Ontological Data Sharing of Open Government Data for Data Curation/Le partage ontologique de données gouvernementales ouvertes aux fins de la conservation des données

Richard P. Smiraglia, Hyoungjoo Park

Canadian Journal of Information and Library Science, Volume 41, Number 4, December / décembre 2017, pp. 285-307 (Article)

Published by University of Toronto Press

Ontological Data Sharing of Open Government Data for Data Curation

Le partage ontologique de données gouvernementales ouvertes aux fins de la conservation des données

Richard P. Smiraglia
School of Information Studies, University of Wisconsin—Milwaukee
smiragli@uwm.edu

Hyoungjoo Park
School of Information Studies, University of Wisconsin—Milwaukee
park32@uwm.edu

Abstract: The purpose of this study is to inform ontological data-sharing practices in the cultural heritage community by using open government data for data curation. Open government data have grown rapidly to enhance civic engagement and the transparency of governmental authorities in many parts of the world. We used a mixed methods approach, including co-word analysis of a population of open government data followed by qualitative analysis of five carefully selected records mapped using the CIDOC CRM. A final step was to use FRBRoo, an extension of the CIDOC CRM, to map the instantiation of data records in a typical data-sharing scenario.

Keywords: data curation, open government data, cultural heritage, CIDOC CRM, FRBRoo

Résumé : Le but de cette étude est de faire connaître les pratiques ontologiques de partage de données dans la communauté du patrimoine culturel en utilisant des données gouvernementales ouvertes, aux fins de la conservation. Les données gouvernementales ouvertes ont augmenté rapidement dans de nombreuses régions du monde, renforçant l'engagement civique et la transparence des autorités gouvernementales. Nous avons utilisé une approche par méthodes mixtes, dont une analyse des cooccurences de termes au sein d'une population de données gouvernementales ouvertes, et une analyse qualitative de cinq notices soigneusement sélectionnés en utilisant le modèle conceptuel de référence CRM CIDOC. La dernière étape a consisté à utiliser FRBRoo, une extension du CRM CIDOC, pour cartographier l'instanciation des enregistrements de données dans un scénario de partage de données typique.

Mots-clés : conservation des données, données gouvernementales ouvertes, patrimoine culturel, CIDOC CRM, FRBRoo

Introduction

The purpose of this study is to inform ontological data-sharing practices in the cultural heritage community by using open government data for data curation. The research problem stems from the current rapid increase in the quantity of open government data in the cultural heritage domain. However, many cultural heritage open government data cannot be shared because those data are "stuck" in siloed, unstructured, and heterogeneous file formats, data formats, or Web addresses. For example, heterogeneous file formats range from MS Excel and HWP (Hangul Word Processor), to XML (eXtensible Markup Language), RDF (Resource Description Framework), and PDF (Portable Document Format) and to Web addresses simply linked to other open government websites. Governmental investments in making open government data available to the general public are increasing. But data sharing of cultural heritage open government data, which could create new value for the general public, has not been actively studied. Also, there has been almost no research on knowledge organization with regard to ontological data sharing using open government data. Knowledge organization as a science has recently been defined as "The evolving science of the order of knowledge, which informs the application of functional knowledge organization systems … Knowledge Organization works with shared perceptions represented by conceptual phenomena" (Smiraglia 2015b, 1).

A primary research question, then, is how to share cultural heritage open government data that might be stuck in heterogeneous and unstructured formats by using ontologies for data curation. With this study, we hope to demonstrate a knowledge organization approach to sharing open government data in the cultural heritage domain to create new accessibility for the general public.

Background

Open government data

Open government data refer to "non-privacy-restricted and non-confidential data which is produced with public money and is made available without any restrictions on its usage and or distribution" (Janssen, Charalabidis, and Zuiderwijk 2012, 309). Open government data open to the general public have grown rapidly in quantity to enhance civic engagement and the transparency of national, state, and local governmental authorities in many parts of the world, including Austria, Australia, New Zealand, the Republic of Korea, Sweden, the United Kingdom and the United States.

However, means of enhancing data sharing in the cultural heritage domain using open government data available to the general public have not been actively studied. Among cultural heritage institutions, libraries, archives, and museums usually are included. Open government data sharing has focused on structured data sets such as RDF formats (e.g., linked data). With regard to the increasing necessity of data sharing of open government data relevant to cultural heritage information, only a few studies have been conducted because of unstructured data sets that are found in varying heterogeneous formats. This paper

presents one of a series of case studies in which open government cultural heritage data records are mapped for data sharing using a metalevel ontology.

Data curation

Data curation enables data sharing throughout the data-management life cycle to create new value for new user needs. One promising approach to data sharing in the cultural heritage domain is the metalevel ontology known as the CIDOC-CRM (The International Committee for Documentation [CIDOC] Conceptual Reference Model [CRM]). The CIDOC-CRM is an object-oriented and extensible event-centred ontology for knowledge sharing in cultural heritage (CIDOC-CRM Special Interest Group, 2015). The CIDOC-CRM has been designated an international standard by the International Organization for Standardization (ISO 21127:2014). The CIDOC-CRM uses classes and property definitions in a formal structure to describe concepts and relationships. Because the CIDOC-CRM is empirically derived, it is particularly well suited to intercultural information sharing. Cultural heritage applications of the CIDOC-CRM have demonstrated its usefulness in enhancing heterogeneous cultural heritage data in the form of a semantic framework. Data curation can consist of mapping data descriptions using the ontological structure of the CIDOC-CRM as an equalizing filter for data sharing. Records created in one repository can be shared among all repositories that utilize the CIDOC-CRM.

CIDOC-CRM mapping using structured data

Several studies have been conducted using the CIDOC-CRM ontologies to summarize structure and combine existing data in a generic ontological model and to map and formalize existing knowledge for interoperability with other data sets. Semantic mapping of cultural heritage between Encoded Archival Description (EAD) and the CIDOC-CRM ontology also has been studied (Bountouri and Gergatsoulis 2011; Gergatsoulis et al. 2010). Koutsomitropoulos, Solomou, and Papatheodorou (2009) conducted a case study to implement a prototype with Dublin Core metadata in digital object collections. Lin, Hong, and Doerr (2008) studied inference platforms by using OWL for mapping through the CIDOC-CRM for cultural heritage digital libraries. Stasinopoulou et al. (2007) investigated ontology-based metadata using the CIDOC-CRM; Theodoridou et al. (2010) studied CIDOC-CRM digital (CIDOC-CRMdig) in Resource Description Framework Schema (RDF/S) for modelling and querying provenance. Doerr and Iorizzo (2008) discussed how heterogeneous cultural heritage information could be enhanced using the CIDOC-CRM due to its flexibility for integration, mediation, and interchange. They presented a case study using core ontologies of the CIDOC-CRM. This study demonstrated the difficulties of linking documents and knowledge. For that reason, semi-automatic co-reference detection and correction were needed to match between identifiers and resources, because direct data input into a semantic network revealed ineffectiveness.

Ontological data sharing using open government data

Previous studies regarding open government data sharing using ontologies have focused on structured data sets rather than unstructured data sets in heterogeneous formats. Despite the increasing importance of ontologies in the Semantic Web, open government data -sharing through ontologies for data sharing has not been studied. A few published studies have applied ontologies as a means of enhancing data sharing using open government data (Fragkou, Galiotou, and Matsakas 2014; Park and Smiraglia 2014). Ontologies such as the Greek E-GIF ontology (eGovernment Knowledge Interoperability Ontology) and the CIDOC-CRM were studied using open government data and the nature of each ontology for data sharing was discussed (Fragkou, Galiotou, and Matsakas 2014). Park and Smiraglia (2014) found that current CIDOC-CRM ontologies support addresses only for Western address systems (i.e., assigning addresses based on street-house number system) rather than East Asian address systems (i.e., assigning addresses chronologically rather than geographically).

