**OCP MEL Data Quality Report**

First draft

2018. 08.31.

*Ágnes Czibik, Mihály Fazekas, Nóra Regös*

Government Transparency Institute

**Summary**

In international comparative perspective, Chile has a public procurement regulatory and data system relatively new and less developed in terms of data accessibility and data quality which can be improved with the implementation of OCDS.

National, local data complexity and DIGIWHIST comprehensiveness - During the data collection process we faced many challenges compared to the other countries participated in the Monitoring, Evaluation and Learning project where bulk files were available, unlike in Chile. We had to scrape data from html websites, which procedure took a lot of time. Despite the carefully processed annotation and preparations, several rounds of data validation had to be carried out due to parsing failures. Data were structured into json format and with the help of the DIGIWHIST flatten tool into a single csv file, avoiding the possible mistakes related to merging several data files.

The rate of data completeness - We partially managed to improve the quality of data, still the final database were lacking important variables, such as contract end date, award id, winners, tender and contract status. While missing rates of some other relevant variables, such as tender, buyer and bidder id, award date were quite low, ranged from 0% to 4.3%.

Bid price, that could have been a valid alternative to define winners (i.e. only winners had a valid contract value), had a significantly high missing rate (77%). Final tender price, which represents the total net amount of a tender, had a lower missing rate of 17%, which increased to 23% after removing outliers, values lower than 237 836 CLP and higher than 7 923 065 708 812 CLP.

Most of these data were not filled in the original source, so to increase the level of completeness it would be fundamental to convince publishers about the importance of filling all relevant fields that are key for indicator calculation and analysis.

Other efficient solution would be to regularly monitor simple missing rate indices for each key variable and engage publishers on why missing rates are high or moving in the wrong direction and devise strategies for decreasing missing rates without compromising on truthfulness of entered values.

**A comprehensive data collection exercise**

We carried out a comprehensive search for relevant public procurement and linked datasets centered on national public procurement data as OCDS yet has not been launched in Chile. Below we briefly describe each step.

1. *Identifying all national sources*

The identified sources were official websites of governments and public institutions that are entitled to collect, publish information and follow-up on national public tenders. Tenders published only on local entities’ websites were not collected.

In Chile, the separate publication system for handling public procurement data called [Mercado Publico](https://www.mercadopublico.cl).

1. *Annotation of sources*

Even though we did not aim to fit the Chilean dataset into an OCDS structure, annotation was still inevitable for comparing data points gathered from official local sources, structuring and defining information on contracting in a distinct manner. This process included the preparation of a list of information available in the local sources and the identification of material published there equivalent to DIGIWHIST fields, another project that encourages transparency and accessibility of open data on public procurement. Finding these matches in some cases was not self-explanatory, as there were no clear data description of local sources available. As a consequence, matching and the interpretation of the lists are not easily standardized, opinions on them might differ. In the last step, overlaps between these lists were identified, for finding additional pieces of information to add to DIGIWHIST json from the local datasets (See Annex A1)

1. *Downloading data from the sources*

Data from Chile on public procurement included data mainly from 2018 and some from 2015-2017. Extracting aggregated data in large bulk files was not a possible option, data was scraped from html sites, converted to a single json and then to a csv file.

Our programming team developed a set of codes to scrape data from the local source, transforming them into a new format providing input for subsequent data manipulation tasks. This is a time-consuming process and also prone to mistakes, there were several rounds of data validation and correction.

1. *Converting json outputs into a single flat csv file*

Given that most analysts work with 2-dimensional data files rather than hierarchical data formats such as json, we had to produce the flat representation of the combined json files. We used DIGIWHIST’s flatten tool to produce a set of flat representations of the jsons. This tool has more flexibility on data structure, fields could be fitted easily and it resulted a single csv file, there were no subfiles to merge.

This 2-dimensional table contains some of the necessary information for indicator calculation with contracts in rows, variables in columns.

**Data cleaning**

Before starting to calculate the indicators and to analyse the data, data cleaning is essential to detect, correct and/or remove inaccurate records or fields from the database. Without this step, errors might bias the analysis and so would lead to incorrect or incomplete insights, hence misguided solutions and recommendations.

