Semantic Enrichment of Linked Archival Materials†

Shu-Jiun Chen

Academia Sinica, Institute of History and Philology, Taipei 11529, Taiwan,
<sophy@sinica.edu.tw>

Shu-Jiun Chen is an assistant research fellow at the Institute of History and Philology, Academia Sinica, and also the Executive Secretary of the Academia Sinica Center for Digital Cultures (ASCDC). She received her PhD degree in library and information science from National Taiwan University in 2012. Her research interests include digital libraries, metadata, knowledge organization, linked data and digital humanities. Dr. Chen initiated the Chinese-language Art & Architecture Thesaurus (AAT) research project with the Getty Research Institute in 2008, and she is also the principal investigator of the Linked Open Data Lab in ASCDC.

Chen, Shu-Jiun. 2019. "Semantic Enrichment of Linked Archival Materials." *Knowledge Organization* 46(7): 530-547. 35 references. DOI:10.5771/0943-7444-2019-7-530.

Abstract: By using the metadata for the fonds of "Chen Cheng-po's Paintings and Documents" (CCP) in the database of the Archives of the Institute of Taiwan History (IHT, Academia Sinica, Taiwan), we develop and enhance a semantic data model for converting the data into a linked data project, focusing on data modeling, data reconciliation, and data enrichment. The research questions are: 1) How can we keep the original rich and contextual information of the archival materials during a LOD task?; 2) How can we integrate heterogeneous datasets about the same real-world resources from libraries, archives, and museums, while keeping the different views distinct?; and, (3) How can we provide added value for semantic metadata of archives in terms of instance-based and schema-based types of enrichment? The project adopts the Europeana Data Model (EDM) as the main model and extends the properties to fit the contextual characteristics of archival materials. Various methods are explored to preserve the hierarchical structure and context of the archival materials, to enrich semantic data, and to connect data from different sources and institutions. We propose four approaches to enriching data semantics by: 1) directly using external vocabularies; 2) reconciling local links to other linked data sources; 3) introducing contextual classes for the appropriate contextual entities; and, 4) utilizing named entity extraction. The results can contribute to the best practice for developing linked data for art-related archival materials.

Received: 22 April 2019; Revised: 30 June 2019; Accepted: 29 August 2019

Keywords: linked open data, semantic data, data modeling, archival resource

† The author would like to acknowledge the assistance of Lu-Yen Lu. The work was supported by a grant for AS-ASCDC-108-001 from the Academia Sinica Center for Digital Cultures.

1.0 Introduction

Semantic enrichment in the context of linked open data (LOD) provides both properties and their values in a linked dataset by assigning meaning to it. This makes it easily discoverable, relating it to other datasets, and even giving it different perspectives of data. In the era of the semantic web, it is an excellent method to convert those original archival metadata into LOD to enhance their reusability and data accessibility. However, the basic characteristics of archival data should also be taken into account before developing a suitable data model. First, archival data has its own special hierarchical structure (such as one with four levels: fonds, series, file, and item) based on the principle of provenance (Bearman and Lytle 1985; Guimarães and Tognoli 2015). Second, an archives is not only a collection of documents but also an intensive contextuality between its components that might exist within an archives (Zhang 2012).

This study was designed to explore various methods of semantic enrichment to enhance the accessibility of archival materials on the World Wide Web, asking the following questions: 1) How can we keep the original rich and contextual information of the archival materials during a LOD task?; 2) How can we integrate heterogeneous datasets about the same real-world resources from libraries, archives, and museums, while keeping the different views distinct?; and, 3) How can we provide added value for semantic metadata of archives in terms of instance-based and schema-based types of enrichment?

2.0 The case study of the fonds of "Chen Cheng-po's Paintings and Documents"

By using the metadata for the fonds of "Chen Cheng-po's Paintings and Documents" (CCP) in the database of the Archives of the Institute of Taiwan History (IHT, Academia Sinica, Taiwan), we develop and enhance a semantic data model for converting the data into a LOD-based dataset and preserving its original relations in the meantime. The CCP archive is one of the 136 archival fonds in the Archives of the IHT. Chen Cheng-po (1895–1947) was an important Taiwanese artist. Many materials in this archives, including those about his artworks, reveal a strong influence of both Japanese and post-impressionist art in a cross-cultural encounter as well as the intensive social network between him and his fellows (Lin 2012; Hsiao 2014; Mathison 2012).

The original metadata standard of the CCP Archives is based on the Encoded Archival Description (EAD), which can encode finding aids compatible with General International Standard Archival Description (ISAD(G)) (Higgins and Inglis 2003; International Council on Archives 2000). To enable interoperability between CCP metadata and other metadata such as those for other artists' archival fonds housed in the Institute and those from other fine art museums and their archives, the Europeana Data Model (EDM) has been adopted as the core data model (Clayphan et all 2017). For the present paper we developed the LOD Lifecycle Framework (Table 1), in a total of twelve modules, for converting legacy systems to the linked data format. Among the modules, four of them (i.e., #4 data modeling, #6 data reconciliation, #7 data enrichment, and # 9 data access) contribute to semantic enrichment. In this study we focus on these four modules.

3.0 The ontology design: extensions of the Europeana Data Model (EDM)

This ontology defines the information of an archival resource itself and its relationships to the whole archival hierarchy. It is designed to structure a semantic standard to execute the data conversion of archival metadata into a LOD-based dataset. The ontology is based on the Europeana Data Model (EDM) and is extended for the study in terms of more specific properties to fit the archival context. Since the original content in the archival database is structured in a four-layered hierarchy as "fonds," "series," "file," and "item," an extension of the EDM data model that matches the original essence of that archival hierarchy is needed. As a result, the metadata of each archival resource will be successfully converted into a LOD-based dataset and will keep its original relational position in the whole hierarchy at the same time.

According to the revised EDM data model, each metadata record of an archival resource consists of the three EDM core classes: "Aggregation," "Provided CHO"

Module	Remarks
Requirements Identification	Identify the purposes of metadata conversion into LOD and requirements for future application
Data Licenses Specification	Obtain the license permission for data re-consumption
Source Data Analysis	Analyze the data structure, data content, and data value of the original metadata.
Data Modeling	Design the ontology/semantic data model. Review the currently reusable ontologies and vocabularies, which can be applied as a basis to reuse the classes and properties.
Data Cleaning	Correct the data errors for optimization of the data quality.
Data Reconciliation	Design and create own URIs for the dataset and link to external controlled vocabularies on the web, which are shared entities but not shared URIs.
Data Enrichment	Augment the source metadata with additional terms.
Data Conversion	Transform the legacy dataset into a linked data representation, which is in RDF file format, and upload into the triple store.
Data Access	Establish the SPARQL endpoint and create templates for semantic queries using examples, so that both machines and humans can access the data.
Data Publishing	Publish the linked data set in the worldwide open data platform for data sharing and distribution (e.g., datahob.io).
Data Applications	Develop the application system for providing the linked data-based service system (such as a website).
Data Maintenance	Develop a mechanism for updating linked data sets.

Table 1. The LOD lifecycle framework (LODLiFrame) version 2.3 (Chen 20xx).

and "Web Resource." The essence of the archival resource belongs to the broad family of the Cultural Heritage Objects (edm:ProvidedCHO), their digital representations (edm:WebResource) and an aggregation to represent the result of the provider's activity (ore:Aggregation) (Isaac, 2013). There are 1,522 records in the database and 44,647 triples generated and released in the linked data project. Considering the four-layered hierarchical structure of the original archive metadata, each archival level is treated as a unique core model in the process of data modeling. The ontology consists of nine classes, forty-six properties, and eight external resources as follows in Table 2.

Table 2 shows the basic data model of an archival resource at any level of archival hierarchy and is composed of four parts (see Figure 1):

– Four core models (fonds, series, file, item) – Nine classes reused from three vocabularies (EDM, ORE, SKOS) to structure the core and contextual classes after EDM – Forty-six properties applied from eleven semantic vocabularies (ASCDC, BIBO, CIDOC-CRM, DC, DCTERMS, EDM, LOCAH, ORE, OWL, RDF, RDFS) to describe data content of an archival resource – Eight external resources reused in entities for data enrichment. These include *AAT* (*Art & Architecture Thesaurus*), ASCDC (Academia Sinica Center for Digital Cultures Terms), DBpedia (DBpedia terms), LOCAH (Linked Open Copac and Archives Hub), *TGN* (*Thesaurus of Geographic Names*), *ULAN* (*Union List of Artist Names*), VIAF (Virtual International Authority File).
Properties in ProvidedCHO
ascdc:arrangement, ascdc:chronologicalDescription, ascdc:prefCiteWay, ascdc:type, bibo:distributor, crm:P62_depicts, dc:contributor, dc:creator, dc:date, dc:identifier, dc:rights, dc:subject, dc:title, dc:type, dcterms:abstract, dcterms:alternative, dcterms:created, dcterms:extent, dcterms:hasPart, dcterms:isPartOf, dcterms:medium, dcterms:modified, dcterms:provenance, dcterms:spatial, edm:begin, edm:end, edm:isNextInSequence, edm:isRelatedTo, edm:rights, locah:accessRestrictions, locah:appraisal, locah:bibliography, locah:hasBiographicalHistory, locah:item, locah:level, locah:origination, locah:scopeContent, ore:proxyFor, ore:proxyIn, owl:sameAs, rdf:seeAlso, rdf:type, rdfs:label
Properties in the Aggregation of ProvidedCHO with WebResource
edm:aggregatedCHO, edm:dataProvider, edm:hasView, edm:isShownBy, edm:rights, rdf:type, rdfs:label

Table 2. The ontology of art archives.

