Norme internationale

ISO 24620-5

Première édition
2024-06

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

Language resource management — Controlled human communication (CHC) —

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

Numéro de référence
ISO 24620-5:2024(fr)

ISO copyright office
Case postale 401 • Ch. de Blandonnet 8
CH-1214 Vernier, Genève
Tél.: +41 22 749 01 11
E-mail: copyright@iso.org
Web: www.iso.org

Publié en Suisse

Sommaire

Page

Avant-propos

L'ISO (Organisation internationale de normalisation) est une fédération mondiale d'organismes nationaux de normalisation (comités membres de l'ISO). L'élaboration des Normes internationales est en général confiée aux comités techniques de l'ISO. Chaque comité membre intéressé par une étude a le droit de faire partie du comité technique créé à cet effet. Les organisations internationales, gouvernementales et non gouvernementales, en liaison avec l'ISO participent également aux travaux. L'ISO collabore étroitement avec la Commission électrotechnique internationale (IEC) en ce qui concerne la normalisation électrotechnique.

Les procédures utilisées pour élaborer le présent document et celles destinées à sa mise à jour sont décrites dans les Directives ISO/IEC, Partie 1. Il convient, en particulier, de prendre note des différents critères d'approbation requis pour les différents types de documents ISO. Le présent document a été rédigé conformément aux règles de rédaction données dans les Directives ISO/IEC, Partie 2 (voir www.iso.org/directives).

L'ISO attire l'attention sur le fait que la mise en application du présent document peut entraîner l'utilisation d'un ou de plusieurs brevets. L'ISO ne prend pas position quant à la preuve, à la validité et à l'applicabilité de tout droit de brevet revendiqué à cet égard. À la date de publication du présent document, l'ISO n'avait pas reçu notification qu'un ou plusieurs brevets pouvaient être nécessaires à sa mise en application. Toutefois, il y a lieu d'avertir les responsables de la mise en application du présent document que des informations plus récentes sont susceptibles de figurer dans la base de données de brevets, disponible à l'adresse www.iso.org/brevets. L'ISO ne saurait être tenue pour responsable de ne pas avoir identifié tout ou partie de tels droits de propriété.

Les appellations commerciales éventuellement mentionnées dans le présent document sont données pour information, par souci de commodité, à l'intention des utilisateurs et ne sauraient constituer un engagement.

Pour une explication de la nature volontaire des normes, la signification des termes et expressions spécifiques de l'ISO liés à l'évaluation de la conformité, ou pour toute information au sujet de l'adhésion de l'ISO aux principes de l'Organisation mondiale du commerce (OMC) concernant les obstacles techniques au commerce (OTC), voir www.iso.org/avant-propos.

Le présent document a été élaboré par le comité technique ISO/TC 37, *Langage et terminologie*, sous-comité SC 4, *Gestion des ressources linguistiques*.

Une liste de toutes les parties de la série ISO 24620 se trouve sur le site web de l'ISO.

Il convient que l'utilisateur adresse tout retour d'information ou toute question concernant le présent document à l'organisme national de normalisation de son pays. Une liste exhaustive desdits organismes se trouve à l'adresse www.iso.org/fr/members.html.

Introduction

L'échange des données à caractère personnel entre des acteurs publics et privés, y compris les personnes physiques, les associations et les entreprises, augmente continuellement. L'évolution rapide des technologies et la mondialisation ont créé de nouveaux enjeux pour la protection des données à caractère personnel. L'ampleur de la collecte et du partage de données à caractère personnel a augmenté de manière importante. Les technologies permettent tant aux entreprises privées qu'aux autorités publiques d'utiliser les données à caractère personnel comme jamais auparavant dans le cadre de leurs activités. De plus en plus, les personnes physiques rendent des informations les concernant accessibles publiquement et à un niveau mondial. Néanmoins, la technologie a transformé la vie économique et sociale, et devrait faciliter davantage la libre circulation des données à caractère personnel au sein d'un pays ainsi que leur transfert vers et entre les autres pays et organisations internationales, tout en assurant un niveau élevé de protection des données à caractère personnel. Ces développements requièrent un cadre de protection des données solide et cohérent. Par exemple, l'ISO/IEC 27701 définit les processus et fournit des recommandations pour la protection des informations personnelles identifiables (IPI) de manière continue, en constante évolution.

