

Committee on WIPO Standards (CWS)

Twelfth Session Geneva, September 16 to 19, 2024

PROPOSAL FOR A NEW WIPO STANDARD SUPPORTING THE DATA CLEANING OF NAMES

Document prepared by the Name Standardization Task Force Co-Leaders

SUMMARY

1. The Name Standardization Task Force presents a final draft for a new WIPO standard on supporting the data cleaning of names, for consideration and adoption at the twelfth session of the Committee on WIPO Standards (CWS).

BACKGROUND

2. At its eleventh session in 2023, the CWS approved the revised description of Task No. 55 which reads:

“Prepare a proposal for future actions aimed at achieving the standardization of names in Intellectual Property (IP) documents, with the view to developing a WIPO standard to assist IP offices in providing a better “quality at source” in relation to names.”

(See paragraphs 75 to 78 of document CWS/11/28.)

3. Further details regarding the history of the Task Force and the progress made since the last session of the CWS can be found in document CWS/12/8.

4. At its eleventh session in 2023, the CWS considered a new set of guidelines to support the cleaning of applicant names presented by the Name Standardization Task Force. The CWS agreed to using the term ‘recommendations’ instead of ‘guidelines’ in the name of the proposed new WIPO standard, as it considered it clearer in scope. The CWS also noted the Secretariat’s proposal of the name: “WIPO Standard ST.93” (see paragraph 135 of document CWS/11/28).

5. However, the CWS did not adopt the proposed standard and referred it back to the Task Force for further discussion and improvement. Furthermore, the CWS noted that the Secretariat would investigate the possibility of publishing a collection of transliteration tables on the WIPO website (see paragraphs 136 and 137 of document CWS/11/28).

PROPOSAL FOR A NEW STANDARD

6. Intellectual Property Offices (IPOs) experience issues with being able to identify family members within a patent family, as different applicant names may be used within the same patent family. Additionally spelling or typographical errors may be made when entering the applicant names. The desire to have clean applicant name data for statistical purposes is well accepted.

7. The Name Standardization Task Force, under the framework of Task No. 55, has prepared a final proposal for a new WIPO standard to support the cleaning of names, to achieve clean applicant data. This proposal is provided as the Annex to the present document.

Objectives

8. These recommendations aim to provide general and high-level guidance. Variations in factors such as legal requirements, data practices, the purpose of the clean-up, the intended use of the data, resource requirements, and technical considerations mean that there is no single approach that works best across all IPOs. These recommendations reflect general practices that can be applied at any IPO to support cleaning customer name data, thereby enhancing name standardization and name matching techniques for downstream users.

Scope

9. The proposed 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.

10. The proposed Standard is structured as follows:

- Main body: defines the general recommendations for the processing of applicant names in order achieve clean data; and
- Annex: provides examples on transliteration, transcription and translation to support the recommendations provided in the Main Body.

11. The following name for the new WIPO Standard is proposed:

“WIPO Standard ST.93 – Recommendations on the data cleaning of names”

Changes made since last draft

12. In the light of the discussions on the proposal for name data cleaning and outstanding interventions from several delegations at the eleventh session of the CWS, the Task Force revised original draft of the proposed guidelines (see Annex to CWS/11/23). The following changes were made:

- The Task Force notes that the previous definition of “clean data” as “free from errors and duplication” was problematic because it is unrealistic to guarantee that data is 100 per cent free from errors and duplications. Consequently, the Task Force agreed to modify the definition of “clean data” to: “*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.”

- In the ‘Transformation of names’ section, the Task Force agreed to change term “conversion” to “transformation” to better align with the section title and allow for a more flexible interpretation.
- In the ‘References’ section, the Task Force discussed referencing ISO standards for the romanization of various languages, as suggested by the International Bureau. The Task Force concluded that the proposed Standard should include only relevant WIPO Standards as a general approach, rather than incorporating relevant ISO standards, because IPOs may not consistently follow ISO standards and may change their practices over time.

13. Regarding the transliteration tables used by IPOs, the Task Force notes that the main objective is to provide a reference for reasonable discussions with applicants, not to change the whole database according to the transliteration tables. The Task Force Offices were asked to provide their transliteration tables, if available, so that applicants, representatives, or IPOs can refer to the tables from other IPOs using different languages when submitting names or cleaning name data. It is proposed that the CWS request its Members to provide their transliteration table(s). It is also proposed to publish transliteration tables provided by IPOs in Part 7 of the WIPO Handbook.

14. If the new Standard is adopted by the CWS at the current session, it is proposed that the CWS requests the Secretariat to publish these recommendations in [Part 3 of the WIPO Handbook](#) .

15. *The CWS is invited to:*

(a) take note of the content of the present document and the Annex to the present document;

(b) consider and approve the name of the new WIPO Standard as indicated in paragraph 11 above;

(c) consider and adopt the new WIPO Standard ST.93 as referred to in paragraphs 8 to 10 above and as reproduced in the Annex to the present document;

(d) request the Secretariat to publish the new WIPO Standard ST.93 in Part 3 of the WIPO Handbook as referred to paragraph 14 above; and

(e) request the Secretariat to issue a circular inviting Offices to provide their transliteration table(s) and publish the provided transliteration tables in Part 7 of the WIPO Handbook as referred to in paragraph 13 above.

[Annex follows]

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*

Introduction.....	2
DEFINITIONS.....	2
INTAKE	2
TRANSFORMATION OF NAMES	3
VALIDATION AND DISAMBIGUATION.....	3
MAINTENANCE	3
PUBLICATION AND DATA EXCHANGE.....	4
STATISTICAL PURPOSES	4
References.....	4
ANNEX.....	1
Transliteration examples:.....	1
Transcription examples:.....	2
Translation examples:	2

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-8¹ 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 EXAMPLES²:

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	<i>h</i> is a silent letter (no sound is associated to the letter) and is not transliterated
جان /dʒɒn/	ج	ا		ن	
Arabic to English					
نجيب /nædʒiːb/	ن	ج	ي	ب	short vowel /æ/ on N is normally not written in Arabic script
Najib /nædʒiːb/	Na	j	i	b	
English to Japanese					
Bill /bi:l/	B	i	l	l	each syllable in Japanese is a consonant-vowel sequence
ビル [bi-ru]	\	/	\	/	
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]: sch³

[ru]: Ш → [en]: sh

[ko]: ㅏ → [de]: ja⁴

[ko]: ㅏ → [en]: ya

[gr] : Ω → latin: O⁵

[da]: Æ → [de]: Ä or AE, [en]: AE⁶

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 Силовые машины → [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