informatics

MDPI

Article

Metadata Integration Framework for Data Integration of Socio-Cultural Anthropology Digital Repositories: A Case Study of Princess Maha Chakri Sirindhorn Anthropology Centre

Marut Buranarach [1], Watchira Buranasing [1], Sittisak Rungcharoensuksri [2], Panita Sarawasee [2], Treepidok Ngootip [3] and Wirapong Chansanam [3,*]

[1] National Electronics and Computer Technology Center, Bangkok 12120, Thailand; marut.buranarach@nectec.or.th (M.B.); watchira.bur@nectec.or.th (W.B.)
[2] Princess Maha Chakri Sirindhorn Anthropology Centre, 20 Borommaratchachonnani Rd., Taling Chan, Bangkok 10170, Thailand; sittisak.r@sac.or.th (S.R.); panita.s@sac.or.th (P.S.)
[3] Department of Information Science, Faculty of Humanities and Social Sciences, Khon Kaen University, Khon Kaen 40002, Thailand; treepidok.n@kkumail.com
* Correspondence: wirach@kku.ac.th; Tel.: +66-90956-2727

Abstract: Data integration is one of the most challenging tasks for digital collections whose data are stored across various repositories. Data integration across digital repositories has several challenges. First, data heterogeneity in terms of data schema and data values usually occurs across diverse data sources. Second, heterogeneity in data representation and semantic issues are among the problems. The same data may appear in different repositories with varied data representations, i.e., metadata schema. Recent research has focused on matching several related metadata schemas. In this paper, a metadata integration framework is proposed to support digital repositories in socio-cultural anthropology at the Princess Maha Chakri Sirindhorn Anthropology Centre (SAC), Thailand. The proposed framework is defined based on the Metadata Lifecycle Model (MLM). It utilizes non-procedural schema mappings to express data relationships in diverse schemas. A case study of metadata integration over the SAC digital repositories was conducted to validate the framework. The SAC common metadata schema was designed to support data mapping across 13 digital repositories. The SAC "One Search" system was developed to exemplify the system implementation of the framework. Evaluation results showed that the proposed metadata integration framework can support domain experts in socio-cultural anthropology in unified searching across the repositories.

Keywords: metadata schema; anthropology; metadata integration; digital humanities; schema mapping

check for updates

Citation: Buranarach, M.; Buranasing, W.; Rungcharoensuksri, S.; Sarawasee, P.; Ngootip, T.; Chansanam, W. Metadata Integration Framework for Data Integration of Socio-Cultural Anthropology Digital Repositories: A Case Study of Princess Maha Chakri Sirindhorn Anthropology Centre. *Informatics* **2022**, *9*, 38. https://doi.org/10.3390/informatics9020038

Academic Editor: Hsuanwei Michelle Chen

Received: 10 March 2022
Accepted: 23 April 2022
Published: 27 April 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

For more than 30 years, Thailand's Princess Maha Chakri Sirindhorn Anthropology Centre (SAC) has been developing digital repositories in anthropology, archaeology, history, ethnology, and socio-cultural studies for academic forums and the general public. Currently, the center has 31 digital repositories with more than 140,000 digital resources, including data records, online databases, articles, e-books, newsletters, videos, photos, audio files, etc. [1]. The data provided by the center are academically reliable and cover a wide range of fields, so these databases have become one of the most important online information sources in the anthropology field in Thailand.

Although each digital repository has its own purposes and designs, with some shared common entities such as ethnic groups, the data have been stored in different locations and database systems. This creates a limitation wherein users are unable to search for the desired information in an integrated and unified fashion. Moreover, the user interface (UI) and the data schema of the data records are different in each database. As a result, users who are not familiar with the subjects of all the available repositories can be confused as to where to search for the information. Therefore, to solve the problem of data silos and disparate retrieval systems in which the data in the repositories are not linked, and to implement a unified search UI, the SAC's "One Search" approach is designed and developed to integrate the data of all repositories and to enable users to access and retrieve all data within the SAC digital repository through one search channel.

In this paper, a metadata integration framework is proposed to support data integration across digital repositories. The proposed framework is defined based on the Metadata Lifecycle Model (MLM). The framework consists of five steps: analyzing information content, creating metadata requirements, developing metadata schema, creating metadata schema mapping profiles, and developing metadata service system and evaluation. A case study of metadata integration over the SAC digital repositories was conducted to validate the framework.

In adopting the framework, the SAC common metadata schema was designed based on the existing metadata schema standards, i.e., Dublin Core (DC) and Europeana Data Model (EDM). The design was also based on content analysis of the SAC digital repositories. The metadata schema mapping profiles define non-procedural schema mappings to express data relationships in diverse schemas. Using the mapping profiles, the existing source metadata schemas were subsequently mapped into the target SAC common metadata schema. Thus, the integration of data from different sources can be conducted, and a search system can be developed based on the SAC common metadata schema. A prototype of the SAC "One Search" system over 13 SAC digital repositories exemplified data integration and unified search system development. Based on the evaluation results, the unified search system provides sufficient support for the description and retrieval of the data by domain experts.

2. Background

2.1. Data Heterogeneity

Data heterogeneity is a common phenomenon in distributed information sources and is growing with the development of systems and applications, which has created an enormous amount of data and information [2,3]. When data are used, sharing and integrating data causes a challenge in the implementation process [4–9]. Data non-standardization, diverse data representation, data disputes, and data with related semantic features are some of the issues that may be found inside the data [10].

There are still numerous issues to be overcome in the deployment of data integration. Sharing and integrating data from loosely coupled sources, heterogeneity of data representation, and mapping data from diverse data sources are the most challenging aspects of data integration [11–14]. The semantic characteristics of multiple data forms and sources are particularly problematic when dealing with extensive data, which almost certainly contain heterogeneous data [10,13,15,16].

One of the most vexing issues in data management is automatically identifying proper mappings between various structured data types [17,18]. Data mappings are fundamental in data cleaning [19,20], data integration [21], and semantic integration [22,23]. In addition, they constitute the fundamental connection for the construction of large-scale semantic web and peer-to-peer information systems, which promote the collaboration of independent data sources [24]. As a result, the challenge of data mapping is manifested in different ways, including schema matching [25,26], schema mapping [17,27], ontology alignment [28], and model matching [18,29].

From a semantic perspective, a semantic data mapping procedure is one of the potential approaches to solving heterogeneous data problems from a semantic perspective [30–34]. The main objective of the semantic data mapping process is to produce data format representations from data sources and convert them into an XML data format using a semantic perspective [35–37]. This is an important process in the implementation of data integration technology [38]. The semantic data mapping process is the standardization and mapping process to produce uniformity between data with various data representations, heterogeneity format data, and different semantic aspects between applications in the other data sources [39–41].

2.2. Semantic Data Integration

Integrating datasets or data sources is a major problem for semantic integration because of the complexity of identifying that the data contain semantic information. The semantic information determined from the data refers to real-world concepts and can be integrated. Many technologies are used for semantic integration to fix the challenges it faces. This section will discuss approaches, frameworks, techniques, and related challenges for semantic integration. Schema matching is the task of finding semantic correspondences between elements (or attributes) of two given database schemas [42–46]. This task is essential for enabling data integration and systems interoperability in e-commerce, geospace, biology, health, etc.