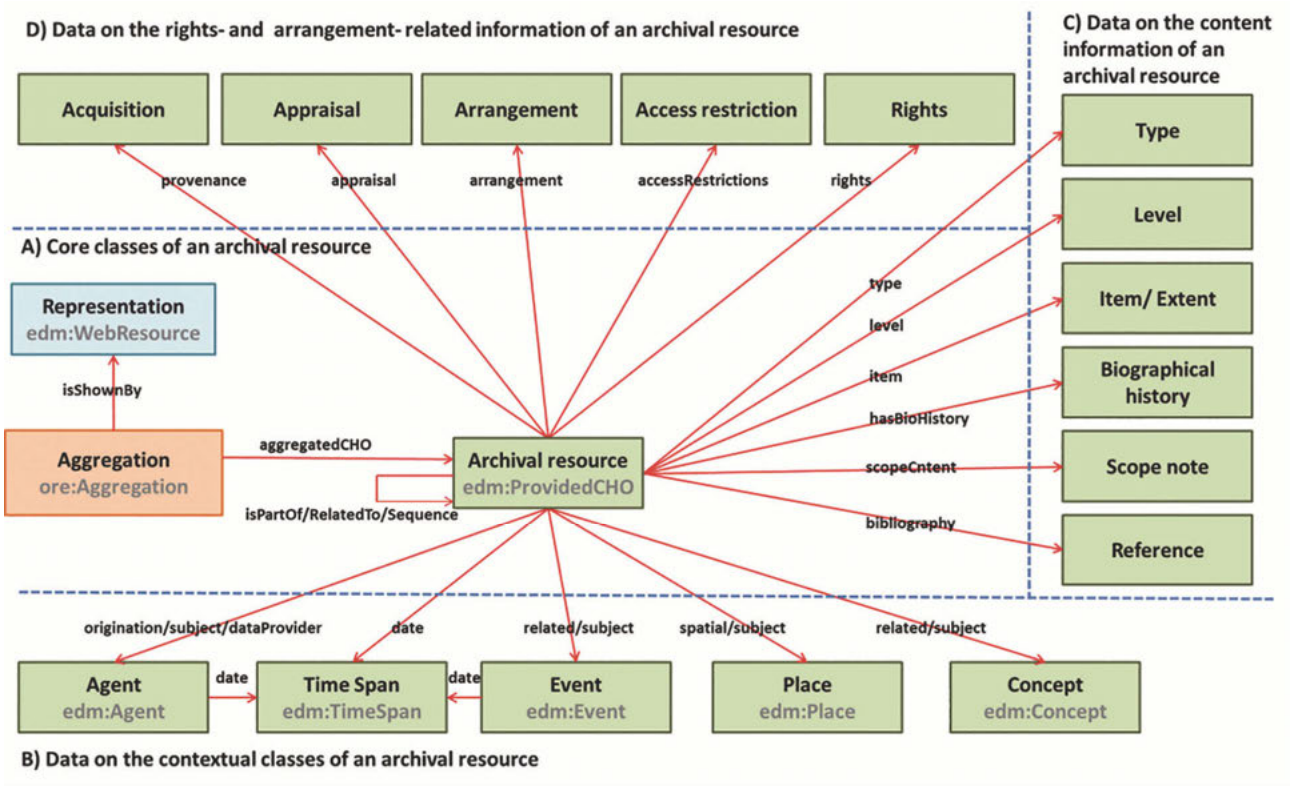

Figure 1. Basic data model for the CCP archives.

a) "Core classes" are the basic components of a metadata record, which describe the real object of an archival resource with its online representation and aggregation.
b) "Contextual classes" are related to an archival resource, including the Agent, Event, Place, Timespan, and Concept. The information can enrich the original data content of an archival resource by linking to external resources.
c) "Data content of an archival resource" is the real information carried by an archival resource (ProvidedCHO in EDM), which describes the level, type item, biographical history scope note, reference, and other information relating to the essence of an archival resource.
d) "Data on rights- and provenance-related information" describes the current rights and status of an archival resource, which mainly contain the statement on the acquisition, appraisal, access restriction, arrangement, and rights of an archival resource. Together with part C, both become the major components of the content of an archival resource.

4.0 Preserving the hierarchical structure and context of an archivse in the design of the data model

In this section we report on the results that preserve various context of the archival materials when designing the data model, including preserving the hierarchical structure and context of archival materials, and keeping the contextuality between archival components.

4.1 Preserving the hierarchical structure of archival materials

Archivists regularly retain the evidential value by providing information about the origins, functions, and activities of creators when arranging and describing archival materials (Pearce-Moses 2005: 152-153). The original database of the Archives of the Institute of Taiwan History follows the finding aids standards of Encoded Archival Description (EAD), capturing the provenance, original order, and contexts of the records, which are crucial to archival users in understanding the significance of the material (Francisco-Revilla, Trace, Li, and Buchanan 2014). To capture the initial meaning of each data field, we extend the EMD by reusing the vocabularies such as Linked Open Copac and Archives Hub (LOCAH), the Bibliographic Ontology, Dublin Core, Dublin Core Metadata Terms, the OWL 2 Schema vocabulary, the RDF Concepts Vocabulary, and RDF Schema vocabulary, which means the hierarchical archival data can be described in more specific meanings that are not in the EDM at present. To keep the original four-layered archival hierarchy with fonds, series, file, and item, we develop a four-level core model. Different hierarchical levels in the archive are made explicit by using the level property (Locah:level), and the hierarchical relations between different levels are assigned by the has-part property (dcterms:hasPart) in the data model (Figure 2).

4.2 Preserving the context of the archival materials

The core model in this study consists of fonds, series, file, and item levels. Each level represents specific contextual

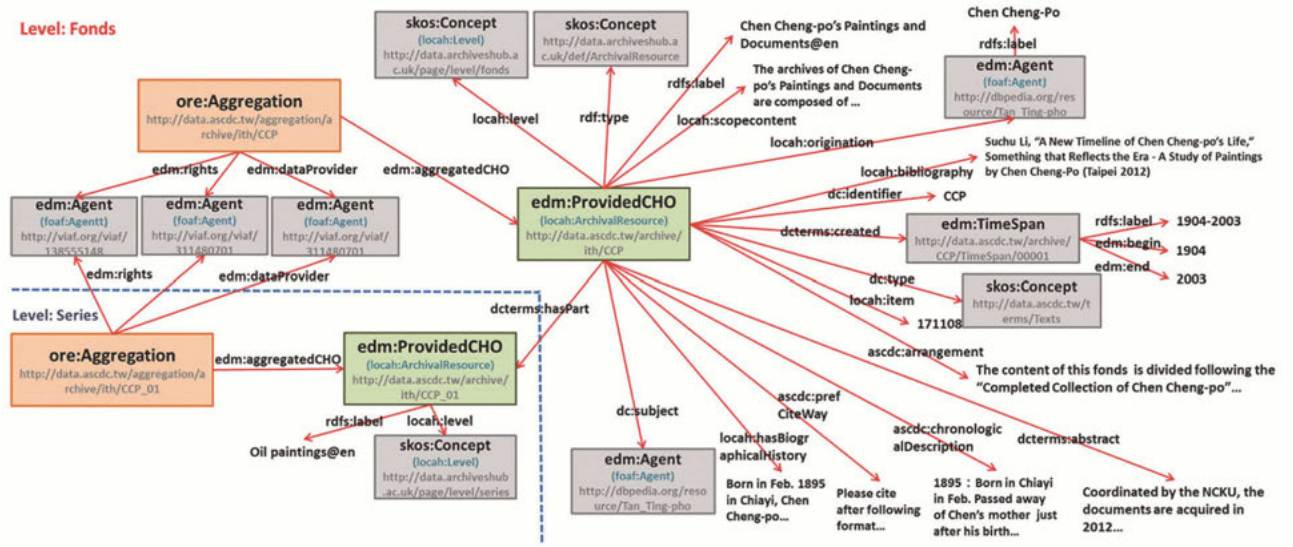

Figure 2. Core model of the fonds level. Example: an archival resource at the fonds level of the CCP Archive.

information of an archive. For instance, the structural view of the fonds level includes its external and internal dimensions, such as origination, time period, biographical history, chronology list, scope content, arrangement, and related materials (Table 3). An archival resource (edm:ProvidedCHO) carries the descriptive information on the fonds and links to external URL resources or the URI resource by contextual class in order to enrich the data content (Figure 2). Each archival item is linked to an aggregation (ore:Aggregation), which is a joint node in the semantic web after the data model.

4.3 Keeping the contextuality between archival components

In archives, the components (e.g., series, files, or items) within a fonds' structure might be related to each other. To demonstrate the contextuality existent in those parts, the "is-related-to" property (edm:isRelatedTo) is used to show relations between these archival components of different archives or different series under the same fonds, while the "is-next-in-sequence-to" property (edm:IsNextInSequence) is applied to present a serial relationship between components in the same series, files, or items, in order to link to another object which logically precedes it (Figure 3).

Compared to the vertical relationship between a whole ProvideCHO and its parts mentioned earlier, the horizontal relationships relate a part in a sequence with the part immediately preceding (Charles and Olensky 2014). Table 4 shows the two types of relationships, their properties, and use cases in this study.

5.0 Enriching data

There are four approaches in this study to enriching data by: 1) direct reuse of external vocabularies; 2) reconciliation; 3) introducing contextual classes; and, 4) utilizing named entity extraction.