Il est nécessaire, pour protéger efficacement les données à caractère personnel, de renforcer et de détailler les droits des personnes physiques en tant que personnes concernées, ainsi que les obligations de ceux qui traitent et déterminent le traitement des données à caractère personnel.

EXAMPLE Le Règlement général sur la protection des données (RGPD) de l'Union européenne (UE)[6][15].

Les principes de protection des données s'appliquent à toute information concernant une personne physique identifiée ou identifiable.

Dans ce contexte, de nombreux secteurs d'activité, organismes gouvernementaux et entreprises ou organisations privées et publiques doivent cacher (masquer),[16] supprimer, anonymiser ou pseudonymiser les données à caractère personnel avant que le texte contenant ces données ne soit traité.[4][8]

Le présent document fournit des principes et une méthodologie permettant de détecter et d'identifier des données à caractère personnel afin qu'elles puissent être cachées ou supprimées, c'est-à-dire protégées avant la transmission et/ou le traitement d'un texte contenant de telles données. La difficulté ne réside pas tant dans la suppression ou le masquage des données, mais dans la reconnaissance des données à caractère personnel dans du texte écrit. Contrairement aux données à caractère personnel contenues dans un texte, les données à caractère personnel contenues dans des données structurées (par exemple présentées dans des tableaux) ne posent pas de réel problème, car ces données sont facilement reconnaissables.[5]

Le présent document s'adresse aux micro-entreprises, aux PME et aux grandes entreprises nationales et internationales, ainsi qu'aux organismes privés et publics qui traitent du texte pouvant contenir des données à caractère personnel dans tous les domaines (par exemple, le droit, la finance, la santé), dans toutes les langues et dans tous les pays.[14] Les principes et la méthodologie sont déjà utilisés par l'industrie et les organismes gouvernementaux.

En vertu des réglementations telles que le RGPD européen, la protection de données à caractère personnel représente un défi considérable pour les micro, petites, moyennes et grandes entreprises, et également pour les organismes privés et publics. Par exemple, le RGPD interdit le transfert de données à caractère personnel de personnes concernées européennes vers des pays situés en dehors de l'EEE, dénommés « pays tiers », à moins que les garanties appropriées ne soient imposées ou que les réglementations du pays tiers concernant la protection des données ne soient formellement considérées comme adéquates par la Commission européenne. En outre, l'État de Californie, aux États-Unis, a adopté le 28 juin 2018 le California Consumer Privacy Act, qui prend effet au 1er janvier 2020 et accorde des droits à la transparence et au contrôle de la collecte d'informations personnelles par les entreprises d'une manière similaire au RGPD (voir la Référence [2] et l'ISO/IEC 27701).

Tous les exemples donnés dans le présent document sont fictifs, mais ils pourraient exister si des données réelles étaient substituées aux données fictives.

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

1 Domaine d'application

Le présent document définit les principes de base et la méthodologie pour reconnaître des données à caractère personnel dans du texte libre, dans différentes langues (qu'elles soient agglutinantes, flexionnelles ou isolantes) et pays.

Le présent document est applicable essentiellement à la protection des données humaines circulant dans les industries nationales et internationales, et dans les organisations privées et publiques.

Le présent document s'applique au traitement par des êtres humains et/ou au traitement automatisé, ainsi qu'à divers domaines (par exemple, le droit, la finance, la santé).

Il ne s'applique pas au traitement automatisé des images.

Le présent document n'utilise que des méthodes formelles, les méthodes statistiques étant de nature très différente.

2 Références normatives

Le présent document ne contient aucune référence normative.

3 Termes et définitions

Pour les besoins du présent document, les termes et définitions suivants s'appliquent.

L'ISO et l'IEC tiennent à jour des bases de données terminologiques destinées à être utilisées en normalisation, consultables aux adresses suivantes:

— ISO Online browsing platform: disponible à l'adresse https://www.iso.org/obp

— IEC Electropedia: disponible à l'adresse https://www.electropedia.org/

3.1
intension
contenu interne d'un terme ou concept qui constitue sa définition formelle

Note 1 à l'article: L'extension est la gamme d'applicabilité d'un concept en nommant les objets particuliers qu'il dénote.

