INTERNATIONAL STANDARD

ISO 24623-2

First edition
2021-12

Language resource management — Corpus query lingua franca (CQLF) —

Part 2: Ontology

Gestion des ressources linguistiques — Corpus query lingua franca (CQLF) —

Partie 2: Ontologie

Reference number
ISO 24623-2:2021(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 ..... **iv**
Introduction ..... **v**
1 Scope **1**
2 Normative references **1**
3 Terms and definitions **1**
4 Motivation and aims **3**
5 Structure and content of a CQLF ontology **4**
5.1 OWL DL formalism 4
5.2 Structure of the ontology 5
5.3 CQLF metamodel 7
5.4 Functionalities 8
5.5 Frames 11
5.6 Use cases 11
5.7 CQLs 12
6 Conformance statements **12**
6.1 Positive conformance statements 12
6.2 Negative conformance statements 13
Annex A (informative) **Illustrative example of a CQLF ontology** **15**
Bibliography **18**

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 4, *Language resource management*.

A list of all parts in the ISO 24623 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

Several families of International Standards codify various aspects of the representation of language data. These standards describe general corpus-oriented data models in the linguistic annotation framework (LAF) (see ISO 24612), various aspects of the semantic representation in the semantic annotation framework (SemAF) (see ISO 24617-1 and others), the representation of lexical data in the lexical markup framework (LMF) (see ISO 24613-1 and others), as well as the representation of metadata in the component metadata infrastructure (CMDI) (see ISO 24622-1 and others). Complementary to the standards concerning the representation of language data, the ISO 24623 series focuses on the exploitation of language data and on ways to satisfy various kinds of information needs targeting these data.

The corpus query lingua franca (CQLF) metamodel, described in ISO 24623-1, is a maximally permissive construct that establishes means of describing the scope of corpus query languages (CQLs) at a general level and with a focus on various kinds of data models assumed by query systems, with conformance conditions meant to be satisfied by a wide range of CQLs. The metamodel provides a "skeleton" for a CQL taxonomy by setting up basic categories of corpus queries (encoded as levels and modules) as well as the dependencies among them.

Consequently, the task of a more concrete characterization of CQLs is meant to be covered in other parts of the ISO 24623 series. This document establishes a framework for an ontology which focuses on the generalized information needs satisfied by corpus queries, and which is structured as a multi-layer taxonomy against which individual CQLs can make positive and negative conformance statements.

Such an ontology allows, on the one hand, a fine-grained comparison of the expressive power of CQLs, and, on the other hand, it serves a practical purpose, i.e. as a foundation for platforms where developers can enter conformance statements, and where end users can see which CQL to turn to in order to ensure that their search needs get satisfied. An example of such a platform is given by Reference [13].

Language resource management — Corpus query lingua franca (CQLF) —

Part 2: Ontology

1 Scope

This document specifies the structure of an ontology for a fine-grained description of the expressive power of corpus query languages (CQLs) in terms of search needs. The ontology consists of three interrelated taxonomies of concepts: the CQLF metamodel (a formalization of ISO 24623-1); the expressive power taxonomy, which describes different facets of the expressive power of CQLs; and a taxonomy of CQLs.

This document specifies:

a) the taxonomy of the CQLF metamodel;

b) the topmost layer of the expressive power taxonomy (whose concepts are called "functionalities");

c) the structure of the layers of the expressive power taxonomy and the relationships between them, in the form of subsumption assertions;

d) the formalization of the linkage between the CQL taxonomy and the expressive power taxonomy, in the form of positive and negative conformance statements.

This document does not define the entire contents of the ontology (see Clause 4).

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 24612, *Language resource management — Linguistic annotation framework (LAF)*

ISO 24623-1, *Language resource management — Corpus query lingua franca (CQLF) — Part 1: Metamodel*

ISO/IEC 10646, *Information technology — Universal coded character set (UCS)*

W3C-OWL 2-SPEC. *OWL 2 Web Ontology Language: Structural Specification and Functional-Style Syntax* (Second Edition). Motik B., Patel-Schneider, P.F., and Parsia, B. eds. W3C Recommendation, 11 December 2012. Available from: http://www.w3.org/TR/owl2-syntax/

3 Terms and definitions

For the purposes of this document, the terms and definitions given in ISO 24612, ISO 24623-1 and the following apply.

ISO and IEC maintain terminological 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
CQLF module
subcomponent of the CQLF metamodel, defined with reference to a specified data-model characteristic

Note 1 to entry: The CQLF metamodel currently distinguishes three modules within CQLF Level 1, Linear (plain-text, segmentation and simple annotation), and three modules within CQLF Level 2, Complex (hierarchical, dependency and containment).

Note 2 to entry: In 5.3, the containment module is formalized by the concept SpanContainment in order to avoid terminological ambiguity.

[SOURCE: ISO 24623-1:2018, 3.8, modified — "the CQLF metamodel" has replaced "a CQLF level" in order to improve clarity outside the context of ISO 24623-1; Note 2 to entry has been added.]

