"SEMANTIC" in a Digital Curation Model

Hyewon Lee[1†], Soyoung Yoon[2], Ziyoung Park[2]

[1]Department of Library & Information Science, Seoul Women's University, Seoul, Republic of Korea
[2]Department of Library & Information Science, Hansung University, Seoul, Republic of Korea

Abstract

Purpose: This study attempts to propose an abstract model by gathering concepts that can focus on resource representation and description in a digital curation model and suggest a conceptual model that emphasizes semantic enrichment in a digital curation model.

Design/methodology/approach: This study conducts a literature review to analyze the preceding curation models, DCC CLM, DCC&U, UC3, and DCN.

Findings: The concept of semantic enrichment is expressed in a single word, SEMANTIC in this study. The Semantic Enrichment Model, SEMANTIC has elements, subject, extraction, multi-language, authority, network, thing, identity, and connect.

Research limitations: This study does not reflect the actual information environment because it focuses on the concepts of the representation of digital objects.

Practical implications: This study presents the main considerations for creating and reinforcing the description and representation of digital objects when building and developing digital curation models in specific institutions.

Originality/value: This study summarizes the elements that should be emphasized in the representation of digital objects in terms of information organization.

Keywords Digital curation model; Semantic enrichment; SEMANTIC model; Representation and description of digital objects

1 Introduction

In recent years, records management has focused on digital curation beyond the concept of digital preservation and digital archiving. Digital preservation centers on technological change and emphasized maintaining integrity, digital archiving focuses on an appraisal that selects and preserves resources for use and access. Digital curation attempts to produce new resources and add new value based on existing records. With the development of information and communication

---

† Corresponding author: Hyewon Lee (E-mail: hwlee@swu.ac.kr).

technology and the expansion of virtual space, the burden of determining the value of resources has been reduced. In the future, the concept of using existing resources to produce new resources for business and increase the value of the organization will be emphasized.

Digital curation, broadly interpreted, is about maintaining and adding value to a trusted body of digital information for both current and future use; in other words, it is the active management and appraisal of digital information over its entire life cycle (Pennock, 2007). Digital curation can be applied in a wide range of fields, from humanities to engineering. Not only administrative agencies dealing with public data but also research institutes and general companies will think about how they use the data they own and what value they should be given.

This study attempts to propose a conceptual model that emphasizes semantic enrichment in a digital curation model. It is the extraction of the most important elements from the description and expression of the resource. An abstract model focuses on redesigning the digital curation model to highlight activities that focus on user services, adding value to the core value of digital objects, and supporting reorganization of institutional functions in response to external changes and challenges.

2   Digital curation model analysis

Digital curation models can be largely divided into two groups; one that covers the whole domain and the other that focuses on a specific domain (Lee et al., 2019). The one that covers the whole domain can be further divided into lifecycle-based and continuum-based models. First, Lifecycle-based digital curation models have spread around the United Kingdom as a frame that views the production, distribution, utilization, and preservation of digital information in an organism-like life cycle. A typical example of such models is the Digital Curation Center's Curation Lifecycle Model (DCC CLM). The Lifecycle of Research Knowledge Creation Model is a model that describes the information processing process, which helps to understand the data generation process sequentially and to figure out the linkage between each process (Humphrey & Hamilton, 2004; Humphrey, 2006; Oliver & Harvey, 2016). Data Curation Continuum is the concept of Australia's record continuum applied to digital curation (Oliver & Harvey, 2016). The concept of record continuum was proposed in 1996 based on the structuration theory of A. Giddens (Upward, 2005). Continuum theory has been less well known than life cycle theory because it initially developed a model to use as a teaching tool to differentiate the work undertaken by the different occupations involved in the management of information in the Monash University, Australia (Oliver, 2010).