There are several reasons that the schema matching challenge is difficult: different schemas might have various names for items, such as attributes that express the same conceptual idea. On the other hand, elements with similar terminology might be referred to differently. It is possible that items that are structurally equal in two schemas vary. Many items from one schema may represent a single element representing a notion from another schema.

Semantic correspondences between items in two schemas are found through schema matching [43]. Database schemas, XML DTDs, HTML form elements, and other types of heterogeneous data sources are all good sources of schemas [44]. Connecting two disparate data sources is an important initial step in any integration [45].

While there have been several approaches to this problem throughout the years, none of them are now regarded as full solutions. When using a technique, it is sometimes necessary for a specialized user to check the results to ensure that they are accurate. Schema matching methods typically use one or more functions to establish a similarity value between pairs of schema items. The elements' similarity increases with the value of the parameter. A pair is referred to as a matching candidate. Between 0 and 1, these matchers evaluate the similarity of two input items. Schema element names, thesaurus-based semantic similarity, data type, cardinality comparisons, and even access data values may all be used by matchers to assess similarities.

It is possible to combine data integration and data semantics in a method known as semantic integration. Using numerous data sources to manipulate them transparently is essential for data integration [46]. It is possible to describe semantics as "the field of linguistics and logic concerned with meaning" [47] while addressing the topic. A technique that employs conceptual models of the bonds or connections and a representation of data conceptually, reducing any heterogeneities, is achieved when semantics and data integration are integrated. The integration of semantically diverse data is a key challenge. Structure and semantic heterogeneity are two forms of data heterogeneity difficulties [48]. Goh summarized the reasons for semantic heterogeneity [49]. The reasons are listed below:

- Naming Conflicts: Consists of synonyms and homonyms among attribute values.
- Scaling and Units Conflicts: Adoption of different unit measures or scales in reporting.
- Confounding Conflicts: Arise from the confounding of distinct concepts.

By achieving data interoperability, ontology is accountable for resolving data heterogeneity. Gruber defined ontology as the "specification of a conceptualization" [50].

2.3. Metadata Ontology

Cverdelj-Fogaraši and colleagues proposed one of the recent techniques for semantic data integration: metadata ontology [51]. The proposed technique is focused on semantic integration for information systems. The method provides semantics to document metadata descriptions and enables semantic mapping between metadata of a domain and metadata of another field. The metadata ontology technique consists of the service layer, data access layer, and persistence layer.

In order to implement the metadata ontology, the ebXML Registry Information Model standard [52] can be utilized to specify the metadata. There are four parts to the metadata ontology: the core, classification, association, and provenance. The major components are illustrated in Figure 1. It was tested and evaluated in real-life data by two independent departments successfully [53]. It is important to remember that the core classes and related attributes, classification, and association all fall under this system's "core" category, as with the higher ontology idea of provenance.

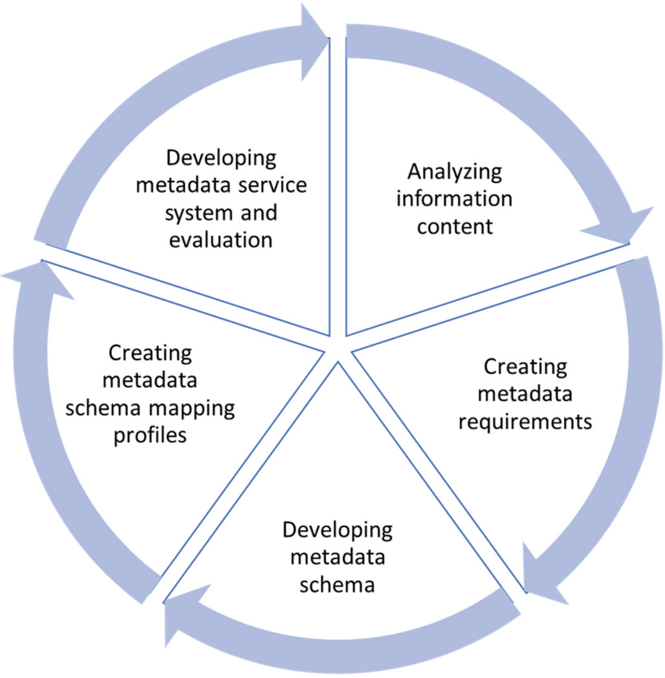

Developing metadata service system and evaluation
Analyzing information content
Creating metadata requirements
Developing metadata schema
Creating metadata schema mapping profiles

Figure 1. Metadata integration framework defined based on the Metadata Lifecycle Model.

One of the fundamental challenges for semantic integration is data heterogeneity. There are three types of data heterogeneity. Syntactic heterogeneity is caused by the use of different models or languages. Schema heterogeneity results from structural differences. Semantic heterogeneity is caused by different meanings or interpretations of data in various contexts [49]. In addition to the challenges mentioned above, there are other challenges to implementing the semantic integration architecture in real life [53]. These challenges may be divided into the following primary categories: scalability with the size of the schema, user interaction, and mapping maintenance [49]. Whereas most methodologies focus on small-sized schema, techniques that work well with large-sized schemas must be investigated. Schema mapping cannot be completely autonomous. Thus, designing interaction with the user in performing a schema mapping task is a significant challenge. Schemas often change. Thus, schema matching techniques must also facilitate mapping maintenance.

3. Methodology

In supporting data integration across digital repositories, a metadata integration framework is defined based on the Metadata Lifecycle Model (MLM) [54]. MLM, proposed by the Metadata Architecture and Application Team, is a methodology involving a ten-step process by which digital library projects can design and implement metadata provision. MLM emphasizes the iterative processes from requirement and content analysis and system specification to metadata system and service evaluation.

The proposed metadata integration framework, shown in Figure 1, is a generic framework that not only can guide the design of metadata schema of digital repositories based on requirement and content analysis but also cover the process of metadata schema mappings across digital repositories and metadata service system development. The framework consists of five steps: analyzing information content, creating metadata requirements, developing metadata schema, creating metadata schema mapping pro-files, and developing metadata service system and evaluation. The steps in adopting the framework for the SAC digital repositories are described as follows.

3.1. Analyzing Information Content

The SAC digital repositories, when considering content, context, and structure, can be classified into five groups as follows: 1. Ethnic Groups; 2. Museums and Archives; 3. Cultural Heritage; 4. Archaeology and History, and 5. Anthropology. Details are shown in Table 1.

Table 1. Groups of anthropology digital repositories of SAC.

