Annex III

To Request for Proposals N° PTD/10/037

Terms of Reference

for the

NATIONAL COLLECTIONS DATA QUALITY IMPROVEMENT PROJECT

Subject: Provisioning of tools to assess, match, correct and enrich the National Collections bibliographic data.

CONFIDENTIAL NOTICE

This document contains information confidential and proprietary to the World Intellectual Property Organization (WIPO). The information may not be used, disclosed, or reproduced without the prior written authorization of WIPO, and those so authorized may only use this information for the purpose of evaluation consistent with the authorization. Reproduction of any section of this document must include this legend.

Table of Content

WIPO CONFIDENTIAL – RESTRICTED ACCESS

This document contain proprietary information of World Intellectual Property Organization and is not to be used for any purpose other than preparation of a response to this RFP.

1 Introduction

1.1 WIPO

The World Intellectual Property Organization (WIPO) is a United Nations agency specialized in the protection of the intellectual property via cooperation among its member states and other international organizations. Its headquarters are in Geneva, Switzerland.

The Organization counts 184 nations as member states. It administers 24 international treaties dealing with different aspects of intellectual property protection including the Patent Cooperation Treaty (PCT) whose mandate is to publish information concerning the protection of intellectual property. In this context, WIPO publishes every week international applications for patents which are visible and searchable via the PATENTSCOPE® search service (http://www.wipo.int/pctdb/en/).

As part of WIPO's strategic goal IV which is Coordination and Development of Global IP Infrastructure, the PATENTSCOPE® Search Service is being extended to publish not only PCT but also published national patent collections of participating offices.

Today, national patents from 8 countries including Mexico, Vietnam, Cuba, Israel, Korea, Singapore, ARIPO and South Africa are available for search in the new PATENTSCOPE® search service (http://www.wipo.int/patentscope/search/en/search.jsf) and a number of countries have expressed their interest into joining this project some of them being in the last stages of the affiliation process.

1.2 National Collections Data Processing - Current Situation

The National Collections extension of the PATENTSCOPE® Search System is being fed with external data originating from the national offices. WIPO receives this data in several different formats such as XML, SGML, TXT, relational database dumps and others and creates transformation processes that interpret the different data formats and load that data in a relational database.

The primary goals in the initial stages of the project were the accurate interpretation of the data received and its accessibility, leaving the more thorough data quality verification and correction for a later stage.

Today, the data quality assessment is done on a very high level and is limited to the identification of:

- Incorrect office codes
- Invalid dates
- Invalid Numbers
- Invalid Kind Codes
- Duplicate Priorities and
- Duplicate patents,

which are then marked as deficiencies and forwarded to the office of origin for follow-up.

The PCT references (numbers in national patent files referencing a PCT application) are the ones that are the closest to being processed in a more complete way. Today the PCT references are being reformatted to the WIPO format, so that they can be connected to their PCT counterparts and this is visible via the detailed page of the PATENTSCOPE® Search System. Additionally, after formatting the PCT reference in the national data the estimation is being verified by matching the actual PCT applications to the national application.

In the first matching exercise, performed on all the history data from AP, MX, ZA, IL, SG, and VN only 2.5% of the applications matched were sent for verification to one knowledge worker who completed the work in 10 days.

Apart from the above priorities, classifications and legal entities are not being formatted in an international format.

Figure 1 shows the Current Data Management process.