The cleaning of the Chilean database was less complicated as most of the data points were already in the desired format, there was no need to change variable names.

Some of the variables included information in Spanish, so the file should be opened with the selected encoding type (e.g. LATIN1 for Spanish language). If this option is not available, even though the text will be displayed in a seemingly incorrect format, R will handle them as normal values, no need to replace those characters..

After opening the file, it is handy to check basic information on the database, such as its class (dataframe), dimensions (number of variables and observations and if they are aligned with the original file’s information, making sure we have all the data imported) and structure of the variables. The latter matters for processing commands in R. At this current data file, the types were assigned to variables correctly, only variables showing the tender prices were converted to numeric.

Most of the variables included in the dataset were irrelevant for the current analysis; out of 99 columns 66 were removed. On the other hand, important information, such as variable that defines the winners (as non-awarded tenders were also included), or contract, award, tender dates were completely missing or inconsistent (e.g. bid\_iswinning had only FALSE values). To find an alternative solution, we defined winner tenders as those for which a bid price was assigned. The dataset lacked contract ID as well, it was generated by the combination of tender ID and row order number.

In the database processed currently by the DFID/DIGIWHIST team, most tenders were submitted in 2018, relatively few observations were available from the previous years, so we took a subset of sample by 2018. More data will be available as web scrapers progress backwards in time. However, the process is taking weeks as we want to avoid being shut out from the servers.

Similarly to variable names, tender procedure type was described with the final terms (open, restricted).

**Variables used for indicator calculation**

The table below includes the summary of missing rates of variables we use for calculating indicators defined by OCP and national counterparts.

*Table 1. Summary of variables used for indicator calculations (fields highlighted in red were missing from the datasets)*

|  |  |
| --- | --- |
| **Variables** | **Missing (%)** |
| **tender/id** | **0%** |
| **buyer/id** | **0.3%** |
| **contract/id** | **0%** |
| **contracts/amendment/date** | **NA** |
| **contracts/period/endDate** | **NA** |
| **contracts/status** | **NA** |
| **contracts/awardID (instead: tender/id)** | **0%** |
| **tender\_publications\_firstdcontractawarddate** | **3.4%** |
| **awards/value/amount (instead: tender/final/price)** | **17%** |
| **bidder/id (awards/suppliers/identifier/id)** | **4.3%** |
| **bidder/name (awards/suppliers/name)** | **4.3%** |
| **tender/status** | **NA** |
| **buyer/name** | **0.4%** |
| **contracts/implementation/milestones/dueDate** | **NA** |
| **contracts/implementation/milestones/dateMet** | **NA** |

*Rate of missing values*

The percentage of missings values can determine if a variable can be reliably used for analysis or not, the data available contains enough valid information. Normally we considered variables for further analysis with a maximum of 10-15% missing values.

There were only few relevant variables included in the database, their missing rate was very low, ranged from 0% to 4.3%, except for price variables. On the other hand, as it has been mentioned already, key variables, such as dates of the different procurement process phases, and status were completely missing.

Amount values (contract, tender, award) are typically central to any analysis one does either for calculating indicators directly or for interpreting other indicators (e.g. bidder number for small and large contracts). Bid price, which would be crucial to define winners, had a significantly high missing rate (77%). Final tender price, which represents the total net amount of a tender, had a lower missing rate at 17%, which increased to 23% after removing outliers, values lower than 237 836 CLP and higher than 7 923 065 708 812 CLP.

It would be crucial to improve the filling rates of these fields as they represent fundamental information on public procurements and potential risks of corruption.

**What have we learned, what can be improved?**

Extracting data from html websites is a very time-consuming procedure that requires a carefully completed data validation process to ensure data quality. The accessibility of public procurement data should be improved, providing a possibility to extract aggregated data in bulk files.

We summarize our experience we gained during the data collection work from annotation to indicator calculation.

1. *Fitting local datasets into more standard structure*

It is a challenging task to transform data tailored to a specific country’s public procurement system into a more consumable, standard outcome, ordinary users might not have the expertise and resources for full investigation.

