

Wikipedia を利用したマンガの書誌データからの ストーリー単位の抽出

何雯凌^{†1} 三原鉄也^{†1} 永森光晴^{†2} 杉本重雄^{†2}

書誌レコードの機能要件 (FRBR) の第 1 グループで定義される実体の概念を利用することで、マンガの探索がより容易になると考えられる。その一方で、マンガにおいては、FRBR の第 1 グループの概念が作品全体や、編などで表される作品を構成するストーリー、あるいはストーリーを構成するより小さな単位 (各話のエピソードや複数のページやコマで表現されるシーン) など様々に考えられる。本研究は、マンガの作品全体を単位とするのとらえ方に加えて、ストーリー単位をとらえることで、探索を効率化することを目指し、書誌データからのストーリー単位の識別を目指す。しかし、マンガの書誌データの中にストーリー単位に関する情報は少ないため、機械的に抽出することは難しい。本研究では、Wikipedia の本文データを利用し、マンガ書誌データからストーリー単位を抽出する手法を提案した。京都国際マンガミュージアムに所蔵しているマンガの書誌データを用いた実験を行い、本手法の有効性を示した。この実験の結果から、本手法は利用した Wikipedia の情報の質に大きく影響されることが分かった。

Identifying Manga Stories from Bibliographic Data of Comic Books Using Wikipedia

WENLING HE^{†1} TETSUYA MIHARA^{†1}
MITSU HARU NAGAMORI^{†2} SHIGEO SUGIMOTO^{†2}

Manga – a Japanese term meaning graphic novels and comics – has been globally accepted. In Japan, there are a huge number of monographs and/or magazines for manga. Functional Requirements of Bibliographic Records (FRBR) provides useful concepts for readers to identify entities of manga, e.g., works of manga, expressions of manga, and so on. However, it is not clear which unit of manga those FRBR entities represent. They can be a whole manga or a story as a part of a manga, or even some smaller units of a manga (i.e. a single scene or an episode). This paper examines how to identify a manga work using a set of bibliographic records maintained by the Kyoto International Manga Museum. This study is aimed to identify a Work instance of manga, which may be a work as a series of stories or as a single story, in order to help readers find a comic book as an instance of a work. It is known that authority data is useful to identify works from bibliographic records. However, there is not enough information to identify a manga story in the bibliographic data. In this study, we used Wikipedia as a dictionary of manga to identify a manga story in the bibliographic records. The result of our experiment shows that using Wikipedia data is possible, but it also shows that the accuracy and efficiency depend on the quality of the Wikipedia data.

1. はじめに

1990 年代まで、マンガを収集対象の一つとするライブラリ (図書館、博物館等) は少なかった。しかし、近年、マンガが日本を代表するポップカルチャーの一つとして広く認知されるにつれて、多くのライブラリがマンガを収集するようになった。日本では、広島市まんが図書館 (1997 年開館) を始め、京都国際マンガミュージアム (2006 年開館) の様なマンガを主要なコレクションとして収集するライブラリが増えてきている。海外においては、アメリカのオハイオ州立大学図書館 (1999 から所蔵開始) や台湾の臺北市立圖書館中崙分館 (1998 から所蔵開始) 等がマンガを収集するライブラリとして挙げられる [1]。

ライブラリでは、ユーザが容易にマンガ資料を利用するために、各自で所蔵しているマンガ雑誌と単行本の書誌データを提供している。しかし、このマンガの書誌データ

はユーザのマンガの認知と探索行為を満足させることができない。なぜなら、マンガを認知及び探索する際にユーザが着眼する単位とマンガの書誌データの記述単位は異なっているからである。

マンガは絵やコマ割り、ページの連続によって、一つ又は複数の連続のストーリーを表すものである。しかし、マンガは雑誌に連載、又は単行本として出版される際、幾つかの話又は本に区切られて出版されることが多い。そのため、ユーザにとって、マンガ又はその中のストーリーは個別の書籍ではなく、それら書籍を集約することによって表現される知的実体、又は個別の書籍の内容の一部によって表現される知的実体である。このような認識はユーザの日常的なマンガの認知と探索行為の中で常に行われている。我々は、ユーザのマンガの認知と探索行為を支援するため、このような書籍の記載単位より大きい又は小さな知的実体をマンガ書誌データから識別することを目指す。

Hickey[2]は、こういった知的実体を同定する際に、典拠データから抽出したタイトルと著者名を利用することが有効であると示した。しかし、ライブラリにとってマンガは新しいコレクションであるため、マンガの典拠データは整

^{†1} 筑波大学大学院図書館情報メディア研究科
Graduate School of Library, Information and Media Studies, University of Tsukuba

^{†2} 筑波大学 図書館情報メディア系
Faculty of Library, Information and Media Science, University of Tsukuba.

