*This project has received funding from the European Union's Horizon 2020research 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: D2.8 Methods Paper

Due date of deliverable:

*Date in amendment to GA: 30/09/2017*

Actual submission date – revised version: 26/02/2018

Author(s): Jan Hrubý, Tomáš Pošepný, Jakub Krafka, Bence Toth (UCAM), and Jiří Skuhrovec (Datlab)

Organization name of lead beneficiary for this deliverable:
University of Cambridge (UCAM)

| 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) | |

# Executive Summary

This document is a final methodological paper of WP2 of the DIGIWHIST project. It describes how the final database (DB henceforth) was developed starting with a high level description of each public procurement source that was processed, continuing with a description of the processes that led to the development of a structured database, followed by the processes involved in linking related data and creating a final database based on the linked data. The last chapter contains the description of performance indicators (transparency, corruption risks and administrative capacity) and the conversion of the DIGIWHIST data template to the Open Contracting Data Standard (OCDS henceforth).

This methodology report describes the following steps in data processing:

- **Data download** - collection of HTML, XML, CSV and other content from government sources
- **Structuring data** - conversion of each publication from its original format to a uniform structured data template
- **Formatting data** - conversion of structured text to standard data types (numbers, dates, enumeration values) including cleaning nonsensical values or ballast information
- **Linking related information** - grouping information which describes one real world tender together
- **Data merging** - putting information from all linked data records together to create one final image of a public tender covering its whole tendering cycle

Within DIGIWHIST, 25 public procurement data sources were processed covering all 34 jurisdictions listed in the Grant Agreement. This total number consists of:

- TED
- 21 national web portals or open data sources
- Archives for UK, CZ
- Project partner's DB of older Hungarian tenders

The table below shows the number of processed publications for each jurisdiction.

| Country | Source | Processed publications | Indicators per tender |
|---|---|---|---|
| **Poland** | TED + national portal | 3253616 | 2.31 |
| **France** | TED + national portal | 3171123 | 2.09 |
| **Portugal** | TED + national portal | 1161396 | 1.93 |
| **Spain** | TED + national portal | 494320 | 1.36 |
| **Czech Republic** | TED + national portal | 437672 | 2.34 |
| **Germany** | TED | 351991 | 2.05 |
| **Romania** | TED + national portal | 322410 | 2.14 |
| **Bulgaria** | TED + national portal | 299574 | 1.40 |

| | | | |
|---|---|---|---|
| **Lithuania** | TED + national portal | 298708 | 1.93 |
| **United Kingdom** | TED + national portal | 281442 | 2.68 |
| **Croatia** | TED + national portal | 252577 | 1.63 |
| **Hungary** | TED + national portal | 248098 | 1.48 |
| **Norway** | TED + national portal | 243435 | 1.74 |
| **Georgia** | TED + national portal | 211863 | 1.67 |
| **Estonia** | TED + national portal | 212225 | 2.09 |
| **Slovenia** | TED + national portal | 203840 | 1.76 |
| **Slovakia** | TED + national portal | 174958 | 1.76 |
| **Latvia** | TED + national portal | 170840 | 1.99 |
| **Switzerland** | TED + national portal | 145808 | 1.36 |
| **Netherlands** | TED + national portal | 120218 | 1.92 |
| **Italy** | TED | 123434 | 1.82 |
| **Ireland** | TED + national portal | 100465 | 1.65 |
| **Sweden** | TED | 81976 | 1.51 |
| **Belgium** | TED | 75264 | 1.98 |
| **Finland** | TED | 43914 | 1.63 |
| **Austria** | TED | 39722 | 1.90 |
| **Denmark** | TED | 38883 | 1.66 |
| **Greece** | TED | 34838 | 1.91 |
| **Luxembourg** | TED | 8136 | 1.77 |
| **Cyprus** | TED | 5914 | 2.55 |
| **Malta** | TED | 4567 | 1.52 |
| **Iceland** | TED | 1332 | 1.31 |
| **Serbia** | TED | 283 | 1.29 |
| **Armenia** | TED | 13 | 1.80 |

A set of indicators was developed to describe the behaviour of contracting authorities within a market. Each indicator is calculated at the level of individual tenders while they are also aggregatable at organisational and sectoral levels. The result can be:

- **1** - given behaviour was confirmed (e.g. a tender is of high corruption risk)
- **0** - given behaviour was disproved (e.g. there is no detectable corruption risk)
- **N/A** - there is not enough information to confirm or disprove given behaviour

| Area | Number of indicators developed | Total number of positively (1) evaluated behaviors |
|---|---|---|
| Corruption risk | 8 | 4036557 |
| Administrative capacity | 6 | 1403466 |
| Transparency | 2 | 529407 |

The final data set is a unique database describing public procurements and buyers' practices across the whole of Europe. It contains detailed information (up to 250 variables) covering the whole life cycle for both above- and below-threshold tenders. It combines all available publications related to a given tender, links them to company and budget databases using complex algorithms and applies various business rules to create one single representative for each tender and subject.
In order to further fine-tune the database (following user feedback from stakeholder workshops) and implement a number of key extensions beyond the Grant Agreement, further data releases are planned throughout autumn 2017 up until February 2018.

**An updated version of this document can be found online at**
https://github.com/digiwhist/wp2_documents/blob/master/d2_8.pdf

# Glossary

### Tender

An object providing information about the whole process of awarding a public contract. At different stages it can contain a different amount of information. For example it can comprise only data from one publication (e.g. Call For Tender form) in the early stages of data processing and compiled information from all publications describing the same contracting process at a master stage.

### Lot

Part of a tender that can be awarded separately.

### Body

General term for e of a public body or a legal entity. This term encompasses more specific terms like buyer, bidder, specifications provider, tender administrator etc.

### Buyer

A contracting authority that uses public procurement to find a provider of a service or supplier of goods.

### Bidder

Potential provider of a service or supplier of goods that participates in a public procurement with the goal of winning a tender.

### Parsing

Process of transformation of raw data (TXT, HTML, XML, CSV) into text data in a DIGIWHIST data standard

### Cleaning

Process of conversion of text data in a DIGIWHIST data standard into typed data (numbers, dates etc.) in a DIGIWHIST data standard

### Matching

Grouping of records describing the same real world entity

### Mastering

A process of a creating one final record from all matched records. It merges data from all records through the application of sophisticated rules with the goal of selecting the best value for each variable in a tender object.

# Tendering data processing overview

This chapter describes from a high level perspective how the final DIGIWHIST database has been created, starting from locating the data in a source and ending with a detailed tender description capturing the whole lifecycle of a particular public procurement, including a set of performance indicators.

## Tender publication cycle

The whole process of creating a final DIGIWHIST database started by devoting a significant amount of time to mapping all possible sources of public procurement data across Europe and selecting a group of sources that is can be processed within a given time scope and covers the biggest set of public procurements, including as many countries as possible and as many historical values as possible. This effort led to processing 25 data sources including web portals, FTP servers, JSON or CSV data dumps. This chapter describes how each publication is treated in the DIGIWHIST data processing system and how it contributes to a final output.

### Data downloading - raw data

Each publication containing information about public procurement starts its life by being published in the source. This means a new web page has been created, an XML file was uploaded to a FTP server, a JSON record appeared in an API or a CSV file is linked from a source. Once the DIGIWHIST data processing system detects such an event it triggers a chain of procedures that leads to the incorporation of the data included in this publication into a final database.

To be able to detect such an event a set of so called source crawlers were implemented. Each crawler draws upon the combined knowledge of DIGIWHIST public procurement experts and developers, gained through a detailed inspection of the source. The knowledge basically consists of two key parts:

- Understanding what data are published in the source and which publications are important to achieve DIGIWHIST's goals
- Understanding the technological aspects of how each source works to be able to: detect, as mentioned above, the publication of new tendering information; download such publications; and store them as so called raw data

The chapter entitled **Sources** describes what was found out about a source, how the source works from a high level perspective (the technological details are documented in the published source codes), what publications are processed and other interesting know-how.

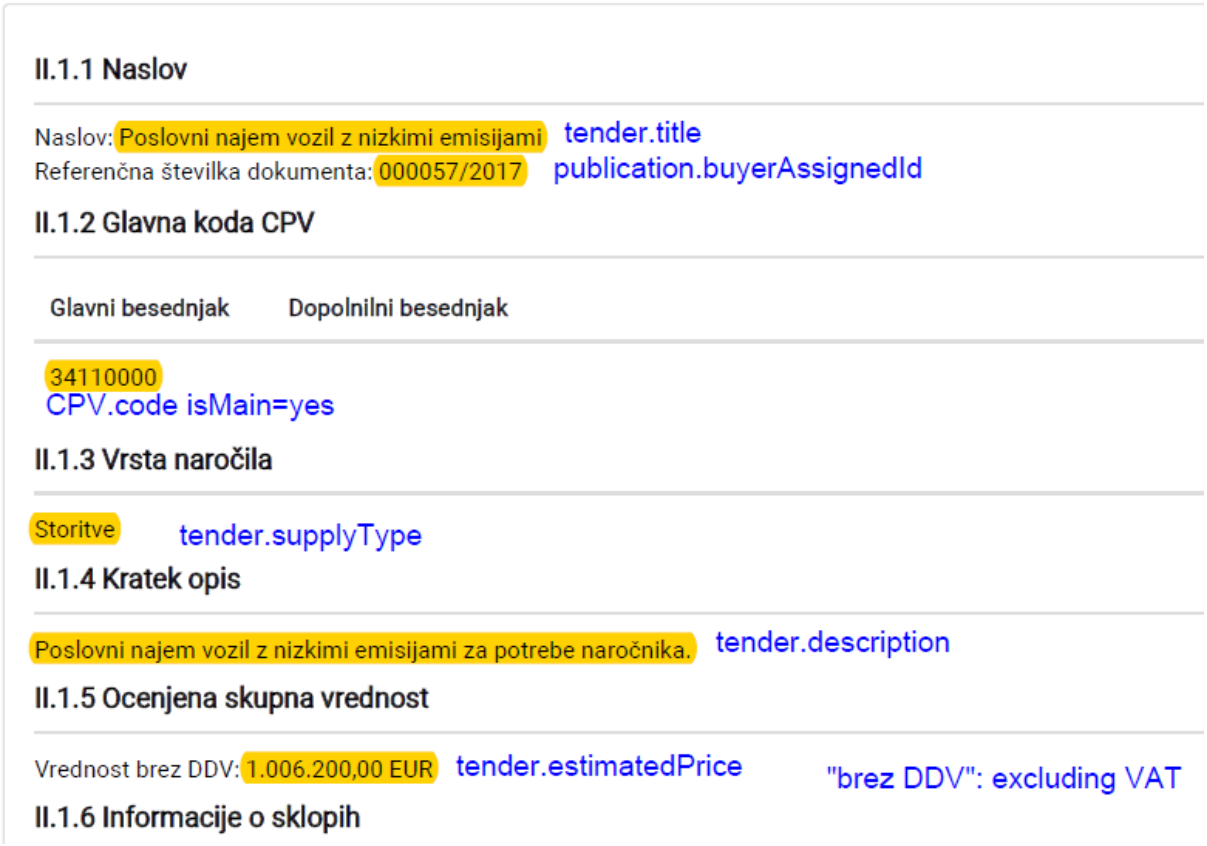
#### Fact to remember

A "raw document" is a structured or unstructured publication including one (title, name, price, ...) or more pieces of information about **one** public procurement.