Knowledge organization systems (KOSs), such as ontologies, can be used as tools for increasing access to and dissemination of government information to wider audiences for the public good (Hodge 2014). Ontological mapping of open government data records into the CIDOC-CRM can facilitate wider access because the CRM's empirically derived, event-based ontology is culturally neutral. Moreover, the CRM's compatibility with the RDF, a backbone of the Semantic Web, can help to generate greater access to open government data through the Semantic Web. The case study reported here is a demonstration of the power of the CRM to contribute to the general public through the network environment, freeing open government data from the silos of its naturally heterogeneous formats.

Methods

We used a mixed methods approach, in which we used quantitative co-word analysis of a population of open government data to empirically determine characteristics of the data records. The overall design of the study is represented visually in figure 1. Co-word analysis often is used as a domain analytical technique in knowledge organization research because it can be used with a large body of textual data to extract a core ontology for a domain. Domain analysis in knowledge organization is formulated as a methodological paradigm for discovering knowledge bases and revealing shared ontologies (Smiraglia 2015a; the methodology is described fully in chapter 6). Thus, our first step was to subject a population of Korean open government data to co-word analysis to improve our understanding of the knowledge base represented by the data elements. This analysis reveals key terms and phrases that co-occur and that recur with high frequency in the data set. Aside from the informative value of the co-word analysis, this is a very pragmatic approach to undertaking CIDOC-CRM mapping by associating key concepts with classes and properties. We followed this with a qualitative analysis of five carefully selected records, which we mapped using the CIDOC-CRM. A final step in the case study was

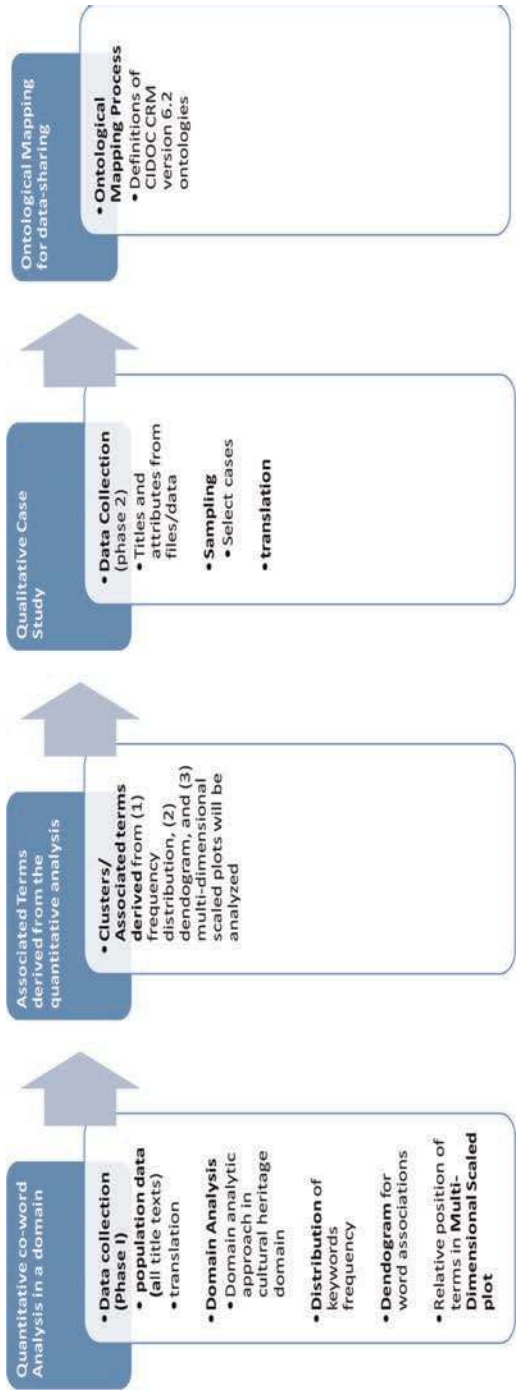

Figure 1: Process of research design

to use FRBRoo, an extension of the CIDOC-CRM, to map the instantiation of data records in a typical data-sharing scenario.

Data collection was conducted from the official Korean Open Government Website (http://data.go.kr/) in two phases. In the first phase of data collection, all records in three cultural heritage categories—libraries, archives, and museums—were downloaded. All records were translated from Korean to English by the researchers before the co-word analysis. Each file or data set in any heterogeneous format was opened to identify representative cases for the qualitative phase by analyzing attributes of distinctive records in each file or data set, again in consultation with the results of the co-word analysis.

Results

Quantitative co-word analysis of titles

Co-word analysis involves the use of software for extracting keywords, terms, and phrases from a body of text for subsequent analysis. The terms extracted can be displayed in frequency distributions that help separate the core of the knowledge base—those most frequently occurring—from the granularity, made up of the traditional "long tail" of a Bradford-type distribution. (This methodology is fully described in Smiraglia 2015b, 74–81.) We used Provalis Research's ProSuite (https://provalisresearch.com/) by first entering all of the translated records into QDA Miner. Then, using the WordStat module, we generated frequency distributions of keywords and frequently occurring phrases. We also used the KWIC (keyword-in-context) function to cross-check the occurrence of multiword terms with the phrase lists. In this manner, we were able to generate basic taxonomies, which can in turn be used to generate visualizations using multidimensional scaling. Goodness of fit requires low stress and high R^2 to develop a visualization that matches the actual term co-occurrence in the knowledge base well. Removing single-occurrence clusters sequentially improves goodness of fit.

We analyzed each segment—archives, libraries, and museums—separately and then in combination to derive diverse views of the data set. The archives data set was the smallest, with only seven records containing only 26 keywords, of which only 3 occurred more than once: archives, digital, and list. No multiword phrases were identified. Using the KWIC feature, we learned that the only recurring term was "digital archives." The libraries data set was much larger, with 42 records containing 3,840 keywords, of which 428 were unique. The phrase-finder revealed 3,375 phrases, of which 116 occurred 3 or more times, and 24 occurred 14 or more times (or more than 2.5% of the total). These were used to generate the visualization shown in figure 2 (stress = 0.22536; $R^2 = 0.9250$).

The visualization thus contains only the most prominent categories in the knowledge base, but reveals the diverse emphases in the data set. For example, library contents are important especially with regard to book holdings and conditions and new book arrivals. Continuing education programs are prominent. Both public libraries and digital libraries are central but distinct. Not shown

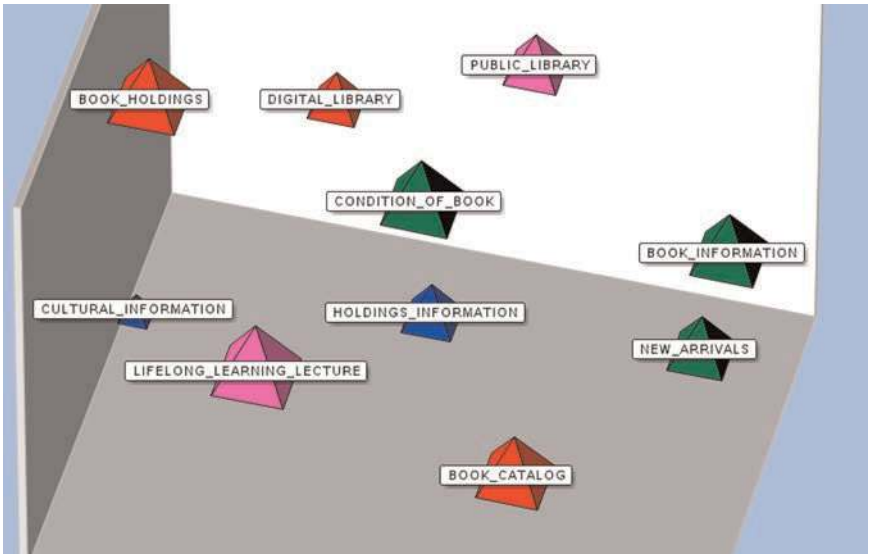

Figure 2: MDS visualization of core terms from the libraries data set

here is the prominence of library information relevant to geography (e.g., province, city, and district), and library size and type in the Republic of Korea. To help provide context, WordStat has a keyword-in-context (KWIC) feature, which can be used to see how the core keywords have been used. For example, "book," the keyword with the highest frequency, is used predominantly in phrases such as "book database [list, information] in [place]" and in conservation statements such as "condition of book at [place]." From this simple analysis, we can see that the next several keywords combine to form a cluster about "book [information or condition] in [place]." This suggests that the CRM classes that will be most applicable are those identifying institutions (place appellation, legal body) and holdings (manmade object, information object) and various actors.

