A

WIPO
المنظمة العالمية
للملكية الفكرية

CWS/12/16
الأصل: بالإنكليزية
التاريخ: 5 أغسطس 2024

اللجنة المعنية بمعايير الويبو

الدورة الثانية عشرة
جنيف، من 16 إلى 19 سبتمبر 2024

اقتراح بشأن معيار جديد للويبو يدعم تنقية بيانات الأسماء

وثيقة من إعداد المشرفَين المشاركَين على فرقة العمل المعنية بتوحيد الأسماء

ملخص

1. تقدّم فرقة العمل المعنية بتوحيد الأسماء مسودة نهائية لمعيار جديد للويبو يدعم تنقية بيانات الأسماء، لكي تنظر فيها اللجنة المعنية بمعايير الويبو (لجنة المعايير) في دورتها الثانية عشرة وتعتمدها.

معلومات أساسية

2. وافقت لجنة المعايير، في دورتها الحادية عشرة المعقودة في عام 2023، على الوصف المنقح للمهمة رقم 55، الذي أصبح نصه الآن كما يلي:

> *"إعداد مقترح الإجراءات المستقبلية الرامية إلى تحقيق توحيد الأسماء في وثائق الملكية الفكرية، بنية وضع معيار للويبو يساعد مكاتب الملكية الفكرية على تحسين "الجودة في المنبع" فيما يتعلق بالأسماء."*
>
> (انظر(ي) الفقرات 75 إلى 78 من الوثيقة CWS/11/28).

3. ويمكن الاطلاع على المزيد من التفاصيل حول تاريخ فرقة العمل والتقدم الذي أحرزته منذ الدورة الأخيرة للجنة المعايير في الوثيقة CWS/12/8.

4. وناقشت لجنة المعايير في دورتها الحادية عشرة المعقودة في عام 2023 مجموعة جديدة من المبادئ التوجيهية لدعم تنقية أسماء المودعين التي قدمتها فرقة العمل المعنية بتوحيد الأسماء. ووافقت لجنة المعايير على استخدام مصطلح "التوصيات" عوضاً عن "المبادئ التوجيهية" في اسم معيار الويبو الجديد المقترح، إذ اعتبرته أكثر وضوحاً من حيث النطاق. وأحاطت لجنة المعايير علماً كذلك باقتراح الأمانة بشأن الاسم: "معاير الويبو ST.93" (انظر الفقرة 135 من الوثيقة CWS/11/28).

5. ولكن لجنة المعايير لم تعتمد المعيار المقترح وأحالته إلى فرقة العمل من أجل مواصلة من المناقشة والتحسين. وعلاوة على ذلك، أشارت لجنة المعايير إلى أن الأمانة ستدرس إمكانية نشر مجموعة من جداول النقل الحرفي على موقع الويبو الإلكتروني (انظر(ي) الفقرتين 136 و137 من الوثيقة CWS/11/28).

اقتراح لمعيار جديد

6. تواجه مكاتب الملكية الفكرية مشاكل في القدرة على تحديد أفراد الأسر داخل أسر البراءات، إذ قد تُستخدم أسماء مختلفة للمودعين داخل أسرة البراءات نفسها. وإضافة إلى ذلك، قد تحدث أخطاء إملائية أو مطبعية عند كتابة أسماء المودعين. والرغبة في الحصول على بيانات واضحة لأسماء المودعين لأغراض إحصائية هي رغبة مقبولة تماماً.

7. وقد أعدت فرقة العمل المعنية بتوحيد الأسماء، في إطار المهمة رقم 55، اقتراحاً نهائياً من أجل وضع معيار جديد للويبو لدعم تنقية الأسماء وتحقيق بيانات نظيفة للمودعين. ويرد هذا الاقتراح في مرفق هذه الوثيقة.

الأهداف

8. يكمن الهدف من هذه التوصيات في تقديم توجيهات عامة رفيعة المستوى. ويعني وجود أوجه تباين في عوامل مثل المتطلبات القانونية وممارسات البيانات والغرض من التنقية والاستخدام المزمع للبيانات والمتطلبات من الموارد والاعتبارات التقنية، أنه لا يوجد نهج واحد يعمل على النحو الأفضل في جميع مكاتب الملكية الفكرية. وتبرِز هذه التوصيات الممارسات العامة التي يمكن تطبيقها في أي مكتب من مكاتب الملكية الفكرية دعماً لتنقية بيانات أسماء العملاء، وتحسّن بالتالي توحيد الأسماء وتقنيات مطابقة الأسماء للمستخدمين النهائيين.