5.1 Data enrichment by direct reuse of external vocabularies

Linking to external vocabularies and data sets is an important method to enhance and enrich archival description. Direct reuse of existent external vocabularies is highly recommended by many researchers and best practices for linked data, although time pressure and other factors, for instance, representing the domain knowledge for a local application, make simply create new identifiers for entities the more expedient route (Hyland and Villazón Terrazas 2011; Schaible, Gottron, and Scherp 2014; Niu 2016; Sanderson 2014). In the present study, the external vocabularies, which are immediately applied and linked to the class of cultural heritage object (edm:ProvidedCHO) or an aggregation of an archival resource without establishing a local-URI identification, are reused in the properties of descriptive information such as "level," "type," "medium," "data provider," and "rights holder" (Table 5).

Original EAD-based Element	**Property***	**Data type**** (Reused Vocabularies)
Level	locah:level	URI (LOCAH)
Title	dc:title/ rdfs:label	String
Originations	dc:creator	URI
Number	locah:item	Integer
Biographical history,	locah:hasBiographicalHistory	String
Scope content	locah:scopecontent	String
Arrangement	ascdc:arrangement	String
Child	dcterms:hasPart	URI (ASCDC)
Acquisition way	dcterms:provenance	String
Appraisal	Locah:appraisal	String
Access restriction	locah:accessRestrictions	String
Copyright	dc:rights	URI (VIAF)
Citation way	ascdc:prefCiteWay	String
Subject/ Person	dc:subject	URI
Related material	edm:isRelatedTo	URI (ASCDC)

Table 3. Mapping of selected original EAD-based elements and triple-based properties.

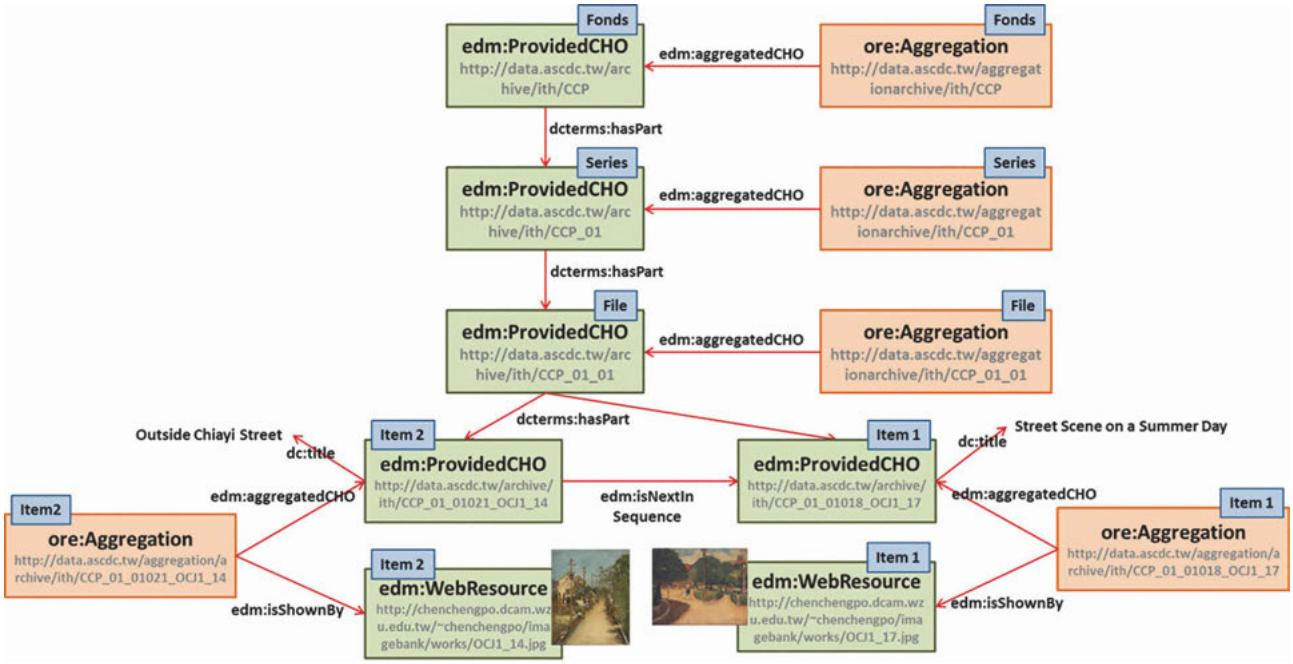

Figure 3. Hierarchical data structure for the metadata of an archival resource based on EDM model.

Types of Relationship	(Ordered)Vertical Relationship		(Ordered) Horizontal Relationship
	Top-down	Bottom-up	
Property	has-part relation (dcterms:hasPart property)	is-part-of relation (dcterms:isPartOf property)	is-next-insequence-to property (edm:isNextInSequence)
Use Case	from fonds to series level of the art archival materials	from series to fonds level of the art archival materials	items under the same series of art archival materials

Table 4. The vertical and horizontal relationships.

Relatively speaking, these have less meaning for further study or queries in an archival resource than the "creator" or "related person" of an archival resource.

For example (Figure 4), in the archival item from the CCP archives "the street scene on a summer day," property values in the "level," "type," "medium," "data provider," and "rights holder" properties are mapped to the external vocabularies. To the property of "level" (locah:level) the vocabulary as "fonds" in LOCAH is applied; to "type" (dc:type, ascdc:type) vocabularies as "JEPG" in AAT and "texts" in ASCDC are reused; to "medium" (dcterms:medium) the vocabulary as "canvas" in AAT; to "data provider" (edm:dataProvider) and "rights holder" (edm:rights) vocabulary as "Chen Cheng-po Cultural Foundation" and "Institute of Taiwan History, Academia Sinica" in VIAF are applied.

In addition, the information on place/location in the "content location" property (dcterms:spatial) of Chen's works is directly linked to the *TGN* vocabulary without endowing a locally-established URI. The reason for not giving a locally-based URI as an instance of Contextual Class for "place" lies in the fact that such geographical information is lacking in the original metadata and later extracted to enrich the original data content for data querying. Other reasons for such direct application of external vocabulary is that the *TGN* is a well-structured thesaurus with the entire information on a geographic name. It, therefore, eliminates the need to establish a place name locally, which already exists in the *TGN*.

5.2 Data enrichment by reconciliation

The purpose of data reconciliation is to make links to other URIs, so they can discover more things, which is one of the linked data principles proposed by Berners-Lee (2006). It means that one can link their data to data from other sources, from things to things, to provide context, access more information, improve research, and create po-

	Property	Type reused vocabulary	Example	Core class in the CCP archive
1	Level	LOCAH	http://data.archiveshub.ac.uk/id/level/fonds (Fonds)	ProvidedCHO
2	Type	AAT	http://vocab.getty.edu/aat/300266224 (JEPG)	ProvidedCHO
3	Type (ASCDC)	ASCDC	http://data.ascdc.tw/terms/Texts (ASCDC)	ProvidedCHO
4	Medium	AAT	http://vocab.getty.edu/aat/300014078 (Canvas)	ProvidedCHO
5	Data provider	VIAF	http://viaf.org/viaf/311480701 (Chen Cheng-po Cultural Foundation)	Aggregation
6	Rights holder	VIAF	http://viaf.org/viaf/138555148 (ITH, Academia Sinica)	Aggregation
7	Content location	TGN	http://vocab.getty.edu/tgn/7468019 (Chiayi)	ProvidedCHO

Table 5. Direct reuse of external vocabularies.

