Schema Resolver as a Solution for Digital Documents Retrieval

Lucky Rachmadeni[a], Detty Purnamasari[b]

[a] lucky.rachmadnei@gmail.com
[a]Magister of Management Information System, Universitas Gunadarma, Depok, 16424, Indonesia
[b]Doctoral Program of Information Technology, Universitas Gunadarma, Depok, 16424, Indonesia

Abstract

Discovering digital documents can sometimes be challenging. A digital document has many attributes to find and has been updated in many versions. In digital government documents, it can become other digital documents in different repositories, so finding the source of digital documents from other digital documents is so uncommon. This paper describes how to find a digital document from other digital document sources from different repositories. Bringing together Xquery and graph theory encourages this research for flexibility and efficiency in developing a tool for the retrieval of documents. Furthermore, to achieve the goal of having many algorithms in this tool, the tool, called the resolver, will develop the knowledge bases for many algorithms that can be used in this tool. Hopefully, the resolver can be used by any algorithm occasionally.

Keywords: Document; Digital; Information Retrieval; Repositories; Xquery.

1. Introduction

Retrieval of government documents digital in Indonesia is a very consolidated task. It has numberless different domain repositories for each librarian system. Moreover, uncovering the document in Indonesia is still carried out with a separate process for each library system. Although there is currently a onesearch application, the search algorithm still follows the application's algorithm. So finding the document digitally using any other algorithm is nearly impossible.

The development of information and communication technology in the digital era is increasing rapidly and influencing the public's need for information. Information becomes a daily commodity for people when doing various activities in particular. This convenience causes information to become more numerous and varied. It formed documents, news, letters, stories, research reports, financial data, simulations, and structured databases (Borgman, 2015). Various types, forms, and media for document storage have also changed to facilitate information storage, management, retrieval, and dissemination. Therefore, it is undeniable that it has become the most significant commodity in today's modern world.

Information has become the most significant commodity, and looking at information becomes massive once done. The information sought is of various forms and types. Seeking information increases not only essential information but also opinion information.

An information retrieval system is a procedure capable of storing, obtaining, and maintaining information (Kowalski, 2000). Information retrieval (IR) represents, searches, and manipulates large collections of electronic text and other human-language data (Büttcher, Clarke, & Cormack, 2016). Furthermore, Information retrieval (IR) is finding material (usually documents) of an unstructured nature (usually text) that satisfies an information need from within large collections (usually stored on computers) (Manning, Raghavan, & Schütze, 2008). In addition, information retrieval deals with the representation, storage, organization of, and access to information items such as documents, Web pages, online catalogues, structured and semi-structured records, and multimedia objects (Baeza-Yates & Ribeiro-Neto, 2011).

Data/documents stored in printed files and in the cupboard are now stored electronically due to technological developments. The entry of the digital transformation era is also a force that forces these changes. Digital Transformation is not solely about digitalizing but rather the fact that digital/electronic technology enables people to solve traditional problems (Caudron J., 2014). Digital Transformation is the highest achievement when the use of digital objects that have been developed allows innovation and creativity in certain domains (Lankshear & Knobel, 2008).

A more specific meaning, one form of digital transformation can be interpreted as a paperless concept to achieve a digital business maturity level that affects individuals and all of society, one of which is the government (Roy, 2006). However, this digital document also leaves various problems that must be solved. Such as searching or retrieval of digital document information. Is because the difficulty of doing good documentation in an organization is as follows (Forcada Matheu, 2005):

- the variety of computer applications used,
- data/documents created by many subsections in the organization,
- data/documents made in different stages according to the sub-division of the maker, and
- searching for data/documents is still demanding because no standardized method exists.

