INTERNATIONAL STANDARD

ISO 12620-2

First edition
2022-07

Management of terminology resources — Data categories —

Part 2: Repositories

Gestion des ressources terminologiques — Catégories de données — Partie 2: Répertoires

Reference number
ISO 12620-2:2022(E)

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/TC 37, *Language and terminology*, Subcommittee SC 3, *Management of terminology resources*.

This first edition of ISO 12620-2, together with ISO 12620-1:2022, cancels and replaces ISO 12620:2019, which has been divided into parts and technically revised. The main changes are as follows:

— ISO 12620:2019 described procedures for defining data categories used in language resources and described requirements for maintaining a pragmatic, consensus-based repository of harmonized data category specifications for use in language resources. ISO 12620-1 has been narrowed to focus on the structure and rationale associated with data category specifications per se.

— The sections of ISO 12620:2019 that dealt with the creation and maintenance of data category repositories have been moved to this document.

A list of all parts in the ISO 12620 series can be found on the ISO website.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

ISO 12620-1 provides requirements and recommendations governing data category specifications for language resources. It specifies mechanisms for creating, documenting, harmonizing and maintaining data category specifications in a data category repository (DCR) with the goal of increasing the interoperability of language resources such as terminological resources, lexicographical resources and annotated text corpora. This document specifies procedures and practices for the creation, management and maintenance of DCRs.

Interoperability of language resources is a key factor for supporting innovation and progress in various focus areas of the language industry, such as terminology management, natural language processing and annotation schemes. These areas support important sectors of the economy and social development such as global communication and trade, knowledge extraction and content management.

Researchers and software developers working with language resources benefit greatly from being able to access a trusted source of information about data categories. Providing a precise description of the data categories that are used within a given data collection allows for a quick diagnosis of its compatibility with other data collections or its suitability for use in computer processes. A DCR containing vetted data category specifications provides users with the information they need in order to implement data categories in a manner that is consistent with other users. Consequently, the interoperability of language resources is greatly enhanced.

Data category specifications are normally stored in electronic format in a specially designed database. This database is called a "data category repository (DCR)". Today, it is essential for DCRs to be sharable for all stakeholders. See, for instance, Reference [3], a DCR for language resource descriptions named DatCatInfo.

Management of terminology resources — Data categories —

Part 2: Repositories

1 Scope

This document establishes criteria for the management of data categories for use in the creation and maintenance of language resources within a given community of practice (CoP). It defines the roles and responsibilities associated with the creation and maintenance of such repositories. It also specifies procedures to establish a governance structure for the management of a data category repository (DCR), including the addition of new data category specifications and continuous quality assurance.

2 Normative references

The following documents are referred to in the text in such a way that some or all of their content constitutes requirements of this document. For dated references, only the edition cited applies. For undated references, the latest edition of the referenced document (including any amendments) applies.

ISO 12620-1, *Management of terminology resources — Data categories — Part 1: Specifications*

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 12620-1 and the following apply.

ISO and IEC maintain terminology databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1
data category specification
DC specification
complete descriptive record of a data category

[SOURCE: ISO 12620-1:2022, 3.5]

3.2
data category repository
DCR
digital collection of *data category specifications* (3.1)

EXAMPLE DatCatInfo, a DCR for language resources (see Reference [3]).

Note 1 to entry: Data category repositories are used as references when specifying language resources.

[SOURCE: ISO 12620-1:2022, 3.6]

3.3
data category selection
DC selection
DCS
set of *data category specifications* (3.1) chosen from a *data category repository* (3.2)

Note 1 to entry: A data category selection can represent the data categories used within a research discipline or a specific application or project.

[SOURCE: ISO 12620-1:2022, 3.7]

3.4
community of practice
CoP
group of people who share a concern or interest in a specific activity or set of activities

EXAMPLE Terminologists.

Note 1 to entry: A CoP can be, but is not necessarily, a formal organization.

3.5
DCR management board
data category repository management board
group of experts selected by a *community of practice (CoP)* (3.4) to oversee the creation, maintenance and administration of a *data category repository* (3.2) for a given subject field

Note 1 to entry: A DCR management board also serves as an information source for members of the CoP in publicizing the DCR.

3.6
profile
selection of *data category specifications* (3.1) within a *data category repository (DCR)* (3.2) that pertain to a specific subfield of the broader subject field covered by that DCR

EXAMPLE Terminology profile, lexicography profile, semantic annotation profile.

3.7
profile management group
group of experts responsible for managing *data category specifications* (3.1) assigned to a particular *profile* (3.6)

4 Data category repositories

4.1 Data category environments

According to ISO 12620-1:2022, a data category is a class of information that forms part of a data collection or annotation scheme for a given language resource. Data categories are most frequently used as field names in data collections, but in some cases take the form of declared values. Data category specifications provide definitions, recognized names and other pertinent information needed for identifying individual data categories.