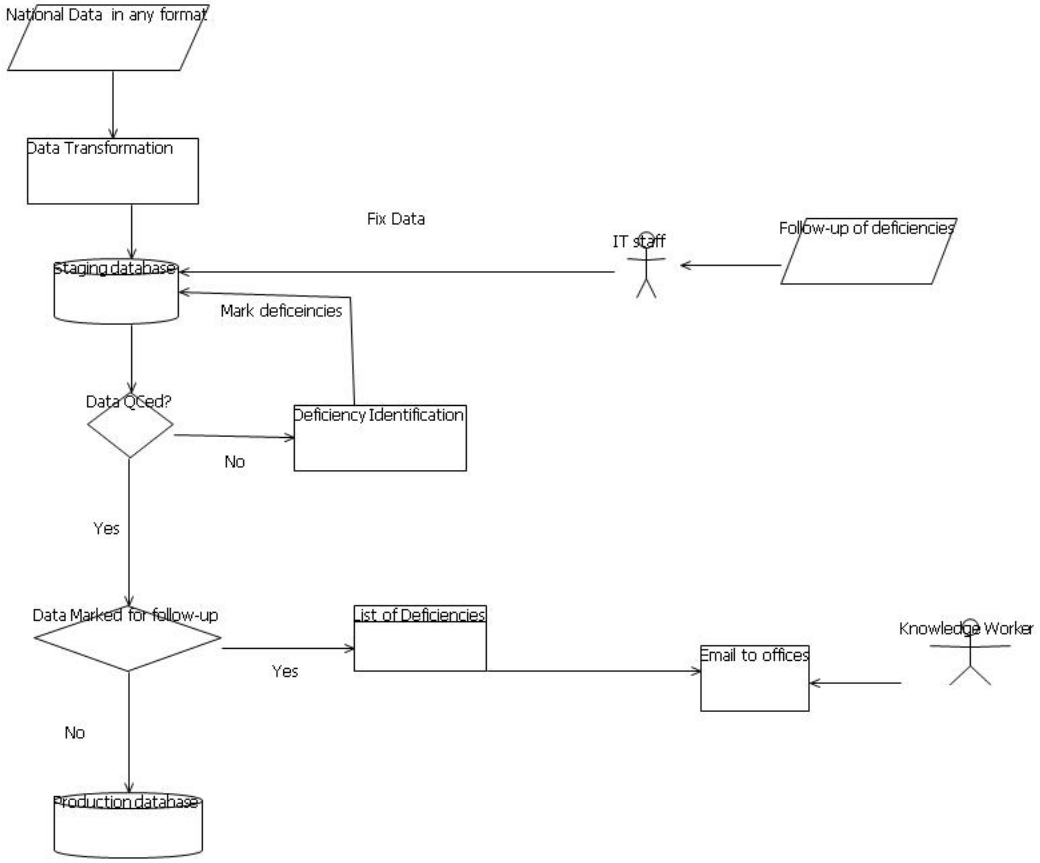

Figure 1: The Current Data Management process

1.3 The Scope of the National Collections Data Quality Improvement Project

This data captured in the different offices has different levels of detail as well as different levels of quality and as such needs to be standardized and eventually enriched in order to be used to its full potential.

There are several areas of the bibliographic data that need to be addressed with the project: Priority data; PCT references; IPC Classifications; Other reference data and optionally Legal Entities.

1.3.1 Priority Data

The priority data is a collection of document references (office, filing number and filing date). All three components of the priority are crucial pieces of information that uniquely identify a priority document, but in most cases at least one component is of poor quality and needs to be standardized, so that it can actually be connected to an existing patent in the same or another collection.

As an illustration, there are two different formats for the US application/priority numbers, yet the priority data in the National Collections originating from the above mentioned 8 offices that references US patents is formatted in 444 different formats, none of them matching the official US format. If used as such to link the priority data to its original document there will be no results.

The immediate benefit of the standardization of the priority numbers would be the possibility to create hyperlinks to either an entry in the national collections or to the office of origin which could make the research more efficient.

Apart from being able to link the patent to its predecessors, the standardization of the priorities is the first and essential step in building data products such as patent families.

1.3.2 PCT References

A number of countries that share their national data with the new PATENTSCOPE® Search System include PCT references (PCT filing number and date) that should if well formatted connect them to their PCT counterpart. Yet, for the majority of patents this is not the case. Until now none for the PCT references in the national data if used as received connects to real PCT numbers. Therefore, this project should also address the matching of the PCT references in the national data to the real PCT application.

