WIPO
世界
知识产权
组织

CWS/12/16 REV.
原文：英文
日期：2024年9月20日

产权组织标准委员会（标准委）

第十二届会议
2024年9月16日至19日，日内瓦

关于支持名称数据清理的产权组织新标准的提案

名称标准化工作队共同牵头人编拟的文件

摘 要

1.. 名称标准化工作队提交了关于支持名称数据清理的新产权组织标准的最终草案，供产权组织标准委员会（标准委）第十二届会议审议和通过。

背 景

2.. 在2023年举行的第十一届会议上，标准委批准了经修订的第55号任务说明：

> “就旨在实现知识产权文献中名称标准化的未来行动编写提案，以期制定一项产权组织标准，帮助知识产权局更好地从源头确保名称的质量。”
>
> （见文件CWS/11/28第75至78段。）

3.. 有关工作队的历史以及自标准委上届会议以来所取得进展的更多详情，见文件CWS/12/8。

4.. 在2023年举行的第十一届会议上，标准委审议了名称标准化工作队提出的一套支持清理申请人名称的新指导方针。标准委商定在拟议的产权组织新标准名称中使用“建议”而非“指导原则”一词，认为这样范围更明确。标准委还注意到秘书处建议的名称：“产权组织标准ST.93”（见文件CWS/11/28第135段）。

5.. 然而，标准委并未通过拟议的标准，而是将其发回工作队作进一步讨论和改进。此外，标准委指出，秘书处将调查在产权组织网站上公布音译表集的可能性。（见文件 CWS/11/28 第 136 段和第 137 段）。

关于新标准的提案

6.. 知识产权局在识别专利族中的成员时会遇到问题，因为同族成员可能会使用不同的申请人名称。此外，在输入申请人名称时可能会出现拼写或排印错误。为统计目的提供清洁申请人名称数据这一愿望已被广泛接受。

7.. 名称标准化工作队在第 55 号任务的框架下，编写了关于支持名称清理工作的新产权组织标准的最终提案，以实现清洁申请人数据。该提案载于本文件附件。

目 标

8.. 这些建议旨在提供一般性和高层次指导。由于法律要求、数据做法、清理目的、数据的预期用途、资源要求和技术考虑等因素各不相同，因此没有对于所有知识产权局都最适用的单一方法。这些建议反映的是可适用于任何知识产权局的一般性做法，以支持清理客户名称数据，进而增进下游用户的名称标准化和名称匹配技术。

范 围

9.. 拟议标准就清洁名称数据的接收、处理、清理和公布提出了一般性建议。该标准不提供与数据清理方法、名称本地化或转换（例如音译、转录或翻译）或名称标准化方法（例如算法的选择、应用转换的位置和时间、频率或合并策略）有关的细节建议。

10.. 拟议标准结构如下：

- 主体：确定处理申请人名称的一般性建议，以实现清洁数据；和
- 附件：提供有关音译、转录和翻译的实例，以支持主体部分提出的建议。

11.. 建议新产权组织标准采用以下名称：

“产权组织标准 ST.93——关于名称数据清理的建议”

自上一版草案以来所作修改

12.. 根据对名称数据清理提案的讨论情况和若干代表团在标准委第十一届会议上的未决发言，工作队修订了拟议指导原则的原始草案（见 CWS/11/23 附件）。修改内容如下：

- 工作队指出，以前将“清洁数据”定义为“无差错、零重复”是有问题的，因为保证数据百分之百无差错和零重复是不现实的。因此，工作队同意将“干净数据”的定义修改为“指的是准确、一致和可靠的数据。由于大型复杂数据集的清洁程度难以衡量，可使用多种衡量标准来替代清洁程度或相关属性，如适用性”。
- 在“名称转换”部分，工作队同意将“转化”（conversion）一词改为“转换”（transformation），以便更好地与该部分标题保持一致，并允许更灵活的解释。
- 在“参考文献”部分，按照国际局的建议，工作队讨论了参考 ISO 标准将各种语言罗马化的问题。工作队得出结论，作为一般性做法，拟议标准应仅包含相关产权组织标准，

而不是纳入相关 ISO 标准，因为知识产权局可能不会始终遵循 ISO 标准，并且可能随着时间的推移改变其做法。

13.. 关于知识产权局使用的音译表，工作队指出，主要目的是为与申请人进行合理讨论提供参考，而不是根据音译表改变整个数据库。要求工作队各局提供音译表（如果有的话），以便申请人、代理人或知识产权局在提交名称或清理名称数据时，可以参考使用不同语言的其他知识产权局的音译表。建议标准委要求其成员提供各自的音译表。还建议在《产权组织手册》第 7 部分公布知识产权局提供的音译表。

14.. 如果新标准在标准委本届会议上获得通过，建议标准委请秘书处在《产权组织手册》第 3 部分公布这些建议。

15.. 请标准委：

(a) 注意本文件及其附件的内容；

(b) 要求名称标准化工作队重新审查并改进产权组织标准 ST.93 草案；

(c) 鼓励主管局和知识产权行业提名专家参加名称标准化工作队；

(d) 要求国际局组织一次名称数据清理讲习班，所有相关方都可以参加；

(e) 要求其成员和观察员支持国际局，宣传这一讲习班。

[后接附件]

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

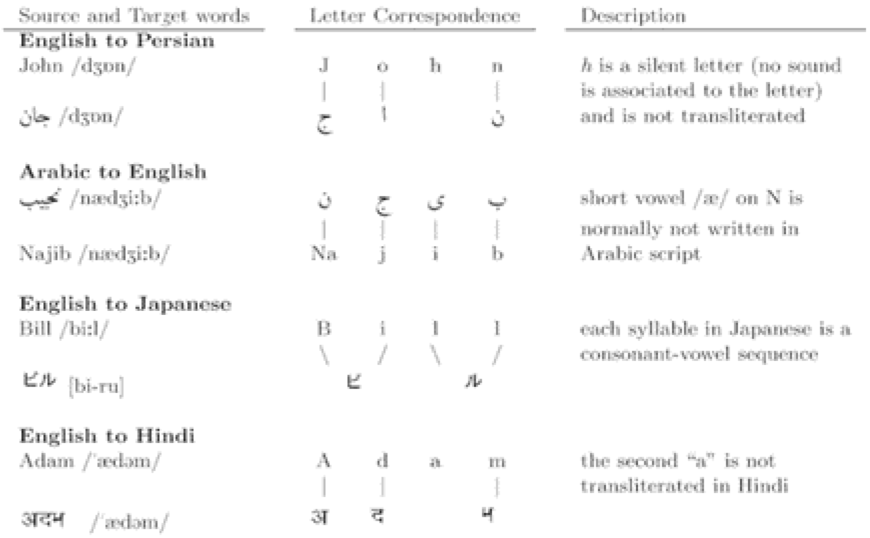
DIFFERENT MEANS OF NAME TRANSFORMATION

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

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:

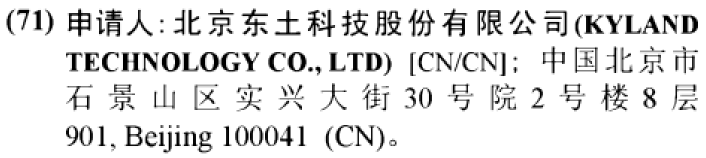
In the first example, it is clear that the direct translation can lead to issues:

[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