BS ISO 24620-5:2024

BSI Standards Publication

Language resource management — Controlled human communication (CHC)

Part 5: Lexico-morpho-syntactic principles and methodology for personal data recognition and protection in text

Documents
Document Sharing Hub

Full document details: BS ISO 24620-5:2024

National foreword

This British Standard is the UK implementation of ISO 24620-5:2024.

The UK participation in its preparation was entrusted to Technical Committee TS/1, Terminology.

A list of organizations represented on this committee can be obtained on request to its committee manager.

Contractual and legal considerations

This publication has been prepared in good faith, however no representation, warranty, assurance or undertaking (express or implied) is or will be made, and no responsibility or liability is or will be accepted by BSI in relation to the adequacy, accuracy, completeness or reasonableness of this publication. All and any such responsibility and liability is expressly disclaimed to the full extent permitted by the law.

This publication is provided as is, and is to be used at the recipient's own risk.

The recipient is advised to consider seeking professional guidance with respect to its use of this publication.

This publication is not intended to constitute a contract. Users are responsible for its correct application.

© The British Standards Institution 2024
Published by BSI Standards Limited 2024

ISBN 978 0 539 21235 8

ICS 01.140.20

Compliance with a British Standard cannot confer immunity from legal obligations.

This British Standard was published under the authority of the Standards Policy and Strategy Committee on 30 June 2024.

Amendments/corrigenda issued since publication

Date	Text affected

Documents
Document Sharing Hub

Full document details: BS ISO 24620-5:2024

International Standard

ISO 24620-5

First edition
2024-06

Language resource management — Controlled human communication (CHC) —

Part 5: Lexico-morpho-syntactic principles and methodology for personal data recognition and protection in text

Gestion des ressources linguistiques — Communication humaine contrôlée (CHC) —

Partie 5: Principes lexico-morpho-syntaxiques et méthodologie pour la reconnaissance et la protection des données à caractère personnel dans du texte

Reference number
ISO 24620-5:2024(en)

© ISO 2024

Documents
Document Sharing Hub

Full document details: BS ISO 24620-5:2024

COPYRIGHT PROTECTED DOCUMENT

© ISO 2024

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
Email: copyright@iso.org
Website: www.iso.org

Published in Switzerland

Documents
Document Sharing Hub

Full document details: BS ISO 24620-5:2024

Contents

Page

Foreword ... **iv**
Introduction ... **v**
1 Scope ... **1**
2 Normative references ... **1**
3 Terms and definitions ... **1**
4 Motivation for controlled human communication ... **2**
5 Basic principles and methodology ... **2**
5.1 General ... 2
5.2 Specific issues ... 3
5.3 Principles ... 3
5.3.1 Overview ... 3
5.3.2 Lexical, morphological and syntactic indicants ... 4
6 Applications ... **6**
6.1 General ... 6
6.2 Different language families ... 6
6.3 Languages and countries ... 6
6.4 Semes in text ... 6
6.5 Applications for personal data recognition ... 6
Annex A (informative) **Examples of text in different languages and different semes** ... **7**
Annex B (informative) **Examples of hidden text with seme indications** ... **13**
Annex C (informative) **Table of semes in context** ... **15**
Bibliography ... **18**

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 24620-5:2024

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 document 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).

ISO draws attention to the possibility that the implementation of this document may involve the use of (a) patent(s). ISO takes no position concerning the evidence, validity or applicability of any claimed patent rights in respect thereof. As of the date of publication of this document, ISO had not received notice of (a) patent(s) which may be required to implement this document. However, implementers are cautioned that this may not represent the latest information, which may be obtained from the patent database available at www.iso.org/patents. ISO shall not be held responsible for identifying any or all such patent rights.

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 24620 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.

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 24620-5:2024

Introduction

The exchange of personal data between public and private actors, including natural persons, associations and undertakings, is continually increasing. Rapid technological developments and globalization have brought new challenges for the protection of personal data. The scale of the collection and sharing of personal data has increased significantly. Technology allows both private companies and public authorities to make use of personal data on an unprecedented scale in order to pursue their activities. Natural persons increasingly make personal information available publicly and globally. Nevertheless, technology has transformed both the economy and social life, and should further facilitate the free flow of personal data within a country as well as the transfer to and between other countries and international organizations, while ensuring a high level of protection of personal data. These developments require a robust and coherent data protection framework. For example, ISO/IEC 27701 defines processes and provides guidance for protecting personally identifiable information (PII) on an ongoing, ever-evolving basis.

Effective protection of personal data requires the strengthening and setting out in detail of the rights of natural persons as data subjects, and the obligations of those who process and determine the processing of personal data.

EXAMPLE The European Union's (EU) General Data Protection Regulation (GDPR).[6][15]

The principles of data protection apply to any information concerning an identified or identifiable natural person.

In this context, numerous industries, governmental bodies, and private and public companies or organizations need to variously hide (mask)[16], remove, anonymize or pseudonymize personal data before text containing such data is processed.[4][8]

This document provides principles and a methodology to detect and identify personal data so that it can be hidden or suppressed, i.e. protected before transmitting and/or processing a text containing such data. The problem is not so much the suppression or hiding of data, but rather the recognition of personal data in a written text. Unlike personal data in text, personal data in structured data (e.g. as presented in tables) does not represent a real problem as such data are easily recognizable.[5]

This document is aimed at national and international micro, small, medium and large enterprises, as well as private/public bodies processing text which can contain personal data in all domains (e.g. law, finance, health) and languages and from different countries.[14] The principles and methodology are already in use in industry and government bodies.

Due to regulations such as the EU's GDPR, personal data protection presents a major challenge for micro, small, medium and large enterprises, as well as private and public bodies. For example, the GDPR forbids the transfer of the personal data of EU data subjects to "third countries" (countries outside of the European Economic Area (EEA)) unless appropriate safeguards are imposed, or the third country's data protection regulations are formally considered adequate by the European Commission. In addition, the state of California in the United States passed the California Consumer Privacy Act on 28 June 2018, taking effect 1 January 2020, granting rights to transparency and control over the collection of personal information by companies in a similar manner to the GDPR (see Reference [2] and ISO/IEC 27701).

All the examples in this document are fictitious but could exist if real data were to be substituted for the fictitious data.

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 24620-5:2024

Documents
Document Sharing Hub

Language resource management — Controlled human communication (CHC) —

Part 5: Lexico-morpho-syntactic principles and methodology for personal data recognition and protection in text

1 Scope

This document establishes basic principles and a methodology to recognize personal data written in free text, in different languages (whether agglutinating, inflectional or isolating) and countries.

This document is applicable to protecting human data circulating in national and international industries, and private and public organizations.

This document is applicable to processing by human beings and/or automated processing, and to various domains (e.g. law, finance, health).

It does not apply to automated image processing.

This document uses formal methods only, as statistical methods are very different in nature.

2 Normative references

There are no normative references in this document.

3 Terms and definitions

For the purposes of this document, the terms and definitions given in 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
intension
internal content of a term or concept that constitutes its formal definition

Note 1 to entry: Extension is the range of applicability of a concept by naming the particular objects that it denotes.

3.2
personal data
any information relating to an identified or *identifiable natural person* (3.6)

[SOURCE: Regulation (EU) 2016/679[6], Article 4 (1)]

© ISO 2024 – All rights reserved

Documents
Document Sharing Hub
Full document details: BS ISO 24620-5:2024