Local Archives-Wikidata: Focusing on NDRM Authority Data

Sangeun Han[1,*] and Seulki Do[2,†]

[1] University of Toronto, 27 King's College Cir, Toronto, Canada
[2] Hansung University, 116 Samseonyo-ro 16-gil, Seongbuk District, Seoul, South Korea

Abstract

This study linked the data of 265 persons and 197 organizations that are related to the National debt Redemption Movement (NDRM), a small local archive, to Wikidata as a way to improve access and usability. In order to improve usability in accordance with the FAIR Principles, we analysed the properties used in the Program for Cooperative Cataloging (PCC) Wikidata linkage pilot project, and mapped them to the elements used in the metadata Application Profile(AP) for the NDRM person/coporate body authority metadata developed in 2023. Based on the mapping table, we donated the data to Wikidata and applied it to a digital archive system to make the person/coporate body data accessible. This study can serve as a guide for small local archives to improve access and usability of their data.

Keywords

National Dept Redemption Movement Digital Archive, Wikidata Linkage, NDRM Person/Corporate Body Authority Data, Metadata Mapping

1. Introduction

In the digital environment, libraries and archives are building digital archives to make their local resources available on the web. With the growing interest in digital humanities and living in a data-driven era, digital archives are exploring ways to make information more accessible and to open and share records at the data level so that anyone can search and utilize archival resources [2, 3, 5, 9]. However, many GLAM (Galleries, Libraries, Archives, and Museums) institutions that have built digital archives still have the limitation that they have systems, structures, and processes that are dependent on their institutions [1, 6]. Especially in small, local archives with small data, it is always mentioned that it is difficult to not only build a digital archive but also to share and utilize the data. One way to address these challenges is to consider linking local archive data with various external institutions through Wikidata to provide various access points for users and manage resources [8].

Wikidata is a free collaborative knowledgebase that provides data in a human- and machine-readable format. It can help search engines find data on the web that is often locked away in libraries and archives [7]. Institutions with their own digital resources of person and corporate

* Corresponding author.
† These authors contributed equally.
sangeun.han@utoronto.ca (S. Han); sinhwask@hansung.kr (S. Do)
0000-0002-0759-5487 (S. Han); 0000-0001-6473-9240 (S. Do)

https://doi.org/10.23106/dcmi.952422992

body, subject, work, etc. can link to Wikidata around the unique IDs of their digital resources. It makes data in the local archive accessible from a variety of sources linked to Wikidata, and vice versa, the local archive can reference the richness of data in Wikidata, giving local archive users a richer data exploration experience.

The National Debt Redemption Movement (NDRM) was a citizen-led national sovereignty restoration movement that from 1907 to 1908 to repay the country's debt through public fundraising. The records of the National Debt Redemption Movement generated in the process are recognized as historical documents with important global significance as records that contain the development of the people's sense of responsibility and peace ideas. 2,475 records were registered as Memory of the World in 2017 [11], and the digital archive provides information services so that anyone can search and utilize the records of the National Debt Redemption Movement on the web.

Despite the limitations of being a local archive with small data, the National Debt Redemption Movement digital archive has been working on linking the digital archive data with Wikidata since 2021. In 2022, 233 archival documents and 25 key persons related to the National Debt Redemption Movement were linked to Wikidata on a pilot basis to create various access points to the data in the National Debt Redemption Movement digital archive.

The National Debt Redemption Movement digital archive lacked a unique ID system for its resources, especially for the names of person and corporate body that could be key subjects. This led to a limited, one-way linkage with Wikidata in 2022. In 2023, the focus was on devising ways to overcome these limitations, so that the archives and person and corporate body authority data, which are the main resources of the institution, could be given unique IDs and linked to Wikidata's global ID to take full advantage of Wikidata's strengths. The metadata Application Profile (AP) was developed to facilitate the internal management of person and corporate body authority data within the local archive and to increase the utility of the data by linking it with external institutional data. Based on the metadata AP, an additional 2,208 manuscripts, 264 persons, and 197 organizations were linked to Wikidata in 2023.