備されていない。一方で、Web 上にはマンガに関する情報が多く存在している。その代表例として、多くのマンガ作品についての情報が記述されている Wikipedia が挙げられる。

そのため、我々は Wikipedia のリンク情報を Linked Open Data (LOD) として公開している DBpedia を利用し、書誌データからユーザの認知に該当するマンガの知的実体を抽出する手法を提案し、実験によって手法の有効性を明らかにした[3]。しかし、実験を通じて、DBpedia は Wikipedia のごく一部の情報(リンク情報)を扱ったものであるため、DBpedia からだけでは断片的な情報しか得られないことが分かった。とりわけ、マンガ作品中のストーリーの区切りに関する情報は殆ど得られなかった。この実験では、書誌データ中の「サブタイトル」という項目をマンガ作品中のストーリーに付けられるタイトル、即ち「ストーリータイトル」として認識することでマンガ作品中のストーリーを抽出した。しかしながら、書誌データ中の「サブタイトル」の項目は単行本毎に、すなわち書誌単位で付けられるサブタイトルが記述されている場合もある。そのため、この「サブタイトル」の項目は、必ずしもマンガ作品中のストーリータイトルを示しているとは言えない。結果として、DBpedia と書誌データを利用したこの実験では、「ストーリー」を上手く抽出できなかった。

そこで、本研究では、Wikipedia の本文データを利用し、マンガ書誌データからストーリーというユーザの認識単位(ストーリー単位)を抽出する手法を提案する。

以下、2 章ではマンガ作品のストーリーモデルの定義を説明する。3 章では、本研究の提案手法について述べる。4 章では、京都国際マンガミュージアムの書誌データを利用した抽出実験を紹介する。5 章では、実験により発見した幾つかの課題を検討する。

2. マンガストーリー単位モデル

2.1 FRBR とマンガメタデータフレームワーク

レコードの機能要件 (FRBR) [4]は書誌データの利用者の関心の対象として、知的・芸術的活動の成果に関する実体の集合を第 1 グループと階層的に定義している。著作 (Work) は個別の知的・芸術的創造を指し、第 1 グループの中の最上位の実体である。さらに、FRBR では著作の内容を表現する知的・芸術的実現を指す実体、著作の表現形の物理的な形態を指す実体、及びその単一の例示となる実体をそれぞれ、表現形 (Expression)、体现形 (Manifestation)、個別資料 (Item) と定義している。そして、ある特定の資料を探すプロセスとして、ユーザは常に一つの著作から、その著作の特定の言語表現である表現形を見つけ、また新書版などの異なる体现形を発見し、最終的にライブラリが所蔵している個別資料に辿り着くという手順を想定している。1 章で述べた、ユーザがマンガを認知・探索する際に

着眼する知的実体は FRBR では著作として定義している。故に、それらの実体を参照可能なマンガの情報を提供することによって、ユーザがより簡単にマンガを探索できると考えられる。

また、FRBR は一つの著作のインスタンスが更に幾つかの著作インスタンスによって構成されることが可能であると指摘している。こういった著作のインスタンスは、集合的実体、または構成的実体であると定義されている。

本研究では長編で且つストーリー性を持つマンガ、即ちストーリーマンガを研究対象とする。ストーリーマンガの場合、一つのマンガが連続性のある複数のストーリーから構成されることも少なくない。そのとき、「ONE PIECE」[a]というマンガのように、ストーリー毎に異なる舞台が設定されて、新たなキャラクターが登場することが多く見られる。一方、「ジョジョの奇妙な冒険」[b]のように、ストーリー毎にメインキャラクターが変更される場合もある。一般的に、ストーリーを構成する重要な要素として背景(舞台)、人物(キャラクター)、事件(出来事)が挙げられる。これより、ストーリー性を持つ著作は、この三要素が各々に異なる可能性が高いと考えられる。「ONE PIECE」や「ジョジョの奇妙な冒険」の様に、作品内に異なる舞台や、異なるキャラクターが登場する部分等があれば、このマンガは作品中に異なるストーリーを持つと考えられる。この様なストーリーもマンガ作品全体と同様に、ユーザがマンガを認知探索する際に、重要な知的実体(著作)として認識される。そして、このようなマンガ作品全体とマンガ作品を構成するストーリーは、FRBR の構成的(集合的)実体の定義と合致していると考えられる。

Morozumi[5]はマンガのためのメタデータについての基本的な要求要件としてマンガメタデータフレームワーク (MMF) を定義し、その中でマンガに関する書誌を記述するための FRBR に基づいたモデルを提案した。MMF は書誌レコードの記述対象となるストーリーの構造を title、story、episode という 3 つの実体とその関係によって定義した。この 3 つの実体は 1 対多の関係、即ち 1 つの title は複数の story から成り、1 つの story は複数の episode から成る関係を持つ。更に MMF はストーリーの構造を示す実体をそれぞれが FRBR の著作、表現形、体现形の実体を持つように定義している。

2.2 マンガのストーリー単位モデル

本研究では、書誌データの記述対象は個別資料レベルの書籍であることを考慮し、書誌データからユーザがマンガを認知・探索する際に必要な知的実体を抽出するために、MMF を基にモデルを提案する。図 1 に示すように、マンガストーリー単位モデルを定義する。

