤エレメント階層関係による部分グラフ（部分グラフ数582）、さらには元のRDA部分グラフとエレメント階層による部分グラフを併合したグラフ（⑥;同115）を準備した。⑤はグラフの併合により、部分グラフのノード数の偏りが増大している。

これらのグラフ分割結果の類似度を、ARI（Adjusted Rand Index）とNMI（Normalized Mutual Information：正規化相互情報量）によって求めた。その結果を表2にまとめてある。エレメント階層による分割⑤が、元のRDAグラフに各種コミュニティ抽出法を適用した結果②～④に、比較的類似した結果となった。

４．２　MARCグラフの分割

MARCグラフについても、同様なグラフ分

割を行い、①元の MARC 部分グラフ（部分グラフ数 165）に加えて、②Louvain アルゴリズム（同 181）、③ラベル伝播アルゴリズム（同 193）、④非同期ラベル伝播アルゴリズム（同 236）のそれぞれによる分割結果を得た。表 1 に個々のグラフ分割結果の特徴量を、表 2 にこれら分割結果の類似度 ARI と NMI を示した。形成された部分グラフを個別に見ていくと、ある程度の意味的まとまりが見えてくるが、それ以上の結論づけは困難と言えよう。

引用文献

1) RDA Registry. https://www.rdaregistry.info/
2) 谷口祥一. 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. 第 71 回日本図書館情報学会研究大会発表論文集. 2023, p.25-28.

hierarchy level	RDA element	inclusion	MARC21 fields				
1	rdam:P30134 title of manifestation [unstructured description]	0	B245 ** $a	B500 ** $a			
2	rdam:P30128 variant title of manifestation	4	+B246 ** $a	+B740 ** $a			
2	rdam:P30131 abbreviated title	3	B500 ** $a	+B210 ** $a			
2	rdam:P30156 title proper	2	B245 ** $a				
3	rdam:P30203 parallel title proper	4	+B246 ** $a	+B245 ** $b			
4	rdam:P30204 parallel title of series	4	+B490 ** $a				
2	rdam:P30157 title of series	3	B500 ** $a	+B490 ** $a			
3	rdam:P30204 parallel title of series	2	B490 ** $a				

図 1　マッピングの階層・包含表示．1

（注：直上位エレメントのマッピング先に含まれない MARC21 フィールドには、記号「+」を前置）

hierarchy level	RDA element	inclusion	MARC21 fields				
1	rdaa:P50305 [IRI] related work of agent	0	A500 ** $1	A510 ** $1	A511 ** $1	A530 ** $1	A672 ** $1
2	rdaa:P50130 [IRI] issuing agent of	1	A500 ** $1	A510 ** $1	A511 ** $1	A530 ** $1	A672 ** $1
3	rdaa:P50546 [IRI] issuing person of	2	A500 ** $1	A510 ** $1		A530 ** $1	A672 ** $1
3	rdaa:P50697 [IRI] issuing collective agent of	2	A500 ** $1	A510 ** $1		A530 ** $1	A672 ** $1
4	rdaa:P50848 [IRI] issuing corporate body of	1	A500 ** $1	A510 ** $1		A530 ** $1	A672 ** $1
4	rdaa:P50999 [IRI] issuing family of	1	A500 ** $1	A510 ** $1		A530 ** $1	A672 ** $1
2	rdaa:P50132 [IRI] dedicator agent of	1	A500 ** $1	A510 ** $1	A511 ** $1	A530 ** $1	A672 ** $1
3	rdaa:P50566 [IRI] dedicator person of	1	A500 ** $1	A510 ** $1	A511 ** $1	A530 ** $1	A672 ** $1
3	rdaa:P50717 [IRI] dedicator collective agent of	1	A500 ** $1	A510 ** $1	A511 ** $1	A530 ** $1	A672 ** $1
4	rdaa:P50868 [IRI] dedicator corporate body of	1	A500 ** $1	A510 ** $1	A511 ** $1	A530 ** $1	A672 ** $1
4	rdaa:P51019 [IRI] dedicator family of	1	A500 ** $1	A510 ** $1	A511 ** $1	A530 ** $1	A672 ** $1

図 2　マッピングの階層・包含表示．2

表 1　プロジェクション後の RDA/MARC グラフの各種分割結果

	部分グラフ数	平均ノード数（SD）	歪度	ジニ係数	部分グラフのノード数（上位 3 つ）
①RDA:ノード連結	165	50.1（SD 366.4）	8.21	0.96	3,723、2,108、2,029
②Louvain	181	45.6（SD 218.0）	6.01	0.93	1,761、1,557、1,116
③ラベル伝播	197	41.9（SD 236.8）	6.81	0.94	1,968、1,839、1,566
④非同期ラベル伝播	201	41.1（SD 223.5）	6.70	0.93	1,761、1,780、1,536
⑤RDA エレメント階層	582	14.2　（SD 50.5）	4.71	0.86	300 が 5 個
⑥ノード連結＋エレメント階層	115	71.8（SD 579.9）	9.26	0.97	5,892、2,053、18
①MARC:ノード連結	165	7.6（SD 40.5）	10.29	0.79	480、185、87
②Louvain	181	7.0（SD 17.7）	5.90	0.72	164、116、79
③ラベル伝播	193	6.5（SD 16.4）	5.87	0.70	148、115、76
④非同期ラベル伝播	236	5.3　（SD 9.0）	4.03	0.61	66、60、49

表 2　各種分割結果の類似度

	②Louvain	③ラベル伝播	④非同期ラベル伝播	⑤RDA エレメント階層
①RDA:ノード連結: ARI / NMI	.161 / .526	.163 / .516	.157 / .507	.032 / .312
⑤RDA:エレメント階層: ARI / NMI	.242 / .619	.239 / .627	.248 / .636	
⑥RDA:ノード連結＋エレメント階層	.158 / .466	.153 / .464	.166 / .463	.071 / .390
①MARC:ノード連結: ARI / NMI	.285 / .812	.262 / .793	.104 / .747	