This study was about the linking of the digital archive of the National Debt Redemption Movement with Wikidata in 2023, and focuses on the systematic management of the digital archive of the National Debt Redemption Movement, as well as its global accessibility and searchability.

2. Overall Research Objectives

The aim of this study was to establish the linking between the authority data of person and corporate body in the National Debt Redemption Movement digital archive and Wikidata. This entailed conducting a case study on the relationship between GLAM and Wikidata, mapping metadata elements, creating a linkage model, and establish an interlinking with Wikidata.

Initially, we will examine instances of interlinking GLAM with Wikidata, analyzing the methods and the essential Properties (attributes) involved. Secondly, we will analyze the authority data model of the digital archive for the NDRM and determine the essential elements required for its interlinking with Wikidata. Thirdly, we will generate a mapping table by conducting case studies and mapping the properties of authority data associated with NDRM, with the objective of establishing an interlinking model. Fourthly, we will establish the interlinking with Wikidata using the model for interlinking that has been developed.

https://doi.org/10.23106/dcmi.952422992

3. Authority Data in Wikidata

We conducted an analysis of the Wikidata PCC (Program for Cooperative Cataloging) Pilot [4], which establishes a connection between the authority data and Wikidata. The PCC is an organization that provides support for the metadata of library and cultural heritage communities and seeks to facilitate the discovery and utilization of knowledge through collaboration. The Wikidata PCC pilot project is being overseen by the task group on identity management in Name Authority Cooperative Program (NACO) and is investigating the potential of utilizing Wikidata to align with the strategic goals of PCC. The Wikidata PCC Pilot Project occurred between 2020 and 2021, involving the collaboration of 43 institutions, which included university libraries and museums. It has been confirmed that a considerable proportion of these institutions continue to actively utilize Wikidata.

We conducted a thorough analysis of the three primary categories of institutions that participated in the Wikidata PCC pilot project. These categories included institutions that created a project page for Wikidata, institutions that utilized data from digital archive collections, and institutions that offered information such as data models on the project page. There is a restricted number of institutions that present data models, and it has been verified that there are 14 cases that provide Wikidata 'Person' property information and 5 cases that provides 'Corporate Body' property information.

3.1. Wikidata Person Property

There were a total of 153 Wikidata properties pertaining to person were utilized in the PCC pilot project, with 67 properties being utilized by two or more institutions out of the 14 institutions participating. The properties used by all 14 institutions include "instance of (P31)", "date of birth (P569)", and "date of death (P570)" (refer to Table 1).

Table 1
Number of Institutions in Use Wikidata Properties – Person

Number of Institutions in Use	Wikidata Properties	Number of Institutions in Use	Wikidata Properties
14	instance of (P31), date of birth (P569), date of death (P570)	13	occupation (P106)
11	place of birth (P19), place of death (P20)	9	sex or gender (P21), Library of Congress authority ID (P244)
8	family name (P734), country of citizenship (P27), given name (P735)	7	award received (P166), languages spoken, written or signed (P1412), pseudonym (P742), on focus list of Wikimedia Project (P5008), archives at (P485)

https://doi.org/10.23106/dcmi.952422992

Among the 14 institutions, there are a total of 15 Wikidata Person properties that are associated with resource IDs. The "Library of Congress authority ID (P244)" is used by 9 institutions, making it the most frequently utilized asset (refer to Table 2).

Table 2
Number of Institutions in Use Wikidata Properties ID

Number of Institutions in Use	Wikidata Properties	Number of Institutions in Use	Wikidata Properties
9	Library of Congress authority ID (P244)	5	VIAF ID (P214)
4	SNAC ARK ID (P3430)	3	Union List of Artist Names ID (P245)
2	Find a Grave memorial ID (P535)		

3.2. Wikidata Corporate Body Property

A total of 41 Wikidata properties were used by 5 institutions participating in the PCC Wikidata pilot, with 13 properties being utilized by two or more institutions. The "instance of (P31)" property was utilized by all 5 institutions, while the "country(P17)", "located in the administrative territorial entity (P131)", and "inception (P571)" properties were confirmed to be used by 4 institutions (refer to Table 3).