النطاق

9. يقدّم المعيار المقترح توصيات عامة بشأن تلقي بيانات الأسماء النظيفة ومعالجتها وتنقيتها ونشرها. ولا يقدم هذا المعيار توصيات بشأن التفاصيل المتعلقة بالنُهج المتبعة في تنقية البيانات أو تحديد الأسماء أو تحويلها مثل النقل الحرفي أو النسخ أو الترجمة، أو النُهج المتبعة في توحيد الأسماء مثل اختيار الخوارزميات، وأين ومتى يتم تطبيق عمليات التحويل، أو التواتر، أو استراتيجيات الدمج.

10. ويتبع المعيار المقترح الهيكلية التالية:

– المتن الرئيسي: يحدد التوصيات العامة لمعالجة أسماء المودعين من أجل تحقيق بيانات نظيفة؛

– المرفق: يقدّم أمثلة على النقل الصوتي والنقل الحرفي والترجمة لدعم التوصيات الواردة في المتن الرئيسي.

11. ويُقترح الاسم التالي لمعيار الويبو الجديد:

"معيار الويبو ST.93 - توصيات بشأن تنقية بيانات الأسماء"

التغييرات الحاصلة منذ المسودة الأخيرة

12. في ضوء المناقشات التي دارت حول الاقتراح المتعلق بتنقية بيانات الأسماء والمداخلات المعلقة المقدمة من عدة وفود في الدورة الحادية عشرة للجنة المعايير، نقحت لجنة المعايير المسودة الأصلية للمبادئ التوجيهية المقترحة (انظر مرفق الوثيقة CWS/11/23). وأُجريت التغييرات التالية:

– تلاحظ فرقة العمل أن التعريف السابق لمصطلح "البيانات النقية" بأنها "خالية من الأخطاء والتكرار" كان إشكالياً لأنه من غير الواقعي ضمان خلو البيانات من الأخطاء والتكرار بنسبة 100%. وبالتالي، وافقت فرقة العمل على تعديل تعريف مصطلح "البيانات النقية" بحيث يصبح كالآتي: *"تعني البيانات الدقيقة والمتسقة والموثوقة. ونظراً إلى صعوبة قياس درجة النقاوة في مجموعة بيانات كبيرة ومعقدة، فقد تُستخدم مقاييس مختلفة كبدائل للنقاوة أو الخصائص ذات الصلة، مثل الملاءمة للغرض".*
– في قسم "تحويل الأسماء"، وافقت فرقة العمل على تغيير مصطلح "القلب" إلى "التحويل" لكي يتوافق بشكل أفضل مع عنوان القسم ويتيح تفسيراً أكثر مرونة.
– وفي قسم "المراجع"، ناقشت فرقة العمل الإشارة إلى معايير المنظمة الدولية لتوحيد المقاييس فيما يخص كتابة اللغات المختلفة بالحروف اللاتينية، كما اقترح المكتب الدولي. وخلصت فرقة العمل إلى أن المعيار المقترح يجب أن يتضمّن فقط معايير الويبو ذات الصلة كنهج عام، عوضاً عن دمج معايير ذات الصلة صادرة عن المنظمة الدولية لتوحيد المقاييس، لأنّ مكاتب الملكية الفكرية قد لا تتبع معايير المنظمة الدولية لتوحيد المقاييس بشكل مستمر وقد تغيّر ممارساتها مع مرور الوقت.

13. وفيما يتعلق بجداول النقل الحرفي التي تستخدمها مكاتب الملكية الفكرية، تشير فرقة العمل إلى أن الهدف الرئيسي هو توفير مرجع للمناقشات المعقولة مع المودعين، وليس تغيير قاعدة البيانات بأكملها وفقاً لجداول النقل الحرفي. وقد طُلب من المكاتب الأعضاء في فرقة العمل تقديم جداول النقل الحرفي الخاصة بها، إذا كانت متاحة، حتى يتمكن المودعون أو الممثلون أو مكاتب الملكية

الفكرية من الرجوع إلى الجداول الصادرة عن مكاتب الملكية الفكرية الأخرى التي تستخدم لغات مختلفة عند تقديم الأسماء أو تنقية بيانات الأسماء. ومن المقترح أن تطلب لجنة المعايير من الأعضاء فيها تقديم جداول النقل الحرفي الخاصة بها. ومن المقترح أيضاً نشر جداول النقل الحرفي التي تقدمها مكاتب الملكية الفكرية في الجزء 7 من دليل الويبو.

