**OCP MEL Data Quality Report-Mexico**

First draft

2018.10.1.

*Ágnes Czibik, Mihály Fazekas, Nóra Regös*

Government Transparency Institute

**Summary**

In international comparative perspective, Mexico has a highly developed public procurement regulatory and data system which has gone one step further with the implementation of OCDS. Clearly these rich and readily downloadable datasets, national and OCDS compliant, will be immensely useful for local and international actors both within and outside government.

National, local data complexity and OCDS comprehensiveness - When trying to arrive at a comprehensive dataset encompassing both local sources and OCDS data, we came across a range of difficulties which could be addressed by publishers increasing the value of the published data at a relatively little cost. The scope of published variables is different in the different data formats and sources, while some variables and values cannot be fitted easily into OCDS structure, even if there are valuable variables in local sources. Fitting those variables to OCDS terms might be difficult as OCDS-language is relatively easy-to-understand for non-professionals, while official national sources often use specific, legalistic terminology bound to the local context. Merging the two datasets could be an option but it is complex and prone to mistakes compared to using a single OCDS json. The flatten tool split the data assigned to the different layers (contract, awards, tenders) into numerous separate sub files (57 in total) without a global unique variable by which to combine all datasets reliably. Having been encountered all these challenges and because the national, local sources included most of the relevant variables and more observations covering a longer period of time, we decided to use national csv publications as the base of the final data frame with adding information from OCDS where it represented added value. The local csvs were merged manually in R by the unique contract id (R scripts can be found on [Github](https://github.com/govtransparency/digiextractor/blob/master/OCP/Mexico.R)). As a broader lesson, we suggest that a possible solution for mismatches and extending the available datasets with additional insightful information could be if OCDS is regularly checked and updated. This would prevent gradual coverage decay (e.g. when national datasets introduce new fields without adding them to OCDS) and would make OCDS the most comprehensive integrator of national datasets both in terms of variable content and coverage of contracts (i.e. time series).

The rate of data completeness - taking into account only the selected variables used for developing indicators such as contract status, contract end date, contract and tender id, procurement method, data related to suppliers - are high, missing rates are close to zero, and very low for contract value, 5.8% after removing outliers. However, essential information were not part of the national datasets on awards, such as award value (83%), start and end date (24-32%) and award ID. Similarly, important data points were not available for tenders such document type (97%). Information on procurement method, milestones and valid bids were completely lacking in both databases.

It would be crucial to engage with publishers in order to convince them to publish all essential data points underpinning indicator calculation and analysis. Provide full and up to date documentation on national procurement datasets facilitating comparison to OCDS and wider data use could improve the rate of completeness. Moreover explaining in depth the various fields would also contribute to the improvement of the process of annotation, and so probably would make the data transition between multiple systems more efficient and less ambiguous.

**Our approach: a comprehensive data collection exercise**

We carried out a comprehensive search for all relevant public procurement and linked datasets centered on OCDS data but then branching out to other sources to provide an as comprehensive as possible combined dataset. Below we briefly describe each step.

1. *Identifying all national sources, OCDS and other national*

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.

Mexico have separate publication system (*[CompraNet](https://sites.google.com/site/cnetuc/contrataciones))* for handling public procurement data.

Full list of sources identified and used is in Annex A1.

1. *Annotation of sources*

In order to be able to compare and merge databases gathered from various sources, structuring and defining information on contracting in a distinct manner, annotation was the starting point. This process included the preparation of a list of information available in OCDS json, a list of information available in local json or csv data sources. In addition, the identification of material published in local data sources equivalent to OCDS fields, which in some cases was not self-explanatory, as there were no clear data description of local sources. 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, to fit pieces of information of national sources into an OCDS structure. (See Annex A2)

1. *Downloading data from the sources*

Our programming team developed a set of codes to extract data from the above mentioned sources, transforming them into a new format providing input for subsequent data manipulation tasks. All codes can be found on [Github](https://github.com/govtransparency/digiextractor).