図 1 の縦方向の矢印はそれぞれ FRBR の第 1 グループに

a) 尾田栄一郎, ONE PIECE, 集英社

b) 荒木飛呂彦, ジョジョの奇妙な冒険, 集英社

基づく階層構造を示している。著作と表現形と体现形の階層内では、個別のマンガ作品全体とあるマンガの一部となっている個別のストーリーをそれぞれ manga-title、story と定義した。manga-title 即ちマンガ作品とは、独立なタイトルを持ち、個別の著者または著者グループによって創作される単一の著作または著作の集合である。story は manga-title を構成する個別のまとまりと定義する。同じ manga-title を構成する story はお互いに共通するメインタイトルを持ちながら、それぞれ独立なタイトル（ストーリータイトル）を持っている。また、manga-title の中に story という区切りが存在しない場合もある。本研究では、story を「ストーリー単位」と称する場合もある。そして、個別資料の階層内では、書誌データに記述されている個別資料を表す実体（a volume）が存在する。また、体现形の階層内では、manga-title、story の他に、単行本（冊子）のクラスである volume 実体を定義した。

各階層間では FRBR の第 1 グループの実体間の関係を継承する。個別の著作実体は一つ以上の表現形によって表現されることを示す「is realized through」の関係、個別の表現形実体は一つまたは複数の体现形によって具体化されることを示す「is embodied in」の関係をそれぞれ持つ。

図 1 の縦方向以外の矢印は各階層内における manga-title、story 間の関係を示している。著作と表現形の階層内では、story が manga-title の一部であることを示す「is part of」という関係が存在している。体现形の階層内では、「is part of」以外に、story と manga-title がそれぞれ幾つかの volume に区切られて出版すること、及び一つの volume の中に幾つかの story が存在することを表す「is contained in」という多対多の関係が存在する。

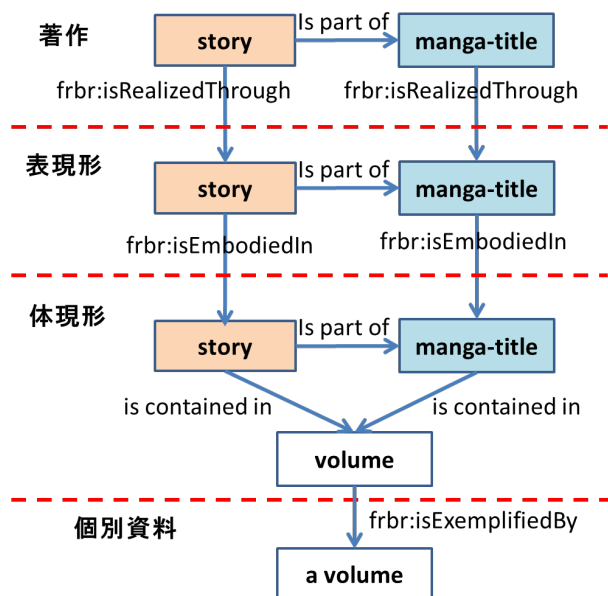


図 1 マンガのストーリー単位モデル

本研究では、マンガ書誌データから図 1 で示した実体及

び実体間の関係を抽出することを目指す。

3. ストーリー単位の抽出手法

3.1 関連研究

書誌データからの著作の同定と FRBR 化については、人手による同定作業のコストが高いため、機械的に同定する取り組みが行われている。Hickey の研究では、MARC のような特定の構造に従って作成された目録レコードの典拠データから抽出された著者とタイトルの組合せが著作を同定するための有効なキーとして利用できるとしている。外部リソースを利用してデータを FRBR 化する研究に関しては、Takhirov [6]と Duchateau [7]が Linked Open Data (LOD) を利用して、Amazon データを FRBR 化する研究を行った。この研究は、LOD リソースから FRBR 化に必要な情報を取得するために、著作の属性である著者、タイトル及びリソースタイプを利用することが有効であることを示した。

また、書誌データから集合的（構成的）な著作を同定する研究として、谷口[8]が挙げられる。谷口は著者とタイトルの組合せは有効な同定キーであるという原則に基づいて、JAPAN/MARC の階層構造を利用し、個々の階層から著者標目、責任表示、記述タイトル、タイトル標目を抽出した。そして、それぞれの階層レベルで同定キーを生成することで集合的（構成的）な著作を同定することができることを示した。

3.2 Wikipedia を用いたマンガの書誌データからストーリー単位の抽出手法

1 章で述べたように、書誌データは書籍の出版情報を主に記述したものである。更に、マンガは図書館にとって新しい種類の所蔵資料であるため、既存のマンガ書誌データから 2 章で定義したような実体及び実体関係を抽出することは難しい。そこで、本研究は Wikipedia の情報を用いて、マンガ書誌データからストーリー単位（story 実体）を抽出する手法を提案する。