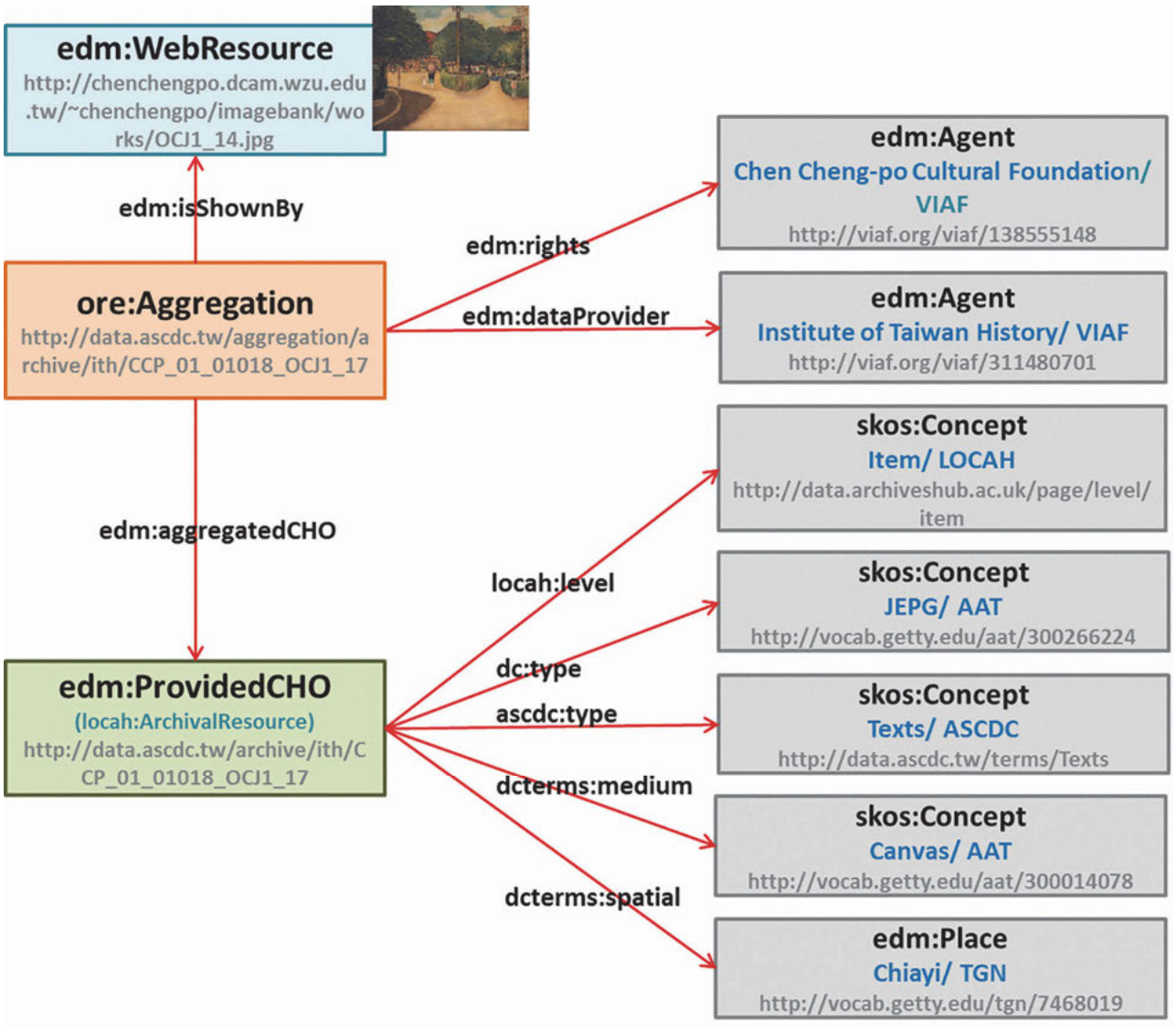

Figure 4. Direct reuse of external vocabularies in the metadata of a CCP archival resource.

tential for new research questions (Sanderson 2016). In terms of people, for example (Table 6), we first use different types of properties, including the "origination" (locah:origination), "depict" (crm:p62_depicts) and "subject" (dc:subject) to connect the CCP's cultural heritage object (edm:ProvidedCHO) to the local URIs, and second enrich information by reconciling the local URIs to external controlled vocabularies on the web by the "same as" property (owl:sameAS) where the local and external URIs are shared entities but not shared URIs.

There are three types of data reconciliation related to people. For example, the second type in Table 6, the metadata of an archival resource (CHO) describes a person, which is related to this archival resource. The semantic relation between them is made by applying the subject property (dc:subject) which leads the resource to link with a locally-based URI representing the entity of the related person. This locally-based URI is further linked to the URI of other external resources (i.e., LTA, ULAN) by the same as property (owl:sameAs). In the third type of the Table 6, the metadata of an archival resource (CHO) defines the person, which is depicted or described in the work content of this archival resource. The semantic relation between them is made by using the depict property (crm:P62_depicts), which creates a linkage of the metadata with a locally-based URI representing the entity of the depicted person in the resource. This locally-based URI is further linked to the URI of the external resource (i.e., LTA, ULAN) by the same as property (owl:sameAs).

For this study we adopt the six criteria developed by Isaac, Manguinhas, Charles and Stiller (2015) to evaluate and select external controlled vocabularies for semantic enrichment, which include availability, access, granularity and coverage, quality, connectivity, and size. The chosen external datasets include the Linked Taiwan Artists (LTA, for names of Asian artists and organizations in Taiwan, China, and Japan, who had contacted with Chen Cheng-po) and Getty's *Union List of Artist Names* (*ULAN*, for names of western artists associated with Chen's collections) (Baca and Gill 2015). Linked Taiwan Artists (LTA) (http://linkedart.ascdc.tw/) is a LOD-based dataset and also an application system developed by the Academia Sinica Center for Digital Cultures (ASCDC) of Taiwan in 2014-15. LTA provides biographical resources on Taiwanese artists in the Japanese colonial period (1895-1945), which include the artist's works, academic training, professional career, participation in associations, and related exhibitions. The objective is to trace the shifts in the development of the fine arts in Taiwan. The dataset contains more than 1,800 records and 17,296 triples, and the semantic data modeling's design is based on CIDOC-CRM (Conceptual Reference Model: http://www.cidoc-crm.org/) (Chen 2017). In this study, LTA is considered an appropriate and unique external vocabulary, which makes the linked data dataset representative of the most important painters in Taiwan modern art history.

In terms of reconciling local vocabularies to the *ULAN* vocabulary in the study, the "same as" property (owl:sameAs) is introduced to assert that the two URIs share the same entity. In total, 638 names are registered in the name list of CCP's archival resources. Ninety of them are names of western artists (Table 7), which mostly appear under the file of "image scrapbook and collection of books," and in the item of "scrapbooking images of western painters' work." More than half (forty-six individuals) are French painters of the nineteenth and early twentieth century, such as Camille Pissarro, Édouard Manet, Paul Cézanne and Paul Gauguin, which also reveals Chen's special attention on the impressionist works and painting style in that period.

5.3 Data enrichment by introducing contextual classes

We adopt the concept of EDM's "contextual classes" to enrich the original data content (Issac 2013). Data within each contextual class can be further extended by applying different ontological models and controlled vocabularies

	Entity/ Subject (CHO)	**Property <element in EAD>**	**Entity/ Object (Subject)**	**Property**	**Entity/ Object**
1	edm:ProvidedCHO Local URI	locah:origination <origination>	edm:Agent Local URI	owl:sameAs	LTA, ULAN Resource URI
2		dc:subject <controlaccess> <perName>			
3		crm:P62_depicts <controlaccess> <perName>			

Table 6. The typology of data related to people.

Nationality	Account of the Western Artist	Selected Example
French	46	Cézanne, Paul, 1839-1906 [http://vocab.getty.edu/ulan/500004793]
Italian	10	Cimabue, 1240-1302 [http://vocab.getty.edu/ulan/500016284]
Belgian	9	Rubens, Peter Paul, 1577-1640 [http://vocab.getty.edu/ulan/500002921]
German	7	Dürer, Albrecht, 1471-1528 [http://vocab.getty.edu/ulan/500115493]
Dutch	6	Gogh, Vincent van, 1853-1890 [http://vocab.getty.edu/ulan/500115588]
American	4	Sargent, John Singer, 1856-1925 [http://vocab.getty.edu/ulan/500023972]
English	4	Langley, Walter, 1852-1922 [http://vocab.getty.edu/ulan/500007053]
Spanish	3	Rembrandt van Rijn, 1606-1669 [http://vocab.getty.edu/ulan/500011051]
Russian	1	Korovin, Konstantin, 1861-1939 [http://vocab.getty.edu/ulan/500026648]
Total	90	

Table 7. The distribution of nationalities.

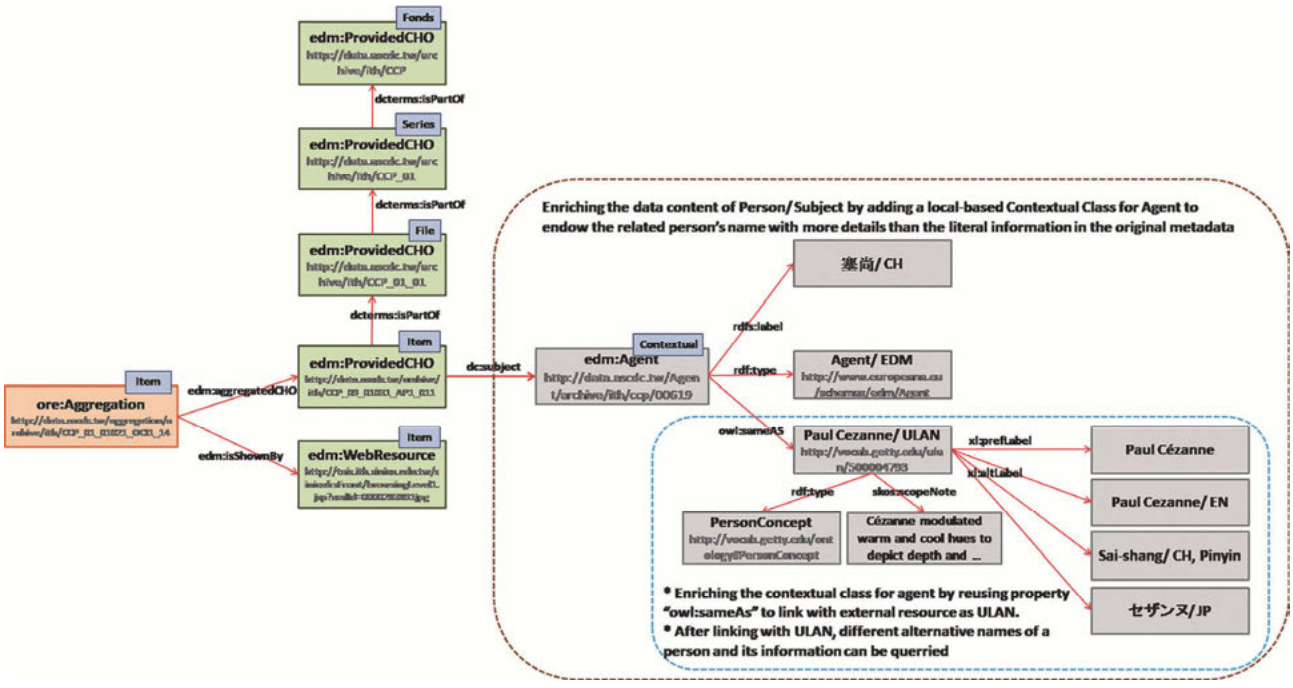

Figure 5. Data enrichment by adding a locally-based entity as a contextual class for agent.

