This project has received funding from the European Union's Horizon 2020 research and innovation Program under grant agreement No 645852

EU Grant Agreement number: 645852

Project acronym: DIGIWHIST

Project title: The Digital Whistleblower: Fiscal Transparency, Risk Assessment and the Impact of Good Governance Policies Assessed

Work Package: 2 - Data Collection and Cleaning

Title of deliverable: D2.7 Data validation results

Due date of deliverable:

Date in amended Grant Agreement: 30/09/2017

Submission date of revised version: 26/02/2018

Author(s): Jiří Skuhrovec, Michal Říha, Miroslav Palanský (Datlab)

Organization name of lead beneficiary for this deliverable:
Datlab

Dissemination Level		
P	Public	x
P	Restricted to other programme participants (including the Commission Services)	
R	Restricted to a group specified by the consortium (including the Commission Services)	
C	Confidential, only for members of the consortium (including the Commission Services)	

The information and views set out in this publication are those of the author(s) only and do not reflect any collective opinion of the DIGIWHIST consortium, nor do they reflect the official opinion of the European Commission. Neither the European Commission nor any person acting on behalf of the European Commission is responsible for the use which might be made of the following information.

1. Introduction

This paper describes the methodology and current output of the data validation process led by Datlab. The purpose of the process has been to check various properties of DIGIWHIST data on public procurement and provide continuous feedback to project partners. Thus the paper will:

1. Describe the validation methodology
2. Provide an overview of the work done
3. Summarize the current status of data quality

Datlab has used its extensive know-how in software development as well as public procurement data analytics to provide timely feedback to project partners. That led not only to a shift in data extraction strategy during the project, but more importantly towards transferring a large portion of procurement expertise towards technical staff. These steps have contributed largely to the good procurement data quality, which is paramount to the project.

However, judging from the experience with Czech and Slovak data, it takes years of work to actually fine-tune data extraction even from a single procurement source. In fact, such a job is never finished because of amendments to the legal framework and consequent changes to the data structure and terminology. Based on this experience, DIGIWHIST's goals are ambitious, even if we only aim to achieve moderate quality data within the project.

Note that since the DIGIWHIST team aims to further improve the quality of the data during the sustainability period, the data quality results are to be taken rather as state-of-the-art. In fact, the methodology has been designed in such a way, that validation is possible on an ongoing basis. That implies, that updated versions of source-specific validation reports can be published upon major data releases.

2. Validation methodology

During the project, the approaches and tools used for data validation have evolved quite rapidly. This has resulted not only from the changing quality of the data (from initial issues with elementary data integrity to the fine tuning of individual extraction of variables), and the change of technology used (database architecture), but mostly from the experience gained during the project. This chapter focuses on describing the final picture of optimal data validation processes, as we see them at present.

The purpose of the validation process is to ensure that procurement data collection provides DIGIWHIST with the correct data. The process has been carried out separately for each data source (after UCAM made the data available in the database), then its individual parts were repeated as needed - reflecting changes announced in the data. The final round of validation has been run prior to the production of this report.

Essentially, the validation process checks for four characteristics:

1. **integrity** = check the data integrity, consistency of its structure and content with the object model. The purpose is to ensure the basic usability of the data for use with further tools such as the opentender.eu portal, and also for further validation checks. In practice this means
    a. Checking that numbers are numbers, dates are dates and enumeration fields contain valid values etc.
    b. Checking the complex data structures to ensure for example we have recorded source publication, bids are nested under tender lots etc.

    Whereas in theory most of these properties should be enforced by the database and software architecture, such approaches fail once NoSQL databases are used or the architecture is adapted (both are real cases).

2. **completeness** = ensure that we have all the announcements (tenders) present in the source system. To do that we use two methods:
    a. Sample of publications is gathered independently of the DIGIWHIST infrastructure, then cross-checked with its data. For sources with previously collected data (CZ,SK,HU) this led to using their samples. For other sources, the manual collection of random samples was found to be the most efficient method.
    b. Random sample data from previous releases is stored, then again used in future releases for cross-checking.

