[E21] ビデオゲームのメタデータ拡張のためのメディア芸術データベースと Wikipedia とのリンキング

○福田一史 1)，三原鉄也 2)
1) 大阪国際工科専門職大学 工科学部，〒530-0001 大阪市北区梅田 3-3-1
2) IT コンサルタント/コネクテッド社会研究機構
E-mail: fukudakz@gmail.com

Linking Media Arts Database to Wikipedia for Video Game Metadata Enhancement

FUKUDA Kazufumi1), MIHARA Tetsuya2)
1) International Professional Unviersity of Technology in Osaka, 3-3-1 Umeda Kita-ku, Osaka-shi, Osaka-fu 530-0001 Japan
2) IT Consaltant / Research Institute for Connected Societies

【発表概要】

検索やデータ利活用など調査・研究のための専門分野の要求に基づく、メタデータのリッチ化は一つの論点であるが、データ作成コストが課題となる。その解決策として、オープンデータとのリンキングは有効な戦略である。ビデオゲーム分野では、オンラインコミュニティが生成するゲーム作品データセットが評価を集めており、同時にアーカイブ機関が生成する現物資料の書誌データ作成も活発化しつつある。本研究は効率的なメタデータの拡張や補完・検証を目的に、メディア芸術データベースのビデオゲームパッケージリソースと、Wikipedia より抽出したビデオゲームリソースのリンキング実践について報告する。Wikipedia のゲームタイトル一覧から情報を抽出し、メディア芸術データベースの関連するリソースを機械的にリンクさせた。その結果、多くのリソースで適切なリンクが確認できたが、Wikipedia の利用者要求や記述ルールの不統一に起因する課題と限界が明らかになった。

1. 背景

検索やデータ利活用など調査・研究のための専門分野の要求に基づく、メタデータのリッチ化はアーカイブ構築における一つの論点である。ただし、リッチ化のためのデータ作成コストの増大は一つの論点である。

低コスト化の方法論として、オープンデータとのリンキングは、継承によるデータ拡張や外部データセットとの接続性向上（データの調和）、相互のデータ比較による検証・網羅性評価などといった観点からメタデータ品質向上のための有効な戦略である。

本研究では、国内のビデオゲームリソースのメタデータを対象に、低コストで機械的なリンキング手法について検討する。ビデオゲーム分野のオープンデータとしては、Wikipedia[1]や IGDB[2]や MobyGames[3]などウェブサービスで、オンラインコミュニティが生成するゲーム作品のデータセットが評価を集めている。

このようなウェブ上での活動と並行して、立命館大学ゲーム研究センターやゲーム保存協会など、アーカイブ機関が生成する現物資料であるビデオゲームパッケージ（以下、パッケージとする）の書誌データ作成も活発化しつつある。これらのメタデータは、文化庁が運営するメディア芸術分野のつなぎ役[4]であるメディア芸術データベース（以下、MADB）への登録が進みつつある。

よく知られるところだが、ビデオゲームでは通常版、限定版、Blu-Ray 版、オンラインリソース版、など異版が存在する場合が多い。オンラインコミュニティにより作成されるゲーム作品リソースは創作的実体すなわち作品を記述対象とし、アーカイブ機関により作成される現物資料は物理的実体を記述対象とする。つまり多くの場合、一つのゲーム作品は一つないし複数の現物資料により体現されるという関連で記述可能であり、アーカイブ機関によるリソースのほうがより細かい粒度だと言える。両者は、機能的には役割分担と捉えることが可能であり、またビデオゲーム分

野のオントロジー[5][6]で定義される一部の実体の記述対象とフィットする。さらに、リンキングによる両者へのメタデータ拡張や補完・検証のメリットは大きく、そのような試作がいくつか展開されてきた[7][8][9]。

2. 研究の目的

本研究は、効率的なメタデータの拡張や補完・検証を目的に、メディア芸術データベースに登録されるビデオゲームパッケージリソースと、Wikipedia 日本語版（以下 Wikipedia）より抽出したビデオゲームリソースのリンキング実践について報告する。オープンデータとして Wikipedia を選択した理由としては、1) 日本のゲームリソースを対象とするサービスであること、2) 各プラットフォームのゲームタイトル一覧として同サービスで記事化されていないリソースを含め網羅性が高いリストが記録されており、3) 記事ごとに生成される Wikidata リソースが多数の外部 ID を持つハブとして機能すること[10]、が挙げられる。MADB では、アーカイブ機関が生成した現物資料すなわちゲームパッケージが主要な記述対象である。未所蔵のものについては、ゲーム雑誌やカタログ、ウェブページなど複数の情報源を元に、2017 年までのものに限られるが、国内で公開された家庭用ゲーム、アーケードゲームの大部分が登録されている。