Particularly, arsip's file has a high use value related to various activities. Apart from the value of information, arsip's file also has a high use value related to evidence for an activity's accountability. Arsip's file has two types of files based on categorized based on their function. The first one is Arsip's dinamis. The other one is Arsip's statis. Arsip's dinamis are files used by the creator for the time being and for a while. Arsip's statis are archives produced by archive creators because they have historical value, have exhausted their retention, and have permanent information directly or indirectly verified by the National Archives of the Republic of Indonesia or archival institutions (RI-UU43, 2009). Because of these reasons, much research has been held on developing retrieval information techniques. The development is to find digital documents from each application or domain. So finding the file will be much easier without changing information from each domain.

2. Research Elaborations

The stage of this research consists of four parts: the requirements analysis stage, the designing Resolver stage, the developing Resolver stage, and the trial stage. The stages of this research can be seen in Figure 1 below.

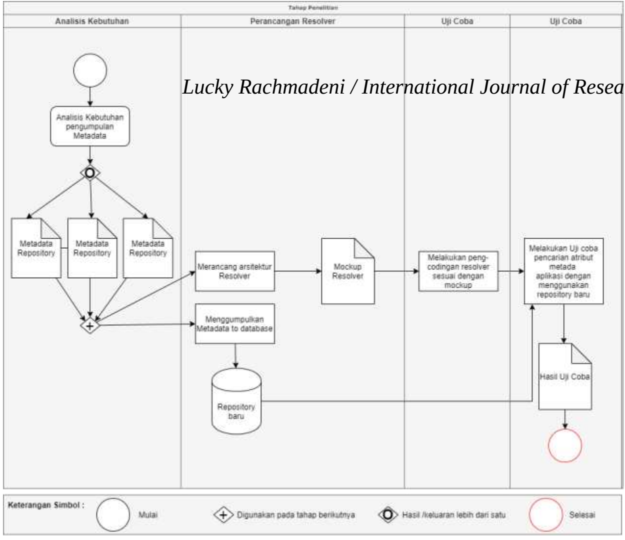

Tahap Penelitian
Analisis Kebutuhan
Perancangan Resolver
Uji Coba
Uji Coba
Analisis Kebutuhan pengumpulan Metadata
Metadata Repository
Metadata Repository
Metadata Repository
Merancang arsitektur Resolver
Mockup Resolver
Melakukan peng-codingan resolver sesuai dengan mockup
Melakukan Uji coba pencarian atribut metada aplikasi dengan menggunakan repository baru
Menggumpulkan Metadata to database
Repository baru
Hasil Uji Coba
Keterangan Simbol :
Mulai
Digunakan pada tahap berikutnya
Hasil /keluaran lebih dari satu
Selesai

Figure 1: The stage of researching digital document retrieval.

In Figure 1, the Resolver was created by analyzing the requirements when making the Resolver. Then, the requirements will determine the metadata used in this research. The metadata selection is based on the domain or repository from which the domain or repository will be obtained from the harvesting results. The harvesting process will be done in each metadata separately. So, this research relies only on the development process of the Resolver. The harvesting metadata will be stored in each domain or repository. The good outcome of that process eventually will be the document and database for this research.

2.1. Composition of requirements analysis from Resolver.

This stage begins by determining the metadata model to be used. The metadata that will be used is grouped into three categories: metadata 1, metadata 2, and metadata 3. The first metadata groups can be seen in Table 1. In Table 1, those groups refer to the metadata function in their respective repository domains. Metadata 1 is used Indonesian terms for the attribute name, such as Tanggal Pembuatan, Pembuat, Jenis, Tingkat Keamanan, and Status. The first metadata is from the initial/draft document.

Table 1: Metadata 1 in indonesia and english term.

Metadata 1 (indonesia)	Metadata 1 (english)
tanggal pembuatan	Created date
Pembuat	Creator
Jenis	Type
Tingkat keamanan	Security Level
Status	Status

An example of the second one of the metadata groups can be seen in Table 2. In Table 2, those groups refer to the metadata function in their respective repository domains. Metadata 2 also uses Indonesian terms for the attribute name, such as KodeSIKN, Sikn, Status, StatusBerkas, TanggalBerkas, KategoriFungsi, NomorBerkas, NomorBerkas, JudulBerkas, BahasaTulisan, MediaArsip, VitalTidakVital, HalJudul, TingkatPerkembangan, JenisNaskah, KlasifikasiAkses, KategoriArsip, and KlasifikasiKeamanan. The second metadata is from an active document (sheet document).