Similarly, the museum data set was analyzed. In this set there were 1,543 keywords, of which 249 were unique. There were 989 multiword phrases, of which 42 occurred three or more times and 16 occurred five or more times (or more than 2.4% of the total). These were used to generate the visualization shown in figure 3 (stress = 0.14649; $R^2 = 0.9636$).

The visualization shows only one cluster, suggesting homogeneity in the data set, and that is dominated by the word "museum." The KWIC feature shows a predominance of statements of the form "[place name] Museum [gallery, information]" or "[place name, city, province, or district]." A small cluster of terms including "cultural" combines that word with "facility," "assets," "heritage," or "space." In addition to the CRM classes identified for libraries, there also is a need for "curation activity," "time span," and many simple "strings" giving actual information concerning the institutions, holdings, and activities.

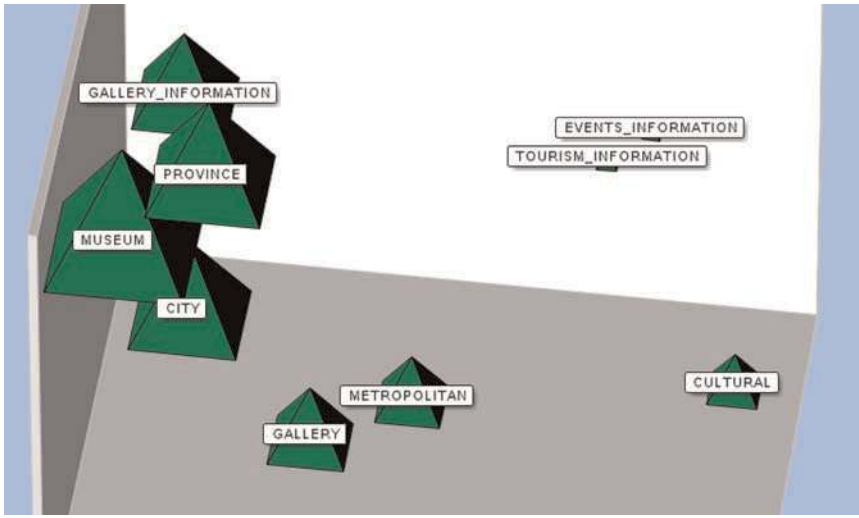

Figure 3: MDS visualization of core terms from the museums data set

Qualitative case study

Five cases were selected from the data downloaded from the Korean Open Government website for qualitative case study. Five cases were selected from the archives, libraries, and museums data sets, with care in the latter two instances to select both examples of management data and holdings data. The five cases are (1) archives of digital advertisement; (2) museum facility; (3) museum artifacts; (4) library of the Korea Research Institute of Bioscience and Biotechnology—booklist (as of March 2014); and (5) library management (August 30, 2015). See figure 4.

The cases then were mapped using the CIDOC-CRM, version 6.2. Mappings were based on those reported by Park and Smiraglia (2014), which were verified by key experienced CRM trainers. These are shown in figures 5–9.

Figure 5 displays the CRM mapping of "Archives of Digital Advertisement," a record from an archive of advertising collections in Korea. The record includes two cartoon advertisements. The first is titled "ChoonHyang and DoRyeong Lee," which is associated with a well-known Korean love story and folk tale that might be compared to the story of Romeo and Juliet in western culture. Just as "Romeo and Juliet" is associated with many iterations ranging from Shakespeare's original play to book titles, musical works, classical characters, and the like, the story of "ChoonHyang and DoRyeong Lee" here is used as the basis of a cartoon advertisement for LakHeui Chemicals. In figure 5, "ChoonHyang and DoRyeong Lee (E35 Title)" is the title of the digital adverting collection. The creator of this digital advertising is cartoonist (E39 Actor) Dal Boo Moon (E82 Actor Appellation). The duration (E54 Dimension) of the digital advertisement is 180 minutes (E52 Time-Span; E58 Measurement Unit),

Archives of digital advertisement

NO	Production Year	Production Month	Seconds	Main Category	Sub category	Minor category	Advertise	Product	Title	Agency	Production company
kj-003	1956		180	group and corporate advertising	Group advertisement	Group PR	Lakheui Chemicals	Total products	Choonhyang and Doryeong Lee	cartoonist Dal Boo Moon	
JEIL 10-27	1982	9	30	group and corporate advertising	Group advertisement	Group PR	SAMSUNG Electronics Limited	SAMSUNG Electronics	For overseas	Cheil Worldwide	

Museum facility

Name	Location	Telephone	Admission fees	Hours	Collections
Museum of Education	Wonmi-gu Sosa-ro 482 (Chooneui-dong)	661-1282	charged	. Open 09:00-18:00 . Closed : Mondays, new year, Chooseok, next day of national holidays	• Size : 672.94㎡ • 4,712 collections from modern times to current such as textbook and reference books • Exhibitions : education in the traditional era→education supplies→classroom in the 60s and the 70s → students' life at school→reference room
Museum of bow	Wonmi-gu Sosa-ro 482 (Chooneui-dong)	614-2678	charged	. Open 09:00-18:00 . Closed : Mondays, new year, Chooseok, next day of national holidays	• Size : 811.67㎡ • Exhibitions, theater, demonstration, experience class, etc. • 467 collections • various programs such as making bows during vacation

Museum artifacts

Artifact name	Size	Holding institution	Artifact number	Description	Nationallity1	Nationality2	Findspot1-1	Material 1-1	Material1-2	Usage function1-1	Usage function1-2	Usage function1-3	Collection 1	Genre1
Banja	Maximum diameter+36.2+Thickness+7.4	National1/Gyeongju	Chrysanthemum 000044-000	Banja is a metal Buddhist percussion. It is used to call the public in the temple or to inform the urgent things, and it is still...	Korea	Goryeo	North Gyeongsang Province	Metal	copper alloys	religious belief	Buddhism	ceremony	National1	study of ancient culture and heritage of the country
Floral design celadon	Height 35.2cm	National1/central	Desheng 000343-000		Korea	Goryeo	Gyeonggi Province-near Kaesong	celadon					National1	

Library of the Korea Research Institute of Bioscience and Biotechnology - Book lists

Registration number	Title	Author	Publisher	Publication Year	Classification	Books	Volume	Copies
7006762	Environmental aspects and management of hyporheic zones	Hyeon Yoon Jeong	Korea Environment Institute	2013	KEI 2013-2			
0000001	Preparation and assay of enzymes	Colowick, Sidney P.	Academic Press	1955	QP601	.M49 1955	v.1	

Library Management (August 31, 2015)

Library name	City/Province	City/Gu/Gun	Library type	Closed	Open	Closed	Seats	Resources-Books	Check-out availability	Checkout length	Address	Management agency	Telephone	Creation date
BaekHab Children' Library	Seoul city	YongSan-gu District	Small-sized private library	Sunday	10:00	17:00	32	8863	3	7 days	19-2 Huam-ro 15-gil	BaekHab Church	755-0525	2015.8.31
Book Café CheongMaRu	Seoul city	YongSan-gu District	Small-sized public library	Saturday, Sundays, Holidays	9:00	19:00	50	4000			Noksapyeong-daero	Yongsan-gu District	2199-8967	2015.8.31

