

ISSN 2096-157X

CN 10-1394/G2

# JDIS

数 据 与 情 报 科 学 学 报

# JOURNAL OF DATA AND INFORMATION SCIENCE (QUARTERLY)

Special Issue on Networked Knowledge Organization Systems (NKOS)  
Guest Editors-in-Chief: Joseph Busch, Douglas Tudhope

**I** Volume 5 Number 1      2020

Supported by Chinese Fund for the Humanities and Social Sciences  
and Action Plan for the Excellence of Chinese STM Journals

National Science Library,  
Chinese Academy of Sciences



## CO-EDITORS-IN-CHIEF

**Xiaolin Zhang**

National Science Library, Chinese Academy of Sciences, China

**Ronald Rousseau**

University of Leuven, KU Leuven, Belgium

**Ying Ding**

University of Texas at Austin, USA

## ASSOCIATE-EDITOR-IN-CHIEF

**Liying Yang**

National Science Library, Chinese Academy of Sciences, China

## EDITORIAL BOARD

**Per Ahlgren**

KTH Royal Institute of Technology, Sweden

**Judit Bar-Ilan**

Bar-Ilan University, Israel

**Christine L. Borgman**

University of California, Los Angeles, USA

**Kevin Boyack**

SciTech Strategies Inc., USA

**Chaomei Chen**

Drexel University, USA

**Dar-Zen Chen**

Taiwan University, Taiwan, China

**Ling Chen**

Peking University, China

**Cinzia Daraio**

Sapienza University of Rome, Italy

**Jane Greenberg**

Drexel University, USA

**Haiyan Hou**

Dalian University of Technology, China

**Xiaojun Hu**

Zhejiang University, China

**Tao Jia**

Southwest University, China

**Renaud Lambiotte**

University of Namur, Belgium

**Guangjian Li**

Peking University, China

**Yuelin Li**

Nankai University, China

**Wei Liu**

Library of Shanghai, China

**Wei Lu**

Wuhan University, China

**Xiaobin Lu**

Renmin University, China

**Ed Noyons**

University of Leiden, the Netherlands

**José Miguel B. Nunes**

Sun Yat-Sen University, China

**Han Woo Park**

Yeung Nam University, South Korea

**Qing Qian**

Institute of Medical Information/Medical Library, Chinese Academy of Medical Sciences, China

**Jian Qin**

Syracuse University, USA

**Allen Renear**

University of Illinois at Urbana-Champaign, USA

**Gunnar Sivertsen**

Nordic Institute for Studies in Innovation, Research and Education, Norway

**Min Song**

Yonsei University, South Korea

**Neil Smalheiser**

University of Illinois at Chicago, USA

**Xinning Su**

Nanjing University, China

**Shigeo Sugimoto**

University of Tsukuba, Japan

**Tan Sun**

Agriculture Information Institute, Chinese Academy of Agricultural Sciences, China

**Jie Tang**

Tsinghua University, China

**Li Tang**

Fudan University, China

**Mike Thelwall**

University of Wolverhampton, UK

**Nees-Jan van Eck**

Leiden University, the Netherlands

**Yuefen Wang**

Nanjing University of Science and Technology, China

**Fang Wang**

Nankai University, China

**Jevin West**

Washington University, USA

**Dietmar Wolfram**

University of Wisconsin-Milwaukee, USA

**Dan Wu**

Wuhan University, China

**Yishan Wu**

Chinese Academy of Science and Technology for Development, China

**Feng Xia**

Dalian University of Technology, China

**Erjia Yan**

Drexel University, USA

**Ying Ye (Fred Y. Ye)**

Nanjing University, China

**Jan Youtie**

Georgia Institute of Technology, USA

**Marcia L. Zeng**

Kent State University, USA

**Lin Zhang**

Wuhan University, China

**Zhixiong Zhang**

National Science Library, Chinese Academy of Sciences, China

**Dangzhi Zhao**

University of Alberta, Canada

**Yuxiang Zhao**

Nanjing University of Science and Technology, China

**Donghua Zhu**

Beijing Institute of Technology, China

## MANAGING EDITOR

**Ping Meng**

National Science Library, Chinese Academy of Sciences, China

## EDITORIAL STAFF

**Ping Meng & Nan Zhou**

National Science Library, Chinese Academy of Sciences, China

**Copyright** ©2019. All rights are reserved by the Editorial Office of *Journal of Data and Information Science (JDIS)*, National Science Library, Chinese Academy of Sciences. Address: No 33, Beisihuan Xilu, Zhongguancun, Haidian District, Beijing 100190, P.R. China.

Tel: 86-10-82624454 or 86-10-82626611 ex. 6628. Fax: 86-10-82624454. Email: [jdis@mail.las.ac.cn](mailto:jdis@mail.las.ac.cn). Website: <http://www.jdis.org>.

**Published by:** National Science Library, Chinese Academy of Sciences, No 33, Beisihuan Xilu, Zhongguancun, Haidian District, Beijing 100190, P. R. China

**Edited by:** Editorial Office of *Journal of Data and Information Science (JDIS)*, No 33, Beisihuan Xilu, Zhongguancun, Haidian District, Beijing 100190, P. R. China

**Printed by:** Beijing KEXIN Printing Co. Ltd., Beijing 102208, P.R. China.

**Tel:** 86-10-62903036. Fax: 86-0-62805493

**Co-Editors-in-Chief:** Xiaolin Zhang, Ronald Rousseau & Ying Ding

**Associate-Editor-in-Chief:** Liying Yang

**Typesetting:** Charlesworth (Beijing) Information Services Co Ltd., Room 1105, Building No. 9, Jianwai SOHO, Chaoyang District, Beijing 100022, P.R. China  
Tel/Fax: 86-10-58698392.

**Distributed by:** Editorial Office of *Journal of Data and Information Science (JDIS)*, No 33, Beisihuan Xilu, Zhongguancun, Haidian District, Beijing 100190, P.R. China

**Subscription:** RMB 200/Issue, RMB 800/Volume in China per year; US\$ 199/Volume outside of China (including air shipping).

**Distributional Code:** 82-563

# *JDIS* Special Issue on Networked Knowledge Organization Systems (NKOS)

Joseph Busch<sup>1†</sup>, Douglas Tudhope<sup>2</sup>

<sup>1</sup>Taxonomy Strategies, Washington D.C, USA

<sup>2</sup>University of South Wales, Pontypridd, UK

NKOS<sup>®</sup> is devoted to the discussion of the functional and data model for enabling knowledge organization systems/services (KOS), such as classification systems, thesauri, gazetteers, and ontologies, as networked interactive information services to support the description and retrieval of diverse information resources through the Internet. These tools help to model the underlying semantic structure of a domain for purposes of information retrieval, knowledge discovery, language engineering, and the Semantic Web. NKOS workshops have been held since 1997 in conjunction with related professional and digital library meetings in the U.S., Europe and Asia. The purpose of the workshops is to bring together KOS researchers and practitioners to share work on projects, good practices and innovations, and to discuss and critique this work. Workshops focus on topics including domain modeling, terminology development, validation, automated indexing, annotation and enrichment, and ethics. This *JDIS* special issue includes a selection of papers developed from presentations at the NKOS Workshop held at the Korean National Library in Seoul on September 26, 2019 as part of the International Conference on Dublin Core and Metadata Applications 2019 (DCMI-2019). In the spirit of the NKOS workshops, these papers include research in process, reports on projects, and “thought experiments”.

Jian Qin’s paper (Knowledge Organization and Representation under the AI Lens) on knowledge organization (KO) and knowledge representation (KR) which was the Workshop keynote talk, includes a KO paradigm which provides a good frame for this selection of papers from the Seoul NKOS Workshop. Figure 1 is based on a KO Paradigm presented in Qin’s paper along dimensions that can be visualized as a 2x2 matrix. In this figure we have contextualized the NKOS papers assigning the authors’ names for each paper to a quadrant.

Citation: Busch, Joseph, and Douglas Tudhope. “*JDIS* Special Issue on Networked Knowledge Organization Systems (NKOS).” *Journal of Data and Information Science*, vol.5, no.1, 2020, pp. 1–2. DOI: 10.2478/jdis-2020-0001



<sup>†</sup> Corresponding author: Joseph Busch (E-mail: jbusch@taxonomystrategies.com).

<sup>©</sup> NKOS [website] <https://nkos.slis.kent.edu/>. Last checked 3/3/20.

Editorial

Pragmatism	<b>DDC, UDC</b> <ul style="list-style-type: none"><li>Golub, Hagelbäck, &amp; Ardö <i>Automatic Classification of Swedish Metadata Using Dewey Decimal Classification: A Comparison of Approaches</i></li></ul>	<b>XML and RDF schemas</b> <ul style="list-style-type: none"><li>Park, Lee, Kim, &amp; Park <i>Improving Archival Records and Service of Traditional Korean Performing Arts in a Semantic Web Environment</i></li><li>Lee, Yoon, &amp; Park <i>“SEMANTIC” in a Digital Curation Model</i></li><li>Zeng, &amp; Clunis <i>FAIR + FIT: Guiding Principles and Functional Metrics for Linked Open Data (LOD) KOS Products</i></li></ul>
	<b>Colon Classification, Integrative Level Classification</b> <ul style="list-style-type: none"><li>Park, Gnoli, &amp; Morelli <i>The Second Edition of the Integrative Levels Classification: Evolution of a KOS</i></li></ul>	<b>Ontologies</b> <ul style="list-style-type: none"><li>Lima, Santos, &amp; Rozestraten <i>The ARQUIGRAFIA project: A Web Collaborative Environment for Architecture and Urban Heritage Image</i></li></ul>
Epistemology	<b>Integration</b>	<b>Disintegration</b>

Figure 1. Characterizing NKOS papers by Qin’s KO paradigm.

We thank the authors and reviewers of these papers as well as the *JDIS* editorial staff for their assistance in quickly producing this special issue on NKOS.



This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivs License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).



# Knowledge Organization and Representation under the AI Lens

Jian Qin<sup>†</sup>

School of Information Studies, Syracuse University, Syracuse, USA

## Abstract

**Purpose:** This paper compares the paradigmatic differences between knowledge organization (KO) in library and information science and knowledge representation (KR) in AI to show the convergence in KO and KR methods and applications.

**Methodology:** The literature review and comparative analysis of KO and KR paradigms is the primary method used in this paper.

**Findings:** A key difference between KO and KR lays in the purpose of KO is to organize knowledge into certain structure for standardizing and/or normalizing the vocabulary of concepts and relations, while KR is problem-solving oriented. Differences between KO and KR are discussed based on the goal, methods, and functions.

**Research limitations:** This is only a preliminary research with a case study as proof of concept.

**Practical implications:** The paper articulates on the opportunities in applying KR and other AI methods and techniques to enhance the functions of KO.

**Originality/value:** Ontologies and linked data as the evidence of the convergence of KO and KR paradigms provide theoretical and methodological support to innovate KO in the AI era.

**Keywords** Knowledge representation; Knowledge organization; Artificial Intelligence; Paradigms

## 1 Introduction

Knowledge organization systems (KOS) are developed to represent knowledge in publications and in natural and societal environments and used for information discovery and retrieval. Depending on the purpose, a KOS may be general and broad, such as the Library of Congress Subject Headings (LCSH), which is used to index books and other publications in library collections, while others may be very specific, such as the National Center for Biotechnology Information (NCBI)

Citation: Qin, Jian.

“Knowledge organization and representation under the AI lens.” *Journal of Data and Information Science*, vol.5, no.1, 2020, pp. 3–17.

DOI: 10.2478/jdis-2020-0002

Received: Jan. 10, 2020

Revised: Mar. 15, 2020

Accepted: Mar. 25, 2020



<sup>†</sup> Corresponding author: Jian Qin (E-mail: jqin@syr.edu).

**Research Paper**

Taxonomy that serves as a nomenclature and classification for organisms (NCBI, 2018). Whether general or specific, traditional KOS are not designed for problem-solving purposes, but rather, as standards to normalize vocabularies and classification systems for organizing and retrieving data and information in different systems. As such, KOS aim at a complete, comprehensive coverage of the knowledge universe and emphasize the use of vocabulary to represent concepts and their relations from linguistic and scope perspective. The capabilities of inference or reasoning are not part of the design of these systems, even though relations in hierarchical and associative systems may imply the possibility for inferencing (e.g., the parent-child class relations in a hierarchical classification or the broader and narrower terms in a thesaurus). By contrast, knowledge representation (KR) in artificial intelligence (AI) applications produces a set of statements that express facts, relations, and conditions in formal languages or schemes upon which reasoning can be performed to determine actions or reach conclusions. The reasoning component is perhaps the most striking difference in KR between traditional knowledge organization and artificial intelligence (AI).

Despite differences between traditional knowledge organization (KO) and AI-style KR, the similarities between the two are perhaps more interesting for the KO community for a number of reasons. First, KR in the digital data era is closely tied to language and technology whether it is for KOS or for AI applications. Extracting or generalizing concepts and relations and expressing them in normalized, encoded formats have been extensively studied over the last 50 years by researchers from information science, computer science and other related disciplines. While pioneering work may be traced as far back as Alan Turing's morphogenesis research (Turing, 1952) and the Weinberg panels and report on scientific information process and transfer (United States. President's Science Advisory Committee, 1963), KR research flourished after computers became more efficient and more readily available. These studies generated and established similar techniques and methods (e.g., natural language processing, machine learning, and algorithms for concept detection/extraction) that have been applied in both KO and AI fields, such as graphical representations of knowledge or linked data (KO) and artificial neural networks (AI). Second, although the KO and AI communities as two research fields have largely developed in parallel in the past, this disconnection is being narrowed and the two fields started converging through advances in semantic web technologies and data science. The search for better and more effective ways to address the challenges that come with digital data and culture have prompted each community to look at the other for new ideas and methods. It is not uncommon, for example, to use machine learning algorithms to extract semantic relations or concepts from



social tags (Castano & Varese, 2011; Chen et al., 2008) or classify concepts from research data (Kubat et al., 1993). Finally, there is an increase in KR convergence in the KO and AI communities, both from members' desires to understand the implications of AI for KO and the urge to utilize state-of-the-art techniques and methods within the KO community, as shown from a recent discussion about AI and information science on the community forum of the Association for Information Science and Technology (ASIST) (Toms, 2019).

This paper will first briefly review the historical background of KR in both KO and AI in the last five decades, during which schematic representations of data and information became the main driving force for modernizing knowledge organization and representation. This brief review is by no means to be exhaustive and complete, but rather, it intends to present evidence to demonstrate the fundamental ideas of KR as a background understanding. While the historical background of KO and KR allows us to see and compare KR paradigms between traditional KO and AI, it is important to understand where paradigmatic similarities exist and how the two parallel fields are converging. Following the brief review and analysis, paradigmatic similarities and convergence of KR in KO and AI are discussed and case studies used to demonstrate the convergence trend.

## 2 Knowledge organization (KO) or knowledge representation (KR)?

What is knowledge representation? The answer to this question can take several different directions depending on which perspective one views KR. In the field of library and information science (LIS), the closest term to knowledge representation is knowledge organization, which in its narrow definition means “activities such as document description, indexing and classification performed in libraries, bibliographical databases, archives and other kinds of ‘memory institutions’ by librarians, archivists, information specialists, subject specialists, as well as by computer algorithms and laymen” (Hjørland, 2008). The focus here is placed on organizing and representing documents that embody knowledge. Hjørland (2008) further articulates that knowledge organization in its broader sense is about how knowledge is socially organized and how such knowledge organization systems (KOS) reflect reality. The activities involved in knowledge organization can be divided into two areas: first, knowledge is organized based on humans' understanding of the world in various systems or tools such as classification schemes and thesauri, and second, these knowledge organization systems are applied by humans or machines to represent the document content through a generalized set of terms as the surrogate for the document. The activities of representing document content by using KOS can be deemed as a form of human-mediated knowledge representation due to the fact that the terms or classes are assigned to documents mainly by



librarians or information specialists, although fully automated document representation does exist, such as the indexing service at LexisNexis, and the purpose of such activities is to organize the documents based on topics either on library shelves or in computer systems as catalogs and indexes for information discovery and use.

The deluge of digital data and information we have experienced in the last 30 years and our need to manage it puts traditional knowledge organization in the forefront. However, while “old tricks” in KO still work and are needed, they cannot keep up with the fast growth in the volume and complexity of digital data and information. Ontologies, as a special type of KOS, blend methods of classification and vocabulary control together with codified expressions and reasoning to handle the increasing complexity and volume of digital data and information. This new style of representing and organizing knowledge quickly attracted the attention of the LIS community. Early ontology models that emerged from the LIS community started with reengineering metadata models into ontologies, for example, the ABC ontology that used *Entity* as the root class and *Artifact*, *Event*, *Situation*, *Action*, *Agent*, *Work*, *Manifestation*, *Item*, *Time*, and *Place* as direct subclasses (Lagoze & Hunter, 2001), and the learning object ontology that remodeled the Gateway to Educational Materials (GEM) metadata schema into an ontological model (Qin & Paling, 2001). It should be pointed out here that these ontological models took an entity-centric, object-oriented view of the information world and presented a departure from term-dominant culture. Ontological models have also been developed and deployed to aggregate metadata from multiple sources and in multiple languages, as in the case of Europeana Data Model (Doerr et al., 2010), as well as in linking and opening datasets at cultural institutions to create broader access to art and archival collections, e.g., the well-known CIDOC Conceptual Reference Model (CRM) and the Linked Art Data Model (<https://linked.art/model/index.html>). After more than a decade’s exploration and testing, the library, archive, and museum (LAM) communities have made significant progress in developing ontology theories and practices. Even though whether linked data models and metadata models are a kind of ontologies is still debatable, the influence of ontological thinking in these modeling efforts is apparent.

The knowledge organization tradition in LIS has been summarized by Qin (2002) as falling into two paradigms: integration and disintegration. The integration paradigm has its root in the theory of “integrative levels” (Feibleman, 1954), which views the physical world as cumulative with increased complexity, and classification systems such as Dewey Decimal Classification (DDC) and the International Classification of Disease (ICD) are typical examples. The disintegration paradigm



stands at the opposite: it does not use levels in organizing knowledge, but rather, focuses on the concept and all aspects related to it, which is also called “polyrepresentation” (Ingwersen, 1994). The spectrum from pragmatism to epistemologism represents the approaches utilizing the integration and disintegration paradigms. In the integration paradigm, both pragmatic and epistemological approaches share the same goal of organizing the universe of knowledge rather than solving problems (the left side of Figure 1), be it a hierarchical classification system, a faceted classification, or a system of controlled vocabulary with covert hierarchical relationships. In the disintegration paradigm, attention is given to data and problems in specific domains that require only relevant knowledge segments to solve the problems. As such, the scope of knowledge goes beyond publications to include data and other forms that embody knowledge. It is fair to say that, in the integration paradigm, the goal is to construct knowledge structures or systems to represent the knowledge universe by means of categorization, synthesis, and generalization, while the disintegration paradigm’s goal is to solve problems by using knowledge organization as a means. In this sense, the disintegration paradigm seems to be more aligned with the AI approaches in KR.

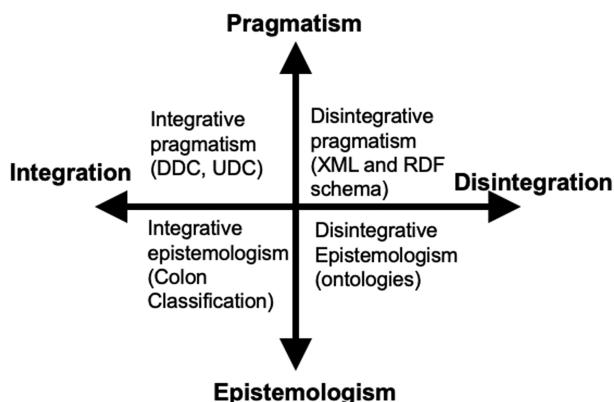


Figure 1. Paradigms in knowledge organization.

Source: Qin, 2002

From a methodological point of view, the integration paradigm is usually in favor of an enumerative (or “just-in-case”) approach, also called “pre-coordination” in KOS construction. The disintegration paradigm tends to represent the concepts in their smallest unit, often in the form of single word or short phrases without any subdivisions. This method allows simple concepts to be coordinated to express complex queries at the time of information retrieval, hence has been called “post-coordination” or “just-in-time” approach. Post-coordination has become the default



method in many information systems for post-search filtering, which is another field of research with a vast body of publications and beyond the scope of this paper.

It is important to point out that KO activities and processes as well as KOS have some properties of KR, but they cannot be equated to KR in AI. Even though ontologies may be considered as the product of KR due to the fact that, in many ways, they resemble KR in artificial intelligence (AI), the objectives, methods through which KR is performed, and the outcomes differ fundamentally. Besides, not all KOS are born equally in purpose, complexity, and function. In this sense, KO is not an interchangeable term for KR. What is knowledge representation then?

### 3 What is knowledge representation?

There are many versions of the definition for knowledge representation in AI field. KR has been defined as “a set of syntactic and semantic conventions that makes it possible to describe things,” in which the syntax refers to “a set of rules for combining symbols so as to form valid expressions,” while semantics are the specification of how such expressions are to be interpreted (Bench-Capon, 1990). A more comprehensive definition is provided by Davies et al. (1993) who specify that a KR is a surrogate, a set of ontological commitments, a fragmentary theory of intelligent reasoning, a medium for efficient computation, and finally, a medium of human expression. Whether simple or comprehensive, these definitions share three core principles. First, knowledge about a domain can be represented systematically “in a sufficiently precise notation that it can be used in, or by, a computer program” and such a systematic representational method can be called a scheme (Hayes, 1974). According to Hayes (1974), such schemes include some programming languages, logical calculi, music notation, or “the systematic use of data structures to depict a world.” The term *scheme* signifies a symbolic paradigm of AI (Bench-Capon, 1990; Hoffmann, 1998).

Another core principle for KR is the formality of representational schemes. The formality of KR refers to the fact that the schemes used for representing knowledge meet the criteria of adequacy and expressiveness. The adequacy criterion relates to things that the representation must have if it is to do what it is required to do. In other words, we need to produce an adequate number of representations of physical objects in the world, and such representations should enable us to both express the facts we wish to express and allow us to perform reasoning by using such representations in problem solving; additionally, these representations should be manipulable by computer systems. Semantically, the representations should be unambiguous, uniform, notationally convenient, relevant, and declarative (Bench-Capon, 1990; Hayes, 1974). The adequacy and expressiveness criteria for



representation naturally lead to a third principle: reasoning or making decisions for solving problems must be based on the facts represented by the schemes. The expert systems that were developed during the late 1980s and 1990s are good examples; for example, the MedIndex system developed at the U.S. National Library of Medicine used the knowledge base frames to guide indexers in completing indexing frames for medical research publications (Humphrey, 1989).

The principles above set the requirements for representing knowledge: they must be sufficiently precise and readable by computer programs to allow for reasoning in problem solving. Over the course of 30 years of KR research, the symbolic and connectionist paradigms have been prevalent in the AI field. While artificial intelligence has its intellectual predecessors from cognitive psychology, mathematics, philosophy of science, and cybernetics, the nature of intelligence and how to develop a formal theory of intelligence became the focus of early scholars in AI (Hoffmann, 1998). This influence is also clear in the study of KR, which has produced some classical works by pioneers such as Patrick J. Hayes, John McCarthy, Brian C. Smith, Ronald J. Brachman, Marvin Minsky, and others (Brachman, 1985). Three paradigms in KR emerged from research: (1) production rules (also called “symbolic paradigm” by Hoffmann (1998)) that are “the representation of knowledge as a set of condition action pairs,” (2) semantic networks (or simply nets) and frames that are rooted in efforts to build systems to understand natural language by structuring objects in graphs (in mathematics) or networks and object-oriented frames, and (3) first-order predicate calculus (also first-order logic) (Bench-Capon, 1990). Hoffmann (1998) named the semantic nets and frames as the “connectionist paradigm.” Whether it is rule-based, object-oriented, or logic-driven, these paradigms fulfill the expressive and adequacy criteria from different approaches and have strengths in different areas (as far as what these different approaches and strengths are, that would need another article to discuss).

#### 4 Similarities between KO and KR paradigms

It is clear that paradigmatic differences exist between KO and KR in terms of the goals, methods, and functions. In general, KO works at the conceptual level and uses language to describe concepts with phrases or terms, while KR focuses on formalizing the expressions in natural language as well as other types of data to enable reasoning as in human intelligence. This seemingly wide gap between the KO and KR paradigms is being bridged by ontologies that have become popular since Berners-Lee et al. (2001) proposed the concept of semantic web. Ontologies by nature are “a formal, explicit specification of a shared conceptualization” (Gruber, 1993). Typically, an ontology defines concepts and specifies relations between them



**Research Paper**

in a formal scheme that can be used for reasoning by computers. According to ISO 25964 Part 2, ontologies defined as such exclude thesauri, classification schemes, and structured vocabularies, even though these are sometimes called “lightweight ontologies” (ISO, 2013). As a specification of conceptualization, ontologies define classes of concepts or entities and relations between classes in a declarative formalism, which is then used to represent a set of objects (or instances). An example is the Gene Ontology (GO) that specifies about ten term elements and four main relations for gene terms from over 600,000 experimentally supported annotations. This central dataset offers “additional inference of over 6 million functional annotations for a diverse set of organisms spanning the tree of life” (Gene Ontology Consortium, 2019). Another example is Schema.org that contains a set of individual ontologies representing creative works, nontext objects, events, health and medical types, organizations, people, and other entities. From both ontologies, one can easily detect the inheritance of paradigms prevailed in the KO and KR communities. Table 1 provides a simplified summary of KO and KR paradigms based on goals, methods, and functions.

Table 1. Similarities in goals, methods, and functions between KO and KR paradigms.

	Paradigm	Goals	Methods	Functions
KO	Integration	Organize the knowledge universe	Categorize, classify, generalize, synthesize	Represent knowledge in publications and organize knowledge about nature and/or society
	Disintegration	Organize the knowledge in a domain	Categorize, classify, generalize, synthesize, model	Represent knowledge in data and publications in a domain
KR	Production rules	Represent knowledge in condition-action pairs to solve problems	Use forward chaining algorithms to execute condition-action pairs	Represent fragmentary knowledge in entity-attribute-value (triple) format
	Semantic networks and frames	Represent semantic relations between concepts	Express semantic relations in triples	Connect knowledge nodes through attributes or slots to form a knowledge graph
	First-order logic	Formalize qualifier construction in natural language	Express declarative propositions using the first-order logic syntax and semantics	Produce a set of axioms for reasoning

The disintegration paradigm in KO as shown in Figure 1 is marked by the adoption of ontology as a methodology in developing knowledge organization systems (KOS). This turned out to be revolutionary for the traditional KO paradigm; it prompted the community to reexamine the KOS structures and explore ways for KOS to fully take advantage of technology advances. One of the most visible efforts



in this area is transforming traditional vocabularies into linked open data to allow for the vocabularies to have some ontology features. It is worth pointing out here that the resulting linked data sets (i.e., transformed controlled vocabularies or restructured bibliographic data) themselves are not ontologies; the model/framework together with the resulting linked data sets have been given the features and functions of ontologies, hence arguably can be considered as ontologies in the broadest sense. Remodeling controlled vocabularies into linked open data has made significant progress, as seen in linked data services offered by U.S. Library of Congress (<http://id.loc.gov/>) for its subject heading list and name authority file, among others, and the Getty Research Institute for its Art and Architecture Thesaurus (AAT), Thesaurus of Geographic Names (TGN), and the Union List of Artist Names (ULAN) (<https://www.getty.edu/research/tools/vocabularies/lod/index.html>).

Whether it is developing ontologies from scratch (as in the cases of Gene Ontology and Schema.org) or remodeling existing vocabularies, it appears that the use of ontologies as a methodology for conceptualizing domain knowledge concentrates on two key features from both KO and KR paradigms:

Formalism in representation schemes: the state-of-the-art encoding languages for ontologies offer a wide range of choices from Web Ontology Language to JSON. These representational languages allow for inference through the creation of axioms.

Structured objects as triples: The use of entity-attribute-value triples, for instance, (Jake age 18), is evidenced in rule-based paradigm as well as the semantic nets and frames. Although semantic nets and frames represent objects in a graph or network and are in a slightly different form from that used in a rule-based paradigm, the base representation is essentially the same triple structure.

Although the goals for KO and KR paradigms vary (Table 1), the similarities lie mainly in methods and functions. This new finding suggests that the KO community may look into the methodology and function similarities further to identify what new opportunities there may be for the KO community to make an impact in AI.

## 5 Challenging issues in KR

The paradigmatic similarities in KR between KO and AI offer not only theory foundations but also practicalities for KO to contribute its unique value for knowledge representation. As mentioned earlier in this paper, KR paradigms in AI follow two distinctive principles—the systematic, sufficiently precise representation schemes that can be processed by computer programs and the adequacy and expressiveness criteria in representing knowledge. Even though KO research has not yet directly discussed these theories and criteria and their implications to KO in the digital era, there have been numerous projects showing the adoption of such principles and criteria. For example, the Simple Knowledge Organization System (SKOS) and the



Research Paper

Semantic Web technology standards suite (Resource Description Framework (RDF), Web Ontology Language (OWL), among others) have been used to convert, and sometimes crosswalk, existing knowledge organization systems into computer processable data structures and offer such structured data (linked data) as knowledge organization services.

While transforming existing KOS into structured data does have great value for open access and reuse of such data, it does not address the challenges in acquiring new knowledge for knowledge organization systems, a well-known bottleneck problem for knowledge representation in AI research. Both KO and KR face the same knowledge acquisition challenges not only because of the complexity of knowledge expressions in texts, data, and multimedia resources, but also due to the fast-changing language and terminologies in modern society and science and technology advances. Whether a logic-based, semantic net, or production-rule-based paradigm is used to represent knowledge, three things must be available for automatic knowledge acquisition: knowledge nodes (k-nodes), relations between the nodes, and rules (which may be labeled differently in different disciplinary fields). To address the bottleneck problem in knowledge acquisition will need an orchestration of natural language processing, machine learning, and clustering and classification techniques to acquire (new) knowledge from texts through clustering and classification (Fisher, 1987).

A recent study by Qin and her team (Qin & Zou, 2017; Qin et al., 2018) is such an attempt to build some foundation work for automatic knowledge acquisition from full-text articles. This study selected a sample of biomedical research papers to identify what types of knowledge nodes and relations may be extracted and expressed in semantic graphs. Table 2 contains example types of k-nodes derived from the sample publications. A further examination of the context of these k-nodes led to the discovery of some relation patterns or types (Table 3).

Table 2. Examples of knowledge nodes derived from the sample publications.

Category	Atomic level (name of things)	Concept level	Cluster level
Gene	Her2, BRCA1, BRCA2, EGFR	Oncogenes	EGFR mutations in lung cancer
Disease	Non-squamous carcinoma, squamous cell carcinoma	Non-small cell lung cancer	Lung cancer
Drug	Pertumzumab, Lmatinib, Crizotinib	Tyrosine kinase inhibitor	Oncogene de-addiction

Source: Qin & Zou, 2017.

These results were obtained by manual analysis of the article texts. A second study was performed to compare differences between manual and automatic detection of knowledge nodes and relations from natural language texts in the domain of biomedical research (Qin et al., 2018). While manual and automatic tools



Table 3. Major relationships types and patterns between knowledge nodes observed in the sample publications.

Relationship	Pattern	Example
has-biomarker	<i>Disease has-biomarker Gene</i>	<i>chronic myeloid leukemia has-biomarker BCR-ABL</i> <i>non-small cell lung cancer has-biomarker EGFR</i>
is-driver-of	<i>Gene is-driver-of Disease</i>	<i>Her2 is-driver-of breast cancer</i> <i>c-Kit is-driver-of chronic granulocytic leukemia</i>
targets	<i>Drug targets Gene</i>	<i>Crizotinib targets ALK</i> <i>Olaparib targets BRCA1/2</i>
has-role-of	<i>Drug has-role-of Treatment</i>	<i>Crizotinib has-role-of oncogene de-addiction</i> <i>Olaparib has-role-of DNA repair</i>

Source: Qin & Zou, 2017.

(MetaMap and SemRep) generated comparable results in identifying knowledge nodes or concepts, the automatic tools either totally missed or did a poor job in detecting relations between k-nodes compared to manually generated relations. In addition, not all relations detected manually had counterparts in the controlled vocabulary Unified Medical Language System (UMLS).

Acquiring and representing relations between k-nodes/concepts remains the most challenging problem in KR. Even powerful machine learning algorithms can fail to detect relations hidden in natural language that are critical for creating machine intelligence. This challenge calls for better algorithms for detecting k-node types and relation patterns for scaling up what human intelligence can achieve in k-node and relation detection. Fortunately, the well-established KOS can save a great deal of ground work for this type of effort. For example, UMLS has already identified a large number of relations as part of its vocabulary. Figure 2 shows part of the search results for the gene BRCA 1 (which causes breast cancer) in UMLS (which provides a long list of relations for this concept and Figure 2 shows only five of them).