また 2 章で述べたように、マンガ作品中にストーリーの区切りが存在するとき、ストーリーが幾つかの単行本に区切られて出版する場合、及び一冊の単行本の中に幾つかのストーリーが存在する場合があるため、論理的な実体関係としては story 実体と volume 実体の関係は多対多であると定義した。しかし、書誌データは書誌単位で記述されている、即ち書誌（単行本）より小さい物理単位（例えば、話数などの情報）が記述されていない。そのため、story 実体と volume 実体の関係が多対一である場合とマンガ作品中にストーリーの区切りが存在しない場合を見分けことは困難であり、この場合のストーリーの抽出は困難であると考えられる。故に、本研究の提案手法は、story 実体と volume 実体の関係は一对多の場合のみを対象とする。

ストーリー単位の抽出プロセスは図 2 のように、以下の 4 つのステップに分けられる。（注：本研究では、2 章のモ

デルで示したような複雑な関係を表現するため、Resource Description Framework(RDF) [c]のグラフ構造でデータ記述する。)

(1) volume 実体を抽出する

マンガの書誌データからタイトル、巻号、叢書名、出版社、著者が同じレコードを一つの volume 実体としてグルーピングし、volume 実体のデータセットを作成する。図 2 は MV1 及び MV3 が volume 実体として抽出されることを示している。

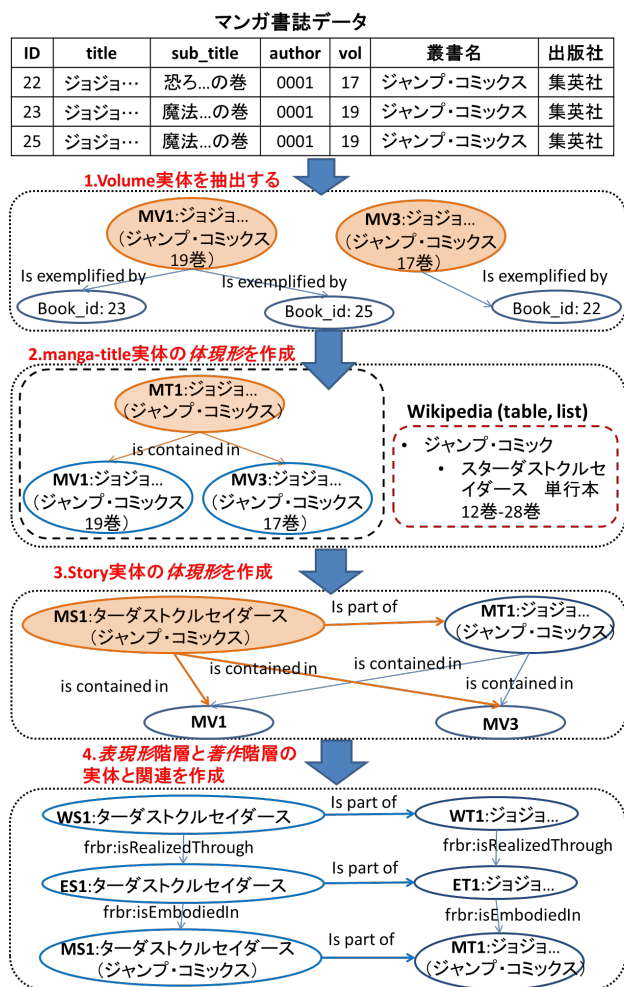


図 2 Wikipedia を利用したマンガ書誌データからのストーリー単位の抽出プロセス

(2) manga-title 実体の表現形を抽出する

(1)で抽出した volume 実体のデータセットからタイトル、叢書名、出版社、著者が同じものを一つの manga-title 実体の表現形としてグルーピングし、manga-title 実体の表現形のデータを作成する。図 2 では、MT1 が manga-title 実体の表現形として抽出された。

(3) story 実体の表現形を抽出する

(2)で抽出した manga-title 実体の表現形の書誌情報から、

タイトルと著者名を取得し、それらの情報を含む Wikipedia の記事を取得する。その記事の本文データから必要な情報を取得し、manga-title 実体の表現形の書誌情報の情報と組み合わせることで、manga-title 実体の表現形のデータから story 実体の表現形を抽出する。図 2 では、MS1 が story 実体の表現形として抽出された。また、抽出のために必要な情報として、タイトル、story 実体の表現形情報、叢書名、出版社が挙げられる。

ここで、「story 実体の表現形情報」は以下の 3 つの要素が含むと考えられる。

- ストーリータイトル
- 出版情報：出版社又は叢書名などの情報
- story の物理的な区切り情報：ストーリーの開始と終了の巻号

図 2 では「ジョジョの奇妙な冒険」の Wikipedia の情報を例として挙げている。

なお、書誌データは場合によって、マンガのストーリーの情報が詳しく構造的に記述されているときがある。一方で、Wikipedia の本文データは自然言語による文章で記述されており、「story 実体の表現形情報」の抽出が難しい。そのため、書誌データによっては、Wikipedia の本文データからより質の高い「story 実体の表現形情報」を取得できる場合もある。しかし、書誌データの構造及び記述内容はライブラリ毎に異なり、「story 実体の表現形情報」を取得する方法は対象となる書誌データ毎に異なると考えられる。本稿では、京都国際マンガミュージアムの書誌データを用いた事例を 4.2 節で述べる。