Figure 4: Five open government records

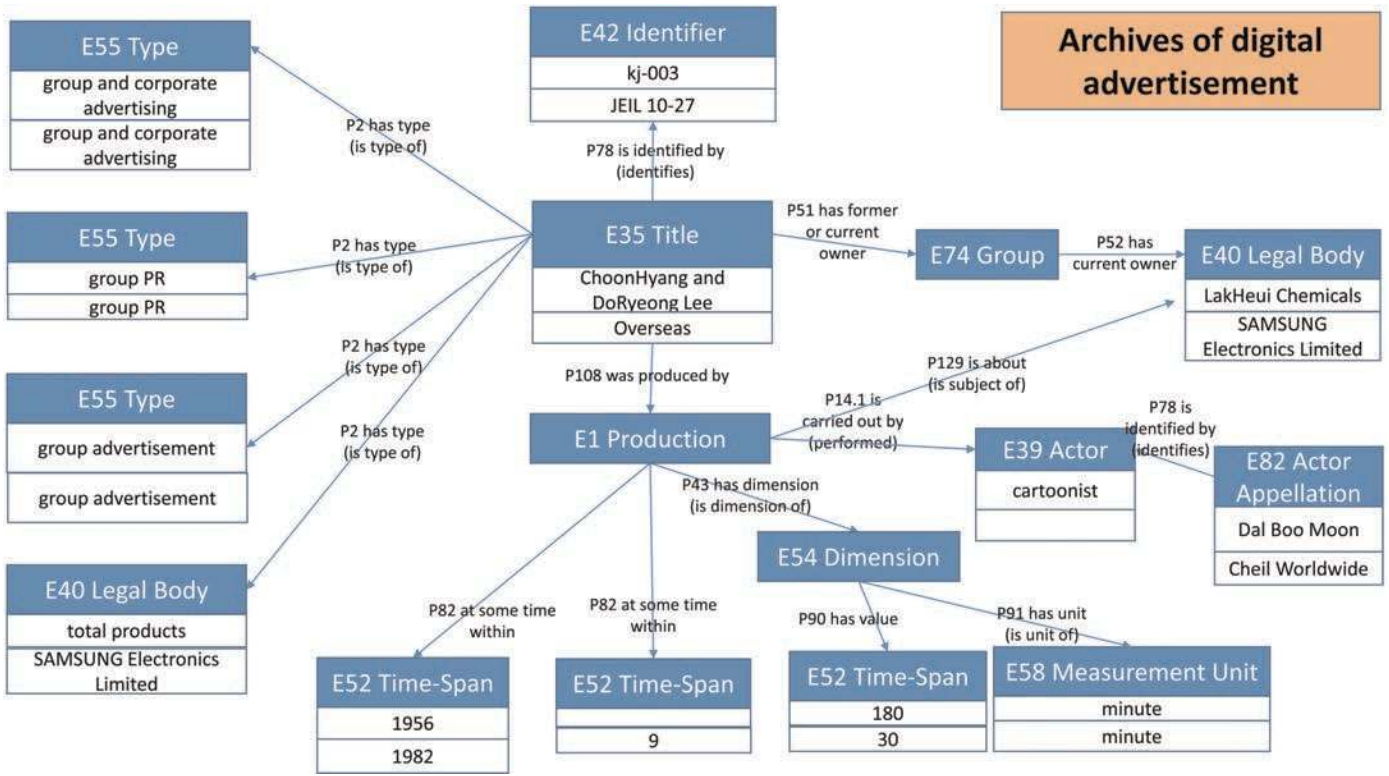

Figure 5: CIDOC-CRM map of ''archives of digital advertisement''

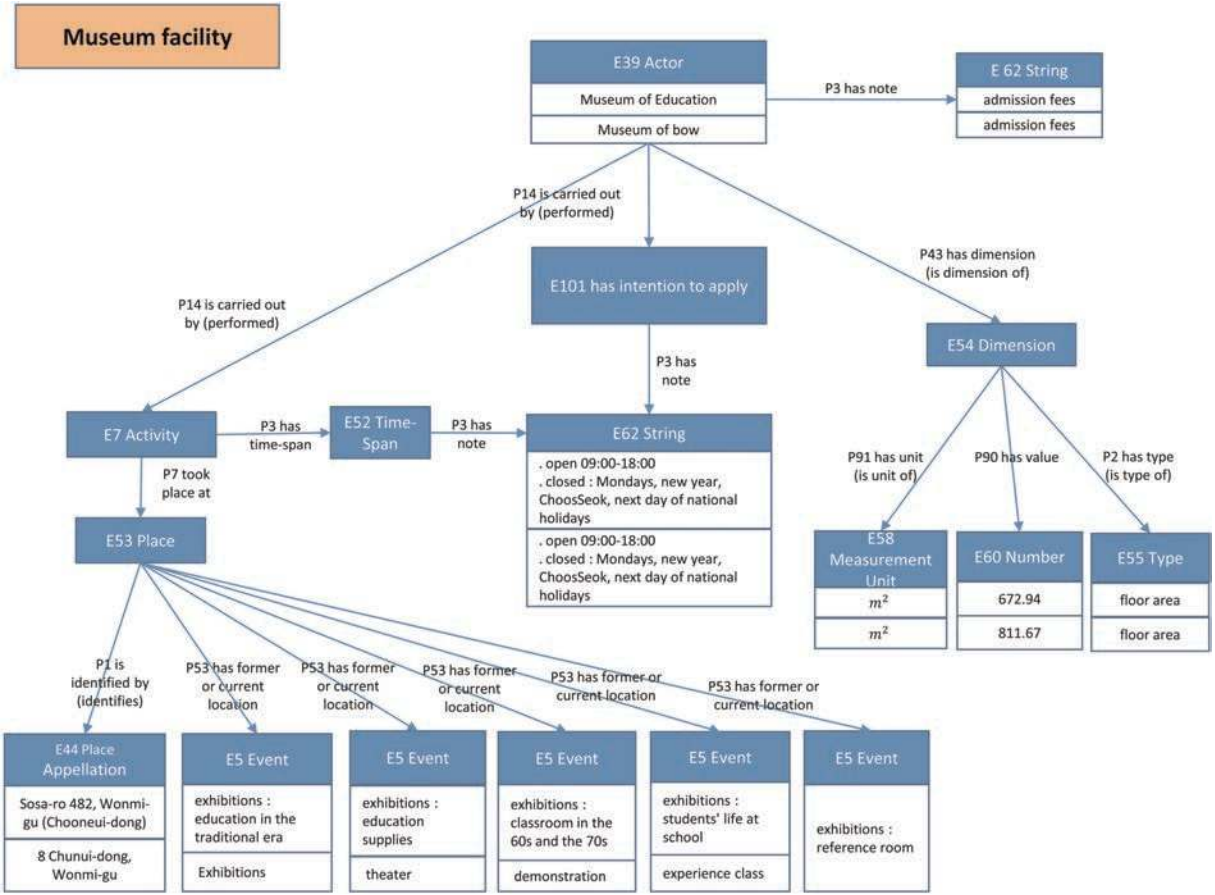

Figure 6: CIDOC-CRM map of "museum facility"

and this digital advertisement was produced (E1 Production) in 1956 (E52 Time-Span). The advertiser (E74 Group) is LakHeui Chemicals (E40 Legal Body). The Archives of Digital Advertisement specify this advertisement as "group and corporate advertising (E55 Type)," "group PR (E55 Type)," and "group advertisement (E55 Type)" for LakHeui Chemicals (E40 Legal Body). The original is accompanied by an instantiation for overseas release.

Figure 6 displays the CRM mapping of "museum facility," a record for use by both museum facility administrators and museum users. For example, the Museum of Education (E39 Actor) opens from 9:00 to 18:00 and closes on Mondays, New Year, ChooSeok, and the day following ("next day"), national holidays (E62 String). The dimension of the museum (E54 Dimension) give a floor area (E55 Type) of 672.94 square metres (E60 Number; E58 Measurement Unit). Museum users need to pay for admission (E62 String). The Museum of Education provides some activities (E7 Activity) to the public, which take place at the facility with the address "Sosa-ro 482, Wonmi-gu (Chunui-dong)" (E44 Place Appellation). The Museum also has events such as "exhibitions: education in the traditional era (E5 Event)," "exhibitions: education supplied (E5 Event)," "exhibitions: classrooms in the 60s and the 70s (E5 Event)," "exhibitions: students' life at school (E5 Event)," and "exhibitions: reference room (E5 Event)." The "Museum of Bow" is the Bucheon Bow Museum in the same location, with similar facility information. In both cases, the record is mapped using the CRM mapping for a "plan" (*E101 Has intention to apply*).