(Tudhope and Binding 2016). For instance, a contextual class for "Time" (edm:Timespan) was enriched with the "time" ontology and a contextual class for "Event" (edm:Event) was enriched with the "event" ontology. The main purpose for designing the "Contextual Class for Agent" (Figure 5) in the study is to enrich the original data content, or the entity, of an archival resource. By converting the string information, which is related to an archival resource or object, in the original metadata into an instance of entity in the semantic web, it could carry more specific

information about an agent, no matter a person or an organization, than the literal words.

There are two methods to construct a contextual class for "agent" in the semantic data model: 1) reusing the existent external vocabularies specialized in personal and corporate names (such as Getty's *ULAN* or VIAF) to enrich the original data content; and, 2) adding a locally-based entity in the contextual class to carry different and detailed information on a related agent. To the first method, a data provider does not have to establish an extra entity for an agent by making a linked data dataset. However, when the related agent of an archival resource does not exist in any external vocabularies, a recommended approach is to add a locally-based entity to describe this related agent. For example, in the CCP archival resource "Portrait of Grandmother" (祖母像) (Figure 6), the pictorial content of this oil painting is the painter Chen's grandmother, Lin Bao-zhu (林寶珠). Since the original metadata do not have any detailed biographical information on Lin, it is reasonable to add a contextual class for agent to the metadata of the archival resource (Provided CHO) of this work by using the "depicts" property (crm:P62_depicts).

After adding the contextual class for "Lin Bao-zhu," detailed information on Lin's life, such as her birth year/place, death year/place, class type, related resource, and description, can be established by the data provider, endowed within the class to enrich the original data content. The external resource on Lin, such as it is from the LOD dataset "Linked Taiwan Artists (LTA)," is also linked to this contextual class by the "same as" property (owl:sameAs).

In actuality, it is often possible that an archival resource could have more than one contextual class for agent by converting the original metadata into LOD-based datasets. For example, for information on the creator and the depicted person of an archival resource, they can both be enriched by giving a contextual class for each of them. These can be linked to the archival resource by using the specific properties, such as "dc:creator" for creator and "crm:P62_depicts" for the depicted person. In this study, it is possible to integrate data provided by different organizations and to enhance the interoperability of the different data properties, which refer to the same meaning of the class but apply different property vocabulary (such as "dc:creator" or "schema:creator"), and also to make a more convenient modus for data querying and management. For this purpose, it is reasonable to create a mechanism for structuring a hierarchy of data properties to the contextual class for agent. Since the properties as "dc:creator," "crm:P62_depicts," "dc:contributor," or others are all defined to link the archival resource with the agent-based contextual classes, reusing the property "edm:hasMet" as a super property for all agent-describing related properties make the data modeling more reasonable in its hierarchical structure. For example, in the metadata for Chen Cheng-po's archival resource "Portrait of Grandmother," two different Contextual Classes for Agent are applied. One of them is to describe the creator of the painting by using the property "dc:creator," while the other describes the depicted person in the painting by using "crm:P62_depicts" (Figure 7).

By using property "edm:hasMet" to cluster those to different properties, which are reused to describe an agent related to an archival resource under the same hierarchy, it is not only beneficial to use data querying for finding out different data values of related type (such as creator, depicted person, or contributors are all related to agent) and helpful for data management, but also useful for maintaining the original meaning of different data fields.

5.4 Data enrichment by named entity extraction

There is rich and useful information in many fields of the source metadata records, which contain unstructured text. To enrich the original data content and enhance the reusability and extensibility of the LOD-based dataset after data

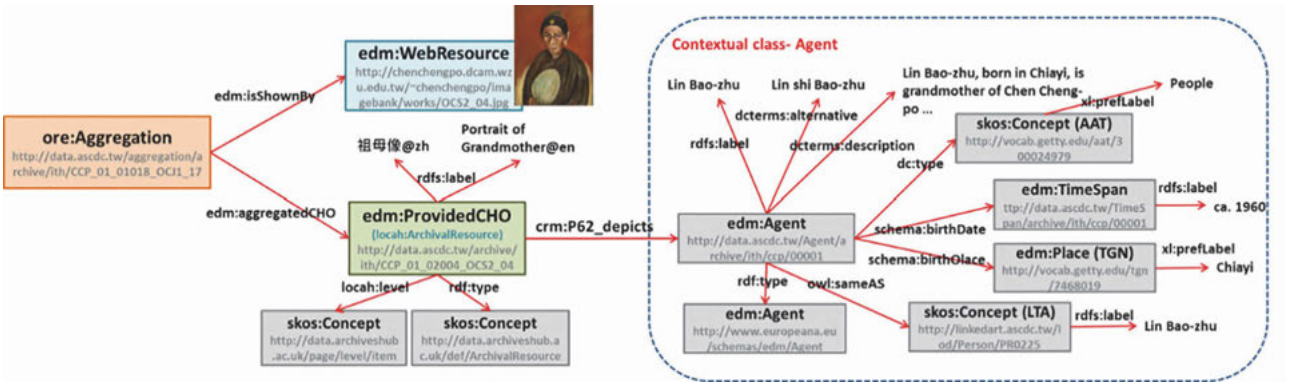

Figure 6. Data enrichment by adding a locally-based entity as a contextual class for agent.

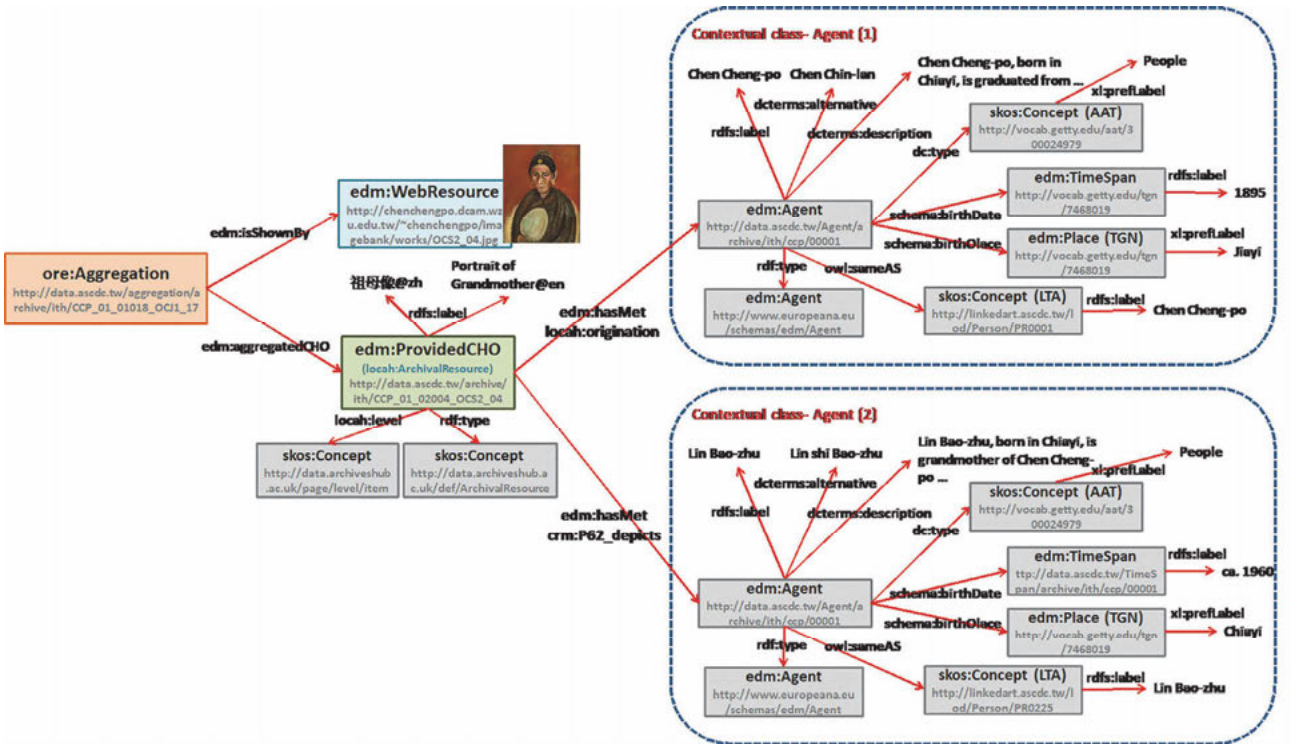

Figure 7. Property "edm:hasMet" as a mechanism to cluster different properties for describing the contextual class for agent.

Figure 8. Extracted place names "related places" from the original data fields "title" and "scope note" (A selected view).

conversion, we use the named entity extraction to pick up different types of names, such as places, persons and organizations, from the metadata records. For example, the study extracts place names in the original metadata records, which refer to the places depicted in Chen Cheng-po's paintings, in the data field "title" or "scope note" and add the extracted names into a new property, named as "related place" (Figure 8).