Utilizing the data from UMLS, computer programs can be written to transform the data into the format suitable for KR. For example, the relations may be transformed into JSON format, one of the popular ontological encoding schemes:

```
{
  "@context": "http://umls.gov/",
  "@concept": "BRCA1",
  "@semanticTypes": "Gene or Genome",
  "@conceptRelations": {
    "@type": "genetic analysis breast cancer gene (lab test)"
    "@id": "C2010863"
  },
}
```



## Research Paper

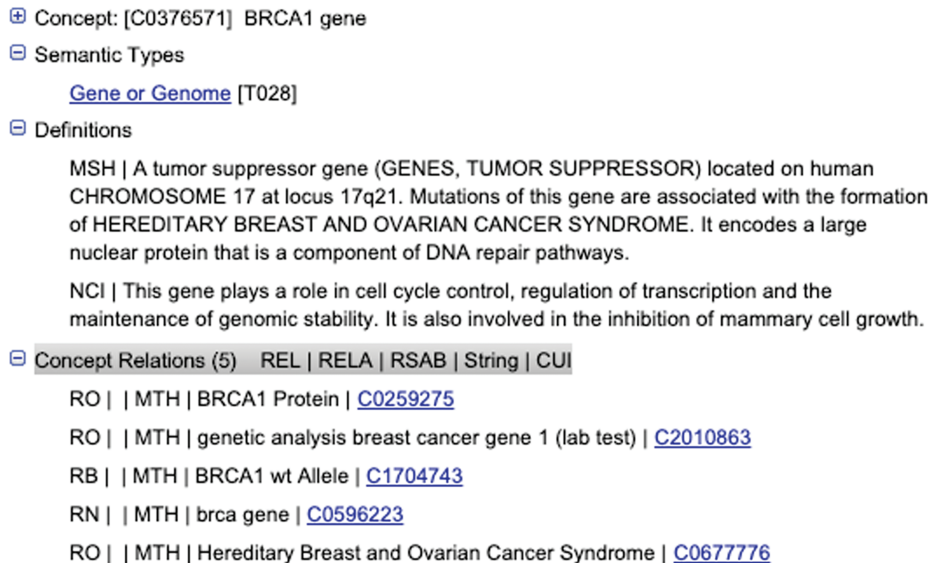


Figure 2. A UMLS example of a concept and relation representation.

This simple example illustrates the potential for well-developed KOS to be restructured and/or remodeled to fit the needs of knowledge representation for artificial intelligence. However, transforming established KOS into KR-feasible structures would not be a trivial task. It is more than likely that new data and information will have to be added in order to make the transformation meaningful. As discussed above, relation detection and representation is still an area requiring further investigation. Manual methods can do a finer job than current tools, but would not be able to scale due to several constraints—the lack of specialized vocabulary sources, rules, and the difficulty in acquiring instances, to name only a few. Semi-automatic indexing, the middle ground between completely manual and completely automatic, has long been practiced to mitigate the problems in manual and automatic methods, e.g., the National Library of Medicine has been using an expert system to assist subject indexing for medical literature (Yang & Chute, 1994). The challenges are two sides: on the one hand, the KO community needs to leverage AI techniques and methods for developing new vocabularies and knowledge organization systems. On the other hand, the vast collections of knowledge organization systems established by years of research and development should also fulfill its value through contributing structured data to the knowledge acquisition and representation solutions.



## 6 Conclusion

The development of knowledge representation in AI provides ample opportunities and new perspectives for the KO community to reexamine the goals, methods, and products from KO activities. It is clear that some of the KO activities have already stepped into the KR domain, such as the linked data services mentioned earlier in this paper. Although the prospects that KOS may contribute to KR can be exciting, there is a need for more research to locate exactly where the trajectory is for KO to converge with KR. The topic of KO and KR paradigms involves much more than this paper has covered. Questions that are important and low-hanging fruit are waiting for exploration, such as How can automatic indexing be deployed by using well-formed KOS given born-digital documents and objects are the new norm today? How can KOS contribute to automatic knowledge acquisition and knowledge base building by using AI methods and techniques? The theory and methodology aspects in such a discussion will benefit both communities by broadening the research horizons in this field.

## References

- Bench-Capon, T.J.M. (1990). *Knowledge representation: An approach to Artificial Intelligence*. London: Academic Press.
- Berners-Lee, T., Handler, J., & Lassila, O. (2001). The Semantic Web. *Scientific American*, Featured article. DOI: 10.1038/scientificamerican0501-34
- Brachman, R.J., & Levesque, H.J. (Ed.). (1985). *Readings in knowledge representation*. Los Altos, CA: Morgan Kauffman Publishers.
- Castano, S., & Varese, G. (2011). Trust-based techniques for collective intelligence through folksonomy coordination. In *Next Generation Data Technologies for Collective Computational Intelligence*, ed. Nik Bessis and Fatos Xhafa, Chapter 4, 87–112. Berlin: Springer-Verlag.
- Chen, M., Liu, X.Z., & Qin, J. (2008). Semantic relation extraction from socially-generated tag: A methodology for metadata extraction. In *Proceedings of the Dublin Core International Conference*, Berlin, Germany. <http://dcpapers.dublincore.org/pubs/article/view/924/920>
- Davis, R., Shrobe, H., & Szolovits, P. (1993). What is a knowledge representation? *AI Magazine*, 14, 17–33. Retrieved from <http://groups.csail.mit.edu/medg/ftp/psz/k-rep.html>
- Doerr, M., Stefan, G., Steffen, H., Antoine, I., Carlo, M., & van de Sompel, H. (2010). The Europeana Data Model (EDM). In *World Library and Information Congress: 76th IFLA General Conference and Assembly 10–15 August 2010, Gothenburg, Sweden*. <https://core.ac.uk/download/pdf/34626222.pdf>
- Feibleman, J.K. (1954). Theory of integrative levels, *The British Journal for the Philosophy of Science*, 5(17), 59–66. <https://doi.org/10.1093/bjps/V.17.59>
- Fisher, D.H. (1987). Knowledge acquisition via incremental conceptual clustering. *Mach Learn* 2, 139–172. <https://doi.org/10.1007/BF00114265>
- Gene Ontology Consortium. (2019). The Gene Ontology and the scientific literature. <http://geneontology.org/docs/literature/>



**Research Paper**

- Gruber, T.R. (1993). A translation approach to portable ontologies. *Knowledge Acquisition*, 5(2), 199–220. <https://doi.org/10.1006/knac.1993.1008>
- Haug, P.J. (1993). Uses of diagnostic expert systems in clinical care. *Proceedings. Symposium on Computer Applications in Medical Care*, 379–383.
- Hayes, P.J. (1974). Some problems and non-problem in representational theory. In *Readings in Knowledge Representation*, ed. Ronald J. Brachman and Hector J. Levesque. Los Altos, CA: Morgan Kaufmann.
- Hjørland, B. (2008). What is knowledge organization (KO)? *Knowledge Organization*, 35(2), 86–101. DOI: 10.5771/0943-7444-2008-2-3-86
- Hoffmann, A. (1998). *Paradigms of Artificial Intelligence: A methodological & computational analysis*. New York: Springer.
- Humphrey, S. (1989). MedIndEx system: medical indexing expert system. *Information Processing & Management*, 25(1), 73–88. [https://doi.org/10.1016/0306-4573\(89\)90092-7](https://doi.org/10.1016/0306-4573(89)90092-7)
- Ingwersen P. (1994) Polyrepresentation of information needs and semantic entities elements of a cognitive theory for information retrieval interaction. In: Croft B.W., van Rijsbergen C.J. (eds) *SIGIR '94*. Springer, London.
- ISO. (2013). *Information and documentation – Thesauri and interoperability with other vocabularies. Part 2: Interoperability with other vocabularies*. Geneva, Switzerland: ISO.
- Kubat, M., Pfurtscheller, G., & Flotzinger, D. (1994). AI-based approach to automatic sleep classification. *Biol. Cybern.* 70, 443–448. <https://doi.org/10.1007/BF00203237>
- Lagoze, C., & Hunter, J. (2001). The ABC ontology and model. *International Conference on Dublin Core and Metadata Applications*, pp. 160–176. Retrieved from <https://dcpapers.dublincore.org/pubs/article/view/655/651>
- NCBI. (2018). *Taxonomy*. <https://www.ncbi.nlm.nih.gov/taxonomy>
- Qin, J., Yu, B., & Wang, L.Y. (2018). Knowledge node and relation detection. In: *Networked Knowledge Organization Systems (NKOS) Workshop at the Dublin Core International Conference DC-2018, Porto, Portugal, September 13, 2018*. <http://ceur-ws.org/Vol-2200/paper3.pdf>
- Qin, J., & Zou, N. (2017). Structures and relations of knowledge nodes: Exploring a knowledge network of disease from precision medicine research publications. *Proceedings of iConference 2017* (pp. 56–65). <https://doi.org/10.9776/17009>
- Qin, J., & Paling, S. (2001). Converting a controlled vocabulary into an ontology: The case of GEM. *Information Research*, 6(2). <http://InformationR.net/ir/6-2/paper94.html> (January).
- Qin, J. (2002). Evolving paradigms of knowledge representation and organization: A comparative study of classification, XML/DTD, and Ontology. *Proceedings of the Seventh International Society for Knowledge Organization Conference*, July 10–12, 2002, Granada, Spain, 465–471. Würzburg, Germany: Ergon. [http://jianqin.metadataetc.org/wp-content/uploads/2019/04/qin\\_isko2002.pdf](http://jianqin.metadataetc.org/wp-content/uploads/2019/04/qin_isko2002.pdf)
- United States. President's Science Advisory Committee. (1963). *The responsibilities of the technical community and the government in the transfer of information: A report of the President's Science Advisory Committee*. <http://garfield.library.upenn.edu/papers/weinbergreport1963.pdf>
- Toms, E. (2019). Artificial intelligence and information science? Discussion thread from ASIST Open Forum. <https://community.asist.org/communities/community-home/digestviewer/viewthread?MessageKey=7cc767bb-dfe3-4ff6-ab5c-959dbe8fcf58&CommunityKey=4eed61bc-fb41-4dd8-9234-fa5fa9f23c20&tab=digestviewer#bm7cc767bb-dfe3-4ff6-ab5c-959dbe8fcf58>



- Turing, A.M. (1952). The chemical basis of morphogenesis. Philosophical Transactions of the Royal Society of London B, 237(641), 37–72.
- Yang, Y., & Chute, C.G. (1994). An application of Expert Network to clinical classification and MEDLINE indexing. Proceedings. Symposium on Computer Applications in Medical Care, 157–161. <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2247915/pdf/procascamc00001-0175.pdf>



This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivs License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).



# Automatic Classification of Swedish Metadata Using Dewey Decimal Classification: A Comparison of Approaches

Koraljka Golub<sup>1†</sup> Johan Hagelbäck<sup>2</sup>, Anders Ardö<sup>3</sup> (emeritus)

<sup>1</sup>Department of Cultural Sciences, Faculty of Arts and Humanities, Linnaeus University, Växjö, Sweden

<sup>2</sup>Department of Computer Science and Media Technology, Faculty of Technology, Linnaeus University, Kalmar, Sweden

<sup>3</sup>Department of Electrical and Information Technology, Lund University, Lund, Sweden

Citation: Golub, Koraljka, Johan Hagelbäck, and Anders Ardö. "Automatic Classification of Swedish Metadata Using Dewey Decimal Classification: A Comparison of Approaches." *Journal of Data and Information Science*, vol.5, no.1, 2020, pp. 18–38.

DOI: 10.2478/jdis-2020-0003

Received: Feb, 4, 2020

Revised: Mar. 20, 2020

Accepted: Mar. 25, 2020

## Abstract

**Purpose:** With more and more digital collections of various information resources becoming available, also increasing is the challenge of assigning subject index terms and classes from quality knowledge organization systems. While the ultimate purpose is to understand the value of automatically produced Dewey Decimal Classification (DDC) classes for Swedish digital collections, the paper aims to evaluate the performance of six machine learning algorithms as well as a string-matching algorithm based on characteristics of DDC.

**Design/methodology/approach:** State-of-the-art machine learning algorithms require at least 1,000 training examples per class. The complete data set at the time of research involved 143,838 records which had to be reduced to top three hierarchical levels of DDC in order to provide sufficient training data (totaling 802 classes in the training and testing sample, out of 14,413 classes at all levels).

**Findings:** Evaluation shows that Support Vector Machine with linear kernel outperforms other machine learning algorithms as well as the string-matching algorithm on average; the string-matching algorithm outperforms machine learning for specific classes when characteristics of DDC are most suitable for the task. Word embeddings combined with different types of neural networks (simple linear network, standard neural network, 1D convolutional neural network, and recurrent neural network) produced worse results than Support Vector Machine, but reach close results, with the benefit of a smaller representation size. Impact of features in machine learning shows that using keywords or combining titles and keywords gives better results than using only titles as input. Stemming only marginally improves the results. Removed stop-words reduced accuracy in most cases, while removing less frequent words increased it marginally. The greatest impact is produced by the number of training examples: 81.90% accuracy on the training set is achieved when at least 1,000 records per class are available in the training set, and 66.13% when too few records (often less than



100 per class) on which to train are available—and these hold only for top 3 hierarchical levels (803 instead of 14,413 classes).

**Research limitations:** Having to reduce the number of hierarchical levels to top three levels of DDC because of the lack of training data for all classes, skews the results so that they work in experimental conditions but barely for end users in operational retrieval systems.

**Practical implications:** In conclusion, for operative information retrieval systems applying purely automatic DDC does not work, either using machine learning (because of the lack of training data for the large number of DDC classes) or using string-matching algorithm (because DDC characteristics perform well for automatic classification only in a small number of classes). Over time, more training examples may become available, and DDC may be enriched with synonyms in order to enhance accuracy of automatic classification which may also benefit information retrieval performance based on DDC. In order for quality information services to reach the objective of highest possible precision and recall, automatic classification should never be implemented on its own; instead, machine-aided indexing that combines the efficiency of automatic suggestions with quality of human decisions at the final stage should be the way for the future.

**Originality/value:** The study explored machine learning on a large classification system of over 14,000 classes which is used in operational information retrieval systems. Due to lack of sufficient training data across the entire set of classes, an approach complementing machine learning, that of string matching, was applied. This combination should be explored further since it provides the potential for real-life applications with large target classification systems.

**Keywords** LIBRIS; Dewey Decimal Classification; Automatic classification; Machine learning; Support Vector Machine; Multinomial Naïve Bayes; Simple linear network; Standard neural network; 1D convolutional neural network; Recurrent neural network; Word embeddings; String matching

## 1 Introduction

Subject searching (searching by topic or theme) is the most common and at the same time the most challenging type of searching in library catalogs and related quality information services, compared to, for example, a known-title or a known-author search. Subject index terms taken from standardized knowledge organization systems (KOS), like classification systems and subject headings systems, provide numerous benefits compared to free-text indexing of commercial search engines: consistency through uniformity in term format and the assignment of terms, provision of semantic relationships among terms, support of browsing by provision of consistent and clear hierarchies (for a detailed overview, see, for example, Lancaster, 2003). However, controlled subject index terms are expensive to produce manually and there is a huge challenge facing library catalogs and digital collections of various types: how to provide high quality subject metadata for increasing numbers



of digital information at reasonable costs. (Semi)-automatic subject classification and indexing represent some potential solutions to retain the established objectives of library information systems.

With the ultimate purpose of establishing the value of automatically produced classes for Swedish digital collections, the paper aims to develop and evaluate automatic subject classification for Swedish textual resources from the Swedish union catalogue (LIBRIS<sup>®</sup>). Based on a data set of 143,756 catalogue records, six machine learning algorithms and one string-matching algorithm were chosen and evaluated.

The paper is structured as follows: Section 2 sets out the rationale for the study and discusses challenges surrounding automatic subject indexing and classification when applied in quality information systems; in Section 3 the data collection, two algorithms and evaluation are described; Section 4 reports on major outcomes; in Section 5 a brief discussion of the impact of the results and implications for operational systems is given.

## 2 Background

Subject searching is a common type of searching in library catalogs (Hunter, 1991; Villén-Rueda et al., 2003) and discovery services (Meadow & Meadow, 2012). However, in comparison to known-item searching (finding an information object whose title, author etc. is known beforehand) searching by subject is much more challenging. This is due to difficulties such as ambiguities of the natural language and poor query formulation, which can be due to lack of knowledge of the subject matter at hand and of information searching. In order to alleviate these problems, library catalogues and related information retrieval systems (could) employ:

1. Hierarchical browsing of classification schemes and other controlled vocabularies with hierarchical structures, which help further the user's understanding of the information need and provide support to formulate the query more accurately;
2. Controlled subject terms from vocabularies such as subject headings systems, thesauri and classification systems, to help the user to, for example, choose a more specific concept to increase precision, a broader concept or related concepts to increase recall, to disambiguate homonyms, or to find which term is best used to name a concept.



The Swedish National Library recently adopted the Dewey Decimal Classification (DDC) to be used as a new national classification system (Svanberg, 2013), replacing SAB (Klassifikationssystem för svenska bibliotek) used earlier since 1921. However, cataloguing with a major classification system, such as DDC, is resource intensive. While fully automatic solutions are not currently feasible, semi-automated solutions can offer considerable benefit, both in assisting the workflow of expert cataloguers and in encouraging wider use of controlled indexing by authors and other users. Although some software vendors and experimental researchers claim to entirely replace manual indexing in certain subject areas (Rotiblat et al., 2010), others recognize the need for both manual (human) and computer-assisted indexing, each with its (dis)advantages (Anderson & Perez-Carballo, 2001; Svarre & Lykke, 2014). Reported examples of operational information systems include BASE by Bielefeld University Library<sup>2</sup>.

NASA's machine-aided indexing which was shown to increase production and improve indexing quality (Silvester, 1997); and the Medical Text Indexer at the US National Library of Medicine, which by 2008 was consulted by indexers in about 40% of indexing throughput (Ruiz, Aronson, & Hlava, 2008).

However, hard evidence on the success of automatic indexing tools in operating information environments, is scarce; research is usually conducted in laboratory conditions, excluding the complexities of real-life systems and situations. It is difficult to compare evaluation results on different systems with widely different datasets. The practical value of automatic indexing tools is largely unknown due to problematic evaluation approaches. Having reviewed a large number of automatic indexing studies, Lancaster concluded that the research comparing automatic versus manual indexing is "seriously flawed" (Lancaster, 2003). One common evaluation approach is testing the quality of retrieval based on the assigned index terms. But retrieval testing is fraught with problems; the results depend on many factors, so retrieval testing cannot isolate the quality of the index terms. Another approach is to measure indexing quality directly. One method of doing so is to compare automatically assigned metadata terms against existing human-assigned terms or classes of the document collection used (as a "gold standard"), but this method also has problems. When indexing, people make errors, such as related to exhaustivity (too many or too few subjects assigned) or specificity (usually because the assigned subject is not the most specific available); they may omit important subjects, or assign an obviously incorrect subject. In addition, it has been reported that different people, whether users or professional subject indexers, assign different subjects to



<sup>2</sup> <https://www.base-search.net/about/en/>

the same document. For a more detailed discussion on these challenges and proposed approach, see Golub et al. (2016).

Research related to automated subject indexing or classification can be divided between three major areas: document clustering, text categorization and document classification (Golub, 2006; Golub, 2017). In document clustering, both clusters (classes) into which documents are classified and, to a limited degree, relationships between them, are produced automatically. Labelling the clusters is a major research problem, with relationships between them, such as those of equivalence, related-term and hierarchical relationships, being even more difficult to automatically derive (Svenonius, 2000). In addition, “Automatically-derived structures often result in heterogeneous criteria for class membership and can be difficult to understand” (Chen & Dumais, 2000). Also, cluster labels, and the relationships between them, change as new documents are added to the collection; unstable class names and relationships are user-unfriendly in information retrieval systems, especially when used for subject browsing. Related to this is keyword indexing whereby topics of a document are identified and represented by words taken from the document itself (also referred to as derived indexing).

Text categorization (machine learning) is often employed for automatic classification of free text. Here characteristics of subject classes, into which documents are to be classified, are learnt from documents with manually assigned classes (a training set). However, the problem of inadequate training sets for the varied and non-uniform hierarchies of the DDC has been recognized. Wang (2009) argues that DDC’s deep and detailed hierarchies can lead to data sparseness and thus skewed distribution in supervised machine learning approaches. Löscher et al. (2011) classified scientific documents to the first three levels of DDC from the Bielefeld Academic Search Engine. They found an “asymmetric distribution of documents across the hierarchical structure of the DDC taxonomy and issues of data sparseness”, leading to a lack of interoperability that was problematic.

In the document classification approach, string matching is conducted between a controlled vocabulary and the text of documents to be classified (Golub, 2006; Golub, 2017). A major advantage of this approach is that it does not require training documents, while still maintaining a pre-defined structure of the controlled vocabulary at hand. If using a well-developed classification scheme, it will also be suitable for subject browsing in information retrieval systems. Apart from improved information retrieval, another motivation to apply controlled vocabularies in automated classification is to re-use the intellectual effort that has gone into creating such a controlled vocabulary. It can be employed with vocabularies containing uneven hierarchies or sparse distribution across a given collection. It lends itself



to a recommender system implementation since the structure of a prominent classification scheme, such as the DDC, will be familiar to trained human indexers.

Automatic document classification based on DDC remains challenging. In early work, OCLC reported on experiments in the Scorpion project to automatically classify DDC's own concept definitions with DDC (Thompson, Shafer, & Vizine-Goetz, 1997). The matching was based on captions. In more recent work, relative index terms from DDC were also incorporated (Khoo et al., 2015); the aim was to investigate automatic generation of DDC subject metadata from English language digital libraries in the UK and USA. The algorithm approximates the practice of a human cataloguer, first identifying candidate DDC hierarchies via the relative index table and then selecting the most appropriate hierarchical context for the main subject. Using a measure called mean reciprocal rank, calculated as 1 divided by the ranked position of the first relevant result, they achieved 0.7 mean reciprocal rank for top 2 levels of DDC and 0.5 for top 3 levels. They considered the results competitive and promising for a recommender system. Golub (2007) and Golub et al. (2007) use a different controlled vocabulary and also report on competitive results.

### 3 Methodology

#### 3.1 Dewey Decimal Classification (DDC)

The DDC was named after its conceiver Melvil Dewey; its first edition was published in 1876. Today the DDC is the most widely used classification system in the world: it has been translated to over 30 languages and is used by libraries in more than 130 countries.

The DDC covers the entire world of knowledge. Basic classes are organized by disciplines or fields of study. At the top level there are 10 main classes each of which is further divided into 10 divisions; each division is further subdivided into 10 sections. As a result, the DDC is hierarchical, and well serves purposes of hierarchical browsing. Each class is represented using a unique combination of Arabic numerals which are the same in all languages, providing the potential for cross lingual integrated search services.

The first digit in the class number represents the main class, the second digit indicates the division, and the third digit the section. For example: 500 stands for sciences, 530 for physics, 532 for fluid mechanics. The third digit in a class number is followed by a decimal point used as a psychological pause since after that the division by 10 continues to a number of other more specific degrees of classification, as needed.



The DDC research permit, the Swedish language version, edition 23, was obtained by the research team from OCLC in 2017. The file received was in MARCXML format<sup>③</sup> comprising over 128 MB. MARCXML is an XML Schema based on MARC (MACHine Readable Cataloguing) format for bibliographic data, derived from the ISO 2709 standard titled “Information and documentation—Format for information exchange” used to exchange electronic records between libraries. For ease of application, relevant data were extracted and re-structured into a MySQL database. The data chosen were the following:

- Class number (field 153, subfield a);
- Heading (field 153, subfield j);
- Relative index term (persons 700, corporates 710, meetings 711, uniform title 730, chronological 748, topical 750, geographic 751; with subfields);
- Notes for disambiguation: class elsewhere and see references (253 with subfields);
- Scope notes on usage for further disambiguation (680 with subfields); and,
- Notes to classes that are not related but mistakenly considered to be so (353 with subfields).

The total of 14,413 unique classes was extracted, of which 819 three-digit classes were found in the LIBRIS data collection described below.

### 3.2 Data collection

The dataset of 143,838 catalogue records was derived from the Swedish National Union Catalogue, LIBRIS, which is the joint catalogue of the Swedish academic and research libraries. It was harvested using the Open Archives Initiative Protocol for Metadata Harvesting (OAIPMH)<sup>④</sup> in the period from 15 April to 21 April 2018. LIBRIS makes its data available in the MARCXML format.

In total 143,838 records with unique id numbers, containing a DDC class (i.e. with MARC field 082<sup>⑤</sup>), were harvested. The records were parsed and all fields and subfields considered relevant were saved in an SQL-database, one field/subfield per row. Relevant fields were the following ones:

- Control number (MARC field 001), unique record identification number;
- Dewey Decimal Classification number (MARC field 082, subfield a);



<sup>③</sup> <https://www.loc.gov/standards/marcxml/>

<sup>④</sup> <https://www.openarchives.org/OAI/openarchivesprotocol.html>

<sup>⑤</sup> For a list and description of MARC fields, see <http://www.loc.gov/marc/bibliographic/>.

- Title statement (MARC field 245, subfield a for main title and subfield b for subtitle); and,
- Keywords (a group of MARC fields starting with 6\*), where available—85.8% of records had at least one keyword.

The records were formatted into an SQL table containing the total of 1,464,046 rows where each row contained 4 columns: ID, field, subfield, and value. The dataset had to be further pruned and cleaned before it could be used for classification experiments. All text features were stripped from special symbols, with the exception of the &-symbol which was replaced by the Swedish word for *and* (*och*), leaving only letters and numbers in the data. For each record, values for title, subtitle and keywords were concatenated into a list of words separated by whitespace, a process known as tokenization.

In the sample, only records containing DDC classes truncated to a three-digit code, ranging from 001 to 999, were used. Records in lower hierarchical levels were reduced to the third-level class to which they belong. Records with other codes as well as those missing both title and subtitle were excluded from the dataset. Duplicates (records with identical title + subtitle) were also removed. This cleaning phase resulted in a total of 143,838 records spread over 816 classes; or, 121,505 records spread over 802 classes when extracting only records which contained at least one keyword.

From the cleaned LIBRIS data a number of datasets were generated, which are presented in Table 1 below. One difficulty with the LIBRIS data is the extreme imbalance between DDC classes, the problem recognized also in previous research (see section 2). The most frequent class is 839 (other Germanic literatures) with 18,909 records, while 594 classes have less than 100 records (70 of those have only one single record). To see how this class imbalance affects classifiers, we have also generated a dataset containing only classes with at least 1,000 records, called major classes below. The latter resulted in 72,937 records spread over 29 classes, and 60,641 records spread over 29 classes when selecting records with keywords.

Table 1. The different datasets generated from the raw LIBRIS data.

Dataset	ID	records	classes
Titles	T	143,838	816
Titles and keywords	T_KW	121,505	802
Keywords only	KW	121,505	802
Titles, major classes	T_MC	72,937	29
Titles and keywords, major classes	T_KW_MC	60,641	29
Keywords only, major classes	KW_MC	60,641	29



### 3.3 Machine learning

Machine learning is the science of getting computers to learn, and improve their learning over time in autonomous fashion, by feeding them data. Instead of explicitly programming a computer what to do, the computer learns what to do by observing the data.

To automatically classify a resource, we need to build models that map input features, i.e. title, subtitle and, optionally, keywords, to a DDC class. These models learn from known, already classified, data (the LIBRIS database) and can later be used to automatically classify new resources. This is referred to as a supervised learning problem; both input features and correct classifications are known.

Machine learning algorithms cannot work with text data directly, so the list of words representing each record in the dataset needs to be encoded as a list of integer or floating point values (referred to as vectorization or feature extraction). The most intuitive way to do so is the “bag of words” representation. The “bag” contains all words that occur at least once in the dataset. A record in the dataset is represented as a vector with the number of occurrences for each word in the title, subtitle and, optionally, keywords. Since the number of distinct words is very high, the vector representing a record is typically very sparse (most values are 0). For the dataset with titles and subtitles, the bag contains a total of 130,666 unique words, and for the dataset with titles, subtitles and keywords, the bag comprises 134,790 unique words. Rare words are later removed, as described below.

When counting occurrences of each single word, all information about relationships between words in the data is lost. This is typically solved using n-grams. An n-gram is a sliding window of size n moving over a list of words, at a pace of one word forward in each step. If a 2-gram is applied, combinations of two words are used as input features instead of, or in combination with, single words (unigrams). For example the text “machine learning algorithm” contains unigrams “machine”, “learning”, “algorithm”, and 2-grams “machine learning” and “learning algorithm”. Using n-grams drastically increases the size of the bag, but can possibly give better classification performance of models. Using unigrams and 2-grams for the datasets with titles, subtitles and keywords as input increases the size of the bag from 134,790 to 828,122 words/word combinations. We have also evaluated higher n-grams (3-grams and 4-grams) but the results did not improve and the computation time of the algorithms increased dramatically.

However, only counting occurrences is problematic: records with longer inputs (title, subtitle and, optionally, keywords) will have higher average count values than records with shorter inputs, even if they belong to the same DDC class. To get around this problem, the number of occurrences for each word is divided by the



total number of words in the record, referred to as Term Frequency (TF). A further improvement is to downscale weights for words that occur in many records and are therefore less informative than words that occur in only a few records. This is referred to as Inverse Document Frequency (IDF). Typically both of these approaches are used, called TF-IDF conversion.

The preprocessing of the text inputs results in high-dimensional, sparse input vectors of either integer values (counting occurrences only) or floating point values (TF-IDF conversion). Many machine learning algorithms are not suited for this type of input data, leaving only a few options left for our task. Historically, good results for different text classification tasks have been achieved with the Multinomial Naïve Bayes (NB) and Support Vector Machine with linear kernel (SVM) algorithms (Aliwy & Ameer, 2017; Trivedi et al. 2015; Wang, 2009). SVM typically gives better results than NB, but is slower to train.

The problem with the bag-of-words approach is that it cannot model any relationships between words. In natural language, some words are related (mango, apple) while others have very different meaning (apple, car). In word embeddings, a numerical representation for words is learned. Each word is represented as a high-dimensional numerical vector, in our case 128 features. The idea with word embeddings is that words that have similar meaning, such as mango and apple, have vectors that are closer to each other (in n-dimensional space) than words with very different meaning. There are pre-learned standard word embeddings that can be used. In our case we used the built-in word embeddings in the Keras machine learning library. Transforming the dataset using the average of all word embeddings results in a set of dense real-valued vectors suitable for neural network algorithms. We have evaluated four different types of neural networks. A simple linear network (Linear), a standard neural network with one hidden layer (NN), a deep neural network using convolutional layers (CNN) (a simplified explanation is that it uses a sliding window over the inputs to reduce the size of the network) and a recurrent neural network (RNN) (can handle relationships between words by allowing recurrent loops in the network, often used for natural language processing tasks).

In addition, of the 143,838 records, 98.6% had one assigned DDC class and 1.4% had more than one assigned class. Because of this, the choice of machine learning algorithms was to apply those producing single output and the 1.4% of records with more than one assigned class were not included in the sample; this also aligned with classification policies of libraries—it is one class per information resources that is typically assigned.

### 3.4 String matching

The approach for string matching is from the Scorpion system by OCLC (Thompson et al. 2017), which implements a ranked retrieval database using terms



and headings from DDC. Introducing text from LIBRIS records as query for such a database produces ranked results that present a list of potential DDC classifications.

As the ranked retrieval database system we used Solr version 8.2.0<sup>®</sup> which is based on the Lucene full-text search engine library<sup>®</sup>. Each document in the database consists of two fields: one with just the DDC class and the second with terms representing that DDC class, extracted as explained in section 3.1 and thus including corresponding DDC heading, relative index term and, if available, any notes. The terms field is just a sequence of concatenated terms ranging in size from 1 word to 168 words depending on the DDC class. These records are indexed in the database using default Solr configuration with a Swedish stop word list consisting of 531 words. Stemming was not used. Some example documents are presented below, with DDC class followed by DDC terms:

- 005.435: Virtuell minne program Minneshantering Program för minneshantering
- 565.38: Decapoda Tiofotade kräftdjur paleozoologi Teuthida Eucarida Lysräkor  
Lysräkor paleozoology
- 760.1: Filosofi och teori
- 781.7101: Allmänna principer för kristen religiös music
- 917.3: geografi Förenta staterna Geografi och resor gällande Förenta staterna

The database field with terms representing a DDC class is queried (relevance-ranked best match using BM25 ranking (Robertson & Zaragoza, 2009)) with queries constructed from title and keywords fields taken from LIBRIS DDC-records. As a result, we get a ranked list of DDC classes.