### Structuring - parsed data

After a raw publication is stored in a database, its lifecycle continues and data included in the publication are extracted and stored as text values in a DIGIWHIST data template. This

procedure requires input from a public procurement expert in the form of a template annotation that defines how to convert a raw publication into DIGIWHIST data structures. There are several methods described in D2.5 about how to communicate expert knowledge to a developer. The developer then implements a program (the so-called data parser) that extracts data from a raw document and stores it as a structured (parsed) document. All values are stored as text values. The first image below shows a visual annotation of an HTML page and the second image shows a structured parsed document created using such an annotation basis.





```
{
  "title":"Poslovni najem vozil z nizkimi emisijami",
  "cpvs":[
    {
      "code":"34110000"
      "isMain":"true"
    }
  ],
  "supplyType":"Storitve"
  "description":"Poslovni najem vozil z nizkimi emisijami za potrebe naročnika."
  "estimatedPrice": {
    "netAmount":"1.006.200,00"
    "currency":"EUR"
  }
}
```

### Fact to remember

One parsed document contains structured information from one raw document. There is a direct relation between one parsed document and the corresponding raw document.

## Formatting - clean data

In the next phase of a publication's lifecycle each piece of information contained in the structured data is converted into a proper data type and format so that all data of the same type have the same structure and are easily processable in the next steps. This can be seen as an understanding of the data content. For example the same date can be published in several formats but at the end all its versions have to be converted into the same value. The following list shows different (but not all) versions of one date:

- 2015-05-02
    - 2015-05-02
    - 2015-02-05
    - 2.5.2015
    - 2/5/2015
    - May 2, 2015

Furthermore, some information that is missing but can be derived from other existing facts can be added in this phase. All procedures that are applied during this stage of data processing are described in the chapter entitled *Data cleaning*

**Parsed document**


```
{
  "title":"Svoz odpadu"
  "cpvs":[
    {
      "code":"34110000"
      "isMain":"Ano"
    }
  ],
  "publications":[
    {
      "publicationDate":"2.5.2015"
      "sourceFormType":"Oznámení o zadání
zakázky"
    }
  ],
  "supplyType":"Služby"
  "estimatedPrice": {
    "netAmount":"1.006.200,00"
    "currency":"cz koruna"
  }
}
```


**Clean document**


```
{
  "title":"Svoz odpadu"
  "cpvs":[
    {
      "code":"34110000"
      "isMain":true
    }
  ],
  "publications":[
    {
      "publicationDate":"2015-05-02"
      "sourceFormType":"Oznámení o zadání zakázky"
      "formType":"CONTRACT_AWARD"
    }
  ],
  "supplyType":"SERVICES"
  "estimatedPrice": {
    "netAmount":1006200
    "currency":"CZK"
  }
}
```


### Fact to remember

The content of one clean document is based on the content of one parsed document. There is a clear relation between each clean and parsed document.

## Linking related information - matched data

When all processes that can be performed with a single publication are completed, data from this publication are linked with data from other publications. Two main processes start at this point

- Body matching
- Tender matching

### Body matching

All objects describing a body (buyers, bidders, tender administrator, tender supervisor, etc.) are extracted from the publication and each of these objects is matched with other bodies separately.

Matching can be seen as a process of assigning one object with a direct relation to a specific publication to a group of objects that describes the same real-world entity.

This process also serves as a tool for the integration of the public procurement database and the external company database.

This process is crucial so that in a final database we can detect all tenders

- from the same buyer
- supplied by the same company
- administered by the same entity
- etc.

The whole process of body matching is described in separate chapter of this document

### Tender matching

Each clean document represents one publication that contains data about a specific public procurement. The most common approach to publishing tendering information is that each piece of information update is published separately. This is also an approach that the TED journal uses. This means there can be 1 to N publications describing the same tender. The tender matching phase involves linking all such publications by adding them into one group.

It requires knowledge of how each source works to be able to define a proper rule to group together all publications from one source describing one tender. Some sources publish a tender specific identificator that is referenced in each publication, other sources publish references to previous publications etc. All methods used for tender matching are described in a separate chapter of this document.

### Example

To demonstrate the process described, let’s imagine a model tender publication:

```
{
  "title":"Coffey supplies"
  buyers[
    {
      "name":"Ministry of pleasant afternoon"
      "street":"Sunny street 1"
      "city":"London"
    }
  ],
  "publications":[
    {
      "formType":"CONTRACT_AWARD"
      "sourceTenderId":"ABC123"
    }
  ]
  "lots":[
    {
      "bids":[
        "bidders":[
          {
            "name":"Peter's coffe house"
            "street":"Abbey Road 5"
            "city":"Liverpool"
          }
        ]
      ]
    }
  ]
}
```


During the body matching process, two objects are extracted from the whole document

**Body 1**


```
{
  "name":"Ministry of pleasant afternoon"
  "street":"Sunny street 1"
  "city":"London"
}
```


**Body 2**


```
{
  "name":"Peter's coffe house"
  "street":"Abbey Road 5"
  "city":"Liverpool"
}
```


Each body is now matched separately. No matching or similar body was found for Body 1; therefore, a new group is created and the matched body record is stored into the database


```
{
  "id":"123123123123"
  "name":"Ministry of pleasant afternoon"
  "street":"Sunny street 1"
  "city":"London"
  "groupId":"group_asdfghjkl123456789"
}
```

On the other hand, for Body 2 the matching process found another record and assigned Body 2 to the same group and stored the new matched body record into the database

**Body 2**


```
{
  "id":"111222333444"
  "name":"Peter's coffe house"
  "street":"Abbey Road 5"
  "city":"Liverpool"
  "groupId":"group_11111111122222222"
}
```


**Body X**


```
{
  "id":"232323232323"
  "name":"Peter's coffe"
  "street":"Abbey Road"
  "bodyIds":[
    {
      "type":"VAT_ID"
      "id":"GB13245678"
    }
  ]
  "groupId":"group_11111111122222222"
}
```


After the body matching is completed, the tender matching starts. The matched publication now does not contain buyer info but only a reference to a matched body record. The matching process found another publication based on the same *sourceTenderId* value and assigned the new publication to the same group and stored the new matched publication into the database

**Publication A**


```
{
  "title":"Coffey supplies"
  buyers[
    {
      "id":"123123123123"
      "groupId":"group_asdfghjkl123456789"
    }
  ],
  "publications":[
    {
      "formType":"CONTRACT_AWARD"
      "sourceTenderId":"ABC123"
    }
  ]
  "lots":[
    {
      "bids":[
        "bidders":[
          {
            "id":"111222333444"
            "groupId":"group_11111111222222222"
          }
        ]
      ]
    }
  ]
}
```


**Publication B**


```
{
  "title":"Coffey supplies"
  "awardDecisionDate":"2015-10-08"
  "estimatedPrice": {
    "netAmount":10500
    "currency":"GBP"
  }
  "publications":[
    {
      "formType":"CONTRACT_NOTICE"
      "sourceTenderId":"ABC123"
    }
  ]
}
```

When all bodies from a publication are matched and the publication itself is matched this phase of data processing ends.

### Facts to remember

- Tender and body matching means finding other objects in the database that describe the same real-world entity.
- Bodies and tender publications are matched separately

## Data merging - master data

When all data describing one real-world public procurement are linked together, a final image of a tender that contains all known information about a procurement's lifecycle is created. At this stage of data processing several publications describing one fact can be linked together and each of these publications can contain the same type of information as some other publication or some completely different type of information. Data mastering is a process of applying complex business rules to linked tender publications. Its goal is to create a final representation of a real-world public procurement that describes it in all aspects as precisely as possible.

A group of bodies are mastered separately to create a final representation of a real world entity.

**Matched body 1**


```
{
  "id":"111222333444"
  "name":"Peter's coffe house"
  "street":"Abbey Road 5"
  "city":"Liverpool"
  "groupId":"group_111111111222222222"
}
```


**Matched body 2**


```
{
  "id":"232323232323"
  "name":"Peter's coffe"
  "street":"Abbey Road"
  "bodyIds":[
    {
      "type":"VAT_ID"
      "id":"GB12345678"
    }
  ]
  "groupId":"group_111111111222222222"
}
```


**Master body**


```
{

  "id":"123456654321"
  "name":"Peter's coffe house"
  "street":"Abbey Road 5"
  "city":"Liverpool"
  "bodyIds":[
    {
      "type":"VAT_ID"
      "id":"GB12345678"
    }
  ]
  "groupId":"group_111111111222222222"
}
```


Also tenders are mastered separately

**Matched tender publication 1**


```
{
 "title":"Supplies of Arabic coffey"
 "publications":[
   {
     "formType":"CONTRACT_AWARD"
     "sourceTenderId":"ABC123"
     "publicationDate":"2015-06-20"
   }
 ]
 "lots":[
   {
     "bids":[
       "isWinning":true,
       "bidders":[
         {
           "id":"111222333444"
           "groupId":"group_11111111222222222"
         }
       ]
     ]
   }
 ]
 "groupId":"group_tender_123"
}
```


**Matched tender publication 2**


```
{
 "title":"Coffey supplies"
 "publications":[
   {
     "formType":"CONTRACT_NOTICE"
     "sourceTenderId":"ABC123"
     "publicationDate":"2015-05-15"
   }
 ]
 "bidDeadline":2015-06-15
 "awardDecisionDate":2015-06-18
 "procedureType":"OPEN"
 "groupId":"group_tender_123"
}
```


**Master tender**


```
{
 "title":"Supplies of Arabic coffey"
 "publications":[
   {
     "formType":"CONTRACT_AWARD"
     "sourceTenderId":"ABC123"
     "publicationDate":"2015-06-20"
   },
   {
     "formType":"CONTRACT_NOTICE"
     "sourceTenderId":"ABC123"
     "publicationDate":"2015-05-15"
   }
 ]
 "lots":[
   {
     "bids":[
       "isWinning":true,
       "bidders":[
         {
           "id":"111222333444"
           "groupId":"group_11111111222222222"
         }
       ]
     ]
   }
 ]
 "bidDeadline":2015-06-15
 "awardDecisionDate":2015-06-18
 "procedureType":"OPEN"
 "groupId":"group_tender_123"
}
```


As a last step the master bodies and master tenders are merged together and a final representation of a specific public procurement is created.

**Final public procurement representative**


```
{
  "title":"Supplies of Arabic coffey"
  "publications":[
    {
      "formType":"CONTRACT_AWARD"
      "sourceTenderId":"ABC123"
      "publicationDate":"2015-06-20"
    },
    {
      "formType":"CONTRACT_NOTICE"
      "sourceTenderId":"ABC123"
      "publicationDate":"2015-05-15"
    }
  ]
  "lots":[
    {
      "bids":[
        "isWinning":true,
        "bidders":[
          {
            "id":"123456654321"
            "name":"Peter's coffe house"
            "street":"Abbey Road 5"
            "city":"Liverpool"
            "bodyIds":[
              {
                "type":"VAT_ID"
                "id":"GB12345678"
              }
            ]
            "groupId":"group_111111111222222222"
          }
        ]
      ]
    }
  ]
  "bidDeadline":2015-06-15
  "awardDecisionDate":2015-06-18
  "procedureType":"OPEN"
  "groupId":"group_tender_123"
}
```


Fact to remember

A master tender is an object describing one real-world public tender compiled from all known information

# The Project in the Broader Context of Open Government Data

The collection, warehousing and use of public data is a quickly evolving domain at the intersection of public policy and data science. Indeed the heterogeneity in quality and structure of the data sources used in this project suggest that there are few established best practices in online, public data science. Nevertheless, the past 10 years have witnessed the emergence of an academic literature (/cite http://onlinelibrary.wiley.com/doi/10.1002/poi3.147/full) on open data and informed collections of standards on the topic. In this section we highlight two such standards which inform the construction of the database, followed by background references to the data mining literature consulted for the computationally intensive parts of the database construction.