	Ethnic Groups	**Museum and Archives**	**Cultural Heritage**	**Archaeology and History**	**Anthropology**
1	Ethnic Groups in Thailand	Anthropology Archive Database	Folk Toys of Thailand	Manuscripts of Western Thailand	Anthropology Concepts
2	Ethnic Groups Research in Thailand	Museums in Thailand	Rituals, Ceremonies and Local Festivals in Thailand Database	The Inscriptions in Thailand Database	Anthropology Clipping
3	Ethnographic Films Database	COVID-19 Digital Archive	Thai Literature Directory	Archaeological Sites in Thailand Database	SAC's Research Database
4	Cultural Ethnographic Map in Sakhorn Buri	Siam Rare Books	Database of Southeast Asian Sociocultural Information	Arts in Thailand Database	Sociologist and Anthropologist in Thailand Database
5			Traditional Objects of Everyday Use	Arts in Southeast Asia	
6			Vernacular Houses in Thailand Database	The Epigraphic Archives of Wat Pho	
7			Folktales Database	Potteries in Thailand Database	
8			Toponym Database	Physical Anthropology in Thailand	
9			Community Archive	Prof. Prasert Na Nagara	
10			Samut Sakhon Religious Sites and Shrines		

3.2. Creating Metadata Requirements

This research used the content analysis and related metadata standards to create the guidelines for identifying metadata by adapting and applying the Dublin Core (DC) metadata [55] to analyze the elements of SAC's repositories. DC was used as a base model to design the metadata schema. The Dublin Core Metadata Element Set comprises 15 elements as follows. (1) Contributor—an entity responsible for making contributions to the resource. (2) Coverage—the spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant. (3) Creator—an entity primarily responsible for making the resource. (4) Date—a point or period of time associated with an event in the lifecycle of the resource. (5) Description—an account of the resource. (6) Format—the file format, physical medium, or dimensions of the resource. (7) Identifier—an unambiguous reference to the resource within a given context. (8) Language—a language of the resource. (9) Publisher—an entity responsible for making the resource available. (10) Relation—a related resource. (11) Rights—information about rights held in and over the resource. (12) Source—a related resource from which the described resource is derived. (13) Subject—the topic of the resource. (14) Title—a name given to the resource. (15) Type—the nature or genre of the resource.

3.3. Developing the Metadata Schema

In developing the common metadata schema for the SAC's repositories, metadata elements were defined based on the Dublin Core Metadata elements. Some DC metadata elements were selected based on their appropriateness in the context of the subjects of SAC repositories. Next, the selected elements from DC were adopted together with new elements to ensure that the developed metadata schema could describe and enable users to access the needed information.

The SAC common metadata elements defined in this research consist of 11 metadata elements. They are based on five metadata elements from the Dublin Core Metadata [55] and six metadata elements from the Europeana Data Model (EDM) mappings of Europeana [56], which is related to a variety of anthropological data and a large amount of image data. The SAC common metadata elements that are based on the Dublin Core Metadata [55] consist of five elements: (1) Title, (2) Description, (3) Creator, (4) Type, and (5) Relation. The elements that are based on the EDM consist of six metadata elements: Properties, Provenance, Time, Location, Rights, and References.

3.4. Creating Metadata Schema Mapping Profiles

In this step, the metadata elements of existing SAC digital repositories, i.e., source metadata elements, are grouped based on the metadata elements of the SAC common metadata schema, i.e., target metadata elements. The mapping can have one to many relationships. Specifically, more than one source metadata element of a repository can be grouped into one target metadata element. For example, the Anthropology Museum repository contains two metadata elements, Title and Alternate Title, which can be grouped into the Title element of the SAC common metadata schema.

The metadata schema mapping profile stores all the mapping information between the source and the target metadata elements. The mapping profile can be represented in the form "Source Repository Name (Metadata Element Names) => Target Metadata Schema Name (Metadata Element Name), e.g., "Museum (Title, Alternative title) => SACCommon (Title)". The use of mapping profiles can facilitate mapping maintenance, i.e., profile updates, when the source metadata element names are added or updated.

3.5. Developing Metadata Service System and Evaluation

In adopting the SAC common metadata schema and mapping profiles, the SAC "One Search" prototype system is developed. The development of the prototype system consists of two major steps: metadata transformation and search system development. The steps are described as follows.

- Metadata transformation. In this step, the metadata schema mapping profiles are added into the search system. The mapping profiles allow the metadata elements of all the repository resources to be transformed into the SAC common metadata elements. Specifically, the resources of the source repositories will be described based on the SAC common metadata elements in an integrated repository. The source metadata element names are also preserved for display purposes.
- Search system development. A prototype search system called SAC "One Search" is developed. The system allows all the repository resources to be displayed and searched in a unified fashion. Specifically, the resources of the source repositories will be displayed based on the SAC common metadata elements. In addition, user queries in terms of SAC common metadata schema can be conducted.

The major components of the SAC's "One Search" prototype system are illustrated in Figure 2.

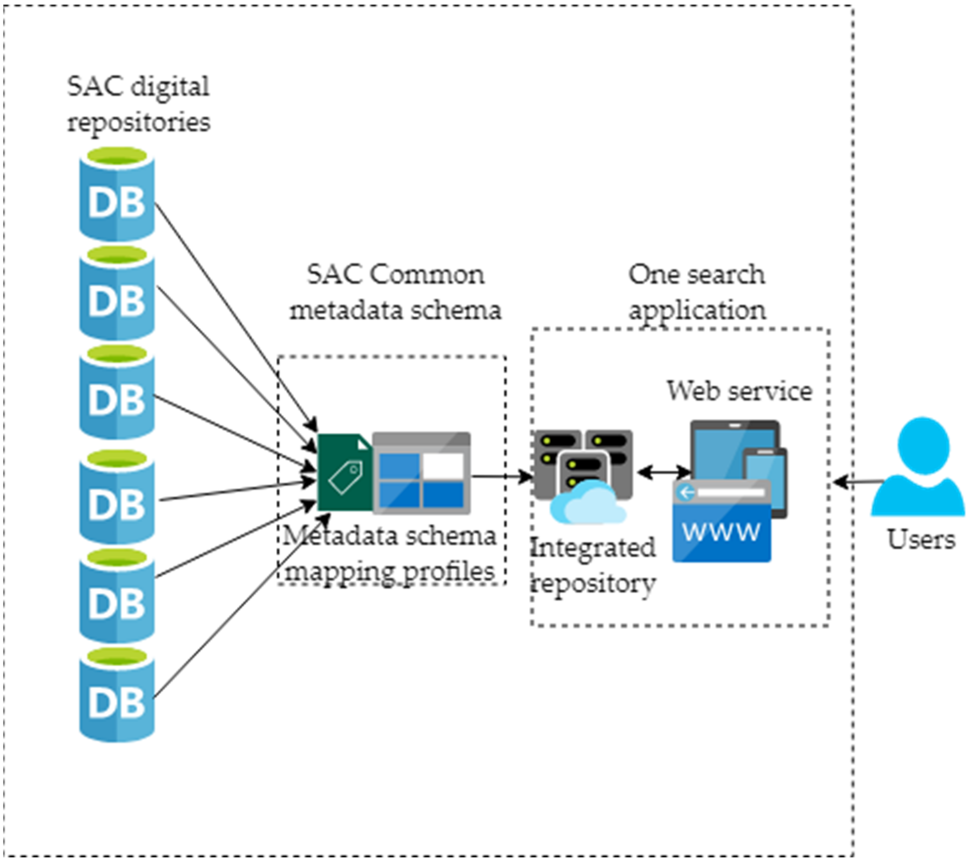


Figure 2. Data integration among SAC's digital repositories based on metadata schema mapping.