## 4 Results

### 4.1 Machine learning on Naïve Bayes and Support Vector Machines

Results on machine learning reported in the remainder of the document refer to the top three DDC levels only (due to lack of training examples, as discussed above). Tables 2 and 3 below show classification accuracy (amount of records classified into the correct DDC class divided by the total number of records) of Naïve Bayes (NB) and Support Vector Machines (SVM) algorithms. The columns labeled “Training set” show results when training and evaluating a classifier on all records in the dataset. This gives an indication of how effectively we can map inputs to classes, but does not show the generalization capabilities of the classifiers, i.e. how



<sup>®</sup> <https://lucene.apache.org/solr>

<sup>®</sup> <https://lucene.apache.org/core/>

good they are at classifying records they have not seen before. Therefore, we have also trained the classifiers on 95% randomly selected records from the dataset, and used the remaining 5% of the records for evaluation (shown in the columns labeled “Test set”).

Table 2. Accuracy of the Multinomial Naïve Bayes classifier on the different datasets.

Dataset	Accuracy, unigrams		Accuracy, unigrams + 2-grams	
	Training set	Test set	Training set	Test set
T	83.54%	34.89%	95.82%	34.15%
T_KW	90.01%	55.33%	98.14%	55.45%
KW	75.28%	59.15%	84.95%	58.11%
T_MC	90.83%	54.21%	98.63%	50.51%
T_KW_MC	95.42%	76.52%	99.66%	75.96%
KW_MC	86.94%	77.25%	94.24%	77.09%

Table 3. Accuracy of the Support Vector Machine classifier on the different datasets.

Dataset	Accuracy, unigrams		Accuracy, unigrams + 2-grams	
	Training set	Test set	Training set	Test set
T	93.74%	40.91%	99.59%	40.45%
T_KW	97.50%	65.25%	99.90%	66.13%
KW	83.09%	64.02%	92.38%	64.09%
T_MC	93.95%	57.99%	99.62%	57.80%
T_KW_MC	97.89%	80.75%	99.93%	81.37%
KW_MC	90.58%	79.56%	96.30%	80.38%

The best results were achieved when combining titles and keywords as input. Using only titles as input results in considerably worse accuracy than when combining titles and keywords or using only keywords as input, a difference around 22–23 percentage units for the major classes datasets. The results show that keywords have much higher information value than titles.

As expected, SVM has higher accuracy scores than NB on all datasets. This is in line with previous research on bibliographic data (Trivedi et al., 2015). The best result for SVM when using all classes was 99.90% accuracy on the training set and 66.13% on the test set, when using both unigrams and 2-grams. When removing all classes with less than 1,000 records, the accuracy on the test set increased to 81.37%.

The overall lower accuracy scores on the test sets compared to the training set even when removing classes with few records may be affected by a phenomenon called *indexing consistency*. A number of studies have shown that humans assigning classes or keywords to bibliographic records often do this in an inconsistent manner, both compared to themselves (intra-indexing consistency) and compared to other humans (inter-indexer consistency) (Leiningner, 2000). Since the classifiers learn



from LIBRIS data categorized by humans, this inconsistency may affect their generalization capabilities leading to difficulties when classifying records they have not seen before. The extreme class imbalance also affects the generalization capabilities negatively.

Combining both unigrams and 2-grams only marginally improved the results on the test sets. The highest accuracy was achieved when using SVM and both titles and keywords as input and only major classes. For this dataset the accuracy only increased with 0.62 percentage units when combining unigrams and 2-grams. For NB, the accuracy scores were for most datasets lower than when using unigrams only. We have done some initial testing with higher n-grams (3-grams and 4-grams) but with slightly worse results and significant increases in training time for the algorithms. This needs however to be explored in more detail.

To summarize, using keywords or combining titles and keywords gives much better results than using only titles as input. SVM outperforms NB on all datasets, and the class imbalance where many DDC classes only have few records greatly affects classification performance. Combining unigrams and 2-grams in the input data only marginally improved classification accuracy but leads to much longer training times.

## 4.2 Stop words, stemming and less frequent words

One approach to improve classification accuracy is to pre-process the input data before feeding it to a machine learning algorithm. We have used three different pre-processing techniques: removing stop words, removing less frequent words and stemming. We have also tested combinations of these techniques.

Stemming is the process of reducing words to their base or root form. There are several stemming algorithms that can be used, for example lemmatization or rule-based suffix-stripping algorithms. To investigate how stemming affects accuracy, we have generated two new datasets where the Snowball stemming algorithm for Swedish was used on titles and subtitles<sup>®</sup>. No stemming was used on keywords as they are typically already in base form. We confirmed that this was a good choice by running some tests which showed that accuracy decreased when using stemming on keywords.

We have used a pre-defined list of Swedish stop words from the ranks.nl website<sup>®</sup> to remove stop words from the titles and subtitles.

When converting text data to bag-of-words and TF-IDF conversion, frequency scores of how often each word appears in the whole dataset is calculated. By setting



<sup>®</sup> <http://snowball.tartarus.org/algorithms/swedish/stemmer.html>

<sup>®</sup> <https://www.ranks.nl/stopwords/swedish>

a minimum threshold value, less frequent words are removed from the bag-of-words. We have used a threshold value of 0.00001, effectively reducing the bag-of-words size to one third.

Tables 4 and 5 below show results for the six algorithms when removing stop words (\_sw) and less frequent words (\_rem) in combination with stemming (\_stm). Removing stop words lead to a small decrease in accuracy for SVM (81.37% to 81.24%) and a small increase for NB (75.96% to 76.62%), using 2-grams. When using stemming, a small increase in accuracy was obtained for both NB (75.96% to 76.36%) and SVM (81.37% to 81.80%), using 2-grams. Removing less frequent words lead to an accuracy increase for both NB (75.96% to 78.21%) and SVM (81.37% to 81.83%). The best result for NB was obtained when combining all three approaches leading to an accuracy of 78.90%. The best results for SVM was when combining stemming and removal of less frequent words, leading to an accuracy of 82.20%. This is slightly better than using no pre-processing which resulted in an accuracy of 81.37%.

Table 4. Accuracy of the Naïve Bayes classifier using different pre-processing.

Dataset	Naïve Bayes			
	Accuracy, unigrams		Accuracy, unigrams + 2-grams	
	Training set	Test set	Training set	Test set
T_KW_MC	95.42%	76.52%	99.66%	75.96%
T_KW_MC_rem	90.17%	76.79%	93.25%	78.21%
T_KW_MC_stm	94.32%	76.36%	99.59%	76.36%
T_KW_MC_stm_rem	89.62%	76.26%	92.95%	78.27%
T_KW_MC_sw	95.50%	76.46%	99.64%	76.62%
T_KW_MC_sw_rem	90.28%	77.09%	92.33%	78.60%
T_KW_MC_sw_stm	94.49%	76.59%	99.53%	76.95%
T_KW_MC_sw_stm_rem	89.79%	76.36%	91.96%	78.90%

Table 5. Accuracy of the Support Vector Machine classifier using different pre-processing.

Dataset	Support Vector Machine			
	Accuracy, unigrams		Accuracy, unigrams + 2-grams	
	Training set	Test set	Training set	Test set
T_KW_MC	97.89%	80.75%	99.93%	81.37%
T_KW_MC_rem	92.51%	80.94%	95.02%	81.83%
T_KW_MC_stm	97.21%	81.07%	99.91%	81.80%
T_KW_MC_stm_rem	92.18%	81.34%	94.89%	82.20%
T_KW_MC_sw	95.44%	80.98%	98.48%	81.24%
T_KW_MC_sw_rem	92.46%	81.04%	94.30%	82.13%
T_KW_MC_sw_stm	94.87%	81.40%	98.72%	81.24%
T_KW_MC_sw_stm_rem	92.17%	81.54%	94.16%	81.90%



### 4.3 Word embeddings

Tables 6 and 7 below show accuracy metrics for word embeddings combined with four different types of neural networks: Simple linear network (Linear), Standard neural network (NN), 1D convolutional neural network (CNN) and Recurrent neural network (RNN).

Table 6. Accuracy of NN and CNN classifiers using word embeddings.

Dataset	NN		CNN	
	Training set	Test set	Training set	Test set
T_KW_MC	96.19%	79.40%	95.33%	79.92%
KW_MC	90.54%	78.23%	90.39%	79.15%
T_KW_MC_stm	95.92%	79.57%	94.60%	80.38%

Table 7. Accuracy of Linear and RNN classifiers using word embeddings.

Dataset	Linear		RNN	
	Training set	Test set	Training set	Test set
T_KW_MC	97.17%	79.99%	92.76%	78.70%
KW_MC	91.30%	78.41%	88.03%	78.74%
T_KW_MC_stm	96.90%	80.81%	92.38%	79.16%

The results show that all the four algorithms perform worse than SVM, but very close—in best example, Simple linear network yields 80.8% compared to 82.2% of best SVM, for main classes and with stemming applied. Like in the case of SVM and NB, stemming slightly improves accuracy. An advantage of word embeddings is having a smaller representation size (then the stored data takes less space); and since differences are not large, these approaches may work sufficiently well when working with large data sets. We have not tried any of the pre-processing techniques (stop words removal, stemming or removal of less frequent words) in combination with word embeddings.

### 4.4 Machine learning and training examples

A problem when using machine learning algorithms on the dataset is the huge number of possible classes (802 when using three digits). Considering that DDC is hierarchical, one approach to increase the performance of the machine learning models could be a hierarchical set of classifiers. A hierarchical classifier first determines values of most specific classes (lowest in the hierarchy) and these outputs are then combined for classification results at a higher level; the process is iterated till top levels. Using two digits instead of three (99 classes instead of 802) increased the accuracy when using all examples from 58.10% to 73.30%, see Tables 8 and 9 below. This indicates that a hierarchical approach could work, but it



needs more investigation as models must be trained and evaluated for each of the ten subsets.

Table 8. Accuracy of the Naïve Bayes classifier when using two digits.

Dataset	Naïve Bayes			
	Accuracy, unigrams		Accuracy, unigrams + 2-grams	
	Training set	Test set	Training set	Test set
T_KW_stm_2d	87.40%	65.64%	93.18%	67.79%
T_KW_2d	88.26%	64.78%	93.55%	66.92%
KW_2d	78.36%	68.12%	82.53%	67.94%

Table 9. Accuracy of the Support Vector Machine classifier when using two digits.

Dataset	Support Vector Machine			
	Accuracy, unigrams		Accuracy, unigrams + 2-grams	
	Training set	Test set	Training set	Test set
T_KW_stm_2d	90.60%	72.68%	96.23%	73.32%
T_KW_2d	91.21%	72.14%	95.48%	73.24%
KW_2d	81.75%	71.86%	86.18%	71.96%

## 4.5 String matching results

First we generated the query by using title and keyword fields; we also tried using a query generated from title, keyword and summary fields in LIBRIS records, but the results were almost identical since only 11,000 records had any summary.

For each query (LIBRIS record) against the database (DDC terms) we analyzed both the best hit and the top three ranked hits for each query. When looking at top-ranked 1000 DDC classes, the results were below 50% correct: in 11.8% of cases were top classes identified accurately and in 32.7% cases was the accurate class found among top 3 results. Reducing the number of classes to top-ranked 100, the best we could achieve was 43.7% correct among top 3 hits; top ranking accuracy increased to 15.4%.

## 4.6 Evaluation based on specific classes

Looking at accuracy of specific classes in all the classes (802), machine learning algorithms in general do worst on class 3xx (social sciences, sociology & anthropology). It often happens that other classes are misclassified as belonging to this class, and what should be 3xx documents, often get misclassified as other classes. Most misclassifications take place between 3xx and 6xx (technology). In the major classes dataset, most problems appear with fiction, which is in a way expected since topics there are mostly about genre, language and country rather than



**Research Paper**

actual themes. In particular, 823 (English fiction) is often misclassified as 839 (other Germanic literatures) and 813 (American fiction in English) get misclassified as 823 and 839. In addition, the other worst example is class 306 (culture and institutions) which is often misclassified as 305 (groups of people)—class most problematic when looking at all classes in the dataset.

In the lack of training documents, string matching may complement machine learning (Wartena & Franke-Maier, 2018), provided that the terms denoting the class at hand are appropriate for the task. Looking at a number of individual classes with high accuracy using the string-matching approach, this is best achieved when terms used to denote a class are unambiguous. Top five performing classes are listed below together with their terms and accuracy scores:

- 324.623: kvinnor kvinnlig rösträtt rösträtt kvinnlig rösträtt; 100% accuracy. A close examination shows that women's voting rights in all the records were mentioned in either title of keywords, which lead to total accuracy of this individual class.
- 597.3: Havsängelartade hajar Carcharhiniiformes Notidanoidei zoologi Såghajartade hajar Hajfiskar Hajar Wobbegongartade hajar Gråhajartade hajar Squatiniiformes Heterodontiiformes Chondrichthyes Broskfiskar Pristiophoriiformes Jättehajar Kamtandhajartade hajar Tjurhuvudhajar Selachii Orectolobiiformes Hexanchiiformes Elasmobranchii Lamniiformes Håbrandsartade hajar Selachii hajfiskar Holocephali helhuvudfiskar Sarcopterygii lobfeniga fiskar; 100% accuracy. Although there are many highly specific terms, only two of them led to complete accuracy for this class: sharks (hajar) and fish (fiskar).
- 616.24: medicin KOL Lungor Pulmonell hypertension Pulmonella sjukdomar Kroniskt obstruktiv lungsjukdom Lungsjukdomar Obstruktiv lungsjukdom Lunghypertoni Lungsjukdomar; 97.7% accuracy. This class is denoted by highly unique terms referring to chronic obstructive pulmonary disease, leading to high accuracy.
- 745.61: Textning Skönskrift Konstnärlig textning Kalligrafi Alfabet konsthantverk Kalligrafi; 91.6% accuracy. This is another example with highly specific and unambiguous terms, leading to good automatic classification results.
- 947.0841: Lenin Vladimir Ryssland Ryska revolutionen 1917 rysk historia Kerenskij Aleksandr 1917–1924 Perioden under revolutionerna Aleksandr Kerenskij Vladimir Lenin 1917–1924; 90.3% accuracy. Words denoting Russian revolution and Lenin are another example of specific terms that rather uniquely represent the unambiguous concepts.



In contrast, examples of classes with 0 accuracy are:

- 005.369: Specific programs. This class will have works on programs like “Word for Windows” but will not be classified rightly since the term used to denote the class is generic and works on specific programs will use titles with words denoting the specific program.
- 510.71: Matematik utbildning. This an example of very short term list which is also rather general; titles are usually more specific, leading often to classification failures.
- 782.42164092: Västerländsk populärmusik Populärmusik biografier västerländska sånger Populärmusik sånger västerländska biografier. In this class misclassifications are due to usage of “1900” or “2000” in the record which then got misclassified as another class that used that year. Or a typical problem with metaphors used in arts and humanities—a work titled “En ung naken kvinna: mitt grekiska drama” (in English: “A young naked woman: My Greek drama”) is misclassified as class 480 for classical Greek. Sometimes the problem with strings is impossible to address, such as in “Adjö det ljuva livet” (in English: “Goodbye the sweet life”) which is misclassified as 236.2 for life after death since “livet” is listed here and it would not have make any sense to list it as a synonym for the class at hand.
- 839.736: Svenska romaner 1809–1909 Svenska romaner och noveller 1800-talet Svenska romaner och noveller 1809–1909. This class is a typical problem of describing fiction with terms denoting genre and periods which will not normally be present in the titles of the works described, leading to many works not having any class assigned or works being misclassified.

## 5 Concluding remarks

State-of-the-art machine learning algorithms require at least 1,000 training examples per class. The complete data set we were able to get access to at the time of research involved 143,838 records for 14,413 classes, meaning that DDC had to be reduced to top three hierarchical levels in order to provide sufficient training data, totaling only 802 classes. Achieving high accuracy of 81% reported when using SVM has proven to be dependent on the availability of a good amount of training data, i.e. at least 1,000 records per class. The lack of training data for a large number of classes is even more severe when looking at more specific classes beyond the top three levels; here out of 14,413 available DDC classes, only about 6% of classes had a sufficient number of training examples.

Previous research has demonstrated value in string matching when applying equivalence, hierarchical and related relationships between terms, built in knowledge



**Research Paper**

organization systems such as classification schemes. This may be a good complement for machine learning approaches, especially when lacking training data. Our results show that this may work for specific classes, while in general machine learning outperforms string matching. Fiction seems a hard problem to address in both approaches, not surprisingly so, due its language which is on purpose often vague and metaphorical.

In all, it seems that automatic approaches could be approved in two main ways: 1) increasing the number of training data for machine learning algorithms, 2) enriching DDC with synonyms to increase performance of string-matching algorithms (e.g., with Swedish thesaurus called Swesaurus®)—the latter would additionally serve another purpose, that of increasing the number of subject access points for end users, which would make classification systems like DDC more end-user friendly and help with term disambiguation and query re-formulation, often lacking in library catalogs. For higher levels, also terms belonging to the subclasses could be taken. Increasing the number of training data may be expected over time; however, having 1,000 records per each of over 14,000 classes may be hard to expect, due to the fact that distribution of materials over all classes is rather skewed (as also seen in previous research, see above). Automatic translation using other languages may also create additional noise and lead to smaller accuracy.

Purely automatic approaches for DDC creation cannot be applied in operative systems. On one hand, because performance is not good enough, and on the other, because evaluation is hard to estimate due to low indexing consistency when applying large classification systems like DDC (with many options to choose from): cannot be used as “the gold standard”: the classes assigned by algorithms (but not human-assigned) might be wrong or might be correct but omitted during human indexing by mistake. A more comprehensive approach to ‘gold standard’ production is needed (Golub et al., 2016). As a result, machine-aided indexing would be the best approach to implement in operative systems, similar to the one used by the National Library of Medicine in the USA (see above).

In conclusion, for operative information retrieval systems applying purely automatic DDC does not work, either using machine learning (because of the lack of training data for the large number of DDC classes) or using string-matching algorithm (because DDC characteristics perform well for automatic classification only in a small number of classes). Over time, more training examples may become available, and DDC may be enriched with synonyms in order to enhance accuracy of automatic classification which may also benefit information retrieval performance based on DDC. In order for quality information services to reach the objective of highest possible precision and recall, automatic classification should never



be implemented on its own; instead, machine-aided indexing that combines the efficiency of automatic suggestions with quality of human decisions at the final stage should be the way for the future.

## Acknowledgments

Thanks are due to OCLC which provided the project with electronic DDC, Swedish version. We are very grateful to Rebecca Green and Sandi Jones for all their advice on how to best process and use the electronic DDC files. Many thanks also to the National Library of Sweden who provided all the training and testing data, especially Harriet Aagaard. Special thanks to anonymous reviewers whose detailed comments significantly helped to improve the paper.

## Author contributions

Koraljka Golub (koraljka.golub@lnu.se) conducted the major part of the literature review, coordinated machine learning and string matching analyses and processes, and wrote the majority of the paper. Johan Hagelbäck (johan.hagelback@lnu.se) implemented the machine learning approaches, contributed to the paper with parts related to machine learning and reviewed the paper. Anders Ardö (anders.ardo@gmail.com) prepared the database, implemented the string-matching algorithm, contributed to the paper with parts related to string matching and reviewed the paper.

## References

- Aliwy, A.H., & Ameer, E.H.A. (2017). Comparative study of five text classification algorithms with their improvements. *International Journal of Applied Engineering Research*, 12(14), 4309–4319.
- Anderson, J., & Perez-Carballo, J. (2001). The nature of indexing: How humans and machines analyze messages and texts for retrieval. Part II: Machine indexing, and the allocation of human versus machine effort. *Information Processing and Management* 37(2), 255–277.
- Chen, H., & Dumais, S. (2000). Bringing order to the web: Automatically categorizing search results. In *Proceedings of the ACM International Conference on Human Factors in Computing Systems*, Den Haag, 145–152.
- Golub, K. (2006). Automated subject classification of textual web documents. *Journal of Documentation*, 62(3), 350–371.
- Golub, K. (2007). Automated subject classification of textual documents in the context of web-based hierarchical browsing: PhD thesis. Lund: Department of Electrical and Information Technology, Lund University.
- Golub, K. (2017). Automatic subject indexing of text. In *ISKO Encyclopedia of Knowledge Organization*, 2017. <http://www.isko.org/cyclo/automatic>.
- Golub, K., Soergel D., Buchanan, G., Tudhope, D., Lykke, M., & Hiom, D. (2016). A framework for evaluating automatic indexing or classification in the context of retrieval. *Journal of the Association for Information Science and Technology*, 67(1), 3–16.



## Research Paper

- Golub, K., Hamon, T., & Ardö, A. (2007). Automated classification of textual documents based on a controlled vocabulary in engineering. *Knowledge Organization*, 34(4), 247–263.
- Hunter, R.N. (1991). Successes and failures of patrons searching the online catalog at a large academic library: A transaction log analysis. *RQ*, 30(3), 395–402.
- Khoo, M. et al. (2015). Augmenting Dublin Core digital library metadata with Dewey Decimal Classification. *Journal of Documentation*, 71(5), 976–998.
- Lancaster, F.W. (2003). *Indexing and Abstracting in Theory and Practice*. Facet: London.
- Leininger, K. (2000). Interindexer consistency in PsycINFO. *Journal of Librarianship and Information Science*, 32(1), 4–8.
- Lösch, M., Waltinger, U., Hortsman, W., & Mehler, A. (2011). Building a DDC-annotated corpus from OAI metadata. *Journal of Digital Information*, 12(2).
- Meadow, K., & Meadow, J. (2012). Search query quality and web-scale discovery: A qualitative and quantitative analysis. *College & Undergraduate Libraries* 19(2–4), 163–175.
- Robertson, S., & Zaragoza, H. (2009). The probabilistic relevance model: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4), 333–389.
- Roitblat, H.L., Kershaw, A., & Oot, P. (2010). Document categorization in legal electronic discovery: Computer classification vs. manual review. *Journal of the American Society for Information Science and Technology*, 61(1), 70–80.
- Ruiz, M.E., Aronson, A.R., & Hlava, M. (2008). Adoption and evaluation issues of automatic and computer aided indexing systems. In *Proceedings of the American Society for Information Science and Technology*, 45(1), 1–4.
- Silvester, J.P. (1997). Computer supported indexing: A history and evaluation of NASA's MAI system. In *Encyclopedia of Library and Information Services*, 61(24), 76–90.
- Svanberg, M. (2013). Slutrapport: Dewey-Projektet. <http://www.kb.se/Dokument/Deweyprojektet%20slutrapport%2020130614.pdf>
- Svarre, T.J., & Lykke, M. (2014). Simulated work tasks: The case of professional users. In *Proceedings of the 5th Information Interaction in Context Symposium*, 215–218.
- Svenonius, E. (2000). *The Intellectual Foundation of Information Organization*. Cambridge, MIT Press.
- Thompson, R., Shafer, K., & Vizine-Goetz, D. (1997). Evaluating Dewey concepts as a knowledge base for automatic subject assignment. In *Proceedings of the Second ACM Int. Conf. on Digital libraries (DL '97)*, 37–46.
- Trivedi, M., Sharma, S., Soni, N., & Nair, S. (2015). Comparison of text classification algorithms. *International Journal of Engineering Research & Technology*, 4(2).
- Villén-Rueda, L., Senso, J.A., & De Moya-Anegón, F. (2007). The use of OPAC in a large academic library: A transactional log analysis study of subject searching. *The Journal of Academic Librarianship*, 33(3), 327–337.
- Wang, J. (2009). An extensive study on automated Dewey Decimal Classification. *Journal of the American Society for Information Science and Technology*, 60(11), 2269–2286.
- Wartena, C., & Franke-Maier, M. (2018). A hybrid approach to assignment of Library of Congress Subject Headings. *Archives of Data Science*, 1(4). <https://publikationen.bibliothek.kit.edu/1000105121>



# The Second Edition of the Integrative Levels Classification: Evolution of a KOS

Ziyoung Park<sup>1†</sup>, Claudio Gnoli<sup>2</sup>, Daniele P. Morelli<sup>3</sup>

<sup>1</sup>Department of Library & Information Science, Hansung University, Seoul, Republic of Korea

<sup>2</sup>Science and Technology Library, University of Pavia, Pavia, 27100, Italy

<sup>3</sup>Kaboom, Torino 10138, Italy

## Abstract

**Purpose:** This paper informs about the publication of the second edition of the Integrative Levels Classification (ILC2), a freely-faceted knowledge organization system (KOS), and reviews the main changes that have been introduced as compared to its first edition (ILC1).

**Design/methodology/approach:** The most relevant changes are illustrated, with special reference to those of interest to general classification theory, by means of examples of notation for individual classes and combinations of them.

**Findings:** Changes introduced in ILC2 include: the names and order of some main classes; the development of subclasses for various phenomena, especially quantities and algebraic structures; the order of facet categories and the new category of Disorder; notation for special facets; distinction of the semantical function of facets (attributes) from their syntactic function. The system can be freely accessed online through a PHP browser as well as in SKOS format.

**Research limitations:** Only a selection of changed classes is discussed for space reasons.

**Practical implications:** ILC1 has been previously applied to the BARTOC directory of KOSs. Update of BARTOC data to ILC2 and application of ILC2 to further information systems are envisaged. Possible methods for reclassifying BARTOC with ILC2 are discussed.

**Originality:** ILC is a newly developed classification system, based on phenomena instead of traditional disciplines and featuring various innovative devices. This paper is an original account of its most recent evolution.

**Keywords** Freely faceted classification; Fundamental categories; Knowledge organization system; Phenomenon-based classification

Citation: Park, Ziyong, Claudio Gnoli, and Daniele P. Morelli. "The second edition of the Integrative Levels Classification: Evolution of a KOS." *Journal of Data and Information Science*, vol.5, no.1, 2020, pp. 39–50.

DOI: 10.2478/jdis-2020-0004

Received: Jan. 17, 2020

Revised: Mar. 6, 2020

Accepted: Mar. 10, 2020



<sup>†</sup> Corresponding author: Ziyoung Park (E-mail: zgpark@hansung.ac.kr).

## 1 Introduction

The Integrative Levels Classification (ILC) is a recently developed knowledge organization system (KOS) that includes innovative features. Gnoli (2020) provides a general description of ILC with further bibliography.

While being inspired by traditional bibliographic classification schemes, it lists phenomena instead of disciplines, which makes it suitable for organizing any kind of documents, including museum specimens or digital information. Classes of phenomena are arranged according to the theory of levels of reality (Hartmann 1952), claiming that a series of levels of increasing organization can be identified in the real world, such that higher levels (e.g. consciousness) depend on lower ones (e.g. organisms) for their existence, but at the same time have emergent properties that cannot be found in lower levels (e.g. self-awareness). The series of ILC main classes is thus one of increasing organization, from the most primitive mathematical and physical entities to the most evolved achievements of human cultures (Gnoli, 2017a; Kleineberg, 2017):

- a forms
- b spacetime
- c branes
- d energy; wave-particles
- e atoms
- f molecules
- g continuum bodies
- h celestial bodies
- i rocks
- j land
- k genes
- l bacteria; prokaryotes
- m organisms (eukaryote)
- n populations
- o instincts
- p consciousness
- q language
- r rituals
- s communities
- t polities
- u enterprises
- v technologies
- w artefacts
- x artworks
- y knowledge

The first edition (ILC1) was published in 2011 and included 7,052 classes and facets. A new edition (ILC2) has been published in 2019, which includes 10,851 classes and facets. ILC2 is available at the ILC project website (<http://www.iskoi.org/ilc/>) through a browsable PHP interface as well as in SKOS format.



Like any KOS in its first years, ILC evolves at a relatively quick pace, as a result of progress in research as well as feedback from testing it with actual applications. This paper focuses on the transition from ILC1 to ILC2, by illustrating the main changes that have been introduced in ILC2 and the theory behind them. Sections 2 and 3 concern changes in the list of main classes and relevant developments within some of them, especially mathematics. Section 4 concerns some changes in the fundamental categories on which facets are based. Section 5 concerns the notation for special facets as opposed to that for common facets, and their distinction from “attributes”. Section 6 briefly informs about publication in SKOS format, which will be the object of other papers. Section 7 discusses ongoing application to the BARTOC directory.

## 2 Rearrangement of some main classes

The sequence of ILC main classes has not changed much in its second edition. It is now acknowledged that such 25 levels can be grouped in some major “strata” as in Nicolai Hartmann’s version of levels theory: information (*a-c*), matter (*d-j*), life (*k-n*), mind (*o-p*) and culture including society (*q-v*) and cultural products (*w-y*).

A major change is that ILC1 class *z* “wisdom”, which included spirituality and religions, has been moved to *r* with the new label “rituals”, also including other celebrations and traditional expressions of human societies. The alternative between an anthropological and a spiritual-intellectual placement of religions, whose nature include aspects of both, is an old dilemma in classification: it is indeed acknowledged explicitly in the Bliss Bibliographic Classification, 2nd edition, which leaves users free to choose either option. Placement of “wisdom” at the end of ILC1 main classes, viewed as the most integrated achievement of human spirit, had an equivalent in Dahlberg’s Information Coding Classification.

After its move to *r* in ILC2, *z* is not used anymore as a main class, and this digit remains reserved only to the function of expanding notation for other classes (*emptying digit*). This device can also be applied to main classes that are not yet prominent though being conceptually fundamental, as done for *czp* “preons” which follows *c* “branes”: both of these entities are theorized in fundamental physics and are in wait of future developments in research. It seems reasonable to provide for free notational space in this region, as substantial developments in knowledge are expected in the coming decades.

In ILC1, *r* was used for languages, which are now moved to *q*, with individual languages listed under *qv*. Notation *q* was previously used for other forms of animal communication, much used in the early application to the BioAcoustic Reference Database, which are now subsumed under animal behaviour (“instincts”) in *o*.



Another local move is that of minerals and rocks from subclasses of  $j$  “land” in ILC1 to their own class  $i$  in ILC2. Notation  $i$  was previously occupied by weather which is now an attribute of  $j$ , in an attempt to have main classes better reflecting actually different levels of organization (land is now seen as a level emerging from rocks) rather than just aspects of a same level (weather is seen as an attribute rather than an autonomous level).

### 3 Development of class $a$ “forms”

Among others, class  $a$  “forms” has been largely developed with the addition of various mathematical entities and their facets. Classes of algebraic structures have thus been outlined in greater detail, and several new classes have been introduced, mostly pertaining to the domains of abstract algebra, topology and combinatorics. The new class  $ak$  “spaces” includes various kinds of geometrical spaces like vector, affine or topological spaces. The class  $al$  “algebraic structures” gathers objects ranging from general algebraic systems to groups, ring-like structures and categories, which were previously scattered into various subclasses of  $a$ . Also, a new class  $am$  “combinatorial structures” has been defined in order to contain discrete mathematical objects, such as partially ordered sets and graph-like structures.

The main efforts in the expansion and development of the classification of mathematical structures have been conducted in order to reach an equilibrium between conflicting necessities, trying to maximize usability (both for the indexer and the user) while keeping the whole system as mathematically sound as possible.

Regarding mathematical soundness, we shall recall that all objects in mathematics are, ultimately, just sets. An ideal KOS for mathematical objects, reflecting the real technical nature of abstract structures in terms of sets, would become an extremely complicated system, which would appear practically unusable: in such a system, even common mathematical concepts would have extremely long and complex notations. Therefore, we had to acknowledge the unsuitability of such a perfect representation and accept some compromises in our classification.

Quite often, a given type of mathematical structure may be defined as a set endorsed by a family of properties: this leads to some problems when trying to include such objects in a KOS, as there is no a priori principle indicating in which order those properties shall be taken into account when defining their respective branchings. Occasionally, some properties can be seen as mathematically more fundamental than others, thus having a bigger priority and causing a branching at a prior rank of specificity, but most often there is no clear preference.

A related issue regards the uncertainty of the place of definition, as some structures may be well defined in different places in the tree. ILC endorses the principle of unique place of definition, so we had to choose only one spot for each structure



type. By means of semantic factors it is nonetheless possible to indicate the semantical dependence of a concept with respect to another one defined in a different place in the schedules, thus keeping track of the ambiguity without loss of information.

Construction of classes for quantities and for dates is now possible by letters that stand for negative or positive digits in the same array and produce a correct ordering:

<i>anad</i>	numerals, decimal digits
<i>anade</i>	-9
<i>anadf</i>	-8
<i>anadg</i>	-7
...	
<i>anadm</i>	-1
<i>anadn</i>	-0 [zero in negative numbers]
<i>anado</i>	0
<i>anadp</i>	1
<i>anadq</i>	2
<i>anadr</i>	3
<i>anads</i>	4
<i>anadt</i>	5
<i>anadu</i>	6
<i>anadv</i>	7
<i>anadw</i>	8
<i>anadx</i>	9
<i>anq [anad]</i>	thousands
<i>anqp</i>	one thousand
<i>anqpxor</i>	1903

These digits can be reused to construct dates, as well as to identify persons (in a way similar to Colon Classification) by their birth time:

<i>rab [anad]</i>	historical periods
<i>rabpxor</i>	year 1903 Common Era
<i>U [anad]</i>	persons by birth time
<i>Upxor</i>	persons born in 1903 CE; Konrad Lorenz; John von Neumann

## 4 Facet categories

Facets of ILC are based on a system of ten fundamental categories, expressed with digits 0 to 9, analogous to Colon's "PMEST" or to Vickery's (1975) standard citation order. As compared to these classical orders in faceted classification, the system of ILC1 was original in several respects, like expressing agents in an "origin" category (6) that filed after processes (3). After experience with ILC1 and reconsideration in light of general systems theory (Bertalanffy 1968; Foskett 1980), in ILC2 the citation order of categories is somehow more traditional, with agents in 3 and processes in 5 (now labeled by the more general term "transformation", which also holds for mathematical and geometrical entities). Details of these changes are discussed in Gnoli (2017b).