In this study, we intensively analyzed the DCC CLM based on the life cycle. DCC CLM provides a graphical, high-level overview of the stages required for successful curation and preservation of data from initial conceptualization or receipt through the iterative curation cycle. It is important to note that the model is ideal. In reality, users of the model may enter at any stage of the lifecycle depending on their current area of need (DCC Homepage). The DCC&U Model is a fusion of the UK Digital Curation Center's Curation Lifecycle Model (DCC CLM) and the Digital Curation Unit (DCU) model proposed by the Athens Research Centre in Athens (Constantopoulos et al., 2009). Designed for digital humanity, mainly in the cultural heritage field, the DCU model emphasizes the characteristics of the cultural heritage domain and context management. Moreover, the Athens Research Centre attempted to connect it with the DCC model to expand domain-oriented DCU from a universal perspective. The DCU Digital Curation Process also added the authority management part to the DCC CLM. An authority can reflect all the key concepts, attributes, relationships, and regulations used for a particular domain, making it is useful for knowledge management. The DCU team considered it as an essential part of improving existing knowledge by using annotations, rules, and ontology to link digital information resources themselves as well as real-world objects, situations, and events mentioned in the resources.

The DCC&U model, as shown in Figure 1, added user experience, authority, and semantic web technology to the DCC CLM. "Knowledge Enhancement" was added to the existing curation preservation action and "Authority" to the information technology and expression action. "User Experience" was added between "Access, Use & Reuse" and "Transform" in the subsequent action.

In this study, we attempted to derive user-oriented data value by applying Knowledge Enhancement, the main concept of the DCU and the DCC&U model, based on the DCC CLM.

The digital curation models centered on specific domains were mainly developed by research institutions. In particular, they focused on the diffusion and reuse of their resources.

As a digital curation model considering semantic interoperability, we looked at the University of California Curation Center (UC3) of the California Digital Library and the Data Curation Network (DCN), which was built by several libraries including the University of Minnesota. Within the UC system the UC Curation Center (UC3), one of five programmatic areas of the California Digital Library (CDL), has a broad mandate to ensure the long-term usability of the University's digital assets (CDL, 2010). UC3 proposed a service model capable of independent but interoperable micro-services while having time strategically segmenting complex curation functions (refer to Figure 2). In Figure 2, the actions above "Curate" and "Preserve"
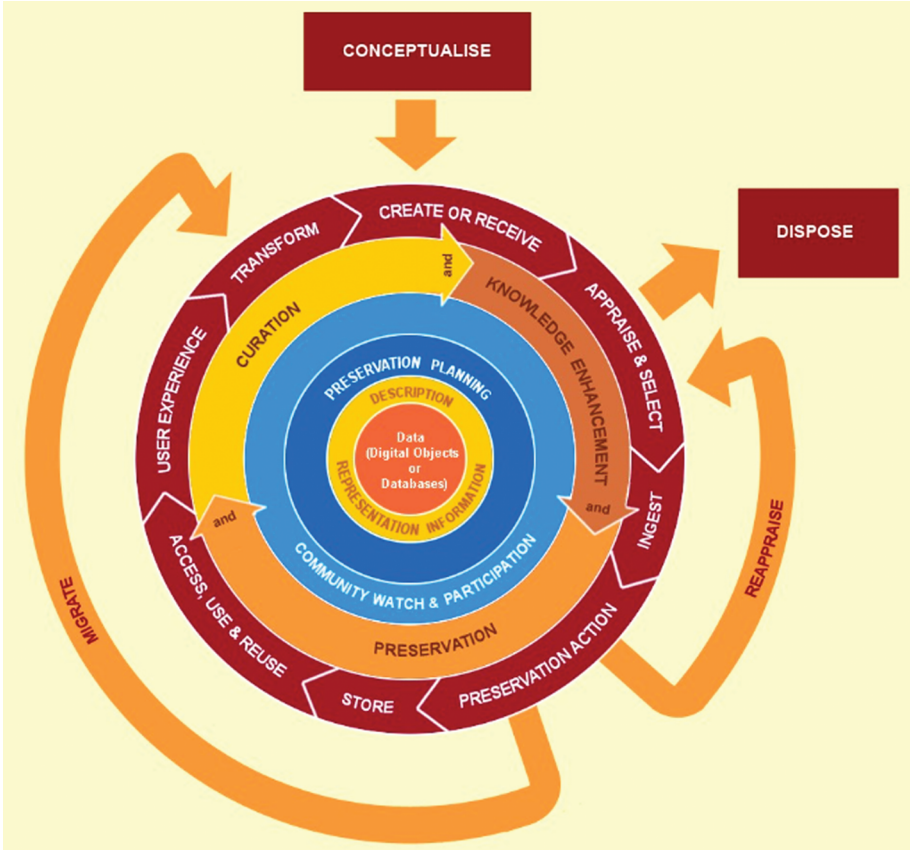
Research Paper

Figure 1.   The DCC&U Curation Lifecycle Model (Constantopoulos et al., 2009, p.42, Fig. 3).

describe services that fit the curation life cycle, and below them is the list of the services that UC3 intends to provide (CDL, 2010).