Figure 7 displays the CRM mapping of "museum artifacts," a record from one of Korea's national museums. This record provides specific information about specific museum artifacts: artifact name, dimensions, materials, provenance (i.e., "find spot"), physical details, purpose of the artifact, holding museum, and accession number. The first example is a banja, a metal Buddhist percussion instrument, donated by low-ranking local officials to a temple to wish for the long life and good health of the king. The second example is a floral design celadon, a celadon jar with floral medallion design in relief. The banja (E22 Man-Made Object) has identification number "chrysanthemum 000044-000" (E42 Identifier) and consists of a metal (E57 Material), specifically a copper alloy (E57 Material). The dimension of the banja is its diameter (E54 Dimension; E55 Type), 36.2 cm (E60 Number; E58 Measurement Unit), and its thickness (E55 Type), 7.4 cm (E60 Number; E58 Measurement Unit). The banja (E22 Man-Made Object) was created during the Goryo dynasty (E4 Period) and is a national cultural asset (E55 Type) of Korea (E74 Group). The national museum of Gyeongju (E87 Curation Activity), which is located in the North Gyeong-Sang Province (E53 Place), currently curates the banja. The banja was used for Buddhist religious ceremony (E17 Type Assignment) and represents the "study of ancient culture and heritage of the country" (E62 String). The banja is "is used to call the public in the Buddhist temple or to inform the urgent things" (E62 String). Fewer details are given about the floral design celadon, which is held by the same museum.

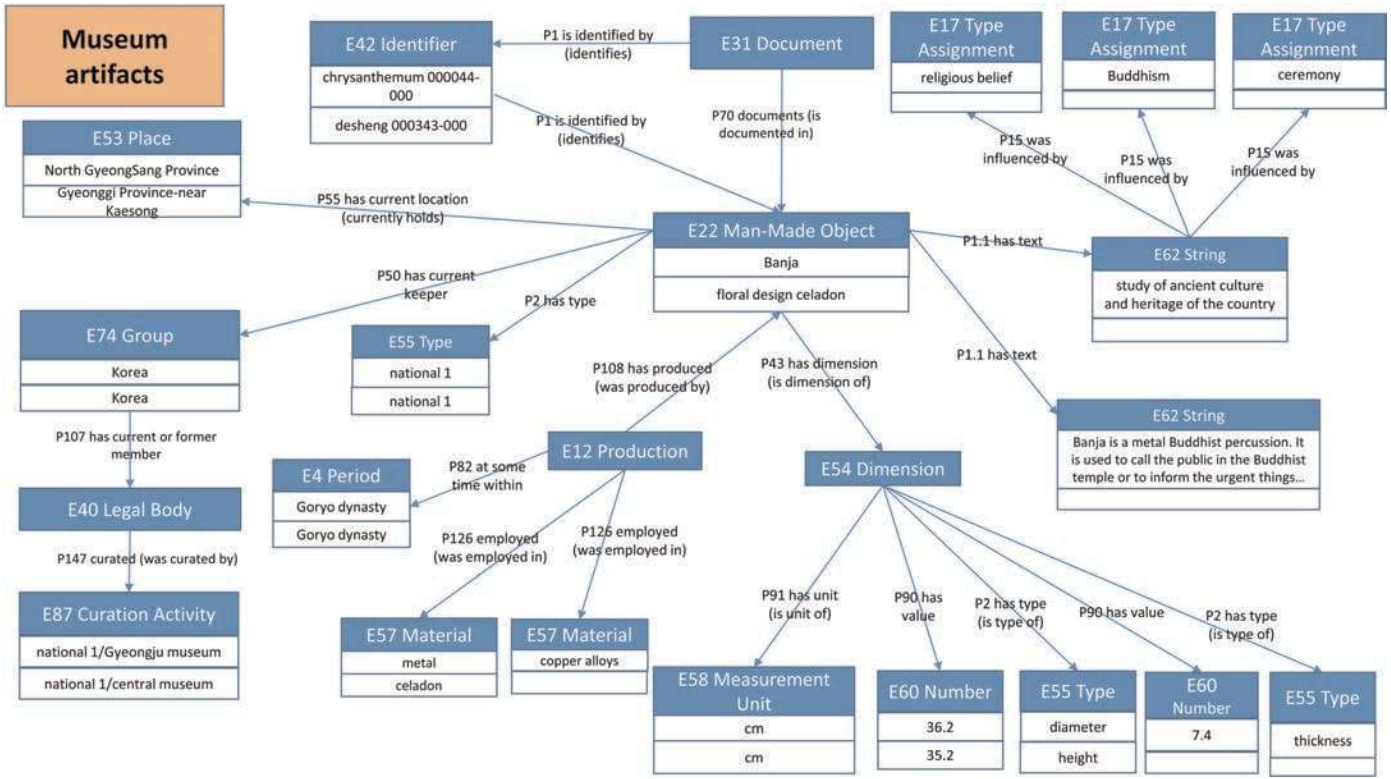

Figure 7: CIDOC-CRM map of "museum artifacts"

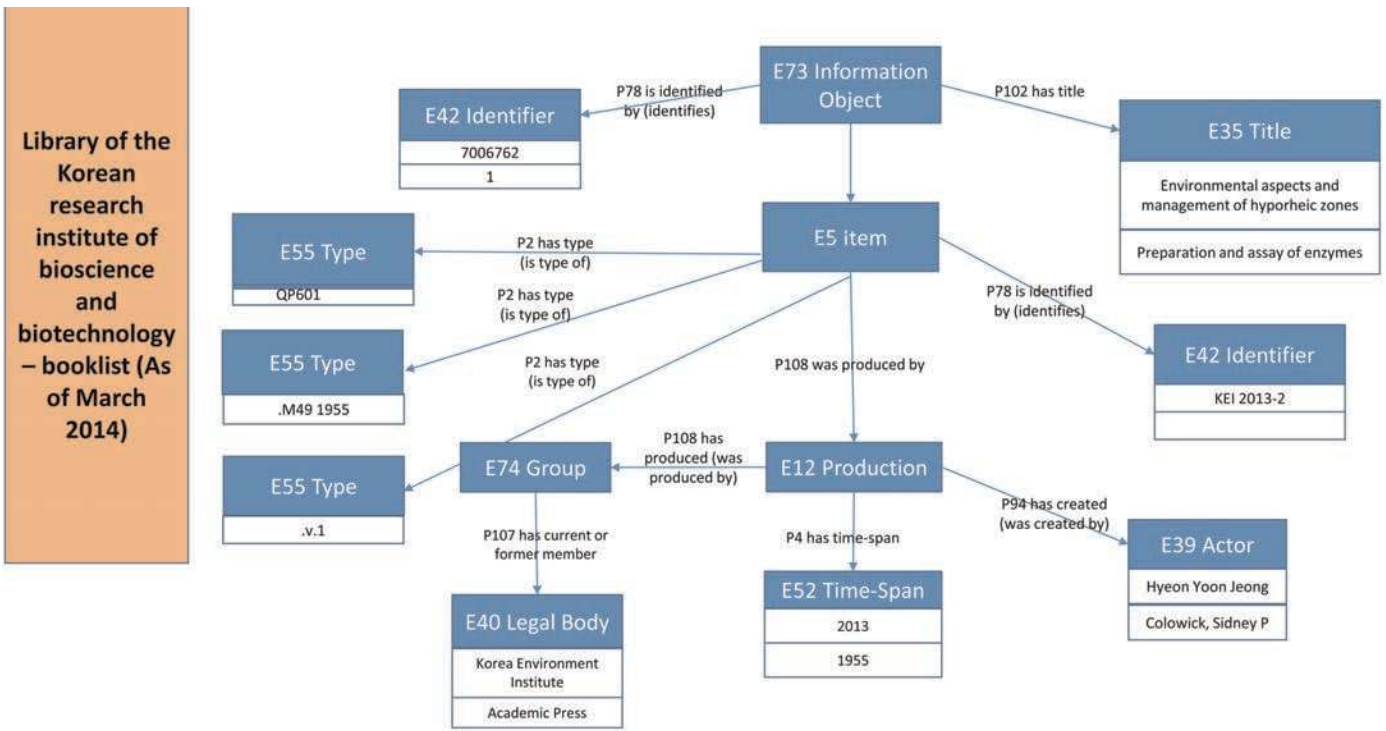

Figure 8: CIDOC-CRM map "Library of the Korea Research Institute of Bioscience and Biotechnology—booklist (as of March 2014)"