Data from Mexico on public procurement procedures of the Administración Pública Federal (APF), about contracts (2010 - 2018), contracting authorities, suppliers registered in CompraNet or in SIEM (Mexican digital information system) were mainly downloaded in local structure, csv format, except the first, which was available in OCDS json and csv formats.

1. *Data conversion and merging*

Due to the different structure of data sources, data conversion and merging had to be carried out carefully. At first, we attempted to convert local sources into OCDS json format to be able to combine those with the original OCDS json files with the help of the flatten tool (if it is of any interest, we are happy to share the combined json files).

This attempt had to be rethought as we could not find the OCDS equivalent for some additional pieces of information of the national sources, thus despite the local csv structures were converted into json files, the tool could not unite the various data sources in a coherent way, as it tried to fill up the information gaps with randomly generated data that led to an output unusable for analysis.

The national publication system covered a longer period of time, included more contract details than the original OCDS and most of the variables we used for developing the indicators. For this reason and in the hope to solve the above mentioned issue, we decided to use the national publication data sets as the core of the final data frame, adding additional necessary data points from the OCDS structure.

Data overlap of sources, meaning the success rate of finding the same contract both in the original OCDS publication and in the local data sources can be considered high: we were able to link 93% of the contracts in the OCDS json to the local source. That is 217 633 contracts from the total of 233 144. In the meantime the local datasets contain 976 253 additional contracts which cannot be found in the OCDS json publication predominantly due to the longer time period covered by local sources than the OCDS.

This also means that most probably data were uploaded to OCDS from the national sources starting in 2017, which did not include information on the different phases as they are distinguished in OCDS (tender period, award period, contract, implementation period), and fields were probably filled in many cases with the same data (contract data).

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

The national local files were in excel format and were converted into csv. We only had to produce flat representation of the OCDS json files to be able to merge them. For that we used OCP’s Flatten Tool. The tool splits the information into many separate sub files (in the case of Mexico, 57 files in total), from which we only used some of the relevant, most recent compiled ones (7 files).

Preparing the dataset for analysis from different csv tables produced by the Flatten Tool and using the local csvs was done using R. This set of data merges produced a 2-dimensional table containing all necessary information for indicator calculation with contracts in rows, variables in columns. Final dataset is available [here](https://drive.google.com/open?id=1zpzIXKgjarvf5mYT8-UiBX2dN7XZHHBT).

**Data cleaning and standardizing formats**

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.

Some problems can already occur while importing data in R (or other software program). Data we used were converted into comma separated files (.csv), which - as its name indicates - separates and splits elements of a database into columns and rows after commas. Contratos, the local files were originally in excel format and saved as a csv file, for that reason while importing them in R, a semicolon (;) needs to be defined as a separator. Once all cleaning steps are completed and the file has been saved as a csv file, the separator can be comma, which is the default in read.csv/read\_csv commands.

If the datasets include information in a language other than English which is the default in R, the file can 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.

Normally, R takes string variables (such as names, addresses, contract titles, etc.) as one piece of information even if they include commas, however it can treat those elements as a distinct variable and place it in a separate column, shifting further other components of the row and might lead to data loss by generating parsing failure due to variable type inconsistency within a column if in the csv it is not put in quote mark. A quick way to check this and that the columns include data points in the expected format is to read the first and last e.g. 100 rows of the database. An alternative, bit more time-consuming option is the use of ‘table’ command that lists all elements of a variable and their occurrences. If this problem occur, source files needed to be extracted again, or if it only affects a limited number of observations, those can be replaced with NA.

After opening the file, it is recommended to check basic information, 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 the structure of the variables. The latter matters for processing commands in R. For example, if R defines contract\_value\_amount as character, we will not be able to use it for calculations, it has to be converted to a numeric variable. Same for date variables, where short date format (%Y-%m-%d) is recommended. In case of the Mexican merged files, R defined dates incorrectly as factors or characters. Factors should be turned into characters first and then characters can be easily modified by the ‘as.Date’ command, giving the desired format.