## Standards and Best Practices

Standards offer two dimensions of value to DIGIWHIST. First, they inform the collection, construction and dissemination of the deliverable dataset. In other words, they highlight how the dataset should be created to give maximum value to stakeholders, especially the broad population of digital users. Second, perhaps less obviously, they provide a framework or checklist for understanding the project's diverse data sources. The deficiencies in various national procurement portals, from the perspective of a comprehensive list of standards such as those highlighted below, guided the development of the database. For example, data was not available in multiple formats from several national portals, necessitating the selection of a common "mother" data format to map all data to. After constructing the database in a unified format, it can be made available in multiple formats, in line with best practices as discussed below.

A variety of standards for storing and sharing data on the web have been developed in recent years. Perhaps the most famous is the World Wide Web Consortium's (W3C) continuously updated *Data on the Web Best Practices* (https://www.w3.org/TR/dwbp/). This document aims to set best practices for data publishing, reuse, machine and human testing, framed in normative terms using intended outcomes. At the time of writing,, 35 best practices are listed in this reference, covering everything from how to document data to how to share it. As suggested above, these practices inform both the construction of the DIGIWHIST data deliverables, and the evaluation and translation of the various data sources.

W3C Best Practices:

| | |
|---|---|
| Best Practice 1: Provide metadata | Best Practice 19: Use content negotiation for serving data available in multiple formats |
| Best Practice 2: Provide descriptive metadata | Best Practice 20: Provide real-time access |

| | |
|---|---|
| Best Practice 3: Provide structural metadata | Best Practice 21: Provide data up to date |
| Best Practice 4: Provide data license information | Best Practice 22: Provide an explanation for data that is not available |
| Best Practice 5: Provide data provenance information | Best Practice 23: Make data available through an API |
| Best Practice 6: Provide data quality information | Best Practice 24: Use Web Standards as the foundation of APIs |
| Best Practice 7: Provide a version indicator | Best Practice 25: Provide complete documentation for your API |
| Best Practice 8: Provide version history | Best Practice 26: Avoid Breaking Changes to Your API |
| Best Practice 9: Use persistent URIs as identifiers of datasets | Best Practice 27: Preserve identifiers |
| Best Practice 10: Use persistent URIs as identifiers within datasets | Best Practice 28: Assess dataset coverage |
| Best Practice 11: Assign URIs to dataset versions and series | Best Practice 29: Gather feedback from data consumers |
| Best Practice 12: Use machine-readable standardized data formats | Best Practice 30: Make feedback available |
| Best Practice 13: Use locale-neutral data representations | Best Practice 31: Enrich data by generating new data |
| Best Practice 14: Provide data in multiple formats | Best Practice 32: Provide Complementary Presentations |
| Best Practice 15: Reuse vocabularies, preferably standardized ones | Best Practice 33: Provide Feedback to the Original Publisher |
| Best Practice 16: Choose the right formalization level | Best Practice 34: Follow Licensing Terms |

| | |
|---|---|
| Best Practice 17: Provide bulk download | Best Practice 35: Cite the Original Publication |
| Best Practice 18: Provide Subsets for Large Datasets | |

Another guide, more specialized, though not exclusively so, to the aspects of government data is the *Open Data Handbook* *http://opendatahandbook.org/guide/en/)*, compiled by Open Knowledge International. This guide is especially applicable to the DIGIWHIST project, as it considers how government data may be organized and shared to be maximally useful to a wide variety of users and stakeholders. It suggests that data scientists and engineers disseminating processed public data cannot anticipate all intended uses or applications of the data, and should therefore adopt a general and modular approach to their data delivery pipeline. The emphasis is on the usability and extendability of open data. The *Open Data Handbook* defines open knowledge in the following way:

> "Knowledge is open if anyone is free to access, use, modify, and share it — subject, at most, to measures that preserve provenance and openness."

Practically it highlights three details relevant to data projects to ensure their "openness": availability and access, re-use and redistribution, and universal participation. The DIGIWHIST project heeds these points both philosophically and practically in the database described in this document.

## Related Data Mining methods

The field of data mining is a relatively mature field. In this section, we briefly review references in the literature that are helpful to understand our solutions to the two primary computational challenges in our pipeline: record matching/linking and deduplication.

A naïve approach to record matching or deduplication requires the comparison of all pairs of records, which grows quadratically in the number of records. In the case of one million records, a naïve pairwise approach requires one trillion comparisons. If one hundred thousand records can be compared in one second, one trillion comparisons would take over one hundred days. Besides the computational difficulty intrinsic to any large-scale linking or deduplication task, care in handling non-uniformly messy data is essential to achieving an accurate, useful result.

Peter Christen (Christen, Peter. 2012. *Data matching: concepts and techniques for record linkage, entity resolution, and duplicate detection*. Springer Science & Business Media.) suggests the following workflow for record deduplication and record linkage processes:

1) Cleaning and Preprocessing: in which records are standardized by, for example, removing unwanted characters, expanding abbreviations, or correcting spelling errors.

2) Indexing: reducing the total pairs of records to compare by *blocking,* or grouping records in a way that matches between blocks are impossible or improbable.
3) Comparing: comparing candidate matches across a variety of data dimensions, using similarity measures appropriate the nature of each dimension. For example, we use string similarity measures like the *Levenshtein distance* to quantify the similarity of two names.
4) Classifying: once candidate matches have been compared across several dimensions, one must create a classification algorithm to decide which candidate matches are to be identified (respectively, which records are to be classified as duplicates).
5) Evaluation: the resulting classification must be evaluated by comparison with ground truth. A variety of measures are available to compare the accuracy of different classifications. At this step it is crucial to consider both false positives and false negatives in the evaluation of the classification.

The ideal data mining pipeline for tasks like record matching and deduplication can only be built iteratively and with domain-specific expertise informing the wide range of choices to be made throughout the process. We took an approach that was cognisant of these two key ingredients.

# Sources

This chapter describes all public procurement data sources that were processed within the DIGIWHIST project, what data were downloaded and how each particular source is handled. It also contains a description of how different tender publications (describing one tender) are matched together because it's always a source specific algorithm

## Public procurement data

### Bulgaria

#### Source credentials

**Source search url** http://rop3-app1.aop.bg:7778/portal/page?_pageid=93,662251&_dad=portal&_schema=PORTAL

#### Source structure

- HTML web portal
- Contains a search form
- A day by day search is allowed only to registered users, therefore, a data update requires a download of the complete history
- Result HTML page contains paged list of tenders
    - Each page allows user to click the next page link
        - If this link is not visible, the end of a list is reached and there is nothing to crawl

    - one result item (on result page) contains a link to a tender detail
        - Tender detail page contains links to all publications
        - Even if there are more publications for a specific tender it appears only once in a search result

The grouping of the all tender records describing the same public procurement is based on a buyer assigned ID.

## Croatia

### Source credentials

**Source url** https://eojn.nn.hr/Oglasnik/
**Source search url**
https://eojn.nn.hr/SPIN/application/ipn/PreglediFrm.aspx?method=ReducedObjavljeniDokumenti

### Source structure

- HTML source
- Page includes search form
- Search form allows daily updates based on the search form field **Datum objave**
- Data since 4.1.2008
- Result set page displays only first 200 items (20 pages, 10 items per page)
- Each link points to a web page that contains a publication summary and a download button for a detailed publication information
- The structure of HTML page changed three times in its history so there are three different templates

The grouping of all the tender records describing the same public procurement is based on information on related publications.

## Czech Republic

### Source credentials

**Source url** https://www.vestnikverejnychzakazek.cz/
**Archive url** https://old.vestnikverejnychzakazek.cz/
**Source search url**
https://www.vestnikverejnychzakazek.cz/SearchForm/Search?SearchFormGrid-sort=PublishDate-desc&SearchFormGrid-pageSize=%1$d&FormDatePublishedFrom=%2$d&FormDatePublishedTo=%2$d
**Archive search url**
https://old.vestnikverejnychzakazek.cz/en/Searching/FullTextSearch?dateTimePublicationFrom=%1$s&dateTimePublicationTo=%1$s&size=%2$d&orderBy=PublishDate-asc

### Source structure

Both the archive portal and the new portal work on the same principle which we use for our crawling strategy.

- HTML web portal

- Contains a search form
- Search form allows to search by publication date
- Search parameters can be sent via a GET HTTP request
- GET parameters allow to define a size of a returned page
- Response is a HTML page that contains paged list of tenders
    - Each page allows user to click next page link
        - If this link is disabled we have reached the end of a list and there is nothing to crawl
    - Each row contains a link to a publication detail and to a list of all publications related to a particular tender
        - For each row in a list we download a notice detail page together with a list of related publications

### Source data

Both portals contains many different publications from which we process following

- Archive
    - Prior information notice
        - Source form type - 15
    - Contract notice
        - Source form type - 2, 5, 9, 11, 12
    - Contract award
        - Source form type - 3, 6, 13, 18
    - Contract implementation
        - Source form type - 54
    - Contract cancellation
        - Source form type - 51
- New portal
    - Prior information notice
        - Source form type - F01, CZ01
    - Contract notice
        - Source form type - F02, F04, F05, F12, CZ02
    - Contract award
        - Source form type - F03, F06, F13, F15, CZ03
    - Contract update
        - Source form type - F14, CZ04
    - Contract amendment
        - Source form type - F20

These notices cover the significant majority of all published notices. Some of published notices are not even related to public procurement and are not relevant for purposes of DIGIWHIST.

All notices on both portals contain information about a source tender id. This is an identificator assigned to a tender in a portal. This means all publications describing one tender can be matched together using this identificator.

## Estonia

### Source credentials

**Source url** https://riigihanked.riik.ee/
**Source search url** https://riigihanked.riik.ee/register/RegisterTeated.html

### Source structure

- HTML web portal
- Contains a search form
- Search form allows to search by publication date
- Each notice appears in a search result for each day when it was published or modified
- Result HTML page contains paged list of publications
    - Each page allows the user to click a next page link
        - If this link is disabled the end of a list is reached and there is nothing to crawl
    - Each row contains a link to a publication form and two important pieces of information, which are not in the notice:
        - "contracting authority" - there is similar information, but not the same, information in the notice
        - "Notice type"
- The maximum number of search results is 500, this limit was not reached when crawling daily

### Source data

From all the data published we process all publications of source form type

- *Contract notice* - "Hanketeade", "Hanketeade (Võrgustik)", "Hanketeade - kaitse- ja julgeolekuvaldkond"
- *Contract award* - "Riigihanke aruanne", "Ehitustööde kontsessiooni teade", "Ideekonkursi tulemused"
- *Prior information notice* - "Eelteade", "Eelteade - kaitse- ja julgeolekuvaldkond"
- *Contract implementation* - "Riigihanke aruande lisa"

For the other form types only publication meta information (form type, publication date etc.), title, and the buyer information are parsed

The grouping of all tender publications describing the same public procurement together is based on the tender ID. A backup strategy is based on matches of URLs of related publications.

## France

### Source credentials

**FTP source url** ftp://echanges.dila.gouv.fr/BOAMP/
**Publication web url** http://www.boamp.fr/avis/detail/xxx , where *xxx* = publication source ID (e.g. “16-80936”)

### Source structure

FTP:

- The FTP structure changed during the time of implementation. Originally, there were daily packages going back to 2008 but now it contains daily packages only for 2017. Historical data are also part of a DIGIWHIST procurement DB but the crawler is only able to download updates.
- The folder with current data updates contains *.taz archives. There are multiple archives for data from one day. Each stores different form type data.
- Daily updates are detected based on the modification date of each file.
- Each archive contains many XML and HTML files. A single publication is represented by both XML and HTML file and its name is the publication source ID plus extension

Web:

- There is the same information as on the FTP + references to related publication. For each FTP publication, a web publication is also downloaded and the data are merged together. Based on a related publication information all tender publications can be grouped together.