Table2 : Metadata 2 in indonesia and english term

Metadata 1 (indonesia)	Metadata 1 (english)
KodeSIKN	Code SIKN
Sikn	Sikn
Status	Status
StatusBerkas	FileStatus
TanggalBerkas	FileDate
KategoriFungsi	Function Categories
NomorBerkas	File Number
JudulBerkas	File Title
BahasaTulisan	Writing Language
MediaArsip	Archive Media
VitalTidakVital	Vital/NoVital
HalJudul	Title Matter
TingkatPerkembangan	Level of Development
JenisNaskah	Type of Manuscript
KlasifikasiAkses	Access Classification
KategoriArsip	Archive Categories
KlasifikasiKeamanan	Safety Classification

An example of the third one of the metadata groups can be seen in Table 3. In Table 3, those groups refer to the metadata function in their respective repository domains. Metadata 3 is already in English terms. The last is from an archived document. These three metadata have gone through the harvesting and ontology processes.

Table3 : Metadata 3 in english term

Metadata 3	
titleInfo	title
name	namePart
role	typeOfResource
genre	originInfo
Place	placeTerm
Publisher	Language

In the ontology process, the metadata will reveal the similarity of one metadata's attributes to another. The amount of the first metadata attribute has five attributes. In comparison, the second metadata has 43 attributes. The second metadata has the most significant number of attributes because it contains the needed child attributes. However, this will be fine for the searching mechanism from the first metadata to the second metadata and vice versa because the attributes of the first metadata are similar to those of the second. Because of that, the search process can still be processed. The same thing applies to the second metadata with the third metadata, whether the number of attributes in the third metadata is greater than in the second. The quantity of the third attribute is 33 attributes. Similar to the first metadata, the third metadata also has similarities with the attributes of the second metadata. So, the retrieval process can be carried out from the second to the third and vice versa. The similarities between this metadata will be used as a reference in determining the rules, especially the rules engine. This Rules Engine will be a rule of searching documents from the source document to find the destination document.

2.2. The designing Resolver stage.

The resolver design comprises several parts: the interface, query engine, rules engine, and rules selector. The basic design of this resolver can be seen in Figure 2.

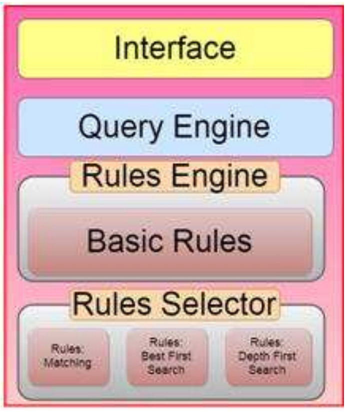


Figure. 2. The basic design of this resolver.

Those parts have different functions. The explanation of each function is as follows,

1. The interface layer is the layer that deals directly with users (people or systems) who use the resolver for conducting searches by using utilizing the contained metadata that generates by itself.
2. The query layer processes queries the user enters to seek the document. Queries are written in semantics that is not only consist of SQL semantics.
3. The rules Engine layer contains the rules used in the searching process in the resolver. The rules of this engine consist of: (i). basic rules that store general rules.
4. The Rules Selector layer determines the rules that will be used to perform the searching mechanism from the algorithm search being put in this layer, such as matching, Bread First Search (BFS), and Depth First Search (DFS).

The resolver application has a query layer useful for processing queries and storing metadata it wants to combine. The metadata combined into the query layer is brought from the process described at the requirements analysis stage. Combining metadata is done separately and goes through the translation process from each metadata to become ontology. The ontology will be translated into basic rules. This research will have three engine rules from three ontologies. These basic rules will then be combined with the search techniques that have been prepared. The search technique used is a graph-search technique. The process can be seen in Figure 3 below.