ILC1 categories	ILC2 categories
<i>0</i> under <i>aspect</i> <i>1</i> at <i>time</i> <i>2</i> in <i>place</i> <b>3</b> through <i>process</i> <i>4</i> made of <i>element</i> <i>5</i> with <i>organ</i> <b>6</b> from <i>origin</i> <i>7</i> to <i>destination</i> <i>8</i> like <i>pattern</i> <i>9</i> of <i>kind</i>	<i>0</i> as for <i>perspective</i> <i>1</i> at <i>time</i> <i>2</i> in <i>place</i> <b>3</b> by <i>agent</i> (ILC1: from <i>origin</i> ) <b>4</b> disturbed by <i>disorder</i> (newly added) <b>5</b> with <i>transformation</i> (ILC1: through <i>process</i> ) <i>6</i> having <i>property</i> <i>7</i> with <i>part</i> <i>8</i> as <i>form</i> <i>9</i> of <i>kind</i>

A remarkable innovation of ILC2 is the original category of “disorder” introduced at 4, covering e.g. the disease facet in organisms and the failure facet in artifacts. This has been inspired by Edgar Morin’s (1977) philosophy of complexity, according to which all real systems are the result of an equilibrium between constructive and destructive factors – a fact often ignored in other views of levels, that depict them as a triumphal, one-directional march towards organization that is probably exceedingly idealistic. Indeed, much of human knowledge to be classified concerns problems and ways to face them, like in medicine or in management. This can be seen as an aspect of general systems theory, that considers all entities as systems composed of parts and links between them, which is another source for the identification of facet categories. Indeed, systems theory and levels theory can be combined fruitfully in consistent treatments, like that of Bunge (2003).

5 Free facets, special facets and attributes

In its most simple application, ILC allows for relevant concepts to be simply juxtaposed, as in mq nyr “animals: forests”. This application can be useful for quick indexing of a general collection of, for example, webpages, videos, or books.

Specialized collections like those of a domain bibliography may need more detailed expression of relationships between concepts, which can be performed by ILC facets. ILC has both *free facets*, that is, facets that can be applied to any class, like “large” or “in Korea”, and *special facets*, which make sense only within certain classes, like “walking” or “wooden”.

In ILC1, digits 0 to 9 introduced both free facets and special facets, and their meaning depended on whether a special facet for the digit on hand was defined in the schedules, and on the complex use of *V* as a facet neutralizer. In ILC2, the distinction has been made more clear in notation, at the price of some longer symbols. Indeed, all indicators of special facets now begin by 9, so that 91-99 introduce special facets, like mq926 “animals, living in *habitat*”, while 0-9 introduce the more general free facets, like 26 “within *environment*”.



Another improvement in the predictability of meaning from notation is that only facets that include a 9 have a shortened notation, that is have foci parallel to those of a given class as specified in the schedules (*mq926r* “animals, living in forests” takes its *-r* from *nyr* “forests”). This can happen not only with special facets, but also with *common facets*, that is free facets that include a 9 in a position other than initial: *29* “in country” takes its foci from *tt* “countries”.

The foci of a special facet are often *context-defined*. For example, *wv97* “vehicles, with *part*” only takes its foci from the context of *w* artefacts themselves, as a part of a vehicle can be a gear or a wheel but not a forest or a country. Now, in ILC1 the only way to express such a vehicle part as “wheels” was *wv97hh*, which literally means “vehicles, with wheels”. There was no way to express wheels as an autonomous phenomenon, like occurring e.g. in the store catalogue of an auto parts seller.

In ILC2, the latter meaning can now be expressed in the form of *attributes* of a class. Attributes are always introduced by a *-a-* which is not used to express subclasses anymore. Thus, *wa* “artefact attributes” include *wahh* “wheels” meant as separate from the artefacts to which they belong.

Attributes usually consist of processes, properties or parts, that are conveniently listed in this order in schedules, though not marked in any particular way for now. However, processes can often be used as foci of process facets (95), properties as foci of property facets (96) and parts as foci of part facets (97, as in *wv97hh* “vehicles, with wheels”).

Expression of wheels as a part using a facet (in ILC1):

<i>w97</i>	with <i>organ, component, part</i>
<i>wv</i>	vehicles
<i>wv97hh</i>	<u>vehicles, with wheels</u> [specifications of a class]

Expression of wheels as a phenomenon itself (newly added in ILC2):

<i>w</i>	artifacts
<i>wah</i>	mechanical components
<i>wahh</i>	<u>wheels as auto parts alone</u> [attribute class]

This distinction between attributes (“wheels”) and faceted compounds (“vehicles, with wheels”) is an innovative feature of ILC as compared to other faceted classifications such as Colon or Bliss. It recognizes Brian Vickery’s (1975) conception, as reported by Coates (1988), that “facets may be characterised alternatively as categories of concepts or as a class of relationships between concepts”. The same consideration was also made by Jacques Maniez, as recently discussed by Hudon and Fortier (2018).



6 Publication as SKOS

The structure of ILC as a freely faceted classification (Gnoli et al., 2011) has been analyzed in order to publish the scheme in SKOS format. SKOS has been chosen because it currently is the standard format for representing knowledge organization systems, despite other formats like OWL that allow for greater expressivity.

While not all structural elements can be represented accurately in SKOS, most classes, captions and synonyms can indeed be published as SKOS, which has been done in collaboration with University of South Wales Hypermedia Research Unit (Binding et al., 2020). The SKOS version of ILC2 is available on the ILC project website (<http://www.iskoi.org/ilc/skos.php>). Using the AllegroGraff 3.3, the SKOS version of ILC2 can be visualized as follows:

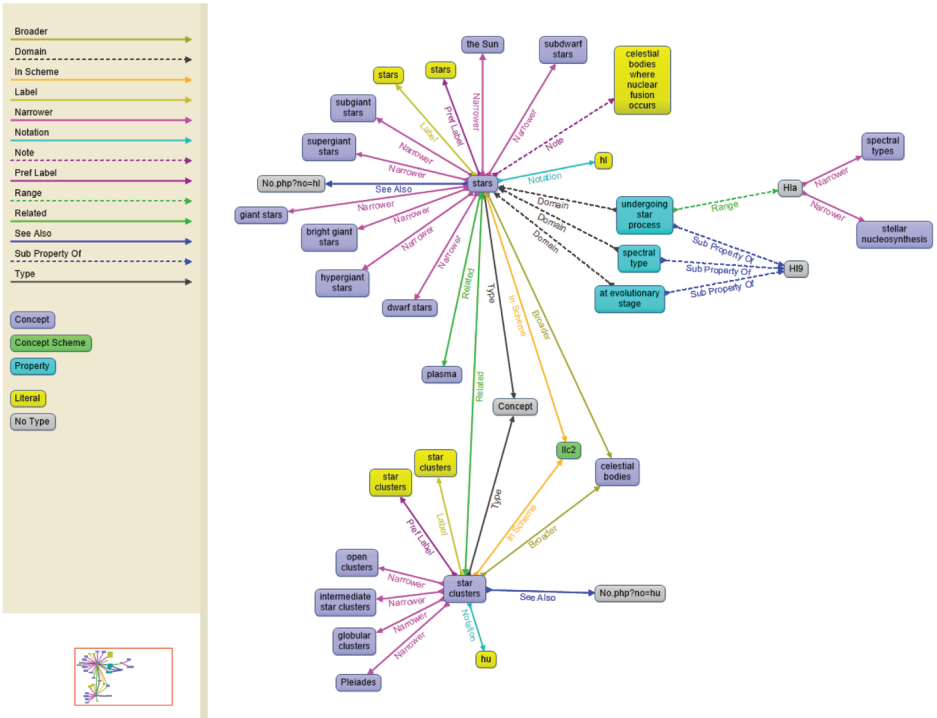


Figure 1. ILC2 as SKOS visualization using AllegroGraff <<http://www.iskoi.org/ilc/skos.php>>.

Visualization in Skosmos is also available on BARTOC Website by Andreas Ledl at Basel University Library.

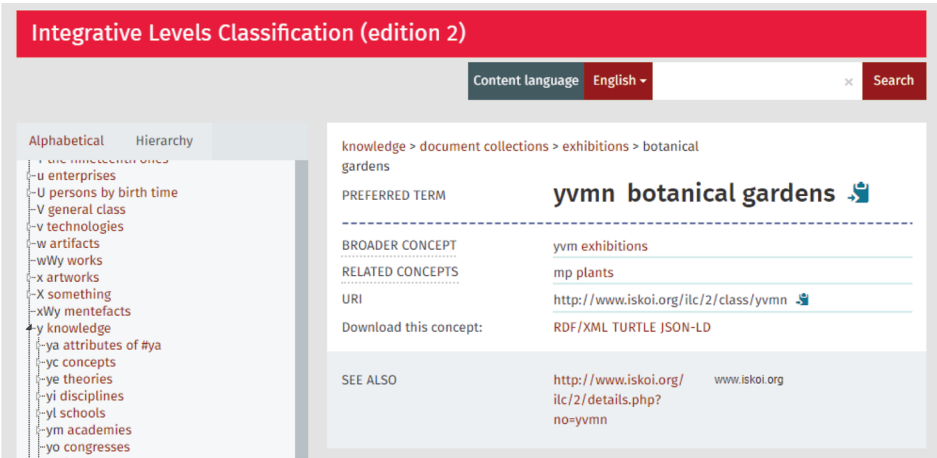


Figure 2. BARTOC Skosmos browser <<https://bartoc-skosmos.unibas.ch/ilc2/en/>>.

## 7 Application to BARTOC

Recent applications of ILC include indexing of the Basel Register of Thesauri, Ontologies and Classifications (BARTOC). In a previous study (Gnoli et al., 2018), ILC1 was used to classify a sample of KOSs in BARTOC. We assigned ILC1 class numbers in a freely combined way to 200 top-rated KOSs plus KOSs of the healthcare domain. By classifying the KOSs in BARTOC with ILC, we could compare the discipline-based DDC with the phenomenon-based ILC and analyze the knowledge dimensions of KOSs in BARTOC (Gnoli, Park, & Ledl, 2019). In accordance with the revision of ILC, ILC2 will be applied to the classification of KOSs in BARTOC.

Reclassification of KOSs in BARTOC can be done in two ways. The first way is to use a mapping table of ILC1 and ILC2. The class numbers of ILC1 applied to BARTOC are basic numbers listed in the ILC1 schedule, without facet indicators or combination of class numbers. Therefore, automatic reclassification is basically possible using mapping information between ILC1 and ILC2 classification scheme. In this process, not only the subclass numbers but also the main class numbers can be changed. For example, the KOS, Thesaurus of Clinical Signs is classified at *mq30* “disease” and *sh* “health care” in ILC1. In ILC2, the class number becomes *mqad* “diseases” and *vm* “health care.”

ILC1 classmark	ILC2 classmark
<i>mq30</i> diseases <i>sh</i> health care ( <i>s</i> civil society)	<i>mqad</i> diseases <i>vm</i> health care ( <i>v</i> technology)



## Research Paper

The first method is effective when the classmarks in ILC1 and in ILC2 are mapped in a 1:1 way. However, as ILC2 is revised, new classes are often created or subdivided. Therefore, it may be necessary to manually assign new classmarks in accordance with ILC2. The classmarks below are some examples of 1: N mapping between ILC1 and ILC2.

ILC1 classmark	ILC2 classmark
<i>n3m</i> “migration”	<i>nam</i> “animal migration” (or) <i>sar</i> “human migration”
<i>s53</i> “occupied as <i>job</i> ”	<i>s975</i> “occupation” (or) <i>uatpb</i> “employment”
<i>sg</i> “cultural services”	<i>spf</i> “cultural services” (or) <i>yv</i> “document collections”

Therefore, in this case, we can reclassify the KOS, International Migration and Colonization, by assigning new ILC2 classmarks after the analysis of the KOS characteristics. The KOS is classed at *n3m* “migration” and *sx* “organized civil society” in ILC1. In ILC2, it will be classed at *sar* “human migration” and *so* “organized civil society.”

ILC1 classmark	ILC2 classmark
<i>n3m</i> migration <i>sx</i> organized civil society	<i>sar</i> <u>human</u> migration <i>so</i> organized civil society

## 8 Discussion

We have reviewed the main changes introduced in ILC2 as compared to ILC1. These concern various areas, from specific classes and subclasses to general syntactical devices for expressing attributes and faceted combinations. Finally, we have shown how changes in classes can affect reclassification of BARTOC items, and evaluated some possible methods to perform it.

We believe that these cases can provide a useful example of how a general KOS with a yet young history can evolve in time. One challenge of this process clearly is conciliating freedom in experimenting new solutions with need of stability for test applications. Indeed, one can not wait until the system is “finished” before applying it, as applications are part of the feedback process that informs the evolution of the system itself. Our paper has tried to document various such changes that have been introduced in ILC2 in the very last years, both to keep track of them and to illustrate a process of KOS evolution between experimentation and stability.



## Acknowledgments

We are grateful to Keiichi Kawamura for reference to Coates and stimulating discussion, and to Mauro Bertani for suggestions concerning negative digits. This research was financially supported by Hansung University.

## Author contributions

Ziyoung Park (zgpark@hansung.ac.kr) analyzed the application ILC to BARTOC, derived reclassification implications of ILC2 development, reviewed the paper. Claudio Gnoli (claudio.gnoli@unipv.it) conducted the introduction and theoretical foundation, analyzed the characteristics of ILC2, and made a major revision to the draft. Daniele P. Morelli (netherself@gmail.com) conducted a major revision to the field of mathematics in ILC2 and review the paper.

## References

- Bertalanffy, L. von. (1968). *General systems theory: Foundations, development, applications*. Rev. ed. New York: Braziller.
- Binding, C., Gnoli, C., Trzmielewski, M., & Tudhope, D. (2020). Integrative Levels Classification as a networked KOS: A SKOS representation of ILC2. *Proceedings of 16th ISKO Conference*, Aalborg, July 2020. Baden-Baden: Ergon.
- Bunge, M.A. (2003). *Emergence and convergence: Qualitative novelty and the unity of knowledge*. Toronto: University of Toronto Press.
- Coates, E.J. (1988). The role of classification in information retrieval: Action and thought in the contribution of Brian Vickery. *Journal of Documentation*, 44(3), 216–225.
- Foskett, D.J. (1980). Systems theory and its relevance to documentary classification. *International Classification*, 7(1), 2–5.
- Gnoli, C. (2017a). Classifying phenomena part 2: Types and levels. *Knowledge Organization*, 44(1), 37–54. DOI: 10.5771/0943-7444-2017-1-37.
- Gnoli, C. (2017b). Classifying Phenomena Part 3: Facets. In Smiraglia, R. & Lee, H. (Eds.) *Dimensions of Knowledge: Facets for Knowledge Organization* (pp. 55–67). Würzburg: Ergon.
- Gnoli, Claudio. (2020). Integrative Levels Classification. In Birger Hjørland and Claudio Gnoli (Eds) *ISKO Encyclopedia of Knowledge Organization*, <https://www.isko.org/cyclo/ilc>.
- Gnoli, C., Ledl, A., Park Z., & Trzmielewski, M. (2018). Phenomenon-based vs. disciplinary classification: Possibilities for evaluating and for mapping. In Ribeiro, F. & Cerveira, M.E. (Eds.) *Challenges and opportunities for knowledge organization in the digital age. Proceedings of the Fifteenth International ISKO Conference*, 9–11 July 2018, Porto (pp. 635–662). Baden Baden: Ergon.
- Gnoli, C., Park, Z., & Ledl, A. (2019). Dimensional analysis of subjects: Indexing KOSs in BARTOC by phenomena, perspectives, documents and collections. *Proceedings of the First ISKO LC Conference*, 20–21 June 2019, Brussels.
- Gnoli, C., Tom, P., Philippe, C., Gabriele, M., & Rick, S. (2011). Representing the structural elements of a freely faceted classification. *Proceedings of the International UDC Seminar*.



**Research Paper**

Classification and ontology: Formal approaches and access to knowledge. 19–20 September 2011 The Hague, Netherlands.

Hartmann, N. (1952). *New ways of ontology*. Westport: Greenwood Press.

Hudon, M., & Fortier, A. (2018). Facet: Itself a multifaceted concept. In Ribeiro, F. & Cerveira, M.E. (Eds.) *Challenges and opportunities for knowledge organization in the digital age. Proceedings of the Fifteenth International ISKO Conference*, 9–11 July 2018, Porto (pp. 204–211). Baden Baden: Ergon.

Kleineberg, M. (2017). Integrative levels. *Knowledge Organization*, 44(5): 349–379. Also available in Birger H. & Claudio G. (Eds.) *ISKO Encyclopedia of Knowledge Organization*. Retrieved on January 11, 2020, from [https://www.isko.org/cyclo/integrative\\_levels](https://www.isko.org/cyclo/integrative_levels).

Morin, E. (1977). *La méthode*. Paris: Seuil.

Vickery, B.C. (1975). *Classification and indexing in science*. 3rd ed. London: Butterworths.



This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivs License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).



# The ARQUIGRAFIA project: A Web Collaborative Environment for Architecture and Urban Heritage Image

Vânia Mara Alves Lima<sup>1†</sup>, Cibele Araújo Camargo Marques dos Santos<sup>1</sup>,  
Artur Simões Rozestraten<sup>2</sup>

<sup>1</sup>Department of Information and Culture, University of São Paulo, São Paulo, Brazil

<sup>2</sup>Department of Architecture Technology, University of São Paulo, São Paulo, Brazil

## Abstract

**Purpose:** This paper presents the ARQUIGRAFIA project, an open, public and nonprofit, continuous growth web collaborative environment dedicated to Brazilian architectural photographic images.

**Design/methodology/approach:** The ARQUIGRAFIA project promotes the active and collaborative participation among its institutional users (GLAMs, NGOs, laboratories and research groups) and private users (students, professionals, professors, researchers), both can create an account and share their digitized iconographic collections in the same Web environment by uploading their files, indexing, georeferencing and assigning a Creative Commons license.

**Findings:** The development of users interactions by means of semantic differentials impressions recording on visible plastic-spatial aspects of the architectures in synthetic infographics, as well as by the retrieval of images through an advanced system search based on those impressions parameters. By gamification means, the system often invites users to review images' in order to improve images' data accuracy. The pilot project named Open Air Museum that allows users to add audio descriptions to images in situ. An interface for users' digital curatorship will be soon available.

**Research limitations:** The ARQUIGRAFIA's multidisciplinary team gathering professors-researchers, graduate and undergraduate students from the Architecture and Urbanism, Design, Information Science, Computer Science faculties of the University of São Paulo, demands continuous financial resources for grants, for contracting third party services, for the participation in scientific events in Brazil and abroad, and for equipment. Since 2016, significant budget cuts in the University of São Paulo own research funds and in Brazilian federal scientific agencies can compromise the continuity of this project.

**Practical implications:** The open source template called +GRAFIA that can freely help other areas of knowledge to build their own visual Web collaborative environments.

Citation: Lima, Vânia Mara Alves, Cibele Araújo Camargo Marques dos Santos, and Artur Simões Rozestraten. "The ARQUIGRAFIA project: A web collaborative environment for architecture and urban heritage image." *Journal of Data and Information Science*, vol.5, no.1, 2020, pp. 51–67.

DOI: 10.2478/jdis-2020-0005

Received: Jan. 17, 2020

Revised: Mar. 2, 2020

Accepted: Mar. 10, 2020



<sup>†</sup> Corresponding author: Vânia Mara Alves Lima (E-mail: [vamal@usp.br](mailto:vamal@usp.br)).

**Research Paper**

**Originality/value:** The collaborative nature of the ARQUIGRAFIA distinguishes it from institutional image databases on the internet, precisely because it involves a heterogeneous network of collaborators.

**Keywords** Metadata; Architectural images; Collaborative web environment; Digital repository

## 1 Introduction

Knowledge Organization (KO) in the context of the new epistemological challenge called Digital Humanities—defined as the articulation of knowledge and methods used in the human sciences with the digital world (Guerreiro & Borbinha, 2014)—needs to establish new methods and procedures, as well as develop increasingly efficient tools, to represent and retrieve the iconographic information, which has grown exponentially as a result of increasing digitization of physical collections of images to be made available on the web.

For some years now we have had initiatives to preserve and guarantee access to cultural heritage, such as Europeana (<http://www.europeana.eu/portal/pt>) and the Getty Research Portal (<http://portal.getty.edu/>). Europeana is a free curatorial image platform that gathers several European institutions aiming to transform the world through culture, sharing high-quality iconographic collections. The Getty Research Portal (<http://portal.getty.edu/>) is a free online search platform providing worldwide access to an extensive collection of digitized art history texts and images from a worldwide range of institutions. It is an inter-institutional collaborative project initiated by the Getty Research Institute responsible for the most well-known tool for describing architecture and urban cultural heritage, the Art & Architecture Thesaurus.

Besides these large scale GLAMs (Galleries, Libraries, Archives, and Museums) many smaller institutions also run digitization centers and a steadily growing number of platforms to render their digital contents globally accessible for research and the interested public as the ETH Library (Gasser, 2017). ETH Zurich's main library with its shift in strategic focus towards the “digital library” has implemented crowdsourcing, a form of user participation used very successfully by a wide variety of institutions all over the world to enhance digital collections and raise their profile. In the ETH Library initiative, the volunteers have free access to the digitized images and they have their names included in the comments field on the Image Database, which is an incentive to continue collaborating with the growing amount of improved metadata. Similar to the ETH Library initiative, but expanding the collaborative approach, we present the ARQUIGRAFIA project.



ARQUIGRAFIA is a web collaborative environment for the preservation, research and dissemination of images of Brazilian architecture and urban spaces, which enables interactions between people and institutions. This digital environment contributes to research on architectural and urban heritage, as well as allowing the organization of Brazilian architectural images on the web (Lima et al., 2016). The collaborative nature of the ARQUIGRAFIA project distinguishes it from institutional image databases on the internet, precisely because it involves a heterogeneous network of collaborators: institutional users such as GLAMs, NGOs, Universities and Research Groups, together with private users such as students, teachers, photographers and people in general.

Since 2010, the ARQUIGRAFIA project has been facing scientific and technological challenges for creating apps and system features that promote active and collaborative participation among its users. Institutional and private users can create an account and share their digitized iconographic collections in the same Web environment by uploading their files, indexing, georeferencing and assigning a Creative Commons license. However, collaboration goes beyond uploading images and deals with user interactions—exchanging assessments, impressions, judgements on the architectural qualities represented in the photographs.

Most of the ARQUIGRAFIA users are architecture students (38.6%); but there are also architects (23.4%); undergraduate students in other areas (10.3%); teachers of architecture (4.3%); photographers (4.1%); and interested lay people (19.3%). Most of these users are between 20 and 30 years old. Together, they collaborate to tag, georeference and assign specific licenses (Creative Commons) to each image of their collections. Being an environment of teaching and researching, we can also include it in the field of Digital Humanities, since this field is defined by research in collaboration with teaching activities, combining computing and information technologies with academic practices in the field of humanities (Lima, Rozestraten, & Orth, 2016).

ARQUIGRAFIA has several technological and scientific integrated fronts: from the cleaning and conservation of original photographic images to its digitization, from the training of researchers to the development of a perpetual beta open source software. From the point of view of knowledge organization systems (KOS), ARQUIGRAFIA comprises both the broader and the narrower meanings of Knowledge Organization (KO) as defined by Hjørland (2008). The more general and broader meaning, because it establishes an inevitable relationship between knowledge and its organization in the society, since its images also represent the social organization of knowledge and its tags represent the conceptual structure in the field of the architecture and urbanism. The more specific and narrower meaning,



because the construction of its controlled vocabulary seeks to organize knowledge and information intellectually and cognitively, facilitating its management and retrieval, including, in addition to institutional indexing, social indexing as well. Therefore, ARQUIGRAFIA provides an opportunity to look accurately at contemporary issues, specifically regarding terminologies and controlled vocabularies for the representation of images. Other issues include the sharing of metadata between systems; as well as the consolidation of standards that respond both to international interoperability requirements and to local needs for information access and organization.

From a technological perspective, ARQUIGRAFIA plays the role of a pilot program for a template called +GRAFIA that was based on PHP Laravel and can offer free help for other areas of knowledge to build their own visual collaborative environments, such as, for example, a hypothetical BOTANYGRAFIA dedicated to the flora, or an ARTGRAFIA, dedicated to visual arts.

Conceptual and technological challenges concerning the design and the operation of ARQUIGRAFIA allow us to characterize it as an online experimental laboratory and a case study on the opportunities and the risks of digital projects on the humanities based on image collections. ARQUIGRAFIA proposes a new interface for the information environment of the institutional collection and, in so doing, it confronts controlled procedures of knowledge representation with the semantic capability of organizing the information for the web and networks.

## 2 First steps

The main source of the images for ARQUIGRAFIA is the photographic collection of the Iconographic Material Sector of the Library of the School of Architecture and Urbanism of the University of São Paulo (FAUUSP). A set of 42,000 images (34,000 slides and 8,000 black and white photographs) were scanned. Of this total of digitized images, more than 8,000 images were uploaded to the ARQUIGRAFIA system, precisely those that already have the authorizations of the copyright holders as a Creative Commons license. The remaining images are in the curatorial stage of requesting authorization and analysis for upload. In addition, 3,000 images were uploaded and cataloged by private users as well, and other 1,782 belonging to other institutional collections such as the Republican Museum of Itu (<http://www.mp.usp.br/museu-republicano-de-Itu>) and the QUAPÁ, the Panorama of Brazilian Landscaping Project (<http://quapa.fau.usp.br/wordpress/>). Figure 1 shows ARQUIGRAFIA's home and login page.

Between 2012 and 2017, a team of scholarship-funded undergraduate students carried a wide effort towards cleaning, identifying and organizing the original images, as well as digital files and backups. Having to deal with FAUUSP Library's



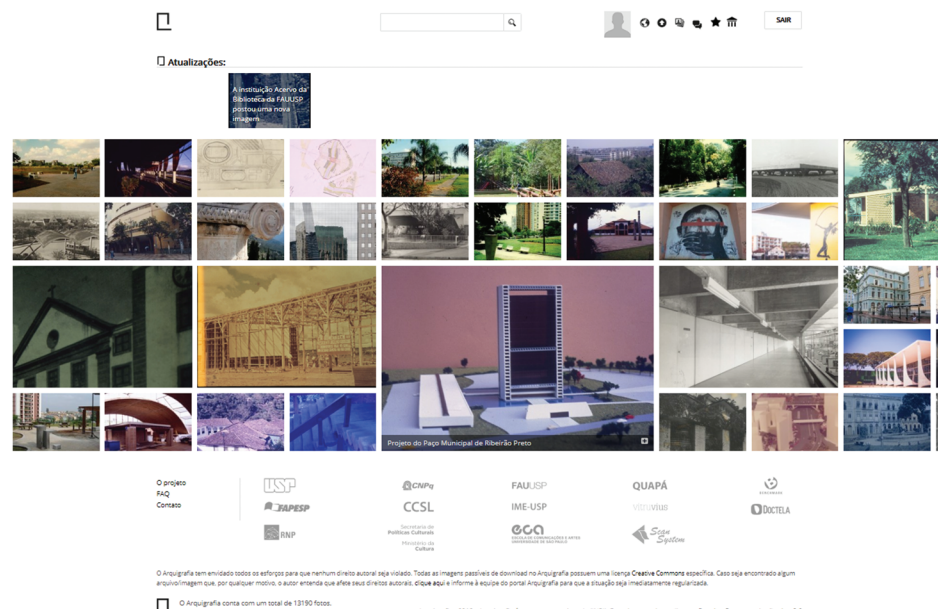


Figure 1. ARQUIGRAFIA homepage.

Source: <http://www.arquigrafia.org.br/home>

sizeable collection of photographs, ARQUIGRAFIA chose to focus on Brazilian Architecture and its urban spaces, letting aside foreign architecture images for now. This set was then cleaned and stored with proper materials, such as folders and mounting corners, which helps the conservation of the originals.

At the same time, cataloguing results were revised based on a survey of procedures and standards for the descriptive and thematic representation of images in order to define the set of metadata which best suits the organization and retrieval of information.

For the digitization of the institutional collection, a third-party company was hired using Plustek Optic 120 film and Silver Fast Ai Studio 8 (64-bit) scanning software that helps in the removal of dust and scratches. Each image was scanned without color correction in order to preserve the original appearance of the photographs and slides, keeping the time stamps (color changes, smudges, saturation, etc.) and its historical aspect. Each generated file is 5 MB with 4,000 dpi resolution and is saved in TIFF, JPEG, and PDF formats and recorded on DVDs and external hard drives. After the scan, backups of the set of images were created and the images were uploaded to ARQUIGRAFIA.

Each image received a registration number, which allows the association with the metadata of its cataloguing and description. A program transformed the metadata



**Research Paper**

into content objects used by the system. In order to make this transformation, the Apache ODF Toolkit software (<http://incubator.apache.org/odftoolkit>) was used to create a communication interface between the metadata and the ARQUIGRAFIA system for information mining and transformation.

Then, the information storage activity allows the creation of associations between the content objects and their representation in the database. Once the object association is made (an author is associated to an image and this image to an address), the system uses the Hibernate persistence library (<http://www.hibernate.org>) to store the database in Mysql (<http://www.mysql.com>).

### 3 Metadata

To define the metadata, some cataloging standards were analyzed, such as the Anglo American Cataloging Code—AACR2; the International Standard for Bibliographic Descriptions for Non-Book Material—ISBD (NBM) and content standards such as Cataloging Cultural Objects—CCO. From the analysis of these standards and the identification of the information required in ARQUIGRAFIA, a spreadsheet was developed to integrate the metadata necessary for the representation of the images and the data administration in the collaborative web environment. In this way, a set of metadata was established according to Table 1.

Table 1. ARQUIGRAFIA Metadata.

Image Metadata Level	Type of Information
Descriptive metadata	Title, Number of the classification, Name, Country, State, City, District, Street, Image author, Tags, Image date, Project author, Construction date, Notes, Date of registration number, Date of cataloging
Structural metadata	Dimensions, width, height, resolution, color depth, color model
Administrative metadata	License ( <i>Creative Commons</i> ), Harvesting, Donors, Authorization form for web distribution

Source: elaborated by the authors.

Regardless of whether they have personal or institutional access users must fill in at least the title, the name of the author of the image, the country and some tags that represent the subject to upload an image to ARQUIGRAFIA. This procedure can be performed on the website through any device, from notebooks to smartphones (Rozestraten, Lima, & Santos, 2017).

In order to encourage users to closely observe images and formulate judgments about buildings and urban spaces represented in the photographs, from the point of view of Architecture, ARQUIGRAFIA proposes to its users the recording of impressions based on pairs of opposing plastic-spatial qualities, called binomials (Figure 2).



SAIR

Residência

Manuel Antônio

Mendes André

Tags:

fachada, residência, concreto, muro

Imagens interpretadas com média similar

(2) Imagens

Suas impressões da Residência Manuel Antônio Mendes André

☐ Eu conheço pessoalmente esta arquitetura.
 ☐ Estou no local.

Para cada um dos pares abaixo, quais são as qualidades predominantes na arquitetura que são visíveis nesta imagem?

Aberta ( 10 %)

Fechada ( 90 %)

Interna ( 20 %)

Externa ( 80 %)

Complexa ( 75 %)

Simple ( 25 %)

Simétrica ( 89 %)

Assimétrica ( 11 %)

Translúcida ( 33 %)

Opaca ( 67 %)

Horizontal ( 85 %)

Vertical ( 15 %)

ENVIAR

VOLTAR

Figure 2. ARQUIGRAFIA binomials.

Source: <http://www.arquigrafia.org.br/>

The binomials are organized as semantic differentials such as: open/closed; internal/external; translucent/opaque; complex/simple; symmetrical/asymmetrical; horizontal/vertical. The conceptual underpinnings for these pairs of opposite qualities come from Henrich Wölfflin (1864–1945) “Principles of Art History” (1950), adapted to Architecture, organized as Charles E. Osgood’s (1916–1991) semantic differentials (1990). ARQUIGRAFIA invites users to record their impressions of the architecture represented in an image, based on six pairs of opposing qualities. Gathering multiple individual impressions as collective interpretations, the system can guide cross-system navigations by means of interactions between images with similar and/or mirrored profiles.

To do so, the average of interpretations is calculated and shown in a chart (Figure 3) and compared with the average of other images that had their binomials



Research Paper

analyzed in the system, allowing the identification and retrieval of images with similar patterns.