Table 3
Number of Institutions in Use Wikidata Properties- Corporate Body

Number of Institutions in Use	Wikidata Properties	Number of Institutions in Use	Wikidata Properties
5	instance of (P31)	3	industry (P452), dissolved, abolished or demolished date (P576), on focus list of Wikimedia project (P5008), archives at (P485)
4	country (P17), located in the administrative territorial entity (P131), inception (P571)	2	field of work (P101), headquarters location (P159), street address (P6375), official website (P856), part of (P361)

For corporate body identification of Wikidata properties, namely "Library of Congress authority ID (P244)", "Union of Artist Names ID (P245)", "VIAF ID (P214)", and "SNAC ARK ID (P3439)", have been utilized by the 5 institutions, with each property being used only once.

https://doi.org/10.23106/dcmi.952422992

3.3. Summary of Authority data in Wikidata

After analyzing the PCC Wikidata pilot, it was confirmed that multiple data from participating institutions are being contributed to Wikidata. It was observed that the emphasis in linking authority data with Wikidata is on person authority data rather than corporate body data. Furthermore, it has been noted that university libraries also contributed researchers' information to Wikidata.

The analysis of the case study entailed the examination of the Wikidata properties that were collected, in relation to their mapping with Describing Archives: A Content Standard (DACS), ISSAR-CPF: International Standard Archival Authority Record For Corporate Bodies, Persons and Families, and EAC-CPF (Encoded Archival Context for Corporate Bodies, Persons, and Families) (refer to table 4). The mapping has been confirmed that the Wikidata properties provide more detailed attributes than the archival data models, and it was observed that most of the properties are well mapped.

The "Library of Congress ID" has been established as a linking point, as confirmed by the fact that the PCC Wikidata pilot was primarily conducted in the United States.

Table 4

Mapping DACS-ISSAR-EAC(CPF) Wikidata Properties Example

DACS	ISAAR(CPF)	EAC(CPF)	Person – PCC Wikidata	Cooperation body- PCC Wikidata
10.1 Authorized Form of the Name (Required)	5.1.2 Authorized form(s) of name	<nameEntry> or <nameEntryParallel> with <authorizedForm>	family name (P734), given name (P735)	official name (P1448)
10.2 Type of Entity (Required)	5.1.1 Type of entity	<entityType>	instance of (P31)	instance of (P31)
10.3 Variant Forms of Names	5.1.3 Parallel forms of name	<nameEntryParallel>	pseudonym (P742), birth name (P1477)	
10.3.2 Standardized form of the name according to other rules	5.1.4 Standardized forms of name according to other rules	<nameEntry> or <nameEntryParallel> with <authorizedForm>	name in native language (P1559)	
10.3.3 Other forms of name	5.1.5 Other forms of name	<nameEntry> or <nameEntryParallel> with <alternativeForm>	noble title (P97), Honorific Prefix (P511), Nickname (P1449)	native label (P1705)
10.4 Identifiers for Corporate Bodies	5.1.6 Identifiers for corporate bodies	<entityID>	-	-

https://doi.org/10.23106/dcmi.952422992

10.5. Form of the Name Area of an Archival Authority Record	-	-	archives at (P485)	archives at (P485)

4. NDRM Authority Data Model

The data model of NDRM Authority has undergone a recent review and redesign in 2023. An analysis was conducted on the standard models used to manage person and corporate body in the library and archival sectors. The analysis primarily focused on the descriptive metadata derived from the authority – Wikidata model, which was developed during the NDRM-Wikidata pilot test. The aim was to identify the key components. In order to redesign the authority data Model, NDRM conducted an analysis of the guidelines for person/corporate body authority data and examined various cases of its construction and utilization to perform mutual mapping. Metadata elements were derived through cross-validation with the research team and digital archive manager of NDRM [10].