4. Results

4.1. SAC Common Metadata Schema

The SAC common metadata schema, developed as a common metadata schema for the repositories of the Princess Maha Chakri Sirindhorn Anthropology Centre, consists of 11 metadata elements, as shown in Table 2. The description of each metadata element consists of a name, definition, format, and example.

Table 2. Description of SAC's common metadata elements.

Element 1	
Name	Title
Definition	A name given to refer to the resources, originally derived from the name of the source material. It is a distinctive name that expresses an identity, a place, or a cultural expression. Its scope of elements includes Title, Alternative title, Autonym, Exonym, Common name, etc.
Format	Text
Example	- Karen (autonym of an ethnic group) - Chan Sen Museum (museum name) - Anan Ganjanaphan (title of archive collection) - Wat Mai Nakhon Ban (title of manuscript collection)

Table 2. *Cont*.

Element 2	
Name	Description
Definition	Details about the resources, including descriptions related to the resource in terms of its physical characteristic, content, and context. Its scope of elements includes Abstract (book abstract), Settlement pattern (ethnic group's settlement), Material (list of material sources), Art (description about art objects), Scope and content (archival description), etc.
Format	Text
Example	The collection consists of 754 photographs, which were recorded by Anan Ganjanapan during his anthropological fieldwork in Thailand. There are 3 series of material in his collection: Research at Ban Sanpong, Sanpathong, Chiengmai in 1980–1981, Ancestral Ritual, Prae and Lampang Province in 1986, and Wedding of Lau, Mae Hong Son Province in 1987.
Element 3	
Name	Creator
Definition	Creator(s) and people involved in the creation of the original and digital resources. Its scope of elements includes Creator, Contributor, Translator, Cataloger name, Administrator, etc.
Format	Text
Example	- Charles F. Keyes (Creator—Donor) - Thanwadee Sookprasert (Contributor—Administrator) - Siwapong Wongkoon (Contributor—Photographer)
Element 4	
Name	Type
Definition	The type and subject of the resources. Its scope of elements includes Collection, Type of Materials, Type of artwork, Subject, Keyword, etc.
Format	Text
Example	- Articles - Moving Images - Ethnography - Pre-historic pottery - Hunting tools
Element 5	
Name	Properties
Definition	The format and language of the resources. Its scope of elements includes Format, Language, Shape, etc.
Format	Text
Example	- Karen - THA - Stamped, cord-marked, applique - Earthenware
Element 6	
Name	Provenance
Definition	The current location(s) of the resources and its original source before being relocated. Its scope of elements includes Identifier, Location, Publisher, LC Call No., etc.
Format	Text
Example	- Kaothai No.33 November–December 2012 - By the collaboration between the Princess Maha Chakri Sirindhorn Anthropology Centre and the University of Washington (UW) under the Digital Archive of Research on Thailand (DART) project, 2010. - Nong Bua Kok Local Museum, Buriram

Table 2. *Cont.*

Element 7	
Name	Time
Definition	The time period in which the resource was created, including the time period related to the content and context of the resource and the time period in which the digital resource was created. Its scope of elements includes Period, Last updated, Year, Publication year, etc.
Format	Text and date
Example	- 20 May 2018 - Bronze Age, Early Ban Chiang Period
Element 8	
Name	Location
Definition	The location(s) related to the creation of the resource or the current location(s) of the resource. Its scope of elements includes Region, Province, Latitude, Longitude, Coverage, etc.
Format	Text
Example	- NE of Thailand (region) - Udon Thani (province) - JuiTui Tao Bo Keang Foundation, 283 Soi Phuthorn, Ranong rd., Muang District, Phuket Province 83000
Element 9	
Name	Reference
Definition	Details of reference(s) used to create the resources. Its scope of elements includes Reference, ISBN, Bibliography, etc.
Format	Text
Example	- Thamnu Worathongchai. 2008. Hunt, Snare, Catch with Isan's Animal Traps. Bangkok: Princess Maha Chakri Sirindhorn Anthropology Centre (Public Organization)
Element 10	
Name	Relation
Definition	Location(s) of other related resources. Its scope of elements includes Dataset, Related Information, Relation, etc.
Format	Text, URL
Example	- Book/Journal Title: Ban Chiang World Heritage - News about the Bang Kloi Karen (url)
Element 11	
Name	Rights
Definition	The right holder(s) and the types of licenses to use the resource. Its scope of elements includes Rights, Type of license, Conditions governing access and reproduction.
Format	Text
Example	- SAC - CC BY - No restrictions on access

4.2. Metadata Schema Mappings

The results of metadata element mappings between the 13 source digital repositories and the SAC common metadata schema are shown in Table 3. For brevity, only partial lists of the metadata schema mapping profiles are shown.

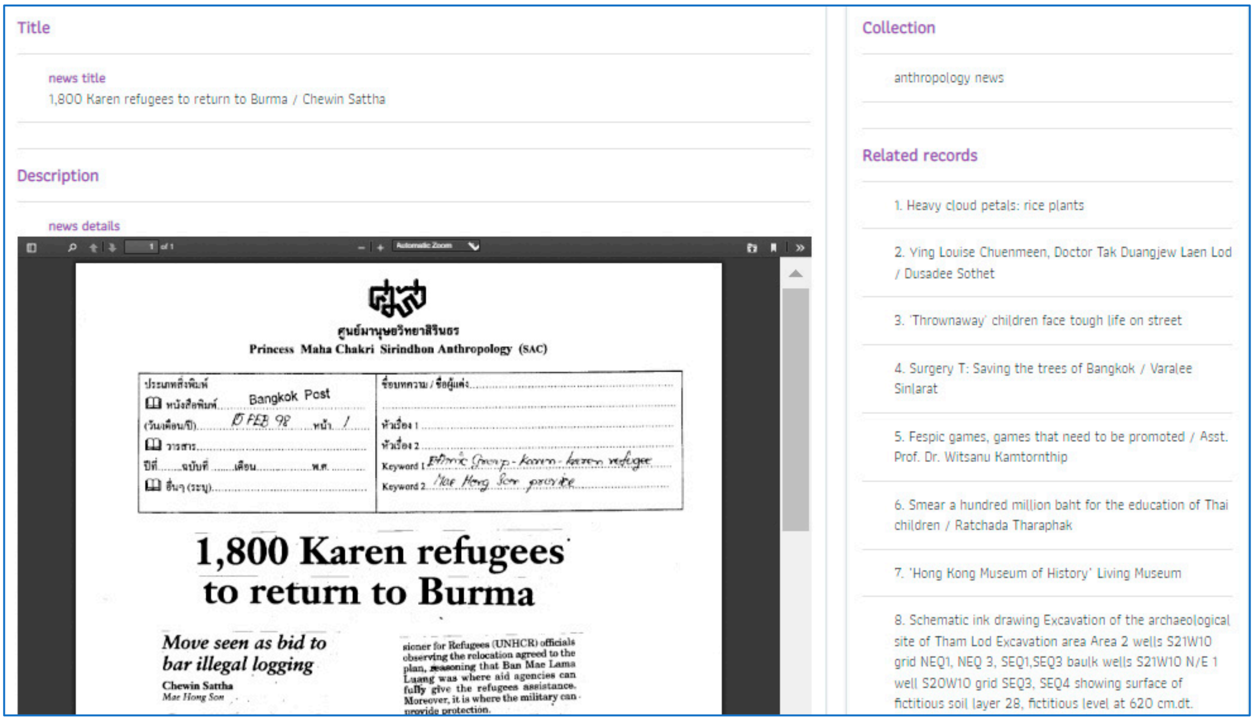
Table 3. Metadata element mapping between source digital repositories and SAC common metadata schema.