## Georgia

### Source credentials

**Source url** https://tenders.procurement.gov.ge/
**Source search url** https://tenders.procurement.gov.ge/public/?lang=en

### Source structure

- HTML web portal
- Contains a search form
- Search form allows to search by publication date but status date is used otherwise an updated tender is not a part of a filtered result set
- Search parameter (Status date from/to) is entered in search form
- Results are shown as a paged list. When the “next page” button is disabled the end of search result is reached and there is nothing to crawl
- Oldest record was published on 12.11.2010
- Tender information is not divided into types of publication (contract notice, contract award) but each tender detail contains compiled information about the whole tender from its announcement to current state.
- Source tender id is used to detect an updated tender record.

## Hungary

### Source credentials

**Source url** http://kozbeszerzes.hu/
**Source search url** http://kozbeszerzes.hu/adatbazis/keres/hirdetmeny/

### Source structure

- HTML web portal

- Contains a search form
- Search form allows searching by publication date
- Search parameter (publication date) is entered in a search form
- Results are shown as a paged list containing links to particular publications
- Publications since 1.1.2013 are processed. All are structurally the same. Previously published data were imported from AKKI's (DIGIWHIST partner) database.

Every publication, has a reference to the first notice (id) published for a particular public procurement, which is used for grouping publications together.

## Ireland

### Source credentials

**Source url** https://irl.eu-supply.com
**Source search url** https://irl.eu-supply.com/ctm/supplier/publictenders

### Source structure.

- HTML web portal
- Contains a search form
- Search form allows searching by publication date
- Search parameter (publication date) is entered in search form
- Results are shown as a paged list. When the "next page" button is disabled the end of search result iss reached and there is nothing to crawl
- Link from a result list does not always point directly to a publication detail page. Sometimes one more click is needed. The easiest way to get to publication detail is to get its PID from the url (not the same as System ID in first column of a result list) and use it to create url to detail which always has the same format
- The tender detail also includes references to TED-like publications which are also crawled by DIGIWHIST software

Tender ID is used for grouping publications together.

## Italy

### Source credentials

**Source url** http://portaletrasparenza.avcp.it
**Source search url** http://portaletrasparenza.avcp.it/microstrategy/html/index.htm

### Source description

- HTML web portal
- Search form that allows a day by day incremental approach is under the **RICERCA AVANZATA** tab
- The number of search results is limited to 300
- Oldest records from 01.01.2011

This source was not processed for two major reasons

- The non-standard and idiosyncratic technological structure of the web portal prevents this web portal from being crawled.
    - The web uses a combination of iframes, javascript, and AJAX and probably some type of javascript framework for displaying of results, page content loading etc.. This makes crawling much more difficult. It is not completely impossible but the end of the crawling process is unpredictable.
- Search form requirements.
    - It Is necessary to fill one of the required fields (Oggetto del bando - tender title, Amministrazione - buyer, Aggiudicatario - bidder), to select active or inactive tenders (Cerca in bandi).
    - The number of displayed results is 300, therefore the next criteria have to be added when the limit is reached for a day by day search (e.g. type of contract (e.g. Tipo Contratto).
    - Usage of Amministrazione required field was elaborated.
        - It allows a substring of subject name to be entered. The list of potential entities includes 462 007 entries, therefore a set of 125 trigrams that allows all entities to be found was detected and used to search tenders
        - This approach produces an excessive number of requests to the server which combined with a server's slow response time will end as an infinite crawling process (from a project's perspective) and might be considered as a DoS (denial of service) attack

## Latvia

### Source credentials

**FTP url** ftp://open.iub.gov.lv

### Source structure

- Structured XML files packed in .tar.gz archive
- Oldest package is from 1.1.2013
- Daily package path has format <year>/<month>_<year>/<day>_<month>_<year>.tar.gz (e.g. 2013/01_2013/01_01_2013.tar.gz)
- All publications are structurally the same

Tender matching is primarily based on the combination of buyer organization id (system ID) and tender id assigned by this buyer if exist otherwise, the related publications are used.

## Lithuania

### Source credentials

**Source (search) url** http://cvpp.lt/

### Source structure

- HTML web portal

- Contains a search form
- The search form allows searching by publication date. The search form searches tenders, not forms (publications). Each tender has one publication date. It means crawling of daily updates is not possible and the whole history has to be always searched from the beginning
- Search parameters can be sent via a GET HTTP request
- GET parameters allow the size of a returned page to be defined
- The oldest tender was published on 19.09.2008
- The response is an HTML page that contains paged list of tenders
    - Transition to the next page of a result set is based on clicking on a next page's number
    - Each row contains a link to a tender detail
        - The linked page does not contains detailed information about public procurement but a related publications list and another link to the notice detail

Matching of all the tender records describing the same public procurement is based on a buyer assigned ID.

## Netherlands

### Source credentials

**Source url** https://www.tenderned.nl
**Source search url**
https://www.tenderned.nl/tenderned-web/aankondiging/overzicht/aankondigingenplatform

### Source structure

- HTML web portal
- Contains a search form
- Search form allows searching by publication date
- The result HTML page contains paged list of publications
    - Each page allows the user to click a next page link
        - If this link is disabled the end of a list is reached and there is nothing to crawl
    - Each row contains a link to a publication detail. The publication detail page has four tabs:
        1. *Overview*
        2. *Publication* - this page is structured like TED
        3. *Documents* - this page contains all the documents for the tender.. The tab does not exist when no document is attached
        4. *Questions and answers*
- For each row in a result list the publication detail page is visited and the following tabs are downloaded:
    - Overview tab
    - Publication tab
    - Documents tab
    - Some publications cannot be downloaded because the page is broken

### Source data

Calls for tender and Contract awards are processed. The form structure differs according to the publication date. For each source form type there can be found three different historical templates. All of them are processed

"TenderNed attribute" ("TenderNed-kenmerk" in Dutch) on the overview page is a system ID for a specific public procurement. This is used to group different publications together.

## Norway

### Source credentials

**Source url** https://www.doffin.no
**Source search url** https://www.doffin.no/Notice

### Source structure

- HTML web portal
- Contains a search form
- The search form allows searching by publication date. It is necessary to search day by day, because each search displays a maximum of 1000 results
- "Include expired notices" check box has to be checked to get expired notices
- Result HTML page contains paged list of publications
    - Each page allows users to click the next page link
        - If this link is disabled the end of a list is reached and there is nothing to crawl
    - Each row contains a link to a HTML publication form and **Doffin reference** number
- The source also contains publication forms in XML format. The URL of the file looks like https://www.doffin.no/Eps.Searching/UnsupportedNotice/NoticeXml/xxx where *xxx* is Doffin reference
    - These XML publications are processed

Grouping of publications describing the same public procurement is based on

1. source IDs of previous publications
2. URLs of related documents which are referred from a  publication

## Poland

### Source credentials

**Source url** ftp://ftp.uzp.gov.pl
**User name** -
**Password** -

### Source structure

- FTP source
- Structured XML since 2007

    - Packages folder bzp/xml/<year>/
    - 2007/
        - year package 2007_xml.rar which includes folders 2007-<month>-<day> with XML files.
    - 2008/
        - year package 2008_xml.rar which includes EXE archives 2008<month><day>.exe with XML files.
        - some archives for 2008 are broken and it's not possible to extract data from them
    - 2009-01-01 and newer
        - daily packages as EXE archives <year><month><day>.exe
        - each package contains one XML file per tender publication

### Source data

All provided types of publication are processed.

- Contract award notice
    - Source form type - ZP-403, ZP-405, ZP-408
- Contract notice
    - Source form type - ZP-400, ZP-401, ZP-402, ZP-404
- Contract update
    - Source form type - ZP-SPR, ZP-406
- Others
    - Parsed only included publication.

Publications contain ids of previous and related publications. Publication ids are used to match various publications describing one tender together.

## Portugal

### Source credentials

**Source ur**l http://www.base.gov.pt

### Source structure

- HTML web portal
- Contains a search form
- Search form allows searching by publication date
- Two search forms exist (in English it is called "Contracts" and "Notices"):
    - Contracts - one tender can have many contracts. The detail page has a link to the notice when the tender is awarded (see row "Notices" on http://www.base.gov.pt/Base/en/Search/Contract?a=2065659).
    - Notices - one tender has one notice. The detail page has a link to the list of its contracts when the tender is awarded (see row "Link to contracts" on http://www.base.gov.pt/Base/en/Search/Notice?a=75223).
- Result HTML page contains a paged list of publications
    - Each page allows users to click the next page link

        - If this link is not visible, the end of a list is reached and there is nothing to crawl
    - Each row contains a link to a publication detail.

Grouping of all tender publications describing the same public procurement together is based on its list of related publications' URLs.

## Romania

### Source credentials

**Source url** http://data.gov.ro/
**Source search url** http://data.gov.ro/dataset/achizitii-publice-2007-2016-contracte6

### Crawling strategy

- Data are stored in CSV files
- All files has the same structure
- All links to CSVs are on one page
- All data are downloaded always when the crawler starts

Buyer assigned ID is used for grouping of related publications.

## Serbia

### Source credentials

http://portal.ujn.gov.rs/OpenData.aspx

### Source description

This is an open data source that provides data in CSV format. After a detailed examination, it became clear that it contains too few variables for supporting DIGIWHIST analytical and indicator building goals.

## Slovakia

### Source credentials

**Source url** https://www.uvo.gov.sk
**Source search url** https://www.uvo.gov.sk/dolezite/vestnik-a-registre/vestnik-590.html

### Source structure

- HTML web portal
- Contains a search form
- The search form allows searching by publication date
- The search parameter (publication date) is sent in a url
- All result are shown on the results page, so there is no need for additional navigation

### Source data

Publications structure was changed twice in time; our working names are ancient, old and new. So there are three different templates for each of following source form types:

- Contract notice
    - Source form type - MDP, MDS, MDT, MNA, MRP, MRS, MRT, MSP, MSS, MST, MUP, MUS, MUT, POT, WYP, WYS, WYT
- Contract award
    - Source form type - IPP, IPS, IPT, VBP, VBS, VBT, VDP, VDS, VDT, VEP, VKP, VKS, VNA, VNS, VRP, VRS, VRT, VSP, VSS, VST, VUP, VUS, VUT, ICP
- Contract cancellation
    - Source form type - ZBP, ZBS, ZBT, ZDP, ZDS, ZDT, ZNA, ZRP, ZRS, ZRT, ZSP, ZSS, ZST, ZUP, ZUS, ZUT, ZWP, ZWS, ZWT
- Contract implementation
    - Source form type - VZP, VZS, VZT

Each publication contains ids of previous and related publications, which is used.

## Slovenia

### Source credentials

**Source url** http://www.enarocanje.si
**Source search url** http://www.enarocanje.si/?podrocje=pregledobjav

### Source structure

- HTML web portal
- Contains a search form
- The search form allows searching by publication date
- The result HTML page contains a paged list of publications
    - Each page allows user to click the next page link
        - If this link is disabled the end of a list iss reached and there is nothing to crawl
    - Each row contains a link to a publication form
- The maximum number of search results is 1000. This limit was never reached when crawling the source information day by day.
- A minimal number of notices cannot be processed because they refer to non existing PDF files (see http://www.enarocanje.si/Obrazci/?id_obrazec=31611 or http://www.enarocanje.si/Obrazci/?id_obrazec=31651)

### Source data

Old publications have a  different structure from new ones. From all the data published, all publications of the listed form types are processed

- *Contract notice* - "EU 2 - SL", "EU 5 - SL", "NMV1", "PZPPO1 - ZJNVETPS", "PZPPO1 - ZJN-2", "PZP"

- *Contract award* - "EU 3 - SL", "EU 6 - SL", "NMV2", "EU 18 - SL", "PZPPO2 - ZJN-2", "PZPPO2 - ZJNVETPS"
- *Contract implementation* - "OS - ZJN-2", "OS - ZJNVETPS"

The grouping of publications describing the same public procurement is based on related publications' URLs.