14. وإذا اعتمدت لجنة المعايير المعيار الجديد في دورتها الحالية، فمن المقترح أن تطلب لجنة المعايير من الأمانة نشر هذه التوصيات في الجزء 3 من دليل الويبو.

15. *إن لجنة المعايير مدعوة للقيام بما يلي:*

(أ) الإحاطة بمضمون هذه الوثيقة ومرفقها؛

(ب) والنظر في اسم معيار الويبو الجديد، على النحو المذكور في الفقرة 11 أعلاه، والموافقة عليه؛

(ج) والنظر في معيار الويبو الجديد ST.93 واعتماده، على النحو المشار إليه في الفقرات من 8 إلى 10 أعلاه، وعلى النحو الوارد في مرفق هذه الوثيقة؛

(د) والطلب من الأمانة نشر معيار الويبو الجديد ST.93 في الجزء 3 من دليل الويبو، على النحو المشار إليه في الفقرة 14 أعلاه؛

(هـ) والطلب من الأمانة إصدار تعميم يدعو المكاتب إلى تقديم جداول النقل الحرفي الخاصة بها، ونشر هذه الجداول في الجزء 7 من دليل الويبو، على النحو المذكور في الفقرة 13 أعلاه.

[يلي ذلك المرفق]

WIPO STANDARD ST.93

RECOMMENDATIONS ON THE DATA CLEANING OF NAMES

Proposal presented for approval by the Committee on WIPO Standards (CWS)
at its twelfth session

WIPO STANDARD ST.93

RECOMMENDATIONS ON NAME DATA CLEANING

Proposal presented for adoption by the Committee on WIPO Standards (CWS) at its twelfth session

INTRODUCTION

1. This Standard provides general recommendations on the intake, processing, cleaning, and publication of clean name data. This Standard does not provide recommendations on details in relation to approaches to data cleaning, name localization or transformation such as transliteration, transcription or translation, or approaches to name standardization such as selection of algorithms, where and when transformations are applied, frequency, or merging strategies. Decisions on those details will vary greatly depending on the party applying them, the purpose of transformations, and the quickly evolving nature of matching algorithms.

2. WIPO Standard ST.20 should be referred to for recommendations to produce indexes to patent documents giving names of applicants and other customers, and to promote a uniform presentation of names occurring in name indexes as well as a uniform method of ordering the names in the index itself.

DEFINITIONS

3. In the context of this document:

(a) "IPO" refers to an Intellectual Property Office, which manage the application and registration process for intellectual property rights.

(b) "Customer data" means data on applicants, registrants, owners, legal representatives, or other parties held by an IPO in connection with an IP right, application, registration, or other instrument. This standard is primarily concerned with customer name data: personal names, business names, and related information such as city, address, or email that can be used to disambiguate potential name matches.

(c) "Clean data" means data that is accurate, consistent and reliable. As the degree of cleanness in a large complex data set is difficult to measure, various metrics may be used as proxies for cleanness or related properties, such as fitness for purpose.

(d) "Transliteration" means the mapping of source language character(s) to target language (phonetic) character(s).

(e) "Transcription" means the mapping of a source language character/logogram/syllable/phoneme to something that corresponds to the sound in the respective system of the target language.

(f) "Translation" represents the meaning of a word or concept in the source language with something that corresponds to the meaning in the target language.

INTAKE

4. IPOs may provide the ability for customers to create and manage electronic customer records containing published name information: personal names, business names, names of legal representatives, and related information such as city, address, or email.

5. IPOs should allow a customer record to be associated with multiple applications or registrations for IP rights, so that customers may reuse the same name information for multiple applications or registrations and update their name information in one place.

6. IPOs may provide a form(s) which customers use to request the IPOs to create or change their name or related information. IPOs may also allow customers to enter and update their name or related information themselves, or may require a designated party such as employees, contractors, or an external service to enter and update customer records at the customer's request.

7. Multiple records for one customer may be created and managed by different entities, such as different legal representatives. IPOs should consider this when designing their customer record systems, as multiple records for a single customer may contain slight variations of the same data or be updated at different times by different representatives.