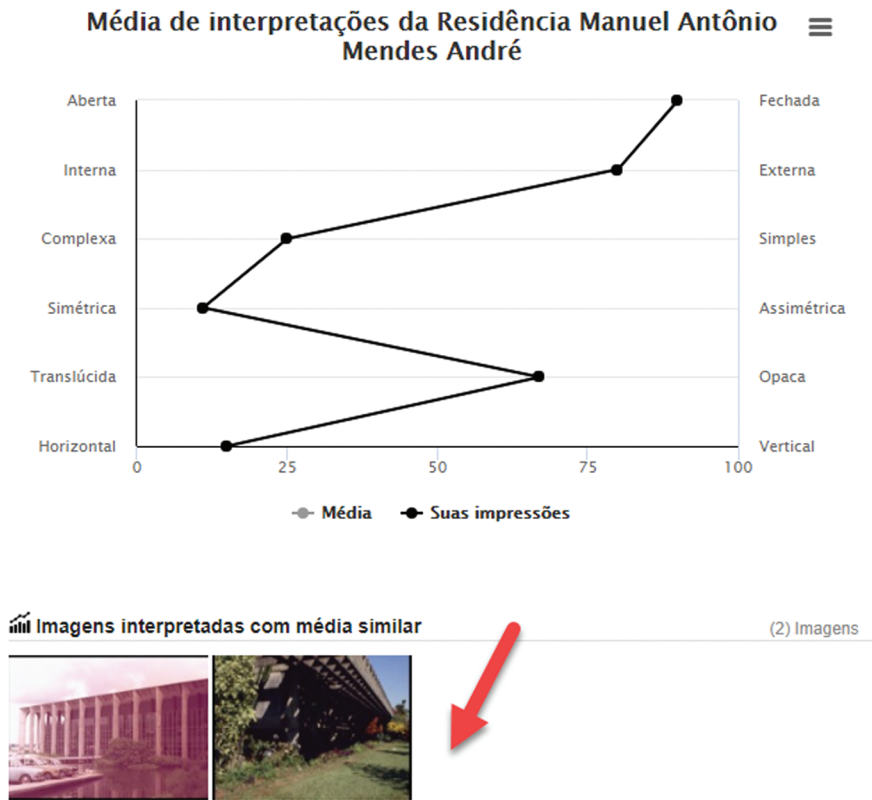


Figure 3. An example of an image interpretation average with suggestions of possibly similar images.  
Source: <http://www.arquigrafia.org.br/>

In addition, it allows the system to establish, for every image, a comparative perspective between its original interpretation, the later interpretations, and the average of all interpretations already made (Rozestraten et al. 2010). The various interpretations are recorded and can change the classifications of similarities. Additionally, the interpretations of each single user can also establish its profile and preferences over the years.

#### 4 Tags vs controlled vocabulary for a Knowledge Organization System

The inclusion of tags or markers by users with natural language is called folksonomy, social tagging or social indexing. Tagging can greatly contribute to the



creation and management of digital collections, as it is carried out collaboratively, distributing resources, activities, and reducing costs. Its importance for the organization, retrieval and access to digital information is demonstrated by the studies of Angus et al. (2010) and Bradley (2011) on Flickr, a photo sharing and social networking site.

Specifically, the study by Angus et al. (2010) explored the potential use of the Flickr image site as an academic image resource by identifying tagged images belonging to academic subject categories. Image content analysis and term frequency analysis provided information from the context of the image. The results of this study showed the possibility of using the tool as a resource for specific images in some areas of knowledge and for individual academic studies, which reinforces the relevance of using social indexing in an image repository developed for teaching, research and extension purposes such as ARQUIGRAFIA.

The tagging performed at ARQUIGRAFIA is related to Vander Wal's specific folksonomy model, where one or a few people insert the tags (Moreiro Gonzalez, 2011; Rafferty, 2018). Thus, we find in this collaborative environment a mixed knowledge organization system where any user can upload an image and tag it using the suggestion list (controlled vocabulary) or adding terms, but only this user can edit the information and the tags that he inserts. However, other users may contribute by reviewing the information and indicating additions and corrections via the contributor's own editing system. Despite the ease of use of social indexing and the approximation with the active vocabulary of users, the lack of language control can present difficulties for information retrieval.

Rafferty (2018), in an entry on the ISKO Encyclopedia of Knowledge Organization, refers to tagging as the practice in which web users use keywords to describe, categorize or comment on digital content.

This marking allows an individual response to information objects by configuring a triad formation: the user, the information object and the keyword, keeping them connected as observed in Figure 4. The tags are such as: "*representar2015*" and "*#representar2015*" for the Architecture event held in 2015, "*library*" the type of building; "*ufu*" the abbreviation of the institution where the event happened; "*uberlândia*" the city; "*minas gerais*" the province; "*paulo zimbres*" the architect; "*cobogó*" the type of the wall with hollow elements and "*tijolo de barro*", mud brick.

For Font, Serra & Serra (2013), collaborative tagging has emerged as a solution for labeling and organizing digital content on the web. However, these collaborative tagging and social indexing systems present problems regarding tag ambiguity, synonyms and the amount of content words used, and it can be inferred that the organization and navigation of content marked in this way may present difficulties for the efficient retrieval of information.



## Research Paper



Figure 4. Central Library of Universidade Federal de Uberlândia.

Source: <https://www.arquiografia.org.br/photos/6298>

The ANSI/NISO Z39.19 “Guidelines for the construction, format, and management of monolingual controlled vocabularies” defines controlled vocabulary as a list of explicit and controlled terms, and these terms may not be ambiguous and must contain definitions which are not redundant (National Information Standards Organization, 2010).

Bearing in mind that in ARQUIGRAFIA both personal and institutional users can upload and index the images of collections of photos, then it is possible to understand the inherent tension between the use of free-form user terminology and the control of indexed information by the institutional users. It happens because this information system needs to fulfill the demands of organizing the institutional images, for the purpose of academic retrieval and preservation, and at the same time count with social indexing and personal user participation for system feedback with images and markers.

Therefore, it was necessary to do the terminology standardization between the lists of subjects used by the library for the indexing of the photographs and slides; the terms of architecture and urbanism of the Controlled Vocabulary of the Integrated Library System of the University of Sao Paulo; the list of default tags of ARQUIGRAFIA, based on the expertise of its team and the tags employed by the users which were harvested in the database.

The 1,145 terms of the list of tags assigned by personal users (Santos & Santos, 2017) were analyzed between 2017–2019 with the purpose of allowing their inclusion in the controlled vocabulary under construction, considering their definitions and the equivalence relationships between them and the terms of the lists of the library, the VOCAUSP and the default list of ARQUIGRAFIA.

The tags that belong to the users' semantic universe surely enrich the vocabulary, showing how they think and retrieve the information. The result constitutes the first version of a collaborative controlled vocabulary that acts as a suggestion of terms for all users, maintaining the possibility of inserting new tags later.

Besides the user guarantee, an indexing tool needs to obtain the literary guarantee too. The literary guarantee of the ARQUIGRAFIA controlled vocabulary is based on research of the terms in dictionaries, glossaries, encyclopedias and specific terminologies and in the terminological method.

The terminological method consists in defining the term from the characteristics of the concept and its definition in the domain based in the works of Dahlberg (1978, 2009, 2011, and 2014), Cabré (1995), the ISO 25964-1 (2011) and their application in the construction of the controlled vocabulary of arts by Lima, Costa, and Guimarães (2017). The characteristics indicate the extension and the intension of the term, the extension being understood as the class of all things that the term applies to, and the intension as the properties that an object must have to be in the scope of the term definition. Each true statement about a certain property or characteristic of an object delivers a knowledge element about it. The sum of the statements about such an object forms the whole of characteristics of its concept. These statements also form its definition, such as in Dahlberg's example: a museum is a public building; it serves for the exhibition of objects; it possesses collections of certain fields of study; it presents collections thematically; it has certain times for visitors and controls visitors (in general) by means of tickets (Dahlberg, 2009).

So far, it has been possible to standardize and define 1,300 terms which have been included in five categories: form (building type), function (use of the building, past or present), materials (materials used in the construction), technique (construction technique used) and history. These categories make up the first level of the controlled vocabulary structure and after that we have been working to establish its logical and ontological relations based on the relationships between their characteristics.

This list was shared in a Google Drive spreadsheet (Figure 5) with students and teachers from the research group containing: definition; source (dictionary or thesaurus from which the definition originated); proposed definition (based on characteristics and according to terminological method); references and synonyms; if the term is in the USP Controlled Vocabulary; hierarchy in USP Vocabulary; possible relationships in the ARQUIGRAFIA vocabulary; suggested hierarchy and consistency with the indexed images.

Finally, these relations are established by using the terminological procedures and the procedures for the construction of controlled vocabularies indicated in the ISO standards (ISO, 2000; 2011). At the same time, the categories and terms were included in a mind map software (Figure 6) to visualize the hierarchical relations



75% - R\$ % 0.00 123 - Arial - 10 - B I A

fx | LISTA DE TAGS

	A	B	C	D	E	F	G	H	I
	LISTA DE TAGS	Definição	Fonte	Definição Proposta	REMÍSSIVAS / SINÓNIMOS	VOCAUSP	Hierarquia VOCAUSP	Relações possíveis no Vocabulário Arquigrafia	Hierarquia st
1		[ACESSO] - O ato de chegar ou entrar [METRÔ] - 1. Sistema de estrada de ferro, geralmente subterrâneo, destinado ao transporte rápido de passageiros em meios urbanos.	Priberam	Espaço interno ou externo por onde é possível entrar no Metrô (VL)		não			
3	Acesso ao metrô								
4	aço corten	É um tipo de aço patinável, ou seja, com pequenas adições de elementos, como cobre, fósforo, níquel e cromo, que em condições ambientais contribuem para a formação de uma película que protege esses aços de aço corrosiva na atmosfera oxidante de muitos ambientes urbanos. Apresenta-se com uma cor laranja-avermelhada.	E-civil	Tipo de aço patinável (OP)		não		Não tem no Arquigrafia	
5	aço inoxidável	Aço resistente a oxidação e corrosão pela inclusão na sua composição de alto teor de cromo. É adequada sua utilização em locais próximos ao mar, pois resiste bem à maresia. É empregado em diversos elementos da construção, como ESQUADRIAS, CORRIMÕES e LUMINÁRIAS. É também chamado aço-cromo.	Albemaz	Tipo de aço resistente a oxidação, também chamado de aço-cromo (OP)	Aço cromo	sim	Engenharia metalúrgica > Metalurgia extrativa > Ferrosos > Aço > Aço inoxidável	Não tem no Arquigrafia	

+ Planilha1 Explor

Figure 5. Spreadsheet of the controlled vocabulary under construction.  
Source: ARQUIGRAFIA research team.

and make decisions about the position of the term in the controlled vocabulary structure.

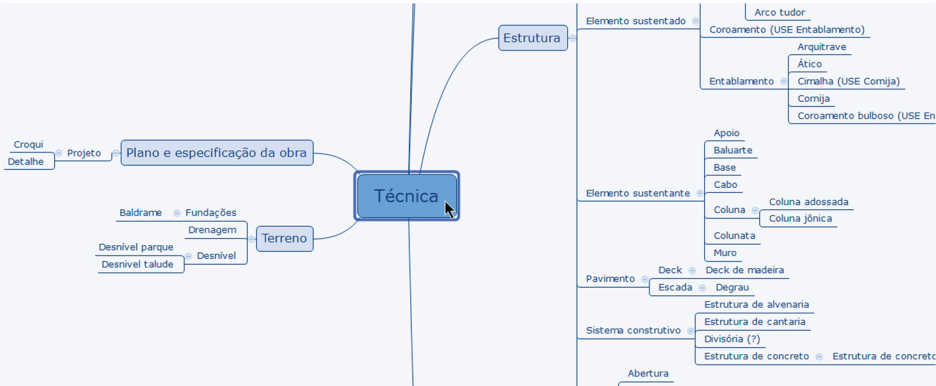


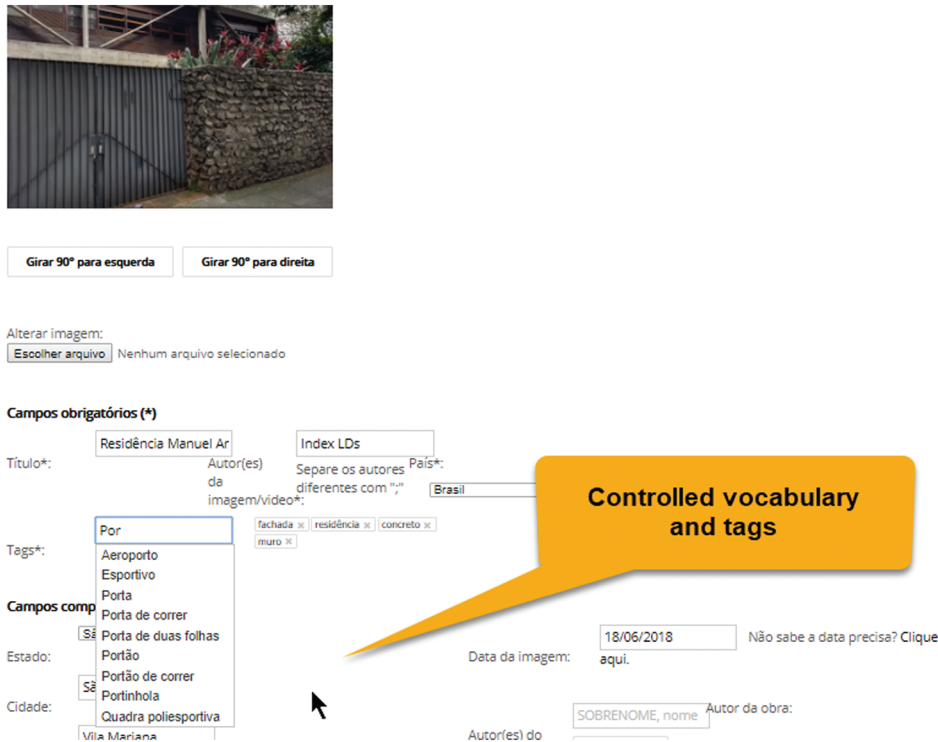
Figure 6. Mind map by categories.  
Source: ARQUIGRAFIA research team.

Periodically, it will be necessary to harvest new tags uploaded by the users and submit them to this process of standardization for further inclusion in the controlled vocabulary, avoiding synonymy (two words with the same meaning) and polysemy (a word with several meanings), contributing to the improvement of information retrieval as shown in Figure 7.

For the consistency of the indexing, it is necessary to have knowledge of the subject area, to analyze the characteristics of the support and its contents, and to develop clear rules for the use of the controlled vocabulary.

ARQUIGRAFIA’s indexing policy indicates that it is advisable to tag the materials used in the construction of the work that are visible in the foreground of the image





The screenshot displays the ARQUIGRAFIA web interface. At the top, there is a photograph of a building with a corrugated metal roof and a stone wall. Below the photo are two buttons: "Girar 90° para esquerda" and "Girar 90° para direita". Underneath these buttons is a section for "Alterar imagem:" with a button "Escolher arquivo" and the text "Nenhum arquivo selecionado".

The main section is titled "Campos obrigatórios (\*)". It contains several input fields and a dropdown menu. The "Título:" field has the text "Residência Manuel Ar". The "Autor(es) da imagem/video\*:" field has the text "Index LDs". The "País\*:" field has the text "Brasil". The "Tags\*:" field has a dropdown menu with the following options: "Por", "Aeroporto", "Esportivo", "Porta", "Porta de correr", "Porta de duas folhas", "Portão", "Portão de correr", "Portinhola", "Quadra poliesportiva", and "Vila Mariana". The "Estado:" field has the text "Sã". The "Cidade:" field has the text "Vila Mariana".

Below the "Tags\*" dropdown, there is a section for "Controlled vocabulary and tags" with a yellow speech bubble. It contains a list of tags: "fachada x", "residência x", "concreto x", and "muro x".

At the bottom, there are fields for "Data da imagem:" (18/06/2018), "Autor da obra:" (SOBRENOME, nome), and "Autor(es) do:".

Figure 7. ARQUIGRAFIA controlled vocabulary and tags.

Source: <http://www.arquigrafia.org.br/>

as well as the architectural elements present which are identified from the type of building and/or urban space and its functions (Rozestraten, Andrade, & Figueiredo, 2018).

## 5 Results

In the beginning of 2019, a responsive version of the ARQUIGRAFIA was implemented to encourage its users to upload georeferenced photographs using their smartphones. Currently, an interface is being developed for the creation of exhibitions that will allow digital curatorship by users, as well as the prototype of an Open Air Museum with audio descriptions of the images thanks to the partnership with the Smart Audio City Guide (Rozestraten, 2013), a project supported by the National Council for Scientific and Technological Development (CNPq).

On the user's profile, there is the possibility of chatting with other users of ARQUIGRAFIA, which then works as a social network, with the creation of photo



## Research Paper

albums, and insertion of contributions (Figure 8), which encourage the users to review the cataloguing of the images by means of gamification processes.

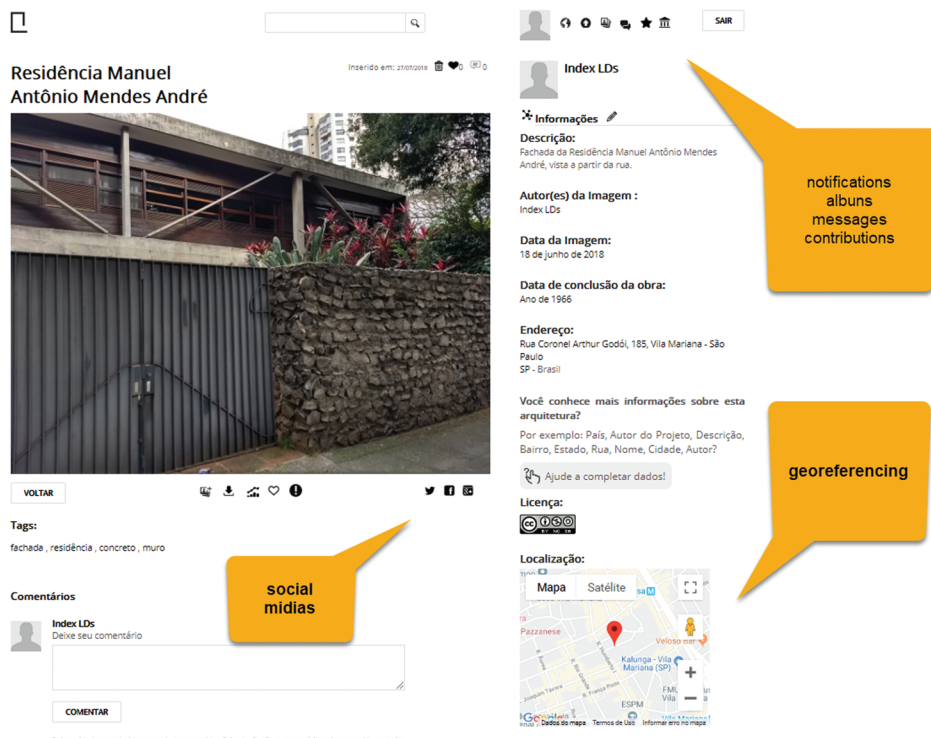


Figure 8. ARQUIGRAFIA social interaction resources.

Source: <http://www.arquigrafia.org.br/>

The User-Centered Design (UCD) procedures were included based on gamification elements aimed at a greater user engagement especially with interface elements related to collaboration, such as: notifications; posts; complementing information about images and comments; the possibility to follow and be followed by other users.



## 6 Summing up

Due to the collaborative nature and the characteristic of enriching its metadata with user-generated content, ARQUIGRAFIA fits in the definition of the Digital Humanities (DH), understood as a new epistemological challenge where we have the articulation of knowledge and methods used in the human sciences with the digital world. At the same time, it faces challenges related to the sustainability of a system in continuous growth, which includes development and programming;

storage and preservation; management and insertion of data and images. The ending of the digitization of the set of 42,000 images in addition to the digital preservation of the information uploaded into the system still brings us to a challenge regarding its digital curatorship, as well as issues related to copyright and the obtaining of licenses for insertion in ARQUIGRAFIA.

Currently, the ARQUIGRAFIA research team deals with new short and medium term objectives such as:

- the implementation of a first version of a moderation system integrated with gamification;
- the development and dissemination of the +GRAFIA template;
- the studies related to the plastic-spatial qualities of the binomials arranged as semantic differentials, seeking to define an image evaluation model based on visual similarities that may be useful for information retrieval;
- the engagement of an interactive community around the images and their information, seeking long-term sustainability for the system.
- the evaluation of the descriptive, technical and administrative metadata to make them more interoperable in a web collaborative environment.
- the improvement of the controlled vocabulary as a visual knowledge organization system

Finally, we must conduct usability studies to evaluate the current beta version and redesign the system from the critical observations made by the users, to expand the iconographic base of ARQUIGRAFIA, including digital video, drawings and other audiovisual resources, as well as to deepen the research into new relevant topics and future developments.

## Author contributions

Vânia Mara Alves Lima (vamal@usp.br) conducted the knowledge organization literature review, analyzed the available materials generated by the stakeholder-driven research project and wrote the paper. Cibele Araújo Camargo Marques dos Santos (cibeleac@usp.br) analyzed the available materials generated by the stakeholder-driven research project and wrote the paper. Artur Simões Rozestraten (artur.rozestraten@usp.br) analyzed the available materials generated by the stakeholder-driven research project and reviewed the paper.

## References

- Angus, E., Stuart, D., & Thelwall, M. (2010). Flickr's potential as an academic image resource: An exploratory study. *Journal of Librarianship and Information Science*, 42(4), 268–278.
- Bradley, P. (2011). From Flickr to Playbills: How to find the right images. *CILIP UPDATE with gazette*, (Jul), 23–23.



**Research Paper**

- Cabré, M.T. (1995). La terminologia hoy: Concepciones, tendências y aplicaciones. *Ciência da Informação*, 24(3), 289–298.
- Dahlberg, I. (2009). Brief communication: Concepts and terms—ISKO’s major challenge. *Knowledge Organization*, 36(2/3), 169–177.
- Dahlberg, I. (2011). Brief Communication: How to improve ISKO’s standing: Ten desiderata for knowledge organization. *Knowledge Organization*, 38(1), 69–74.
- Dahlberg, I. (2014). Brief communication: What is knowledge organization? *Knowledge Organization*, 41(1), 85–91.
- Dahlberg, I. (1978). Teoria do conceito. *Ciência da. Informação*, 7(2), 101–107.
- Font, F., Serra, J., & Serra, X. (2013). Folksonomy-based tag recommendation for collaborative tagging systems. *International Journal on Semantic Web and Information Systems*, 9(2), 1–30.
- Gasser, M. (2017). Research partnerships, user participation, extended outreach—some of ETH Library’s steps beyond digitisation. DH. Opportunities and Risks. Connecting Libraries and Research, Berlin, Germany: fihal-01660814f.
- Guerreiro, D., & Borbinha, J.L. (2014). Humanidades digitais novos desafios e oportunidades. *Revista Internacional del Libro, Digitalización y Bibliotecas*, 2(2), 63–75.
- Gil-Leiva, I. (2002). Consistence of document indexation involving novel indexers. *Anales de Documentación*, 5, 99–111.
- Hodge, G. (2000). Systems of knowledge organization for digital libraries: Beyond traditional authority files. Washington, DC: The digital Library Federation.
- Hojrland, B. (2008). What is knowledge organization (KO)? *Knowledge Organization*, 35(2/3), 86–101.
- Hughes, A.V., & Rafferty, P. (2011). Inter-indexer consistency in graphic materials indexing at the National Library of Wales. *Journal of Documentation*, 67(1), 9–32.
- International Standard Organization (2011). ISO 25964-1: Information and documentation—thesauri and interoperability with other vocabularies—part 1—thesauri for information retrieval. Geneva: ISO.
- Lancaster, F.W. (2004). Indexação e resumos: Teoria e prática. (2nd ed.). Brasília: Briquet de Lemos Livros.
- Lima, V.M.A., Costa, I.G., & Guimarães, M.O. (2017). A Organização do Conhecimento no domínio das Artes: o fazer terminológico na gestão do vocabulário controlado. Recife: Ed. UFPE.
- Lima, V.M.A., Rozestraten, A.S., Santos, C.A.C.M., Marques, E.A., & Sampaio, L.A. (2016). Arquigrafia: um repositório digital de imagens em ambiente colaborativo web. RBBB. *Revista Brasileira de Biblioteconomia e Documentação (Online)*, 12, 103–107.
- Moreiro González, J.A. (2011). Linguagens documentárias e vocabulários semânticos para a web: elementos conceituais. Salvador: EDUFBA.
- National Information Standards Organization. (2010). Guidelines for the construction, format, and management of monolingual controlled vocabularies. Baltimore: NISO.
- Osgood, C.E. (1990). The nature and measurement of meaning. In Osgood, C.E., & Tzeng, O.C.S. (Eds.). *Language, meaning and culture: The selected papers of C. E Osgood*. New York: Praeger Publishers.
- Rafferty, P. (2018). “Tagging”. *Knowledge Organization*, 45(6), 500–516.
- Rozestraten, A.S., Andrade, B.M., & Figueiredo, F.G. (2018). Manual de Procedimentos Técnicos do Projeto ARQUIGRAFIA. 2nd. Ed. São Paulo: FAU/USP.



- Rozestraten, A.S., Chou, A., Valentini, S., Gerosa, M.A., Reganati, G., Valente, C., & Claro, R. (2013). Smart audio city guide: Um sistema colaborativo para apoio ao deslocamento urbano de pessoas com deficiência visual. *Proceedings of the SBSC Conference* (pp. 175–178). Porto Alegre: Brazilian Computer Society.
- Rozestraten, A.S., Lima, V.M.A., & Santos, C.A.C.M. (2017). ARQUIGRAFIA: Digital images in the collaborative environment on the Web. *IFLA Satellite Meeting: Digital Humanities*. Berlin. Hague: IFLA.
- Rozestraten, A.S., Martinez, M.L., Gerosa, A.M.A., Kon, F., & Santos, A.P.O. (2010). Rede social ARQUIGRAFIA-Brasil: Estudos iconográficos da Arquitetura Brasileira na web 2.0. *Seminário Nacional de Documentação do Patrimônio Arquitetônico com o uso de Tecnologias Digitais*. Salvador: UFBA.
- Santos, L., & Santos, C.A.C.M. (2017). Arquigrafia: Um projeto de indexação de fotografias no meio digital, São Paulo. In: 25. *Simpósio Internacional de Iniciação Científica e Tecnológica da USP*. São Paulo: USP.
- Wölfflin, H. (1950). *Principles of art history: The problem of the development of style in later art*. New York: Dover Publications.



This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivs License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).



# Improving Archival Records and Service of Traditional Korean Performing Arts in a Semantic Web Environment

Ziyoung Park<sup>1†</sup>, Hosin Lee<sup>1</sup>, Seungchon Kim<sup>2</sup>, Sungjae Park<sup>1</sup>

<sup>1</sup>Department of Library & Information Science, Hansung University, Republic of Korea

<sup>2</sup>Department of IT Convergence Engineering, Hansung University, Republic of Korea

Citation: Park, Ziyoung, Hosin Lee, Seungchon Kim, and Sungjae Park. "Improving archival records and service of traditional Korean performing arts in a semantic Web environment." *Journal of Data and Information Science*, vol.5, no.1, 2020, pp. 68–80.

DOI: 10.2478/jdis-2020-0006

Received: Jan. 17, 2020

Revised: Mar. 17, 2020

Accepted: Mar. 25, 2020

## Abstract

**Purpose:** This research project aims to organize the archival information of traditional Korean performing arts in a semantic web environment. Key requirements, which the archival records manager should consider for publishing and distribution of gugak performing archival information in a semantic web environment, are presented in the perspective of linked data.

**Design/methodology/approach:** This study analyzes the metadata provided by the National Gugak Center's Gugak Archive, the search and browse menus of Gugak Archive's website and K-PAAN, the performing arts portal site.

**Findings:** The importance of consistency, continuity, and systematicity—crucial qualities in traditional record management practices—is undiminished in a semantic web environment. However, a semantic web environment also requires new tools such as web identifiers (URIs), data models (RDF), and link information (interlinking).

**Research limitations:** The scope of this study does not include practical implementation strategies for the archival records management system and website services. The suggestions also do not discuss issues related to copyright or policy coordination between related organizations.

**Practical implications:** The findings of this study can assist records managers in converting a traditional performing arts information archive into a semantic web environment-based online archival service and system. This can also be useful for collaboration with record managers who are unfamiliar with relational or triple database system.

**Originality/value:** This study analyzed the metadata of the Gugak Archive and its online services to present practical requirements for managing and disseminating gugak performing arts information in a semantic web environment. In the application of the semantic web services' principles and methods to an Gugak Archive, this study can contribute to the improvement of information organization and services in the field of Korean traditional music.



**Keywords** Gugak archive; Korean traditional music; Performing arts archive; Linked semantic data; K-PAAN

## 1 Introduction

This research project aims to organize the archival information of traditional Korean performing arts in a semantic web environment. Key requirements, which the archival records manager should consider for publishing and distribution of gugak performing archival information in a semantic web environment, are presented in the perspective of linked data. Specifically we analyze the characteristics of the archival records in the Gugak Archive of the National Gugak Center with the cooperation of the records managers in the archives. Based on the analysis, we make suggestions for improving the organization of the archival information in terms of identifiers, data schema, and authority data. Technical aspects of the semantic web and linked data were also examined for the web service of the performing arts archival records. This is because the records manager should to understand and accept the proposals of study. We hope that the records manager could extend the legacy archival records management process and actively request for the improvement of their informationa organization practices.

In Section 2, the scope of research is defined, and in Section 3, the methodology is presented based on the semantic web and linked data principles. Section 4 describes the characteristics of gugak, traditional Korean music, and the Gugak Archive's performing arts archives. Section 5 discusses the use of metadata and funcations of web services for searching and browsing of the performing arts information on the Gugak Archive website. Lastly, in Section 6, implications are derived from the analysis results.

## 2 Scope of analysis: Gugak archival records and online finding aids

This study analyzes the metadata provided by the Gugak Archive, the search and browse menus of Gugak Archive's website and K-PAAN, the performing arts portal site from the perspective of the semantic web and linked data. Traditionally, an archivist analyzed the bibliographic attributes and subject materials of the archival records and produced their metadata. He/she also designs the tools for providing metadata search and the user interface. Under the traditional finding aids, the metadata structure produced by an archivist was virtually identical to the search results provided to users. The assignment of headings (access points) to enable searching or the ordering of shelf arrangement was also at the archivist discretion.

However, with the widespread use of online catalogues published through websites, it is becoming increasingly difficult for an archivist to personally design



the search or browse functions and services to re-organize search results of online catalogue. Unlike in a traditional environment, with online search tools, the metadata structure stored in a server is not identical to what's on the users' screens, including the search results. Depending on the system, functions and interfaces—which are distinct from the original data structure's intended search and browse functions—may be added. When organizing archival information, it is important to consider how information is displayed on the user's web interface. Accordingly, this study extended the analysis of the Gugak Archive's metadata structure to include how data are provided in the online environment on both the Gugak Archive's own website and the Korea Performing Art Archives Network (K-PAAN), performing arts search portal.

### **3 Framework for analysis: Semantic web and linked data principle**

We share large amounts of information with others through web, and the senders and receivers of information vary depending on the type of information. Now, the most commonly used unit of information is the web page, which contains text, images, and videos. Every time we type a keyword into a search box, the search engine compares the query term with strings that were pre-extracted from web documents. A web search engine does not reveal the exact internal processes the search word goes through for thousands and even millions of hits to come up. The semantic web differs from the legacy worldwide web in many ways, mainly in the following characteristics (Antoniou et al., 2012):

- (1) make structured and semi-structured data available in standardized formats on the web;
- (2) make not just the datasets, but also the individual data-elements and their relations accessible on the web;
- (3) describe the intended semantics of such data in a formalism so that this intended semantics can be processed by machines.

The semantic web and linked data complement each other in terms of their purpose and usage. In the semantic web, for a machine (or robot) to access, understand, and analyze data, several technologies are required. These include three interconnected technologies: a resource description framework (RDF) data model to describe resources in triples, uniform resource identifiers (URI) for web resource identification, and an ontology for linking and inference of data. When datasets compliant with these technologies are created, data can be searched or arranged and then restructured using SPARQL. In other words, the semantic web publishes data in a way that the computer understands web documents, and when necessary, new documents can be produced.



Data built using semantic web technologies can be directly used as linked data without further processing. “Linked data” refers both to technologies for publishing certain web data and the data that are published using these technologies. In a technical context, “linked data” is a concept associated with semantic web technologies, which also include resource identifiers, such as URIs and IRIs, and RDF—the data structure. It is necessary to build an ontology, which provides the basis for data linking and inference, to increase the linked data’s usability. Equally important is interlinking, which expands data within the web space by interlinking URIs. The linked data principles outlined by Berners-Lee make it clear that they include the semantic web’s basic principles (Berners-Lee, 2006; Bizer, Heath, & Berners-Lee, 2009).