Place names are shown in literal forms and hidden in titles or descriptive sentences in the metadata. To ensure data accuracy, the extracted place names from the metadata will also be rechecked and referenced by confirming with the related bibliographic books and "Starting Out from 23.5°N: Chen Cheng-po," a website on the digital collection and curation of Chen's works and collections initiated by the Academia Sinica Center for Digital Cultures (ASCDC). After parsing out place names in the metadata, those names will be mapped to the external resource Getty's *TGN* and enriched in the converted LOD dataset. In total, ninety-three records of twenty-one place names in Taiwan, Japan, and China were parsed out, which also reflects the places the painter travelled to over his entire career.

5.5 Connect data from different sources and institutions

In the practice of working on the LOD, a dataset might be enriched or integrated with data from different sources, like an organic structure. However, how to integrate data created by different institutions and preserve the original content's meaning from different creators under the same model structure is a central issue in the design of the semantic data model, especially when the data of different sources are actually referring to the same object. This study includes the "proxy" mechanism in the model design to satisfy such possible need in the process of data integration, allowing aggregation of multiple records for the same item. The "proxy" mechanism refers to reuse the Aggregation class and Proxy class of the Object Reuse and Exchange ontology (ORE), and their related properties (i.e. proxy-for, proxy-in) for representing aggregations of digital objects from different sources and institutions. Each proxy represents a set of the original data from each different institution, which refers to the same single object (edm:ProvidedCHO) by using the proxy-for property (ore:proxyFor) and can be further integrated into the aggregation of an object (ore:Aggregation) in application of "ore:proxyIn." With this special mechanism, the cross-domain or cross-institute data (such as data from libraries, archives, and museums) can be harmonized in the same data structure, while the originality of each data can be separately maintained in its own proxy (see Figure 9).

For example, Chen Cheng-po's oil painting "Street scene on a summer day" created in 1927 is currently part of the collection in the Taipei Fine Arts Museum (TFAM). However, the data information on the same work is found in the "Taiwan Archival System," the online database of "Starting Out from 23.5°N: Chen Cheng-po, 23.5°N," and the LOD dataset "Linked Taiwan Artists" in addition to the metadata database of the TFAM. To integrate the data referring to the same work from the abovementioned different databases, we can add four "Proxy-classes" to carry information on the same work for those four sources one by one, following the model. Since those data mention the same work, all those "Proxies" can represent the same cultural heritage object (Provided CHO) by applying the "proxy for" property (ore:proxyFor) and link to its own Aggregation class and online representation (Web Resource class). By using the "Proxy-classes," data provided from diverse sources will not be merged, making data querying of different content easier under an integrated framing of data resources that come from different databases.

6.0 SPARQL template-based question answering

Linked data is built on standard web technologies such as RDF and URIs. SPARQL is the standard language for querying RDF data (Pérez, Arenas, and Gutierrez 2009). After data conversion into a LOD-based dataset, which can be accessed online via a SPARQL query endpoint, we furthermore design a template-based SPARQL query generator,

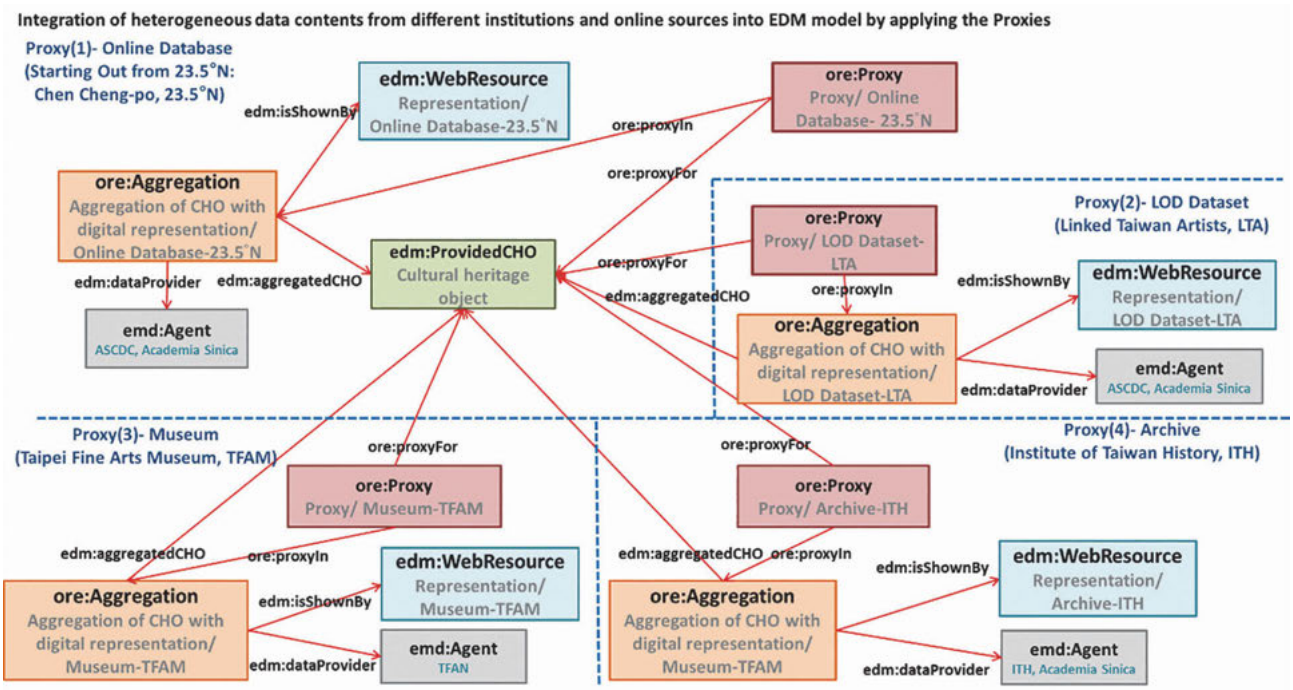

Figure 9. "Proxy" as a mechanism to integrate data referring the same object from different data providers in the model.

which benefits novice users by making it easier to give multiple and different possible semantic queries on one's interest within the designed template. With the example of CCP's archival materials, a template-based query can be formulated as follows: Which "oil painting" by Chen Cheng-po was created during his "study period at the Imperial Art Academy in Japan?" In such a question, information as "type of works" (e.g. oil paintings, drawings, etc.) and "periods in one's life of creation" (e.g. period in Japan, period in China, study period, etc.) are changeable depending on user interest. The following is the preliminary list, which provides scenarios and SPARQL template-based questions based on different categories of queries.

Since the archives' metadata records have been transformed into triple-based linked data (i.e., subject-property-object), this project provides a SPARQL query template that allows users to explore specific questions in a more flexible way (Table 9). For instance, users can ask for the subject based on a given property (e.g., the level of an archive) and object (e.g., series) (Table 9, first row-1), asking for the subject and property based on a given object (Table 9, second row-3), and asking for the subject and object based on a given property (Table 9, second row-8).

In number seven in the SPARQL query in Table 8, for example, the question is "What places did Chen paint? How often do they occur?" The procedures for SPARQL querying are as follows (Figure 10).

1. Locate all data fields in dcterms:spatial at the "item" level.
2. Apply statistics to the searched results.
3. Search results will appear in the fields of "item" and "place."

The results of the query show that ninety-three works are found as below (Figure 11) in the whole oeuvre of Chen

Category of query	**Scenario for query**	**SPARQL template-based question**
Data structure query	To understand the data structure	1. What are the items contained in each series in the archive of the "Chen Cheng-po's paintings and documents?"
Data content query- basic	To query the basic information of a painter's work	2. What are the watercolor paintings of Chen Cheng-po? Please list the title, alternative title, and the date of creation of these paintings.
	To study the material	3. Which of Chen's paintings are created on canvas? What are the titles of his paintings on canvas?
	To query a specific subject in Chen's artworks and their hierarchy within the collection	4. Which of Chen's paintings contain female nudity? What are the titles of these paintings and their files?
	To explore a painter appeared in which item level and what property/relation to the archival materials	5. In which archival item and which property could we find information on Paul Cezanne?
Data content query- advanced	To study the temporal- spatial scenario of a work's creation	6. During Chen's studentship at the Tokyo School of Fine Arts, which of his paintings depict the landscape of Chiayi? What are their dates of creation?
	To research content locations of the works	7. What (actual) places did Chen paint? How often do they occur in the body of his paintings?
Cross-dataset query	To find event information related to the artist	8. In Chen's postcard collections, which annual edition(s) of the "Taiwan Art Exhibition" (Taiten) postcards can be found? Please locate the data about the relevant editions of Taiten in Linked Taiwan Artists.
	To find related persons, who were influential to the artist	9. Which western artists appear in Chen's collage? Please locate the name of these artists and their data in ULAN.

Table 8. SPARQL template-based questions.

Number in Table 8	**Subject**	**Property**	**Object**	**Questions**
1	?	Level	Series	What are the items contained in each series in the archive of the "Chen Cheng-po's paintings and documents?"
5	?	?	Paul Cezanne	In which archival item and which property field could we find information on Paul Cezanne?
8	?	Related resource	?	Which western artists appear in Chen's collage? Please locate the names of these artists and their data in ULAN.

Table 9. Asking questions based on triple-based linked data.

```
prefix locah:<http://data.archiveshub.ac.uk/def/>
prefix dcterms:<http://purl.org/dc/terms/>
select *
from <http://data.ascdc.tw/archive/ith/ccp/>
where{
?item locah:level <http://data.archiveshub.ac.uk/id/level/item>;
dcterms:spatial ?spatial.
}
```

SUBMIT

Figure 10. SPARQL querying language for question seven.

