生成 AI を用いたメタデータマッピングの効率的作成法
RDA-BIBFRAME マッピングの作成試行

谷口 祥一　（元・慶應義塾大学文学部）
s.taniguchi@keio.jp

メタデータスキーマのマッピング作成における生成 AI 活用の有効性検証を目的にして、最新 RDA エレメントから BIBFRAME へのマッピング作成を ChatGPT を用いて試行し評価した。体現形・著作・表現形エレメント群を対象に、それらの定義データに加えて、階層表現、「意味的な分解表現」を追加した上で、マッピング先を示した一定数の訓練用データを与えた評価実験とした。併せて、全件マッピング先同定済みデータを用いた層化 k 交差検証を追加して実行し評価した。

１．はじめに

メタデータスキーマのマッピング作成と管理には多大な労力を通常必要とする。本研究は、汎用的な生成 AI である ChatGPT 5 をマッピング作業支援ツールとしてどのように活用することができ、どの程度の性能を示すのかを実験により検証することを目的とする。取り上げる事例は、最新 RDA エレメントから BIBFRAME（以下、BF）へのマッピングという、両者間に大きなギャップがある事例とし、human-in-the-loop による効率的なマッピング作成支援の構図を意図した。

最新 RDA は「オフィシャル RDA」と呼ばれ、2020 年末に完了した 3R プロジェクトによって、それ以前の RDA（オリジナル RDA）から大幅に変更されており、数年後には欧米等において移行が予定されている。他方、BF は 2016 年にバージョン 2 が公開されてからも微調整が続いており、その後に実体 Hub が導入されるなど最終形は未だ見えない状況にある。

これら両者間のマッピングとして、現在、公認されたものはない。RDA 運営委員会は、最新 RDA から MARC21 や DCMI メタデータ語彙などへのマッピングを策定し公開しているが、BF へのマッピングは対象とされていない。代わって、BF へのマッピングをフィンランド国立図書館は 2023 年時点で試作し公開したが、マッピング先の BF 記述が粗く、大まかなマッピングと言わざるをえない [1]。また、複数のマッピング先 BF 記述があるものは除外されている。他方、オリジナル RDA から BF へのマッピングはワシントン大学図書館による 2020 年時点のものがあるが、その後の更新はなされていない [2]。

２．マッピング作成の試行実験
２．１　対象データと前処理

実験には 2025 年 7 月時点の最新 RDA 語彙 v5.4.5[3] と BF オントロジーv2.6.0[4] の定義データを用いた。RDA エレメントは、体現形を定義域とする 409 個、著作を定義域とする 608 個、表現形を定義域とする 516 個を採用した。これら以外の個別資料や Agent（行為主体）を定義域とするエレメントは対象にしていない。

個別エレメントの定義データには、URI、エレメント名であるラベル、定義文、直上位エレメントが示されている。

その後の試行を見据えて、各エレメントの意味と位置づけの理解を容易にするために、エレメントの上位下位関係に基づき階層的に整列させた。ただし、最大 2 つの上位エレメントをもつ多重階層であるため、同一エレメントの複数回出現が伴う。また、上位エレメントのうち、定義域が実体 RDA entity であるものは階層木に加えていない。

さらに、最上位エレメントからの経路を順に並べて表した「階層表現」を定義データに追加した。例えば、「related entity of work > related RDA entity of work > related agent of work > creator agent of work > artist agent > artist collective agent > artist corporate body > calligrapher corporate body」となる。同一エレメントでも出現箇所に依存して、この階層表現は異なるものとなる。

他方、BF オントロジーは RDF+OWL を用いて定義されたクラスとプロパティからなる。

２．２　実験 1：RDA エレメントの「意味分解表現」の付与

個々の RDA エレメントの意味と位置づけを明瞭に表すものとして、新たにエレメントの「意味分解表現」を追加した。エレメント calligrapher corporate body の意味分解表現は「related agent - creator agent - corporate

body - calligrapher」、エレメント preferred title of work のそれは「related nomen - title of work - preferred title of work」のようにした。こうした意味分解表現の方式は複数考えられ、ここではそのうちの一つを採用した。厳密な構成規定等は設けておらず、他の表現方式を含めて詳細な検討は行っていない。

