JAPAN/MARC レコードから自動構築可能な著作識別子の提案

宮田洋輔（慶應義塾大学大学院） miyayo@slis.keio.ac.jp

はじめに

近年，図書館目録を概念レベルから再設計する試みが議論され，議論に基づいていくつかの概念モデルが提案されている。それらのなかの著名なものに，IFLA が『書誌レコードの機能要件　最終報告』[1]で提案した概念モデル，いわゆる FRBR モデルがある。FRBR モデルでは，実体-関連分析技法を用いて，記述の対象として 4 つの実体とその属性，そして各実体間の関係を定義し，より経済的で，集中機能に優れた設計が成されている。橋詰[2]や宮田[3]の調査から，日本の図書館目録においても，FRBR などの概念モデルを用いた図書館目録の再設計の有効性が明らかになっている。

FRBR に基づいて図書館目録を再構築することを「FRBR 化」と呼ぶ。FRBR 化に際して，これから作成される書誌レコードを FRBR 的に作成するのとともに，これまでに作成されてきた書誌レコードの，FRBR 化した形式への変換が必要である。

本研究では，著作レベルの実体を目録中で表現するために必要な著作識別子を，現行の JAPAN/MARC 形式の書誌レコードから自動的に構築することを目的として，日本で出版された 4 著作の分析をおこなった。

FRBR 実体と識別子

書誌レコードで表現される実体をデータベース中で扱うためには，それぞれの実体を識別する識別子が必要である。これまでの図書館目録にすでに組み込まれた実体，つまり FRBR モデルにおける体現形レベルと個別資料レベルのレコードに対しては，それぞれ ISBN と，請求記号や BOOK ID という形で識別子が利用されてきている。これらの識別子を用いることで，それぞれの実体がほかの実体と別の実体であることを表現できる。

しかし，これまでの図書館目録に実体として組み込まれていない著作レベルと表現形レベルとの実体に対しては，このような識別子が準備されておらず，今後，図書館目録への FRBR モデルの実装において，それらの実体に対する識別子の構築が必要である。

識別子の構築において，人手で書誌レコードや資料自体を精査し，表現された実体を識別し，なんらかの形の識別子を付与していくことは 1 つのアプローチである。信頼性や確実性の観点からは，標準化された規則に基づいて，このアプローチが望ましい。しかし，これまでに作成されてきた書誌レコード数は膨大であり，それらすべての性差という作業負荷を考慮すると現実的なアプローチとはいえない。

作業負荷の問題に対して，なんらかのアルゴリズムを用いて，機械的にそれぞれの実体を識別できる識別子を自動構築するアプローチが考えられる。現在，このようなアルゴリズムを用いた実体の自動識別が欧米を中心に研究がなされている。

先行研究

欧米では，既存の MARC レコードから，FRBR で設定された実体を自動的に同定・識別するアルゴリズムが開発されている。OCLC は，著作のレベルの実体にクラスタ化する Work-set Algorithm を開発し，公開している[4]。Work-set Algorithm は，WorldCat や，FictionFinder，xISBN など OCLC が提供する，FRBR 化されたサービスの基盤となっている。LC は，MARCXML と XSLT とを用いて，FRBR 化した検索結果表示を提供する FRBR Display Tool を開発し，公開し

smollett, tobias george\1721 1771/expedition of humphry clinker

図1　Work-set Algorithm での著作識別子[4)]

ている[5)]。

OCLC の Work-set Algorithm は，O'Neill がおこなってきた一連の Tobias Smollett 著"Humphly Clinker"の分析が基礎となっている[6)]。その分析に基づいて，著作を識別する同定子として，MARC21 フォーマットから，1XX（著者基本記入）と 24X（書名）あるいは 130（統一書名）を連結して，著作を識別する識別子を作成している。図1にその例を示した。