For the management and value creation of digital contents, which are reliable from a long-term perspective, UC3 micro-service digital curation provided a total
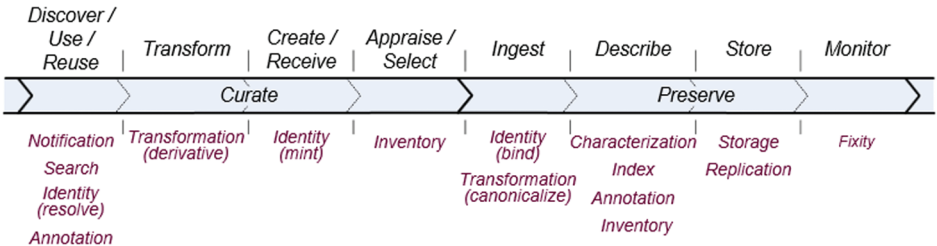
Figure 2.   UC3 Micro Services (from CDL, 2010, p.17, Fig. 7).

of 12 services, such as identity, storage, fixity, replication, catalog, characterization, ingest, index, search, transformation, publication, and annotation (CDL, 2010). In this study, we identified the need to develop a digital curation life cycle model focusing on services through the UC3 curation model.

The emphasis on the "human layer" (Johnston et al., 2017) in the local data repository, which provides expert services, collaboration incentives, standardized curation cases and professional development training for the data curator community, is represented by the DCN model. DCN participating institutions include the University of Michigan, Washington University in St. Louis, the University of Illinois at Urbana Champaign, Cornell University, and Pennsylvania State University. The DCN model is designed to make it easier to find multiple academic datasets; access, interoperability, and reuse them, and further enhance the expertise of the institutions that collectively provide data curation services. The DCN curation workflow based on this is shown in Figure 3 (Johnston et al., 2017). "Augment Metadata" step is also represented semantic augmentation of the data. The step includes metadata enhancement to facilitate discoverability, etc. Data curation enables data discovery and retrieval, maintains data quality, adds value, and provides for re-use over time through activities including authentication, archiving, management, preservation, and representation (Johnston et al., 2017).
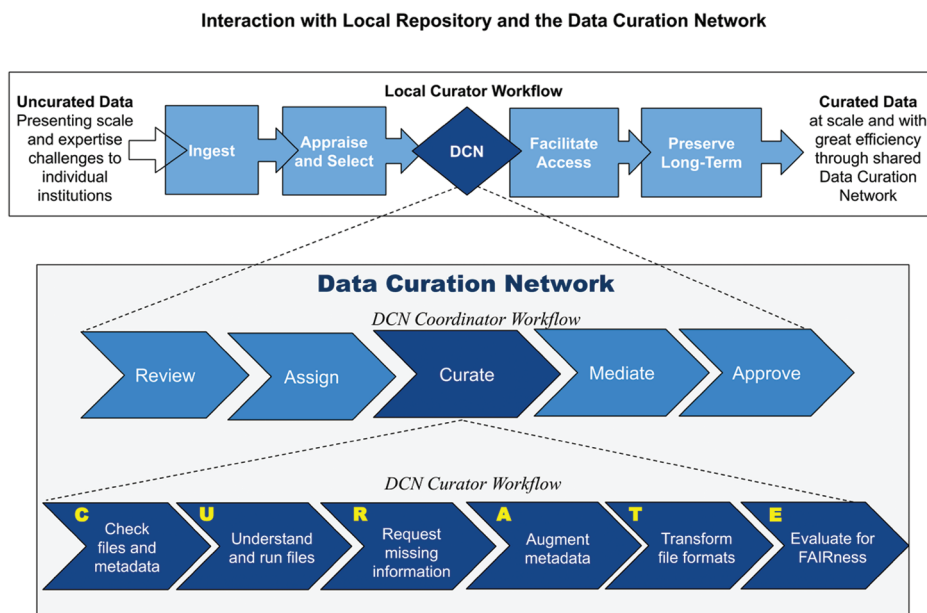

Figure 3.   Curation Workflow for the Data Curation Network (Johnston et al., 2017, p.29, Fig. 8).