Triples | Map | Chart | Relation chart

RESULTS ⬇ JSON XML TTL CSV

item	Works created by Chen Cheng-po	spatial	Place/ TGN
http://data.ascdc.tw/archive/ith/ccp/CCP_01_01002_OCJ1_10		http://vocab.getty.edu/tgn/7468019	
http://data.ascdc.tw/archive/ith/ccp/CCP_01_01009_OCJ1_27		http://vocab.getty.edu/tgn/1001032	
http://data.ascdc.tw/archive/ith/ccp/CCP_01_01012_OCJ1_16		http://vocab.getty.edu/tgn/1001032	
http://data.ascdc.tw/archive/ith/ccp/CCP_01_01013_OCJ1_08		http://vocab.getty.edu/tgn/1001032	
http://data.ascdc.tw/archive/ith/ccp/CCP_01_01014_OW5_05		http://vocab.getty.edu/tgn/1001032	
http://data.ascdc.tw/archive/ith/ccp/CCP_01_01015_OCJ1_05		http://vocab.getty.edu/tgn/1001032	

< 1 | 2 | 3 | 4 | 5 >

Figure 11. Results of queried works with TGN information on the depicted place of each work.

Cheng-po, each of them is shown with geographic information in TGN URI on the place depicted in the artwork, which demonstrates how external vocabularies have been reused in the study.

To better present the data, we further visualize the query results as follows (Figure 12).

The map (Figure 12) shows quantity and name of places, which are depicted in the entire opus of Chen Cheng-po. The GIS (Geographic Information System) map reveals the places in points, where the place is depicted by Chen Cheng-po in his work. After clicking one of a certain point, it also shows the related number of items and even the work titles (see Shanghai as an example in Figure 12).

7.0 Discussion

Yakel (2003: 1-2) proposed the concept of archival representation to refer to both the processes of arrangement and description, which is viewed as a fluid, evolving and socially constructed practice. The process of arrangement is about respecting order, and the process of description includes the creation of access tools (finding aids, bibliographic records) or systems (card catalogs, bibliographic databases, EAD databases). As technology is changing fast, it makes finding aids an ongoing advancement, and linked data principles are a promising method for extending the lifecycle of archival resources, in order to integrate,

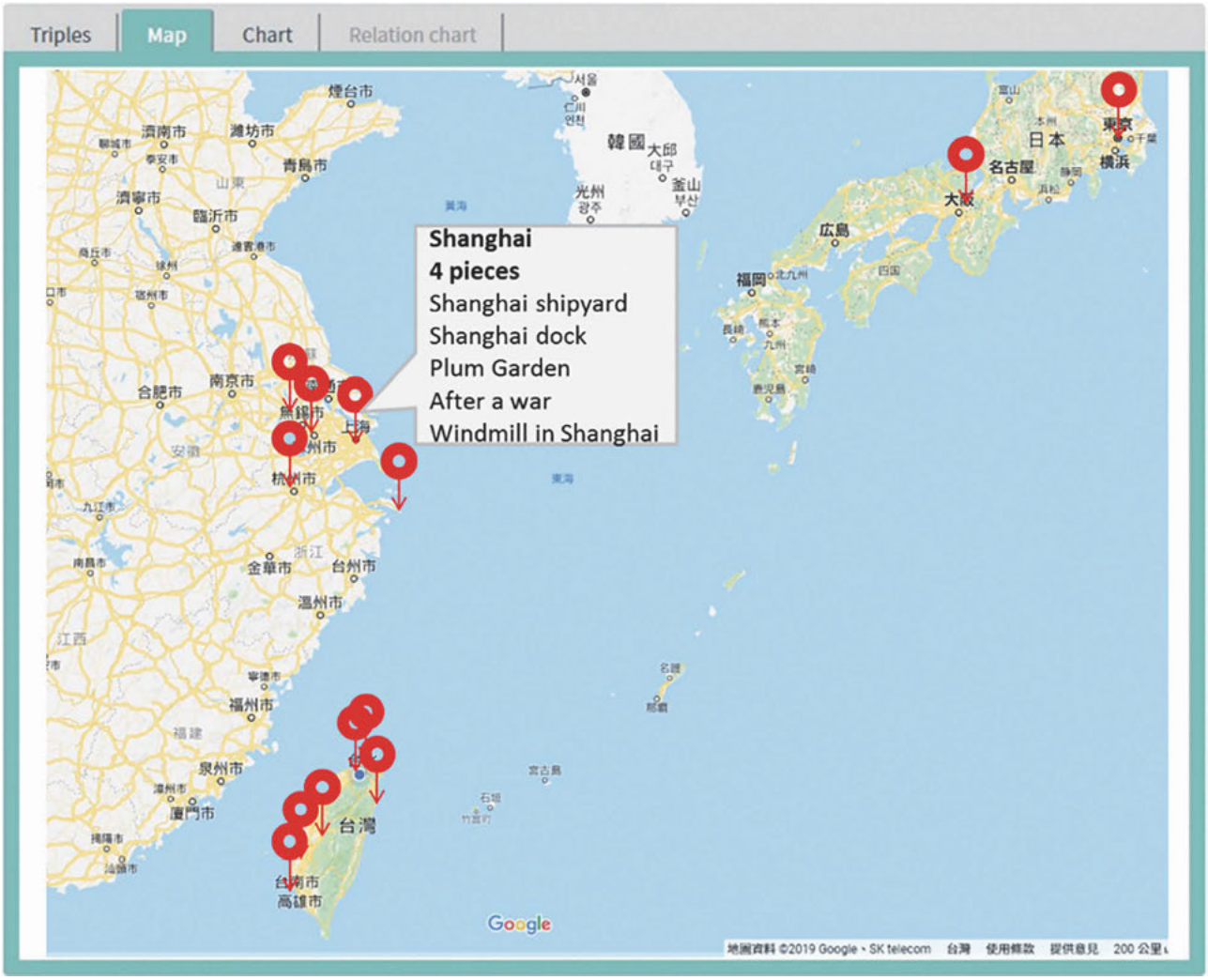

Figure 12. Places depicted in Chen's paintings shown in the GIS map with marks of corresponding locations.

represent and connect resources in a common model. An understanding of the new paradigm of archival representation in the environment of the semantic web can help develop workable methodologies, a variety of tools and good practices. These, in turn, can help archival organizations transform and manage their collections and facilitate a better experience for users with archival discovery. As an example, the Encoded Archival Description (EAD) is a commonly accepted standard and representational tool for encoding the descriptive information of the archival records and has been since the 1990s (Ruth 2001). It could benefit the management of the archival data in the past decades as follows: 1) in demonstrating the hierarchical structure of these collections, better characterizing them as a whole and their respective individual parts; and, 2) by utilizing XML as a structural and preservation format which will allow finding aids to be easily converted into different formats for ease of display and access (Society of American Archivists 2019).

However, the representational tools developed in the 1990s are confronting the potential limitations of data interoperability where the user can navigate resources regardless of their provenance, especially in the modern day semantic web environment, including: 1) the traditional representational tools are only human-readable, not machine-processable data formats which are RDF-based (The Resource Description Framework) semantic web technologies and enable computers to better manipulate information on our behalf; 2) the organizational representational system makes it hard to link related collections that are scattered among their different repositories, as the data are delivering documents containing the data themselves. It is important to shift the current document-oriented web into a web of interlinked data where the archival metadata records are deconstructed in triples (statements) and published in the linked data cloud, so archivists and users can reuse, consume, and repurpose these fundamental units of knowledge from different collections according to their

own purposes and research needs and connect the different data silos; and, 3) the content in the majority of current archival representation does not have a mechanism to unambiguously identify what a specific string of characters means and what the content is about. For example, if users search for the word "Venus" within collections, it is possible that the results will show the planet, the Botticelli painting (i.e., The Birth of Venus), or the tennis player (i.e., Venus Williams) (Képéklian, Curé, and Bihanic 2015, 61). To identify resources in a unique and universal manner, the URI (uniform resource identifier) is introduced to name a thing (i.e., creators, works, places, concepts, etc.) for the resources, so that the archivists and users are able to integrate data across diverse systems.

This study has demonstrated how to take advantage of linked data to enhance archival representation. To enhance the interchange, sharing, and enrichment of the archival data in the world of the semantic web, archival data structured as a LOD-based data might be a better method with following benefits:

1) LOD is a dataset of a knowledgebase in RDF standard, which could be readable and processed by machine and create linkage with external resources. This study has shown how to enrich the existing archival materials by reconciling external vocabularies which can contribute by providing context, accessing more information, and creating potential for new research questions;
2) The data are clustered based on several units of triple set and can be enriched or integrated with other triples of data from different institutions, domains, and topics. We have illustrated in this research how to reuse the Aggregation class and Proxy class of the Object Reuse and Exchange ontology (ORE) and their related properties (i.e., proxy-for, proxy-in) for representing aggregations of digital objects from different sources and institutions. It means that the associated resources, which are scattered in libraries, archives, and museums, are able to aggregate while keeping the different views distinct; and,
3) The data are open to free access and reusable as controlled vocabularies to enrich the data of other LOD-based datasets. We contributed data enrichment for the current archival collection by direct reuse of external vocabularies, such as *AAT*, *TGN* and VIAF. This will lower the cost of an archives' arrangement and description process and maintenance expenses in the long run.