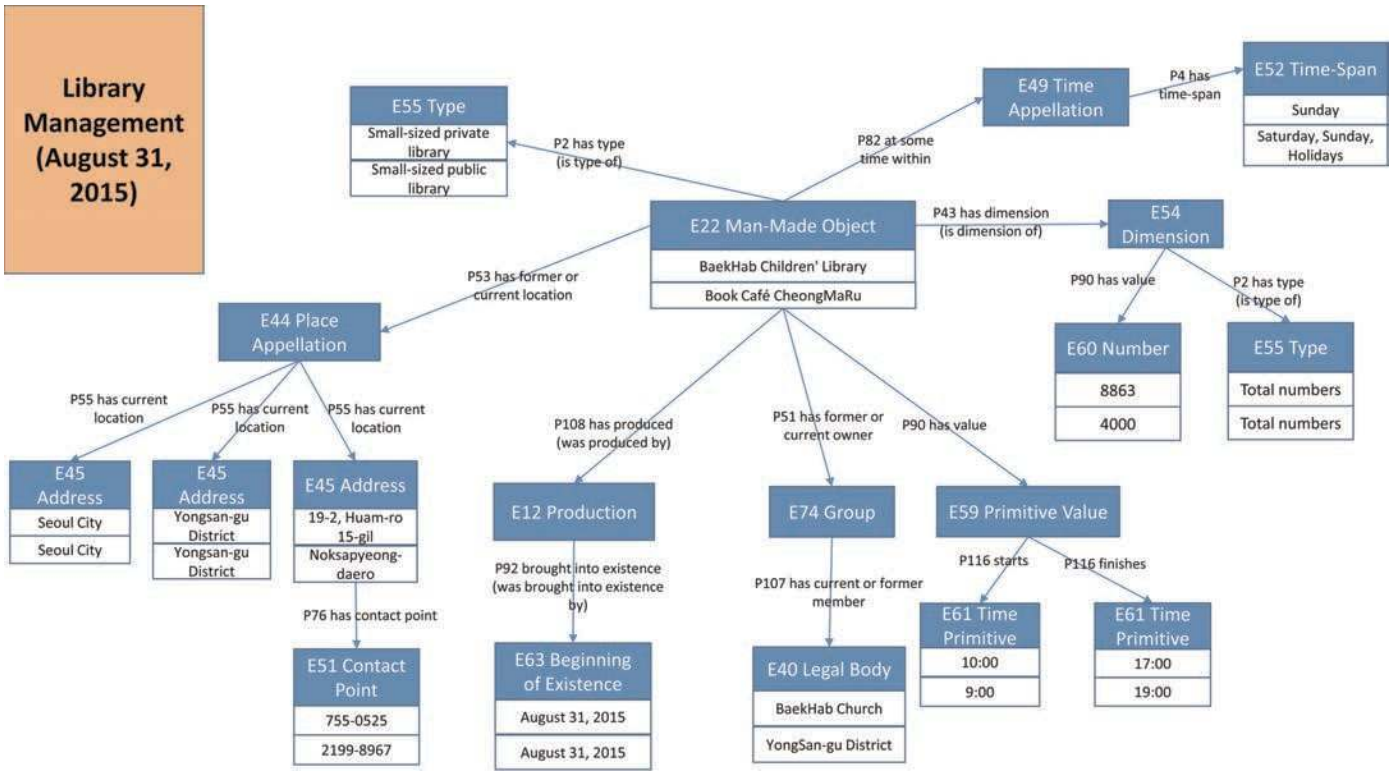

Figure 9: CIDOC-CRM map of "library management"

Figure 8 displays the CRM mapping of the "Library of the Korea Research Institute of Bioscience and Biotechnology—booklist (as of March 2014)," a special library in Korea. The record contains details about a journal article. The record for the article (F5 Item) has a record identification number, 7006762 (E42 Identifier), and a library article identification number, KEI 2013–2 (E42 Identifier). The title (E35 Title) of this article is "Environmental Aspects and Management of Hypothetic Zones," the author (E39 Actor) is Hyeon Yoon Jeong, and its publisher (E40 Legal Body) is the Korea Environment Institute (KEI). The library call number for this article uses the Library of Congress classification symbol QP601 (E55 Type; general works on enzymes within animal biochemistry), a local title number M49, date 1955 (E44 Type), and v. 1 (E55 Type).

Figure 9 displays the CRM mapping of "library management." The record is useful for managerial purposes because the data set includes library type, library hours, total holdings, total numbers of seats, organization that operates (or owns) the library, location, and construction year. For instance, the BaekHab Children's Library (E22 Man-Made Object), a small public library (E55 Type), was constructed (E12 Production) on August 31, 2015 (E63 Beginning of Existence) and is operated by Yong San-gu District (E40 Legal Body). The address (E44 Place Appellation) of this library is "19-2 Huam-ro 15-gil" (E45 Address), "Yongsan-gu District" (E45 Address), and "Seoul City" (E45 Address). The telephone number is 755–0525 (E51 Contact Point). Library hours are from 10:00 to 17:00 (E61 Time Primitive) and the library is closed on Sundays (E52 Time-Span). The total number of seats in the library is 8,863 (E54 Dimension; E60 Number) seats (E55 Type). Similar details are given for the Book Café CheongMaRu.

Discussion

CIDOC-CRM and FRBRoo ontologies in KO

The mappings shown above are simple and straightforward. CIDOC-CRM is an event-centred ontology for cultural heritage. In our data sets, books (as artifacts) are controlled according to their acquisition, holdings availability, or condition in a diverse but related set of government repositories. The data map quite well to the CRM, which tells us that scalability—extending the mapping to the entire set of repository data—is possible.

Of course, an interesting problem then emerges, which is how to control the instantiation of the data. Instantiation is a phenomenon of information objects that takes place when the same content has several carriers (Park and Smiraglia 2014). For example, the text of this article exists in manuscript in Microsoft Word, but also in the PDF format. If we map our data sets using CRM, we create an instantiated set of mapped data. How, then, can an information system disambiguate mapped data?

An extension of the CIDOC-CRM is the FRBRoo ontology (Functional Requirements for Bibliographic Records—object-oriented; http://www.cidoc-crm.org/frbr/). The FRBR conceptual model is based on understanding that an

Table 1: FBRRoo/CRM Mapping of Archives of Digital Advertisement

	FRBRoo	CRM	Value 1	Value 2
Number	F13 Identifier	E42 Identifier	kj-003	JEIL 10–27
Production year	F30 Publication Event	E52 Time Span	1956	1982
Production month	F30 Publication Event	E52 Time-Span		9
Seconds	F30 Publication Event	E52 Time Span	180	30
Main category		E55 Type	Group and corporate advertising	Group and corporate advertising
Subcategory		E55 Type	Group advertisement	Group advertisement
Minor category		E55 Type	Group PR	Group PR
Advertise	F11 Corporate Body F12 Nomen	E40 Legal Body	Lakheui Chemicals	Samsung Electronics
Product	F11 Corporate Body F12 Nomen	E40 Legal Body	Total products	Samsung Electronics
Title	F12 Nomen F35 Nomen Use Statement	E35 Title	Choonhyang and Doryeong Lee	For overseas
Agency	E10 Person	E39 Actor E82 Actor Appellation	Cartoonist Dal Boo Moon	Cheil Worldwide

abstract "work" takes intellectual form in its "expressions," published form in its "manifestations," and physical form in "items" (which, of course, may be digital). These entities, sometimes represented as W-E-M-I, are FRBR's explication of instantiation in the library community. The object-oriented version of FRBR has been harmonized with the CRM, and therefore may be used to disambiguate mapped data such as those in this case study.