The program cannot interpret original variable names (e.g. records/0/compiledRelease/contracts/0/items/0/id), thus for code processing and in general, it is advisable to modify variable names in a short, coherent format that is easier to handle and identify (e.g. rec\_com\_ca\_items\_id). This is important, as the final database was created by merging several smaller files into a single dataframe and so coherent data points must have identical names in each sub files, otherwise the program will treat them as distinct variables. Combination of different data frames is only possible by at least one unique ID that each sub file holds (we used *contract id*). After merging the files, checking and removing completely duplicated rows, if there is any, is necessary.

We excluded all fields from the final database that were not important for this particular analysis (e.g. contact points name, email address, etc.). Release files were also not included.

Some string variables might include special characters, typos (e.g. @&%!) that need to be removed to process with string variables more easily and to avoid misleading results. This is important especially when we want to rename, recode those variables, as we did in case of contract status, procurement method and category, changing the Spanish definitions to OCDS terms.

As a general advice, R cannot undo already completed processes, so to avoid that an incorrect command deletes a whole variable, it is better to generate a copy of that variable as a first step, do the operation with that, replace the original one with the correct values and remove the generated copy of the variable.

**Variables used for indicator calculation**

Table 1 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; in some cases NA also indicates the variables that might be available in that source but were not added to the final database)*

|  |  |  |  |
| --- | --- | --- | --- |
| **Variables** | **Missing (%)** | **Missing % Contratos** | **Missing % OCDS** |
| **tender/id** | **0%** | **0%** | **NA** |
| **tender/startDate** | **32,7%** | **31,4%** | **NA** |
| **tender/endDate** | **24,3%** | **24,8%** | **NA** |
| **award/supplier/id** | **71%** | **70%** | **NA** |
| **award/supplier/name** | **0%** | **0%** | **NA** |
| **contract/id** | **1,8%** | **0%** | **NA** |
| **contracts/awardID** | **81%** | **NA** | **25%** |
| **contract/status** | **1,85%** | **0%** | **NA** |
| **contract/endDate** | **1,85%** | **0%** | **NA** |
| **tender/document/type** | **NA** | **NA** | **NA** |
| **contract/document/type** | **97%** | **NA** | **88%** |
| **award/value/amount** | **83.3%** | **NA** | **35%** |
| **contract/value/amount** | **4.6%** | **5.8%** | **NA** |
| **tender/procurement/method** | **1.8%** | **0%** | **NA** |
| **award/status** | **Database only contains awarded tenders** | | |
| **tender/values/amount (instead:contract/amount)** | **4.6%** | **5.8%** | **NA** |
| **planning/budget/amount** | **99%** | **99%** | **NA** |
| **contracts/implementation/transactions/amount** | **99%** | **99%** | **NA** |
| **number/of/tenderers** |  | **NA** | **0%** |

*Rate of missing values*

The percentage of missing values can determine if a variable can be reliably used for analysis or not, the data available contains enough valid information. We replaced nonsensical values that do not belong to the given variable, empty cells, observations with 0 value with ‘NA’.

Normally we considered variables for further analysis with a maximum of 10-15% missing rate. As this quality standard reflects our analytical needs (e.g. data will be used for auditing individual contracts or only for providing broad policy advice over longer periods of time) different cut-points can be set for different goals. As we used distinct sources for the final dataset and the availability of each variable varied in those source files, we provide a brief summary of variable description and missing rates by source (See Table 1).

