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

Marcia Lei Zeng†, 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

† Corresponding author: Marcia Lei Zeng (E-mail: mzeng@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®, BioPortal®, and Linked Open Vocabularies (LOV)®) 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®. 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® 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.

---

© http://datahub.io and https://old.datahub.io/
© https://bioportal.bioontology.org/
© https://lov.linkeddata.es/dataset/lov
© https://lod-cloud.net/
© 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 and 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.
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® 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® and/or Dublin Core Metadata Terms®, (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.)

---

* https://nkos.slis.kent.edu/nkos-type.html
* https://www.dublincore.org/specifications/dublin-core/dces/
* https://www.dublincore.org/specifications/dublin-core/demi-terms/
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:

| 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) |

* 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.
• 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).

| Search Type of KOS/DATASET | 2016 # found | 2017 # with SPARQL endpoints | 2019 # with SPARQL endpoints |
|---------------------------|--------------|-----------------------------|-----------------------------|
| Thesaurus                 | 67           | 39                          | 79                          |
| Classification            | 458          | 29                          | 476                         |
| Taxonomy                  | 26           | 8                           | 35                          |
| Terminology               | 35           | 7                           | 39                          |
| List                      | 665          | 52                          | 821                         |
| Total                     | 1,251        | 135                         | 1,450                       |

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
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

| 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     |       |
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.

| 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
needs (e.g. Wikidata query service®). 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
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)). 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).
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)\(^\circledast\), indicate the available RDF formats (e.g. at BARTOC)\(^\circledR\) 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 21st 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 20th 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\(^\circledast\)).

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

\(^\circledast\) https://lov.linkeddata.es/dataset/lov/
\(^\circledR\) https://bartoc.org/
\(^\circledast\) http://bioportal.bioontology.org/ontologies/MESH
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.

![Vocabulary table](image)

| Vocabulary                                      | URL                                      | Type of entity         |
|------------------------------------------------|------------------------------------------|------------------------|
| The Getty - Art & Architecture Thesaurus (AAT) | http://vocab.getty.edu/aat/              | skos:Concept           |
| The Getty - Union List of Artist Names (ULAN)    | http://vocab.getty.edu/ulan/             | edm:Agent              |
| Getty Thesaurus of Geographic Names (TGN)        | http://vocab.getty.edu/tgn/              | edm:Place              |
| Virtual International Authority File (VIAF)     | http://viaf.org/viaf/                    | edm:Agent              |
| Geonames                                        | http://sws.geonames.org/                 | edm:Place              |
| IconClass                                       | http://iconclass.org/                    | skos:Concept           |
| Gemeinsame Normdatei (GND)                      | http://id-nb.info/gnd                    | edm:Agent, edm:Place,  |
| Israel Museum Jerusalem Concepts                 | http://www.imj.org.il/imagines/thesaurus/objects/ | skos:Concept           |
| data.europeana.eu WWI Concepts from Library of Congress Subject Headings (LCSH) | http://data.europeana.eu/concept/lc | skos:Concept           |
| Europeana Sounds Genres                         | http://data.europeana.eu/concept/soundgenres/ | skos:Concept           |
| UDC                                             | http://loddata.info/rdfl                | skos:Concept           |
| UNESCO Thesaurus                                | http://vocabularies.unesco.org/thesaurus/ | skos:Concept           |
| YSO - General Finnish ontology                  | https://finno.lyso/en/                   | 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.) 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 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). 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

---

http://aims.fao.org/standards/agrovoc/linked-data
https://tools.wmflabs.org/mix-n-match/#/
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\textsuperscript{18} held at TPDL (Theory and Practice of Digital Libraries) conferences and DCMI International Conferences on Dublin Core and Metadata Applications, the LODLAM Summit unconferences\textsuperscript{19}, and events held by ISKO (International Society for Knowledge Organization) and ISKO-chapters\textsuperscript{20}.

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\textsuperscript{st} 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

\textsuperscript{18} https://nkos.slis.kent.edu/#workshop
\textsuperscript{19} https://lodlam.net/
\textsuperscript{20} 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\(^2\) 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)\(^2\). 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\(^{th}\) 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

\(^{2}\) http://vocab.getty.edu/queries#Finding_Subjects  
\(^{2}\) http://browser.agrisemantics.org/gacs/en/
domain. 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.

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 the 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

---

Example: http://vocab.getty.edu/aat/300196975
Example: in the RDF/XML raw data of http://id.worldcat.org/fast/35588/, view-source: http://id.worldcat.org/fast/35588.rdf.xml
be involved in a disease” and “Select all triples that relate to the taxon that describes \textit{Homo sapiens} in the named graph for taxonomy” (UniProt Consortium, 2002–2020)\(^\circ\). 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\(^\circ\) (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,

\(^{\circ}\) Examples provided by UniProt https://sparql.uniprot.org/

\(^{\circ}\) 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\(^\circ\).

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.

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

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

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
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., Schultz, E., Doorn, P., & Dumontier, M. (2018). A design framework and exemplary 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. Hjorland & C. Gnoli (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/).