WIPO
WORLD
INTELLECTUAL PROPERTY
ORGANIZATION

RFP N° PTD/10/037 – BIDDERS' QUESTIONS AND WIPO'S RESPONSES

1. Is the data quality improvement work expected to be an ongoing activity, or one-time?

   It is going to be an ongoing activity.

2. What is WIPO's approach towards executing the data quality improvement work? There could be the following options, or more:

   (a) WIPO procures the tool/solution via this RFP; actual data quality improvement work done by WIPO IT team. or,

   (b) WIPO procures the tool/solution; identifies a partner responsible for doing the services delivery. or,

   (c) WIPO contracts with a services delivery provider for the data quality improvement work as a turn-key job; tool/s to be selected by the services provider in accordance with WIPO's requirement.

   (d) WIPO procures the tool/solution via this RFP; actual data quality improvement work done mainly by data stewards and/or WIPO IT staff.

   Options (a), (b) and (d) are considered.

3. The RFP's Terms of Reference mentions in point (1.3.2) a poor matching between the PCT references as used by PATENTSCOPE® and as used in the countries national databases. How is it envisioned to address this issue: by federating the countries to use the common standard reference or by providing some kind of automated centralised conversion system?

   PCT references follow a standard which is: PCT/CCYYYY/NNNNNN where CC is the PCT receiving office country/region code, YYYY the filing year, and NNNNNN a unique number per year attributed by the receiving office.

4. Point (1.3.3) expresses the wish to replace the current creative practice in classification patterns by a common reference to a standard table. Does this table already exist or is the building and feeding of the IPC classification part of the project?

   The IPC reference table exists but is regularly changed to follow the revisions of the IPC. See http://www.wipo.int/classifications/ipc/en/support/ for more details

5. In point (1.3.4) the lookup table to de-duplicate legal entity is restricted to companies. Is the reason for this restriction due to volume, or could the proposed solution also include individuals, even if this means to tolerate a certain level of residual duplicates?

   The proposition may include individuals as an option but the short term business need is to de-duplicate companies.

6. The deliverables of Phase I (2010) of the project includes a running DQ solution. Does WIPO need the DQ already loaded with the appropriate set of rules? In other words, will the "training" of the DQ tool be performed by the solution provider or by WIPO team?

   A pre-customization adapted to the problems expressed in the proof of concept is expected to be delivered. However, further customizations/training of the DQ tool should ideally be conducted by WIPO staff.

7. About Type II error (false positives): a priority that would be matched to a wrong predecessor or an ICP classification that would be assigned to the wrong category. What is the error threshold tolerated by WIPO?

   The very least minimum of false positives should be obtained via automatic procedures. Ideally, when there are several possible matches for a correction, the best option should be selected according to the correction rules and the correction occurrence should be marked for further check by a human, possibly with a confidence estimate.

8. How do you currently make the corrections in the following processes: « Data Transformation » and « Deficiency Identification » (Data Qced)?

   We use an open source ETL tool to perform the required data transformations, as well as ad hoc JAVA developments. The "data QCed" part is mostly achieved by checking simple rules like mandatory field, value against a reference table, invalid date field which trigger exceptions in the SQL loading procedures. Corrections are so far done using SQL and ad-hoc programming.

9. In which language is stored the information concerning IPC Classifications?

   The textual description of the classification codes is available in English as well as other languages. See http://www.wipo.int/classifications/ipc/en/support/ for more details

10. Would it be possible to get a representative list of the different databases and file formats? For instance, is the EBCDIC format involved?

    All input formats are converted into UTF-8. Database engines used are ORACLE and MySQL.

11. Could you be more specific about what you mean by 'internal' or 'external' to existing database?

    The profiling tool should be able to work with data stored in databases or in other storage formats (SGML, XML, flat files,…).

12. What is the external data?

    See answer to no. 11 above.

13. Could you be more specific concerning the visualization of TIFF documents?

National patent documents are obtained from offices in two forms: a structured format that contains the captured bibliographic data and an unstructured format consisting of a series of scanned TIFF A4 pages of the patent document. As a result, the information available in the TIFF images can be used to correct typos or uncertainties in the corresponding bibliographic data information.

14. Could you provide us with some examples of data structure types concerning ‘Legacy and non-relational databases’?

XML, SGML, TXT, relational database dumps and others.

15. It is evident that data quality assessment is done periodically to determine invalid and duplicate items; what are the tools used currently for data quality assessment? What is the mode of employing corrections to data?

See answer to no. 8 above.

16. As a part of this project, we envision that a thorough data profiling exercise will be an important step in determining the state of data quality. Has WIPO undergone data profiling on the entire data scope that includes Priority data, PCT references, Legal entities, etc? If yes, please share details.

No, only basic loading checks have been performed so far. The thorough profiling exercise will be performed using the tool/solution to be acquired.

17. We expect that the scope of data standardization and data quality improvement will not be merely restricted to the inbound data collection mechanism; but also to the historical data in the relational database. Please confirm.

This is correct.

18. What is the estimated volumetric for the identified in-scope systems?
- Amount of entities?
- Records (Total & Daily)?
- Size in GB's

Here are some rough estimates:

- 10 to 20 entities (sharing similar types)

- 5 millions records (to grow up to a maximum of 100 millions records in the next 5 years)

- A few dozens of GBs (database)

19. What is the approximate number of XML, SGML, text files or relational dumps received by WIPO daily or monthly?

Today, maximum 10 data transfers by month.

20. Could you please provide approximate existing volumes by entities in question?
- Priority Data
- PCT References
- IPC Classifications
- Legal Entities
- Duplicate Patents

- Priority Data: 15 millions
- PCT References: 5 millions
- IPC Classifications: 40 millions
- Legal Entities: 15 millions
- Duplicate Patents: 1 million

21. One of the general requirements expected from the data quality tool is support for Unicode standard and multibyte character sets. We assume that inbound data comes in multiple language formats. Could you please specify the extent of data coming in different language formats?

It could be any language. Today, we have English, French, Spanish, Portuguese, German, Chinese, Japanese, Russian, Korean, Hebrew and Vietnamese.

22. There is a requirement expected from the data quality tool to be to enrich national data with external data to improve its completeness and accuracy. Could you please specify the kind of external data in question here?

Database or flat files extracts from other patent offices, any kind of reference files.

23. The data quality tool is expected to provide functionality to visualize documents in TIFF format. Please clarify whether the documents in question here are priority documents. Is there any other operation or functionality desired from the DQ tool in terms of working with documents?

Yes, they are the priority documents, as well as the patent applications. The DQ tool shall be able to visualize the pages and browse through them.

24. Please clarify whether WIPO has users/branches in multiple regions/territories/geographies. If yes, please clarify the regions in scope? Is data distributed and maintained in different locations or is all data centralized?

WIPO has some representation offices in a few countries (see http://www.wipo.int/about-wipo/en/what_is_wipo.html for general information about WIPO). The data for this project is currently centralized in WIPO's headquarters in Geneva.

25. Could you please share details on the data governance model and data mgmt and improvement related activities in WIPO?

This RFP is specific to the Patentscope® National Collections project and its requirements may differ from other data mgmt and improvement related activities in WIPO.