実験では、階層的に整列したエレメント群の一定範囲に対して、人手で意味分解表現を付与した。ある範囲のエレメント群に適用されるであろうパターンをもつ部分を選んで人手で意味分解表現を付与した。つまり、ランダムに付与対象エレメントを選択することをしていない。これは ChatGPT による付与実験において、訓練用（事前学習用）の正解データとして用いる、すなわち自動付与処理において適用可能な事例が一定程度存在し、それによって有効に機能することを意図したからである。

実験 1 では表 1 に示した通り、先ず人手で順次、体現形エレメント 409 のうち 213 個（階層表示では 224 行）、著作エレメント 608 のうち 133 個（149 行）、表現形エレメント 516 のうち 63 個（63 行）に意味分解表現を付与した。それらを訓練用正解データとし、ChatGPT にその作成ルールを導出させ、かつ未付与分のエレメントに対して意味分解表現を付与させた。

その付与結果を人手で確認したところ、体現形で未付与の 196 エレメントのうち 52 が正しい表現であった（正解率 0.265）。同様に、著作の 475 エレメントのうち 232、表現形の 453 エレメントのうち 222 がそれぞれ正しい表現であった（正解率 0.488 と 0.490）。正解率に幅があるとはいえ、ChatGPT による意味分解表現の付与は比較的良好な性能を示したと考えられる。なお、著作エレメントの実験、表現形エレメントの実験では、それぞれ先行して処理した体現形と著作のエレメント意味分解表現全件を追加データとして参照させている。

自動付与した結果の誤りデータはすべて人手で修正し、以降の実験に用いた。後の BF 記述の付与実験において、この意味分解表現を参照対象データに含める場合と含めない場合との相違を調べた。

２．３　実験 2：マッピング先 BF 記述の付与

1) ステップ 1：RDA エレメントのマッピング先 BF 記述を人手で一定数のエレメントに対して付与し、訓練用データを準備した。ランダムに対象エレメントを選ぶのではなく、一定程度の適用が期待されるパターンをもつ部分を選んでおり、ChatGPT による付与処理が有効に機能することを意図した。体現形では 226 個（全体の 55.3%）、著作では 133（21.9%）、表現形では 111（27.4%）のエレメントに対する正解付与とした（表 2）。

マッピング先 BF 記述の同定においては、フィンランド国立図書館とワシントン大学図書館によるものの両者を参考にしたが、それらとは異なるマッピング先とした事例もある。BF の説明文書などは不十分であり、曖昧さが残る段階での判断とした。例えば、RDA エレメント calligrapher corporate body のマッピング先は「W - contribution - [Contribution ; agent - Organization ; role - rdf:resource= "http://id.loc.gov/vocabulary/relators/cll"]」、エレメント preferred title of work のマッピング先は「W - title - Title」とした。これらは Turtle の略記形式を採用しており、最初の事例は「bf:Work bf:contribution [a bf:Contribution ; bf:agent bf:Organization ; bf:role rdf:resource= "http://id.loc. gov/vocabulary/relators/cll"]. 」を略記したものであり、Relator の MARC コード（"http://id.loc.gov/vocabulary/relators/"）を組み合わせて使用している。

なお、体現形間・著作間・表現形間という同一実体間での関連、あるいはこれら異種実体間の関連については、プロパティ relatedTo とその下位プロパティのみマッピング先として採用した（「W - relatedTo - W」、「W - hasDerivative - W」など）。クラス Relationship に属する MARC コード値（"http://id.loc.gov/vocabulary/relationship/"）を採用して「W - relation - Relation - relationship - rdf:resource="http://id.loc.gov/vocabulary/relationship/***"]」のようにすることも可能と見られるが、BF プロパティとして定義済みのものとの持ち分けなどが不明なため、今回のマッピング先には採用していない。