SAC Common Metadata Elements	Source Repository Name (Metadata Element Names) (Partial Lists)
Title	Museum (Title, Alternative title); Inscription (Title, Alternative title); Ancient documents (Title, Alternative title); Tools (Title, Alternative title, Local title); Ethnographic research (Title, Translated title); Ethnic groups (Title, Autonym, Exonym)
Description	Inscription (Description, History, Script, Age determination, Reference naming list, Transliteration-translation); Ancient documents (Description, Script); Tools (Description, Material); Anthropologists' archives (Scope and content); Ethnic groups (Compellation of ethnonym, Overview, Settlement pattern, Socio-cultural context, Demographic context, Current situation)
Creator	Museum (Founder, Curator); Inscription (Creator, Editor); Ancient documents (Creator, Data curator); Tools (Creator, Contributor); Ethnographic research (Author, Translator, Text analysist); Ethnic groups (Creator, Editor, Data curator)
Type	Museum (Collection, Type of museums, Museum categories, Type of management); Inscription (Collection); Ancient documents (Collection, Type of manuscript, Type of source); Tools (Collection, classification); Anthropology news (Collection); Ethnic groups (Collection, Script); Folk toys (Collection); Archaeological site (Collection, Type of archaeological site); Thai art (Collection, Type of artwork)
Properties	Inscription (Size, Face/line, Material, Form); Ancient documents (Size, Quality, Language); Ethnographic research (Original language of text, Total pages); SAC's research (Original language, Total pages); Thai art (Size)
Provenance	Ancient documents (ID); Tools (ID); Anthropologists' archives (Identifier); Anthropology news (Publisher); Ethnographic research (Location of documents, Funding organization); SAC's research (LC Call No., Provenance, Publisher, Publication place); Thai cultural encyclopedia (SAC Label); Thai art (Identifier)
Time	Museum (Founded, Survey date, First published, Last updated); Ancient documents (Period, Creation date, Digital creation date, First published); Tools (Creation date, Survey date, First published, First updated); Anthropologists' archives (Creation date, First published, Last updated); Anthropology news (Date, Volume, Issue); Ethnographic research (Year, Year of ways, Study period, First published); Ethnic groups (First published, Last updated);
References	Museum (References); Inscription (source, bibliography, photo source); Tools (References); Ethnographic research (Sources); Ethnic groups (References); Thai cultural encyclopedia (Reference, SAC item, ISBN); Folk toys (References); Archaeological site (Bibliography); Thai art (Bibliography)
Relations	Inscription (News, Dataset); Ethnographic research (Related information); SAC's research (Relation); Thai cultural encyclopedia (Relation); Folk toys (Relation); Archaeological site (Related record); Thai art (Related information)
Location	Museum (Region, Province, Address Latitude, Longitude); Inscription (Founding location, Founding province, Current preserved location, Latitude, Longitude); Ancient documents (Current preserved location, Address, Province, District, Sub-district, Zip code, Latitude, Longitude); Tools (Source Coverage, Latitude, Longitude); Anthropologists' archives (Coverage, Address, Region, Province, District, Sub-district, Latitude, Longitude)
Rights	SAC's research (Rights, Type of license); Thai cultural encyclopedia (Rights, Type of license)

4.3. Unified Search System Development

A prototype search system was developed as a unified metadata service system. The system organizes and presents data from 13 SAC digital repositories based on the SAC common metadata schema. The Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH) [57] is a standard used to retrieve the data from the source repositories.

The SAC "One Search" system for demonstrating the unified metadata approach of the Sirindhorn Anthropology Centre can be accessed at: https://onedb.sac.or.th/, accessed on 9 March 2022. The system allows the unified representation and searching of 140,000 digital resources from SAC's 13 repositories based on the SAC common metadata schema. An example of unified resource representation on the SAC One Search system is shown in Figure 3.



Title

news title

1,800 Karen refugees to return to Burma / Chewin Sattha

Description

news details

ศูนย์มานุษยวิทยาสิรินธร

Princess Maha Chakri Sirindhon Anthropology (SAC)

ประเภทสิ่งพิมพ์	ชื่อบทความ / ชื่อผู้แต่ง
หนังสือพิมพ์ Bangkok Post	
(วันเดือน/ปี) 10 FEB 98 หน้า 1	หัวเรื่อง 1
วารสาร	หัวเรื่อง 2
ปีที่ ฉบับที่ เดือน พ.ศ.	Keyword 1 Ethnic Group - Karen - karen refugee
อื่นๆ (ระบุ)	Keyword 2 Mae Hong Son provice

1,800 Karen refugees to return to Burma

Move seen as bid to bar illegal logging

Chewin Sattha
Mae Hong Son

sioner for Refugees (UNHCR) officials observing the relocation agreed to the plan, reasoning that Ban Mae Lama Luang was where aid agencies can fully give the refugees assistance. Moreover, it is where the military can provide protection.

Collection

anthropology news

Related records

1. Heavy cloud petals: rice plants
2. Ying Louise Chuenmeen, Doctor Tak Duangjew Laen Lod / Dusadee Sothet
3. 'Thrownaway' children face tough life on street
4. Surgery T: Saving the trees of Bangkok / Varalee Sinlarat
5. Fespic games, games that need to be promoted / Asst. Prof. Dr. Witsanu Kamtornthip
6. Smear a hundred million baht for the education of Thai children / Ratchada Tharaphak
7. 'Hong Kong Museum of History' Living Museum
8. Schematic ink drawing Excavation of the archaeological site of Tham Lod Excavation area Area 2 wells S21W10 grid NEQ1, NEQ 3, SEQ1,SEQ3 baulk wells S21W10 N/E 1 well S20W10 grid SEQ3, SEQ4 showing surface of fictitious soil layer 28, fictitious level at 620 cm.dt.

Figure 3. An example of unified resource representation on the SAC One Search system.