This study confirmed the importance of the curation function in analyzing various works within the life cycle and adjusting stakeholders using DCN models.

3   Connection of resources and users through Semantic Enrichment

The digital curation model is highly significant because it links the collected and accessible resources with users. In other words, it focuses on resource utilization. Resources represent the core values of individuals or institutions. These core values of the individuals and institutions are linked to their works, which are the process of performing the functions and roles they pursue. As such, the objects that are needed or used to perform their tasks or the results of performing the tasks are all their resources and exist as data or contents. Meanwhile, users are divided into active users and potential users. An active user is someone who belongs to or needs the resources of an institution. On the other hand, a potential user is someone who will express the need for resources in the future, and he/she may belong to the present but may also be future users. Even for the same resources, the value of the resources can change depending on how users use them. Therefore, a new value is imposed on resources. Hence, users demand various services of resources because they use the resources differently according to the society they belong to or to an information technology environment. They are also interested in techniques for finding information, how to use the information or insight into the resources. Figure 4 describes how resources and users are connected through digital curation. Besides, resources are further enhanced and evolved by users. Conversely, users perform their works through resources and solve the problems by finding the information they need. They also identify evidence of values and information values. Thus, digital curation is building a resources management plan according to their life cycle. Ultimately, it logically demonstrates the process that can constantly create new resources and provide advanced services to users.

Before constructing a life cycle–based digital curation model, this study attempted to derive the important points in the connection of resources and users presented in Figure 4. Meanwhile, Figure 5 shows "Create" and "Ingest" stages for resource acquisition and indicates an "Organize" stage to emphasize the act of entering the meta-information of resources. Lastly, it emphasized a user "Service" to utilize the resources and meta-information.

In Figure 5, the connection between "Create" and "Organize" focused on expressing meta-information together in the resource creation stage. Hence, specific resources are organized at the point of their production. On the contrary, existing resources get new forms and contents through integration among resources, reuse, and meta-information is added. Thus, the connection of "Ingest" and "Organize"
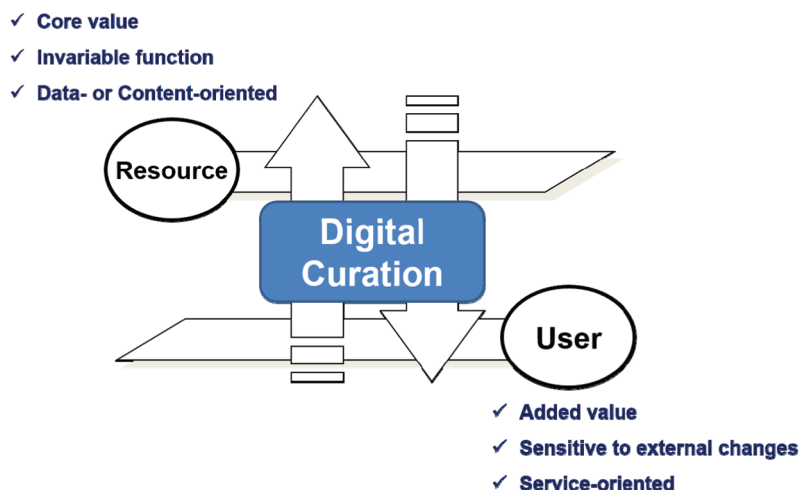
Figure 4.   Connection of resource and user through digital curation (modified by Lee et al., 2019. p.218, Fig. 9).