(4) 表現形階層及び著作階層のデータセットを作成する

文献[3]で示した実験の手法を利用し、Wikipedia のリンク情報などのデータを LOD として公開している DBpedia Japanese[d]から他言語のタイトル情報を取得し、表現形の階層のデータから表現形階層及び著作階層を作成する。

最終的に、作成した 4 つの階層を整合し、2 章で定義したモデルに沿ったデータを作成する。

4. 京都国際マンガミュージアムの書誌データを利用した実験

4.1 実験データ

本研究では 3 章で示した Wikipedia を利用したマンガの書誌データからストーリー単位を抽出する手法の有効性を検証するために、京都国際マンガミュージアム (KMM) の OPAC 用データを用いて実験を行った。

KMM の書誌データは 41 個のテーブルによって構成されるリレーショナルデータベースに保存されている。これらのテーブルは雑誌、単行本、著者典拠の 3 つのグループに関する情報として分類できる。雑誌と単行本のレコードは

c) Resource Description Framework (RDF): <http://www.w3.org/RDF/>

d) DBpedia Japanese: <http://ja.dbpedia.org/>

主に個別資料レベルの情報についての記述である。本実験では、5000 件の単行本レコード及び著者典拠レコードのみを利用する。また、KMM の書誌データを処理し易くするための前処理を行った。具体的には、単行本のタイトルとサブタイトルについて空白や特殊記号の削除などの正規化を行い、更に正規化されたデータセットと著者典拠データを結合させた。

また、本実験ではストーリー単位を抽出するため、補足の情報リソースとして Wikipedia 日本語版の本文データ (2013 年 1 月 25 日版) [e]を全文検索システム Apache Solr [f]に投入して利用した。これにより、Wikipedia の記事タイトル及び記事内容に対し、全文検索を行えるようにした。Wikipedia の記事は大勢のユーザの編集によって作成されたものであるため、全ての記事の記述構造は一定な形に整っていない。一方で、「story の表現形実体情報」のような構造を持つ情報の多くは、Wikipedia の記事の中に項目付きのテーブル、又は箇条書きのリストの形で記述されている。そのため本実験では、容易な取得が可能なテーブル及びリストで記述された情報のみをストーリー情報の抽出に用いた。

4.2 KMM の書誌データの場合の「story 実体の表現形情報」を抽出する手法

3.2 節で述べたように、ステップ 3 の manga-title 実体の表現形のデータから story 実体の表現形を抽出する手法の一例として、KMM の書誌データを用いた「story 実体の表現形情報」の取得手法を説明する。

(1) KMM の書誌データにおける「story 実体の表現形情報」の記述パターン

KMM の書誌データでは、既に「story 実体の表現形情報」の 3 つの要素の内の 2 つである単行本の巻号及び出版情報が記述されている。また、KMM の書誌データのレコード毎に「サブタイトル」という記述項目が存在する。書誌データの記述は個別の単行本に対するものであるため、「サブタイトル」の項目は単行本毎に付けられた「単行本タイトル」と考えられる。一方で、マンガが出版される際に、「ストーリータイトル」を単行本の「サブタイトル」として付ける場合も多い。この場合、KMM の項目「サブタイトル」の記述はストーリータイトルであると考えられる。

そのため、KMM の書誌データの場合、「story の表現形実体情報」の抽出は項目「サブタイトル」の記述パターンに大きく関わると言える。我々は KMM の書誌データの項目「サブタイトル」の記述パターンを図 3 のように 4 つのパターンに分ける。それぞれのパターンを A、B、C、D と表記する

パターン A は、マンガが区切られずまとめて単行本一冊

として出版された場合、つまり manga-title 実体の表現形と volume 実体の関係が 1 対 1 になっている場合である。パターン B、C、D はパターン A と違い、マンガが一冊以上の単行本として出版された場合、つまり manga-title 実体の表現形と volume 実体の関係は 1 対多になっている場合である。パターン B は、単行本毎に異なるサブタイトルが付けられた場合である。パターン C は、反対に、単行本毎に「サブタイトル」項目に情報が記述されていない場合である。パターン D は、単行本にサブタイトルが付けられ、異なるサブタイトルの数が単行本の数より少ない場合である。このパターンに該当する典型的な例として、図 3 のパターン D #a のように、最初の数巻にはサブタイトルが付いていないが、その後の単行本にサブタイトルが付けられた状況、または、図 3 のパターン D #b のように、継続する複数の単行本に同じサブタイトルが付けられた状況が挙げられる。

