R

WIPO
ВСЕМИРНАЯ ОРГАНИЗАЦИЯ
ИНТЕЛЛЕКТУАЛЬНОЙ
СОБСТВЕННОСТИ

CWS/11/23
ОРИГИНАЛ: АНГЛИЙСКИЙ
ДАТА: 6 НОЯБРЯ 2023 ГОДА

Комитет по стандартам ВОИС (КСВ)

Одиннадцатая сессия
Женева, 4–8 декабря 2023 года

ПРЕДЛАГАЕМЫЕ ОБЩИЕ РЕКОМЕНДАЦИИ ПО ОЧИСТКЕ ДАННЫХ ОБ ИМЕНАХ

Документ подготовлен руководителями Целевой группы по стандартизации имен

СПРАВОЧНАЯ ИНФОРМАЦИЯ

1. На пятой сессии Комитета по стандартам ВОИС, состоявшейся в 2017 году, в программу работы Комитета была включена новая задача: задача № 55. На этой же сессии была учреждена Целевая группа по стандартизации имен, которой была поручена задача № 55, имевшая следующее описание:

> «В рамках изучения возможности разработки стандарта ВОИС, который бы помог ведомствам промышленной собственности (ВПС) обеспечить более высокое качество исходной информации, касающейся имен заявителей, подготовить предложение, касающееся дальнейших действий по стандартизации имен заявителей в документах по ИС, и представить его на рассмотрение КСВ».
>
> (См. пункты 85–88 документа CWS/5/22).

2. На своей седьмой сессии в 2019 году КСВ принял к сведению, что в ноябре 2018 года Секретариат распространил среди ведомств промышленной собственности (ВПС) обследование, посвященное использованию идентификаторов заявителей, как было согласовано на шестой сессии КСВ. Сославшись на то, что посвященное использованию идентификаторов обследование завершено, Целевая группа предложила пересмотреть описание задачи № 55 с отсылкой к данному обследованию. КСВ утвердил пересмотренное описание задачи № 55 в следующей редакции:

«В рамках изучения возможности разработки стандарта ВОИС, который бы помог ведомствам промышленной собственности (ВПС) обеспечить более высокое качество исходной информации, касающейся имен заявителей, подготовить предложение, касающееся дальнейших действий по стандартизации имен заявителей в документах по ИС, и представить его на рассмотрение КСВ».

(См. пункты 72–79 документа CWS/7/29).

3. В 2019 году Международное бюро провело практикум, на котором обсуждались вопросы стандартизации имен. Для обмена опытом и возможными решениями были приглашены докладчики, представлявшие как ведомства, так и предприятия. На этом практикуме Международное бюро отметило, что ВПС имеют множество различных подходов к очистке данных для целей стандартизации имен.

4. Более подробная информация о ходе реализации плана работы Целевой группы после последней сессии КСВ содержится в документе CWS/11/22.

ПРЕДЛАГАЕМЫЕ ОБЩИЕ РЕКОМЕНДАЦИИ

5. В рамках задачи № 55 Целевая группа по стандартизации имен подготовила окончательное предложение касательно набора общих рекомендаций для оказания поддержки ВПС, осуществляющим очистку данных об именах. Данные общие рекомендации приводятся в приложении к настоящему документу.

6. Эти общие рекомендации подготовлены с целью дать общие инструкции высокого уровня. Различия в таких факторах, как требования согласно законодательству, практика работы с данными, цель очистки, предполагаемое использование данных, требования к ресурсам и технические соображения, означают, что не существует единого подхода, который был бы оптимальным для всех ВПС. Настоящие общие рекомендации призваны отразить общую практику, которая может быть применена в любом ВПС для целей очистки данных об именах клиентов, что, в свою очередь, поспособствует совершенствованию процедуры стандартизации имен и методов их сопоставления последующими пользователями.

7. Целевая группа отмечает, что стандарт ВОИС ST.20 содержит рекомендации по подготовке указателей патентной документации, раскрывающих имена заявителей и других клиентов, а также по содействию единообразному представлению имен, помещаемых в этих указателях, равно как и по обеспечению единообразного расположения имен в самом указателе. Поэтому доработанные общие рекомендации будут включать ссылку на этот стандарт и поддерживать его использование.

8. В случае одобрения КСВ на текущей сессии предлагается, чтобы КСВ поручил Секретариату опубликовать эти общие рекомендации в части 3 Справочника ВОИС по адресу: https://www.wipo.int/standards/en/part_03_standards.html.

9. *КСВ предлагается:*

(a) принять к сведению информацию, содержащуюся в настоящем документе и приложении к нему; и

(b) рассмотреть и утвердить общие рекомендации, упомянутые в пунктах 5–7 выше и содержащиеся в приложении к настоящему документу; и

(c) поручить Секретариату опубликовать эти общие рекомендации в части 3 Справочника ВОИС, как сказано в пункте 8 выше.

[Приложение следует]

GUIDELINES FOR THE DATA CLEANING OF NAMES

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

GUIDELINES FOR THE DATA CLEANING OF NAMES

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

INTRODUCTION

1. This set of guidelines covers general considerations for the intake, processing, cleaning, and publication of clean name data. It does not address the many complex issues with 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. These decisions will vary greatly depending on the party applying them, the purpose of transformations, and the quickly evolving nature of matching algorithms.

2. It should be noted that WIPO Standard ST.20 provides 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 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, free from errors and duplication. 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 language(s) that the language 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 many 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 these guidelines.

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 conversion of names, there are three ways referred to in this set of guidelines: 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 set of guidelines.

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 clean 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 set of guidelines:

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

[Annex follows]

ANNEX

DIFFERENT MEANS OF NAME TRANSFORMATION

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

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.

With English as the Lingua Franca of the global(ized) economy, 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
Najīb /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 /ˈædəm/	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 a 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 and of guidelines]

[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