*Outlier detection - contract values*

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). We considered those values as outliers which were lower than 6700 MXN (= 300 €) or higher than 222 600 000 000 MXN (= 10000000000 €). While these values are imperfect, that is some erroneous values will remain in the dataset while some correct values will be removed, they nevertheless represent our best guess at typical outlier values based on European examples we encountered (i.e. DIGIWHIST EUR outlier cut-points exchanged into MXN). The low cut-point is designed to remove those values which are unit prices rather than global/total prices hence would bias the value distributions. While the high value threshold is designed to remove those values which are magically high, very likely resulting from typos rather than actual money spent.

*National public procurement data - CompraNet*

The publicly available national procurement data in Mexico covers a relatively long period of time: from 2010 until 2018. In general, data completeness - considering only the selected variables we used for developing indicators, such as contract status, contract end date, contract and tender id, procurement method, data related to suppliers - is high, the missing rates are close to zero.

The missing rate for amount values (contract, tender, award) even after removing the outliers (values lower than 6 700 MXN and higher than 222 600 000 000 MXN) remained low: 5.8%.

Variables we defined as tender start and end date had higher missing rates of 31.4% and 24.8%. There were basically no data on estimated contract values or contract values characterising the implementation phase (e.g. final amount spent on a contract upon completion).

Variables kept but not directly used for calculation had various missing rates. 65% of information on contract signature date, buyer identification were not attainable; around one fifth (21%) of data points on buyer name and procurement category were also missing.

The ‘Contratos’ files /means ’contracts’ in English/, mainly contain data on awarded contracts. However, unfortunately information on tender and award status, value amounts, document types, number of tenderers, award date were not included. These variables are essential to compare what rate the financial sources and timing had been exceeded the starting estimations and execution plans. These fields hold valuable information on public procurement efficiency, value for money and potential risks of corruption. Data publishers must be aware of the importance of these variables when filling out data in the publication system.

*OCDS - Mexico*

In OCDS file, three quarter (75%) of data on award id and two thirds (65%) on award value amount were available. Information on contract document type was only rarely available (12%). There was a separate sub file including data on tenderers, where the filling rate was 100%, however it had only observations on few tenders (170528 data points on 45470 unique tenders) compared to the other files (around 1.5 million in total in Contratos, and 3.5 hundred thousands in OCDS). This file was handled separately as it would make the structure of the main database convoluted that only included information on awarded contracts. The average number of tenderers per tender was 4. As it was mentioned above, award and tender data points are completed with contract data. Thus, it is hard to obtain trustful, reliable information on these stages. Data on tender document type and value amount were completely lacking in the sources.

To be able to carry out solid, robust analysis with meaningful insights, the improvement of completeness of dates, value amounts and status of the different contracting stages (tender, award, contract, implementation) with real-time data should be encouraged and emphasized among data publishers by providing clear explanation on the importance of these variables.

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

First of all, it is important to emphasize how useful it is that Mexico publishes rich, detailed, downloadable datasets on public procurement. This goes both for the OCDS json publication and the local csv datasets. In many countries of the world, including Europe, it requires much harder work to access rich data like this.

However, the publication process still can be improved. We summarize our experience we gained during the data collection work from annotation to indicator calculation.

1. *The issue of parallel publications - OCDS and local datasets*

The scope of published contracts differs, however this discrepancy and the precise degree of overlap among different datasets are unclear without detailed investigation. We tested overlap by downloading all data: 233 144 records in OCDS and more than 1,5 Million in local csv publication. Some explanations are clear such as the budgets covered by OCDS (federal agencies’ spending) and national datasets (spending from federal funds including those spent by local agencies); and time periods covered are different too: OCDS (2017-2018) and national data (2010-2018). Still, we don’t know exactly what other reasons might be. At any rate, ordinary users will not have the expertise and resources for full investigation or even to realize there is a discrepancy; so any discrepancy should be clearly and precisely explained

The scope of published variables is different. On the one hand, some information cannot be fitted easily into OCDS structure - this is not necessarily a problem, because OCDS does not have to reflect all local specificities, but to capture general features of public procurement which are understandable in many countries. On the other hand, there are valuable variables in local sources which could have been fitted into OCDS but they are missing from the OCDS json for some reason. Thus, users might feel the urge to use both datasets - OCDS and local publication - but linking datasets is complex and prone to error compared to using only the OCDS json.