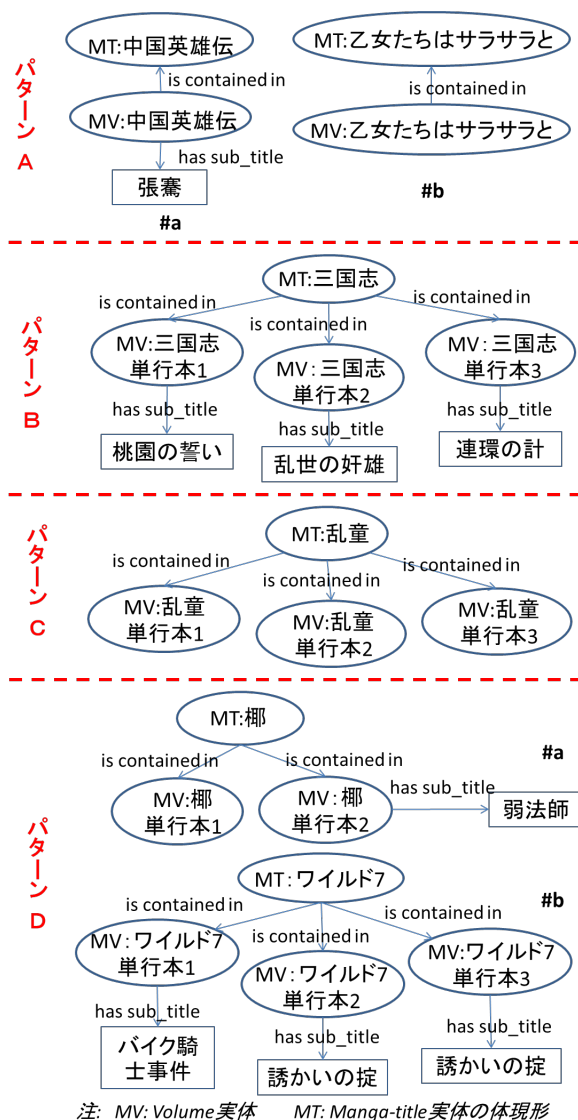


図 3 KMM の書誌データの項目「サブタイトル」の記述パターン

そして、ここでは manga-title 実体の表現形を持つ volume

e) Wikipedia 日本語版の本文データ: Wikipedia 日本語版のダンプファイル (2013 年 1 月 25 日版)。ダウンロード先: <http://dumps.wikimedia.org/jawiki/>
 f) Apache Solr: <http://lucene.apache.org/solr/>

実体の数を m 、volume 実体に付けられた異なるサブタイトル
の数を n として、この 4 つのパターンの特徴を以下の様
にまとめる。

- パターン A : $m=1 \cap m \geq n$
- パターン B : $m>1 \cap m=n$
- パターン C : $m>1 \cap n=0$
- パターン D : $m>1 \cap m>n$

(2) 各記述パターンに対応する「story 実体の体現形情報」の抽出方法

上述の 4 つの記述パターンのそれぞれに対応した「story
の体現形実体情報」の抽出方法は以下の様になる。

まず、2 章の定義により、著作階層の manga-title 実体と
story 実体の関連は 1 対多だと定義した。従って、manga-title
実体の体現形と volume 実体の関係は 1 対多でなければなら
ない。そのため、パターン A に該当するレコードが表すマ
ンガ作品中には、ストーリーの区切りが存在しない可能性
が高いと考えられる。実際に、パターン A の場合、story
実体の体現形と volume 実体の関連は多対 1 である可能性も
ある。しかし 3.2 節で述べたように、この実験は story 実体
の体現形と volume 実体の関連が 1 対多のみを考慮するため、
多対 1 の場合を扱わない。

また、パターン A に該当する manga-title 実体の体現形と
関連付ける唯一の単行本 (volume 実体) にサブタイトルが
付いている場合 (図 3 のパターン A #a) と付いていない場
合 (図 3 のパターン A #b) があるが、この場合では、ストー
リー単位を持たないため、「サブタイトル」の記述は「story
実体の体現形情報」と無関係であると考えられる。

パターン B とパターン C の場合は KMM の書誌データから
「story 実体の体現形情報」を取得できないと考えられる。
パターン B は「サブタイトル」項目に記述された情報を「単
行本タイトル」と判断したためである。パターン C は「サ
ブタイトル」項目に情報が記述されていないからだ。この
2 つの場合では、3.2 節で説明したステップ 3 のように、改
めて Wikipedia から「story 実体の体現形情報」を取得し、
manga-title 実体の体現形の書誌情報と組合せ、story 実体の
体現形を抽出する。

そして、パターン D の場合、項目「サブタイトル」の記
述はストーリータイトルと考えられるため、KMM の書誌
データから「story 実体の体現形情報」を取得できると判断
する。この場合、文献[3]で示した実験の手法を利用し、
KMM の書誌データから、タイトル、サブタイトル、著者
の組を同定キーとして作成し、manga-title 実体の体現形の
データから、story 実体の体現形を抽出する。