5. Evaluation

5.1. Comparison with Existing Socio-Cultural Anthropology Digital Repositories

In order to verify the coverage of the SAC common metadata schema, the metadata elements of the SAC common metadata schema are compared with those of the common metadata schema of two existing socio-cultural anthropology digital repositories: the Smithsonian Learning Lab (https://learninglab.si.edu/search, accessed on 12 April 2022) and the National Institutes for the Humanities, Japan (https://int.nihu.jp/?lang=en&, accessed on 12 April 2022). Both repositories were selected because they have provided access to digital resources in socio-cultural anthropology collections and provided a unified search system in exploring various resource types and collections. The comparison of the SAC common metadata elements and those of the reference systems is shown in Table 4.

Table 4. Comparison of the SAC common metadata elements and those of the reference systems.

SAC Common Metadata Elements	Smithsonian Learning Lab	National Institutes for the Humanities, Japan
Title	Title	Title
Description	Description	-
Creator	Name	Creator
Type	Object Type	-
Properties	Measurements, Dimensions	Format, Language
Provenance	Source	Publisher, Display Original DB Records
Location	Places	-
Time	Dates	Date
Reference	Additional Resource Information	-
Relation		Open Similar Record
Rights	Copyright, Record Information	-
Subject (future work)	Keyword	Subject

The comparison results show that the coverage of the SAC common metadata elements is comparable with that of the reference systems. Specifically, most of the metadata elements of the reference systems can be mapped with the SAC common metadata elements.

However, there is one metadata element of the reference systems that has no equivalence in the SAC common metadata elements, which is "Subject/Keyword". The element is currently planned for future work to support unified subject classification among the SAC digital repositories using the domain ontology approach.

5.2. Evaluation of Search Application

The prototype system was subsequently evaluated by anthropology domain experts and information management experts from SAC. The assessment was carried out on 12 December 2021, based on Bruce and Hillmann's Continuum of Metadata Quality [58], comprising four dimensions: integrity, validity, accessibility, and compliance with expectations (completeness, accuracy, accessibility, and conformance to expectations). Experts were satisfied with the four dimensions of metadata on the highest level (mean above 3.50), which was most in line with expectations (mean = 4.78) (Table 5). Data at some point is the addition of an element of "Provenance", which provides the feature to add new elements for system users and provides the English version of the metadata. The researchers modified the metadata schema in response to discussion with group experts to ensure that quality improvements were made as advised.

Table 5. Result of metadata evaluation.

List	Mean Score	Std. Deviation
Completeness	4.78	0.38
Data cover all the necessary elements of an anthropological database as data objects.	4.70	0.45
The data can comprehensively describe the anthropological database.	4.80	0.35
Data elements can describe all types and formats of anthropological databases.	4.85	0.33
Accuracy	4.77	0.41
The name of each data element is correct and appropriate.	4.90	0.32
The definitions of each data element are clear and accurate.	4.65	0.47
The symbols or abbreviations used in metadata are easy to understand and accurate.	4.75	0.45
Accessibility	4.77	0.38
Using metadata to help find information in anthropology databases to match their needs.	4.85	0.34
Various search options provide access to a wide range of anthropology databases.	4.80	0.36
The search filters are sufficient and useful.	4.65	0.44
Conformance to expectations	4.78	0.44
The effectiveness of the search yields results that are in line with expectations.	4.70	0.46
Metadata can be useful for studies in anthropology.	4.75	0.44
The system is friendly and easy to use.	4.90	0.41
Total	4.78	0.40

6. Conclusions

Data heterogeneity among various digital repositories of a data provider, i.e., data silos, has often led to inconsistency and inefficiency in users' data access. In this paper, a metadata integration framework based on metadata schema mapping is proposed to resolve such a challenge. The framework was designed as a generic framework based on the Metadata Lifecycle Model. Based on the framework, the common metadata schema of Thailand's Princess Maha Chakri Sirindhorn Anthropology Centre (SAC Common Metadata) was developed. The SAC common metadata schema consists of 11 metadata elements designed based on the Dublin Core (DC) and the Europeana Data Model (EDM) metadata elements. The mapping procedure between the source metadata elements from 13 SAC's anthropology digital repositories and the target SAC common metadata schema was described. Metadata integration of the existing digital repositories increases the likelihood of the resources being discovered and accessed via a unified search system.

Finally, the SAC "One Search" system was developed as a prototype search system. It has provided a web-based portal for representing and searching digital resources from different repositories based on the metadata elements of the common metadata schema. The metadata schema mapping profiles have supported the process of metadata transformation from the source repositories into the target integrated repository. An evaluation of the metadata schemas found that they can sufficiently support the description and retrieval of the data by domain experts. The coverage and comparison of the SAC common metadata elements with those of two existing socio-cultural digital repositories are also provided. The implications of this research include (1) the elaboration and description of a metadata integration framework defined based on the Metadata Lifecycle Model (MLM) and (2) the design and adoption of a common metadata schema and metadata schema mapping profiles to support the development of a unified search system for heterogeneous digital repositories in the socio-cultural anthropology domain. Future work includes extending the common metadata schema and mappings to support unified subject classification across digital repositories using the domain ontology approach. Software tools and implementation based on the framework are planned to be released to benefit other digital repositories with similar requirements. One of the limitations of the proposed framework is that it relies on experts in creating metadata schema mapping profiles. Future research should investigate combining a semi-automated mechanism in simplifying experts' mapping tasks.

