

Linked Open Data Enabled Bibliographical Data (LODE-BD) 3.0 – A practical guide on how to select appropriate encoding strategies for producing LOD-enabled bibliographical data

Imma Subirats
Food and Agriculture Organization of
the United Nations

Marcia Lei Zeng
Kent State University
USA

Abstract

The Linked Open Data Enabled Bibliographical Data (LODE-BD) practical guide aims to support the selection of appropriate encoding strategies for producing meaningful LOD-BD (directly or indirectly). The latest version (3.0) of the LODE-BD provides recommendations applicable for structured data, describing bibliographic resources and research datasets. The core component of LODE-BD contains a set of recommended decision trees for common properties used in describing a resource instance, enabling data providers to make choices according to their development stages, internal data structures, and the reality of their practices.

Keywords: Linked Open Data; bibliographic resources; research datasets; decision trees; metadata crosswalk; DC-terms; Schema.org.

With the paradigm of Linked Open Data (LOD) and emerging technologies in the 21st Century, it has become a general strategy to liberate data from their silos that are framed by proprietary database schemas. However, simply transforming database schemas into RDF does not create Linked Data. The idea of assisting information professionals in deciding what metadata terms and standard vocabularies to use when encoding existing bibliographic data for the purpose of exchanging and sharing across data providers was born under the umbrella of Virtual Open Access Agriculture & Aquaculture Repository (VOA3R), an European research consortium project for sharing scientific and scholarly research related to agriculture, food, and environment, conducted during the early 2010s. The VOA3R Federation was composed of 17 institutions from 13 countries which contributed bibliographic data to eight open repositories. With the goal of communicating across the data repositories of these institutions, turning the bibliographic data from an ad-hoc modeled database in a silo to the data in a standardized metadata repository, and heading to the LOD universe, it is essential to encode bibliographic data using properties from standardized namespaces, use well-established authority data and controlled vocabularies that are available as Linked Data in agriculture and aquaculture, and exchange and publish bibliographic data in RDF triples.

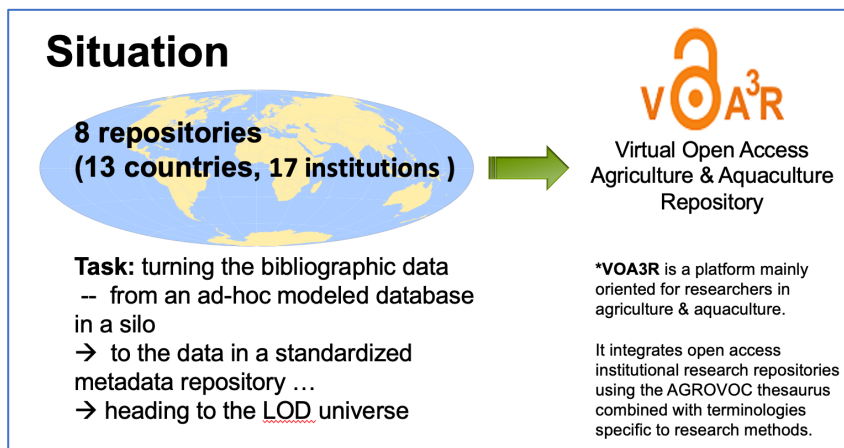


FIG 1. The initial situation and task.

The approach of employing well-accepted metadata element sets and value vocabularies has already shown great benefits and potentials in terms of resource discovery, data reuse, data sharing, and the creation of new content based on Linked Data. However, deciding to take this approach is only the first step for the data providers and service providers of an open bibliographic repository. In the context of producing LOD-enabled bibliographical data, data and service providers are likely to have many specific questions related to the encoding strategies, for example:

- What metadata standard(s) should be followed in order to publish any bibliographic data as Linked Data?
- What is the minimal set of properties that a bibliographic dataset should include to ensure meaningful data sharing?
- Is there any metadata model or application profile that can be directly adopted for producing bibliographical data (especially from a local database)?
- If the controlled vocabulary that has been used is available as Linked Data, what kind of values should be exchanged through the repository, specifically, the literal form representing a concept or the URI identifying the concept?
- How should data be encoded to move from a local database to a Linked Data dataset?

All these led to the efforts of the creation of LODE-BD, a practical guide on how to select appropriate encoding strategies for producing. The guide was born with the purpose of assisting data providers in selecting appropriate encoding strategies for producing LOD-enabled bibliographical data (directly or indirectly). LODE-BD was initially issued in 2011 and updated in 2015 (LODE-BD 2.0) with a new crosswalk to Schema.org vocabulary which was founded by several search engines in 2011. Following both DCMI Metadata Terms (DCTERMS)'s up-to-date specification and W3C's recommendation Data Catalog Vocabulary (DCAT) Version 2, released in early 2020, LODE-BD conducted a full modification and extension in the 3.0 edition. As the movement of open research data has become more mainstream, this new version also includes metadata describing research data resources, based on the experience of FAO's AGRIS pilot project of integrating research datasets metadata from United States Department of Agriculture (USDA) which was successfully conducted in 2019.

LODE-BD 3.0, published in December 2020, includes properties from dc, dcterms, bibo, agls, ags, eprint, prov, and dcat namespaces.

