This is a preview of "ISO 24623-1:2018". Click here to purchase the full version from the ANSI store.

First edition
2018-04

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

Part 1: **Metamodel**

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

Partie 1: Métamodèle

Reference number
ISO 24623-1:2018(E)

© ISO 2018

This is a preview of "ISO 24623-1:2018". Click here to purchase the full version from the ANSI store.



COPYRIGHT PROTECTED DOCUMENT

© ISO 2018

All rights reserved. Unless otherwise specified, or required in the context of its implementation, no part of this publication may be reproduced or utilized otherwise in any form or by any means, electronic or mechanical, including photocopying, or posting on the internet or an intranet, without prior written permission. Permission can be requested from either ISO at the address below or ISO's member body in the country of the requester.

ISO copyright office
CP 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Geneva
Phone: +41 22 749 01 11
Fax: +41 22 749 09 47
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

© ISO 2018 – All rights reserved

This is a preview of "ISO 24623-1:2018". Click here to purchase the full version from the ANSI store.

Contents

Page

Foreword ..... **iv**
Introduction ..... **v**
1 Scope **1**
2 Normative references **1**
3 Terms and definitions **1**
4 Aims **4**
5 Assumptions **4**
6 CQLF Metamodel **4**
7 Conformance **7**
Annex A (informative) **Example CQLF conformance statements** **8**
Bibliography **12**

© ISO 2018 – All rights reserved

This is a preview of "ISO 24623-1:2018". Click here to purchase the full version from the ANSI store.

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 on 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 the following URL: 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. Additional parts on single-stream and multi-stream ontology architectures are planned to be developed in the future.

 © ISO 2018 – All rights reserved

This is a preview of "ISO 24623-1:2018". Click here to purchase the full version from the ANSI store.

Introduction

A range of standards relating to language resource management, with the Linguistic Annotation Framework (ISO 24612) at the centre, have been developed. These standards are mostly designed to regulate the representation aspect of language data – they look at the data from the point of view of preparation and curation. This document complements this perspective by that of the end-user, that is to say, from the point of view of processing and querying.

The corpus linguistic community has, by now, developed several corpus query languages (QLs), and there is a particularly large number of them if "dialects" and forks are included. There are two main reasons for this abundance. Firstly, there are socio-economic and organizational factors, with separate query systems having been created by isolated projects with un-coordinated funding, many of them eventually developing their own set of followers. Secondly, query systems are typically sensitive to the format of the data and are often designed with a specific purpose in mind. For example, systems for querying parallel audio and transcription streams with multiple speakers have different characteristics from systems designed to query purely textual data with a single layer of morphosyntactic description. Dependency and hierarchical annotations demand yet another set of solutions. All of this results in the richness of alternatives or near-alternatives on the one hand, and in the lack of interoperability among the variants on the other. As a consequence, a "wrong" choice at the beginning of a project can bury months of research by exposing inadequacies in the initial decision after the project has become mature enough to move to new extended functionality and towards addressing more complex information needs.

This document codifies, in a modular way, the best existing practices followed in the design of corpus query languages. Its theoretical aim is to provide a basis for the investigation of the relationships between language resource architecture and corpus query language properties. The practical aim of the Corpus Query *Lingua Franca* (henceforth CQLF) is to provide linguists and language technology practitioners with a clear and coherent basis for making informed choices concerning data architectures and the query languages appropriate to them.

© ISO 2018 – All rights reserved