3.2
functionality
label for a concept in a *CQLF ontology* (3.15) that represents a family of *CQL capabilities* (3.12) contributing to the expressive power of a *CQL* (3.5), formulated at a general level and linked to one or more *CQLF modules* (3.1)

3.3
frame
label for a concept in a *CQLF ontology* (3.15) that represents a typical *search need* (3.6) of *end users* (3.7), understood as one facet of the expressive power of *CQLs* (3.5)

Note 1 to entry: Most frames arise from the specialization of a *functionality* (3.2) and/or the combination of multiple functionalities.

3.4
use case
label for a concept in a *CQLF ontology* (3.15) that represents a concrete instantiation of a *frame* (3.3), for which it can be determined unambiguously whether a given *query expression* (3.8) satisfies the *search need* (3.6) or not

Note 1 to entry: Use cases are often parameterized, i.e. they contain variable elements. Parameterized use cases are satisfied by parameterized query expressions.

3.5
CQL
corpus query language
formal language designed to retrieve specific information from (large) language data collections, and thereby incorporate certain abstractions over commonly shared data models that make it possible for the *end user* (3.7) (or user agents) to address parts of those data models

Note 1 to entry: A CQL defines a syntactic notation for *query expressions* (3.8) and the corresponding search semantics, i.e. an intensional specification of the intended result set. For most current CQLs, semantics are implicitly defined by a particular implementation.

[SOURCE: ISO 24623-1:2018, 3.4, modified — "CQL" has been added as preferred term, "end user" has replaced "user" in the definition and Note 1 to entry has been added.]

3.6
search need
information pattern that an *end user* (3.7) wants to locate in a corpus, based on the primary data stream and/or simple or complex annotation

3.7
end user
agent who uses a *CQL* (3.5) to satisfy his or her *search needs* (3.6)

Note 1 to entry: This can be done via an interactive graphical user interface (GUI), a command-line tool, programmatically via some application programming interface (API) or by a software program developed by the end user.

3.8
query expression
string that is syntactically valid in a given *CQL* (3.5) and can be executed to return a result set

Note 1 to entry: Query expressions are often parameterized with variable elements. No formal specification of the parameter substitution procedure is attempted, but entries for parameterized query expressions in the ontology are required to include informal descriptions of the range of admissible values and any transformations required.

3.9
parameter
variable element in a *query expression* (3.8) or in the description of a *search need* (3.6)

3.10
positive conformance statement
assertion that a given *CQL* (3.5) supports a given *use case* (3.4) by means of a *query expression* (3.8)

3.11
negative conformance statement
assertion that a given *CQL* (3.5) cannot support a given *use case* (3.4), *frame* (3.3) or *functionality* (3.2)

Note 1 to entry: Negative conformance is due to technical unavailability of specific capabilities in the respective CQL or limitations on the complexity of *query expressions* (3.8).

3.12
CQL capability
capability
corpus query language capability
facility provided by *CQLs* (3.5) to meet a specific aspect of *search needs* (3.6)

3.13
layer
totality of concepts at the same level of abstraction in a *CQLF ontology* (3.15)

EXAMPLE *Functionalities* (3.2), *frames* (3.3), *use cases* (3.4).

3.14
token
non-empty contiguous sequence of graphemes or phonemes in a document

[SOURCE: ISO 24611:2012, 3.21, modified — Note 1 to entry has been deleted.]

3.15
CQLF ontology
ontology for a fine-grained description of the expressive power of *CQLs* (3.5) in terms of *search needs* (3.6), which adheres to the structure specified in this document

4 Motivation and aims

CQLs differ widely in their basic sets of capabilities. Whereas some are restricted to rather specific application scenarios, others are able to cover a wider variety of applications and search needs. It is therefore both the quality and the quantity of CQL capabilities – as well as the degree to which they can be combined freely – that determine the expressive power of a CQL. A CQLF ontology as specified

in this document is not intended to articulate all the possible combinations of capabilities unless these are justified by genuine usage. Its aim is to provide representative categories for typical search needs within a taxonomy of CQL capabilities. These typical search needs evolve with general progress in the fields of corpus linguistics and digital humanities, and with the discovery of new challenges, new methods and new research questions. In order to accommodate the dynamic nature of the evolving search needs, most of the content of such an ontology is outside the scope of standardization. This document provides a structural framework for this dynamic information (by specifying the three-layer structure of the expressive power taxonomy, the content of the topmost layer of functionalities, and the relationships between different layers and taxonomies), ensuring that the ontology can adapt to new search needs that emerge as the relevant disciplines evolve.