Carlyle らは，OCLC などの自動構築アプローチが実証的な評価がおこなわれていないことに対して，4 つの小説著作の分析から構築した著作識別子の識別性能の評価実験をおこなった[7)]。Charles Dickens 著『荒涼館』，Robert Louis Stevenson 著『宝島』，Alexandre Dumas 著『三銃士』，Louisa May Alcott 著『若草物語』の 4 著作の分析をおこない，書名，著者名，LCC からなる著作識別子を構築した。また，異なる著作に属するレコードの排除する除去基準も設定した。

これらのアルゴリズムは，欧米の記述規則と MARC21 及び MARCXML による符号化からなる書誌レコードに基づいている。そのため，目録規則も MARC フォーマットも異なる日本の図書館目録にこのアルゴリズムを適用しようとした場合，欧米と同程度の結果が得られるかどうかは定かではない。

そこで，本研究では，日本の書誌レコード作成の実践に適応した，著作識別子の自動構築を目的として，ある程度の大きさをもった著作を用いた分析をおこなった。

実体同定の流れ

機械的なアルゴリズムに基づく，書誌的実体の自動同定の流れを，図2に示した。,書誌レコードから識別子を構成する要素である同定子を抽出する。フィールドによる表記の相違を修正するために同定子に対して，変換処理をおこなう。本研究では，書名に対して拗音と促音処理を，著者名に対してフィールドによる表記法の

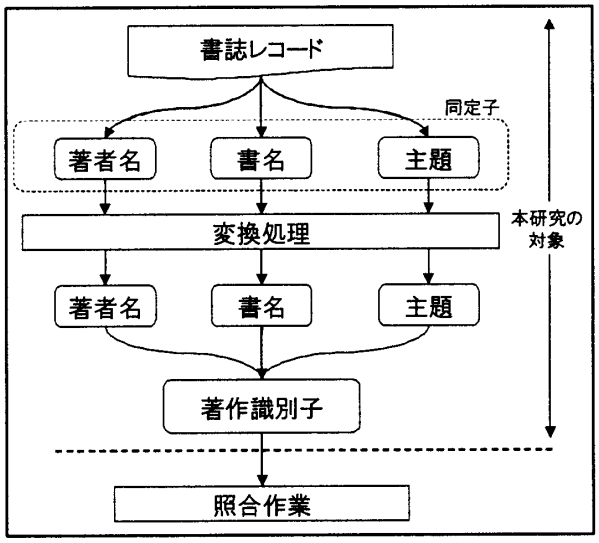

図2　実体同定の流れ

違いの修正をおこなった。変換された同定子に基づいて，識別子の構築し，構築した識別子の照合によって，実体の同定をおこなう。本研究では，実体同定のための照合をおこなう前の，著作識別子の構築までを主な対象としている。

データ集合

先行研究では小説著作の分析に基づいて識別子の構築がおこなわれてきた。本研究では，著作のジャンルでの限定はおこなわず，小説以外の著作も標本も含めた。

日本で出版された著作の中で，ある程度の大きさを持っていると判断したものとして，夏目漱石著『坊っちゃん』，徳富蘆花著『不如帰』，石川啄木著『一握の砂』，河上肇著『貧乏物語』の 4 著作を標本として選択した。

そして標本に含まれる各著作に対して，J-BISC(明治期-2007 年)を用いて，対象著作とその関係著作を示すレコードを収集した。レコードの検索には，「キーワード」フィールドに書名，「著者名」フィールドに著者名を入力し，検索した。現行の JAPAN/MARC レコードから検索できないレコードに対しては機械的なアルゴリズムを用いた識別子の構築が不可能と考えたため，それ以上の網羅的な検索はおこなわなかった。返された検索結果の JAPAN/MARC 形式のレコードをダウンロードし，分析に用いるデータ集合とした。データ集合中の著作に含まれる集合のみを特に著作集合と呼ぶ。表1に各著作

表 1　分析対象著作の概要

書名	著者名	ジャンル	データ集合の大きさ	著作集合の大きさ
坊っちゃん	夏目漱石	小説	129	127
不如帰	徳富蘆花	小説	25	21
一握の砂	石川啄木	歌集	74	65
貧乏物語	河上肇	評論	17	8