- (1) Use URIs as names for things
- (2) Use HTTP URIs so that people can look up those names
- (3) When someone looks up a URI, provide useful information, using the standards (RDF, SPARQL)
- (4) Include links to other URIs, so that they can discover more things

Brunetti et al. (2012) stated that most data exposed on the web are difficult to analyze or use as they are often in raw tabular form and argued that the use of linked data could improve data’s structure and semantics. They, however, added that it is not easy for end-users to intuitively understand the structure of a dataset that is published as linked data. As a matter of fact, to make a SPARQL query on exposed linked data, a user needs to understand the overall data structure, such as classes, attributes, and even values. Accordingly, for our Gugak Archive data and web services analysis, we used the four following criteria:

- (1) Whether URIs are assigned in such a way that archival records can be continuously identified and accessed on the web;
- (2) Whether the archival records’ metadata can be converted into an RDF format in a consistent way;
- (3) Whether the Gugak Archive website’s information can be provided with information about related performances or similar performances from other organizations; and
- (4) Any differences between Gugak Archive information provided through the global search portal and Gugak Archive’s own website.



#### 4 Gugak Archive and performing arts information management

“Gugak” is a term that refers to traditional Korean music. The National Gugak Center (NGC) was established in 1951 for the preservation and legacy of traditional

**Research Paper**

Korean performing arts. The NGC started its archiving project in 2007 and has emphasized the need to expand the Gugak Archive. The Gugak Archive began collecting and cataloging materials related to Korean classical performance, and in 2008, it began to develop a metadata and classification scheme for traditional arts archive. The Gugak Archive analyzes, collects, provides, and preserves Korean musical performances, ranging from traditional music, dance, plays, and Contemporary Korean music. At present, the Gugak Archive has produced and collected over 380,000 records, including videos, sounds, images, and textual materials. The NGC aims to provide Gugak-related information to the government and public sector, as well as foreign institutions (Kweon, 2016; NGC, 2019).

An important characteristic of performing arts is that they are not tied to specific media and are complete by and of themselves, by virtue of their expression. For this reason, while it is impossible to preserve performances as such, ironically, they tend to promote preservation through the two most common methods of recording—filming and photography. Moreover, the media environment’s recent advancements have opened more possibilities in using recorded performances (Reason, 2006; Lee, 2018).

However, searching for information about performing arts—a field where aesthetic qualities are important—using keywords is still difficult because they are often in image or video form rather than text. For the same reason, browsing through a performing arts archive by text like a book is not possible because the performance posters, flyers, and program brochures are often in image files with special design elements. Therefore, searching and browsing, in this case, requires information with detailed metadata. Rather than conducting known-item searches, users may prefer versatility in browsing records by inputting their artistic preferences, genres, or themes, which the keyword-based search function or vertical hierarchical classification cannot adequately support.

Linked data can address these limitations in searching and browsing performing arts records. Waitelonis and Sack (2009) tested whether using linked data in multimedia search can provide unexpected yet interesting results for users. Suppose a user’s query about a certain dancer’s videos had no results, but through the linked data’s link information, the website can suggest alternative videos featuring dancers with a similar style. This test making use of linked data’s key characteristics demonstrates that linked data produces greater results when there are abundant relationships between the data—which are defined in detail—than when the data exists in an isolated storage (like islands) or when there is a limited network of hierarchical or equal relationships between the data (like a tree’s structure).



## 5 Issues of using Gugak Archival information

### 5.1 Resource identification

When a new record is added into a library or an archive, it is assigned a shelf number or a call number. Likewise, when a new record is uploaded onto the web, it must be assigned a number according to a widely used and understood identification system in the web environment. Furthermore, these record numbers must be consistently managed so that they remain unchanged even when a website is redesigned or reorganized. URI-type numbers that are shown to users must be continuously valid to ensure uninterrupted user access to records. If the record numbers in a offline system can identify the collection to which a record belongs or describes a record's attributes, URIs provided on the web must have a similar or equivalent system.

Currently, information on traditional Korean performing arts and related performing arts archival records is provided not only in the NGC and Gugak Archive but also on APIs of the Cultural Data Plaza and integrated metasearch sites, such as K-PAAN. In this system, the Gugak Archive and the other sites create a link using a webpage URL, not permanent URIs. However, permanent URIs for the performing arts and related archival materials are needed. It is difficult to clearly identify and link the performing arts record on the web with a URL that only shows the page address. Also, in case of website renewal, the URL link could be broken. In order to improve the utilization of the archival information through the website, it is vital to give the URIs access to the archive's own data and provide the URI-granted data to external organizations.

Depending on policy or copyright management of the performing arts institution, there is a chance that the external connection to the archival information via direct URL will be blocked and provide encrypted URIs (do include). This must be improved because, if not, the data on the website may be isolated and users would be severely inconvenienced. It is also worth noting that when assigning a URI, a large number of performing arts archival materials are produced in a single performance. It is important to maintain the relationships between archival materials and the performance itself, too. Therefore, using permanent URIs to identify and link archival information among various web services is needed.

At present, we can estimate the data format of archival records on the web site through URL. Gugak metaterials are managed by item and clip unit in Gugak Archive. If the data format is video, audio, or image, describe the video, audio, and the image as a whole in the item level, and describe the individual performance recordings as a clip. Unlike live performance recordings, items such as leaflets, pamphlets, banners or programs are not expanded by clip. The URL of the Gugak archive



materials contains the clip number, so we can guess the data format by URLs. In this URLs, the item number is “V013358”, which contains the character of the data format, and the clip number is “20939”, which is the serial number excluding the data format designation.

[Example of Item URL]

- Korean Court Music: Royal Ancestral Ceremonial Music-Chon Pye Hee Mun (eng)
- “한국의 궁중음악: 의식음악-종묘제례악 중 전폐희문” (kor) [1966]
- URL for this item “V3358”:  
[http://archive.gugak.go.kr/portal/detail/searchVideoAiDetail?clipid=V013358&system\\_id=AI&recording\\_type\\_code=V](http://archive.gugak.go.kr/portal/detail/searchVideoAiDetail?clipid=V013358&system_id=AI&recording_type_code=V)  
 (or <http://archive.gugak.go.kr/portal/detail/searchVideoAiDetail?clipid=V013358>)

[Example of Clip URL]

- Korean Court Music: Royal Ancestral Ceremonial Music-Chon Pye Hee Mun - 01. Korean Court Music-Royal Ancestral Ceremonial Music ‘Chon Pye Hee Mun’ (eng)
- “한국의 궁중음악: 의식음악-종묘제례악 중 전폐희문 [1966 추정] - 01. 한국의 궁중음악-종묘제례악, 전폐희문” (kor)
- URL for this clip “20939” of the item “V3358”:  
[http://archive.gugak.go.kr/portal/detail/searchVideoDetail?clipid=20939&system\\_id=AV&recording\\_type\\_code=V](http://archive.gugak.go.kr/portal/detail/searchVideoDetail?clipid=20939&system_id=AV&recording_type_code=V)

In addition, within the Gugak Archive, items are grouped once more by folder numbers for management purposes. Until now, they are not reflected in browsing or searching on the website of Gugak Archive.

## 5.2 Data schema based on ontology

The Gugak Archive designed its own classification and metadata scheme to categorize and organize archival records (Noh, 2017), the “Traditional Arts Archive Classification Scheme,” which is linked to the Korean Decimal Classification (KDC). The KDC is widely used in Korean library classification, with the entry “679 Korean Traditional Music” used to link two classification schemes. The Gugak Archive also has its own descriptive metadata standard, “Traditional Arts Archive Metadata,” which is based on International Standard Archival Description (ISAD (G)) and Dublin Core Metadata. It may not be scalable in the semantic web environment, even though a descriptive standard, such as ISAD (G), can guarantee interoperability in the records management community. Therefore, upper ontology



is needed to exchange and share data with various fields and institutions. ISAD (G) was also developed as an XML-based encoding schema, Encoded Archival Description (EAD) and considered as a basis for archival linked data (Tillman, 2018). There is also a research project that builds linked data based on the EAD such as Locah project (LOCAH Linked Archives Hub (2013)).

Physically and Logically organized performing arts records have been linked with other units of related records through vertical and horizontal relationships. Link information between records, which need to be maintained for searching and browsing through performing arts records, can be given various forms such as a network structure established through RDF-type expressions and linking methods. Moreover, link information can be created between the Gugak Archive's records and other institutions' performing arts records using linked data techniques. Linked data makes it possible for archival information to be more systematically and effectively provided.

In Archival Science, a study has been carried out to map the Encoded Archival Description, an XML version of ISAD (G), to CIDOC CRM (Bountouri & Gergatsoulis, 2011; Park, 2018). For example, a part of the current descriptive elements can be mapped into the CIDOC CRM class and attributed as follows.

Descriptive Elements	Matching to Upper Ontology
Title	→ E19 Physical Object (P102 has title) <u>E35 Title</u> ( <u>The name</u> of an archival resource)
Time	→ E19 Physical Object (P4 has time-span) <u>E2 Time-Span</u> ( <u>The time</u> when an archival resource is created)
Place	→ E19 Physical Object (P55 has current location) <u>E53 Place</u> ( <u>The name</u> of a place where an archival resource is created)
Type of Records	→ E19 Physical Object E55 Type {Fonds, Series, Files, Items} ( <u>The name</u> of a type of material)

### 5.3 Interlinked performing arts information network design

In the semantic web environment, the link between resources with URIs, or those that interlink, increases data scalability and quality. To increase interlinked data, the value of descriptive elements must have URIs, not strings. This requires authority-controlled vocabularies. In the Gugak Archive, the vocabulary and meaning of Korean traditional music can be controlled using a Gugak dictionary and thesaurus. However, at present, mainly uncontrolled keywords are listed in the archival description of the Gugak Archive.

At present, the Gugak Archives provide detailed classification scheme and keywords (tags) on Korean traditional performances. The Korean traditional music



Research Paper

classification scheme, based on KDC, is marked with headings instead of class numbers. The gugak thesaurus is also built and provided on the Gugak archive website. Thesaurus has synonym clusters centered on descriptors, and provides BT, NT, and RT relationship information. The Korean traditional thesaurus consists of key terms that describe Korean traditional music in detail.

However, in the Gugak Archive website, the classification scheme and the thesaurus does not used sufficiently for navigation or reciprocal linking of Gugak performance records. The below is the screen of the Gugak thesaurus for the term, “Jongmyo rite music”. The term cluster is formed around the descriptor.



Figure 1. Gugak Thesaurus term “Jongmyo jeryeak” <<https://archive.gugak.go.kr/portal/division/thesaurusMain>>.

However, the term relationships in the thesaurus are not yet applied to the Gugak Archive website when searching and browsing gugak performing records. On the screen for the individual records, classification information and keywords are displayed only. We cannot navigate the gugak term network by clicking the headings in the classification information. The figure below is an example of gugak record display. There is no link in the classification headings. If you click on a keyword, only the same keyword can be searched or not. In linked data, individual headings or index terms can be semantically linked and searched in conjunction with URIs.

In particular, performance works, performances, and performing arts records need to be interlinked to structure performing arts information. Because most works are performed several times, it is important to link them with the higher-level unit “container work” as well, rather than just interlinking individual performances and performance records. It is also necessary to clearly distinguish performance works from individual works contained within a certain work and manage them separately.



## Classification information

**Court music > Religious music > Jongmyo jeryeak > Botaepyeong > Chon Pye Hee Mun [ENG]**

정악>종교음악>종묘제례악>보태평>전폐희문

Korean Court Music: Royal Ancestral Ceremonial Music-Chon Pye Hee Mun [1966 추정] - 01. 한국의 궁중음악-종묘제례악'전폐희문' (Korean Court Music 'Chon Pye Hee Mun')

CHON PYE HEE MUN

ROYAL ANCESTOR'S MUSIC

ORCHESTRA ON THE TERRACE

클립 1 건

01. 한국의 궁중음악-종묘제례악' (Korean Court Music-Royal Ancestral Ceremonial Music 'Chon Pye Hee Mun') - 동영상

재생 시간 : 0:00:11

#1966

#로버트가피어스

#로버트가피어스

#R. Garfias

#희귀영상

#Korean Court Music

#종묘제례악

#보태평

#전폐희문

## Keywords

#1966, #Korean Court Music, #R.Carfias...

Figure 2. Gugak archival records “Jongmyo jeryeak” <<http://archive.gugak.go.kr/portal/detail/searchVideoDetail?clipid=20939>>.

Instead of simply entering search keywords related to individual works into the system, the data must be modeled in a way that treats them as separate works by also assigning a unique URI so that they may be linked to various authority data.

## 5.4 Browsing and searching Gugak archival information

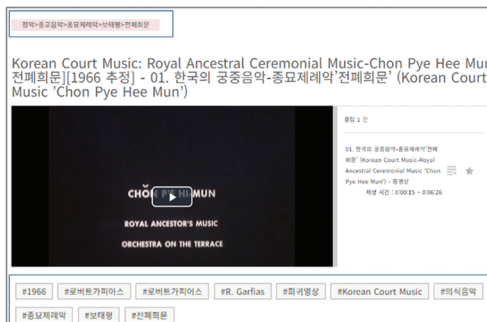
The Gugak Archive’s information is not only distributed through its own website but also through K-PAAN—a performing arts search portal. Besides Gugak Archive, K-PAAN also includes information from the Museum of Performing Arts, Arko Arts Archive, and the National Intangible Heritage Center, making it unrealistic to redesign the portal according to the Gugak Archive’s requirements. Moreover, K-PAAN does not have its own identification system for the performing arts item as it indexes these organizations’ data and performs a parallel search to display the results by organization successively. When a user clicks on a K-PAAN search result,



## Research Paper

they are redirected to the source website because the portal does not have or manage its own link information on these four organizations' information. Going forward, for K-PAAN to provide LOD search services, it will need to implement required elements such as URIs, RDF, and interlinking information.

In addition, the classification and keyword information provided on the website of Gugak Archives are not equally reflected in K-PAAN. There is a difference in the unit of classification of records among the institutions that make up K-PAAN service, which may not be distinguished on the portal performing arts archive website. To improve this, the records manager and the website architecture should cooperate concretely the building criteria for the structure of records provided to K-PAAN. In most cases, the records managers of individual institutions have less authority over data structure or services of the web site in the case of portal sites than their own websites. It is also convenient for users to use the portal site as a starting point for searching. Therefore, the quality control of the portal site should also be considered.



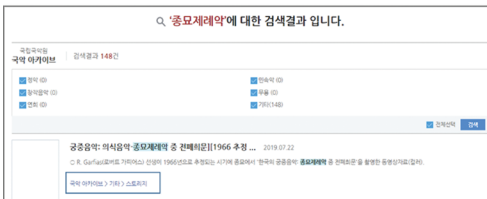
Korean Court Music: Royal Ancestral Ceremonial Music-Chon Pye Hee Mun  
전폐회문[1966 추정] - 01. 한국의 궁중음악-종묘제례악 '전폐회문' (Korean Court Music 'Chon Pye Hee Mun')

URL for this video clip:  
<http://archive.gugak.go.kr/portal/detail/searchVideoDetail?clipid=20939>

Detailed classification for this video clip:  
정악 > 종교음악 > 종묘제례악 > 보태평 > 전폐회문 [KOR]  
Court music > Religious music > Jongmyo jeryeak > Botaebyeong > Chon Pye Hee Mun [ENG]

Tags for this video clip:  
#1966, #Korean Court Music, #R.Carfias...

Figure 3. Individual archival records in Gugak Archive website <<http://archive.gugak.go.kr/portal/detail/searchVideoDetail?clipid=20939>>.



국악 아카이브 검색결과 148건

No fixed URL for this video clip  
Classification information become blurred  
국악 아카이브 > 기타 > 스토리지 [KOR]  
Gugak Archive > Etc. > Storage [ENG]

No tags for this video clip:

Figure 4. Individual archival records in K-PAAN website. <<https://www.iha.go.kr/k-paan/main>>.

## 6 Discussion

The importance of consistency, continuity, and systematicity—crucial qualities in traditional record management practices in a library setting—is undiminished in a semantic web environment. However, a semantic web environment also requires new tools such as web identifiers, data models, and link information. We therefore propose the following suggestions to improve the openness and quality of the traditional Korean performing arts record in the semantic web environment:

- Identify traditional Korean performing arts and related archival records with URIs permanently and consistently.
- Design data schema for archival records based on the formal ontological point of view for data linking and sharing among various institutions.
- Refine name and subject authority data to strengthen the linkages between archival data, as well as interlinking with external web data.
- The records manager should be aware of the difference between providing archival services on individual websites and providing services on performing arts search portal, and design and support the user's search and browsing experience.

## Acknowledgments

We wish to give our special thanks to Dasom Jung, Seunghee Son, and Yoonwhan Kim for their help and support throughout this project. This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2016S1A5A2A03927725).

## Author contributions

Ziyoung Park (zgpark@hansung.ac.kr) led the analysis of gugak archival records and web services, and write a draft of the paper. Hosin Lee (leehs@hansung.ac.kr) was responsible for the analysis of the performing arts information and review the paper. Seungcheon Kim (kimsc@hansung.ac.kr) was in charge of open data system requirements and review the paper. Sung Jae Park (sungjae.p@gmail.com) was in charge of the quality control of linked data and review the paper.

## References

- Antoniou, G., Groth, P., Harmelen, F.V., & Hoekstra, R. (2012). *A Semantic Web Primer*, 3<sup>rd</sup> edition. Cambridge/London: The MIT Press.
- BelliniPaolo, P., & Nesi, P. (2013). A linked open data service for performing arts. In Nesi, P., & Santucci, R. (Eds.) *Information Technologies for Performing Arts, Media Access, and Entertainment: proceedings of Second International Conference, ECLAP 2013, Porto, Portugal, April 8–10, 2013, Revised Selected Papers* (pp. 13–25). Berlin, Heidelberg: Springer.



## Research Paper

- Berners-Lee, T. (2006). Linked Data—Design Issues. Retrieved from: <https://www.w3.org/DesignIssues/LinkedData.html>.
- Bizer, C., Heath, T., & Berners-Lee, T. (2009). Linked data: The story so far. *International Journal on Semantic Web and Information System*, 5(3), 1–22.
- Bountouri, Lina & Gergatsoulis, Manolis. (2011). Mapping encoded archival description to CIDOC CRM. *Advances in Water Resources—ADV WATER RESOUR*.
- Brunetti, J.M., Rosa, G., Gimeno, J.M., & Garcia, R. (2012). Improved linked data interaction through an automatic information architecture. *International Journal of Software Engineering and Knowledge Engineering*, 22, 325–343.
- Kweon, Hyekyung. (2016). Current status and future of Gugak archive for collecting and providing Gugak information resources. In *proceedings of the Seminar on Gugak Archive: Role and Value Creation in the Digital Age*, Performing Arts Archive, Seoul, Korea.
- Lee, H. (2018). Performing arts videos: Beyond the record. *Journal of National Gugak Center*, 40, 59–79. <http://doi.org/10.29028/JNGC.2018.37.059>
- LOCAH Linked Archives Hub. (2013). Retrieved from <http://data.archiveshub.ac.uk/>
- National Gugak Center. (2019). Introducing Gugak Archive. <https://archive.gugak.go.kr/portal/main/introduction>
- Noh Yea-ri. (2017). The preservation and utilization of Korean traditional music records—Focusing on the records housed at Gugak Archive. *The Society for Korean Historico-Musicology*, 59(0), 43–82.
- Park, Z. (2018). An exploratory study on linking ISAD(G) and CIDOC CRM using KARMA. *Korean Society of Archives and Records Management*, 18(2), 189–214.
- Park, Z. (2019). Performing arts information: User-friendly knowledge publication and service design on the Web: With focus on linked semantic data. In *proceedings of 2019 Gugak Archive Academic Seminar*. Seoul: National Gugak Center.
- Park, Z., Lee, H., Kim, S., Park, S., Jung, D., Son, S., & Kim, Y. (2019). Improving archival records of traditional Korean performing arts in a semantic web environment. Presentation at the DCM2019. September 23–26, 2019. Seoul, Korea.
- Park, Z., Lee, H., Kim, S., Park, S., Jung, D., Son, S., & Kim, Y. (2020). Organizing performing arts records of Korean traditional music in a semantic web environment—Focusing on the case of Gugak archive. In *proceedings of 16th ISKO Conference*, Aalborg, July 2020. Baden-Baden: Ergon. [In progress].
- Reason, M. (2006). *Documentation, disappearance and the representation of live performance*. New York: Palgrave Macmillan.
- Tillman, R.K. (2018). Opportunities for encoding EAD for linked data extraction and publication. *Journal of Archival Organization*, 19, 19–36.
- Waitelonis, J., & Sack, H. (2009). Towards exploratory video search using linked data. In *proceedings of 11th IEEE International Symposium on Multimedia* (pp. 540–545). San Diego, CA, USA.
- Wikipedia. (2019). Music of Korea. [https://en.wikipedia.org/wiki/Music\\_of\\_Korea](https://en.wikipedia.org/wiki/Music_of_Korea)



# “SEMANTIC” in a Digital Curation Model

Hyewon Lee<sup>1†</sup>, Soyoung Yoon<sup>2</sup>, Ziyong Park<sup>2</sup>

<sup>1</sup>Department of Library & Information Science, Seoul Women’s University, Seoul, Republic of Korea

<sup>2</sup>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

Citation: Lee, Hyewon, Soyoung Yoon, and Ziyong Park. “SEMANTIC” in a digital curation model.” *Journal of Data and Information Science*, vol.5, no.1, 2020, pp. 81–92  
DOI: 10.2478/jdis-2020-0007  
Received: Jan. 18, 2020  
Revised: Mar. 6, 2020  
Accepted: Mar. 10, 2020

## 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



<sup>†</sup> 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”



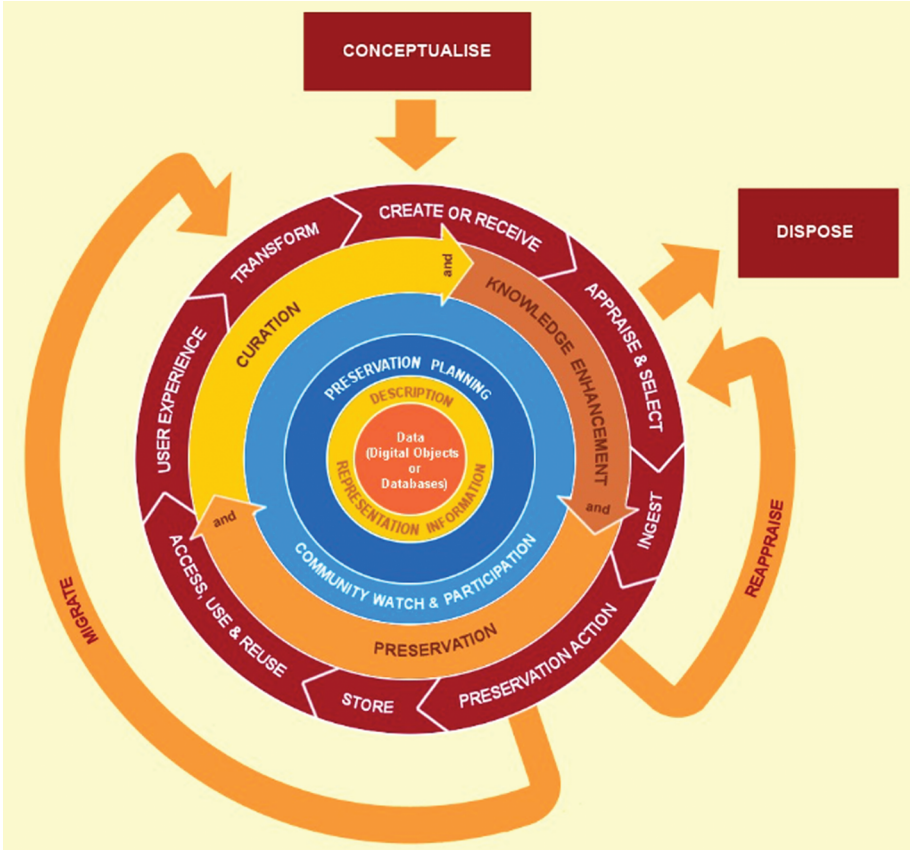


Figure 1. The DCC&U Curation Lifecycle Model (Constantopoulos et al., 2009, p.42, Fig. 3).

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

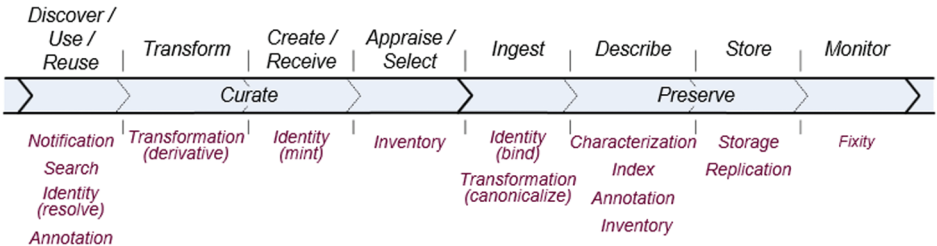


Figure 2. UC3 Micro Services (from CDL, 2010, p.17, Fig. 7).

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).

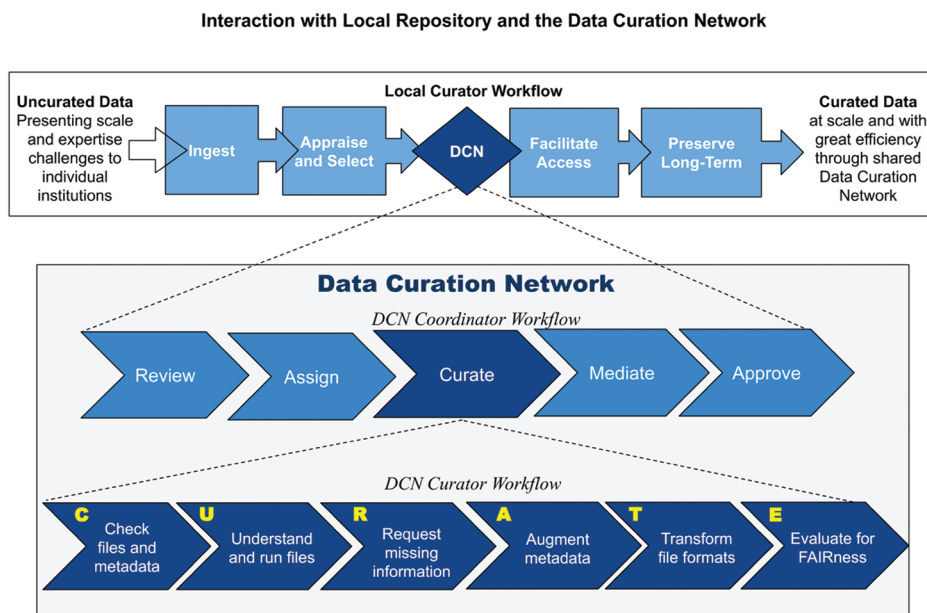


Figure 3. Curation Workflow for the Data Curation Network (Johnston et al., 2017, p.29, Fig. 8).

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”



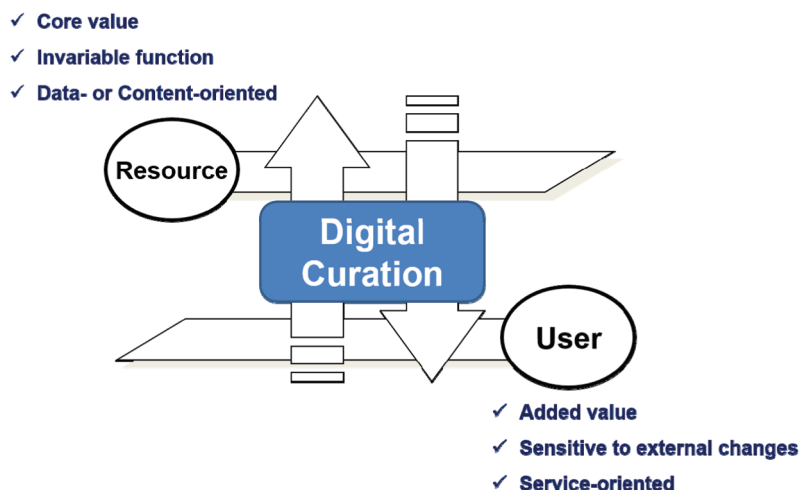


Figure 4. Connection of resource and user through digital curation (modified by Lee et al., 2019, p.218, Fig. 9).

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

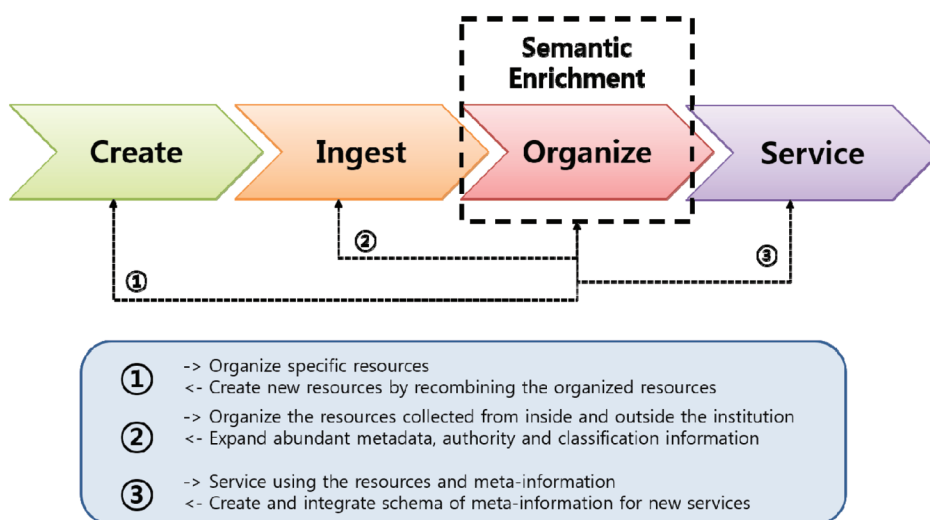


Figure 5. Semantic Enrichment’s functions.



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).

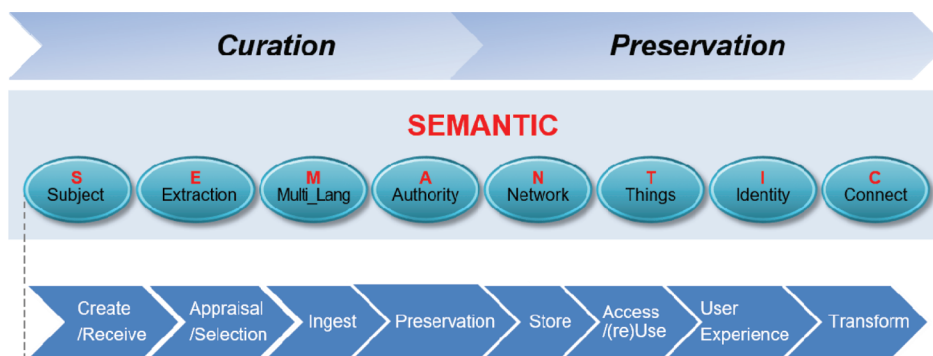


Figure 6. “SEMANTIC” in a digital curation model.

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.

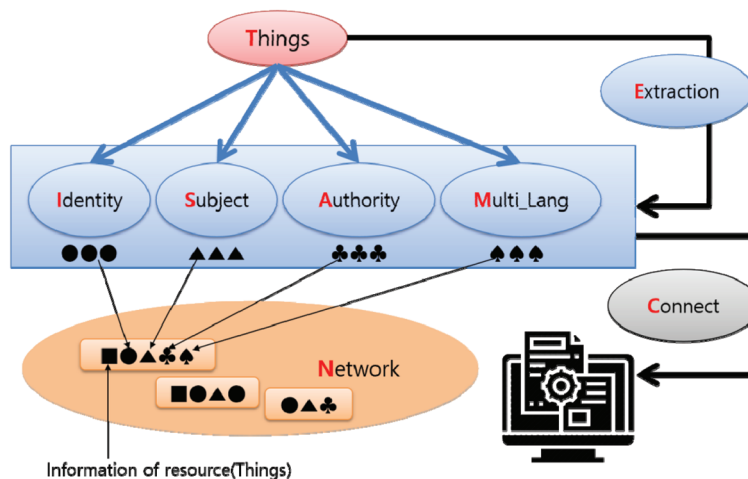


Figure 7. SEMANTIC elements and their relationships (Lee et al., 2019, p.220, Fig. 11).

“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

We wish to extend our special thanks to the Korea Institute of Science and Technology Information (KISTI) Curation Center for the help and support they provided throughout this project. This research was supported by a research grant from Seoul Women’s University (2020). This research was financially supported by Hansung University.

## 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.



## 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., Gavrili, 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](http://www.ukoln.ac.uk/ukoln/staff/m.pennock/publications/docs/lib-arch_curation.pdf)



This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivs License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).



