FINAL DRAFT
International Standard

ISO/FDIS 24620-5

Language resource management — Controlled human communication (CHC) —

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

ISO/TC **37**/SC **4**

Secretariat: **KATS**

Voting begins on:
2024-03-07

Voting terminates on:
2024-05-02

Reference number
ISO/FDIS 24620-5:2024(en)

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 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.

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 (UN) 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.

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)]

3.3
pseudonymization
processing (3.4) of *personal data* (3.2) in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organizational measures to ensure that the personal data are not attributed to an identified or *identifiable natural person* (3.6)

[SOURCE: Regulation (EU) 2016/679[6], Article 4 (5)]

3.4
processing
any operation or set of operations which is performed on *personal data* (3.2) or on sets of personal data, whether or not by automated means, such as collection, recording, organization, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction

[SOURCE: Regulation (EU) 2016/679[6], Article 4 (2)]

3.5
seme
Saussure's signified with its different signifiers (instantiations) in text

Note 1 to entry: Saussure was the first person to use the terminology "signified" and "signifier". Saussure offered a "dyadic" or two-part model of the sign. He defined a sign as being composed of a "signifier" (signifiant) and a "signified" (signifié) (see References [17] and [18]).

3.6
identifiable natural person
data subject
person who can be identified, directly or indirectly, in particular by reference to an identifier

Note 1 to entry: An identifier can be a name, an identification number, location data or an online identifier of a natural person. Further examples which are excluded from the examples in this document are references to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of the natural person.

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

3.7
indicant
significant occurrence of interaction between lexical, morphological and syntactic phenomena or of one of these phenomena across a wide spectrum of languages or in few languages or in just one language that is suited to identify *personal data* (3.2)

4 Motivation for controlled human communication

The first step in protecting personal data is being able to recognize such data automatically, especially when they are not structured but rather occur in free text, as shown in Example 1 in Clause A.1.

Once data are detected or recognized as personal data, different ways can be used to hide them in the text: they can be hidden (masked), removed, anonymized (see References [9] and [10]) or pseudonymized (see Reference [7]), as shown in Example 2 in Clause A.2.

Examples 3 and 4 in Clauses A.3 and A.4 show a similar example in French.

5 Basic principles and methodology

5.1 General

For the basic principles, various lexical, morphological and syntactic linguistic phenomena shall be used, in particular concerning the way in which personal data are represented in free text. For example, addresses

not respecting the English format in the UK as seen in Example 1 in Clause A.1, i.e. “Stoneham-le-Willows at 24 Britanny Park, F2 7AN (GB29 NWBK 6016 1331 9268 19)”.

The methodology specifies formal representations designed in intension (see References [11], [12] and [13]) based on lexical, morphological and syntactic phenomena that shall apply in a sequential order at each of the levels of linguistic analysis which have an impact on the recognition of personal data.

The basic principles and methodology are specifically formulated to provide an explanatory power to show how and when each of the linguistic phenomena (lexical, morphological and/or syntactic) and/or their combinations and interactions embedded in context shall be used and applied according to different semes recognition in the analysis of text. In consequence, the methodology uses linguistics phenomena conforming to basic principles for the recognition of personal data (instead of a lexicon), and specifies a system of constraint rules completed with an algorithm, which, when applied, results in extracting personal data.

5.2 Specific issues

The basic problem is the recognition of personal data in free text in different languages, from different countries and from different domains (e.g. law, finance, health).

The problem also concerns the use of the same language within different countries as, for example, addresses are not written the same way in France, Switzerland, Belgium and Canada, or in Austria and Germany (see Examples 5 and 6 in Clauses A.5 and A.6). Text in one language can also include personal data in other languages and from different countries.

The problem is thus fourfold:

a) personal data in free text;

b) personal data in different countries;

c) personal data in different languages;

d) personal data in different domains.

The methodology described in this document uses indicants to detect personal data such as a telephone number, identification or bank account number, an address, etc.

The methodology works in intension. For this reason, it is rule-based: it establishes a system of ordered constraint rules based on linguistic phenomena, which conform to the basic principles that shall be followed (see 5.3). Indicants shall be established, and they shall be lexical, morphological or/and syntactic.

5.3 Principles

5.3.1 Overview

The principles addressing the four specific issues listed in 5.2 formulate rules for an explanatory linguistic formal representation with its own meta-language and grammar accompanied with examples so that new semes and new languages can be processed.

The user shall establish lexico-morpho-syntactic indicants. The indicants, if they do not already exist in the grammar of the meta-language, shall be added.

5.3.2 Lexical, morphological and syntactic indicants

5.3.2.1 General

Some indicants shall be sets valid for different semes in different languages and different countries such as N or A which very often appear in the text but can also be optional. They shall be formally represented as follows in the meta-language grammar (extract):

— N = digit (0...9)

— L = A,B...Z (without diacritic signs)

— A = 0,1...9,L (L signifies an element of L above)

— listA = 01,02,03,04,05,06,07,08,09,10,11,12

— ListB = ListA, 13...100

— () = element(s) between parentheses optional

— M = any word starting with an uppercase or lowercase letter followed by zero or more uppercase or lowercase letters as well as ‘‘°-

— UM = M starting with a majuscule letter

— n = new line followed or not by space(s) and/or tab(s)

— listC = 1...9999

— l = a,b...z (without diacritic signs)

— |any character(s) with character | coded as || terminated by| = literal.

NOTE Sequence of literals are illegal (ambiguity of ||).

Some indicants shall be added according to the seme to be recognized, or to the country or language used (see the examples in Annex A).

Example 3 in Clause A.3 contains the following addresses written in different formats:

— 1, rue des échelles

25620 Besançon

— 43 Bd du 11 novembre 1918 à Dijon dans la Côte d’Or 21280

— Saint-Germain-en-Laye, au n° 25 rue du Château d’If 78100

The algorithmic formal representation (diagram representing the algorithm) to recognize any addresses shall be:

semeAddress = (UM( )(UM)(,)( )(M)( )(M)( ))(listC)( )(listE)(,)( )listL( )(M)( )(M)( )(n)(listNC)( )(listC)( )(listmois)( )(listC)( )(M)( )(M)( )(M)( )(M)( )(M)( )(M)( )(M)( )(M)(,)(n)( )listB( )NNN( )(UM)( )(UM)

Using elements (indicants) of the general meta-language grammar and the following complementary indicants:

— listE = bis,ter,Bis,Ter

— listL = rue,avenue,allée,impasse,place,chemin,square,boulevard,Bd,ruelle

— listNC = 1ER,1er,1ERE,1ère,2ème,2EME,3...999|ème|

— listmois = mai,juillet,novembre