~~Date: 2023-12-28~~

ISO/FDIS 24620-5:~~2023(E)~~2024

Date: 2024-02-21

ISO/TC 37/SC 4/WG 5

Secretariat: KATS

Language resource management — Controlled human communication (CHC) — Part 5: Lexico-morpho-syntactic principles and methodology for personal data recognition and protection in ~~texts~~text ~~(DataPro)~~

Gestion des ressources linguistiques — Communication humaine contrôlée (CHC) — Partie 5: Principes lexico-morpho-syntaxiques et méthodologie pour la détection et protection des données personnelles dans les textes ~~(DataPro)~~

ISO Copyright Office
CP 401 • CH-1214 Vernier, Geneva
Phone: + 41 22 749 01 11
Email: copyright@iso.org
Website: www.iso.org
Published in Switzerland.

Contents

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 ~~(see 3.2)~~ 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 ~~same~~ country as well as the transfer to and between other countries and international organizations, ~~whilst~~while ensuring a high level of ~~the~~ 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~~(see ISO/IEC 27701 in Normative references).~~.

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 ~~(see 3.4)~~ of personal data.

~~NOTE An example of this is the~~ EXAMPLE The European Union's (UN) General Data Protection Regulation (GDPR).~~[5],[16]~~[15]

The principles of data protection apply to any information concerning an identified or identifiable natural person. ~~(also called "data subject") (see 3.6). An identifiable natural person is one who can be identified, directly or indirectly,~~[2] ~~in particular within the context of this document, by reference to an identifier. Examples of identifiers can include, a name, an identification number, a bank account number, location data such as residence address or an online identifier. Further examples which are excluded from the examples in this document can include references to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. This document will apply only for written personal data in free texts.~~

In this context, numerous industries, governmental bodies, and private and public companies or ~~organisms~~organizations need to variously hide (mask~~),~~[17]~~)~~[16], remove, anonymize or pseudonymize ~~(see 3.3)~~ personal data before ~~texts~~text containing such data ~~are~~is processed.~~[3],[94]~~[8]

~~The purpose of this~~This document ~~is to provide~~provides principles and a methodology to detect and identify personal data so that it can be hidden or suppressed, ~~that is,~~i.e. protected before transmitting ~~or/~~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 ~~for example~~) does not represent a real problem as such data ~~is~~are easily recognizable.~~[4],~~[5]

~~The present document deals with formal methods only, as statistical methods are very different in nature.~~

~~The intended market for this~~This document is ~~that of~~aimed at national~~/~~ and international micro, small, medium and large~~-sized~~ enterprises, ~~and also~~as well as private/public bodies processing text which ~~could~~can contain personal data in all domains (~~for example,~~e.g. law, finance, health~~, etc.~~) and languages and from different countries.~~[15]~~[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.

~~This document is the 5th part in the ISO 24620 series on Controlled Human Communication (CHC). ISO/TS 24620-1 is the introductory Part for the series.~~

Language resource management — Controlled human communication (CHC) — Part 5: ~~lexico~~Lexico-morpho-syntactic principles and methodology for personal data recognition and protection in ~~textstext (DataPro)~~

1 Scope

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

~~The applications are essentially for~~ This document is applicable to protecting human data circulating in national and international industries, and private and public organizations.

This document is ~~directed at~~applicable to processing by human beings and/or automated processing~~. It is applicable~~, and to various domains (~~for example,~~e.g. law, finance, health~~, etc.~~)

~~but excludes.~~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~~Extension is the range of applicability of a concept by naming the particular objects that it denotes.

~~[SOURCE: and also References [12]]~~

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

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

3.3
~~pseudonymisation~~
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 ~~organisational~~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, ~~organisation~~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 ~~texts~~text

Note 1 to entry: Saussure was the first person to use ~~this~~the terminology ~~'signified'~~"signified" and ~~'signifier' which was later, between others, explained in Chandler 'Saussure~~"signifier". Saussure offered a ~~'dyadic'~~"dyadic" or two-part model of the sign. He defined a sign as being composed of~~:~~ a ~~'signifier'~~"signifier" (signifiant) and a "signified" (signifié)~~'.~~ (see References ~~[18~~ [17] and [~~19~~]18]).

~~[SOURCE:]~~

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

Note 1 to entry: An identifier can be a name, an identification number, location data~~,~~ or an online identifier of ~~that~~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

~~Due to regulations such as the European Union's GDPR, personal data protection presents a major challenge for micro, small, medium and large-sized enterprises, and also 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. The U.S. state of California also 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 [1] and ISO/IEC 27701 in Normative reference).~~

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 ~~in the following example:~~shown in Example 1 in Clause A.1.

~~**Example 1 in an English text:**~~

~~English, England~~

~~27, Milton Road,~~
~~Fredlington,~~
~~North Yorkshire PJ7 4HT~~

~~12 December 2019~~

~~John Alfred St John Esq.~~
~~7b Brians Close,~~
~~Upper Avon FY7L 2PQ~~

~~Dear John,~~

~~We are pleased to invite you to our daughter's wedding which will take place at Wexham Register Office at 3 pm 15 July 2020. If you wish to give a present to the future happy couple, either send your gift to my brother, now living in Stoneham-le-Willows at 24 Britanny Park, F2 7AN (GB29 NWBK 6016 1331 9268 19), or directly to NO93 8601 1117 947. In case you need more information, you can contact us at 0121 496 0998.~~

~~On another matter, as you may know, my sister is at the hospital. As a doctor you may access her medical data with her NHS number 485 777 3456. In respect of any emoluments, these will be sourced from her old age pension (her NINO being ZP 24 47 29 B), to be debited from LC55 HEMM 0001 0001 0012 0012 0002 3015.~~

~~Very sincerely yours,~~

~~Anon Nona Other~~

Once data ~~is~~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 ~~also~~ References ~~[10~~ [9] and [~~11~~10]) or pseudonymized (see ~~definition in 3.~~Reference [7]), as shown in Example 2 in Clause A.2.