# FAIR + FIT: Guiding Principles and Functional Metrics for Linked Open Data (LOD) KOS Products

Marcia Lei Zeng<sup>†</sup>, Julaine Clunis

School of Information, Kent State University, Ohio, USA

## Abstract

**Purpose:** To develop a set of metrics and identify criteria for assessing the functionality of LOD KOS products while providing common guiding principles that can be used by LOD KOS producers and users to maximize the functions and usages of LOD KOS products.

**Design/methodology/approach:** Data collection and analysis were conducted at three time periods in 2015–16, 2017 and 2019. The sample data used in the comprehensive data analysis comprises all datasets tagged as types of KOS in the Datahub and extracted through their respective SPARQL endpoints. A comparative study of the LOD KOS collected from terminology services Linked Open Vocabularies (LOV) and BioPortal was also performed.

**Findings:** The study proposes a set of Functional, Impactful and Transformable (FIT) metrics for LOD KOS as value vocabularies. The FAIR principles, with additional recommendations, are presented for LOD KOS as open data.

**Research limitations:** The metrics need to be further tested and aligned with the best practices and international standards of both open data and various types of KOS.

**Practical implications:** Assessment performed with FAIR and FIT metrics support the creation and delivery of user-friendly, discoverable and interoperable LOD KOS datasets which can be used for innovative applications, act as a knowledge base, become a foundation of semantic analysis and entity extractions and enhance research in science and the humanities.

**Originality/value:** Our research provides best practice guidelines for LOD KOS as value vocabularies.

**Keywords** Knowledge Organization Systems; Linked Open Data; FAIR; FIT; Semantic web

Citation: Zeng, Marcia Lei, and Julaine Clunis. "FAIR + FIT: Guiding principles and functional metrics for Linked Open Data (LOD) KOS products." *Journal of Data and Information Science*, vol.5, no.1, 2020, pp. 93–118.

DOI: 10.2478/jdis-2020-0008

Received: Jan. 18, 2020

Revised: Mar. 10, 2020

Accepted: Mar. 16, 2020



<sup>†</sup> Corresponding author: Marcia Lei Zeng (E-mail: mzensg@kent.edu).

## 1 Introduction

Semantic technology standards advanced in the Semantic Web era have enabled open access to well-structured and well-curated Linked Open Data (LOD) datasets. Among the most useful LOD products are the Knowledge Organization Systems (KOS) that were originally published as thesauri, classifications, taxonomies, name authorities, or picklists and now are available as LOD datasets. At the tenth anniversary of the W3C (2009) formal recommendations Simple Knowledge Organization System (SKOS) and SKOS eXtension for Labels (SKOS-XL), the number of KOS datasets available through open data registries (e.g. Datahub<sup>①</sup>, BioPortal<sup>②</sup>, and Linked Open Vocabularies (LOV)<sup>③</sup>) reached nearly two thousand. We use the umbrella term “LOD KOS” to refer to all these value vocabularies (distinguishable from the “property vocabularies”) and lightweight ontologies (not the same as “reference ontologies”) within the Semantic Web framework (Zeng & Mayr, 2018). These value vocabularies are invaluable engines for all 5-star LOD datasets, as can often be seen in the LOD Clouds<sup>④</sup>. The objective of the project reported in this article is to develop a set of metrics and identify criteria for assessing the functionality of LOD KOS products while providing common guiding principles that can be used by LOD KOS producers, publishers, and users to maximize the functionality and added values for LOD KOS.

The “FAIR Guiding Principles for scientific data management and stewardship” (Wilkinson et al., 2016) provides guidelines for the publication of digital resources such as datasets, code, workflows, and research objects, in a manner that makes them Findable, Accessible, Interoperable, and Reusable (FAIR). The FAIR principles<sup>⑤</sup> have been widely implemented in the open data environment, in an effort to achieve FAIRification of certain types of data metrics, conduct FAIRness assessment of specific datasets, and turn FAIR into reality (FORCE11, 2014; European Commission Expert Group on FAIR Data, 2018; Wilkinson et al., 2018).

Consequently, we were inspired to implement and assess FAIR principles for KOS open datasets and in particular to assemble relevant functional metrics for a specific type of product—the LOD KOS vocabularies. This paper reports a set of metrics developed through a comprehensive data analysis and a comparative study. The project extends a previous study of “KOS in the Semantic Web” (Zeng & Mayr, 2018) which examined the functions of LOD KOS based on a set of collected cases



<sup>①</sup> <http://datahub.io> and <https://old.datahub.io/>

<sup>②</sup> <https://bioportal.bioontology.org/>

<sup>③</sup> <https://lov.linkeddata.es/dataset/lov>

<sup>④</sup> <https://lod-cloud.net/>

<sup>⑤</sup> <https://www.go-fair.org/fair-principles/>

through the viewpoints of LOD dataset producers, KOS vocabulary producers, and researchers who are the end-users of LOD KOS. The study reveals the remarkable potential of LOD KOS while also highlighting obstacles and issues. It is believed that the main barrier for maximizing the usage of LOD KOS resides in communication about these KOS through their delivery services rather than their structure, format, or contents. Undoubtedly, common guiding principles and assessment metrics are needed for the LOD KOS. By using these metrics, any assessment performed on LOD KOS products can lead to actions addressing any or all identified issues case-by-case, collectively or independently.

Our research has also confirmed that, semantic technologies have brought KOS vocabularies into a new era with many technologically advanced use cases made possible. KOS' functions have been extended far beyond being controlled vocabularies and taxonomies. They have become knowledge bases, the trustable resources for knowledge graphs, and fundamental components for the contextualization of data-driven and AI-dominated processes. It is essential that the LOD KOS be measurable by developers and users for enhanced and effective usage, while encouraging innovative approaches for the LOD KOS to be FAIR (Findable, Accessible, Interoperable, and Reusable) as open data, plus to be FIT (Functional, Impactful, and Transformable) as value vocabularies.

## 2 Methodology

In this section, we present the methodology of an investigation designed to gather descriptive data of LOD KOS datasets. This study involved a series of steps for collecting and analyzing LOD KOS datasets to assess their functionality. Data collection and analysis were conducted at three time periods in 2015–16, 2017, and 2019. The sample data comprises all datasets tagged as kinds of knowledge organizations systems in the Datahub ([www.datahub.io](http://www.datahub.io) and [www.old.datahub.io](http://www.old.datahub.io)) which is a data management platform from Open Knowledge International. The first review of the released LOD KOS products in 2015 helped us to narrow down targeted research topics and discover significant challenges. The scope of the study was extended to BioPortal and LOV after the FAIR principles became mainstream.

The steps described below involve data collected through the SPARQL endpoints of LOD KOS products registered in the Datahub. We further analyzed and recorded the presence of certain features and extracted LOD KOS properties used via a SPARQL query.

**Step 1.** To search for qualified KOS datasets to study, we used the following terms provided and used as tags by providers in the Datahub: *authority file*, *list*, *terminology*, *thesaurus*, *taxonomy*, *classification*, *classification scheme*, and *ontology*. The results were manually transcribed and stored in a spreadsheet (Refer



to Table 1 in Section 4.1). The data then went through data cleaning processes for removing duplicates.

**Step 2.** To evaluate the structure and content of the datasets in detail, verifications were performed manually to ensure that the datasets found aligned with the definition of each type of KOS. In some cases, we found that providers tagged their products with terms such as “terminology” or “taxonomy”. However, the datasets tagged as terminology may just have been a list of terms not crafted for information retrieval purposes, arranged in no order, or have no definitions. In the case of those tagged as taxonomies, some datasets were neither groups of objects based on any particular characteristics as commonly understood, nor did they have any kind of hierarchical arrangement. After validation, those datasets that did not fit the commonly understood definitions of various types of KOS were removed from consideration. Therefore, as showing in Figure 2, datasets “found” refers to all datasets that were tagged as a kind of KOS showed at the initial search, while “verified” refers to those that were checked and align with commonly understood definitions of these terms. (Refer to Figure 2 in Section 3.3 for the initial search and verified results.)

**Step 3.** To study those being confirmed as real KOS vocabularies in each of the categories, the names and features of all datasets were documented in a spreadsheet. For each dataset, we checked and recorded whether the SPARQL API was available and working, was moved, unavailable or returned a type of error message. (Refer to Table 1 in Section 4.1.)

For each dataset, additional features of the SPARQL API are also recorded, including: (1) the name of the editor facilitating deployment of the SPARQL endpoint (e.g. Virtuoso, Fuseki, PoolParty, etc.); (2) the default query, if available; (3) the enabled operations (e.g. SELECT, CONSTRUCT, ASK, DESCRIBE); (4) the maximum number of possible results from queries; (5) the enabled HTTP methods; (6) the available formats in which results can be downloaded; and (7) the number of example queries provided. (Refer to Table 2 and 3 in Section 4.1.)

**Step 4.** To investigate the datatype properties of each dataset, a specific query was run at each endpoint across all datasets to extract the first 100 properties (*SELECT DISTINCT ?p WHERE {?s ?p ?o} LIMIT 100*). This query selects 100 distinct properties from the triples in the dataset. The results in an HTML table format were dumped into a spreadsheet for further analysis.

A complete assessment of the datasets’ interoperability level requires an examination of the property vocabularies used. We created a list of common vocabularies, including Dublin Core, FOAF, RDFS, OWL, SKOS, DBPedia, Schema.org, etc. The use of standard vocabularies was recorded along with the occurrence of specialized and locally created terms. Extracted element properties were then color-marked to provide an overview of the sources of the properties



reused by a dataset. Note the frequency of non-standard/local terms compared to standard controlled terms as evidenced by the color codes. (Refer to Figure 3 and Figure 4 in Section 4.1.)

**Step 5.** To give a simple quantitative analysis of the collected data, we obtained counts of the collected datasets for each KOS type, the frequencies of various property elements, as well as an understanding of how many KOS overall use those properties. Furthermore, the number of occurrences of various document formats was captured. Finally, the number of endpoints that provide SPARQL query examples or templates for exploration of the dataset was also assessed.

**Step 6.** In addition to the data collected from Datahub, another data collection and analysis was added in 2019, aimed at gathering real cases that demonstrate areas that could be enhanced. The data were collected from BioPortal and LOV.

Beside the processes mentioned above, a parallel project “KOS in the Semantic Web” has been running concurrently, focusing on case studies and content analysis of the LOD KOS products. Many specific cases have been traced and studied in order to discover best practices and innovative approaches in KOS creation, connection, production, and transformational usages. In the following sections, selected cases will be referenced from our reports and publications from this “KOS in the Semantic Web” project.

### 3 Research findings and recommendations for LOD KOS as open datasets: FAIR

As explained by the European Commission Expert Group on FAIR Data (2018), “FAIR” or “FAIR data” should be understood as shorthand for a concept that comprises a range of scholarly materials that surround and relate to research data. This includes the algorithms, tools, workflows, and analytical pipelines that lead to creation of the data and which give it meaning. It also encompasses the technical specifications, standards, metadata, vocabularies, ontologies and identifiers that are needed to provide meaning, both to the data itself and to any associated materials. We recommend that LOD KOS follow the FAIR principles, to improve findability, accessibility, interoperability, and to enable reuse of any digital assets owned.

Since FAIR Guiding Principles have been studied and reported in multiple locations, we will not repeat the explanations and implementation cases in this paper. Instead we emphasize situations which align with FAIR and can be enhanced with FAIRification. Each of the FAIR principles is presented and described below with our findings which come from two main sources.

- First, in Figure 1, various descriptive elements used on the Datahub for the datasets are highlighted, found in various views, they indicate how the LOD KOS have been registered/described and made available in the Datahub.



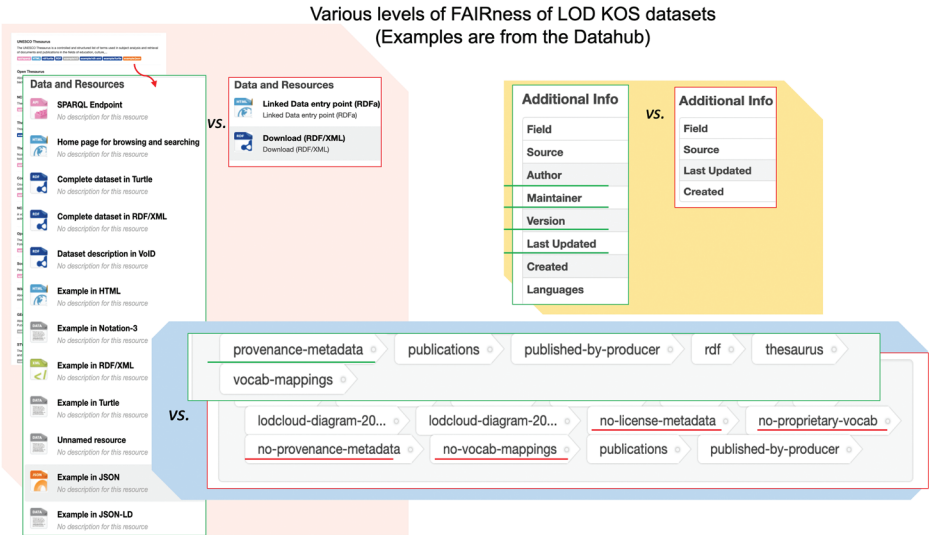


Figure 1. Various levels of FAIRness of LOD KOS datasets as seen from the Datahub.

- Second, the properties describing the LOD KOS collected through their own SPARQL endpoints have been captured and analyzed.

### 3.1 Findable

Findable requires that, metadata and data should be easy to find for both humans and computers.

Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an indispensable component of the FAIRification process. Among the “F” approaches, when considering the LOD KOS datasets registered in the Datahub; we found that the metadata provided by those providers could benefit from being enriched. Furthermore, while many met requirements for “F”, some missed basic information about the KOS dataset, such as languages, creators, and history. (See example on the upper right of Figure 1.) Thus, we have a specific additional recommendation for Findable:

- Enrich metadata about KOS as much as possible to enable data discovery processes.

### 3.2 Accessible

Accessible requires that, once the user finds the required data, she/he needs to know how they can be accessed, possibly including authentication and authorization.



Access to a LOD KOS includes various paths: query access, entity-level access, and access via data dump. There are a wide range of types or formats in which a KOS vocabulary dataset can be delivered. A highly accessible KOS would provide not only a SPARQL endpoint for query access, but also common entity-level browsing and searching access. It would be ideal to deliver datasets in certain RDF serialization formats, and provide examples in varying formats (e.g. left in Figure 1). Yet, in reality, some KOS products only have one mode of downloadable access. This leads to a specific additional recommendation for Accessible:

- Provide multiple pathways for accessing the KOS datasets.

### 3.3 Interoperable

Interoperable requires that the data usually needs to be integrated with other data. In addition, the data needs to interoperate with applications or workflows for analysis, storage, and processing.

The preliminary findings of this study reveal that, the metadata that have been used in describing the vocabulary types vary at different registries. For example, the way vocabulary types are categorized in the Datahub is un-standardized, even though the terms to use are suggested (refer to Figure 2).

The situation is similar in other KOS registries. The tags could be mis-used when indicating the types of KOS vocabularies. One can imagine the amount of time spent on verifying them before any mapping. The issue can be resolved by applying the *KOS Types Vocabulary*<sup>®</sup> generated by the DCMI NKOS Task Group which also impacts the KOS' findability and reusability. Thus, our additional recommendation for Interoperable:

- Utilize the *KOS Types Vocabulary* to standardize the way vocabulary types are categorized thereby supporting mapping and interoperability.

Metadata about the LOD KOS found at individual endpoints showed that many have used Dublin Core Metadata Element Set<sup>®</sup> and/or Dublin Core Metadata Terms<sup>®</sup>, (49/135 in 2016, 44/140 in 2017, and 66/160 in 2019). Some LOD KOS indicate whether vocabulary mapping exists for a particular vocabulary. This is important since the interoperability of KOS products involves more than the metadata that describes the dataset. We discuss this key principle further in Section 4.1 and 4.2. (Refer to the findings in Figure 3 and Figure 4 in Section 4.)



<sup>®</sup> <https://nkos.slis.kent.edu/nkos-type.html>

<sup>®</sup> <https://www.dublincore.org/specifications/dublin-core/dces/>

<sup>®</sup> <https://www.dublincore.org/specifications/dublin-core/dcmi-terms/>

Initial search and verified results of types of KOS (2019)		
Search Type of KOS/DATASET	# found (initial)	# found (verified)
Authority Files	164	18
List	825	71
Terminology	39	35
Thesaurus	80	91*
Taxonomy	37	22
Classification	478	43
**Ontology	531	266
Totals	1623 (+531 ontologies)	280 (+ 266 ontologies)
<p>* A thesauri marked as an ontology is considered as thesauri.  **An ontology is designed as an ontology, not a product converted from another type of KOS.</p>		

Figure 2. Initial search and verified results of types of KOS.

### 3.4 Reusable

Reusable indicates that the ultimate goal of FAIR is to optimize the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.

License and provenance metadata are critical to the dissemination of data. There has been a paradigm shift in the digital age for the publication of KOS products. Before the Web, authorized and relevant versions of a thesaurus or taxonomy were clear due to the existence of a single release source. However today, multiple versions, formats, and locations for a single KOS vocabulary can be found ranging from formally updated releases to project-based releases at various hubs or registries. Deciding which source to use is in the hands of the users and not the producers of a vocabulary. Lack of license and provenance metadata may cause confusion and negatively impact the KOS, especially those being constantly updated and with high quality control. Therefore, as alluded to in Figure 1 (those items listed as not present, e.g. no-license-metadata), to enable dataset reusability we additionally recommend for Reusable:



- Adequately supply license and provenance metadata to enable datasets' reusability.

## 4 Research findings and recommendations for LOD KOS as value vocabularies: FIT

The main aim of FAIR principles is to enable and advance the reuse of data. All data is not created equally and as such we wanted to examine KOS data in light of these principles. We discovered that while the FAIR principles could significantly improve the quality of these datasets, additional considerations were necessary for KOS as value vocabularies. This section presents the main research findings for LOD KOS as value vocabularies with a set of functional metrics and recommendations. We use the term “metrics” in the sense of key performance indicators which can be used to assess the efficiency, performance, and quality of LOD KOS datasets. Tiemensma (2010) discusses this usage of the term and describes it as a shift from measuring what you can count to measuring what counts. As such, we have identified three critical indicators here defined: Functional, Impactful, and Transformable, and coined the acronym FIT to reference them. These recommendations are supported by our research findings.

As introduced at the beginning of this article, the FAIR guiding principles inspired us to assemble relevant functional metrics for this specific type of product—the KOS vocabularies, in addition to implementing FAIR for KOS as open datasets. Our “KOS in the Semantic Web” study (Zeng & Mayr, 2018) highlighted the need for common guiding principles and assessment metrics for the LOD KOS and further that these should be used by LOD KOS producers and users to maximize the functions and usages of LOD KOS products and enable better research and application development utilizing them. In the following sections, when presenting the FIT metrics, selected examples will be presented as evidence of the reality of today's LOD KOS products.

For this section, a significant portion will be devoted to the research findings and recommendations related to the Functional metric. If a dataset is assessed according to these guidelines and fails to adhere to the principles indicated in terms of its functions, it would not be useful to further consider its Impacts and potential Transformable usages.

### 4.1 Functional

*Functional means that the vocabulary is made available in ways that enhance its inherent purpose.*

Our findings suggest that this is the most critical metric for datasets to align with. No matter how a LOD KOS dataset meets the FAIR principles, if a LOD KOS is



not made available in ways that enhances its inherent purpose, it would not be relevant to consider it as a good value vocabulary. To be Functional, we recommend that a dataset be assessed using four major criteria.

### Functional–1. The vocabulary is delivered in consumable formats

Table 1 presents the number of datasets of the original KOS types, including thesauri, classifications, taxonomies, terminologies, and lists, plus the number of SPARQL endpoints provided. Table 2 shows the formats as well as the number of datasets which make data available in that format. The major findings reveal that: (1) Many serialization formats have been used for KOS' deliverables, such as JSON, HTML, Turtle, N-triples, RDF/XML, CSV, and more. The default /auto format is usually a html table. (2) The number of KOS products offering operational SPARQL endpoints is still very limited, despite a gradual increase. We should understand that, enabling a SPARQL endpoint allows for the targeting of specific bits of data, and the content types of query results can be selected based on the intended usage in applications. (3) The majority of KOS datasets made available via SPARQL endpoints have been implemented with tools including Virtuoso, PoolParty, Fuseki, and ARC SPARQL+. Others have used their own custom-built implementation or do not indicate what tool they are using. (4) Supported SPARQL features differ based on the service implementation, yet we found that all endpoints offered a number of serialization formats for query results.

Table 1. Number of SPARQL endpoints provided (data collected in 2016, 2017, and 2019 from the Datahub).

2016			2017			2019		
Search Type of KOS/ DATASET	# found	# with SPARQL endpoints	Search Type of KOS/ DATASET	# found	# with SPARQL endpoints	Search Type of KOS/ DATASET	# found	# with SPARQL endpoints
Thesaurus	67	39	Thesaurus	79	40	Thesaurus	80	41
Classification	458	29	Classification	476	31	Classification	478	31
Taxonomy	26	8	Taxonomy	35	8	Taxonomy	37	10
Terminology	35	7	Terminology	39	8	Terminology	39	8
List	665	52	List	821	58	List	825	59
<b>Total</b>	<b>1,251</b>	<b>135</b>	<b>Total</b>	<b>1,450</b>	<b>145</b>	<b>Total</b>	<b>1,459</b>	<b>149</b>



Our recommendation is that a KOS vocabulary should be delivered in consumable formats: available in various data serialization formats and accessible through a SPARQL endpoint.

### Functional–2. Provided SPARQL endpoints are operational

Our findings have shown that nearly 80% of endpoints reviewed are operational. Though encouraging, it is still evidence that more than 20% are no longer working. Being able to rely on an endpoint or to anticipate when it might be unavailable will

Table 2. Available serialization formats of KOS datasets (sorted based on data collected 2019).

Format	2016	2017	2019
JSON	54	42	74
HTML	47	37	71
XML	55	42	69
TSV	44	30	63
RDF+XML	40	30	61
DEFAULT/AUTO	37	27	51
TURTLE	30	26	39
CSV	34	20	39
N-TRIPLES	26	18	36
JAVASCRIPT	23	11	31
SPREADSHEET	22	3	30
PLAIN/TEXT	20	21	28
QUERY STRUCTURE	15	15	23
SERIALIZED PHP	15	15	22
JSON-LD		3	1

be critical for some users. It is our recommendation that institutions should commit to ensuring the sustainability of access to their KOS dataset deliverables by providing a persistently available SPARQL endpoint.

### Functional–3. Dataset properties and structures are informed effectively

An assessment of the properties used inside a LOD KOS dataset is necessary because even though the datasets are made available, it is not immediately obvious what classes and properties are involved and what links exist between datasets. Understanding the properties used can help us evaluate how suited a dataset may be for reuse in other contexts; it also allows users to better understand and integrate it in various applications. One of the metrics of dataset performance quality when evaluating a data structure is assessing whether it has complied with W3C standards and if independent specialized properties are used. Our assumptions were that LOD KOS datasets would rely heavily on W3C recommended standard properties from SKOS, OWL, and RDFS. Yet, the initial findings in 2015 alerted us to the fact that a noticeable number of independent specialized properties are used. These are represented by the orange-colored rows in the following partial example from the 2015–16 sample. A similar property check was applied to the 2019 data collected to assess changes over time. In figure 4, the independent properties used in the datasets are colored in grey.

The findings reveal that properties from standard vocabularies are from SKOS, Dublin Core and Dublin Core Terms, OWL, RDFS, DBPedia ontology, FOAF, and Schema.org. SKOS was highly used especially among datasets tagged as thesauri. RDFS and OWL are increasingly being used. The majority of datasets have included independent specialized properties (as additional or as primary) to represent their



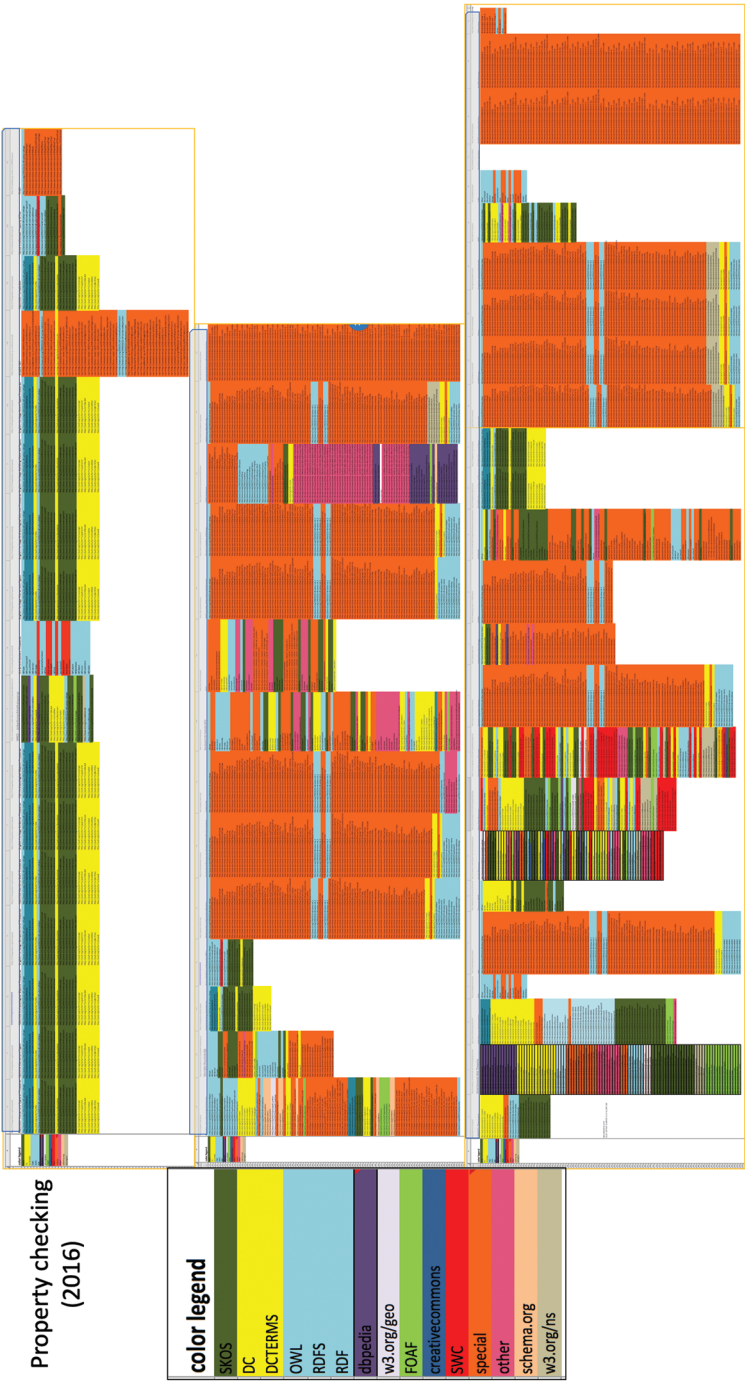


Figure 3. Property checking (2016).  
Study conducted in 2015–2016. Independent special properties used in the datasets are orange colored.

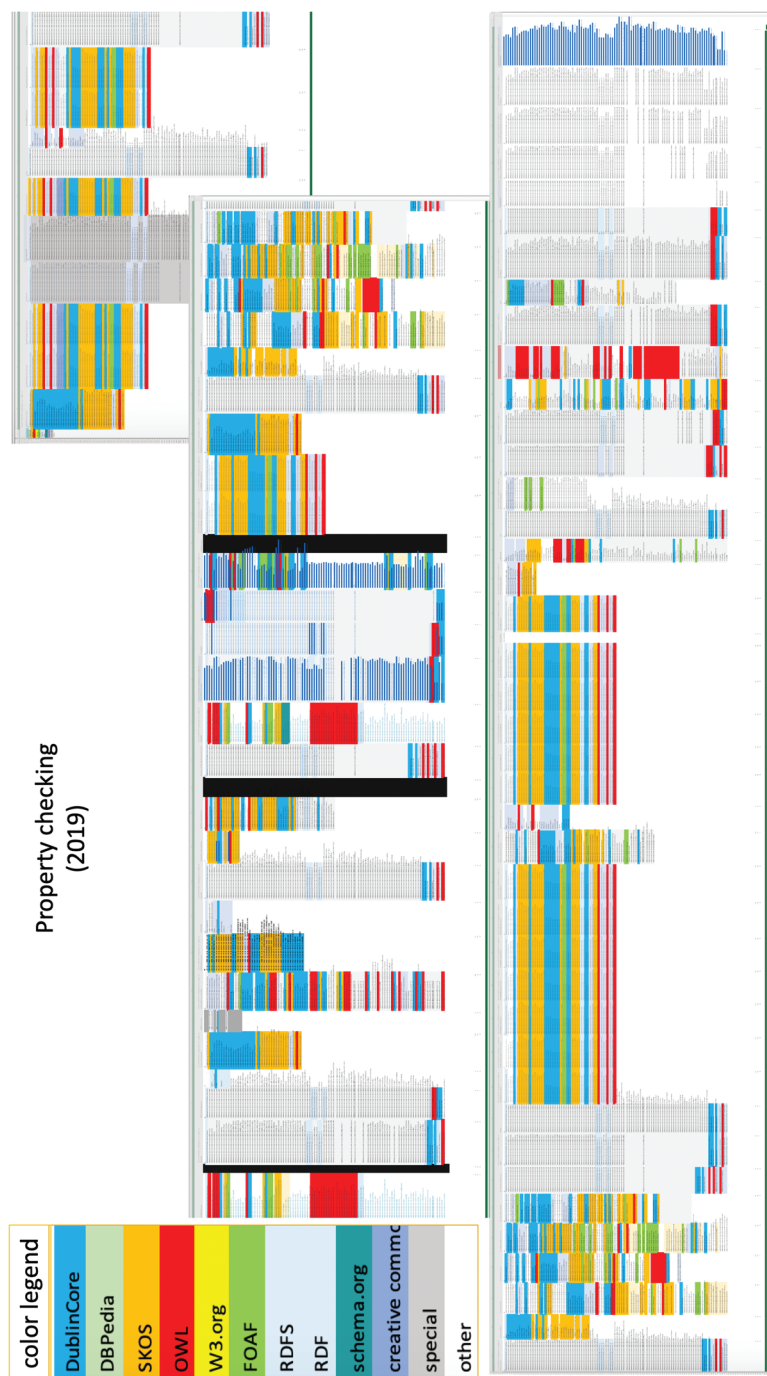


Figure 4. Property checking (2019).  
Study conducted in 2019. Independent special properties used in the datasets are grey colored.



structures, which could directly impact their interoperability and reusability. The situation is especially troublesome when there is no hint of how to query and use these properties through an endpoint. Even for users who know the SPARQL query language, they still must know the internal properties and data structures, in order to use the products.

We recommend dataset properties and structure information be more effectively and readily available. A SPARQL service should at least contain refined query examples to reveal the internal structures of the datasets. This will serve to make vocabulary contents reachable, as the usability and reusability of LOD KOS products is a major hurdle yet to be overcome.

#### **Functional–4. Services are user-friendly, making vocabulary contents reachable**

A common and increasingly challenging issue for the full usage of LOD products is that, end-users may have difficulty accessing and using the LOD datasets since they might not have been trained to access data dumps or SPARQL endpoints. Therefore, although SPARQL gives users the opportunity to design unique queries, a non-technical user may find themselves at a loss as they encounter endpoint implementations loaded with a default query and no other information. Among the available endpoints checked, less than a quarter of SPARQL endpoints would load with a default query in the query window. From Table 3, for example, it discloses that, in 2019, about 41% of them loaded with a default query; less than 20% provided query examples for users to explore the data; and only a few endpoints made more than three example queries available.

Table 3. Query examples available by year.

Year	# of datasets	Endpoint provided	Endpoint no longer available	# providing default query	# providing example queries	# providing more than 3 example queries
2019	1,459	149	74	66	26	9
2017	1,450	145	63	33	21	10
2016	1,251	135	29	-	16	6

Figure 5 shows an ideal example for endpoint providers to emulate. In the UNESCO vocabularies SPARQL service, multiple query examples are provided such as those for obtaining lists of all concepts, microthesauri, or data values in certain languages.