8. IPOs may support entry of the customer's name in native characters of the customer's language, in addition to the customer's name in the language(s) of operation for an IPO, which should be stored using UTF-8 [1]encoding. For instance, an IPO that works in English could allow separate fields for an applicant name in English and the original applicant name in Korean.

9. IPOs may optionally use identification numbers to identify customers. Identification numbers may be created by the IPO or used from an external source, such as a registered business number or passport number. Identification numbers alone do not resolve issues with clean customer data, such as duplicate entries, name changes, and outdated or incorrect information. IPOs using identification numbers should continue to pay attention to and address the considerations in other parts of this Standard.

TRANSFORMATION OF NAMES

10. For data exchange and processing, including the receipt of international applications or registrations, IPOs may consider the name transformation (see the Annex to this document). It is recommended that IPOs should send and receive name data using UTF-8 encoding.

11. It should be noted that the localization or conversion of customer names is extremely error prone as there are no generally accepted or uniformed standards. For localization or transformation of names, there are three ways referred to in this Standard: transliteration, transcription and translation. If IPOs transliterate, transcribe or translate characters from one language (such as Greek) to another (such as English), they should publish their scheme of transliteration, transcription or translation. The transliterated, transcribed or translated document, or parts of the document, should be made available to the customer for review and customers should have a way to submit corrections if the transliteration, transcription or translation is flawed.

12. Reverse transliteration should be avoided if possible, instead it is recommended to use the original name instead. For instance, an application filed by "Phony Corp" might be transliterated to Greek characters as "Φονι Κορπ" in an IPO system, and on publication might be reverse transliterated from Greek back to Latin characters as "Foni Corp", leading to mismatches. Examples of common issues arising from reverse, or re-transliteration, re-transcription or re-translation are available in the Annex to this Standard.

VALIDATION AND DISAMBIGUATION

13. Validation and disambiguation approaches should be designed to meet specific objectives, either administrative or statistical, and appropriate methods applied given the objectives. Approaches to name matching and disambiguation should be appropriately scoped and risk assessed given their design objective to ensure appropriate levels of disambiguation are achieved for the use case.

14. IPOs may choose to perform validation of submitted customer information, including automated checks. Validation results should be made available to the customer, and corrections accepted by the customer if needed, including ways to bypass an automated validation mechanism, in case it provides incorrect or incomplete results.

15. IPOs attempting to disambiguate name records (i.e., find duplicate entries) may wish to consider more than just the customer names. Names are not inherently unique. For example, there may be multiple individuals named "John Smith" or multiple companies named "Data Corp". Comparing related data points such as city, post code, birthdate, or other information, where available, can increase the likelihood of successful matches.

16. Any validation or disambiguation process initiated by the IPO that potentially could have legal effects, such as correcting or standardizing the name of the registered owner of an IP right, should be confirmed by the customer before the change is made in the IPO's system.

MAINTENANCE

17. IPOs should develop a strategy to periodically clean data in customer name databases, including searching for and attempt to resolve duplicate records, i.e., multiple records for the same entity. In some instances, the duplicates may be merged or combined, for instance, records with slight unintentional differences in spelling such as "ABC Corp" and "ABC Corp.". In other instances, maintaining separate records might be preferable. Each IPO should decide what approach fits best for their own name record management system. The strategy may include the involvement of the concerned customers of the records in the data cleaning process and the responsibility of the cleaned data.

18. IPOs should provide a mechanism for customers to update their name information on multiple applications or IP rights by entering the information once. For instance, this could be achieved by associating each application or IP right with a single customer record containing name information, or by allowing customers to select multiple applications or IP rights and submit one instance of updated name information to be applied to all of them.

[1] UTF-8 is an encoding system for Unicode.

19. IPOs may designate someone to be responsible for cleaning data issues, including development of metrics for measuring clean data, regular monitoring and reporting of those metrics, and taking action to improve customer data when needed.

PUBLICATION AND DATA EXCHANGE

20. IPOs should make available updates to name information that are made after an IP right has published. For instance, if “ABC Corp” changes their name to “XYZ Corp” in their customer record, then the name “XYZ Corp” should be associated with the IP right in online publications. The original name may also appear on the published IP right, according to legal requirements of the IPO.

21. If an IPO has other forms of a customer name, such as original name expressed using native characters, these should be included in published data and the data exchanged with other IPOs.

22. If an IPO uses identification numbers to identify entities, the numbers should be included in published data and data exchanged with other IPOs. If the identification numbers are sensitive and cannot be shared, then the IPO should indicate which customer data uses these identification numbers, such as by replacing the sensitive numbers with generated unique numbers for publication.