Valuable variables missing from OCDS are for example:

* tender/tenderPeriod/startDate - PROC\_F\_PUBLICACION
* tender/tenderPeriod/endDate - FECHA\_APERTURA\_PROPOSICIONES
* tender/mainProcurementCategory - TIPO\_CONTRATACION

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 the OCDS code list without expert knowledge on the local public procurement procedure. OCDS-language is relatively easy-to-understand for non-professionals, while official national sources often use specific, legalistic terminology.

Such problems arose for example when trying to match the values below:

|  |  |
| --- | --- |
| Values of ‘Plantilla\_expediente’ | Values of tender/Procurement/Method |
| Licitación Pública LAASSP/LOPSRM | OPEN |
| Invitación a Cunado Menos Tres Personas LAASSP/LOPSRM | SELECTIVE |
| Adjudicación Directa LAASSP/LOPSRM | DIRECT |
| Reporte de otras contrataciones y contrataciones con crédito externo | LIMITED |
| Licitación Pública Internacional Bajo TLC, para OSD LAASSP/LOPSRM |  |
| Adjudicación Directa Nacional Simplificada LAASSP/LOPSRM |  |
| Proyecto de Convocatoría a la Licitación Pública |  |
| Licitación Pública Internacional Abierto LAASP/LOPSRM |  |
| Licitación Pública Internacional de Adquisiciones (Legislación Estatal) |  |
| Adjudicación Directa Nacional Abierta Art 42. LAASP/LOPSRM |  |
| Licitación Pública Internaciona Abierta, para OSD LAASP/LOPSRM |  |

RECOMMENDATION:

Make sure OCDS serves as an integrator of all relevant national data sources both in terms of variable content and coverage of contracts (e.g. years and budgets). Regularly check if OCDS 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)

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

*2.1 Non-available key data points*

Essential information were not part of the datasets on awards, such as award value, start and end date and award ID. Similarly, important data points were not available for tenders such document type and dates. Information on procurement methods, milestones and valid bids were completely lacking.

RECOMMENDATION:

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

*2.2 Missing values*

As we pointed out already above, extremely high rate of missing values were detected for some key variables such as tender and award values (83%), documentation (97%), periods (24-32%), signature date (57%) that can be considered as poor quality of the data in the identified areas and indicates to look for alternative solutions. Information on implementation and planning were almost completely missing. In most of the cases, supplier and buyer identification were not available (71% and 57% respectively).

One possible explanation can be related to structural issues, that is fitting multiple fields from a national data source into a single OCDS field might be a difficult and error-prone task. In addition, those fields with high missing rate were not filled or not available at all in the original source to start with (this would chyme with many European experiences of DIGIWHIST). On the other hand, rich data, without almost any missing values were available on suppliers’ name, contract values, procurement categories, contract status, contract id and end date. Some of those can be valid alternatives to or the above mentioned fields with high missing rates.

Some of the value missing rates were exacerbated by our cleaning procedures which might not be perfect. We considered value amounts lower than 6700 MXN (= 300 €) and above 222 600 000 000 MXN (= 10000000000 €) as outliers, hence removed them. However, after removing such outliers, contract value only increased from 0% to 4.6%, where 0% percent also included all the 0 values which were removed with the outliers; contract implementation budget were already 99% missing.

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.

*2.3 Documentation, annotation*

Detailed description of data extracted from local sources, explaining in depth the various fields would improve the process of annotation, and so probably would make the data transition between multiple systems more efficient and less uncertain. This is valid almost for 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. Another common problem is the categorization of status. Sometimes it is challenging to classify ten or more different status categories into the four categories of OCDS. All these difficulties might be frustrating for public officials who are in charge to complete the required fields and will decrease their willingness to do so, leaving key parts empty, and losing important information. Inappropriate or lack of such documentation also impedes insightful analyses.