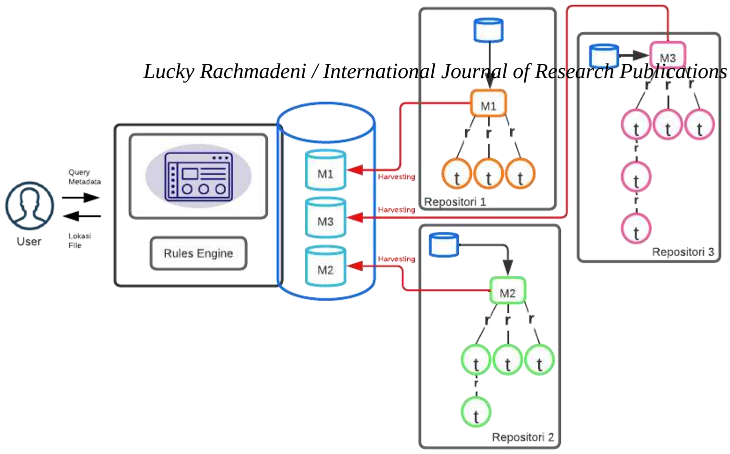


Figure. 3. the association between ontology and rules engine.

Metadata harvested from the source repository will be stored in a database resolver. The resolver database uses Xbasequery, which can be applied to graph searches. The database of each metadata is stored in different databases. There is no merging process in the database, only the metadata collection. After metadata collection, the metadata will be entered into the knowledge base graphs. Accordingly, the resolver has knowledge graphs of each metadata in this process.

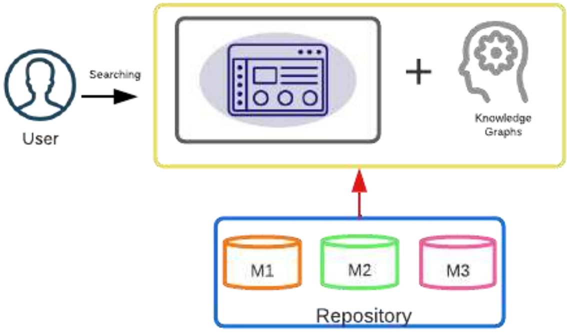
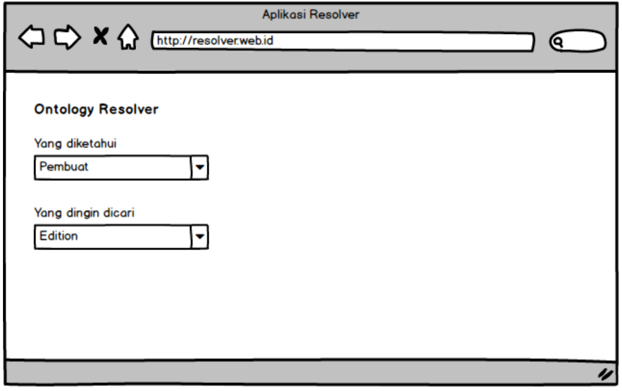


Figure. 4. the association between Knowledge graphs and metadata.

2.3. The developing Resolver stage.

The developing Resolver stage starts at the top layer, namely the interface layer.


Aplikasi Resolver

http://resolver.web.id

Ontology Resolver

Yang diketahui

Pembuat

Yang dingin dicari

Edition

Figure. 5. The view of Resolver.

On this layer, it will be given the option to conduct a retrieval document process that wants to find the relationship from each metadata.

2.4. The Trial stage.

In this stage, the research is to try the node of the mechanism search. It should be the same as the node from the algorithm in The Rules Selector. The result nodes become metadata elements. The trial's steps are as follows:

1. Define the data, which is the query input into the resolver.
2. Run a search based on the query.
3. Inspect the results and compare them with the query.
4. Check the search results' node and compare them with the rules in this test.