## Spain

### Source credentials

**Source url** https://contrataciondelestado.es/
**Source search url**
https://contrataciondelestado.es/wps/portal/!ut/p/b1/jY5JDoJAFETP4gn-pydh2QINTVBQBqU3hIUxGIaN8fy2xq1I7Sp5L1VgoHEoJUi5yxAuYKbu2d-6Rz9P3fDuRrQszHxfxQTdggZI0qCqRGxrxC3QLAFknc-pz-qkzkWhI0QdqyCtHG51sc7HH5H4zz-DWUbIF1i6-AEWPhziebxCY7FtK-vwKLVHMdud7FCS78s8Ig4igxIaDaMZlPL0nXVy8wlYYbKy/dl4/d5/L2dBISEvZ0FBIS9nQSEh/pw/Z7_AVEQAI930OBRD02JPMTPG21004/act/id=0/p=javax.servlet.include.path_info=QCPjspQCPbusquedaQCPFormularioBusqueda.jsp/321178471136/-/

### Source structure

- HTML web portal
- Contains a search form
- The search form allows searching by publication date
- Result HTML page contains a paged list of tenders
    - Each page allows user to click next page link
        - If this link is missing the end of the list is reached and there is nothing to crawl
    - Each row contains a link to a tender (not only publication) detail
        - For each row in a list the tender detail page has to be visited and the following information is downloaded:
            - the tender detail page - an HTML page containing a summary of a tender compiled from all publications
            - referenced XML files representing publications of the tender
- Attaching a new form to a tender causes that tender to appear repeatedly in a daily search. It can happen that some tenders have 3 different XMLs published on 3 different days. In such a case the tender can be found for 3 different days and the crawler detects 9 files to download = 2 (the first publication is added, the detail page and first XML publication is downloaded) + 3 (the second publication is added, the detail page and first two XML publications are downloaded) + 4 (3 XMLs and 1 tender detail). This causes duplicities in crawled data that has to be detected later.

Matching of the tender publications describing the same public procurement is based on its XML URL. The tender detail page knows the URLs of all its XML publications which allows the grouping of all tender publications together.

## Switzerland

### Source credentials

**Source (search) url**
https://www.simap.ch/shabforms/COMMON/search/searchresultDetail.jsf

### Source structure

- HTML web portal
- Contains a search form
- Search form allows to search by publication date
- Result HTML page contains a paged list of tenders
- All pages have the same url
- The publication detail is not accessible via a permanent link
- Two different templates are processed *Invitations* and *Awards*

The grouping of the all tender records describing the same public procurement is based on a buyer assigned ID.

## TED

### Source credentials

**Source url** ftp://ted.europa.eu/
**User name** guest
**Password** guest

### Crawling strategy

- FTP source
- Structured XML data since 01.01.2011
- Daily or monthly packages
    - Daily package repository structure root/year/month/package_name
        - daily-packages/2011/01/20110104_2011001.tar.gz
    - Each daily package contains multiple files, one file per notice
    - Daily packages are crawled day by day and all notices are downloaded

### Source data

XML data in TED have two different structures; one follows the old directive and one follows the 2014 directive. The visual format for all notices is described on the TED website[1]. The structured format of both old and new data is documented on the TED FTP server[2].

From all data published in TED we process all structured XML publications of form type

- 2 - Contract notice
- 3 - Contract award

[1] http://simap.ted.europa.eu/standard-forms-for-public-procurement
[2] ftp://ted.europa.eu/Resources/

- 5 - Contract notice - utilities
- 6 - Contract award - utilities

These publications cover roughly 75% of all TED publications

The matching of the all tender records describing the same public procurement is based on related publications information. Each form also contains a list of previously published notices, therefore, all tender records that have at least one common publication in tender.publications list describes one public procurement.

### United Kingdom

**Source ur**l https://www.contractsfinder.service.gov.uk/
**Archive ur**l https://data.gov.uk/data/contracts-finder-archive/static/files/
**Source search url** https://www.contractsfinder.service.gov.uk/Search
**Archive search url** https://data.gov.uk/data/contracts-finder-archive/static/files/notices_<year>_<month>.xml

#### New portal

- HTML web portal
- Contains a search form
- The search form allows searching by publication date
- The search parameter (publication date) is entered in the search form
- A daily XML package is downloaded by pressing the button 'Download as XML'
- Data are published starting from 17.12.2014
- All data are structurally the same.

#### Archive

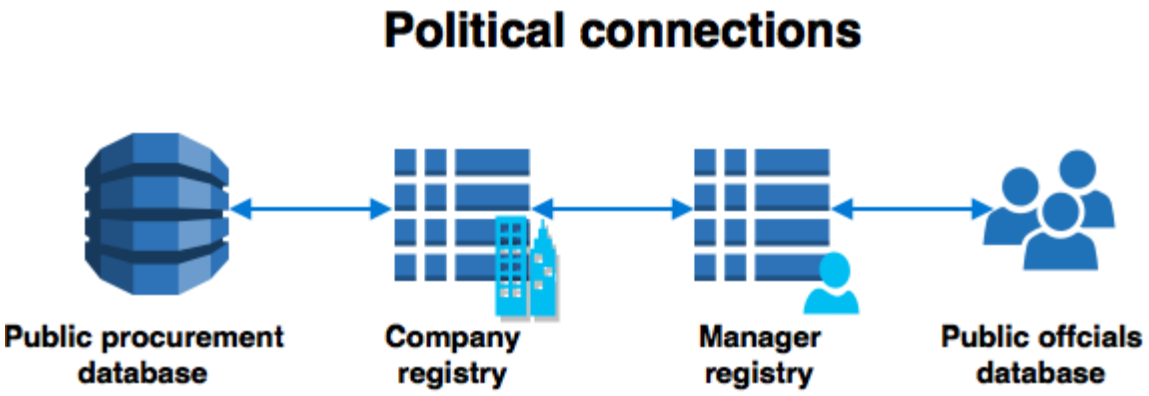
- The month package is published on the appropriate URL (see ‘Archive search url’ above)
- It contains data in XML format published since 01.01.2011 until 01.01.2015, all are structurally the same.

The source id is used for matching.

## Company data

The company database consists of four parts
- Company registry
- Financial information
- Links
- Manager registry

As an output of the integration of the public procurement database, company database and public officials database two indicators were designed.
- Tax haven
- Political connections of suppliers




## Company registry

This contains basic information about companies, public institutions etc. In DIGIWHIST terminology it's an external register of bodies. It serves two different purposes:

- Improves the results of the body matching process because it contains useful data like structured addresses, names, different IDs (VAT number, statistical number )
- Is an intermediary between the public procurement database and the manager information DB

More about integration of the public procurement database and company database can be found in the chapter "Body matching".

## Financial information

Financial information contains general company financial measures like yearly revenue, assets, before or after tax profit or losses etc. This information was used to test and generate advanced corruption risk indicators[3]. Financial information can be used to analyse whether companies winning public contracts are more or less profitable than the market average. Furthermore, financial performance can be also connected to public procurement integrity (i.e. whether low integrity contracts are won by companies with extreme profitability or not) or simply to the level of competition of a given market.

## Links

The links database is a core of company ownership database. It contains links between records from the registry information database. In other words it says which company owns which company.

## Manager information

This database contains information about relations between people and companies. It says who was in which position, in which company from the company register. By integrating these data into the public officials database the chain from public procurements to public officials is finished.

[3] In order to avoid the potential ethical and legal concerns raised by the use of the word "corruption" and to reduce the risk of stigmatisation of a company or individual associated with a high risk of corruption, the Consortium intends to use the term "procurement integrity indicator" in the portals to denote corruption risk as defined by the statistical approaches to indicator development. The underlying conceptual and methodological innovations remain unchanged.

## Budget data

For the collection of the budget data, potential sources were evaluated on the basis of their usability for DIGIWHIST purposes. Usability was primarily determined by whether the budget data has the potential to be meaningfully linked with with procurement data and to be used for the calculation of risk indicators. In practice, that requires:

1. Machine-readable format
2. Granularity on a level of the contracting authority
3. Level of detail sufficient for identifying potential procurement expenditures

As it turned out, very few countries ultimately passed the test. This is mainly because of the granularity requirement: State budgets are generally available online, but they are often structured by budget chapters rather than by individual organisations. To account for this possibility, we prepared country-level buyers lists to identify such entities and cross-referenced these lists with available sources (including BvD data, which also appeared not to contain 95% of the buyers). The success rate of this strategy varied significantly. For some countries, the budget contained only broadly-stated programs that did not directly correspond with specific institutions/public procurement buyers. In other countries, the budget figures did not go beyond plain turnover. In still other cases, the institution or department names in the budget itself did not match entries on the list of known procurement buyers. The extensive log of work done while mapping country resources can be found online at https://github.com/digiwhist/wp2_documents/blob/master/country_mapping_notes.pdf.

In order to at least showcase the potential of linking budget and procurement data, countries with the highest potential to fulfil the above-mentioned criteria were chosen: UK (government), Spain (municipalities, government) and Czech Republic (municipalities, government).

All three resources were scraped into a single data template, further paired with buyers available in the public procurement database. This was typically done using exact matches of national ids, names or both (this differs by availability on source).

## Public officials data

Four sources were processed for the purposes of obtaining a database of public officials:

- http://everypolitician.org/
- http://www.politicaldatayearbook.com/
- http://rulers.org/
- https://www.cia.gov/library/publications/world-leaders-1/index.html

To integrate public officials information with the manager registry, all records have to be preprocessed to standardize names.

- trailing and ending whitespaces are removed
- white spaces between first and last name are replaced by a single whitespace
- titles like Mr., Ms, Ing., Dr., are also removed

Then an equality test on the whole name value is processed. If two names matches we judge them to be the same person. Additional criteria can be used but this is not available for all records. The following table shows how many records from the public officials database were linked to a manager information registry using a different combination of criteria.

| **Fields used in comparison** | **Unique persons identified** |
|---|---|
| name | 5286 |
| name + gender | 4109 |
| name + gender + year of birth | 910 |
| name + gender + date of birth | 503 |
| name + gender + (year of birth OR date of birth) | 910 |
| name + year of birth | 928 |
| name + date of birth | 506 |

# Data cleaning

Data cleaning is a process that transforms structured text information into an understandable set of information. This comprises of

- Data types conversion
- Imputation of missing information that can be derived from observed data

## Data types conversion

Basically we process several fundamental data types

- Texts
- URLs
- True/False values
- Dates
- Numbers
- Enumeration values

### Text

Text cleaning consists of several modification rules that are applied on data extracted from raw HTML, XML or CSV data. Different rules are applied on short and long strings.

### Short string

1. All occurrences of the Unicode spaces[4] are replaced with ordinary space and all occurrences of Unicode invisible characters are removed
2. All trailing white spaces are removed
3. All HTML4 entities[5] are replaced by a proper character
4. All white spaces are replaced by a single space character

Fields cleaned as a short strings are listed below

| Entity | Fields |
|---|---|
| Tender | title, titleEnglish, buyerAssignedId, eligibleBidLanguages, excessiveFrameworkAgreementJustification, nationalProcedureType, acceleratedProcedureJustification |
| Lot | contractNumber, title, titleEnglish |
| Address | city, street, postcode, nuts, state, country |
| AwardCriterion | name |
| Body | contactName, contactPoint, email, name, phone |
| CPV | code |
| Corrigendum | sectionNumber |
| Document | format, language, title |
| Funding | programme, source |
| Publication | buyerAssignedId, sourceFormType, language, sourceId, sourceTenderId |

### Long string

1. All occurrences of the Unicode spaces are replaced with ordinary space and all occurrences of Unicode invisible characters are removed
2. Specific HTML tags are replaced by the new line character
    a. <br>, <p>, <ul>, <li>
3. All HTML4 entities are replaced by a proper character

Fields cleaned as a long string are listed below

| Entity | Fields |
|---|---|

[4] https://www.cs.tut.fi/~jkorpela/chars/spaces.html
[5] https://www.w3schools.com/charsets/ref_html_entities_4.asp

| | |
|---|---|
| Tender | description, descriptionEnglish, personalRequirements, economicRequirements, technicalRequirements, nationalProcedureType, deposits, appealBodyName, mediationBodyName, cancellationReason, modificationReason, modificationReasonDescription, eligibilityCriteria, additionalInfo |
| Lot | cancellationReason, description, descriptionEnglish, eligibilityCriteria |
| Bid | disqualificationReason |
| Address | rawAddress |
| AwardCriterion | description |
| Corrigendum | placeOfModifiedText, original, replacement |
| Document | description |

## True/False value

All values that are defined as a boolean values (true or false) are cleaned from all unicode spaces and invisible characters in the same way as short or long string. Then we convert it's text representation to true or false value using a library function[6] implemented by Apache software foundation.