1.3.3 IPC Classifications

The classification data also arrives with different level of detail, starting from clear separation of the section, class, subclass, main group and subgroup in some cases, via well structured but incomplete data for others to a very complex data where several different classifications are joined in a single classification text. Currently there are about 500 different patterns of classification text. It would be in the scope of this project to connect the classifications of the national patents to a unique table of commonly understandable classifications that already

exist. In a first instance this would make the search for specific technologies more powerful and exact.

1.3.4 Legal Entities

Today there is no connection made among the different legal entities even though it is obvious that (especially in the cases where the legal entity is a company) one legal entity plays a role in more than one patent. Deduplicating all of the legal entities might be out of the scope of this project, but deduplicating companies might be considered. While identity matching is much more complex, and the existing data for legal entities in most cases is either insufficient or outdated, there is still some potential to set-up rules and procedures for the deduplication of companies.

1.3.5 Other Reference Data

Other parts of the bibliographic data also come with a number of errors that need to be addressed, but they are partly out of the scope of this project. For example, not all countries use ISO codes to identify offices, and in most cases they do not provide reference data so that their data can be mapped to ISO codes, which doesn't leave a lot of space for fixing the issue. In cases like this to fix the data quality issues input on the reference data from the office of origin is needed. Other issues such as the mapping of WO to IB and vice versa can be also addresses via the DQ solution in order to centralize the data quality processing.

1.3.6 Duplicate Patents

In its lifecycle a patent goes trough several phases and as such it can be received at WIPO more than once each time caring slightly different information. Having in mind that there should be no duplication of patents, it is imperative that a duplicate survival technique is put in place to cater for these kinds of situations.

WIPO CONFIDENTIAL – RESTRICTED ACCESS

This document contain proprietary information of World Intellectual Property Organization and is not to be used for any purpose other than preparation of a response to this RFP.

2 Objectives

WIPO has issued this RFP in order to collect offers for Data Quality Tools that would meet the functional and technical requirements as described below. WIPO's aim is to obtain an **off-the-shelf** solution for its Data Quality Improvement project, but bespoke solutions based on open-source components shall also be taken into consideration.

3 Functional and technical requirements of the Data Quality Tool

The Data Quality Tool shall integrate all the inherent components of a Data Quality project such as profiling, standardization and cleansing, matching/relationship identification, and enrichment. Additionally the Data Quality Tool shall provide means for management, monitoring and debugging of the data quality business processes built.

3.1 General Requirements

Imperatively the Data Quality Tool shall:

1) Interact with a wide spectrum of data structure types such as relational databases, XML and other file formats. Optionally the Data Quality Tool shall work with industry-standard message formats such as EDI.

2) Support the Unicode standard and be able to handle multibyte character sets.

3.2 Core Data Quality Processing Requirements

The Data Quality Tool shall:

3) Provide functionality and methods for assessment of the quality of the data. This profiling functionality shall be able to profile the data both internal and external to existing databases. The tool shall provide as a minimum column-based analysis, dependency(relationships) analysis and custom made analysis as well as graphical and textual representation of the profiling results. Optionally, the tool shall provide graphical dashboards.

4) Be able to decompose text data to its integral parts and match the result to corresponding knowledge bases of reference data.

5) Provide functionality for creating user-defined parsing rules.

6) Include built-in functionalities for standardization and cleansing of the data. These functionalities shall include operations such as decomposition and concatenation; data type conversions; look-up and replace operations; as well as free text matching to reference data. The tool shall also seamlessly integrate customized standardization and cleansing operations.

7) Provide matching functionality to identify relationships both between entities in the source data and the source and reference data. The matching functionality should include both deterministic and probabilistic matching with the possibility of tuning the matching rules for optimal results. The tools should also provide customizable duplicate management functionalities.

8) Provide built-in monitoring functionalities that shall ensure that the data meets pre-defined quality benchmarks, alert and report in case of violations.

9) Be able to enrich the national data with external data to improve its completeness and accuracy.