4.3 実験結果

サンプルとした KMM の 5000 件の単行本レコードに記述
された内容に対して、空白や特殊記号の削除などの正規化
を行った。更に、正規化されたデータと KMM の著者典拠
データを結合させた。その後、本実験の第一と第二段階と

して、3.2 節で述べたプロセスのステップ 1 とステップ 2
を行った。その結果、5000 件のサンプルレコードから 4976
件の volume 実体が抽出された。また、4976 件の volume 実
体から、2686 件の manga-title 実体の体現形が抽出された。
そして、ステップ 3 として、4.2 節で述べたパターン処理
を行った。その結果を表 1 に示す。ステップ 3 により、17
件の manga-title 実体の体現形から、54 件の story 実体の体
現形を抽出した。

表 1 パターン処理の結果

パターン 表記	該当する manga-title 実 体の体現形 の数	Story 実体の体現形 の抽出に成功した Manga-title 実体の 体現形の数	抽出した story 実体の体現形 の数
A	1926	0	0
B	44	1	3
C	700	0	0
D	16	16	51

最後に、ステップ 4 を行い、2686 件の manga-title 実体の
体現形と 54 件の story 実体の体現形から、2609 件の
manga-title 実体の表現形と 54 件の story 実体の表現形を抽
出し、更に 2609 件の manga-title 実体の著作と 54 件の story
実体の著作を抽出した。

しかし、手作業で Amazon と Wikipedia のデータを確認
した結果、パターン B とパターン C に該当する 744 件の
manga-title 実体の体現形の内、実際にストーリー単位を持
つものが 19 件あると分かった。そして、表 1 が示すよう
に、その 19 件の manga-title 実体の体現形の内、story 実体
の体現形の抽出に成功した Manga-title 実体の体現形は
1 件 (「ジョジョの奇妙な冒険」) のみだった。

5. 考察

4.3 節で述べたように、パターン B とパターン C に該当
する且つ実際にストーリー単位を持つ manga-title 実体の体
現形は 19 件があるが、その内の 18 件は story 実体の体現
形の抽出に失敗した。その原因を分析するため、Wikipedia
にその 18 件の実体の「story 実体の体現形情報」が取得可
能な条件を満たしているのか、即ちストーリータイトル、
出版情報、ストーリーの区切り情報の 3 つの要素がテーブ
ルやリストのような簡単に情報を取得できる形式で提供さ
れているかどうかについて調査した。(表 2 参照)

表 2 が示す様に、19 件の内 16 件 (84.2%) のリソース
は Wikipedia の記事として存在した。しかし、それらは「story
実体の体現形情報」を取得するための条件を満たしていな
かった。故に、「story 実体の体現形情報」を Wikipedia から
取得することに失敗し、この 19 件の manga-title 実体の体
現形から story 実体の体現形を作成することが出来なかつ
た。このことから、本研究における提案手法は Wikipedia
の情報の質に大きく影響されることが分かった。

表 2 「ストーリー実体の体現形情報」の取得に失敗した
19 件の実体の Wikipedia の記述状況 [g]

タイトル	Wikipedia に存在する リソースは 該当する	story 実体の体現形情報				
		出版 会社	書名 タイトル	ス トー リー	巻 号	リス ト (テ ー ブル 形式 存在 形式)
野球狂の詩	T	F	F	F	F	F
ハロー張りネズミ	T	T	F	T	T	T
犬神	T	F	F	F	F	F
機動警察パトレイバー	T	F	F	F	F	F
キャプテン翼	T	F	T	F	F	F
孔雀王	T	F	T	F	F	F
ドカベン	T	F	F	F	F	F
Swan：白鳥	T	F	F	F	F	F
ロンタイ baby	F	F	F	F	F	F
なんて素敵にジャパネスク	T	T	F	T	T	T
遊・戯・王	T	F	F	F	F	F
三つ目がとおる	T	F	F	F	F	F
課長島耕作	T	F	T	F	F	F
僕はムコ養子	T	T	F	T	T	T
花きゃべつひよこまめ	F	F	F	F	F	F
それでも地球は回ってる	F	F	F	F	F	F
麒麟館グラフィティ	T	F	F	F	F	F
新編風の輪舞	T	F	T	F	F	F
Papa told me	T	F	F	F	F	T
計数	16	4 項目が同時に T になる数:0				

実際に Wikipedia 上で「story 実体の体現形情報」を取得するための条件を満たす、即ちストーリータイトル、出版情報、ストーリーの区切り情報の 3 つの要素がテーブルやリストのような簡単に情報を取得できる形式で提供されているリソースの数を調べるための調査を行った。調査結果として、Wikipedia に存在する 11780 件のマンガに関する記事の内、「story 実体の体現形情報」を取得するための条件を満たす記事の数は 1293 件 (11.0%) であると分かった。また、要素毎の存在状況は以下の通りである。

- 「ストーリータイトル」を含む記事数: 2789 件 (23.7%)
- 「出版情報」を含む記事数: 2475 件 (21.0%)
- 「ストーリーの区切り情報」を含む記事数: 2176 件 (18.5%)