のデータ集合と著作集合の概要を示した。

同定子

識別子の構築に用いる要素のことを同定子と呼ぶ。FRBR モデルで示された著作の持つ属性と JAPAN/MARC フォーマットを比較し，対応付けをおこなった。表 2 にその結果を示した。なお，377（内容に関する注記）は，記述内容が構造化されておらず，機械による処理が困難なため，以降の分析には用いなかった。

結果

データ集合での同定子の出現率を算出した（表 2）。同定子の出現率は下式により定義した。

$$出現率 = \frac{タグの出現回数}{データ集合の総レコード数} \times 100$$

出現率の高さから，書名では，251A と 291A，著者名には 251F，291F，751B，791B を同定子として用いた。主題では，677A の NDC 以外は出現率が低かったため，同定子としなかった。

各同定子フィールドに対して変換処理を施し，各同定子フィールドに記述されたデータの一致率を算出した（表 3）。一致率は下式で定義した

$$一致率 = \frac{記述されたデータの一致件数}{データ集合に含まれる件数} \times 100$$

著作の書名では 291 を 251 に優先した。著作の著者は，791，291，751，25n の順に優先度を下げた。いずれの同定子も複数のデータを持った場合は，1 つ目のデータのみを考慮した。

一致率を見ると，書名は，典拠形が作成されていないためばらつきがあった。また全集，選集，合冊などとして出版され，書名が 2 つ以上の著作の書名を含んだ形になっているものも少なくなく，なんらかの処理が必要であろう。著者名は，記述形式を合わせる処理をおこなうことで,6 割から 8 割程度の高い一致率が得られた。主題は，『坊っちゃん』と『不如帰』との文学作品では，高い一致率を示した。

書名と，著者名，主題を組み合わせた著作識別子を構築し，著作集合に対する精度と再現率

表 2　同定子の出現率

属性	タグ	坊っちゃん データ集合	坊っちゃん 著作集合	貧乏物語 データ集合	貧乏物語 著作集合	不如帰 データ集合	不如帰 著作集合	一握の砂 データ集合	一握の砂 著作集合
書名	251A	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
	252A	0.8%						4.1%	4.5%
	253A							1.4%	1.5%
	291A	20.9%	21.3%	11.8%	33.3%	28.0%	25.0%	18.9%	15.2%
	377	34.1%	33.9%	29.4%	42.9%	36.0%	37.5%	48.6%	50.0%
著者名	251F	107.0%	106.3%	88.2%	85.7%	104.0%	75.0%	100.0%	103.0%
	252F	1.6%						5.4%	6.1%
	253F							2.7%	3.0%
	291F	2.3%	2.4%	5.9%	14.3%	12.0%	12.5%	12.2%	6.1%
	377	34.1%	33.9%	29.4%	42.9%	36.0%	37.5%	48.6%	50.0%
	751B	114.7%	112.6%	111.8%	104.8%	112.0%	100.0%	113.5%	116.7%
	791B	5.4%	5.5%	11.8%	76.2%	64.0%	25.0%	14.9%	12.1%
主題	650B	2.3%	0.8%	5.9%			12.5%	4.1%	4.5%
	658B	2.3%	1.6%	82.4%	9.5%	8.0%	75.0%	5.4%	4.5%
	677A	81.4%	81.1%	117.6%	100.0%	100.0%	112.5%	94.6%	93.9%
	685A	46.5%	44.1%	41.2%	81.0%	76.0%	50.0%	62.2%	62.1%