2) ステップ 2：人手で付与した BF 記述を訓練用データとし、併せてエレメントのラベル、定義文、階層表現、先に付与した意味分解表現、そして BF オントロジーであるクラスとプロパティの定義ファイル、Relator の MARC コードファイルを ChatGPT に入力し、そこから Relator マッチング手順書を生成させた。これは上述の事例のように「I/W - contribution - [Contribution ; role - Role]」（I は Instance）となるマッピング先において、クラス Role に該当する MARC コード値をいかに同定するかという手順をまとめたものである。意味分解表

現の箇所を活用した規則が導出されているバージョンと、意味分解表現を用いないバージョンとを生成させた。

同様に、著作エレメントについてのみ、プロパティ subject が用いられるパターンを同定する Subject マッチング手順書を導出させた。

3) ステップ 3：BF 記述未付与分の RDA エレメントに対して、ChatGPT による BF 記述付与を実験した。実験は、2-①先に付与した RDA エレメントの「意味分解表現」とそれを活用したマッチング手順書を使用した場合と、2-②意味分解表現を参照せずに BF 記述付与を試行した場合との両者を行った。それぞれの実験において、ChatGPT に順位を付けて 3 つまで候補を回答させた。最終的には、付与されたものを人手で点検し、必要な修正を行った。

実験 2-①：「意味分解表現」とそれに対応する Relator のマッチング手順書、さらに著作エレメントの場合には Subject マッチング手順書を入力している。体現形・著作・表現形エレメントの順に実行し、著作の実験では先行して同定した体現形エレメントの全件正解データを、表現形の実験では著作エレメントの全件正解データを、それぞれ訓練用データとして追加入力している。

実験結果は、体現形エレメントでは評価用データである 183 エレメントのうち 84(45.9%)、著作は 475 エレメントのうち 175（36.8%)、表現形は 405 エレメントのうち 188（46.4%）が正解であった（表 3)。また、正解となったエレメントのマッピング先 BF 記述パターンを集計したところ、その殆どはパターン「I/W - contribution - Contribution」に該当するもの、すなわち実体 Agent との関連に属するエレメントであった。体現形エレメントでは正解 84 個のすべてがこれに該当し、著作では正解 175 のうち 171、表現形では正解 188 のうち 186 がこれに該当した（表 4)。

一方、著作エレメントにおける Subject マッピング手順書は結果的に殆ど機能せず、正解数の増加に寄与しなかった。表 4 に示した「正解ーそれ以外」の 4 件がこの Subject パターンに該当したが、同じ著作エレメントにおいても、Subject に該当する 23 件が不正解であった。

なお、ChatGPT からの第 2 候補および第 3 候補の回答は、いずれのエレメントにおいてもすべて不正解であった。

実験 2-②：「意味分解表現」を参照せず、併せて Relator マッチング手順書もそれに合ったものを採用して実験した。実験結果は、いずれのエレメント群においてもかなり低い正解率となった（表 3)。

ChatGPT による第 2 および第 3 候補の回答は、体現形エレメントにおける 3 件の正解にとどまった。表 3 ではこの 3 件を便宜的に正解数に含めて示した。

以上の実験 2-①および②の結果を得たが、ChatGPT を含めて生成 AI による回答は、プロンプトの微細な調整、あるいは同一プロンプトの再実行によっても、その回答が変化しうるという特性がある。本研究で今回得られた結果も安定したものとは言えないことは留意しなければならない。

３．交差検証による性能評価実験（実験 3）

先の実験によりマッピング先 BF 記述の同定と確認が完了した 3 つの RDA エレメント群に対して、再度、BF 記述付与の性能評価実験を、層化抽出 k 分割交差検証によって実行した。ここではエレメントの階層的な整列は用いず、エレメント数に該当するデータとした。

2 つの方法による層化抽出を行っており、1 つめはマッピング先 BF 記述のパターンに依拠した 4 分割を行い、2 つめの方法ではマッピング先 BF 記述を参照させず、残りの要素に基づき 4 分割を行った。それぞれの層化抽出後に、訓練用・テスト用のデータを順次切り替え、4 回の独立した BF 記述付与実験を行った。実験は、前述の実験 2-①に相当する、「意味分解表現」とそれを活用した Relator マッチング手順書・Subject マッチング手順書を用いた実験のみとした。ただし、先の実験と異なり、著作エレメントの実験において体現形エレメントの全件正解データを、表現形エレメントの実験において著作エレメントの全件正解データを訓練用データとして追加入力することはしていない。