3. Wikipedia におけるビデオゲームのタイトルリストの記述

Wikipedia にはゲーム作品やゲーム機など様々なビデオゲームに関する記事が公開されている。とりわけゲーム作品の記事には詳細な記述と、前章で述べたデータ連携における有用性が期待できる。ただし記事が作成されその記述が充実している作品に偏りがあり、網羅性の点では充分ではない。また記事の記述対象も個別の作品以外に、続編を含むシリーズ、原作を含む作品群など多様で、粒度は一定しない。これらは網羅性・記述対象の厳密性に焦点化する MADB とは異なる性質で、そのリンキングを困難にするものである。

そこで本研究では Wikipedia のゲームタイトル一覧の記事に着目した。Wikipedia では「バンダイナムコエンターテインメント発売のゲームタイトル一覧」「CERO レーティング 18 才以上対象ソフトの一覧」といった多様なゲーム一覧を主たる内容とする記事および記事を集約するカテゴリが提供されている。本研究ではその内「カテゴリ:ゲーム機別ゲームタイトル一覧」[11]のゲームプラットフォーム毎に作成されている一覧（以下、PF 一覧とする）からのデータ抽出を試みた。日本国内で公開された多くのゲームプラットフォームの PF 一覧が既に公開されており、収録の基準が明確で、さらに高い網羅性が期待できる。また、一覧中のリソースの内、作品の個別記事を持つものについてはそれらへのリンクが記録され、先に述べた作品記事の情報の活用やデータ連携にも有用である。

PF 一覧は記事毎に記載形式や記載項目が異なるが主に HTML のリスト形式またはテーブル形式で公開されている。主要な項目はタイトルと発売日であり、その他に発売元や付属品・限定版等の異版体に関する情報が記録されることも多い。その他にメディア形式（例：SNK が販売していたネオジオでは ROM カートリッジと CD-ROM が存在し、同タイトル複数のパッケージが頒布される）やダウンロード版の有無、レーティングなど、ゲームプラットフォームの特徴に即した項目が存在し、一覧毎に特有の記法・ルールで示される。また、オンライン配信以前のプラットフォームについては物理パッケージ主体でリストが作成されているが、オンライン配信以後のものについては、それらは一体として記述されており両者で記述の粒度が異なる。

記事の作成単位は概ねプラットフォーム毎であり記事タイトルからそれらを判別することが可能だが、プラットフォームそのものの解釈に曖昧さが残るため MADB のプラットフォーム記述とは一部異なる。またタイトル数が多数の場合、発売年毎に個別の一覧が作成されているものもある。

4. リンキングの実践と結果

本研究では、3章で述べたPF一覧からWebスクレイピングを用いてビデオゲームリソースを抽出した。抽出したリソースの各項目のうち、タイトル・発売日・プラットフォームを主要なキーとしてMADBのパッケージリソースと同一のパッケージに関するリソースを同定、リンキングした。

データ抽出に際しては、「カテゴリ:ゲーム機別ゲームタイトル一覧」に含まれる PF 一覧 84 ページを対象に、各ページの記述ルールに従って記述された項目全てを取得した。取得した項目は、全ての PF 一覧に共通する項目（ただしデータが欠損し不完全なタイトルを含む）に「タイトル」「発売日」「発売者」、一部の一覧の項目として「CERO レーティング」「ジャンル」「対応周辺機器」「ゲーム容量」「ゲームモード」「規格」「備考」「開発元メーカー」「プレイヤーモード」「脚注」「基盤」があった。これらについて文字列の他、ゲーム作品記事へのリンクも個別の項目として取得した。また「プラットフォーム名」や前述の項目の内、一覧のコンテンツとして示されないが、記事名や見出しなどの HTML 要素として記述された情報も適宜項目値として加えた。さらに、タイトル及び備考に「〇〇版」といった異版の記述が含まれる場合にそれらを値とする「異版体」の項目も設けた。

抽出したリソースと MADB とのリンキングに際しては、抽出したリソースの内、MADB に登録されている家庭用ゲーム機のプラットフォームのパッケージリソースのうち、物理パッケージを対象とし、ダウンロード版やオンライン配信で提供されているリソースについては除外した。結果リンキングを行ったプラットフォームは 34 件、それらに関する PF 一覧は 59 ページとなった。

リンキングのアルゴリズムは次の通りである。まず、プラットフォームと発売日の完全一致を条件として、抽出したリソース 1 件に対して候補となる MADB のリソースを絞り込む。発売日については年月日の内、与えられている値のみの一致を条件とする。次にその候補についてレーベンシュタイン距離を用いてタイトルの文字列の類似度を算出して比較し、20%以上類似した候補の内、最も類似度が高いものをリンキングする。タイトルの比較に際しては抽出したリソースがタイトル 1 項目に記述されていることに対し、MADB のリソースではタイトルと別タイトルなど構造化されているため、これらも結合して評価する。また評価対象の抽出したパッケージリソースが異版体である場合には MADB のリソースの版表示についてもテキスト一致をリンキングの条件とした。最後に、発売日のずれを許容するためにリンキングできていない抽出したリソースと MADB のリソースを対象に、日付の条件を 1 段階緩和し（年月日まで値のあるものは年月、年月しか値のないものは年のみ)、タイトルの類似度によるリンキングを行った。