The metadata elements of the authority data of NDRM are classified into four categories: 'Identification', 'Description', 'Related Corporate Bodies, Persons, and Families', and 'Authority Recorded Management' (refer to Table 5). The metadata for the Identification category consists of 10 elements: GAC_identifier (NDRM identification), QID (Wikidata ID), OtherIdentifiers, ObjectType, Name, ChineseName, EnglishName, AlternativeName, Artname, and CoursteyName. The metadata for the Description Category comprises 14 elements: Nationality, BirthPlace, DeathPlace, Residence, Address, EstabilishedDate, TerminatedDate, BirthDate, DeathDate, Occupation, FieldOfActivity, History, RoleInNDRM, and References. The Metadata for Related Corporate Bodies, Persons, and Families Category includes 8 elements: FormerName, LatterName, ParentOrganization, SubsidiaryOrganization, RelatedPerson, Affiliation, CreatedReseource, and ContributedResource. The metadata for the authority recorded management consists of 4 elements: Creator, CreatedDate, ModifiedDate, and Note.

The redesign of the authority data model has proposed practical elements for a small-scale local archive. These elements aim to provide standardized access points for contextual information retrieval, ensuring interoperability with the data standard structure.

Table 5
Metadata for NDRM Authority Data

Category	Elements
Identification	GAC_identifier, QID, OtherIdentifiers, ObjectType, Name, ChineseName, EnglishName, AlternativeName, Artname, CourtesyName
Description	Nationality, BirthPlace, DeathPlace, Residence, Address, EstabilishedDate, TerminatedDate, BirthDate, DeathDate, Occupation, FieldOfActivity, History, RoleInNDRM, References

https://doi.org/10.23106/dcmi.952422992

Related Corporate Bodies, Persons, and Families	FormerName, LatterName, ParentOrganization, SubsidiaryOrganization, RelatedPerson, Affiliation, CreatedReseource, ContributedResource
Authority Recorded Management	Creator, CreatedDate, ModifiedDate, Note

5. NDRM Authority-Wikidata Mapping

A mapping table was developed to establish an interlinking between the NDRM authority data and Wikidata, using Wikidata properties obtained from PCC Wikidata pilot and NDRM authority data. The properties that were used more than twice in the PCC Wikidata pilot were assigned to the NDRM authority metadata elements (refer to Table 6). For the NDRM metadata elements that were not mapped to properties used more than once in the PCC Wikidata plot project, we prioritised mapping to properties that appeared once. Some elements were mapped after further searching and identifying properties within Wikidata. The "NDRM authority ID" was added as a Wikidata property because it was not previously existing as a property in Wikidata. The mapping does not include metadata in authority recorded management category because it is used for internal management.

Table 6
NDRM Authority-Wikidata Property Mapping Table

Category	Person		Corporate body	
	NDRM	Wikidata	NDRM	Wikidata
Identification	GAC_identifier	NDRM authority ID*	GAC_identifier	NDRM authority ID*
	OtherIdentifiers	Library of Congress authority ID (P244), VIAF ID (P214)*, National Library of Korea ID (P5034)*, Encyclopedia of Korean Culture ID (P9475)*	OtherIdentifiers	Library of Congress authority ID (P244), VIAF ID (P214), National Library of Korea ID (P5034)*, Encyclopedia of Korean Culture ID (P9475)*
	ObjectType	instance of (P31)	ObjectType	instance of (P31)
	Name	family name (P734), given name (P735), name in native language (P1559)	Name	official name (P1448)
	ChineseName, EnglishName,	alternative name (P4970)	ChineseName, EnglishName,	alternative name (P4970)