Author Contributions: Conceptualization, M.B. and W.C.; methodology, M.B. and W.C.; validation, S.R. and P.S.; formal analysis, W.B., S.R. and P.S.; investigation, W.C., S.R. and P.S.; resources, S.R. and P.S.; data curation, T.N.; writing—original draft preparation, M.B., S.R. and W.C.; writing—review and editing, M.B., S.R. and W.C.; visualization, W.B. and T.N.; supervision, M.B. and W.C.; project administration, M.B.; funding acquisition, M.B. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the Program Management Unit for Human Resources & Institutional Development, Research and Innovation-CU (grant number B05F630072).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. SAC. Anthropology database of the Anthropology's Princess Maha Chakri Sirindhorn Anthropology Centre. 2021. Available online: https://db.sac.or.th/search_api/main.php?keyword=&page=&db= (accessed on 15 October 2021).
2. Kim, W.; Seo, J. Classifying schematic and data heterogeneity in multi database systems. *IEEE Comput.* **1991**, *24*, 12–18. [CrossRef]
3. Kashyap, V.; Sheth, A. Semantic heterogeneity in global information systems: The role of metedata, context and ontologies. In *Cooperative Information Systems*; Papazoglou, M.P., Schlageter, G., Eds.; Academic Press: San Diego, CA, USA, 1997; pp. 139–178.
4. Yunianta, A.; Yusof, N.; Bramantoro, A.; Haviluddin, H.; Othman, M.S.; Dengen, N. Data mapping process to handle semantic data problem on student grading system. *Int. J. Adv. Intell. Inform.* **2016**, *2*, 157–166. [CrossRef]
5. Yunianta, A.; Othman, M.S.; Yusof, N.; Yusuf, L.M.; Selamat, N.S. Solving the complexity of heterogeneity data on learning environment using ontology. *Telkomnika* **2015**, *13*, 341–348. [CrossRef]
6. Ekaputra, F.J.; Serrai, E.; Winkler, D.; Biffl, S. A semantic framework for data integration and communication in project consortia. In Proceedings of the 2014 International Conference on Data and Software Engineering (ICODSE), Bandung, Indonesia, 26–27 November 2014; pp. 1–6.
7. Yunianta, A.; Yusof, N.; Othman, M.S.; Aziz, A.; Dengen, N.; Ugiarto, M.; Haeruddin, H.; Angelina, J. Semantic data mapping on E-learning usage index tool to handle heterogeneity of data representation. *J. Teknol.* **2014**, *69*, 1–6. [CrossRef]
8. Yunianta, A.; Yusof, N.; Othman, M.S.; Aziz, A.; Dengen, N. Analysis and Identification of Data Heterogeneity on Learning Environment Using Ontology Knowledge. In Proceedings of the International Conference on Electrical Engineering, Computer Science and Informatics (EECSI 2014), Yogyakarta, Indonesia, 14 January 2014; pp. 154–160.
9. Yunianta, A.; Yusof, N.; Jayadianti, H.; Othman, M.; Suhaimi, S. Ontology Development to Handle Semantic Relationship between Moodle E-learning and Question Bank System. In *Recent Advances on Soft Computing and Data Mining*; Herawan, T., Ghazali, R., Deris, M.M., Eds.; Springer International Publishing: Cham, Switzerland, 2014; Volume 287, pp. 691–701.
10. Sandborn, P.; Terpenny, J.; Rai, R.; Nelson, R.; Zheng, L.; Schafer, C. Knowledge representation and design for managing product obsolescence. In Proceedings of the NSF Civil, Mechanical and Manufacturing Innovation Grantees Conference, Atlanta, GA, USA, 9 December 2011.
11. Nadal, S.; Romero, O.; Abelló, A.; Vassiliadis, P.; Vansummeren, S. An integration-oriented ontology to govern evolution in Big Data ecosystems. *Inf. Syst.* **2019**, *79*, 3–19. [CrossRef]

12. Ma, Z.; Zhao, Z.; Yan, L. Heterogeneous fuzzy XML data integration based on structural and semantic similarities. *Fuzzy Sets Syst.* **2018**, *351*, 64–89. [CrossRef]
13. Blazquez, D.; Domenech, J. Big Data sources and methods for social and economic analyses. *Technol. Forecast. Soc. Change* **2018**, *130*, 99–113. [CrossRef]
14. Benedikt, M.; Grau, B.C.; Kostylev, E.V. Logical foundations of information disclosure in ontology-based data integration. *Artif. Intell.* **2018**, *262*, 52–95. [CrossRef]
15. Munir, K.; Anjum, M.S. The use of ontologies for effective knowledge modelling and information retrieval. *Appl. Comput. Inform.* **2018**, *14*, 116–126. [CrossRef]
16. Zheng, L.; Terpenny, J. A hybrid ontology approach for integration of obsolescence information. *Comput. Ind. Eng.* **2013**, *65*, 485–499. [CrossRef]
17. Kolaitis, P.G. Schema Mappings, Data Exchange, and Metadata Management. In Proceedings of the ACM PODS, Baltimore, MD, USA, 13 June 2005; pp. 61–75.
18. Melnik, S. *Generic Model Management: Concepts and Algorithms*; LNCS 2967; Springer: Berlin, Germany, 2004.
19. Carreira, P.; Galhardas, H. Execution of Data Mappers. In Proceedings of the ACM SIGMOD Workshop IQIS, Paris, France, 18 June 2004; pp. 2–9.
20. Raman, V.; Hellerstein, J.M. Potter'sWheel: An Interactive Data Cleaning System. In Proceedings of the VLDB Conference, Roma, Italy, 11–14 September 2001; pp. 381–390. Available online: http://www.vldb.org/conf/2001/P381.pdf (accessed on 15 October 2021).
21. Lenzerini, M. Data Integration: A Theoretical Perspective. In Proceedings of the ACM PODS, Madison, WI, USA, 3–5 June 2002; pp. 233–246.
22. Doan, A.; Noy, N.; Halevy, A. (Eds.) Special Section on Semantic Integration. *SIGMOD Rec.* **2004**, *33*, 11–70. [CrossRef]
23. Noy, N.F.; Doan, A.; Halevy, A.Y. (Eds.) Special Issue on Semantic Integration. *AI Mag.* **2005**, *26*, 7.
24. Ives, Z.G.; Halevy, A.Y.; Mork, P.; Tatarinov, I. Piazza: Mediation and Integration Infrastructure for Semantic Web Data. *J. Web Sem.* **2004**, *1*, 155–175. [CrossRef]
25. Rahm, E.; Bernstein, P.A. A Survey of Approaches to Automatic Schema Matching. *VLDB J.* **2001**, *10*, 334–350. [CrossRef]
26. Shvaiko, P.; Euzenat, J. A Survey of Schema-Based Matching Approaches. *J. Data Semant. IV* **2005**, *3730*, 146–171. [CrossRef]
27. Miller, R.J.; Haas, L.M.; Hern´andez, M.A. Schema Mapping as Query Discovery. In Proceedings of the VLDB Conference, Cairo, Egypt, 10–14 September 2000; pp. 77–88. Available online: http://www.vldb.org/conf/2000/P077.pdf (accessed on 15 October 2021).
28. Euzenat, J.; Le Bach, T.; Barrasa, J.; Bouquet, P.; De Bo, J.; Dieng, R.; Ehrig, M.; Hauswirth, M.; Jarrar, M.; Lara, R.; et al. *D2.2.3: State of the Art on Ontology Alignment; KWEB/2004/D2.2.3/v1.2*; Technical Report for Knowledge Web Project IST-2004-507482; Knowledge Web Consortium: Singapore, 2004.
29. Melnik, S.; Bernstein, P.A.; Halevy, A.; Rahm, E. Supporting Executable Mappings in Model Management. In Proceedings of the ACM SIGMOD, Baltimore, MD, USA, 14 June 2005.
30. Huang, D.M.; Du, Y.L.; Zhang, M.H.; Zhang, C. Application of ontology-based automatic ETL in marine data integration. In Proceedings of the 2012 IEEE Symposium on Electrical & Electronics Engineering (EEESYM), Kuala Lumpur, Malaysia, 24–27 June 2012; pp. 11–13.
31. Bittner, T.; Donnelly, M.; Winter, S. Ontology and Semantic Interoperability. In *Large-Scale 3D Data Integration: Challenges and Opportunities*; Prosperi, D., Zlatanova, S., Eds.; CRC Press (Tailor & Francis): Boca Raton, FL, USA, 2005; pp. 139–160.
32. Calero, J.M.A.; Pérez, J.M.M.; Bernabé, J.B.; Clemente, F.J.G.; Pérez, G.M.; Skarmeta, A.F.G. Detection of semantic conflicts in ontology and rule-based information systems. *Data Knowl. Eng.* **2010**, *69*, 1117–1137. [CrossRef]
33. Schulz, S.; Martínez-Costa, C. How Ontologies Can Improve Semantic Interoperability in Health Care. In *Process Support and Knowledge Representation in Health Care*; Riaño, D., Lenz, R., Miksch, S., Peleg, M., Reichert, M., Teije, A.T., Eds.; Springer International Publishing: Berlin/Heidelberg, Germany, 2013; Volume 8268, pp. 1–10.
34. Sonsilphong, S.; Arch-int, N. Semantic Interoperability for Data Integration Framework using Semantic Web Services and Rule-based Inference: A case study in healthcare domain. *J. Converg. Inf. Technol.* **2013**, *8*, 150–159.
35. Cyganiak, R.; Bizer, C.; Garbers, J.; Maresch, O.; Becker, C. The D2RQ Mapping Language (v0.8–2012-03-12 ed.). (2012, 2). Available online: http://d2rq.org/d2rq-language (accessed on 15 October 2021).
36. Bizer, C.; Cyganiak, R. D2RQ—Lessons Learned. 2007. Available online: http://www.w3.org/2007/03/RdfRDB/papers/d2rq-positionpaper (accessed on 15 October 2021).
37. Bizer, C.; Seaborne, A. D2RQ-Treating Non-RDF Databases as Virtual RDF Graphs. In Proceedings of the ISWC2004 (Posters), Hiroshima, Japan, 7–11 November 2004. Available online: https://files.ifi.uzh.ch/ddis/iswc_archive/iswc/ab/2004/iswc2004.semanticweb.org/posters/PID-SMCVRKBT-1089637165.pdf (accessed on 15 October 2021).
38. Kienast, R.; Baumgartner, C. Semantic Data Integration on Biomedical Data Using Semantic Web Technologies. In *Bioinformatics-Trends and Methodologies*; Mahdavi, D.M.A., Ed.; InTech: Singapore, 2011.
39. Melik-Merkumians, M.; Zoitl, A.; Moser, T. Ontology-based fault diagnosis for industrial control applications. In Proceedings of the 2010 IEEE Conference on Emerging Technologies and Factory Automation (ETFA), Bilbao, Spain, 13–16 September 2010; pp. 1–4.
40. Kovalenko, O.; Moser, T. Using explicit and machine-understandable engineering knowledge for defect detection in automation systems engineering. In Proceedings of the International Doctoral Symposium on Software Engineering and Advanced Applications (IDoSEAA), Oulu, Finland, 15 March 2011; pp. 1–5.