RECOMMENDATION:

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

**Annex A1 List of data sources in Mexico**

|  |  |  |
| --- | --- | --- |
| Description | URL | Format |
| Bulk download of public procurement procedures of the Administración Pública Federal (APF) | <https://datos.gob.mx/busca/dataset/concentrado-de-contrataciones-abiertas-de-la-apf> | OCDS, json and csv |
| Contracts 2010-2018 (Contratos ingresados a CompraNet) - Administrative datasets on the Compranet website | <https://sites.google.com/site/cnetuc/descargas> | Local structure, csv |
| List of contracting authorities ([Unidades Compradoras](https://upcp.funcionpublica.gob.mx/descargas/UC.zip)) - Administrative datasets on the Compranet website | <https://sites.google.com/site/cnetuc/descargas> | Local structure, csv |
| List of suppliers registered in the Compranet system (CompraNet-RUPC) - Administrative datasets on the Compranet website | <https://sites.google.com/site/cnetuc/descargas> | Local structure, csv |
| List of companies in the SIEM (Sistema de Información Empresarial Mexicano,digital business information system) | <https://datos.gob.mx/busca/dataset/sistema-de-informacion-empresarial-mexicano-siem> | Local structure, csv |

Annex A2 List of variable annotations

|  |  |  |
| --- | --- | --- |
| source: https://sites.google.com/site/cnetuc/descargas  Contratos ingresados a CompraNet | |  |
|  |
|  |
| **Original variable name** | **OCDS equivalent** |  |
| Gobierno | none |  |
| Siglas | none |  |
| Dependencia | buyer/name |  |
| Claveuc | buyer/identifier/id |  |
| Nombre\_de\_la\_uc | none |  |
| Responsable | buyer/contactPoint/name |  |
| Codigo\_expediente | tender/id |  |
| Titulo\_expediente | tender/title |  |
| Plantilla\_expediente | tender/procurementMethod |  |
| Numero\_procedimiento | ocid |  |
| Exp\_f\_fallo | none |  |
| Proc\_f\_publicacion | tender/tenderPeriod/startDate |  |
| Fecha\_apertura\_proposiciones | tender/tenderPeriod/endDate |  |
| Caracter | none |  |
| Tipo\_contratacion | tender/mainProcurementCategory |  |
| Tipo\_procedimiento | none |  |
| Forma\_procedimiento | none |  |
| Codigo\_contrato | contracts/id |  |
| Titulo\_contrato | contracts/title |  |
| Fecha\_inicio | contracts/period/startDate |  |
| Fecha\_fin | contracts/period/endDate |  |
| Importe\_contrato | contracts/value/amount |  |
| Moneda | contracts/value/currency |  |
| Estatus\_contrato | contracts/status |  |
| Archivado | none |  |
| Convenio\_modificatorio | none |  |
| Ramo | none |  |
| Clave\_programa | planning/budget/project |  |
| Aportacion\_federal | planning/budget/amount/amount |  |
| Fecha\_celebracion | contracts/dateSigned |  |
| Contrato\_marco | none |  |
| Identificador\_cm | none |  |
| Compra\_consolidada | none |  |
| Plurianual | none |  |
| Clave\_cartera\_shcp | none |  |
| Estratificacion\_muc | none |  |
| Folio\_rupc | awards/suppliers/identifier/id |  |
| Proveedor\_contratista | awards/suppliers/name |  |
| Estratificacion\_mpc | none |  |
| Siglas\_pais | awards/suppliers/address/countryName |  |
| Estatus\_empresa | none |  |
| Cuenta\_administrada\_por | none |  |
| C\_externo | none |  |
| Organismo | planning/budget/description |  |
| Anuncio | tender/documents/datePublished |  |