https://doi.org/10.23106/dcmi.952422992

	AlternativeName		AlternativeName	
	ArtName	art name (P1787)*	-	-
	CourtesyName	courtesy name (P1782)*	-	-
Description	Nationality	country of citizenship (P27)	Nationality	country of citizenship (P27)
	BirthPlace	place of birth (P19)	-	-
	DeathPlace	place of death (P20)	-	-
	Residence	residence (P551)	Address	location (P276)
	BirthDate	date of birth (P569)	EstablishedDate	date of official opening (P1619)
	DeathDate	date of death (P570)	TerminatedDate	dissolved, abolished or demolished date (P576)
	Occupation	occupation (P106)	-	-
	FieldOfActivity	field of work (P101)	FieldOfActivity	field of work (P101)
	History		-	-
	RoleInNDRM	subject has role (P2868)*	RoleInNDRM	subject has role (P2868)*
	References	-	-	-
Related Corporate Bodies, Persons, and Families	Affiliation	affiliation (P1416)	RelatedPerson	significant person (P3342)*
	CreatedResource ContributedResource	has works in the collection (P6379)	CreatedResource ContributedResource	has works in the collection (P6379)
	-	-	ParentOrganization	parent organization (P749) **
	-	-	SubsidiaryOrganization	has subsidiary (P355)
	-	-	FormerName	alternative name (P4970)
	-	-	LatterName	alternative name (P4970)

* is a Wikidata property that has never appeared in PCC Wikidata pilot, while ** is a property that has appeared once.

https://doi.org/10.23106/dcmi.952422992

6. Interlinking: NDRM-Wikidata

The process consisted of three stages: 1) data preparation, 2) donation and quality control, and 3) interlinking NDRM digital archive (refer to Figure 1). Cross-validataion was conducted among the researchers to ensure the accurate input data at all stages.

Figure 1: Wikidata Interlinking Process

The data in the "preparing donation data" was organized in a spreadsheet format using the NDRM authority-Wikidata Mapping table (refer to Figure 2). The data was prepared with a focus on descriptors for 264 person and 197 corporate body, and subsequently donated to Wikidata.

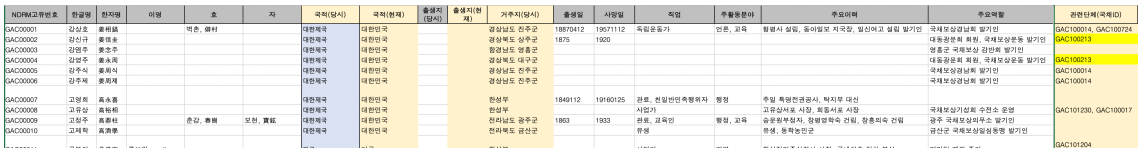

Figure 2: Preparing Donation Data-Spread sheet example

During the "donation and quality control" phase, the data collected in spreadsheet was transferred to Wikidata through the use of OpenRefine (See Figure 3). Additionally, a thorough examination of the source data pertaining to NDRM was carried out. In order to guarantee the dependability of the source data for the NDRM, we have made revision to the data and included it as a reference by comparing it with databases created by national research institutions, such as "Encyclopedia of Korean Culture" and "History Net", as well as refer to academic papers, conference proceeding, and online resource.

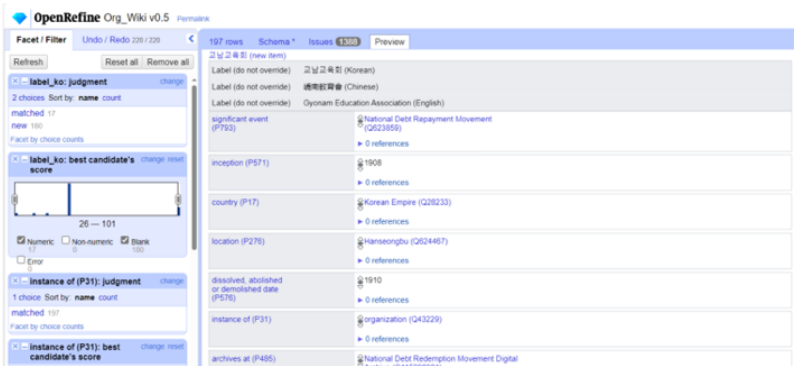

Figure 3: Organization Authority Data uploading example by OpenRefine

https://doi.org/10.23106/dcmi.952422992

During the Interlinking with NDRM archive phase, the Wikidata ID was integrated into the NDRM digital archive, which allowed NDRM digital archive user to access external data within the archive (as shown in Figure 4).