実験結果は、交差検証のマイクロ平均正解率と、その正解率の Wilson 95%信頼区間をもって表 5 に示した。層化抽出法 1 による実験では、体現形エレメントの正解率 0.210、著作エレメント 0.321、表現形エレメント 0.343 であった。なお、ChatGPT の第 2 または第 3 候補の回答が正解であった事例も、ここでは正解に含めている。結果的には大まかに言って、BF 記述のうちどの程度が「I/W - contribution - Contribution」のパターンに該当するかが性能を決定している模様である。

層化抽出法 2 による実験結果は、体現形エレメントの正解率 0.249、著作エレメント 0.326、

表現形エレメント 0.192 であった。表現形エレメントを除いては、層化抽出法 1 による結果と同程度の正解率であった。ただし、十分に有効とは言い切れない性能値である。

引用文献

1) Finland National Library. *RDA_BIBFRAME_mapping_2022_2023_FI*. https://docs.google.com/spreadsheets/d/16JWdhQcFMybVQtEsWV1RIiOh2URl1JT0/edit?usp=sharing&ouid=107462956795239912593&rtpof=true&sd=true
2) University of Washington Libraries. *RDA/RDF to BIBFRAME Mapping*. 2020-11-04. https://lib.uw.edu/wp-content/uploads/cams_swr_rda-rdf-to-bibframe-mapping.pdf
3) *RDA Registry*. https://www.rdaregistry.info/. v5.4.5 の定義データは、次の URL でアクセス可能：https://github.com/RDARegistry/RDA-Vocabularies/releases/tag/v5.4.5
4) Library of Congress. *BIBFRAME Ontology*. v2.6.0. https://id.loc.gov/ontologies/bibframe.rdf

表 1　実験 1：RDA エレメントの意味分解表現の付与実験結果

		訓練用データ		テスト用データ					
	エレメント数	エレメント数	%	エレメント数	%	正解数	正解率	不正解数	不正解率
体現形エレメント	409	213	52.1	196	47.9	52	.265	144	.735
著作	608	133	21.9	475	78.1	232	.488	243	.512
表現形	516	63	12.2	453	87.8	222	.490	231	.510

表 2　実験 2：マッピング先 BF 記述の付与実験

			訓練用データ			テスト用データ		
	エレメント数	行数	エレメント数	%	行数	エレメント数	%	行数
体現形エレメント	409	865	226	55.3	408	183	44.7	457
著作	608	1,765	133	21.9	491	475	78.1	1,274
表現形	516	2,731	111	21.5	173	405	78.5	2,558

表 3　実験 2：マッピング先 BF 記述の付与実験結果 1

		実験 2-①：意味分解表現の活用あり				実験 2-②：意味分解表現の活用なし			
	エレメント数	正解数	正解率	不正解数	不正解率	正解数	正解率	不正解数	不正解率
体現形エレメント	183	84	.459	99	.541	12	.066	171	.934
著作	475	175	.368	300	.632	40	.084	435	.916
表現形	405	188	.464	217	.536	42	.104	363	.896

表 4　実験 2：マッピング先 BF 記述の付与実験結果 2（実験 2-①：意味分解表現の活用あり）

	訓練用データ			テスト用データ				
				合計	正解数		不正解数	
	合計	contribution	それ以外		contribution	それ以外	contribution	それ以外
体現形エレメント	226	65	161	183	84	0	1	98
著作	133	41	92	475	171	4	65	235
表現形	111	13	98	405	186	2	100	117

表 5　実験 3：マッピングの層化 k 交差検証結果

		層化交差検証 1		層化交差検証 2	
	エレメント数	正解率	95%信頼区間	正解率	95%信頼区間
体現形エレメント	409	.210	[.174, .252]	.249	[.210. .294]
著作	608	.321	[.285, .359]	.326	[.290, .364]
表現形	516	.343	[.303, .385]	.192	[.160, .228]