Fields cleaned as a true/false values are listed below

| Entity | Fields |
|---|---|
| Tender | hasLots, hasOptions, areVariantsAccepted, isCentralProcurement, isCoveredByGpa, isDps, isEInvoiceAccepted, isElectronicAuction, isFrameworkAgreement, isJointProcurement, isOnBehalfOf, isWholeTenderCancelled, documentsPayable, isDocumentsAccessRestricted |
| Lot | isAwarded, isCoveredByGpa, isDps, isElectronicAuction, isFrameworkAgreement |
| Bid | isConsortium, isDisqualified, isSubcontracted, isWinning, |

[6] https://commons.apache.org/proper/commons-lang/apidocs/org/apache/commons/lang3/BooleanUtils.html#toBoolean-java.lang.String-

| | |
|---|---|
| | wasFinishedOnTime, wasForEstimatedValue, wasInRequestedQuality |
| Body | isLeader, isPublic, isSectoral, isSme, isSubsidized |
| CPV | isMain |
| AwardCriterion | isPriceRelated |
| Funding | isEuFund |
| Publication | isIncluded, isParentTender, isValid |

## URL

The most common typos are being fixed and replaced in published data if the original value is not in a proper URL form. If even after these fixes URL is not in proper form we erase it from a clean DB.

Fields cleaned as URL are listed below

| Entity | Fields |
|---|---|
| Tender | courtInterventions, courtProceedings |
| Publication | humanReadableUrl, machineReadableUrl |
| Address | url |
| Document | url |

## Date

Each source uses a different date format based on local conventions. Some sources even use multiple date formats. When developing programs for data extractions developers detected all possible formats used in a specific source. When converting text values to date values all possible date formats are used for transformation. When a transformation is successful a particular field is stored as a date. If all transformations fail, a value for particular field is not stored

As an example date formats used in

- Czech procurement journal
    - d/M/yyyy, d. M. yyyy, d.M.yyyy, yyyy/M/d
- TED
    - yyyy-MM-dd, yyyy-M-d

Fields cleaned as a dates are listed below

| Entity | Fields |
|---|---|
| Tender | awardDeadline, awardDecisionDate, cancellationDate, contractSignatureDate, enquiryDeadline, estimatedCompletionDate, estimatedStartDate, |
| Lot | awardDecisionDate, cancellationDate, completionDate, contractSignatureDate, estimatedCompletionDate, estimatedStartDate |
| Publication | dispatchDate, lastUpdate, publicationDate |
| Corrigendum | replacementDate |
| Document | signatureDate |
| Payment | paymentDate |

## Numbers

As well as dates, numbers are also handled differently and may be published in different formats in different countries based on local rules. This means various number formats are tested to make a transformation from text to number value. Before the transformation can start the text value is preprocessed as a short text. This means all ballast information like trailing empty spaces, new line characters, multiple empty spaces etc. are replaced or removed from the test. If all transformations fail, a value for particular field is not stored

Fields cleaned as a numbers are listed below

| Entity | Fields |
|---|---|
| Tender | estimatedDurationInDays, estimatedDurationInMonths, estimatedDurationInYears, maxFrameworkAgreementParticipants, maxBidsCount, awardDeadlineDuration, envisagedCandidatesCount, envisagedMinCandidatesCount, envisagedMaxCandidatesCount |
| Lot | bidsCount, electronicBidsCount, estimatedDurationInDays, estimatedDurationInMonths, foreignCompaniesBidsCount, lotNumber, maxFrameworkAgreementParticipants, nonEuMemberStatesCompaniesBidsCount, otherEuMemberStatesCompaniesBidsCount, |

| | |
|---|---|
| | positionOnPage, smeBidsCount, validBidsCount |
| Bid | annualPriceYearsCount, monthlyPriceMonthsCount, subcontractedProportion |
| Publication | version |
| Price | amountWithVat, maxAmountWithVat, minAmountWithVat, maxNetAmount, minNetAmount, netAmount, netAmountEur, vat |
| AwardCriterion | weight |
| Corrigendum | lotNumber |
| Document | order |
| Funding | proportion |
| UnitPrice | unitNumber |

### Enumeration values

To be able to provide analysis of the final data we need to convert some fields from national or source specific values to uniform enumeration values. Mapping tables were created for these purposes manually and cleaning programs only applies these mappings to a source data.

**Updated version of mapping files is available online at**:
https://github.com/digiwhist/wp2_documents/blob/master/country_mapping/

## Calculation of missing information

On top of data type conversion we are also trying to impute missing information from existing values where it's not present

- Lot status update (based on publication.formType)
- Framework agreements lot merge
- Contract implementation handling
- Award criteria update
- Completion of price object
- Removal of nonsense objects

See Annex 1 for planned improvements

### Lot status update

Each tender lot can be in several stages depending on a tender's progress

| | |
|---|---|
| **PREPARED** | lot is prepared |
| **ANNOUNCED** | lot is publicly announced, bids are accepted or negotiated |
| **AWARDED** | lot is awarded and being fulfilled |
| **CANCELLED** | lot has been cancelled |
| **FINISHED** | lot was fulfilled and paid |

This information is very often missing in the published data but can be derived from a publication type or from other information.

We set lot.status as a

- CANCELLED - if tender.isWholeTenderCancelled = TRUE

Base on a publication type we set lot status as a

- PREPARED if publication.formType=PRIOR_INFORMATION_NOTICE
- ANNOUNCED if publication.formType=CONTRACT_NOTICE
- AWARDED if publication.formType=CONTRACT_AWARD
- CANCELLED if publication.formType=CONTRACT_CANCELLATION

## Framework agreements lot merge

Framework agreements usually don't have lots, so if there are lots, it's most likely just more winners of one lot and either someone published the data in a wrong way or it’s simply an imperfection of a source system that does not allow the publication of more winners for one tender lot.

If the tender

- has multiple lots and
- is a framework agreement (tender.isFrameworkAgreement = TRUE) and
- is CONTRACT_AWARD or CONTRACT_IMPLEMENTATION and
- has same tender.lot.bidsCount for all lots

Then move all bids under the first lot and delete all other lots.

## Contract implementation handling

The DIGIWHIST data model works with a word payment which is considered to be a value paid by a buyer to a supplier. It can be either

- The proportion of total contract value (e.g. one installment)
- The value of a contract based on a previous framework agreement

Contract implementations based on framework agreements are often published in the same way as contract awards, therefore, we have to clean the source data so that it correctly fits into the DIGIWHIST methodology

First, if a publication’s type is set to contract award and procedure type is set to MINITENDER we consider such publication to be a contract implementation instead of being a contract award.

Second, if a bid has no payments set we create one payment for each winning bid as

- bid.payment.price = bid.price

    - if bid.price is missing we take tender.finalPrice value
- bid.payment.date = publication date

To have payment information for contract implementation is very important for the data mastering process. See the Contract Implementation subchapter in the Mastering matched data chapter.

## Award criteria update

For each lot if there is information about the selection method and the selection method is set to Lowest price then we set a single award criterion for this lot as

- awardCriterion.name = PRICE
- awardCriterion.weight = 100
- awardCriterion.isPriceRelated = TRUE

## Completion of price object

During the data type conversion phase all fields *netAmount*, *netAmountEur*, *minNetAmount*, *maxNetAmount*, *minNetAmountWithVat*, *maxNetAmountWithVat*, *amountWithVat*, *vat* are either converted to a numeric value or set to an empty value. Also the currency field is converted into ISO 4217 code value if possible. Otherwise it's set to empty value.

If the price object contains *vat* information and does not have *netAmount*, *minNetAmount* or *maxNetAmount* value set then it's calculated if the price object contains corresponding amount with VAT.

If the *currency* is EUR and *netAmount* value is present then it's copied also into *netAmountEur* variable.

## Removal of nonsense objects

As a last step of data cleaning all invalid data are removed from a tender.

- Empty list
- Empty text, date, number and other non-complex data type fields
- Object that has all fields empty

Some objects have more strict rules defined otherwise it's not considered valid

- Publications - *source* field mustn't be empty
- Funding - *source* or *isEuFund* has to be filled in
- Award criterion - *name* has to be filled in
- Body identifier - *id* has to be filled in
- Payment - *price* has to be valid
- Price - at least one of *amountWithVat, maxAmountWithVat, maxNetAmount, minAmountWithVat, minNetAmount, netAmount, netAmountEur, netAmountNational* has to be filled in
- CPV - *code* has to be filled in

# Tender matching

Various publications from the same source describing one tender can be linked together using several different approaches. Sometimes more approaches can be used for one source but mostly one is suitable. Even if there are more suitable approaches only one that is considered to be most appropriate is used.

## Related publications

Each cleaned tender record also contains information about all or at least all previously published tender publications related to the same tender. In such cases, the intersection of related publications sets can be tested between each two records and if this set is not empty then both records belong to the same group. This comparison can be based on several fields of Publication object. It can be either equality of

- humanReadableUrl
- machineReadableUrl
- sourceId
- buyerAssignedId

| Field | Sources |
|---|---|
| sourceId | TED, Poland, Hungary, Latvia, UK, Norway, France, Croatia |
| humanReadableUrl | Slovakia, Estonia, Slovenia, Portugal |
| machineReadableUrl | Spain |
| buyerAssignedId | Romania |

## Source tender ID

Some source systems directly link each publication to a tender. This means that variable *tender.publication[isIncluded=true].sourceTenderId* has the same value for each publication related to the same tender. This rule is used for

- Czech Republic
- Netherlands
- Georgia
- Ireland

## Buyer assigned ID

In rare cases the only information that can be used for tender matching is a buyer assigned ID. This is an identifier that a buyer (not the source system) assigned to a whole public procurement  and each tender publication contains this information. This is an option that is susceptible to errors because any typo can happen when inserting a data into a source system. This rule is used for

- Switzerland
- Lithuania
- Bulgaria

# Body matching

Body matching is a name for a complex process that groups all body objects (objects describing buyer, supplier, tender administrator etc.) together. The goal of this process is to assign the same identifier to all objects describing the same real world entity. When this goal is reached aggregate statistics can be calculated using an assigned identificator.

DIGIWHIST team is not the only one team who is dealing with this task, therefore, we also draw upon a knowledge of other studies or field experts, such as team at DG GROW that produced study on cross-border penetration[7], or academics like Johannes Wachs[8] (Center for Network Studies at the Central European University) who employs such methods supporting the study of networks in public procurement.

## Idea

There are several fundamental requirements for the whole process

- it has to be able to work when an external company database is available and also if it's not available
- because the data are downloaded as an increments (e.g. on daily bases) the process has to be able to incorporate new data into an already existing dataset
- the whole solution has to be able to provide results within a day for the most complicated source when calculating from the beginning.