Figure 4: Interlinking NDRM Digital Archive and Wikidata

7. Results and Discussion

The data from NDRM digital archive was interlinking with Wikidata in this study, to improve its usability in line with the FAIR (Findable, Accessible, Interoperable, Reusable) principles. The interlinking with Wikidata is a strategy that enhances the accessibility and usability of the data in digital archive of NDRM on the web. In order to archieve this goal, we examined the PCC Wikidata pilot as a case study to determine applicable Wikidata properties and actively connected them with NDRM authority data through the process mapping. By reusing the properties (terminology), we can improve interoperability with other institutions. In addition, these best practices offer a structured framework for smaller organization, like NDRM, that aim to utilize Wikidata.

The key aspect of Wikidata donation and interlinking is the systematic management of the data in the digital archive for linkage to proceed smoothly. The management of authority data and identifiers in small-scale institutions continues to pose a challenge that has not been completely resolved.

In this study, we have developed a mapping table that is minimally intrusive and is based on case studies. It is intended for use by small-scale institutions and has been interlinking with Wikidata. We anticipate that this will enhance the utilization of Wikidata for small digital archives.

https://doi.org/10.23106/dcmi.952422992

Acknowledgements

This work was supported by the Commemorative Association of the National Dept Redemption Movement.

References

[1] A. Fagerving. Wikidata for Authority Control: Sharing Museum Knowledge with the World. Digital Humanities in the Nordic and Baltic Countries Publications 5.1 (2023): 222–239. doi.org/10.5617/dhnbpub.10665.

[2] A. Pal, P. Mukhopadhyay. Fetching Automatic Authority Data in ILS from Wikidata via OpenRefine. SRELS Journal of Information Management 59.6 (2022): 353-362. doi.org/10.17821/srels/2022/v59i6/170677.

[3] C. Bianchini, Carlo, S. Bargioni, C. C. Pellizzari Di San Girolamo. Beyond VIAF: Wikidata as a Complementary Tool for Authority Control in Libraries. Information Technology and Libraries 40.2 (2021). doi.org/10.6017/ital.v40i2.12959.

[4] Campaign: PCC Wikidata Pilot, URL: https://outreachdashboard.wmflabs.org/campaigns/pcc_wikidata_pilot/overview.

[5] F. Zhao. A Systematic Review of Wikidata in Digital Humanities Projects. Digital Scholarship in the Humatnities 38.2 (2023): 852-874. doi.org/10.1093/llc/fqac083.

[6] G. Prebor. From Authority Data, to Linked Open Data and Wikidata: The Case Study of a Hebrew Manuscript Catalogue. In Proceedings of the iConference 2020, iSchools, Berlin, DE, 2020. https://www.ideals.illinois.edu/items/114151.

[7] J. A. Clark, Helen K. R. Williams, D. Rossmann. Wikidata and Knowledge Graphs in Practice: Using Semantic SEO to Create Discoverable, Accessible, Machine-Readable Definitions of the People, Places, and Services in Libraries and Archives. Information Services & Use 42.3–4 (2022): 377–390. doi.org/10.3233/ISU-220171.

[8] S. Do, H. Park. A Study on Wikidata Linkage Methods for Utilization of Digital Archive Records of the National Debt Redemption Movement. Journal of Korean Society of Archives and Records Management 23.2 (2023): 95-115. doi.org/10.14404/JKSARM.2023.23.2.095.

[9] S. Han, H. Park. A Study on Wikidata Utilization for Digital Archives. Journal of Korean Society of Archives and Records Management 22.1 (2022): 201-217. doi.org/10.14404/JKSARM.2022.22.1.201.

[10] S. Han, S. Do. A Study on Metadata Design for Managing Person and Organization Names in the National Debt Redemption Movement Digital Archive. Journal of the Korean Society for Information Management 41.1 (2024): 509-536. doi.org/10.3743/KOSIM.2024.41.1.509.

[11] S. Park. A Study on the National Dept Redemption Movement and Digital Humanities Contents. Journal of Daegu Gyeongbuk Studies 19.3 (2020): 129-147 doi.org/10.23029/jdgs.2020.19.3.129