The trial's steps are shown in Figure 5 below:

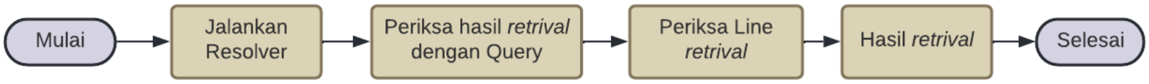


Figure. 6. The trial's steps Method.

3. Results

The metadata is associated with three metadata, namely metadata 1, metadata 2, and metadata 3. This metadata refers to the benefit of the metadata itself. Metadata 1 is an example of an initial/draft document. Metadata 2 is an example of an Arsip's dinamis. Metadata 3 is an example of an Arsip's statis. This metadata will use for determining the rules, especially the rules engine.

The resolver will try to retrieve from the original domain to get digital documents on a different destination domain. Each of these domains has a connection with one another. The relationship of each attribute of this particular metadata is the key to the success of the resolver uncovering digital documents. The map will be generated from the source of the attribute to its destination. It makes digital document retrieval tests from different repository domains become successful.

3.1. Collecting and creating flow of Metadata.

The Metadata was obtained from several sources, including inlislite used by the national library and the Senayan Library Management System (SLiMS). The Metadata obtained is grouped into three types of existing Metadata. The first Metadata is usually used for work documents that will be used or as new work documents from a work unit/part of an institution. Work documents are usually still in the form of draft documents in which ongoing activities will be carried out. The second Metadata is the Metadata of the Archive being used, also known as the active Archive (Archive that is still in process and has yet to be recorded). This second Metadata is usually found in documents currently running or in progress that year. The difference between the first and second Metadata lies in the number of attributes and key attributes of each Metadata. The second Metadata already has other key attributes such as KodeSIKN, creator, manager, and anything else. Whereas in the first Metadata, the attribute does not yet exist. It happened because the first Metadata comes from the draft metadata that has yet to receive the SIKN code and others. At the same time the third is the Metadata of the archives that will be stored/published (the bookkeeping process has been carried out). The third Metadata comes from the second Metadata, which has added more complex attributes. The difference can be seen from the increasing number of derivative attributes.

In the metadata mapping process, firstly, the metadata collection will be carried out on an ontology process, and then it will easily trace the similarities and differences of each Metadata. After the metadata process is successfully ontologized, the ontology will be converted into a more understandable model by the XML language, namely in XSD format. With this process, we will get an XSD file that can be used when running and processing metadata using the XML language.

This resolver uses a unique database. The database used is the BaseX database which can easily recognize XSD files. XSD files from the metadata transformation process can be directly imported into the BaseX database. To perform data processing and manipulate databases and tables, BaseX uses XQuery to query its database. This search process will be better and more optimal for use in the resolver.

3.1.1. Metadata 1.

The ontology of the first metadata has eight class attributes such as DocumentTitle, CreatedDate, Status, SecurityLevel, Type, Format, URI and Creator. The following is shown in Figure 7 below.

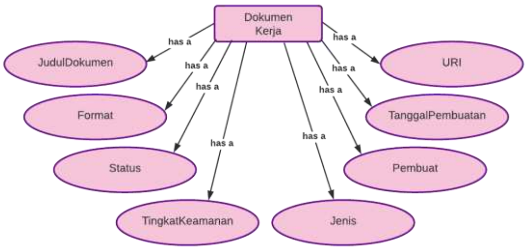


Figure. 7. The ontology of the first metadata.

This ontology is then converted into an XSD template so that it can be translated into XML. After successfully carrying out the transformation process from XML to XSD, the XSD file will be entered into a database. The database used is a database that can map XSD files in graph form. This database is known as the Xquery database. The database is named BaseX.

3.1.2. Metadata 2.

The ontology of the second metadata has 46 attributes, with the main class in this domain being Archive. The Archive has three subclasses: KodeSIKN, File Creator, and Manager.