We highly recommend KOS producers adopt and adhere to best practices like this to enhance usability. Datasets with SPARQL endpoints should provide query examples or forms and templates to enable the easy creation of queries allowing users to interact with the data. This could be accomplished through a showcase of example queries which loads in the query window and is adjustable to user



UNESCO vocabularies - SPARQL service

Contact us

Default graph (IRI)

Query

```

1 PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
2 PREFIX isothses: <http://purl.org/iso25964/skos-thes#>
3 SELECT (CONCAT(?mtCode, ' - ', ?mtEnglishLabel) AS ?microThesaurus) ?concept
  (STR(?englishLabel) AS ?english) WHERE {
4   ?collection isothses:superGroup <http://vocabularies.unesco.org/thesaurus/domain1> .
5   ?collection skos:member ?concept .
6   ?collection skos:notation ?mtCode .
7   ?collection skos:prefLabel ?mtEnglishLabel .
8   FILTER(langMatches(lang(?mtEnglishLabel), 'en')) .
9   ?concept skos:prefLabel ?englishLabel.
10  FILTER(langMatches(lang(?englishLabel), 'en'))
11 }
12 ORDER BY ?microThesaurus ?englishLabel

```

Query examples

- Explore a sample of the data
- List all concepts of a micro-thesaurus in french
- List all concepts of a domain
- List all concepts
- List all micro-thesauri
- List all the translations english-french
- List the translations english-french
- List concepts created after
- Get the list of countries
- Select all the properties of
- Make a search on all the co
- Get all concepts in english a
- Get the hierarchical table o
- Get the hierarchical table o

HTML

Text

XML

JSON

NTriples

RDF/XML

CSV

TSV

Result format:

HTML

Run Query

Reset

Source: UNESCO vocabularies (<http://vocabularies.unesco.org/sparql-form/>).Image captured 2020-01-12.

Figure 5. User friendly SPARQL service providing multiple templates for obtaining data.

needs (e.g. Wikidata query service<sup>®</sup>). Another method could be via ready-made query templates such as those provided by the UNESCO Thesaurus (see Figure 5 above) and Getty LOD vocabularies (see Figure 7 in Section 4.3). The ideal one

<sup>®</sup> <https://query.wikidata.org/>

would be code-free visual queries generated by selecting values from various name authorities and picklists, removing the need for coding or query language knowledge (e.g. Online Coins of the Roman Empire (OCRE)<sup>®</sup>). In general, we recommend that a LOD KOS construct a user-friendly environment for SPARQL querying, offer SPARQL query examples for different data topics based on users' needs, give instruction for step-by-step SPARQL query reasoning, apply tools for data visualization, and carefully consider string values used for SPARQL retrievals.

Although such user-friendly products are rare, they illustrate how LOD KOS datasets can be potentially useful to researchers and eventually become knowledge bases (i.e. not just published RDF triple stores of value vocabularies). This will be the primary strategy that will enable LOD KOS to be Transformable (to be discussed in section 4.3).

In summary, this whole section on Functional presents four major criteria and addresses the importance of delivering the vocabulary in consumable formats, ensuring a product is accessible through persistently operational SPARQL endpoint. At the minimal level, a KOS SPARQL endpoint should contain refined query examples to inform the dataset properties and internal structures. The usability and user friendliness can be enhanced by providing default or refined query examples that enable users to explore the data structure and contents of the KOS. When a KOS is Functional, it could further its Impacts and potential Transformable usages.

## 4.2 Impactful

*Impactful means maximizing the impact of a LOD KOS vocabulary.*

A KOS vocabulary requires tremendous amounts of investments. How can we measure the investment worthiness of the task and additionally maximize its impact? The following section illustrates the best practices found among well-known KOS vocabularies.

### Impactful–1. Exposed through terminology services

The first recommended approach is to expose the LOD KOS vocabulary through terminology services such as vocabulary registries and repositories. The direct result of this action would be an increase of visibility. In our “KOS in the Semantic Web” study, major vocabulary registries and services are listed and explained, including (a) vocabulary registries and (b) vocabulary repositories/portals (Zeng, 2018)<sup>®</sup>.

<sup>®</sup> <http://numismatics.org/ocre/><sup>®</sup> <https://www.isko.org/cyclo/interoperability.htm#app2>

a). Vocabulary *registries* offer information about vocabularies (i.e. metadata); they are the fundamental services for locating KOS products. The metadata usually contain both the descriptive contents and the management and provenance information. The registry may provide the data about the reuse of ontological classes and properties among the vocabularies, (e.g. at LOV)<sup>®</sup>, indicate the available RDF formats (e.g. at BARTOC)<sup>®</sup> and the mappings (e.g. at the Datahub).

b). Vocabulary *repositories* are services hosting the full content of a KOS vocabulary as well as the management data for each component, updated regularly on time. One prominent example of such a service is BioPortal, the world's most comprehensive repository of biomedical KOS vocabularies. These terminology services' primary functions include registering, publishing, and managing diverse vocabularies and schemas, as well as ensuring they are cross-linked, cross-walked, and searchable (Golub et al., 2014). KOS content such as concepts, classes, and relationships will become available in different kinds of tools via terminology services and may be used by humans or between machines. The impacts they bring to a KOS is clear as they facilitate KOS discovery, reuse, harmonization, and synergy across disciplines and communities. When exposing data to the terminology services, it is essential to fully follow the FAIR principles and the additional recommendations we provided in Section 3.

## Impactful–2. Used by data providers

The impact of a KOS can be measured through its usage by data providers in two categories: (a) used as a primary value vocabulary and (b) used in semantic enrichment processes.

a). In the 21<sup>st</sup> century, the number of users of, and variety of applications for KOS as primary value vocabularies by data providers have increased in comparison to the original usages by cataloging, indexing, and abstracting services in the 20<sup>th</sup> century. LOD KOS vocabularies have become a fundamental component of the LOD building blocks because they enable datasets to become 4- and 5-star Open Data, that depend on KOS value vocabularies as the sources of URIs/IRIs. Individual KOS services may accumulate daily site statistics (visits, downloads, and sharing). KOS vocabularies served by BioPortal are shown with statistics of visits and a list of the projects using this vocabulary (see example about *Medical Subject Headings* at BioPortal<sup>®</sup>).

b). A noticeable new usage of LOD KOS is related to the semantic enrichment process. Enriching metadata has been used to improve data quality while providing

---

<sup>®</sup> <https://lov.linkeddata.es/dataset/lov/>

<sup>®</sup> <https://bartoc.org/>

<sup>®</sup> <http://biportal.bioontology.org/ontologies/MESH>



Research Paper

more contextual and multilingual information. Metadata enrichment from select KOS vocabularies is now an integral part of Europeana and its data providers’ strategy to enrich millions of data values related to concepts, places, and agents.

Europeana Dereferenceable vocabularies

File Edit View Insert Format Data Tools Add-ons Help

100% View only

Vocabulary		
A	B	C
Vocabulary	URL	Type of entity
The Getty - Art & Architecture Thesaurus (AAT)	<a href="http://vocab.getty.edu/aat/">http://vocab.getty.edu/aat/</a>	skos:Concept
The Getty - Union List of Artist Names (ULAN)	<a href="http://vocab.getty.edu/ulan/">http://vocab.getty.edu/ulan/</a>	edm:Agent
Getty Thesaurus of Geographic Names (TGN)	<a href="http://vocab.getty.edu/tgn/">http://vocab.getty.edu/tgn/</a>	edm:Place
Virtual International Authority File (VIAF)	<a href="http://viaf.org/viaf/">http://viaf.org/viaf/</a>	edm:Agent
Geonames	<a href="http://sws.geonames.org/">http://sws.geonames.org/</a>	edm:Place
IconClass	<a href="http://iconclass.org/">http://iconclass.org/</a>	skos:Concept
Gemeinsame Normdatei (GND)	<a href="http://d-nb.info/gnd">http://d-nb.info/gnd</a>	edm:Agent, edm:Place, skos:Concept
Israel Museum Jerusalem Concepts	<a href="http://www.imj.org.il/imagine/thesaurus/objects/">http://www.imj.org.il/imagine/thesaurus/objects/</a>	skos:Concept
data.europeana.eu WWI Concepts from Library of Congress Subject Headings (LCSH)	<a href="http://data.europeana.eu/concept/loc">http://data.europeana.eu/concept/loc</a>	skos:Concept
Europeana Sounds Genres	<a href="http://data.europeana.eu/concept/soundgenres/">http://data.europeana.eu/concept/soundgenres/</a>	skos:Concept
UDC	<a href="http://udcdata.info/rdf/">http://udcdata.info/rdf/</a>	skos:Concept
UNESCO Thesaurus	<a href="http://vocabularies.unesco.org/thesaurus/">http://vocabularies.unesco.org/thesaurus/</a>	
YSO - General Finnish ontology	<a href="https://finto.fi/ysa/en/">https://finto.fi/ysa/en/</a>	skos:Concept

Source: Europeana semantic enrichment (<https://pro.europeana.eu/page/europeana-semantic-enrichment>)  
→ link to several vocabularies. Image captured 2019-12-27.

Figure 6. Vocabularies used by Europeana for semantic enrichment.

Institutions that use name authority data to semantically enrich their digital collections can easily embed the identifiers within individual datasets. By using specific properties of established schemas, for example, *owl:sameAs*, datasets can link to the HTTP URIs from value vocabularies such as ULAN, TGN, GeoNames, LCSH, and portals such as VIAF and Wikidata. Successful cases can be seen in libraries, archives and museums (LAMs) as well as project-based digital collections (see details at Zeng, 2019). LOD KOS that can be used for semantic enrichment of originally structured, semi-structured, and unstructured data have directly impacted the quality and effectiveness of those data’s delivery on the web, greatly enhancing their FAIR compliance.

Impactful-3. Mapped with other KOS vocabularies

To achieve the semantic interoperability of existing KOS vocabularies, activities establishing relationships between the contents of one vocabulary and those of another have seen increased engagement. Mapping is a common process of establishing relationships between the concepts of one vocabulary and those of



another. A top-down or centralized full vocabulary mapping could be initiated by one source vocabulary (e.g. AGROVOC) and mapped to other target vocabularies. Alignments require interoperability in syntax & structure. The levels of mapping might be clearly defined based on SKOS, such as the *skos:exactMatch* and *skos:closeMatch*. (See example of AGROVOC Alignments report<sup>®</sup>.) Thesauri are the most common KOS type utilizing KOS vocabulary alignment due to the standardized model interpreting thesaurus structure using SKOS. Examples include EuroVoc, AGROVOC, LC Subject Headings (LCSH), STW Thesaurus for Economics, Medical Subject Headings (MeSH), etc. Wikimedia items are increasingly being included in the alignments, among them Wikidata and Wikipedia are the main targets.

The bottom-up alignment product Mix'n'match<sup>®</sup> is the largest mash-up effort by volunteers to manually map the entries of selected KOS vocabularies (full or sections) to Wikidata items. As a tool, Mix'n'match lists entries of hundreds of external databases in a variety of categories; the Authority Control category has over 100 listed (as of the end of 2019). The scale and diversity of the KOS datasets involved are very notable, covering many languages, domains, events, and regions of the world. The inclusion of a KOS vocabulary in the Mix'n'match is one significant way to demonstrate interoperability and improved quality of a KOS product.

Another approach distinct from vocabulary-based mapping is value-based mapping. A similar volunteer-contributed outcome is the “Authority Control” section in Wikipedia pages for agents, places, works (distinct intellectual or artistic creation), and historical events where identifiers of a THING are provided (e.g. for Leonardo da Vinci<sup>®</sup>). Each name authority has a namespace and reveals details of the original KOS vocabulary, in addition to allowing direct exploration of the identifier. Name authorities for agents and places have dozens listed, including those globally used and those used only in certain national- and language-domains. A few KOS vocabularies for concepts can also be found. The mapping identifiers in Wikidata authority records are double or triple the numbers found in the Wikipedia. The KOS exposed through Wikipedia are increasingly impactful and may lead to further exploration of the Wikipedia entries.

#### **Impactful—4. Showed/discussed at professional conferences and publications**

Cases of KOS showed and discussed at professional conferences and academic publications provides another way to disseminate, discover, and measure the impact

<sup>®</sup> <http://aims.fao.org/standards/agrovoc/linked-data>

<sup>®</sup> <https://tools.wmflabs.org/mix-n-match/#/>

<sup>®</sup> [https://en.wikipedia.org/wiki/Leonardo\\_da\\_Vinci](https://en.wikipedia.org/wiki/Leonardo_da_Vinci)



of a LOD KOS. Established methods such as content analysis and bibliometrics would be appropriate for studying their impact. (Since those methodologies are pretty mature, we will not explain them here in detail.) Notable professional conferences include the NKOS workshops<sup>®</sup> held at TPD (Theory and Practice of Digital Libraries) conferences and DCMI International Conferences on Dublin Core and Metadata Applications, the LODLAM Summit unconferences<sup>®</sup>, and events held by ISKO (International Society for Knowledge Organization) and ISKO-chapters<sup>®</sup>.

In summary, the discussions about the Impactful metric in this section reveal ways to measure the new impacts of KOS vocabularies brought by their advanced LOD products in the 21<sup>st</sup> century. Any vocabulary can be exposed through terminology services following the FAIR criteria, used as a primary value vocabulary as well as in semantic enrichment processes, mapped with other KOS vocabularies (whole or part) and aligned with entry-level entities of Wikimedia. Research projects and usages showed or discussed at professional conferences and publications provide evidence of the impacts. All these will help to maximize the impact of a LOD KOS vocabulary, which may affect the investment decisions of the vocabulary itself as an open resource.

### 4.3 Transformable

*Transformable means extending the functionality and impact through innovative adaptations.*

During the last decade, encouragingly, a handful of LOD KOS products have extended the functionality of original KOS resources through publishing into LOD. Among the transformable approaches, we would like to highlight the great potential when LOD KOS datasets become knowledge bases (i.e. rather than existing solely as published RDF triple stores). A LOD KOS is Transformable when it extends its functionality and impact through innovative adaptations.

#### **Transformable–1. Allows special KOS products to be derived from the original data**

LOD brings effective new features to KOS vocabularies, enabling derivation of components from the original datasets in a few seconds. For example, from the UNESCO Thesaurus, dozens of micro-thesauri (e.g. “Geography and oceanography”, “Culture”, “Religion”, “Social policy and welfare”, “International relations”, “Finance and trade”) can be obtained quickly. A micro-thesaurus is a



<sup>®</sup> <https://nkos.slis.kent.edu/#workshop>

<sup>®</sup> <https://lodlam.net/>

<sup>®</sup> <https://www.isko.org/events.html>

designated subset of a thesaurus that is capable of functioning as a complete thesaurus (ISO 25964-2:2013). Their individual concepts can be also obtained in other languages.

### **Transformable–2. The user is given autonomy to determine what structure and information is desired and can be reproduced from the vocabulary**

Fully benefiting from the original faceted and hierarchical structures of a KOS vocabulary, a LOD KOS gives users the autonomy to determine what structure and information is desired and can be reproduced. One case worth sharing is the Art and Architecture Thesaurus (AAT). AAT's Linked Data SPARQL endpoint<sup>®</sup> makes it possible for anyone to generate a micro-thesaurus dataset (e.g. Object Genres or a smaller unit of Object Genres by Function), encompassing concept URIs, labels, scope notes, and semantic relationships represented as linked data datasets. Since it is easily obtainable through pre-prepared query templates and is downloadable in multiple formats for both human and machine applications, this transformable feature gives a shortcut to any digital collection that needs standardized value vocabularies.

Another illustration of this point is the Global Agricultural Concept Scheme (GACS)<sup>®</sup>. By selecting and mapping among three selected sets of frequently used concepts from three large KOSs, the GACS created a shared LOD KOS hub that includes interoperable concepts related to agriculture, providing 15,000 concepts and over 350,000 terms in 28 languages (Baker et al., 2016).

These cases demonstrate the benefits of giving users the autonomy to determine what and how they will use the data provided, which acts as incentive to reproduce it in unique applications.

### **Transformable–3. Enables extensibility to fit diverse needs**

The above cases also apply to T3, as KOS are being extended to fit the diverse needs of language, culture, domain, and structure. This concept is not new for KOS, since several have been internationally adopted and used worldwide in the 20<sup>th</sup> century. For KOS to be appropriately adopted, reused, and reproduced in these contexts, the provenance data of the whole KOS or parts are essential for its quality and trustworthiness. Such data can be very well documented and used in the LOD version. Commonly used properties such as *dcterms:created*, *dcterms:modified*, *skos:changeNote*, *prov:wasGeneratedBy*, and *prov:used* have been applied to a KOS's entry level to express changes made to fit language, culture, structure, and



<sup>®</sup> [http://vocab.getty.edu/queries#Finding\\_Subjects](http://vocab.getty.edu/queries#Finding_Subjects)

<sup>®</sup> <http://browser.agrisemantics.org/gacs/en/>

domain<sup>⊗</sup>. In addition to deriving a new vocabulary from existing LOD KOS datasets, some vocabularies may be extended to align with other resources, i.e. virtual harmonization through linking. In these situations, the correct indication of relationships becomes critical. For example, *foaf:focus* vs. *owl:sameAs* tells if a *skos:Concept* instance is connected to the external URI of a real-world entity or a name authority of this thing<sup>⊗</sup>.

As we consider these cases, especially the last example, we can further ask the question: how can LOD KOS products become something beyond a value vocabulary? Next, we will outline our final T: Supports innovative and transformative uses beyond normal “value vocabularies.”

#### **Transformable—4. Supports innovative and transformative uses beyond normal “value vocabularies”**

Through various case studies, we found the newest and most important function of KOS datasets which should be considered as “knowledge bases,” beyond being normal “value vocabularies” (Zeng & Mayr, 2018). With the advancement of the RDF model, a graph data model is considered to be one of the most flexible formal data structures. Among the knowledge bases, “knowledge graphs” have increasingly become a more widely used concept and label.

A Knowledge Graph (KG) is a graph-theoretic knowledge representation that (at its simplest) models entities and attribute values as nodes, and relationships and attributes as labeled, directed edges (Kejriwal, Sequeda, & Lopez, 2019). Prior to coinage of the term “Knowledge Graph”, proponents of the Semantic Web pressed for the use of graph-theoretic models, pattern-matching query languages, graph data management and use of publicly available KGs like DBpedia, GeoNames and Wikipedia for information retrieval as well as knowledge acquisition and alignment (Kejriwal, Sequeda, & Lopez, 2019). Among the many benefits of knowledge graphs, one of the most noticeable is the potential for discovery of hidden knowledge. The contextual information which enables this discovery is provided by the RDF triples which follow the Linked Data principles and are embedded in trustable value vocabularies and property vocabularies.

Consider the SPARQL query examples provided by UniProt (Universal Protein Resource). One of them is “Select all bacterial taxa and their scientific name from the UniProt taxonomy” which is obviously a function that a value vocabulary provides. It also provides over 20 other query examples, such as “Select the preferred gene name and disease annotation of all human UniProt entries that are known to

<sup>⊗</sup> Example: <http://vocab.getty.edu/aat/300196975>

<sup>⊗</sup> Example: in the RDF/XML raw data of <http://id.worldcat.org/fast/35588/>, view-source: <http://id.worldcat.org/fast/35588.rdf.xml>



Getty Vocabularies: LOD		SPARQL	Queries
4	TGN-Specific Queries	5	ULAN-Specific Queries
4.1	Places by Type	5.1	Agents by Type
4.2	Places, with English or GVP Label	5.2	Associative Relations of Agent
4.3	Places by Direct and Hierarchical Type ←	5.3	Female Artists ←
4.4	Breakdown of Sovereign States by Type	5.4	Female Artists as a Hobby
4.5	Inhabited Places That Were Sovereign States	5.5	Native American Painters ←
4.6	Places by Type and Parent Place ←	5.6	Names of Native American Painters
4.7	Places by Type, with placeTypePreferred ←	5.7	Architects Born in the 14th or 15th Century ←
4.8	Places by Triple FTS	5.8	Indian and Pakistani Architectural Groups
4.9	Places by FTS Parents	5.9	Non-Italians Who Worked in Italy ←
4.10	Capitals by Association	5.10	Artists Associated to a Given Patron or His ←
4.11	Members of the European Union	Family	
4.12	Members of the United Nations	5.11	German, Dutch, Flemish printmakers, listed with ←
4.13	Geo Chart with sgvizler	their teachers	
4.14	Column Chart with sgvizler	5.12	Artists Whose Identity May be Associated or ←
4.15	Countries and Capitals By Type and ←	Confused With Another	
Containment		5.13	Ordered Hierarchy of Given Subject
4.16	Places by Coordinate Bounding Box ←	5.14	Ancient Artists or Groups by Nationality ←
4.17	Places Within Bounding Box	5.15	Art Repositories in the USA by State ←
4.18	Places by Type Within Bounding Box ←	5.16	Popes and Their Reigns
4.19	Places Outside Bounding Box (Overseas ←	5.17	Pope Reign Durations
Possessions)		5.18	Life Events
4.20	Places Nearby Each Other ←	5.19	Artists with Name, Bio, Nationality, Type

Source: Getty Vocabularies LOD—Queries (<http://vocab.getty.edu/queries>). Image captured 2019-12-29.

Figure 7. Query templates for ULAN and TGN (portion).

be involved in a disease” and “Select all triples that relate to the taxon that describes *Homo sapiens* in the named graph for taxonomy” (UniProt Consortium, 2002–2020)<sup>9</sup>. Prior to meeting the FIT requirements these comprehensive questions could not be answered. Since these innovative uses are enabled and FITted, the benefits to researchers become obvious. Another case that set high standards for others to follow is the Getty Vocabularies LOD SPARQL endpoints<sup>10</sup> (see Figure 7). The templates reveal potential outcomes that researchers will be very interested in and can use to obtain contextualized datasets and knowledge graphs in a few seconds. These facts demonstrate that LOD KOS can be used for obtaining special graphs or datasets that answer complicated questions, revealing unknown and unrecorded relationships and facts, and bringing new discovery of non-obvious relationships. Additionally, new knowledge could be formally abstracted in several forms: New links between entities; a potential new important entity in the domain; and changing significance of an existing entity (Taylor, 2018).

The important roles of KOS in the creation of knowledge graphs have been emphasized in the past two years, as knowledge graph development has been considered a major strategy for corporations including Google, Apple, Amazon,

<sup>9</sup> Examples provided by UniProt <https://sparql.uniprot.org/>

<sup>10</sup> <http://vocab.getty.edu/queries>

Alphabet, Microsoft Corp, Facebook, and more. The Microsoft Academic Knowledge Graph (MAKG) set has over eight billion triples with information about scientific publications and related entities as of 2018–11<sup>®</sup>.

In summary, the discussions about the Transformable in this section imply great potential for KOS vocabularies to extend their functionality and impact through innovative adaptations. Allowing special KOS products to be derived from the original data, giving users autonomy in reproduction, and enabling the extensions to fit diverse needs are major transformable approaches. More importantly, the innovative use of the originally constructed high quality, contextualized data entries enable the LOD KOS to generate large or specialized knowledge graphs, which function as knowledge bases; and to become foundations of semantic analysis and entity extractions. They consequently become the building blocks of a framework for research in humanities and science. LOD KOS products are thus transformed beyond being just “value vocabularies.”

## 5 Summary and conclusion

The motivation for this research was to encourage more productions of LOD KOS products. It addresses the major issues and challenges encountered with LOD KOS as well as offers suggestions for improving their quality and the impacts of their contribution to the Semantic Web. It is our passion to share best practice approaches identified through our multiple years of investigations and present a set of recommended metrics. By using these metrics, any assessment performed on LOD products can lead to actions addressing any or all identified issues, case-by-case, from the top-down or the bottom-up, collectively or independently.

In conclusion, without FAIR principles, FIT metrics have no foundation. Therefore, as an open dataset, a LOD KOS should be Findable, Accessible, Interoperable, and Reusable, plus implementing these additional recommendations for KOS as FAIR datasets:

- Findable recommendation – Enrich metadata as much as possible to enable data discovery processes.
- Accessible recommendation: Provide multiple pathways for access to the data.
- Interoperable recommendation: Utilize the KOS types vocabulary to standardize the way vocabulary types are categorized and thereby support mapping and interoperability.
- Reusable recommendation: Adequately supply license and provenance metadata to enable dataset reusability.



As a value vocabulary, a LOD KOS should be Functional, Impactful, and Transformable, as outlined in Table 4.

Table 4. FIT – Metrics for LOD KOS (as value vocabularies).

Functional	Impactful	Transformable
<p>[The vocabulary is...]  <b>Made available in ways that enhance its inherent purpose</b></p> <p><b>Metrics:</b>            F1. The vocabulary is delivered in consumable formats            F2. Provided SPARQL endpoints are operational            F3. Dataset properties and structures are informed effectively            F4. Services are user-friendly, making vocabulary contents reachable</p>	<p>[The vocabulary...]  <b>Maximizes the impact of a LOD KOS vocabulary</b></p> <p><b>Metrics:</b>            I1. Exposed through terminology services            I2. Used by data providers                a) as a primary value vocabulary                b) in semantic enrichment            I3. Mapped with other KOS vocabularies            I4. Showed/discussed at professional conferences and publications</p>	<p>[The vocabulary...]  <b>Extends the functionality and impact through innovative adaptations</b></p> <p><b>Metrics:</b>            T1. Allows special KOS products to be derived from the original data            T2. The user is given autonomy to determine what structure and information is desired and can be reproduced from the vocabulary            T3. Enables extensibility to fit diverse needs            T4. Supports innovative and transformative uses beyond normal “value vocabularies”</p>

The metrics of FAIR and FIT for LOD KOS need to be further tested and aligned with the best practices and international standards of both open data and various types of KOS. Discussions with the KOS community and further enhancement of these metrics will be ongoing.

## Acknowledgements

College of Communication and Information (CCI) Research and Creative Activity Fund, Kent State University.

## Author Contributions

Both authors contributed equally to this work.

## References

- Baker, T., Caracciolo, C., Doroszenko, A., & Suominen, O. (2016). GACS Core: Creation of a global agricultural concept scheme. In E. Garoufallou, Coll. I. Subirats, A. Stellato, & J. Greenberg (Eds.), *Metadata and Semantics Research* (pp. 311–316). Communications in Computer and Information Science, vol 672. Springer, Cham. [https://doi.org/10.1007/978-3-319-49157-8\\_27](https://doi.org/10.1007/978-3-319-49157-8_27)
- European Commission Expert Group on FAIR Data. (2018). Turning FAIR into reality: Final report and action plan from the European Commission expert group on FAIR data. Publications Office of the European Union. <https://op.europa.eu/s/nF4t>



**Research Paper**

- FORCE11. (2014). Guiding principles for findable, accessible, interoperable and re-usable data publishing version b1.0. FORCE11. <https://www.force11.org/fairprinciples>
- Kejriwal, M., Sequeda, J., & Lopez, V. [in press]. Knowledge graphs: Construction, management and querying. *Semantic Web Journal*, 10(6). <http://www.semantic-web-journal.net/content/knowledge-graphs-construction-querying-and-management>
- Golub, K., Tudhope, D., Zeng, M.L., & Žumer, M. (2014). Terminology registries for knowledge organization systems: Functionality, use, and attributes. *Journal of the Association for Information Science and Technology*, 65(9), 1901–1916.
- ISO 25964-2:2013. Information and documentation—Thesauri and interoperability with other vocabularies—Part 2. Interoperability with other vocabularies International Organization for Standardization.
- Taylor, J., Gao, Y., Narayanan, A., Patterson, A., & Jain, A. (2018). Panel: Enterprise-scale knowledge graphs. *International Semantic Web Conference 2018*, Monterey, CA, United States. <http://www.iswc2018.semanticweb.org/wp-content/uploads/2018/10/Panel-all.pdf>
- Tiemensma, L. (2010). Quality metrics in academic libraries: Striving for excellence. In *Qualitative And Quantitative Methods In Libraries: Theory and Applications* (pp. 219–232).
- Uniprot Consortium (2002–2020). UniProt. <https://sparql.uniprot.org/>
- Wilkinson, M., Dumontier, M., Aalbersberg, I. ..., & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data* 3, 160018 (2016). doi:10.1038/sdata.2016.18
- Wilkinson, M., Sansone, S., Schultes, E., Doorn, P., & Dumontier, M. (2018). A design framework and exemplar metrics for FAIRness. *Scientific data* 5, 180118 (2018) doi:10.1038/sdata.2018.118
- W3C. (2009). SKOS simple knowledge organization system reference. W3C Recommendation. W3C Recommendation 18 August 2009. <https://www.w3.org/TR/skos-reference/>
- Zeng, M.L. & Mayr, P. (2018). Knowledge Organization Systems (KOS) in the Semantic Web. *International Journal on Digital Libraries*. 20(3), 209–230. <https://doi.org/10.1007/s00799-018-0241-2>
- Zeng, M.L. (2019). Interoperability. *Knowledge Organization*, 46(2), 122–146. DOI: 10.5771/0943-7444-2019-2-122. Also available in B. Hjørland & C. Gnanli (Eds.), *ISKO Encyclopedia of Knowledge Organization (IEKO)*. <http://www.isko.org/cyclo/interoperability>
- Zeng, M.L. (2019). Semantic enrichment for enhancing LAM data and supporting digital humanities. *El profesional de la información*, 28(1), Article e280103. <https://doi.org/10.3145/epi.2019.ene.03>



This is an open access article licensed under the Creative Commons Attribution-NonCommercial-NoDerivs License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

# SUBMISSION GUIDELINES

## AIMS & SCOPE

*JDIS* devotes itself to the study and application of the theories, methods, techniques, services, infrastructural facilities using big data to support knowledge discovery for decision & policy making. The basic emphasis is big data-based, analytics centered, knowledge discovery driven, and decision making supporting. The special effort is on the knowledge discovery to detect and predict structures, trends, behaviors, relations, evolutions and disruptions in research, innovation, business, politics, security, media and communications, and social development, where the big data may include metadata or full content data, text or non-textual data, structured or non-structured data, domain specific or cross-domain data, dynamic or interactive data.

## REFEREING PROCESS

Articles covering the topics or themes mentioned above will be refereed through a double-blind peer review process.

## MANUSCRIPTS CATEGORIES

**Research Articles** represent original research work or a comprehensive and in-depth analysis of a topic. More than 3,000 words are considered as a proper length for such manuscripts, with a structured abstract ca. 200 words.

**Application Articles** cover the latest development and application in any segment of data analysis and information science field work. The length of the manuscript is preferred to be more than 3,000 words, with a structured abstract ca 200 words.

**Review Articles** summarize the status of knowledge and outline future directions of research within the scope of the journal.

**Perspectives** are forward-looking viewpoints that advocate important future directions in the theory or application of data and information science.

## MANUSCRIPTS REQUIREMENTS

All papers are submitted in English with a double-line space. For the assurance that all the materials of the to-be submitted are included, please check the following:

**Title page.** The title, author list, affiliations, and author contribution statement should all be included on a title page as the first page of the manuscript file.

**Structured abstract & Keywords.** A structured abstract is required for all research papers in order for readers to acquire the key points of a manuscript.

**References.** Be sure all the references used should be cited properly in both in text and in bibliography. For the detailed information, please request a copy of Reference Citation Format.

## COPYRIGHT

All submitted papers only normally should not have been previously published nor be currently under consideration for publication elsewhere. For all the materials translated or obtained from other published resources, they should be properly acknowledged. All copyright problems should be cleared without any legal entanglements prior to the publication.

## NOTES FOR INTENDING SUBMISSIONS

A guide for authors and other relevant information, including submitting papers online, is available at <http://www.jdis.org>. For any questions, you can e-mail the Office at [jdis@mail.las.ac.cn](mailto:jdis@mail.las.ac.cn).

The JDIS Editorial Office

National Science Library, Chinese Academy of Sciences

No 33, Beisihuan Xilu, Zhongguancun, Haidian District, Beijing 100190, P.R. China

Website: <http://www.jdis.org>

# JDIS

## JOURNAL OF DATA AND INFORMATION SCIENCE

(QUARTERLY)

Special Issue on Networked Knowledge Organization Systems (NKOS)

Guest Editors-in-Chief: Joseph Busch, Douglas Tudhope

| Volume 5 Number 1 2020

### CONTENTS

#### EDITORIAL

- 1 | *JDIS* Special Issue on Networked Knowledge Organization Systems (NKOS)  
Joseph Busch, Douglas Tudhope

#### RESEARCH PAPERS

- 3 | Knowledge Organization and Representation under the AI Lens  
Jian Qin
- 18 | Automatic Classification of Swedish Metadata Using Dewey Decimal Classification: A Comparison of Approaches  
Koraljka Golub Johan Hagelbäck, Anders Ardö (emeritus)
- 39 | The Second Edition of the Integrative Levels Classification: Evolution of a KOS  
Ziyoung Park, Claudio Gnoli, Daniele P. Morelli

- 51 | The ARQUIGRAFIA project: A Web Collaborative Environment for Architecture and Urban Heritage Image  
Vânia Mara Alves Lima, Cibele Araújo Camargo Marques dos Santos, Artur Simões Rozestraten

- 68 | Improving Archival Records and Service of Traditional Korean Performing Arts in a Semantic Web Environment  
Ziyoung Park, Hosin Lee, Seungchon Kim, Sungjae Park

- 81 | "SEMANTIC" in a Digital Curation Model  
Hyewon Lee, Soyoung Yoon, Ziyoung Park

- 93 | FAIR + FIT: Guiding Principles and Functional Metrics for Linked Open Data (LOD) KOS Products  
Marcia Lei Zeng, Julaine Clunis

ISSN 2096-157X



ISSN 2096-157X  
CN 10-1394/G2