3.3 Operational Requirements

On the operational level the Data Quality Tool shall:

10) Provide an interface and underlying functionality for performing data stewardship functions such as defining custom rules for decomposing or merging records, profiling and monitoring as well as resolving duplication issues and providing incomplete data. Ideally the interface should be web based to allow outsourcing of the data stewardship functions.

11) Be able to capture, reconcile and interoperate metadata relating to the data quality process.

12) Provide a graphical configuration environment for the use of IT and business professionals. This configuration environment shall incorporate a graphical representation of all repository objects, rules and built workflows. Additionally to this development view of the environment the tool shall provide team-based capabilities such as version control as well as role-based security. The tool shall also provide testing, debug and troubleshooting functionalities.

13) Be deployed in WIPO's premises and shall be able to run both in batch and real-time mode. The tool should provide support for Linux and Windows environments and ideally be Java based.

14) Provide runtime error handling, monitoring and statistics.

15) Ideally it shall be a single package integrating all the required functionalities from profiling to enrichment. The data Quality Tool shall also provide interfaces for interoperability with other integration tools.

16) Scale to perform the full data quality process for the history data of big national collections with several millions of records.

17) Provide functionality to visualize TIFF documents.

3.4 Final Process Scenario

The rationale behind the integration of the data quality tools in the data management process would be to maximize the efficiency and correctness and minimize the cost of processing. While the human mind is apt to process complex information which is not always taken into consideration by automatic processes, involving human beings can be very error prone as well. On the other hand automatic processing is by far quicker then processing by human beings, produces uniform results and errors are easily correctable.
Therefore, for optimal results the ideal approach would be to use a hybrid solution combining data quality tools and a limited number of knowledge workers that would concentrate only on the cases where there is a high risk of misinterpretation.

WIPO CONFIDENTIAL – RESTRICTED ACCESS

This document contain proprietary information of World Intellectual Property Organization and is not to be used for any purpose other than preparation of a response to this RFP.

The project is envisioned to be fully implemented in three phases stretching over a period of three years.

1) Phase 1(2010): In phase one there are two tasks that have to be executed at the same time:
    a) gathering of essential reference data and
    b) selection of an appropriate DQ tool

There will be three deliverables at the end of phase 1:

a) Complete reference data available for use for a number of preselected countries;
b) A running DQ solution in WIPO's premises and
c) Trained staff both on the IT and business side

2) Phase 2(2011):

    a) Enrich the current reference data with new collections
    b) Improve the quality on all data in the National Collections and
    c) Seamlessly apply the data quality improvement methodology to the standard data processing

The aim is to achieve a final process as presented in figure1:

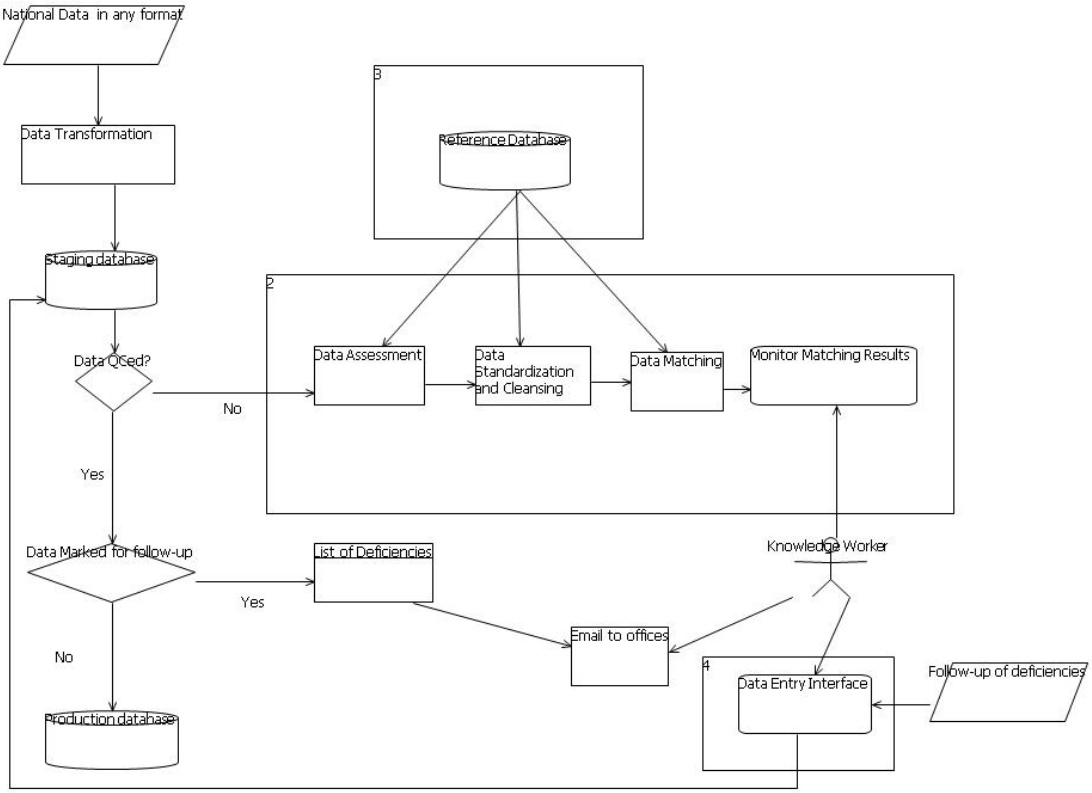

Figure 2 A scenario for the future data management process

3) Phase 3(2012): Design data products such as patent families for use with PATENTSCOPE® or as separate data services.

4 Proof of concept/Proof of value

The proof of concept shall be an integral part of the bidding process and shall be part of the bidder's presentations. WIPO shall provide data and request results for the following tests:

1. Matching of priority numbers to filing numbers.

   Figure 1 and Figure 2 show parts of the bibliographic data of two patents, one filed in Mexico (PA/a/1993/000200) on 15-01-1993 and another one filed in South Africa (1994/00236) on 13-01-1994. The South African patent lists as a priority a Mexican patent filed on 15-01-1993 with a number 93 0200. Looking further at the applicants and the title of the two patents it is becoming obvious that the Mexican document listed under the number 93 0200 in the South African patent is actually the patent in figure 1 and should be reformatted as PA/a/1993/000200 which is the official Mexican format of filing numbers.

   Both applications can be seen via the PATENTSCOPE® Search Service:

   1. http://www.wipo.int/patentscope/search/en/detail.jsf?docId=MXPA/a/1993/000200&recNum=1&docAn=PA/a/1993/000200&queryString=ALLNUM:(PA/a/1993/000200)&maxRec=1

   2. http://www.wipo.int/patentscope/search/en/detail.jsf?docId=ZA1994/00236&recNum=1&docAn=1994/00236&queryString=ALLNUM:(1994/00236)&maxRec=1

   Therefore the first proof of concept exercise should prove that the DQ tool can identify the matches between the filing and priority numbers among the national data that WIPO currently has as well as between WIPO's national collections data and any other publicly available reference data and standardize that data to conform to the official format(s) of the filing numbers of the respective offices.

   WIPO will provide the patent data in XML format, one file per patent document, and including the filing data, priority data, title(s), and applicants.