In our study, each record is a work, each version of it is an expression (so that we have the open government data expression and our CRM-mapped expression), and their occurrence in available data sets constitutes manifestation. In addition, several of the records identify objects that can be interpreted as works, and in one case textual works, which can be mapped to FRBRoo classes as a means of enriching the mapping. Tables 1–5 display parallel mapping of our five cases using FRBRoo as an extension of CRM mapping. In the tables, to the best of the capacity of the tabular format, we have attempted to draw parallels between the CRM and FRBRoo categories. For example, in several cases, E40 Legal Bodies (CRM) are the parallel mappings for bodies that also serve as F11 Corporate Bodies (FRBRoo). Similarly, where FRBRoo has inherited the CRM category, such as E35 Title, no parallel is shown.

Table 1 displays parallel mapping between FRBRoo and CIDOC-CRM for the record "Archives of Digital Advertisement." This mapping shows the alignment between FRBRoo and CIDOC-CRM, which is useful for extending the

Table 2: FRBRoo/CRM Mapping of Museum Facility

	FRBRoo	CRM	Value 1	Value 2
Name	F11 Corporate Body	E39 Actor	Museum of Education	Museum of Bow
Location	F9 Place	E44 Place Appellation	Wonmi-gu Sosa-ro 482 (Chooneui-dong)	Wonmi-gu Sosa-ro 482 (Chooneui-dong)
Telephone		E51 Contact Point	661–1282	614–2678
Admission fees		E62 String	Charged	Charged
Hours		E5 Event E55 Type	• Open: 09:00–18:00	• Open: 09:00–18:00
		E58 Measurement Unit	• Closed: Mondays, New Year, Chooseok, next day of national holidays	• Closed: Mondays, New Year, Chooseok, next day of national holidays
		E60 Number E62 String	• 4,712 collections from modern times to current such as textbook and reference books	• 467 collections • Size: 811.67 m2 • Exhibitions, theatre, demonstration, experience class, etc.
			• Size: 672.94m2 • Exhibitions: education in the traditional era → education supplies → classroom in the 60s and the 70s → students' life at school → reference room	• Various programs such as making bows during vacation

two ontologies together for digital records of archival materials. Noticeable are points such as E35 Title where FRBRoo refers to CRM classes, and the overlap between F11 Corporate Body and F12 Nomen for the two corporations that are both producers and subjects of the cartoons, and F12 Nomen and E35 Title for the characters that are also subjects of the cartoons.

Table 2 displays parallel mapping between FRBRoo and CIDOC-CRM for the record "museum facility." In this case, the FRBRoo mapping provides an intersection between the appellations for the museums and their place of location.

Table 3 displays parallel mapping between FRBRoo and CIDOC-CRM for the record "museum artifacts." Once again, we see an overlap between the FRBRoo F12 Nomen and F35 Nomen Use Statements for the objects identified in the record as CRM E22 Man-Made Objects. This points out the dual usage wherein artifacts are represented both as objects and as names of kinds of things. Also, FRBRoo allows the characteristics of the objects themselves to be represented as F3 Manifestation Product Types. In this way the mapping is enriched by the use of the two ontologies together.

Table 3: FRBRoo/CRM Mapping of Museum Artifacts

	FRBRoo	CRM	Value 1	Value 2
Artifact name	F7 Object F12 Nomen	E22 Man-Made Object	Banja	Floral design celadon
Size		E55 Type E 58 Measurement Unit E60 Number	Maximum diameter + 36.2 + Thickness + 7.4	Height 35.2 cm
Holding institution		E87 Curation Activity	National 1/Gyeongju	National 1/Central
Artifact number	F13 Identifier	E42 Identifier	Chrysanthemum 000044–000	Desheng 000343–000
Description	F35 Nomen Use Statement	E62 String	Banja is a metal Buddhist percussion instrument. It is used to call the public in the temple or to inform the urgent things, and it is still used in . . .	
Nationality 1		E74 Group	Korea	Korea
Nationality 2		E4 Period	Goryeo	Goryeo
Findspot 1–1	F9 Place	E53 Place	North Gyeongsang Province	Gyeonggi Province—near Kaesong
Material 1–1	F3 Manifestation Product Type	E57 Material	Metal	Celadon
Material 1–2	F3 Manifestation Product Type	E57 Material	Copper alloys	
Usage function 1–1	F3 Manifestation Product Type	E17 Type Assignment	Religious belief	
Usage function 1–2	F3 Manifestation Product Type	E17 Type Assignment	Buddhism	
Usage function 1–3	F3 Manifestation Product Type	E17 Type Assignment	Ceremony	
Collection 1		E55 Type	National 1	National 1
Genre 1		E62 String	Study of ancient culture and heritage of the country	

Table 4 displays parallel mapping between FRBRoo and CIDOC-CRM for the record "Library of the Korea Research Institute of Bioscience and Biotechnology (as of March 2014)." The specific objects identified in the record are texts available in the library; thus, FRBRoo is very useful for extending the mapping to the bibliographic characteristics of the texts.

Table 5 displays parallel mapping between FRBRoo and CRM for the record "library management (31 August 2015)." The record contains standard information for the administrative purposes of managing libraries in the Republic

Table 4: FRBRoo/CRM Mapping of Library of the Korea Research Institute of Bioscience and Biotechnology—Booklist (as of March 2014)

	FRBRoo	CRM	Value 1	Value 2
Number		E60 Number	1	20779
Registration number	F13 Identifier	E42 Identifier	0000001	7006762
Title				
Author	F10 Person	E21 Person	Colowick, Sidney P.	Hyeon YoonJeong
Publisher	F24 Publication Expression	E40 Legal Body	Academic Press	Korea Environment Institute
Publication year	F30 Publication Event	E50 Date	1955	2013
Category			QP601	
Books			M49 1955	
Volume				Vol. 1
Cpies		E84 Information Carrier		

of Korea. As was the case in table 2, the representation of "museum facility," there is little overlap between the two ontologies because this record contains little bibliographic information. However, F44 Bibliographic Agency supplies a potential point of connection should the libraries in this record also be represented elsewhere in the data set as producers of bibliographic data for their holdings.

Finally, figure 10 shows how we used FRBRoo to map each instantiated record. Each record in a data set is an information object (F73), which is expressed (F2) in the form of the tangible data in the data set. Researchers (E39 Actor) translated the original records from Korean to English, creating a second set of information objects, which are manifestation product types (F3) that also, at least in this case, are manifestation singletons (F4). Each has metadata concerning the actors and the expression creators such as the original, the translated, and the CRM-mapped. The end object is inherent in a set of information carriers (E84) that requires disambiguation. We have added value by extending the mapping of the instantiated records with FRBRoo.

Conclusions

We applied a metalevel ontology for cultural heritage information sharing to a set of selected Korean open government data. Our analysis of the population of data sets from which the cases were drawn demonstrates the compatibility of these cultural heritage open data sets with the event-based CIDOC-CRM ontology. Further, our analysis extends the mapping by the use of FRBRoo to provide disambiguation for the mapped data. In this way, we have demonstrated one

Table 5: FRBRoo/CRM Mapping of Library Management (August 31, 2015)

	FRBRoo	CRM	Value 1	Value 2
Library name	F44 Bibliographic Agency	E41 Appellation	BaekHab Children's Library	Book Café CheongMaRu
City/province name	F9 Place	E44 Place Appellation	Seoul city	Seoul city
City/gu/gun	F9 Place	E44 Place Appellation	Yongsan-gu district	Yongsan-gu district
Library type		E55 Type	Small-sized public library	Small-sized public library
Closed		E52 Time-Span	Sunday	Saturday, Sunday, holidays
Open		E61 Time Primitive	10:00	9:00
Closed		E61 Time Primitive	17:00	19:00
Seats		E60 Number	32	50
Resources—books		E55 Type	8863	4000
Resources—serials		E55 Type		
Checkout availability		E60 Number	8863	
Checkout length		E52 Time-Span	7 days	
Address	F9 Place	E45 Address	19–2, Huam-ro 15-gil	Noksapyeong-daero
Management agency	F11 Corporate Body	E40 Legal Body	BaekHab Church	Yong San-gu district
Telephone		E51 Contact Point	755–0525	2199–8967
Creation date (as of)		E61 Time Primitive	2015.8.31	2015.8.31

approach to the use of classical facet analytical theory with the CRM and FRBRoo ontologies by using CRM to represent event-based data and FRBRoo to add dimensionality in a manner similar to the use of auxiliary devices employed in many classification systems (Smiraglia 2017, 6). This study is original—to the best of our knowledge, FRBRoo mapping with CRM using open government data has not been discussed previously.