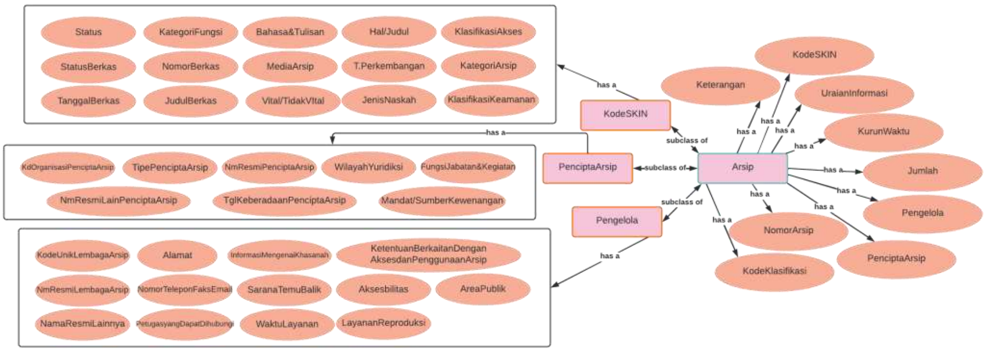


Figure. 8. The ontology of the second metadata.

This ontology is then converted into XSD files so that it can be translated into XML. Then, the XSD files will be entered into a database.

3.1.3. Metadata 3.

The ontology of the third Metadata consists of 34 attributes, following the Metadata in the publications/library section. The following is shown in Figure 9 below.

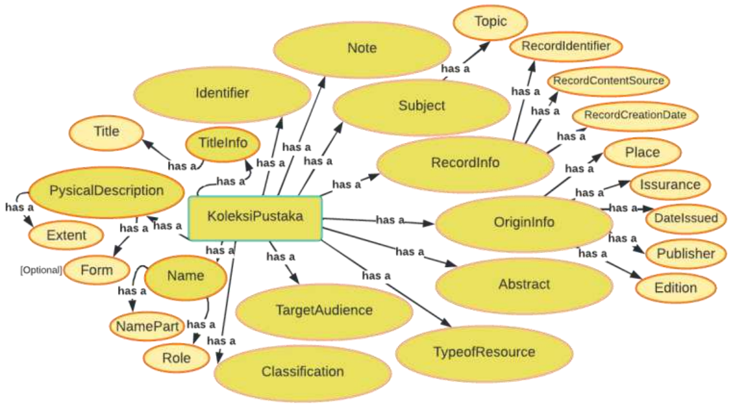


Figure. 9. The ontology of the third metadata.

This ontology is then converted into XSD files so that it can be translated into XML. Then, the XSD files will be entered into a database.

3.2. Mapping Metadata.

After successfully mapping and transforming these metadata, the Metadata will create a connection between each Metadata. The connection that is made comes from the knowledge base owned by each of these metadata. For example, we can map archive documents (metadata 2) with Library Documents (metadata 3). The relationship is to map the file number and Title in the archive document with the Title and Edition in the Library Collection document. So, every search made on the Archive for FileTitle can also be traced to the Library Document for the same Title.

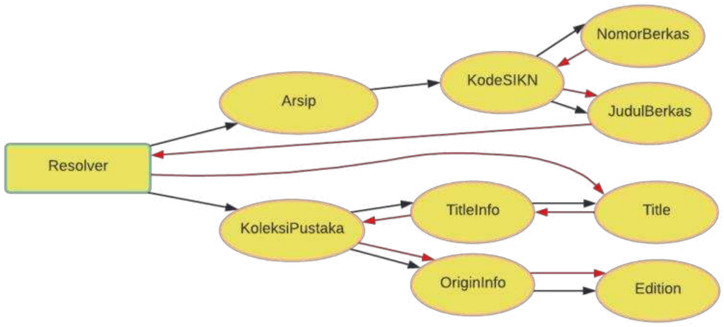
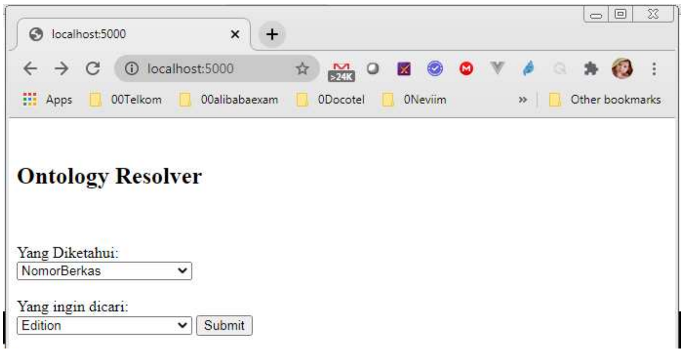