    Cases of missing publications are manually reviewed, either to adjust the detection methods themselves or to report an issue to the programmer team.
    The original idea was also to check the tender sums against national accounts, to cross-check the extent of government procurement spending. This exercise however turned out to provide more of an insight into the differences in legislative framework rather than data incompleteness (for details see[1]), and was thus dropped from this report.

3. **data availability and correctness** = The availability of individual tender attributes is monitored, on a semi-automatic basis.
    For basic form types - CONTRACT AWARD and CONTRACT NOTICE, the variables that should be available are defined and their real availability in the source is checked. Both cases of present and missing variables are then manually verified against source data by Datlab and partner experts on individual sources, to confirm either an error in a source or in the DIGIWHIST database.

4. **data matching quality** = Ensure that individual publication of the same tenders are correctly matched together as well as the same bodies (suppliers, buyers). That includes detection of false positives (overmatching) as well as false negatives (undermatching). To do that, various practical heuristics with consequent manual inspection were used:
    Tender matching quality:
    a. Checking of unmatched awards and notices (singletons)

[1] https://medium.com/datlab/the-elephant-in-the-room-acb12a0908da

b. Checking of large groups of matched awards and notices
c. Checking for inconsistent dates (award < notice)

Body matching quality:

d. Checking of conflicting buyers in the same tender
e. Checking of bodies not matched with the same name

Additionally, in the Czech and Slovak Republics and Hungary, the imported data from previous projects that were manually cleaned were used as an objective benchmark of body matching algorithms. This provided timely feedback for the fine-tuning of similarity functions and setting appropriate thresholds.

Whereas the **core output of the validation process was the internal feedback leading to improvements** in data, we also produced a single report for each country that will provide users of the data with basic information on its completeness and reliability. Thus reports using this methodology for each processed source are provided in the last chapter.

3. Validation prioritization

By far the largest task in the validation process was goal 3 - checking data availability and correctness. Let us illustrate the requirements for a thorough inspection of the data produced: so far, the DIGIWHIST team has designed software to extract data from 22 sources structured in 642 different templates (most of which required a tailored approach). From each of those templates up to 250 variables present in the DIGIWHIST data template might have been theoretically extracted. In our practical experience the amount of data in a source is only 40-80 variables per tender. This implies, that only checking extraction of each variable once per each template would mean checking 60 * 642 = 38 520 values. Even that would however not suffice, as in order to eliminate the most serious mistakes, a manual check of 10+ random publications per template is needed. **In order to ensure reasonable quality of all templates, conducting about 385 200 manual data checks would be needed**. Neither this number of manual data checks nor corrections of any identified errors would be feasible within the project scope.

In order to maximize impact on data quality, a more rational approach was chosen. Since the majority of the most relevant data is contained within a smaller portion of templates, the validation focused on the most frequent ones at each source. To take such an approach further, DIGIWHIST partners agreed on a priority list of variables[2], ensuring that issues of prices, dates and the identification of buyers and bidders will be addressed more thoroughly than other variables. Finally automated checks were employed to pre-select tenders with potentially missing variables, which were also manually checked.

[2] The priority variables were identified through mapping of requirements for intended risk indicators as well as other research goals.

4. Continuous error reporting and peer review

Datlab organized several rounds of data review by consortium partners (UCAM, AKKI, HSOG & OKFDE) as well as external experts (Open Contracting Partnership, TI Slovakia, Oživení Czech Republic, opentender.eu beta users). Together with internal validation efforts coordinated closely with the UCAM team this resulted into **1175 documented error reports and tasks, out of which 903 have been already resolved to date**[3]. The Redmine issue tracking system maintained by Datlab has been used and tuned for purposes of the project.

All the error reports were checked and expanded by the Datlab team to provide clear explanations of what the programmers should fix. After an error has been reported as fixed, Datlab team members again re-checked whether this was correct or whether the task should be re-assigned to the programmers' team.