Beyond these aforesaid advantages, a major challenge for the further developing of LOD-based datasets needs to be overcome. The development of the retrieval and presentation of the LOD-based data on a human readable interface is still at the initial stage. Although there are a few studies of exploring marrying data visualization to online archival finding aids (Kramer-Smyth, Nishigaki, and Anglade 2007; Lemieux 2012; Bahde 2017), few have been tackled in the LOD context. How to combine the semantic data with different, meaningful, and suitable models of data visualization and present the retrieved results is still an issue even after structuring a LOD-based application system.

8.0 Conclusion

This study was based on a fonds archive to develop and construct its linked data and provide SPARQL template-based question answering. The EDM has been adopted as the fundamental data model design, and data originated from heterogeneous providers or sources can be integrated by using mechanisms such as "proxy" and relational properties (edm:hasPart or edm:isRelatedTo), through which the original contextuality of different data are harmonized. We extend the EDM and create more specific properties to meet the characteristics of the CCP archival context. We demonstrated that the EDM model is a high-level and flexible model, which allows different data providers from different communities to interoperate in the model. We applied four approaches to enriching data by: 1) direct reuse of external vocabularies; 2) reconciliation of local links to other people's data; 3) introducing contextual classes that develop the appropriated contextual entities; and, 4) utilizing named entity extraction. To facilitate data access, we design a template-based SPARQL query generator that benefits novice users by making it easier to give multiple and different possible semantic queries. The outcome of this study can contribute to the methodology and best practice for developing linked data of nationwide art-related archival materials.

References

Baca, Murtha and Melissa Gill. 2015. "Encoding Multilingual Knowledge Systems in the Digital Age: The Getty Vocabularies." *Knowledge Organization* 42: 232-43.

Bahde, Anne. 2017. "Conceptual Data Visualization in Archival Finding Aids: Preliminary User Responses." *portal: Libraries and the Academy* 17: 485-506. doi:10.1353/pla.2017.0031

Bearman, David A. and Richard H. Lytle. 1985. "The Power of the Principle of Provenance." *Archivaria* 21: 14-27.

Berners-Lee, Tim. 2006. "Linked Data – Design Issues." http://www.w3.org/DesignIssues/LinkedData.html

Charles, Valentine and Antoine Isaac. 2015. "Enhancing the Europeana Data Model (EDM)." http://pro-beta.europeana.eu/files/Europeana_Professional/Publications/EDM_WhitePaper_17062015.pdf

Charles, Valentine and Marlies Olensky. "Report on Task Force on EDM Mappings, Refinements Extensions." http://pro-beta.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/Mapping_Refinement_Extension/EDM%20%20Mapping%20refinement%20extension%20Report.pdf

Chen, Shu-Jiun. 2017. "A Study of Linked Data for Digital Collections: A Case of the Painter Chen Cheng-Po." *Journal of Library & Information Science* 43: 71-96.

Clayphan, Robina, Valentine Charles and Antoine Isaac. 2017. "Europeana Data Model: Mapping Guidelines v2.4" http://www.bibnat.ro/dyn -doc/biblioteca%20digitala/EDM_Mapping_Guidelines_v2.4_102017.pdf

Daquino, Marilena, Mambelli, Francesca, Peroni, Silvio, Tomasi, Francesca, and Fabio Vitali. 2017. "Enhancing Semantic Expressivity in the Cultural Heritage Domain: Exposing the Zeri Photo Archive as Linked Open Data." *Journal on Computing and Cultural Heritage* 10: 21.

Douglas, Jennifer, Greg Bak, Evelyn McLellan, Seth van Hooland, and Raymond Frogner. 2018. "Decolonizing Archival Description: Can Linked Data Help?" *Proceedings of the Association for Information Science and Technology* 55: 669-72. doi:10.1002/pra2.2018.14505501077

Francisco-Revilla, Luis, Ciaran B. Trace, Haoyang Li, and Sarah A. Buchanan. 2014. "Encoded Archival Description: Data Quality and Analysis." *Proceedings of the American Society for Information Science and Technology* 51: 1-10. doi:10.1002/meet.2014.14505101043

Guimarães, José Augusto Chaves and Natália Bolfarini Tognoli. 2015. "Provenance as a Domain Analysis Approach in Archival Knowledge Organization." *Knowledge Organization* 42: 562-9.

Higgins, Sarah and Gavin Inglis. 2003. "Implementing EAD: The Experience of the NAHSTE Project." *Journal of the Society of Archivists* 242: 199-214.

Hsiao, Chong Ray. 2014. "The Historical Significance of Chen Cheng-po's 120th Birthday Anniversary Touring Exhibition." In *Surging Waves*, ed. T.S. Yeah. Tainan, Taiwan: Tainan City Government, 10-25.

Hyland, Bernadette and Boris Villazón Terrazas. 2011. "Linked Data Cookbook. World Wide Web Consortium." http://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook.

Isaac, Antoine. 2013. "Europeana Data Model Primer." http://travesia.mcu.es/portalnb/jspui/bitstream/10421/5981/1/EDM_primer.pdf

Isaac, Antoine, Hugo Manguinhas, Valentine Charles and Juliane Stiller. 2015. "Selecting Target Datasets for Semantic Enrichment: Companion Document to the Report of the EuropeanaTech Task Force on Enrichment and Evaluation." https://pro.europeana.eu/files/Europeana_Professional/EuropeanaTech/EuropeanaTech_taskforces/Enrichment_Evaluation/EvaluationEnrichment_SelectingDatasets_102015.pdf

International Council on Archives. 2000. *ISAD(G): General International Standard Archival Description*. 2nd ed. Ottawa: International Council on Archives.

Gabriel, Képéklian, Olivier Curé and Laurent Bihanic. 2015. "From the Web of Documents to the Linked Data." In *Business Intelligence: 4th European Summer School, eBISS 2014, Berlin, Germany, July 6-11, 2014, Tutorial Lectures*, ed. Esteban Zimányi and Ralf-Detlef Kutsche. Lecture Notes in Business Information Processing 205. Cham: Springer, 60-87. doi:10.1007/978-3-319-17551-5_3

Kramer-Smyth, Jeanne, Morimichi Nishigaki and Tim Anglade. 2007. "ArchivesZ: Visualizing archival collections." http://archivesz.com/ArchivesZ.pdf

Lemieux, Victoria. 2012. "Envisioning a Sustainable Future for Archives: A Role for Visual Analytics." Paper presented at the International Council on Archives Conference Brisbane, Australia, August 20-24 2012. http://ica2012.ica.org/files/pdf/Full%20papers%20upload/ica12Final00239.pdf

Lin, Yuchun. 2012. "The Real Scenes and Images Utopia in Chen Cheng-po's Journey of Life." In *Journey Through Jiangnan: A Pivotal Moment in Chen Chen-po's Artistic Quest*. Taipei, Taiwan: Taipei Fine Arts Museum, 6-15.

Mathison, Christina Burke. 2012. "Identity, Hybridity, and Modernity: The Colonial Paintings of Chen Cheng-po." In *Journey Through Jiangnan: A Pivotal Moment in Chen Chen-po's Artistic Quest*. Taipei, Taiwan: Taipei Fine Arts Museum, 50-63.

Miguez, Matthew Roland. 2018. "Linked Data for Archivists: Graphs and Rhizomes." *Society of Florida Archivists Journal* 1: 6-12.

Niu, Jinfang. 2016. "Linked Data for Archives." *Archivaria* 82: 83-110.

Pearce-Moses, Richard. 2005. *A Glossary of Archival and Records Terminology*. Chicago, IL: Society of American Archivists.

Pérez, Jorge, Marcelo Arenas and Claudio Gutierrez. 2009. "Semantics and Complexity of SPARQL." *ACM Transactions on Database Systems (TODS)* 34, no. 3: Article 16. doi:10.1145/1567274.1567278

Ruth, Janice E. 2001. "The Development and Structure of the Encoded Archival Description (EAD) Document Type Definition." *Journal of Internet Cataloging* 4, no. 3/4: 27-59.

Sanderson, Robert. 2014. "Reconciliation of Linked Data in the Cultural Heritage Sector." https://mellon.org/media/filer_public/7f/8e/7f8e4ddf-97c4-44a2-a1d6-cf333acf1916/sanderson_lod_reconciliationreport_12-2014.pdf

Sanderson, Robert. 2016. "The Linked Data Snowball or Why We Need Reconciliation." Slides of paper presented at the AAC/Getty Workshop on Reconciliation of Linked Open Data April 4, 2016. https://www.slideshare.net/azaroth42/linked-data-snowball-or-why-we-need-reconciliation

Schaible, Johann, Thomas Gottron, and Ansgar Scherp. 2014. "Survey on Common Strategies of Vocabulary Reuse in Linked Open Data Modeling." In *The Semantic Web: Trends and Challenges; 11th International Conference, ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014. Proceedings*, ed. Valentina Presutti, Claudia d'Amato, Fabien Gandon, Mathieu d'Aquin, Steffen Staab and Anna Tordai. Lecture Notes in Computer Science 8465. Cham: Springer, 457-72.

Society of American Archivists. 2019. "Frequently Asked Questions about EAD and EAD3." https://www2.archivists.org/groups/encoded-archival-standards-section/frequently-asked-questions-about-ead-and-ead3

Tudhope, Douglas and Ceri Binding. 2016. "Still Quite Popular After all Those Years: The Continued Relevance of the Information Retrieval Thesaurus." *Knowledge Organization* 43: 174-9.

Yakel, Elizabeth. 2003. "Archival Representation." *Archival Science* 3: 1-25.

Zhang, Jane. 2012. "Archival Context, Digital Content, and the Ethics of Digital Archival Representation." *Knowledge Organization* 39: 332-9.