表3　同定子の一致率

同定子	タグ	坊っちゃん				不如帰				貧乏物語				一握の砂			
		データ集合		著作集合		データ集合		著作集合		データ集合		著作集合		データ集合		著作集合	
		種類数	最大一致率	種類数	最大一致率	種類数	最大一致率	種類数	最大一致率	種類数	最大一致率	種類数	最大一致率	種類数	最大一致率	種類数	最大一致率
書名	251	40	47.3%	39	48.0%	12	52.0%	10	52.4%	6	41.2%	4	62.5%	28	24.3%	25	27.3%
	251+291	37	45.7%	32	53.5%	11	60.0%	8	61.9%	6	41.2%	4	62.5%	28	27.0%	25	30.3%
著者名	251+751	13	85.3%	13	85.0%	6	60.0%	5	61.9%	3	82.4%	2	87.5%	8	74.3%	7	75.8%
	251+291+751+791	10	76.7%	10	76.4%	7	64.0%	6	66.7%	3	82.4%	2	87.5%	10	79.7%	8	80.3%
主題	677	11	43.4%	11	43.3%	10	44.0%	8	47.6%	9	29.4%	7	25.0%	15	32.4%	14	31.8%

表4　著作識別子の精度と再現率

識別子	坊っちゃん			不如帰			貧乏物語			一握の砂		
	種類数	精度	再現率	種類数	精度	再現率	種類数	精度	再現率	種類数	精度	再現率
書名+著者名	33	100.0%	52.0%	12	100.0%	61.9%	7	100.0%	62.5%	38	100.0%	28.8%
書名+主題	43	100.0%	28.3%	15	100.0%	47.6%	10	100.0%	25.0%	45	100.0%	9.1%
書名+著者名+主題	47	100.0%	28.3%	15	100.0%	47.6%	11	100.0%	25.0%	49	100.0%	9.1%

著者名+書名
　夏目漱石/坊つちやん
著者名+書名+主題
　夏目漱石/坊つちやん:913.6

図3　著作識別子の例

を，下式により算出した。

$$\text{精度} = \frac{\text{識別子}\ i\ \text{の著作集合に含まれる　件数}}{\text{識別子}\ i\ \text{が構築された件数}}$$

$$\text{再現率} = \frac{\text{識別子}\ i\ \text{の著作集合に含まれる　件数}}{\text{著作集合に含まれる件　数}}$$

再現率のもっとも高かったものを表に示した。いずれの著作でも精度はきわめて高かったが，再現率では，6割から2割と，著作によって，性能に大きな差があった。『一握の砂』,『坊っちゃん』で性能が低かったことは，これらの著作には，ほかの著作との合集として出版されているものが多く，書名による同定があまり機能しなかったことが考えられる。

同定子の組合せの中では，書名+著者名が最大の性能を示した。構築された識別子の例を図3に示した。

まとめ

JAPAN/MARC 形式の書誌レコードの分析を行い，自動構築可能な著作識別子の提案をおこなった。より有効な識別子の構築のためには，2番目以降のデータの利用や注記の処理も必要であろう。今後，著作識別子の評価実験及び大規模データへの適用などが考えられる。

引用文献

1)書誌レコードの機能要件：IFLA 書誌レコード機能要件研究グループ最終報告. 和中幹雄, 古川肇, 永田治樹訳. 東京, 日本図書館協会, 2004, 121p.

2) 橋詰秋子. FRBR からみた日本の図書館目録における著作の傾向：慶應義塾大学 OPAC を例として. Library & Information Science. 2007, No.58, p.33-48

3) 宮田洋輔. 日本の図書館目録における書誌的家系. 2008, 2008 年日本図書館情報学会春季研究集会.p 95-98.

4) Hickey, Thomas B; O'Neill, Edward. T; Toves, Jenny.. Experiments with the IFLA Functional Requirements for Bibliographic Records (FRBR). 2002, D-Lib Magazine, Vol.8, No.9,< http://www.dlib.org/dlib/september02/hickey/09hickey.html>[2008-09-09]

5) Library of Congress. FRBR Display Tool Version2.0< http://www.loc.gov/marc/marc-functional-analysis/tool.html>[2008-09-04]

6) O'Neill, Edward. The FRBRization of Humphry Clinker. <http://www.oclc.org/research/projects/frbr/clinker/>[2008-09-03]

7) Carlyle, Allyson; Ranger, Sara; Summerlin, Joel. Making the pieces fit: little women, works, and the pursuit of quality. 2008, Catalogin & Classification Quatery. Vol, 48, No.1, p.35-63.