Data categories together with their data category specifications can be collected in DCRs. These DCRs serve as references when specifying language resources, designing databases for lexicographical or terminological data, or developing data exchange formats. For a specific field of knowledge, research discipline, application or project, a subset of data category specifications can be identified to form a data category selection.

4.2 Requirements for a DCR

A DCR shall meet the following requirements in order to support the effective use of data categories for language resources:

— be available in electronic form, either online or to stakeholders in a limited environment;

— provide a collection of data category specifications for reference;

— provide an automated mechanism to avoid the creation of multiple specifications containing the same data category name;

— provide a mechanism whereby users can submit new data category specifications and provide feedback on existing data category specifications;

— provide search filters allowing subsets of the DCR to be searched and retrieved using various search criteria (for instance, by date, by creator, by the content of a field);

— provide user access controls to limit write-access to persons authorized by the DCR management board;

— allow data category selections to be defined for various applications or user groups;

— subset the data category specifications based on a rigorous ontology of the data category concepts;

— allow export of data category selections, for instance, to a CSV file or XML file.

4.3 Data category profiles

A DCR shall comprise a central resource, but for management purposes it can be subset into groups of related data category specifications, each associated with a profile that represents one of the subsets. Each profile should be administered by a profile management group comprised of experts and/or specialists in the subject matter of the profile. These profiles can reflect the needs of smaller communities of practice within the context of a larger community. Members of a profile management group can also serve as members of the DCR management board.

Profiles should also provide a framework for submitting and maintaining new data category specifications. Each individual data category specification can be assigned zero or more profile values. In this regard, the management of the repository should not be fully centralized, but should be based on a structure that draws on the relevant expertise distributed throughout various subfields of language resources.

Figure 1 illustrates a DCR created to meet the needs of several closely related communities of practice, all working with similar data category subsets. The overlapping diagrams reflect shared data category specifications.

However, a specific profile should only be required if its related CoP has a need for distinct data category specifications. The set of profiles in the DCR should reflect a rigorous typology of semantic or application-oriented collections of data category specifications.

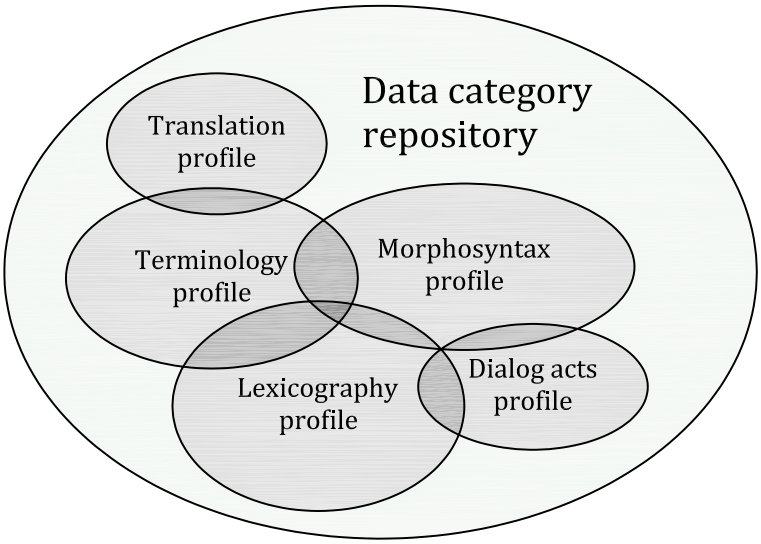

Figure 1 — Data category repository indicating sample profiles

4.4 Data category selections

Users of a DCR should be able to subset collections of data category specifications from the whole DCR or from a profile for application-specific uses in order to support research and development in language resources. These subsets are referred to as data category selections (DC selections). A data category selection shall define, in combination with a data model and optionally additional constraints (see ISO 30042), a given application-specific language resource.

EXAMPLE According to ISO 30042, a TBX data category module shall consist of a selection of terminology-related data category specifications, the metamodel defined in ISO 16642, and additional specified constraints. Modules shall be combined to define a TBX dialect, which is a TML (terminological markup language) as described in ISO 16642.

5 Roles with governance responsibilities

5.1 Communities of practice

5.1.1 Description

Members of a CoP should be active in professional fields that are represented in the DCR. Consequently, they should have applicable expertise in relevant subject fields. They should have a natural vested interest in the development of the DCR. In order to participate actively, they should express their interest by applying to participate officially. Once recognized as members, they can propose new data categories and provide feedback regarding existing resources. A CoP is strengthened by the involvement of official organizations such as professional associations.

The group can evolve naturally because of the members' common interest in a particular subject field or area, or it can be created specifically with the goal of gaining knowledge related to their field. It is through the process of sharing information and experiences with the group that the members learn from each other and have an opportunity to develop themselves personally and professionally.

5.1.2 Procedures for establishing a DCR

A CoP should observe the following steps to establish a DCR:

— demonstrate the need for a DCR and provide support for the creation of requisite infrastructure;