3.2
données à caractère personnel
toute information relative à une *personne physique identifiable* (3.6) ou identifiée

[SOURCE: : Règlement (UE) 2016/679,[6] Article 4 (1)]

3.3
pseudonymisation
traitement (3.4) de *données à caractère personnel* (3.2) de telle façon que celles-ci ne puissent plus être attribuées à une personne concernée précise sans avoir recours à des informations supplémentaires, pour autant que ces informations supplémentaires soient conservées séparément et soumises à des mesures techniques et organisationnelles afin de garantir que les données à caractère personnel ne sont pas attribuées à une *personne physique* identifiée ou *identifiable* (3.6)

[SOURCE: : Règlement (UE) 2016/679,[6] Article 4 (5)]

3.4
traitement
toute opération ou tout ensemble d'opérations effectuées ou non à l'aide de procédés automatisés et appliquées à des données ou des ensembles de *données à caractère personnel* (3.2), telles que la collecte, l'enregistrement, l'organisation, la structuration, la conservation, l'adaptation ou la modification, l'extraction, la consultation, l'utilisation, la communication par transmission, la diffusion ou toute autre forme de mise à disposition, le rapprochement ou l'interconnexion, la limitation, l'effacement ou la destruction

[SOURCE: : Règlement (UE) 2016/679,[6] Article 4 (2)]

3.5
sème
signifié de Saussure avec ses différents signifiants (instanciations) dans le texte

Note 1 à l'article: Saussure a été le premier à utiliser les termes « signifié » et « signifiant ». Saussure a proposé un modèle « dyadique » ou en deux parties du signe. Il a défini le signe comme étant composé d'un « signifiant » et d'un « signifié » (voir les Références [17] et [18]).

3.6
personne physique identifiable
personne concernée
personne qui peut être identifiée, directement ou indirectement, notamment par référence à un identifiant

Note 1 à l'article: Un identifiant peut être un nom, un numéro d'identification, des données de localisation ou un identifiant en ligne d'une personne physique. D'autres exemples exclus des exemples du présent document sont des références à un ou plusieurs facteurs spécifiques à l'identité physique, physiologique, génétique, mentale, économique, culturelle ou sociale de la personne physique.

[SOURCE: : Règlement (UE) 2016/679,[6] Article 4 (1)]

3.7
indicateur
présence significative d'une interaction entre des phénomènes lexicaux, morphologiques et syntaxiques ou l'un de ces phénomènes dans un large éventail de langues ou dans un petit nombre de langues ou dans une seule langue permettant l'identification des *données à caractère personnel* (3.2)

4 Raisons en faveur d'une communication humaine contrôlée

La première étape de la protection des données à caractère personnel consiste à pouvoir reconnaître automatiquement ces données, en particulier lorsqu'elles ne sont pas structurées, mais se présentent sous forme de texte libre, comme le montre l'Exemple 1 de l'Article A.1.

Une fois que les données sont détectées ou reconnues comme étant des données à caractère personnel, différents moyens peuvent être utilisés pour les dissimuler dans le texte : elles peuvent être cachées (masquées), supprimées, anonymisées (voir les Références [9] et [10]) ou pseudonymisées (voir la Référence [7]), comme le montre l'Exemple 2 de l'Article A.2.

Les Exemples 3 et 4 des Articles A.3 et A.4 montrent un exemple similaire en français.

5 Principes de base et méthodologie

5.1 Généralités

Pour les principes de base, divers phénomènes linguistiques lexicaux, morphologiques et syntaxiques sont utilisés, notamment en ce qui concerne la manière dont les données à caractère personnel sont représentées sous forme de texte libre. Par exemple, les adresses ne respectant pas le format anglais au Royaume-Uni, comme indiqué dans l'Exemple 1 de l'Article A.1, c'est-à-dire « Stoneham-le-Willows at 24 Britanny Park, F2 7AN (GB29 NWBK 6016 1331 9268 19) ».

La méthodologie spécifie des représentations formelles conçues en intension (voir les Références [11], [12] et [13]) basées sur des phénomènes lexicaux, morphologiques et syntaxiques qui doivent s'appliquer dans un ordre séquentiel à chacun des niveaux d'analyse linguistique qui ont un impact sur la reconnaissance des données à caractère personnel.