Figure. 10. The association of the second metadata & the third metadata.

3.3. Testing Metadata.

Tests were carried out on the rules generated in this study. Rules will be made based on the mapping that has been obtained. This mapping will later become Basic Rules or Basic Knowledge. An example of visualizing the search results path for the query "Searching for an Edition that is published by FileNumber" in this test is shown in Figure 10 above.

In Figure 11, a search process for the Library Collection Edition will be carried out by only knowing the file number of the data in the Archive Document.



Ontology Resolver

Yang Diketahui:
NomorBerkas

Yang ingin dicari:
Edition

Figure. 11. Example of a search request by query.

4. Results

This resolver tool can connect many metadata from each domain of the library system. It is also equipped with a Rules selector to change the algorithm technique of searching for each repository. Moreover, it can do a more comprehensive search and various search algorithms. Searching-algorithm can append to the rule's selector layer.

The resolver has been used in three diverse library domains and different Searching-algorithm during this study. The library domains are Unit-Kerja, Arsip, and Koleksi-Pustaka. The algorithms are matching algorithm, Best-First-Search, and Depth-First-Search.

In conclusion, this initial study explains how the resolver development process begins, starting from the process of harvesting metadata from each domain, the process of extracting Metadata into an XML database, the process of searching for metadata using the XML database engine, and the process of searching-algorithm by selector rules that can change according to the requirements.

Acknowledgements

The authors are grateful to the Magister of Management Information Systems and Doctoral Program of Information Technology, Universitas Gunadarma.

References

Baeza-Yates, R., & Ribeiro-Neto, B. (2011). Modern Information Retrieval: The Concepts and Technology Behind Search 2nd Edition. Addison-Wesley Professional; 2nd edition.

Borgman, C.L.2015. Big Data, Little Data, No Data: Scholarship in the Networked World.

The MIT Press, Cambridge.

Borgman, C.L., Wallis, J.C., Mayernik, M.S. 2012. Who’s got the data? Interdependencies in science and technology collaborations. Comput. Support. Coop. Work 21(6), 485–523

Büttcher, S., Clarke, C., & Cormack, G. 2016. Information Retrieval: Implementing and Evaluating Search Engines. The MIT Press; Illustrated edition.

Connoly, T., Begg, C., & Strachan, A. (2003). Database Systems : A Practical Approach to Design, Implementation and Management (3rd edition). Addison Wesley.

G. J. Kowalski, 2000. Information storage and retrieval systems: theory and implementation. United States of America.

J. Roy, 2006. E-Government in Canada: Transformation for the Digital Age. University of Ottawa Press.

Manning, C., Raghavan, D., & Schütze, D. (2008). Introduction to Information Retrieval. Cambridge University Press; Illustrated edition.

M. Lankshear, C. and Knobel, 2008. Digital Literacies: Concepts, Policies and Practices. Peter Lang.

N. Forcada Matheu, 2005. Life cycle document management system for construction. Universitat Politècnica de Catalunya.

P. RI-UU43, 2009. tentang kearsipan. Undang-undang No. 43.

P. V. Caudron J, 2014. Digital Transformation: A Model to Master Digital Disruption. BookBaby.