これにより、Wikipedia の記事中には、ストーリー情報がある程度記述されているが数は少ない、また、たとえ記述されていても機械的に扱いやすい形で提供されているものは少ないということが分かった。従って、今後ストーリー単位に関する情報の抽出のためには Wikipedia 自体の情報がより充実することが求められる。また、Wikipedia 上の情

g) 表内の記号の説明:

- T: 該当する項目の情報が Wikipedia に存在する。
- F: 該当する項目の情報が Wikipedia に存在しない。

報の利用性を向上させるため、日本語 Wikipedia オントロジーや DBpedia のような Wikipedia の情報を Linked Open Data (LOD) として公開するデータセットや Wikipedia の様に多くのマンガの情報が一箇所に集まるリソースの増加が望まれる。

また、本研究では、2 章で述べたように、ストーリー単位と単行本の関係を $n:m$ であると定義した。しかし、提案手法は書誌データの記述単位に依存し、ストーリー単位と単行本の関係が $1:n$ である場合のみの対応に留まり、例えば、一冊の単行本に複数のストーリーが掲載される場合、またはストーリーの中で更に小さい物語の単位が含まれている場合には対応出来ていない。今後の課題の一つとして、こうしたストーリー単位の構造のパターンに対応できるように提案手法を改善することも挙げられる。

6. おわりに

本研究は、マンガの作品全体を単位とする捉え方に加えて、ストーリー単位を捉えることで、探索を効率化するために、書誌データからストーリー単位を識別することを目指した。しかし、マンガの書誌データの中にはストーリー単位に関する情報があまり記述されていないため、機械的に抽出することは難しい。本研究では Wikipedia の本文データを利用し、マンガ書誌データからストーリー単位を抽出する手法を提案した。京都国際マンガミュージアムに所蔵しているマンガの書誌データを用いた実験を行い、本手法の有効性を示した。一方で実験の結果から、本手法は利用した Wikipedia の情報の質に大きく影響されることが分かった。今後の課題として、Wikipedia の情報がより充実し、LOD のような利用しやすい形で提供されることが望まれる。また、Wikipedia の様に多くのマンガの情報が一箇所に集まるリソースの増加も課題の一つとして挙げられる。また、一冊の単行本に複数のストーリー単位が掲載される様な他のマンガ作品のストーリー単位のパターンも考慮し、抽出手法を改善することも課題の一つとして挙げられる。

謝辞 本研究に取り組むにあたって実験データを提供してくださった京都国際マンガミュージアムに感謝致します。また、実験環境の整備及び論文の日本語のチェックをしてくださった本研究室学生の本間維、落合香織の各氏に感謝致します。

なお、本研究は一部、科学研究費補助金（挑戦的萌芽研究 #25540153）による。

参考文献

- 1) カブリン、三原 鉄也、永森 光晴、杉本 重雄. DBpedia を利用したマンガの書誌データからの work の同定. デジタル図書館. no.44, pp.11-19, 2013-03-14, <https://www.tulips.tsukuba.ac.jp/dspace/handle/2241/119863>
- 2) Thomas B. Hickey, Jenny Toves.2009.for OCLC Research FRBR

- Work-Set Algorithm (Version 2.0),
<http://www.oclc.org/content/dam/research/activities/frbralgorithm/2009-08.pdf>
- 3) Wenling He, Tetsuya Mihara, Mitsuharu Nagamori, Shigeo Sugimoto. Identification of works of Manga Using LOD Resources - An Experimental FRBRization of Bibliographic Data of Comic Books -, JCDL '13 Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries, P253-256,
<http://dl.acm.org/citation.cfm?id=2467731>
- 4) IFLA Study Group on the Functional Requirements for Bibliographic Records.2009.Functional Requirements for Bibliographic Records Final Report,
http://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf
- 5) Ayako Morozumi, Satomi Nomura, Mitsuharu Nagamori, Shigeo Sugimoto.2009.Metadata Framework for manga: A Multi-paradigm Metadata Description Framework for Digital Comics.
<http://dcpapers.dublincore.org/pubs/article/view/979/952>
- 6) Naimdjon Takhirov, Fabien Duchateau, Trond Aalberg.2011.Linking FRBR Entities to LOD through Semantic Matching.Research and Advanced Technology for Digital Libraries Lecture Notes in Computer Science.6966.284-295.
<http://liris.cnrs.fr/~fduchate/papers/duchateau-tpdl11-linking.pdf>
- 7) Fabien Duchateau, Naimdjon Takhirov, Trond Aalberg.2011.FRBRPedia: a Tool for FRBRizing Web products and Linking FRBR Entities to DBpedia. In: Proc. of JCDL.
<http://dl.acm.org/citation.cfm?id=1998183>
- 8) 谷口祥一.2009.FRBR OPAC 構築に向けた work の機械的同定法の検証--JAPAN/MARC 書誌レコードによる実験.Library and information science.61.119-151,
<http://lis.mslib.jp/pdf/LIS061119.pdf>