STATISTICAL PURPOSES

23. For statistical purposes, IPOs may attempt to match customer data with variations in customer names, or other fields, to achieve counts that are more accurate. In such cases, IPOs should publish their matching strategy or algorithm along with the statistical results so others can understand the methodology used.

REFERENCES

24. References to the following Standard are of relevance to this Standard:

WIPO Standard ST.20	Preparation of name indexes to patent documents

[Annex to the proposed Standard follows]

ANNEX

DIFFERENT MEANS OF NAME TRANSFORMATION

Although transliteration and transcription are different concepts from a linguistic perspective, the result is usually very similar for character-based writing systems. However, transcription provides a more practical result, because only standard characters from the target language are required for the conversion.

As English is a language that is adopted as a common language between speakers whose native languages are different, it is generally overlooked that transcription is rarely standardized between any pair of languages. In the best case there are official definitions for [xx] -> [en] leading to the assumption that [xx] -> [en] -> [yy] is equal to [xx] -> [yy], which is usually not correct.

TRANSLITERATION EXAMPLES[2]:

Figure 1 shows below an example of letter correspondence and remarks regarding this transliteration.

Source and Target words	Letter Correspondence				Description
English to Persian					
John /dʒɒn/	J	o	h	n	*h* is a silent letter (no sound
	|	|		|	is associated to the letter)
جان /dʒɒn/	ج	ا		ن	and is not transliterated
Arabic to English					
نجیب /nædʒi:b/	ن	ج	ی	ب	short vowel /æ/ on N is
	|	|	|	|	normally not written in
Najib /nædʒi:b/	Na	j	i	b	Arabic script
English to Japanese					
Bill /bi:l/	B	i	l	l	each syllable in Japanese is a
	\	/	\	/	consonant-vowel sequence
ビル [bi-ru]	ビ		ル		
English to Hindi					
Adam /'ædəm/	A	d	a	m	the second "a" is not
	|	|		|	transliterated in Hindi
अदम /'ædəm/	अ	द		म	

Figure 1: Transliteration example

[2] Machine Transliteration Survey https://www.researchgate.net/figure/Transliteration-examples-in-four-language-pairs-Letter-correspondence-shows-how-the_fig1_220566444

TRANSCRIPTION EXAMPLES:

Shown below are examples where transcription can lead to inaccuracies:

[ru]: Ш → [de]: sch[3]

[ru]: Ш → [en]: sh

[ko]: ㅑ → [de]: ja[4]

[ko]: ㅑ → [en]: ya

[gr] : Ω → latin: O[5]

[da]: Æ → [de]: Ä or AE, [en]: AE [6]

TRANSLATION EXAMPLES:

In the first example, it is clear that the direct translation can lead to issues:

[de]: Aktiengesellschaft → [en]: corporation, stock co, …

[ru]: ОАО Силовы́е маши́ны → [en] : OJSC “Power Machines” - OR - [en]: Open Joint-stock Company “Power Machines”

A second example below, which demonstrates typical borderline cases of the Romanization of a Chinese company name shown in Figure 2 are:

- [zh]: 北京东土科技股份有限公司 → [en] transliterated (pinyin): běi jīng dōng tǔ kē jì gǔ fèn yǒu xiàn gōng sī ;
- [zh]: 北京东土科技股份有限公司 → [en] transcribed (pinyin): beijing dongtu keji gufen youxian gongsi
- [zh]: 北京东土科技股份有限公司 → [en] translated (English): Beijing, China Science and Technology Joint-stock Limited Company
- [zh]: 北京东土科技股份有限公司 → in reality : Kyland Technology Co., Ltd.

(71) 申请人：北京东土科技股份有限公司(KYLAND TECHNOLOGY CO., LTD) [CN/CN]；中国北京市石景山区实兴大街30号院2号楼8层901, Beijing 100041 (CN)。

Figure 2: Romanization of Chinese company name

[End of Annex to the proposed Standard and of Standard]

[End of Annex and the document]

[3] https://de.wikipedia.org/wiki/Kyrillisches_Alphabet#Russisch

[4] https://de.wikipedia.org/wiki/Koreanisches_Alphabet

[5] https://en.wikipedia.org/wiki/Romanization_of_Greek

[6] https://en.wikipedia.org/wiki/Dania_transcription