WIPO CONFIDENTIAL – RESTRICTED ACCESS

This document contain proprietary information of World Intellectual Property Organization and is not to be used for any purpose other than preparation of a response to this RFP.

```
<?xml version="1.0" encoding="UTF-8" ?>
<mx-patent-document file="mx-patent-document.xml" lang="es" office="MX" office-dtd="MX-bibliographic-d
- <bibliographic-data>
  - <application-reference appl-type="">
    - <document-id lang="es">
        <country>MX</country>
        <doc-number>PA/a/1993/000200</doc-number>
        <date>19930115</date>
      </document-id>
    </application-reference>
  - <priority-claims>
    - <priority-claim sequence="1">
        <country>GB</country>
        <doc-number>92 00993.5</doc-number>
        <date>19920117</date>
      </priority-claim>
    - <priority-claim sequence="2">
        <country>GB</country>
        <doc-number>92 24612.3</doc-number>
        <date>19921124</date>
      </priority-claim>
    - <priority-claim sequence="3">
        <country>GB</country>
        <doc-number>9200993.5</doc-number>
        <date>19920117</date>
      </priority-claim>
    - <priority-claim sequence="4">
        <country>GB</country>
        <doc-number>9224612.3</doc-number>
        <date>19921124</date>
      </priority-claim>
    </priority-claims>
  - <parties>
    - <applicants>
      - <applicant app-type="applicant" sequence="1" designation="all">
        - <addressbook lang="es">
            <name>THE MORGAN CRUCIBLE COMPANY, P.L.C.</name>
          - <address>
              <address-1>Morgan HouseMadeira Walk, Windsor, Berkshire</address-1>
              <postcode>SL41E</postcode>
              <country>GB</country>
            </address>
          </addressbook>
        - <nationality>
            <country>GB</country>
          </nationality>
        - <residence>
            <country>GB</country>
          </residence>
        </applicant>
      </applicants>
    </parties>
    <invention-title lang="es">FIBRA INORGANICA VITREA SOLUBLE EN SOLUCION SALINA</invention-title>
    <invention-title lang="en">SALINE SOLUTION-SOLUBLE VITREOUS INORGANIC FIBRE</invention-title>
  </bibliographic-data>
</mx-patent-document>
```

Figure 1:

```
<?xml version="1.0" encoding="UTF-8" ?>
- <bibliographic-data>
  - <application-reference appl-type="national">
    - <document-id lang="English">
        <country>ZA</country>
        <doc-number>1994/00236</doc-number>
        <date>1994/01/13</date>
      </document-id>
    </application-reference>
  - <priority-claims>
    - <priority-claim kind="national" sequence="0">
        <country>GB</country>
        <doc-number>PCT/GB93/00085</doc-number>
        <date>1993/01/15</date>
      - <office-of-filing>
          <country>GB</country>
        </office-of-filing>
      </priority-claim>
    - <priority-claim kind="national" sequence="1">
        <country>GB</country>
        <doc-number>93 14236.2</doc-number>
        <date>1993/07/09</date>
      - <office-of-filing>
          <country>GB</country>
        </office-of-filing>
      </priority-claim>
    - <priority-claim kind="national" sequence="2">
        <country>MX</country>
        <doc-number>93 0200</doc-number>
        <date>1993/01/15</date>
      - <office-of-filing>
          <country>MX</country>
        </office-of-filing>
      </priority-claim>
    </priority-claims>
  - <parties>
    - <applicants>
      - <applicant sequence="1" app-type="applicant" designation="all">
        - <addressbook>
            <orgname>THE MORGAN CRUCIBLE COMPANY PLC</orgname>
          - <address>
              <address-1>Morgan House~Madeira Walk~Windsor~Bershire SL4 1EP</address-1>
              <country>GB</country>
            </address>
          </addressbook>
        - <nationality>
            <country>GB</country>
          </nationality>
        - <residence>
            <country>GB</country>
          </residence>
        </applicant>
      </applicants>
    </parties>
    <invention-title>SALINE SOLUBLE INORGANIC FIBRES</invention-title>
  </bibliographic-data>
```

Figure2:

2. Matching of classification data to the official classification tables.

The various types of the classification data are shown in Figure3 - Figure 5. The goal of the second proof of concept exercise shall be to uniquely identify the existence of the classification data in the table of classifications as shown in Table1.

```
<?xml version="1.0" encoding="utf-8" ?>
- <il-patent-document>
  - <bibliographic-data country="IL" lang="en">
    - <application-reference appl-type="PCT">
      - <document-id>
          <country>"IL"</country>
          <doc-number>146558</doc-number>
          <date>20011119</date>
        </document-id>
      </application-reference>
    - <classifications-ipcr>
      - <classification-ipcr>
        - <ipc-version-indicator>
            <date />
          </ipc-version-indicator>
          <position>0</position>
          <classification-level>A</classification-level>
          <section>C</section>
          <class>12</class>
          <subclass>Q</subclass>
          <main-group>01</main-group>
          <subgroup>/68</subgroup>
        - <generating-office>
            <country>IL</country>
          </generating-office>
        </classification-ipcr>
      - <classification-ipcr>
        - <ipc-version-indicator>
            <date />
          </ipc-version-indicator>
          <position>1</position>
          <classification-level>A</classification-level>
          <section>G</section>
          <class>06</class>
          <subclass>F</subclass>
          <main-group>17</main-group>
          <subgroup>/18</subgroup>
        - <generating-office>
            <country>IL</country>
          </generating-office>
        </classification-ipcr>
      </classifications-ipcr>
    </bibliographic-data>
  </il-patent-document>
```

Figure 3: Classification Data divided into its integral parts

```
<?xml version="1.0" encoding="UTF-8" ?>
- <bibliographic-data country="SG" lang="en">
  - <application-reference appl-type="SG">
    - <document-id>
        <country>SG</country>
        <doc-number>2004040697</doc-number>
        <date>20010126</date>
      </document-id>
    </application-reference>
  - <classification-ipc>
      <edition>N/A</edition>
      <main-classification>C07H 19/067, A61K 31/403, A61K 31/4045, A61K 31/7042,
        A61K 31/7068, A61K 31/7076, A61K 47/48, A61P 1/00, A61P 1/08, A61P 3/00,
        A61P 3/04, A61P 9/00, A61P 9/12, A61P 15/00, A61P 15/08, A61P 25/00, A61P
        25/04, A61P 25/06, A61P 25/18, A61P 25/22, A61P 25/24, A61P 25/28, A61P
        35/00, A61P 3</main-classification>
    </classification-ipc>
  </bibliographic-data>
```

Figure 4: Multiple classifications concatenated in one string

```
<?xml version="1.0" encoding="UTF-8" ?>
- <bibliographic-data country="SG" lang="en">
  - <application-reference appl-type="SG">
    - <document-id>
        <country>SG</country>
        <doc-number>2007019797</doc-number>
        <date>20040427</date>
      </document-id>
    </application-reference>
  - <classification-ipc>
      <edition>N/A</edition>
      <main-classification>G01R 31/319 (2006.01), G01R 31/3193 (2006.01), G01R 19/10
        (2006.01), G01R 19/165 (2006.01)</main-classification>
    </classification-ipc>
  </bibliographic-data>
```

Figure 5: Multiple classifications concatenated in one string and including dates.

Symbol	Section	Class	Sub Class	Main Group	Sub Group
C12Q 1/60	C	12	Q	1	60
C12Q 1/62	C	12	Q	1	62
C12Q 1/64	C	12	Q	1	64
C12Q 1/66	C	12	Q	1	66
C12Q 1/68	C	12	Q	1	68
C12Q 1/70	C	12	Q	1	70
C12Q 3/00	C	12	Q	3	00

Table 1: IPC Codes

The bidder shall present the following results:

1. The bidder shall standardize the priority numbers and match them to their counterparts in the reference data to verify their existence and correctness.
2. The bidder shall parse the classification text to decompose it to its integral parts and match it to an existing classification instance in the reference data.

The data and more details will be provided to the bidder upon request.

WIPO CONFIDENTIAL – RESTRICTED ACCESS

This document contain proprietary information of World Intellectual Property Organization and is not to be used for any purpose other than preparation of a response to this RFP.

4.1 Glossary of terms

WIPO	World Intellectual Property Organization
PCT	Patent Cooperation Treaty
GUI	Graphical User Interface
SGML	Standard Generalized Mark-up Language
XML	Extensible Markup Language
IPC	International Patent Classification
WO	Short code for WIPO
IB	Short code for WIPO as well