The algorithm design is based on four steps

1. Hash matching
    - Two bodies that can be considered the same are assigned to the same group
2. Manual matching
    - Two bodies that some human user considered the same are assigned to the same group
3. Exact matching
    - Two bodies that are identical in most significant fields are considered the same
4. Approximate matching
    - Additional variables are used to calculate a score for a body to body match. A score for a new body record is calculated using records that are already matched. The body will be assigned to the group for which it has the highest score, as long as the score is above a set threshold

If there is an external company database, steps 3 and 4 can be run first against such a DB. If a match against an external DB is successful such a record is copied to a matched body

[7] https://publications.europa.eu/en/publication-detail/-/publication/5c148423-39e2-11e7-a08e-01aa75ed71a1

[8] https://cns.ceu.edu/people/johannes-wachs

database. If in future some other body record has a successful match with the same record from the external DB than a new copy is not created but the body record is assigned to the same group as the original copy.

# Preprocessing

The whole process starts with data preprocessing. This can be imagined as a standardization of the fields that are used for matching two records. Namely *standardized name* and *standardized address* are calculated and a digest which will be explained later

## Standardized name

Name is one of the key fields that is naturally used to compare two records whether they represent the same entity or not. On the other hand, a match in name does not necessarily mean that two records are the same. To be able to make comparisons based on names as precise as possible we calculate its standardized form by:

- Trimming all trailing white spaces
- Converting it to lowercase
- Replacing all white spaces or group of white spaces by the single space
- Removing accents
- Applying special replacements

The most complicated step in this preprocessing are the special replacements. In most cases it means replacing differently published business entities by the same text value. For example all these variants (the list is incomplete) describes the same thing

- LTD, l.t.d., l t d
- Limited
- Co. limited
- Single private l.t.d.

Therefore, all these text values must be replaced by the same value so that match in name value returns true for companies like *Peter's coffey l.t.d.* and *Peter's coffey, limited*

It is not only a matter of business entities but also other synonyms like *Uni* and *University* that can be found in the data

## Standardized address

Address is also a piece of information that is used for comparing two records. The standardization of addresses runs in very similar way to the standardization of names. All fields are processed by a function that

- Trims all trailing white spaces
- Converts text to lowercase
- Replaces all white spaces or group of white spaces by the single space
- Removes accents

Final a standardized address is a concatenation of standardized *street*, *city* and *country* or if a structured address is not available then a concatenation of standardized *rawAddress* and *country*

### Digest

Digest is used for performance reasons to reduce a pool of bodies for which an approximate matching score is calculated

There is one digest calculated for each record. It's created as a concatenation of name digest + separator + address digest where

- name digest as first two alphanumerical symbols in a standardized name (turned into lowercase); Calculated only if its 3 chars long
- separator is a pipe char |
- address digest as first two alphabetical symbols and a first number (regardless of a number of digits) in a standardized address. If the street has no number, use next two alphabetical symbols instead. Calculated only if its at least 3 chars long.
    - if standardized address is not present because *city* and *country* is missing in the structured address and there is no *rawAddress*, the standardized address which is only for purposes of digest calculation is calculated as
        - standardize(street + city + postcode + country)
        - at least 2 of 4 variables must be present

## Hash matching

The idea behind this step is that two bodies which appear the same should be assigned into the same group. Even another steps in body matching process should ensure this characteristic of the algorithm but comparison of two hashes significantly increase a performance.
Hash function is a cryptographic function that encodes information in the way that the same information always produces the same hash code and different information is never interpreted as the same hash code.

The definition of hash function is the core task here.
If all information from a body record are encrypted then it ensures that no false positive matches will occur. The disadvantage is that only absolutely the same records will be matched together which is not a desired functionality.
The definition of hash function must use as little information as possible but still has to ensure that the information uniquely identifies a real world entity

The decision was made that a combination of standardized name and body identifier defines sufficiently the real world entity. The definition is

- Alphabetically order all body ids values
    - Body id value is a concatenation of *scope* and *id*
- Concatenate *standardized name* with all ordered body ids
- Encrypt with s sha256 [9]algorithm

If two bodies have the same hash code they are considered the same and assigned into the same group.

[9] https://en.wikipedia.org/wiki/SHA-2

In this case, for example, two company records will be considered the same if the name of a company and its VAT tax number won't change but the company can move to another address.

# Manual matching

This is a way for the human user to interfere with a body matching process. It allows a user to explicitly say that two records represent the same entity and can be used to apply a human correction. Each body that enters a body matching process is then assigned to a group based on this database of human corrections. If no match for a body is found based on this method it is processed by the following steps.

# Exact matching

## Company DB exact matching

This proceeds the same way as *Matched body exact matching*. If a match occurs, there is a check, whether the company DB entry doesn't exist in the matched items already. If so, the body is assigned to the existing group. Otherwise, a new group including both company DB entry and matched body is created.

## Matched bodies exact matching

The body is compared to each matched item, matches of following are checked:

- standardized name
- standardized address
- all available identifiers

The following rules are applied to find the best group of bodies for a body that is being matched:

- two identifiers are equal when their *id* and *scope* are equal
- each identifier can be used only for one match.
- A perfect match of at least two non-empty items is considered an exact match (for instance standardized name + identifier, two different identifiers, identifier + standardized address etc.)
- from all exactly matched bodies, the one with the highest matching score (+1 for each above-mentioned match) is selected
- if such a match occurs, the Body is assigned as a member of a group and matching ends.

# Approximate matching

## Matched bodies approximate matching

We identify pool of bodies for approximate matching as union of these two:

A. all records already found in exact matching with one perfectly matched field
B. all bodies within existing groups with the same *digest*

We compute match *S_i* of body with each member of pool i, as a number from 0-1.

- standardized name, weight 1
    - One or both values NULL, return 0.5
    - otherwise return trigram matching value
- standardized address, weight 1
    - One or both values NULL, return 0.5
    - otherwise return trigram matching value
- postcode, weight 0.2
    - perfect match, return 1
    - difference in one digit, return 0.5
    - One or both values NULL, return 0.5
    - otherwise, return 0
- nuts, weight 0.2
    - perfect match, return 1
    - difference in last digit in case of 5 digit code like CZ041, return 0.8
    - One or both lists NULL or empty automatically results in value 0.5
    - 0 otherwise
- ID match, weight 1
    - One or both lists NULL or empty automatically results in the value 0.5
    - If there is at least one comparable pair of IDs return maximum value from all comparisons. Comparable IDs have the same scope and both have not NULL bodyId.id value. The value of each comparison is
        - perfect match, return 1
        - difference in one digit, return 0.8
        - otherwise, return 0
    - If there are no comparable IDs (having the same scope and not NULL bodyId.id value), return 0.5

Finally, the *S_i* = weighted average of all match ratios
If some *S_i* > 0.75 we take the match with *max_i(S_i)* and matching ends.

## Company DB approximate matching

This proceeds the same way as *Matched bodies approximate matching*. The body pool is built as

A. all bodies from company DB having at least one field matched
    a. standardized name
    b. standardized address
    c. all available identifiers (matching only against the same type of identifier minding its scope)
B. all bodies from company DB with the same digest

If a match occurs, a new group including both the record from company DB and matched body is created.

# Mastering matched data

Data mastering stands for a process of applying complex business rules on associated tender publications and linked datasets with a clear goal to create one single representative for each tender

## Variable by variable mastering

Matched data provides several instances representing the same characteristic of the same tender coming from different publications. At this stage of data processing a decision must be taken on which of these values will be a final representation that most probably describes the reality. This means a rule for each variable has to be defined. Several generic rules were developed for this purpose and each of them is applied to multiple variables. Some fields can’t be handled in a simple way and require special treatment, therefore, specific rules were designed and developed for them. All the rules are described in the chapter *Master rules*.

### Entities

This chapter describes which rules are applied to which variable within a specific entity

#### Tender

| Rule | Fields |
|---|---|
| Modus + Last published value | buyerAssignedId, title, titleEnglish, procedureType, nationalProcedureType, isAcceleratedProcedure, description, descriptionEnglish, maxBidsCount, supplyType, size, furtherInformationProvider, specificationsProvider, bidsRecipient, specificationsCreator, appealBodyName, mediationBodyName, maxFrameworkAgreementParticipants, estimatedDurationInMonths, estimatedDurationInDays, estimatedDurationInYears, envisagedCandidatesCount, envisagedMinCandidatesCount, envisagedMaxCandidatesCount, awardDeadlineDuration, country |
| Last published value | bidDeadline, documentsDeadline, estimatedStartDate, estimatedCompletionDate, awardDecisionDate, contractSignatureDate, limitedCandidatesCountCriteria, selectionMethod, cancellationDate, cancellationReason, isWholeTenderCancelled, enquiryDeadline, awardDeadline |
| Logical OR | documentsPayable, isDocumentsAccessRestricted, isCentralProcurement, isJointProcurement, isOnBehalfOf, hasLots, areVariantsAccepted, hasOptions, isCoveredByGpa, isFrameworkAgreement, isDps, isElectronicAuction, isEInvoiceAccepted |

| | |
|---|---|
| Longest | deposits, eligibilityCriteria, personalRequirements, economicRequirements, technicalRequirements, excessiveFrameworkAgreementJustification, |
| Bodies array | buyers, onBehalfOf, administrators, supervisors, candidates, approachedBidders |
| Union | publications, courtProceedings, courtInterventions, npwpReasons, eligibleBidLanguages |
| Price | documentsPrice, estimatedPrice, finalPrice, |
| Address | documentsLocation, addressOfImplementation |

Lot

| Rule | Fields |
|---|---|
| Modus + Last published value | contractNumber, estimatedDurationInMonths, estimatedDurationInDays, estimatedDurationInYears, maxFrameworkAgreementParticipants, envisagedCandidatesCount, envisagedMinCandidatesCount, envisagedMaxCandidatesCount, bidsCount, validBidsCount, electronicBidsCount, foreignCompaniesBidsCount, SMEBidsCount, otherEUMemberStatesCompaniesBidsCount, onEUMemberStatesCompaniesBidsCount |
| Last published value | awardDecisionDate, contractSignatureDate, completionDate, cancellationDate, cancellationReason, selectionMethod, limitedCandidatesCountCriteria, status, estimatedStartDate, estimatedCompletionDate |
| Logical OR | isElectronicAuction, isFrameworkAgreement, isDps, isCoveredByGpa, areVariantsAccepted, hasOptions, isAwardedToGroupOfSuppliers |
| Longest | title, titleEnglish, description, descriptionEnglish, eligibilityCriteria |

Body

| Rule | Fields |
|---|---|
| Modus + Last published value | name, email, contactPoint, contactName, phone, |

| | |
|---|---|
| | buyerType |
| Logical OR | isPublic, isSubsidized, isSectoral, isSme |
| Union | mainActivities |
| Address | address |

### Bid

| Rule | Fields |
|---|---|
| Modus + Last published value | subcontractedProportion |
| Last published value | disqualificationReason |
| Logical OR | isWinning, isDisqualified, wasInRequestedQuality, wasFinishedOnTime, wasForEstimatedValue, isSubcontracted, isConsortium |
| Bodies array | bidders, subcontractors |
| Union | unitPrices, payments |
| Price | price, subcontractedValue |

### Document

| Rule | Fields |
|---|---|
| Modus + Last published value | title, type, signatureDate, version, order, language |
| Last published value | description, format |
| Maximum | publicationDateTime |
| Union | otherVersions, extensions |

## Master rules

### Modus + Last published value

1. Take all values and pick the most frequent.
2. In the case of comparing bodies, two bodies are considered the same if they have identical groupId (ie belonging to the same group of matched bodies)
3. If there are more values of the same frequency then select the latest published

### Last published value

1. Sort all values by publication date
2. Pick the latest published not empty value

### Logical disjunction

This rule makes a logical disjunction and can be applied to fields containing TRUE/FALSE value. It is evaluated in the following steps:

1. if at least one value is TRUE then the master value is TRUE, otherwise
2. if at least one value is FALSE then the master value is FALSE, otherwise
3. the master value is empty

### Longest

This rule selects the longest text value from all considered values

### Maximum

This rule selects the maximum value from all considered values. For example latest date, highest number etc.