In order to provide a normative skeleton for the ontology while at the same time making provisions for keeping its main content (search needs and corresponding query expressions) dynamic, this document does not comprise a normative listing of the middle and bottom layer of the expressive power taxonomy (i.e. frames and use cases). An exhaustive inventory of concepts at these two layers is not possible due to the fact that existing CQLs differ widely in the complexity of the supported combinations of functionalities, that new CQLs can be created offering additional combinations or subtypes of functionalities, and that new search needs emerge from progress in the relevant research fields. The frames and use cases of a CQLF ontology are expected to be supplied by a moderated community process, driven by CQL developers as well as end users (see Reference [13]). For illustration, a sample of frames and use cases together with conformance statements linking them with the CQP[6] and ANNIS[8] query languages is provided in Annex A.

The permissive architecture and terminology defined by this document enables research groups to extend the relevant parts of the ontology with further CQL capabilities and search needs arising in future.

The following application scenarios are thus made possible:

— describing the scope and capabilities of a given CQL, in terms of conformance statements against a CQLF ontology (typically carried out by the CQL developers);

— comparing different CQLs with respect to their ability to meet typical search needs;

— identifying suitable CQLs and query tools that support (combinations of) CQL capabilities required by an end user, together with examples of the respective query syntax;

— guiding the development of new CQLs and query tools by building an inventory of complex search needs that are important for the community (typically carried out by end users).

5 Structure and content of a CQLF ontology

5.1 OWL DL formalism

The taxonomic framework for a CQLF ontology is modelled in OWL 2 DL[7] – a dialect of the Web Ontology Language (OWL) based on the family of description logics (DL) (see Reference [9]) as a formal framework. All definitions and requirements of the W3C OWL 2 specification shall be followed. The normative representation and exchange format for a CQLF ontology is RDF/XML[10][11]. All labels and annotations shall be represented as sequences of Unicode code points, in accordance with ISO/IEC 10646.

W3C OWL 2 DL furnishes developers with a set of tools for:

a) stating concept hierarchies and membership of individuals,

b) defining highly expressive property restrictions.

In particular, this document makes use of the AnnotationProperty construct of OWL DL in order to associate additional information with concepts and individuals.

For better readability, CQLF ontology axioms are provided in DL notation in Clauses 5 and 6 rather than in the RDF/XML exchange format.

Relevant DL notions[9]:

— **concept inclusion** ⊑: This operator asserts a logical subsumption relationship between two concept expressions.

EXAMPLE 1 $A \sqsubseteq B$ asserts that A covers either a subset or the entire set of individuals contained in B. A is also said to be subsumed by B.

NOTE 1 The same notation is sometimes used to express the opposite relation (A subsumes B) for feature structures (see Reference [12], p. 496).

— **concept equivalence** ≡: This operator asserts an equivalence between two concept expressions.

EXAMPLE 2 $A \equiv B$ asserts that A covers exactly the same set of individuals as B.

— **intersection/conjunction** ⊓: This operator denotes the intersection of two concept expressions, i.e. the individuals contained in both concept expressions.

NOTE 2 $A \sqsubseteq B \sqcap C$ asserts that A is subsumed by B as well as C. It is equivalent to the assertions $A \sqsubseteq B$ and $A \sqsubseteq C$.

— **union/disjunction** ⊔: This operator denotes the union of two concept expressions, i.e. the individuals contained in either or both of the concept expressions.

NOTE 3 $A \sqsubseteq B \sqcup C$ does not imply that A is subsumed by either B or C on its own. Some of the individuals covered by A can be contained in B and others in C.

— **top concept** ⊤: denotes the set of all individuals in the domain, i.e. the entire universe. Also referred to as “Thing” or “the root class”.

— **bottom concept** ⊥: denotes the empty set of individuals in the domain. Also referred to as “Nothing” or “the empty class”.

— **concept assertion** ∈: This operator asserts that an individual belongs to a concept. Also known as “class assertion” because the concepts represent classes (see T-Box below).

EXAMPLE 3 $x \in A$ asserts that the individual x is a member of the concept A.

— **A-Box**: The domain of interest is spanned by a universe of individuals which serve as the fundamental atoms for the ontology of what shall be modelled. They become members of concepts through concept assertions (also referred to as “A-Box axioms”) and implicitly through the subsumption relations expressed by concept inclusion assertions (in the T-Box).

— **T-Box**: Concepts are represented within the terminological box (T-Box). They are classes into which individuals are organized by the A-Box axioms. The T-Box thus provides a vocabulary of concepts and a rule set of hierarchical relations between them (“is-a” relations expressed by concept inclusion axioms). Ideally, sibling categories cover a mutually exclusive space of sub-categories and/or individuals.

5.2 Structure of the ontology

The T-Box of a CQLF ontology consists of three separate taxonomies of concepts. The main taxonomy describes different facets of the expressive power of CQLs. It is called “expressive power taxonomy” and is divided into three layers.

Concepts in the top layer are called “functionalities”. They represent (families of) individual search capabilities that can be provided by CQLs at a general level. Functionalities serve as entry points for navigating the main taxonomy. Functionalities belong to the normative part of the ontology and are defined in 5.4.