focused on the granularity of meta-information of resources. Meanwhile, resources may belong to the relevant institution or may be obtained from an external institution, which means that meta-information is needed to connect resources owned by the relevant institution with the external resources that can be accessed. Conversely, metadata, authority and classification information of the institutions should be connected with external data to obtain a more detailed description and semantic information. The connection of "Organize" and "Service" focused on the provision
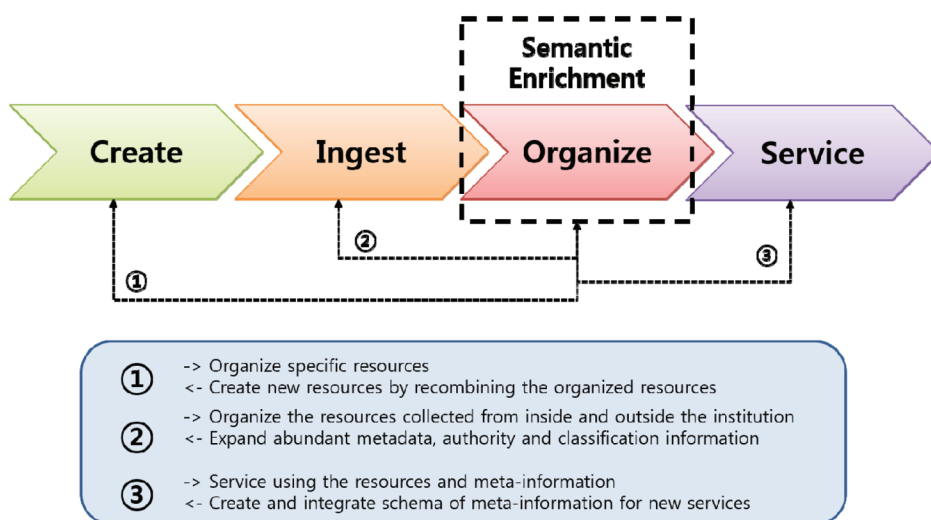

Figure 5.   Semantic Enrichment's functions.

of various services. Therefore, all users, including potential users, can be provided with an identification service, search service, original text service, annotation service, statistical service, and visualization service using meta-information. However, meta-information must be continuously created, integrated, and managed for new services.

Institutions that manage cultural heritage, such as libraries, museums, and archives, as well as data repositories of specific institutions and data centers that independently collect data, focus on using resources that they own and manage. Thus, the primary effort to share and spread resources is to produce and collect them for a quantitative increase in resources. Moreover, sharing and spreading resources require the understanding of different information representation technologies and compliance with the rules that are necessary for the exchange of resources. Meanwhile, the secondary effort is to add meta-information for the qualitative growth of resources. Hence, increasing the granularity of resources also gives various meanings to resources and connect resources and resources. As the resources have more representation, the accuracy of the search becomes higher, and the range of their utilization becomes wider. Besides, the driving force for such quantitative and qualitative growth is to secure meta-information and expand its meaning in consideration of future-oriented services. In Figure 5, it is expressed as semantic enrichment.

4　Semantic Enrichment of digital curation model

The objectives of Semantic Enrichment emphasized in this study are as follows. First, we highlighted the description of the typical curation model and expressed it more precisely. Second, we determined a way to utilize the resources that are produced, managed, and preserved in the life cycle and support user-oriented services. Third and last, we considered the connection with external institutions and other systems as this is crucial in using digital objects and meeting the needs of various users.

Semantic Enrichment in the digital curation model has the following characteristics. First, the concept of DCC CLM, a representative digital curation model, was used to represent the lifecycle-based digital curation model. Second, Semantic Enrichment is one of the full lifecycle actions, and it affects the entire sequential actions. Third, "SEMANTIC" in Semantic Enrichment is a symbolic word that lists the characteristics of the life cycle model. It is a collection of the first letters of the eight terms representing each characteristic. Fourth and last, the order of the letters in SEMANTIC is meaningless. The concept is expressed in a single word, emphasizing the characteristics of the model. The Semantic Enrichment in Digital Curation

