E

WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

CWS/10/17
ORIGINAL: ENGLISH
DATE: SEPTEMBER 7, 2022

Committee on WIPO Standards (CWS)

Tenth Session
Geneva, November 21 to 25, 2022

REPORT BY THE NAME STANDARDIZATION TASK FORCE (TASK NO. 55)

Document prepared by the Name Standardization Task Force Leaders

BACKGROUND

1. At its ninth session in 2021, the Committee on WIPO Standards (CWS) noted the progress made by the Name Standardization Task Force. In particular, the Task Force reported its work gathering information on data cleaning activities in support of name standardization. The Task Force reported plans to present recommendations at the tenth session of the CWS. (See paragraphs 117 to 118 of document CWS/9/25.)

REPORT ON ACTIVITIES

2. The Task Force continued gathering information from Task Force members on their experiences with cleaning data for purposes of name standardization. More detailed questions were asked than previous data collections to elicit additional helpful information for the Task Force. Submissions were provided by six Task Force members in the first quarter of 2022.

3. Using the collected information, the Task Force started work on draft recommendations for best practices. The recommendations cover general considerations for intake, processing, cleanup, and publication of clean name data. They do not address the many complex issues with particular approaches to data cleaning, transliteration, or name standardization, such as choice of algorithms, where and when transformations are applied, frequency, or merging strategies. These types of decisions will vary greatly depending on the party applying them, the purpose of transformations, and the quickly evolving nature of matching algorithms.

4. An initial draft of the recommendations is presented in the Annex to the current document. The draft recommendations are at a very early stage and do not reflect agreement or consensus by the Task Force yet. They are presented to the CWS for information purposes and comments. Final recommendations may change considerably.

5. The Task Force plans to continue work on draft recommendations in 2023 with several rounds of discussion. The Task Force expects to present a final proposal for recommendations to the next session of the CWS.

6. The CWS is invited to:

(a) note the content of this document;

(b) note the progress made on draft recommendations for clean data in support of name standardization, as presented in the Annex to the current document; and

(c) comment on the draft recommendations.

[Annex follows]

RECOMMENDATIONS ON DATA CLEANING FOR NAME NORMALIZATION

Working Draft

Editorial Note:

This working draft is prepared by the Name Standardization Task Force and shared at CWS/10 for information and comments. This draft will be further updated by the Task Force and a final draft submitted for consideration at the next session of the CWS.

SCOPE

1. This Recommendation covers general considerations for intake, processing, cleanup, and publication of clean name data. It does not address the many complex issues with particular approaches to data cleaning, transliteration, or name standardization, such as choice 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.

DEFINITIONS

In the context of this document:

2. Customer data means data on applicants, registrants, holders, owners, legal representatives, or other parties held by an Intellectual Property Office (IPO) in connection with an IP right, application, registration, or other instrument. This Recommendation 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.

3. Clean data means data that is accurate, consistent and reliable, free from errors and duplication. Because 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.

INTAKE

4. IPOs should 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 allow customers to enter and update their name 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 should provide for entry of the customer's name in native characters of the customer's language, in addition to the customer's name in language(s) the IPO works in. For instance, an IPO that works in English could allow separate fields for Applicant Name in English and Original Name in other characters if applicable.

9. IPOs may use identification numbers to identify customers if desired. 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 this Recommendation.

TRANSLITERATION

10. For electronic data exchange including receipt of international applications or registrations, IPOs should send and receive data represented using UTF-8 character encoding.

11. If an IPO transliterates characters from one language (such as Greek) to another (such as English), they should publish their transliteration scheme. The transliterated document should be made available to the customer for review and customers should have a way to submit corrections if the transliteration is flawed.

12. Reverse transliteration should be avoided if possible, preferring 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.

TRANSCRIPTION

Task force to consider recommendations…

TRANSLATION

Task force to consider recommendations…

VALIDATION AND DISAMBIGUATION

13. 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 from the customer if needed, including ways to bypass an automated validation mechanism in case it provides incorrect or incomplete results.

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

15. 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

16. IPOs should develop a strategy to periodically clean data, 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.

17. 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.

18. IPOs should 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

19. IPOs should make available updates to name information that are made after an IP right has been 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.

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

21. 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 the same identification numbers, such as by replacing the sensitive numbers with generated unique numbers for publication.

STATISTICAL PURPOSES

22. For statistical purposes, IPOs may attempt to match customer data with variations in name 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.

[End of Annex and of document]