41. Jirkovsky, V.; Kadera, P.; Obitko, M.; Vrba, P. Diagnostics of distributed intelligent control systems: Reasoning using ontologies and hidden markov models. In Proceedings of the 14th IFAC Symposium on Information Control Problems in Manufacturing (INCOM), Bucharest, Romania, 23–25 May 2012; pp. 1315–1320.
42. Bonifati, A.; Velegrakis, Y. Schema matching and mapping: From usage to evaluation. In Proceedings of the 14th International Conference on Extending Database Technology, Uppsala, Sweden, 21 March 2011; Association for Computing Machinery: New York, NY, USA, 2011; pp. 527–529.
43. Do, H.-H.; Rahm, E. COMA: A system for flexible combination of schema matching approaches. In Proceedings of the 28th International Conference on Very Large Data Bases, Hong Kong SAR, China, 20–23 August 2002; Morgan Kaufmann Publishers: San Francisco, CA, USA, 2002; pp. 610–621.
44. Madhavan, J.; Bernstein, P.A.; Rahm, E. Generic schema matching with cupid. In Proceedings of the 27th International Conference on Very Large Data Bases, 11–14 September 2001; The VLDB Endowment: New York, NY, USA, 2001; pp. 49–58.
45. Doan, A.; Domingos, P.; Halevy, A.Y. Reconciling schemas of disparate data sources: A machine-learning approach. In Proceedings of the 2001 ACM SIGMOD International Conference on Management of Data, New York, NY, USA, 1 May 2001; Association for Computing Machinery: New York, NY, USA, 2001; pp. 509–520.
46. Bernstein, P.A.; Madhavan, J.; Rahm, E. Generic schema matching, ten years later. In Proceedings of the PVLDB, Seattle, WA, USA, 3 June 2011; Volume 4, pp. 695–701.
47. Gal, A. Why is schema matching tough and what can we do about it? *Proc. SIGMOD* **2006**, *35*, 2–5. [CrossRef]
48. Doan, A.; Halevy, A.Y.; Ives, Z.G. *Principles of Data Integration*; Morgan Kaufmann: San Francisco, CA, USA, 2012.
49. Goh, C.H. Representing and Reasoning about Semantic Conflicts in Heterogeneous Information Systems. Doctoral Dissertation, Massachusetts Institute of Technology, Cambridge, MA, USA, 1997.
50. Gruber, T. What is an Ontology? 1993. Available online: http://www-ksl.stanford.edu/kst/what-is-an-ontology.html (accessed on 15 October 2021).
51. Cverdelj-Fogaraši, I.; Sladić, G.; Gostojić, S.; Segedinac, M.; Milosavljević, B. Semantic integration of enterprise information systems using meta-metadata ontology. *Inf. Syst. e-Bus. Manag.* **2017**, *15*, 257–304. [CrossRef]
52. Ballve, D.; Bedini, I.; Breininger, K.; Chiusano, J.; Kacandes, P.; Macias, P.; Mattocks, C.; MacKenzie, M.; Martin, M.; Martell, R.; et al. ebXML Registry Information Model Version 3.0. 2005. Available online: https://www.oasis-open.org/committees/download.php/22323/regrep-rim-3.0.1-cd1.pdf (accessed on 9 April 2022).
53. Meng, R.; Chen, L.; Tong, Y.; Zhang, C. Knowledge base semantic integration using crowdsourcing. *IEEE Trans. Knowl. Data Eng.* **2017**, *29*, 1087–1100. [CrossRef]
54. Chen, Y.-N.; Chen, S.-J.; Lin, S.C. A metadata lifecycle model for digital libraries: Methodology and application for an evidence-based approach to library research. In Proceedings of the 69th IFLA General Conference and Council, Berlin, Germany, 1–9 August 2003.
55. Dublin Core Metadata Initiative. Dublin Core Metadata Element Set, Version 1.1. 2012. Available online: http://dublincore.org/documents/dces/ (accessed on 15 October 2021).
56. Isaac, A. EDM Mappings of Europeana. 2015. Available online: https://pro.europeana.eu/page/edm-profiles (accessed on 15 October 2021).
57. Suleman, H. Introduction to the Open Archives Initiative protocol for metadata harvesting. In Proceedings of the 2nd ACM/IEEE-CS Joint Conference on Digital Libraries, Portland, OR, USA, 14 July 2002; p. 414.
58. Bruce, T.R.; Hillmann, D.I. *The Continuum of Metadata Quality: Defining, Expressing, Exploiting*; ALA Editions: Chicago, IL, USA, 2004.