Model proposed in this study emphasizes both concepts of conservation and curation and considered future service aspects (refer to Figure 6).
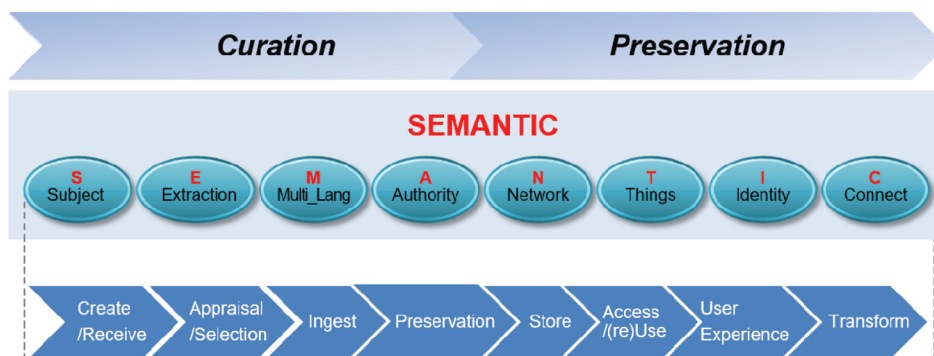

Figure 6.   "SEMANTIC" in a digital curation model.

The first element presented in the Semantic Enrichment Model in digital curation is "Subject," which builds subject authority data and expresses the alternative form and hierarchy structure of digital objects. It can also be used as data to be integrated into a thesaurus or an ontology model. Concepts can be combined or expressed as Subject by sharing the same meaning, using opposite meaning, or connecting with other meanings. "Extraction" means pulling the important attributes of resources, and it supports the process of materializing information that meets the users' needs. For this, it is necessary to comply with resource description standards and express data accordingly. "Multi-Language" constructs a multilingual dictionary and thesaurus, listing and connecting various languages. It identifies the language notations of one concept and suggests a differentiation between terms. "Authority" builds authority data and collects and manages information about names such as persons and organizations. It can also list the characteristics of the alternative forms and explain the hierarchy structure. Authority can be applied not only to the names of persons and organizations but also objects (e.g. books). It is an excellent device for differentiating objects, especially in the Asian region with many homonyms. "Network" structuralizes data connection, enabling connection even in the content unit or data unit. Connections of central ideas are all possible. For example, connections between contents are the "connection of specific R&D report and academic papers that summarized and presented it" and "connection of academic papers and figures and tables included in them." On the other hand, connections between data are "academic papers and their authors" and "figures and the number of downloads." To structuralize the connection of digital objects, identification symbols should be used. Similar to its concept in ontology, "Thing" is used to

Research Paper

describe and express resources, and it includes everything that can be perceived—existent or not. Moreover, "Thing" can be a digital object itself and used as information that explains the object. "Identity" is a string of letters that identify digital objects, which are resources, and it is expressed in special symbols or letters. The identifiers have important implications in a digital curation model. An identification system for digital objects performs the following functions: 1) identify specific digital objects, 2) add a description of the digital objects, and 3) establish linkage with external digital objects. The action manages the name authorities and the subject authorities that can be processed as properties of the digital objects. Besides, it provides terminologies including definitions of entry terms (descriptors) and multilingual expression. "Connect" expresses the relationship between digital objects and information systems. Digital objects can be independently used for the information system. However, they can also be gathered and used as one content.
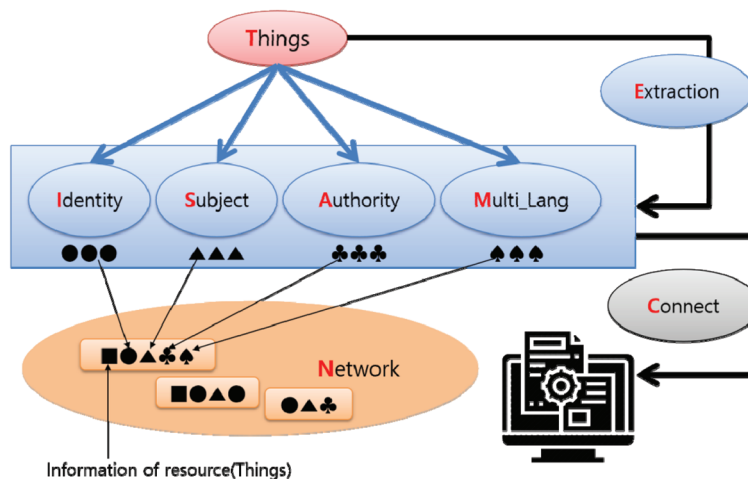

Figure 7.   SEMANTIC elements and their relationships (Lee et al., 2019. p.220, Fig. 11).