Les principes de base et la méthodologie sont formulés spécifiquement pour leur conférer un pouvoir explicatif afin de montrer comment et quand chacun des phénomènes linguistiques, lexicaux, morphologiques et/ou syntaxiques, et/ou leurs associations et interactions, en contexte, doivent être utilisés et appliqués en fonction de la reconnaissance des différents sèmes dans l'analyse des textes. Par conséquent, la méthodologie utilise des phénomènes linguistiques conformes aux principes de base pour la reconnaissance des données à caractère personnel (au lieu d'un lexique) et spécifie un système de règles de contrainte complété par un algorithme qui, lorsqu'il est appliqué, permet d'extraire les données à caractère personnel.

5.2 Aspects spécifiques

La principale difficulté réside dans la reconnaissance des données à caractère personnel dans des textes libres rédigés dans des langues différentes, provenant de pays différents et issus de domaines différents (par exemple, droit, finance, santé).

Le problème concerne également l'utilisation de la même langue dans différents pays. Par exemple, les adresses ne sont pas écrites de la même manière en France, en Suisse, en Belgique et au Canada, ou en Autriche et en Allemagne (voir les Exemples 5 et 6 des Articles A.5 et A.6). Un texte rédigé dans une langue peut également contenir des données à caractère personnel dans d'autres langues et dans d'autres pays.

Le problème est donc quadruple :

a) les données à caractère personnel dans du texte libre ;

b) les données à caractère personnel dans des pays différents ;

c) les données à caractère personnel dans des langues différentes ;

d) les données à caractère personnel dans des domaines différents.

La méthodologie décrite dans le présent document utilise des indicateurs pour détecter des données à caractère personnel telles qu'un numéro de téléphone, un numéro d'identification ou de compte bancaire, une adresse, etc.

La méthodologie fonctionne en intension. De ce fait, elle est basée sur des règles : elle établit un système de règles de contraintes ordonnées fondées sur des phénomènes linguistiques, qui sont conformes aux principes de base à suivre (voir 5.3). Les indicateurs doivent être établis et ils doivent être lexicaux, morphologiques et/ou syntaxiques.

5.3 Principes

5.3.1 Vue d'ensemble

Les principes abordant les quatre problématiques spécifiques listées en 5.2 élaborent des règles pour une représentation linguistique explicative formelle avec son propre métalangage et sa grammaire accompagnée d'exemples afin que de nouveaux sèmes et de nouvelles langues puissent être traités.

L'utilisateur doit établir les indicateurs lexico-morpho-syntaxiques. Les indicateurs, s'ils n'existent pas déjà dans la grammaire du métalangage, doivent être ajoutés.

5.3.2 Indicateurs lexicaux, morphologiques et syntaxiques

5.3.2.1 Généralités

Certains indicateurs doivent être des ensembles valables pour différents sèmes dans différentes langues et dans différents pays tels que N ou A, lesquels apparaissent très souvent dans le texte, mais ils peuvent également être facultatifs. Ils doivent être représentés formellement dans la grammaire du métalangage comme suit (extrait) :

— N = chiffre (0…9) ;

— L = A, B, …Z (sans signes diacritiques) ;

— A = 0,1, …9,L (L signifie un élément du L ci-dessus) ;

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

— ListB = ListA, 13, ...100 ;

— () = élément(s) entre parenthèses facultatif(s) ;

— M = tout mot commençant par une lettre majuscule ou minuscule suivi de zéro ou de plusieurs lettres majuscules ou minuscules ainsi que des caractères ‘’°- ;

— UM = M commençant par une lettre majuscule ;

— n = nouvelle ligne suivie ou non d'un ou de plusieurs espaces et/ou d'une ou de plusieurs tabulations ;

— listC = 1, …9999 ;

— l = a, b, …z (sans signes diacritiques) ;

— |tout caractère avec le caractère | codé comme || terminé par | = littéral.

NOTE Les séquences de littéraux sont illégales (ambiguïté de ||).

Certains indicateurs doivent être ajoutés selon le sème à reconnaître, le pays ou la langue utilisée (voir les exemples de l'Annexe A).

L'Exemple 3 de l'Article A.3 contient les adresses suivantes rédigées dans différents 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

La représentation algorithmique formelle (diagramme représentant l'algorithme) permettant de reconnaître toute adresse est la suivante :

  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)

À l'aide des éléments (indicateurs) de la grammaire générale du métalangage et des indicateurs complémentaires suivants :

— listE = bis,ter,Bis,Ter ;