As a part of the process, OKFDE integrated the data into the national portals. As such this worked as a replication exercise which helped check the integrity and quality of the data and provided further feedback for the validation and data quality improvement.

The internal validation process involved repetitive checking of random samples of tenders across various types of input forms and historical periods. Also, since the data processing involves several stages (see DIGIWHIST deliverable D2.8 for details), the validation needed to look into results of those stages to pinpoint where potential errors could be created. The validation strategy thus evolved jointly with the software architecture of the data itself, and ended up examining various properties of the data at appropriate stages. Thus the checks described in the previous chapter needed to be conducted at different points in time (with main consistency checking at the early stages of source cleaning, and matching checks at the end of the matching phase etc.). This resulted in about 100 back and forth full data processing cycles coordinated with programmers.

5. Current data quality

In this chapter we use the processes described in the methodology chapter to provide a snapshot of the quality of procurement data extracted by DIGIWHIST to date. That means we use methods that were previously used to identify potential extraction software flaws to report on the current data quality. It is vital to stress, that such **results reflect both the quality of the published source data in the national portals and software processing by DIGIWHIST**. If the project results were perfect, then 100% of the identified flaws could be attributed to the source data quality. In practice, we are close to such a state for priority variables in countries with well-structured data and where few data templates are used, for reasons discussed in the Validation prioritization chapter.

In order to provide detailed information to the users of the data (which we consider a main priority of this deliverable), we produced separate reports per source processed. The

[3] We report only the reporting done with current Cambridge team, error reports in 2015 were stored in different system with limited track record)

methods for obtaining such results are summarized in Annex 1, the main results of the reports are summarized in the following table (figures as of 15.9.2017):

Source (country)	Tenders	Data templates [4]	Integrity	Complete-ness	Priority variables quality	Body matching score	Tender matching score
TED	2.68 M	10	100%	100%	65.28 %	42.22 %	78.51 %
Czech Republic	329 K	43	100 %	100 %	82.68 %	92.24 %	84.36 %
Slovakia	132 K	119	100 %	100 %	68.05 %	98.75 %	91.70 %
Latvia	149 K	29	100 %	100 %	66.61 %	97.42 %	95.41 %
Poland	2.88 M	9	100 %	100 %	47.39 %	46.69 %	76.18 %
Hungary	99 K	68	100 %	57.69 %	15.5 %	100 %	96.45 %
Lithuania	261 K	25	100 %	100 %	33.8 %	46.92 %	55.26 %
Croatia	235 K	32	100%	100%	40.75%	95.94%	100 %
Ireland	84 K	44	100%	100%	23.81%	26.43%	51.35 %
Estonia	193 K	17	100%	100%	76.52%	99.37%	93.30 %
Georgia	211 K	10	100%	100%	68.09%	98.91%	100 %
Norway	195 K	61	100%	100%	16.55%	50.24%	50.00 %
Switzerland	112 K	19	100%	100%	31.25%	65.41%	86.45%
Netherlands	56 K	28	100%	100%	49.17%	89.84%	93.41%
Bulgaria	241 K	26	100%	100%	11.09%	10.86%	97.94%
Portugal	796 K	5	100%	100%	33.34%	99.43%	91.08%
France	2.60 M	41	100%	100%	50.34%	84.07%	74.82%
Slovenia	174 K	36	100%	100%	22.34%	69.99%	99.98%
United Kingdom	131 K	6	100%	100%	33.99%	87.88 %	77.34 %
Romania	208 K	1	100%	100%	-[5]	-	-
Spain	355 K	7	100%	100%	65.21%	72.78 %	98.37 %

[4] This number roughly corresponds to difficulties in processing the source. In some sources (like Norway) where the data structure is fairly unified, it may however be overestimating the difficulty.

[5] There was an issue with Romanian data that did not allow stats to be reliably calculated at the time of the report. This should be fixed in near future and updated in the country validation report.