However, this is normal as those particular structures does not intent to reflect all local specificities but to capture general features of public procurement which are understandable in many countries. It was also hard to fit those information to similar layers, such as award and contract periods as most of the variables were related to tender periods.

An additional problem with discrepancies in variable scope is that it can get difficult to match the values of a variable in the local publication to indicators that were defined in advance and that mainly reflects OCDS structure.

RECOMMENDATION:

Regularly check if it still acts as the most comprehensive integrator of national datasets in order to avoid gradual coverage decay (e.g. when national datasets introduce new fields). The framework should be flexible, ensuring the possibility of data extension.

1. *Flatten tool*

The DIGIWHIST flatten tool allows more flexibility with the data structure. To absent fields assigns missing values, keeping the main framework coherent. Furthermore, there were no separate sub files, all information could be processed into a single csv file that was easy to work with and were not prone to mistakes by merging several datasets.

1. *Indicator building: data scope, quality, and documentation*

*3.1 Non-available key data points*

Essential information were not part of the datasets on winners, awards and contracts such as value amount, status, start and end date and ID. We could find alternative solutions only in half of the cases as date and value amount variables were mostly missing completely. Information on amendments and implementation milestones were entirely lacking in the original source.

RECOMMENDATION:

Engage with publishers in order to convince them to publish all essential data points underpinning indicator calculation and analysis and upgrade their data publication system.

*3.2 Missing values*

As we pointed out already above, extremely high rate of missing values were detected for some key variables such as bid price (77%), and others were completely missing, e.g. contract period end date, contract status, tender status, awards id. This can be considered as poor quality of the data in the identified areas and indicates to look for alternative solutions. Those fields with high missing rate were not filled in the original source either.

On the other hand, rich data, without almost any missing values were available on buyers and suppliers, tender award decision date, where missing rates ranged from 0% to 4.3%. Some of those can be valid alternatives to or the above mentioned fields with high missing rates.

Some of the missing rates were exacerbated by our cleaning procedures which might not be perfect. We considered value amounts lower than 237 836 CLP (= 300 €) and above 7 923 065 708 812 CLP (= 10000000000 €) as outliers, hence removed them. After removing such outliers, the missing rate of final tender price increased from 17% to 23%.

RECOMMENDATION:

Regularly monitor simple missing rate indices for each key variable and engage publishers on why missing rates are high or moving in the wrong direction and devise strategies for decreasing missing rates without compromising on truthfulness of entered values.

*3.3 Documentation, annotation*

Detailed description of data extracted from local sources, explaining in depth the various fields would improve the process of annotation. It would probably also make the data transition between multiple systems more efficient and less uncertain. This is valid almost to any field. For example, better understanding of what a certain value amount variable cover exactly: is it net or gross amount?; are they unit/part/total values? Was it part of the main project or an additional cost?, etc.

RECOMMENDATION:

Provide full and up to date documentation on national procurement datasets facilitating comparison to wider data use.

**Annex A1 List of annotation**

|  |  |
| --- | --- |
| **Original name** | **DIGIWHIST equivalent** |
| Número de Adquisición | tender\_buyer\_assigned\_id |
| Nombre de Adquisición | tender title |
| Tipo de Adquisición | tender type |
| Descripción | tender description |
| Tipo de Convocatoria | tender procedure type |
| Moneda | currency |
| Fecha de Publicación | publication date |
| Fecha de Cierre | bid deadline |
| R.U.T. | bidder id |
| Razón Social | bidder name |
| Sucursal | branch |
| Monto | amount |
| R.U.T. Proveedor | supplier id |
| Proveedor | supplier name |
| Total Oferta | bid price |
| Estado | tender/status |
| Número de Licitación | tender/id |
| Organismo Demandante/R.U.T. | buyer/id |
| Clasificación ONU | tender\_cpvs |
| Monto Unitario Oferta | bid\_price |
| Cantidad Adjudicada | number of awarded units |
| Total Neto Adjudicado | lot final price/net amount |
| Estado de la Adjudicación | bid\_iswinning |
| Monto Total Adjudicado | tender final price |
| Criterios de evaluación/Ítem | award/criterion/name |