Table 1. A list of the prefix and namespaces used in LODE-BD

Prefix	Namespace	Standard
dc	http://purl.org/dc/elements/1.1/	<i>Dublin Core Metadata Element Set</i>
dcterms	http://purl.org/dc/terms/	<i>DCMI Metadata Terms</i>
bibo	http://purl.org/ontology/bibo/	<i>Bibliographic Ontology</i>
agls	http://www.agls.gov.au/agls/terms	<i>AGLS Metadata Standard</i>
eprint	http://purl.org/eprint/terms/	<i>Eprints Terms</i>
prov	http://www.w3.org/ns/prov#	<i>PROV-O: The PROV Ontology</i>
dcat	http://www.w3.org/ns/dcat#	<i>Data Catalog Vocabulary</i>

The guide is presented as a whole package, encompassing the important components that a data provider may encounter when deciding to produce sharable LOD-ready structured data describing bibliographic resources and research datasets within a bibliographic service. First, it provides the general recommendations of common properties identified by LODE-BD and the selected metadata terms to be used for describing bibliographic resources. Second, a set of decision trees are used to demonstrate how to select recommended properties according to local needs. A crosswalk of metadata terms used in LODE-BD and Schema.org terms is attached as an appendix.

The core component of LODE-BD contains a set of recommended decision trees for common properties used in describing a bibliographic resource instance. In order to facilitate the understanding and decision making by the VOA3R community, which has professionals who have been dealing with different projects and legacy data, having various levels of data/information

skills, and speaking different languages, the LOD-BD guide presents each of the decision trees in both an image and a table. Starting from the property that describes a resource instance, a flowchart presents decision points and gives a step-by-step solution to a given problem of metadata encoding. Each decision tree is delivered with various acting points and the matching encoding suggestions. At the end of a flowchart, alternative sets of metadata terms for selection are specified. The full range of options presented by this guide will enable data providers to make their choices according to their development stages, internal data structures, and the reality of their practices.

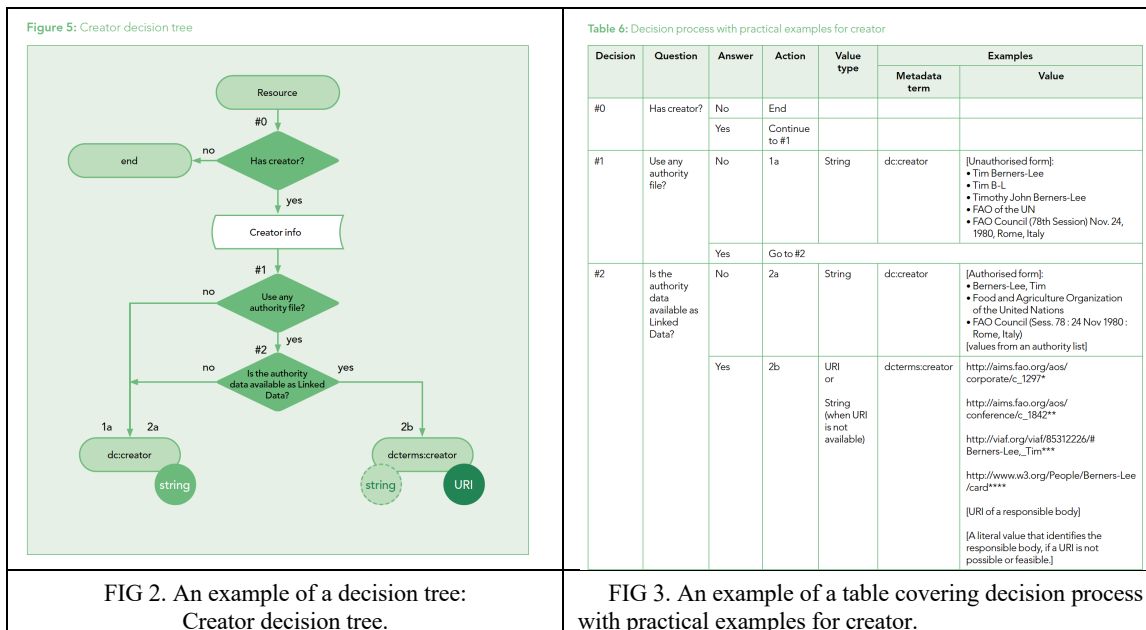


FIG 2. An example of a decision tree: Creator decision tree.

FIG 3. An example of a table covering decision process with practical examples for creator.

In summary, in order to enhance the quality, interoperability, and effectiveness of information exchange, the LOD-BD guide is built on five key principles:

- To promote the use of well-established metadata standards and the emerging Linked Open Data (LOD)-enabled vocabularies proposed by the Linked Data community.
- To encourage the use of authority data, controlled vocabularies, and syntax encoding standards in metadata statements whenever possible.
- To encourage the use of resource Uniform Resource Identifiers (URIs) as data values when they are available.
- To facilitate the decision-making process regarding data encoding for the purpose of enabling the [meta]data being findable, accessible, interoperable, and reusable (FAIR).
- To provide a reference support that is open for suggestions of new properties and metadata terms according to the needs of the Linked Data community.

Although the practical guide is geared towards the agriculture and aquaculture sectors through the VOA3R project, the LOD-BD guide is applicable for any type of metadata describing bibliographic resources and research datasets in any subject domain.

References

DCMI. (2019). Dublin Core User Guide. Retrieved May 18, 2022, from <https://www.dublincore.org/resources/userguide/>.

Subirats, I. and Zeng, M.L. 2020. *Linked Open Data Enabled Bibliographical Data (LODE-BD) 3.0 – A practical guide on how to select appropriate encoding strategies for producing Linked Open Data Enabled Bibliographical Data*. Rome, FAO. ISBN: 978-92-5-133655-7. <https://doi.org/10.4060/cb2209en>.