Examples 3 and ~~References [7]~~ 4 in Clauses A.3 and ~~[8]).~~ A.4 show a similar example in French.

~~**Example 2 with the English text personal data obfuscated:**~~

~~English, England~~

****

~~12 December 2019~~

~~John Alfred St John Esq.~~

****

Dear John,

We are pleased to invite you to our daughter's wedding which will take place at Wexham Register Office at 3 pm 15 July 2020. If you wish to give a present to the future happy couple, either send your gift to my brother, now living in ********** (**********), or directly to **********. In case you need more information, you can contact us at**********

On another matter, as you may know, my sister is at the hospital. As a doctor you may access her medical data with her NHS number **********. In respect of any emoluments, these will be sourced from her old age pension (her NINO being **********), to be debited from **********.

Very sincerely yours,

Anon Nona Other

Example 3 with personal data obfuscated in a French text:

Jean-François Dupont d'Alembert de la Pauserie

[ADD][ADD][ADD][ADD][ADD][ADD][ADD

Cher Monsieur Duconte,

Je vous écris de la part de Madame la Maire de Fougerolles suite à l'accident de mon fils Hervé.

Orange a déjà été contacté par Mme Le Guevel qui habite au [ADD][ADD][ADD][ADD][ADD][ADD][ADD][ADD][ADD][ADD][ADD][A.

Je vous précise le n° de ss de mon fils: [IDN][IDN][IDN].

Afin que vous puissiez verser l'indemnité et les frais d'hôpitaux, veuillez noter mon compte: [BAN][BAN][BAN][BAN][BAN][BAN][BA.

Je vous laisse aussi un numéro pour me contacter:[TEL][TEL][Tou bien vous pouvez me joindre, en cas d'urgence ici[TEL][TEL][TEL][TELLe médecin traitant est à [ADD][ADD][ADD][ADD][ADD][ADD][ADD][ADD][ADD][ADD][ADD][A. Mon fils est contactable au hervedupond@orange.fr

25 Basic principles and methodology

25.1 General

This clause lays out the basic principle and methodology for personal data recognition.

(1) 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 texts, astext. For example, addresses not respecting the English format in Englandthe UK as seen in the example below Example 1 in Clause A.1, i.e. "Stoneham-le-Willows at 24 Britanny Park, F2 7AN (GB29 NWBK 6016 1331 9268 19)".

Stoneham-le-Willows at 24 Britanny Park, F2 7AN (GB29 NWBK 6016 1331 9268 19)

extracted from Example 1 in Clause 4

~~(2)~~ The methodology ~~as proposed in this document~~ specifies formal representations designed in intension (see ~~3.1 and~~ References [11]. [12~~],~~] and [13~~] and [14~~]) 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 ~~specified~~specifically formulated ~~so as~~ 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 ~~(see 3.5)~~ recognition in the analysis of ~~texts~~text. In consequence, the methodology ~~proposed~~ uses ~~the~~ linguistics phenomena ~~conformant~~conforming to ~~the~~ basic principles for the recognition of personal data~~, this~~ (instead of a lexicon~~,~~), and specifies a system of constraint rules completed with an algorithm, which, when ~~applies~~applied, results in extracting personal data.

~~25.2~~5.2 ~~Problem:~~ Specific issues

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

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 4~~Examples 5 and ~~5~~ 6 in Clauses A.5 and A.6). ~~Texts~~Text in one language can also include personal data in other languages and from different countries.

~~**Example 4 in a German text:**~~

~~German, Austria~~

~~Herr Dominik Hornung~~
~~Raxstrasse 32/12~~
~~1100 Wien~~

~~Sehr geehrter Herr Hornung,~~

~~Ich schreibe Ihnen im Auftrag von Herrn Bürgermeister von Grub in Folge des Unfalls meines Sohnes, Ludwig Beethoven.~~
~~Die Versicherung wurde durch Frau Mag. Kratochwill, wohnhaft Hauptstrasse 72, in Sulz Wienerwald, verständigt.~~
~~Die Versicherungsnummer meines Sohnes lautet: 4029.~~
~~Damit die Krankenversicherung die Kosten für den Spitalsaufenthalt zurückzahlen kann, hier meine Kontodaten:~~
~~IBAN: AT25 110067643485~~
~~BIC: BKAUAWWT~~
~~Meine persönliche Telefonnummer: +43 431 718 2356.~~
~~Im Notfall können Sie mich auch unter der Nummer 0043 68187235698 erreichen.~~
~~Der behandelnde Arzt hat seine Ordination Badgasse 16, 2412 Mödling, NÖ.~~
~~Mein Sohn ist zu kontaktieren an folgende Email Adresse: ludwig.beethoven@gmx.at~~

~~Mödling, am 1. März 2020,~~

~~**Example 5 in a German text:**~~

~~German, Germany~~