"Thing" expresses digital objects, and the attributes of individual objects are "extracted" from "identifier," "subject," "authority," and "variants." These attributes are then utilized in new services through a new combination, which is expressed as "NETWORK" in SEMANTIC. These new services are connected to an integrated system of an Institution or other institutions' systems (refer to Figure 7).

Previous researches emphasized the importance of representation and description of digital objects as the additional or explanatory information and contextual information about the data and knowledge but did not present the specific considerations when building a digital curation model in a real information environment. In this study, the "SEMANTIC" model was proposed by summarizing

the concrete concepts of information representation and description. The core concepts can reflect the needs of various users, derive new values for digital objects, and enhance the integrated perspective of managing digital objects to enable sharing and linking with data from internal to external organizations.

5   Conclusions and discussions

In this study, we proposed the semantic enrichment in the digital curation model to emphasize the description and expression of digital objects. Through the literature review, we examined the preceding curation models such as DCC CLM, DCC&U, UC3, and DCN models, derived the advantages of the models. We ultimately suggested an abstract and conceptual model of semantic enrichment. The concept of semantic enrichment is expressed in a single word, SEMANTIC in this study. SEMANTIC has the following advantages. First, it embraces external changes while maintaining the unique values and functions of data and content. Second, it prepares new services that can accommodate the needs of various users and refines the description and expression of digital objects accordingly. Third and last, it suggested the elements that should be considered important to produce and maintain descriptions and expressions of resources when specific research areas or institutions construct and develop digital curation models.

This study focused on the information expression and description which is one of the stages of the DCC model and lacked practical aspects on how to apply the SEMANTIC model in data management. Further research is needed to identify how the SEMANTIC model has a positive effect in the field where the digital curation model is applied.

Acknowledgment

Author contributions

Hyewon Lee (hwlee@swu.ac.kr) designed the research framework and was responsible for writing and revising the manuscript. Soyoung Yoon (corba99@gmail.com) supported the designing of the research framework and conducted writing and revising the manuscript. Ziyoung Park (zgpark@hansung.ac.kr) investigated and analyzed previous studies as well as writing and revising the manuscript.

Research Paper

References

CDL. (2010). UC3 Curation Foundations. Retrieved from https://confluence.ucop.edu/download/attachments/13860983/UC3-Foundations-latest.pdf

Constantopoulos, P., Dallas, C., Androutsopoulos, I., Angelis, S., Deligiannakis, A., Gavrilis, D., Kotidis, Y., & Papatheodorou, C. (2009). DCC&U: An extended digital curation lifecycle model. International Journal of Digital Curation, 4(1), 34–45. https://doi.org/10.2218/ijdc.v4i1.76

DCC. (2019). Retrieved from http://www.dcc.ac.uk/resources/curation-lifecycle- model

Humphrey, C. (2006). E-Science and the life cycle of research. https://doi.org/10.7939/R3NR4V

Humphrey, C., & Hamilton, E. (2004). Is it working? Assessing the value of the Canadian data liberation initiative. Bottom Line, 17(4), 137–146. https://doi.org/10.1108/08880450410567428

Johnston, L.R., Carlson, J., Hudson-Vitale, C., Imker, H., Kozlowski, W., Olendorf, R., & Stewart, C. (2017). Data curation network: A cross-institutional staffing model for curating research data. Retrieved from https://conservancy.umn.edu/handle/11299/188654

Lee, H., Yoon, S., Park, Z., Hwang, H., Kim, J., & Rhee, H.L. (2019). Developing the research contents life cycle model; Based on the Curation Model for KISTI Curation Center Journal of the Korean Society for Information Management, 36(3), 203–228. http://dx.doi.org/10.3743/KOSOM.2019.36.3.203

Oliver, G. (2010). Transcending silos, developing synergies: Libraries and archives. Information Research: An International Electronic Journal, 15(4). Retrieved from http://www.informationr.net/ir/15-4/colis716.html

Oliver, G., & Harvey, R. (2016). Digital curation. 2nd edition. Chicago: ALA Neal-Schuman, an imprint of the American Library Association.

Pennock, M. (2007). Digital curation: A life-cycle approach to managing and preserving usable digital information. Library and Archives Journal, Issue 1. Retrieved from http://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/lib-arch_curation.pdf