この手法により 34,109 件の抽出したリソースと MADB のリソースをリンキングすることができた。一方で 2,749 件の抽出したパッケージリソースについては MADB へのリンクを得られなかった。また、リンクが得られたリソースの内 24,833 件は個別記事へのリンクを有していた。さらに MADB のリソースとリンクしたリソースとその項目については RDF データセットを生成した。作成したデータは 1,225,907 トリプルに及んだ。

リンキングできなかったリソースについては目視による誤リンクを分析したところ、両者のデータそれぞれに登録されていないリソースがあり、これが原因であることが明らかになった。特に Wikipedia では異版体の網羅が十分ではないことが目立った。これに加えて、異版の記述ルールが多くの場合不明確で、リスト及びテーブルでの構造化も十分ではないことも抽出を困難にする要因であった。これらは PF 一覧が現物資料の網羅的なリストではなく、あくまで作品のリストを志向していることの表れと言える。

5. 示唆と今後の課題

本研究では、Wikipedia に記載されている

半構造化リストからビデオゲームリソースの抽出を行い、MADB のビデオゲームパッケージリソースとのリンキング手法を開発し実践した。これにより物理的、コスト的な制約のある現物資料に基づく目録作成では容易ではない大規模なメタデータの拡張が可能であることを示した。あわせて、MADB の高い網羅性を示すことができた。

しかしながら、今回行ったリンキングについては厳密な正確性の評価ができておらず、この実現は大きな課題である。有識者による網羅的なチェックが最も高品質なものの、本手法の利点である効率性の観点からは望ましくない。ゲーム販売者が提供している信頼性の高いデータの利用や複数のデータベースを参照する手法が有効であろう。

また、今回リンキングについては統一的なアルゴリズムを設計し実践することができたが、抽出では一覧毎に個別の対応を要し、品質管理の負担が大きかった。Wikipedia は常に編集されうるため、継続的にデータ抽出・リンキングを行うことで持続的なメタデータの品質向上が期待できるが、そのためにはデータ抽出仕様の汎用的な記述や抽出ルールの管理手法による手間の削減、及びデータ整合性を担保する仕組みが求められる。

参考文献

[1] Wikipedia. https://ja.wikipedia.org/ (参照 2022-09-04).
[2] IGDB. https://www.igdb.com/ (参照 2022-09-04).
[3] MobyGames. https://www.mobygames.com/ (参照 2022-09-04).
[4] デジタルアーカイブの連携に関する関係省庁等連絡会・実務者協議会. デジタルアーカイブの構築・共有・活用ガイドライン. 東京, 内閣府知的財産戦略推進事務局, 2017. https://www.kantei.go.jp/jp/singi/titeki2/digitalarchive_kyougikai/guideline.pdf (参照 2022-09-04).
[5] Lee, Jin ha, Perti, Andrew, Clarke, Rachel Ivy, Windleharth, Travis W., Schmalz, Marc. Video Game Metadata Schema Version 4.0. UW/SIMM, 2017. https://gamer.ischool.uw.edu/releases/ (参照 2022-09-04).
[6] Fukuda, Kazufumi, Mihara, Tetsuya. A development of the metadata model for video game cataloging: for the implementation of media-arts database. 2017.
[7] 大石康介, 豊田将平, 吉岡孝祐, 三原鉄也, 永森光晴, 杉本重雄. [P05] マンガ・アニメ・ゲーム作品の横断的アーカイブのための作品間関係 LOD データセット開発. デジタルアーカイブ学会誌. 2018, vol. 2, no. 2, p. 150–151.
[8] Hoffmann, Tracy. “Developing a Mediated Vocabulary for Video Game Research”. Proceedings of the Doctoral Symposium on Research on Online Databases in History 2019. 2019, p. 11p.
[9] 福田一史. コミュニティ生成データを典拠とした目録作成の試み：ビデオゲーム目録を事例として. 情報の科学と技術. 2022, vol. 72, no. 8, p. 299–306.
[10] 大向一輝. 識別子としての Wikidata. 情報の科学と技術. 2020, vol. 70, no. 11, p. 559–562.
[11] Wikipedia -Category:ゲーム機別ゲームタイトル一覧. https://ja.wikipedia.org/wiki/Category:ゲーム機別ゲームタイトル一覧 (参照 2022-09-24).

この記事の著作権は著者に属します。この記事は Creative Commons 4.0 に基づきライセンスされます(http://creativecommons.org/licenses/by/4.0/)。出典を表示することを主な条件とし、複製、改変はもちろん、営利目的での二次利用も許可されています。