Further, we have borrowed methods and theoretical framing from knowledge organization. In knowledge organization, the concept is the basic element of any ontology, and it is the ordering of concepts that produces a useful knowledge organization system. CIDOC-CRM uses the concepts embedded in cultural heritage metadata, reordering them by event-based categories and properties. To this we have added the FRBRoo categories and properties that disambiguate instantiation sets. We have shown that knowledge organization domain analytical techniques can be applied to the analysis of potential data for ontological

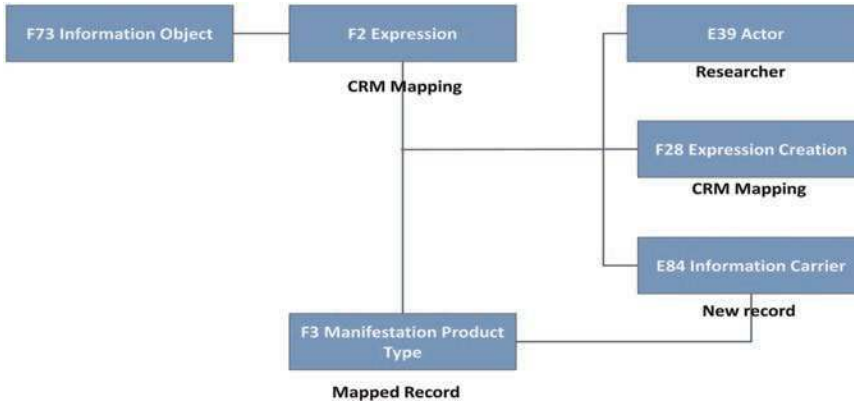

Figure 10: FRBRoo extension

mapping. We have demonstrated the utility of parallel use of the CIDOC-CRM and its FRBRoo extension as valuable ontologies for mapping digital cultural heritage data sets to facilitate information sharing. Also, this study contributes to the cross- or metainstitutional integration of curation across institutional boundaries in cultural heritage institutions such as libraries and museums as the imperative for cultural synergy and the role of information institutions (Park and Smiraglia 2014) by reusing open government data in heterogeneous formats.

Ontologies derived from open government data are culturally neutral because no biases are represented when a resource is described, meaning no biased judgments of culture. This study focuses on the internal records of the open government data, which themselves were generated from the repository. Ontological mappings from this study can be applied to both (1) the virtual catalogue with the data components and (2) the visualization of KOS, which can provide useful navigational maps.

References

Bountouri, Lina, and Manolis Gergatsoulis. 2011. "The Semantic Mapping of Archival Metadata to the CIDOC-CRM Ontology." *Journal of Archival Organization* 9 (3–4): 174–207. https://doi.org/10.1080/15332748.2011.650124.

CIDOC-CRM Special Interest Group. 2015. "Definition of the CIDOC-CRM Conceptual Reference Model." Last modified May 2015. http://www.cidoc-crm.org/Version/version-6.2. Accessed 11 May 2018.

Doerr, Martin, and Dolores Iorizzo. 2008. "The Dream of a Global Knowledge Network: A New Approach." *Journal on Computing and Cultural Heritage* 1 (1): 1–23. https://doi.org/10.1145/1367080.1367085.

Fragkou, Pavlina, Eleni Galiotou, and Michalis Matsakas. 2014. "Enriching the e-GIF ontology for an Improved Application of Linking Data Technologies to Greek Open Government Data." *Procedia: Social and Behavioral Sciences* 147: 167–74. https://doi.org/10.1016/j.sbspro.2014.07.141.

Gergatsoulis, Manolis, Lina Bountouri, Panorea Gaitanou, and Christos Papatheodorou. 2010. ''Query Transformation in a CIDOC-CRM Based Cultural Metadata Integration Environment.'' In ECDL'10: Proceedings of the 14th European Conference on Research and Advanced Technology for Digital Libraries, Glasgow, UK, September 6–10, 2010, 38–45. Berlin: Springer. https://doi.org/10.1007/978-3-642-15464-5_6.

Hodge, Gail. 2014. ''Government Knowledge Organization Systems: Valuing a Public Good.'' *Bulletin of the Association for Information Science and Technology* 40 (4): 23–29. https://doi.org/10.1002/bult.2014.1720400411.

Janssen, Marijin, Yannis Charalabidis, and Anneke Zuiderwijk. 2012. ''Benefits, Adoption Barriers and Myths of Open Data and Open Government.'' *Information Systems Management* 29 (4): 258–68. https://doi.org/10.1080/10580530.2012.716740.

Koutsomitropoulos, Dimitrios A., Georgia D. Solomou, and Theodore S.P. Papatheodorou. 2009. ''Metadata and Semantics in Digital Object Collections: A Case-Sudy on CIDOC-CRM and Dublin Core and a Prototype Implementation.'' *Journal of Digital Information* 10 (6). https://journals.tdl.org/jodi/index.php/jodi/article/view/693/577. Accessed 11 May 2018.

Lin, Chia-Hung, Jen-Shin Hong, and Martin Doerr. 2008. ''Issues in an Inference Platform for Generating Deductive Knowledge: A Case Study in Cultural Heritage Digital Libraries using the CIDOC-CRM.'' *International Journal on Digital Libraries* 8 (2): 115–32. https://doi.org/10.1007/s00799-008-0034-0.

Park, Hyoungjoo. and Richard P Smiraglia. 2014. ''Enhancing Data Curation of Cultural Heritage for Information Sharing: A Case Study Using Open Government Data.'' In Proceedings, *Metadata and Semantics Research: 8th Research Conference, MTSR 2014, Karlsruhe, Germany, November 27–29, 2014,* ed. Sissi Closs, Rudi Studer, Emmanouel Garoufallou, and Miguel-Angel Sicilia. In Communications in Computer and Information Science 478: 95–106. https://doi.org/10.1007/978-3-319-13674-5_10.

Smiraglia, Richard P. 2015a. *Domain Analysis for Knowledge Organization: Tools for Ontology Extraction. Chandos Information Professional Series*. Oxford: Elsevier/Chandos.

——. 2015b. ''The Roles of Ontology in Knowledge Organization.'' In *Ontology for Knowledge Organization*, ed. Richard P. Smiraglia and Hur-li Lee, 1–2. Würzburg: Ergon Verlag.

——. 2017. ''A Brief Introduction to Facets in Knowledge Organization.'' In *Dimensions of Knowledge: Facets for Knowledge Organization*, ed. Richard P. Smiraglia and Hur-li Lee, 1–6. Würzburg: Ergon Verlag.

Stasinopoulou, Thomais, Lina Bountouri, Constantia Kakali, Irene Lourdi, Christos Papatheodorou, Martin Doerr, and Manolis Gergatsoulis. 2007. ''Ontology-Based Metadata Integration in the Cultural Heritage Domain.'' In *Asian Digital Libraries. Looking Back 10 Years and Forging New Frontiers*. ICADL 2007, ed. D.H.L. Goh, T.H. Cao, I.T. Sølvberg, and E. Rasmussen, 165–75. Lecture Notes in Computer Science, vol. 4822. Berlin: Springer. https://doi.org/10.1007/978-3-540-77094-7_25.

Theodoridou, Maria, Yannis Tzitzikas, Martin Doerr, Yannis Marketakis, and Valantis Melessanakis. 2010. ''Modeling and Querying Provenance by Extending CIDOC-CRM.'' *Distributed and Parallel Databases* 27 (2): 169–210. https://doi.org/10.1007/s10619-009-7059-2.