Detailed scores (together with further potential methodology revisions) can be found in the country reports stored at:

https://github.com/digiwhist/wp2_documents/blob/master/validation_reports/

Since the DIGIWHIST team plans further works on improving the data extraction, we plan to update the results jointly with releases introducing major changes. This way we will provide additional feedback to users on data about their quality.

Annex 1 – methodology for country validation reports

The methodology for the calculation of reported indicators was developed to capture principles outlined in chapter 2. Essentially we proposed various statistics to check for various dimensions of data quality.

Checks were calculated over the full dataset (unless stated otherwise) as available to date. They are designed to provide value between 0 % (minimum quality) and 100 % (maximum quality). Still it needs to be noted that they cannot in principle cover all the possible issues with data, they however proved sufficient to monitor most of it.

These indicators are defined as follows:

1. **Integrity** – the score checks the consistency of the gathered clean data with the internal data model. Since the data are not stored in the form of relational tables but json structures, both structural rules and variable types are not strictly enforced by default and need to be checked. The overall figure of the rating denotes the share of individual variables in all tenders, which are in line with the data model in terms of their content and structural placement.

$$I = \frac{\#\, Correct\ variables}{\#\, All\ variables}$$

2. **Completeness** – we use two methods to check if we are not missing any tenders in the clean data (because of crawler, parser errors, but also the possible disappearance of tenders in the source). The result is a linear combination of both sub-results.
    a. Historical sample – checks for the presence of tenders historically obtained by crawlers from the same source (by internal persistent id)
    b. Independent sample – checks for the presence of sample of tenders which were obtained manually from the source (by url)

$$I = \frac{1}{2}\left(\frac{\#\, Historical\ data\ \cap Current\ data}{\#\, Historical\ data} + \frac{\#\, Independent\ sample\ \cap Current\ data}{\#\, Independent\ sample\ data}\right)$$

3. **Priority variables quality** – we check for availability of variables defined as priority variables by the DIGIWHIST consortium. These include:
    a. **Tender level:** Title, cpvs, procedureType, isFrameworkAgreement, publicationDate, selectionMethod, bidDeadline, awardCriteria, estimated price, final price, awardDecisionDate, bidsCount (last two are variables checked only for contract awards)
    b. **Buyer, Bidder level**: name, id, city, street, country, rawAddress

    The indicator is calculated in the master data, that is after the merger of notices, awards etc – so that the variable needs to be present in at least one of these to be deemed available. The benchmark number of possible variables is based on available publications, for awarded contracts final prices are expected, for contract notice bid deadlines etc. This ensures that availability is not underestimated due to missing publications (this is checked by a completeness check). Note that content of variables is also not checked, as that is already ensured by consistency checks.

$$I = \frac{\#\, Avaliable\ priority\ variables}{\#\, Possible\ priority\ variables}$$

4. **Body matching score** – we use two statistics to check for extensive numbers of false positives and false negatives in the results of body matching.
   a. **Buyer conflicts** – measures the share of notice-award pairs, where a different buyer has been assigned by body matching (which is apparently incorrect).
   b. **Name conflicts** – measures the share of bodies, which have not been matched together but have exactly the same name

$$I = \frac{1}{2}\left(\frac{\#\, notice\_award\quad pairs\ without\ buyerconflict}{\#\, notice\_award\ pairs} + \frac{\#\, body\ groups\ with\ name\ conflict}{\#body\ groups}\right)$$

5. **Tender matching score** – we use four different statistics to capture various aspects, flagging both possible overmatching and undermatching of different publications (clean tenders) together, claiming these represent the same tender.
   a. **Inconsistent dates** – measures % share of notice-award pairs, where the award date is preceding the notice.
   b. **Orphan notices** – measures the % share of contract notices with neither award nor cancellation matched.
   c. **Orphan open awards** – measures the % share of awards in an open procedure, which have not been matched to a notice.
   d. **10+ matched publications** – measures the % share of tenders, where more than 10 publications have been matched together (although such cases can happen in reality, their occurrence should be very low)

$$I = \frac{1}{4}(Inconsistent\ dates + Orphan\ notices + Orphan\ awards + 10^{+}publications)$$