### Bodies array

Some variables represent an array of bodies like buyers or bidders. Even the most common case is that all matched arrays from different publications contains only one item the algorithm has to be capable to handle a situation when arrays contain any number of items

- if all arrays contains only 1 body the one with the highest completeness score (described in Body matching chapter) is selected
- if at least one array contains more that one value master value is a union of all published bodies

### Union

This rule is applied to fields that are stored as arrays. A requirement for the application of this rule is that a condition for testing whether two objects equals  is defined for structures stored in an array. If this condition is fulfilled a union of all arrays can be made. This means all published values are present in a master value and each value is present just once.

| Data type | Equality condition |
|---|---|
| Publication | Two publications are considered the same when sourceId, machineReadableUrl, humanReadableUrl, publicationDate and version are equal. Empty value equals whatever |
| URL | Two URLs are equal when the string representation of URLs are the same |

| | |
|---|---|
| Enum | Enumeration value equality |
| String | String value equality |
| Corrections | All corrections are included in a final array of corrections |

### Price

All Price type objects are handled using this rule

- All objects that contains netAmout value are taken into consideration
- For <= 2 prices, use the latest published price
    - If there are two prices without a publication date, use a random value
    - If there is one price object without associated publication date information, pick the one that has publication date information associated as a master value
- For > 2 prices, we find the netAmount MEDIAN (for an even number of prices, the first of the two middle ones is picked).

### Address

The whole address object is selected, individual fields are not merged. For example if there are two matched tender publications and both contain the address of implementation, one of them is picked as a master value. It is the one with the highest scoring where

- NUTS has priority
- otherwise the number of non-empty fields

In case of the same score, the last published address is taken

### Lots

Since each publication can contain multiple lots and each publication can contain a different number of lots (e.g. contract award publication containing information only about awarded lots vs. contract notice announcing all lots) corresponding lots have to be grouped together before variable by variable mastering can start. This chapter describes how lots from matched tenders are grouped together. Each particular field is then mastered using one of the above or below described rules.

- if all the tenders have one lot only, skip the algorithm and put them all into one group
- otherwise calculate the matching ratio MR for each cross tender lot-lot pair:
    - MR = MS / C
        - where
            - MS is the **matching score** - sum of scores from all the comparisons
            - C is the **number of comparisons** - number of comparisons on non-null values (null values are not compared)
        - compare on following attributes:
            - bidsCount (exact match 1, otherwise 0)
            - selectionMethod (exact match 1, otherwise 0)
            - contractSignatureDate (exact match 1, otherwise 0)

        - estimatedPrice.netAmountEur (exact match 1, otherwise 0)
        - main cpv code (exact match 1, otherwise 0)
        - title (exact match 1, otherwise 0)
        - winning bids bidders (match of at least one bidder 1, otherwise 0)
        - contractNumber (exact match 1, otherwise 0)
        - lotNumber (exact match 2, otherwise 0)
        - positionOnPage
            - (1 - ((|lot1.positionOnPage - lot2.positionOnPage|) / (N - 1))) * k
                - **N** is number of lots ( maximum from the matched tenders )
                - **k** is a constant:
                  k = 1 if all the tenders have the same number of lots
                  k = 0.9 otherwise

- sort by MR and match lots with MR >= 0.5 (the higher score wins)
    - groups are created starting from the best match
    - if the next best match creates an invalid group (only one lot from each publication can be present in one group) than it's skipped
    - lots that do not match anything create separate lots

### Bids

Since each lot can contain multiple bids, corresponding bids have to be grouped together before variable by variable mastering can start. This chapter describes how bids from grouped lots are grouped together. Each particular field is then mastered using one of the rules described above or below.

Bids are assigned to groups on a bidder id basis. The logic behind this is that each bidder can participate only in one bid per lot, therefore

- if we find two bids from one lot with the same bidder we consider them the same bid
- if a bid cannot be assigned to any existing group of bids, a new group is created.

### Documents

Before mastering of the document starts all documents from all matched tenders or bids are grouped, each group describing one document of a final tender or bid. Groups of documents are then mastered variable by variable using one of the rules described above or below. The grouping rule is very simple

- all documents with the same URL are considered to be the same document

### CPV

- CPV objects are stored as an array
- A set union of all values is created as a master value
- Two CPVs are equal when their code values are equal
- After a set union there can be more than one CPV marked as main. The following rules are used to set only one main CPV
    - From all CPVs marked as main
        - Pick the most specific one (most digits before first 0 digit)

        - If there are several similarly specific CPVs pick the most recently published one
        - If there are more CPVs of the same age pick a random one

### Fundings

- Fundings objects are stored as an array.
- A set union of all values is created as a master value
- Two values are considered equal for set union calculation when *source* and *isEuFunded* variables have the same content.
- If two or more same funding objects are detected then the one with more non empty values is inserted into a final set.

### Award criteria

Award criteria makes sense when the weight of criteria is 100% in total. Criteria from different tender publications are not combined.

- The latest published combination of criteria which has the sum of weights 100% is selected as a master value.
- If there is no such published combination of award criteria the one with the highest sum of weights is selected
- If two or more combinations have the same sum of weights, a random one is picked

### Body IDs

The Body ID consists of three fields *id*, *type*, *scope.* Multiple body IDs of a same type and scope are not desired in a master body object. If several different IDs of the same type and scope appears

- the one that comes from the company DB record is preferred
- if no ID comes from the company DB record, the most frequent value is preferred
- if there is no most frequent value, the most recently published value is selected.

# Master data postprocessing

## Currency conversion

All prices in the DIGIWHIST data model are being converted into both national currencies (those coming from national portals) and EUR. At the end of data processing, where possible, each price contains three values

- netAmount
- netAmountEur
- netAmountNational

The date that determines an exchange rate is selected by application of following rules. The first applicable rule determines the date

    - minimum publication date of all processed CONTRACT_AWARDs
    - minimum publication date of all processed CONTRACT_NOTICEs
    - minimum publication date of any publication

- if the date that can be used to select a currency exchange rate from exchange rate table cannot be determined then currency conversion cannot continue and proper fields will be empty in a final database
- otherwise if *netAmount* and *currency* is set an exchange rate table is used to calculate *netAmountEur*
- for each national source also *netAmountNational* is calculated if *netAmount* and *currency* are known
    - national sources has *currencyNational* always a local currency (Czech Republic = CZK, Slovakia = EUR, Poland = PLN)
    - for TED *currencyNational* is EUR
- If the exchange rate table does not contain the required value for currency conversion it cannot be performed and proper fields will be empty in a final database

### Contract implementations

Tender publications of the type *contract implementation* provide information on the actual fulfillment of the contract. During the mastering process these tender publications serve to adjust payments, therefore, a different strategy is employed.

- First, the master object is created based on all other tender publications except the contract implementations
- A union of *tender.payments* per bidder from contract implementation publications is created. Duplicate payments (same date and price) are removed.
- Payments are added to corresponding bidders in master object (again with removal of duplicates)
- If there is only one lot and one winning bidder in a master object, payments are added regardless of any bidder conflicts

## Indicators

Each tender has a set of indicators associated. These indicators were designed and tested within WP3 works. Each indicator says that a given fact is either true (1) for a tender, is false (0) or if the indicator does not exist then particular tender does not provide sufficient information and specific fact cannot be confirmed or disproved.

### Single bidder contract (valid/received)

Single bid signals a risk when only one bid is submitted in a tender in a competitive market.

#### Calculation

- If *lot.validBidsCount* is not empty and
    - *lot.validBidsCount* = 1
        - Create indicator and set its value to 1
    - *lot.validBidsCount* > 1
        - Create indicator and set its value to 0
- If *lot.validBidsCount* is empty but *lot.bidsCount* is not empty and
    - *lot.bidsCount* = 1
        - Create indicator and set its value to 1

        - *lot.bidsCount* > 1
            - Create indicator and set its value to 0
- Otherwise don’t create an indicator

## New company

New company signals the risk of a very young company winning a tender (younger than 1 year at the time of winning).

### Calculation

- calculate *date of contract award* as the publication date of the first (oldest) publication of *formType* CONTRACT_AWARD
- for each winning bid check following:
    - for each bidder take a *company foundation date* from BvD company DB:
    - if *date of contract award* - *company foundation date* < 365 days for at least one bidder
        - Create an indicator and set its value to 1
        - Create an indicator just once even when there are more bidders fulfilling such condition
        - Store the relevant bidder ID in the indicator metadata
    - if *date of contract award* - *company foundation date* > 365 days for all bidders
        - Create indicator and set its value to 0
    - if at least one bidder is *missing company foundation date* and if *date of contract award* - *company foundation date* > 365 days for all remaining bidders
        - don’t create an indicator

## Joint of centralized procurement

Centralized procurement suggests good administrative capacity if the tender is managed by a central procuring body.

### Calculation

- if *tender.isCentralProcurement* = true
    - Create indicator and set its value to 1
- if *tender.isCentralProcurement* = false
    - Create indicator and set its value to 0
- Otherwise don’t create an indicator

## Length of advertisement period

Advertisement period length reveals the risk of suspiciously tight bidding deadlines or when advertisement period is excessively long.

### Calculation

- if there is no *tender.bidDeadline* specified

        - search in the table below to find whether it's an indicator. If *Yes* create an indicator and set its value to 1
- if there is *tender.bidDeadline* set
    - calculate *date of contract notice* as a *publication date* of the first (oldest) publication of *formType* CONTRACT_NOTICE
        - if there is no such publication don't create an indicator
    - calculate *advertisement period length* in days as *tender.bidDeadline - date of contract notice*
    - if the result is a non-negative value in the table below search for a given country and day range. If the advertisement period length fits the range, create an indicator and set its value to 1
    - if the result is a non-negative value in the table below search for a given country and day range. If the advertisement period length is out of the range, create an indicator and set its value to 0
    - if the result is a negative value indicator don't create an indicator

| **Country** | **Missing bidding deadline is a risk factor** | **Indicator** |
|---|---|---|
| AT | No | 0-20;34-47 |
| BE | Yes | 18-34;78-1095 |
| BG | No | 0-28;35-1095 |
| CY | No | 0-46;53-60 |
| CZ | No | 0-50 |
| DE | No | |
| DK | No | 52-61 |
| EE | No | 0-32;50-57 |
| ES | No | 39-42;52-1095 |
| FI | No | 0-39;52-1095 |
| FR | No | 0-40 |
| GR | No | 0-54 |
| HR | No | 0-40;49-1095 |
| HU | No | |
| IE | No | 41-1095 |
| IT | No | 0-47 |
| LT | No | 40-42;48-1095 |
| LU | No | 51-54;86-1095 |

| | | |
|---|---|---|
| LV | No | 0-40;51-57 |
| NL | No | 0-38;48-56 |
| NO | No | 36-42;50-56 |
| PL | No | 0-25;43-1095 |
| PT | No | 0-42 |
| RO | No | 41-50 |
| SE | No | |
| SI | No | 51-1095 |
| SK | No | 49-52 |
| UK | No | 0-53 |

## Length of decision period

Length of decision period signals risks when the the decision period length is either suspiciously short or suspiciously long.

### Calculation

- if there is no *tender.bidDeadline* specified
    - search in the table below, to see whether it's an indicator. If *Yes* create an indicator and set its value to 1
- if there is *tender.bidDeadline* set
    - iterate over tender.lots and calculate *award decision date* as a first (oldest) *tender.lot[i].awardDecisionDate* as award_decision_date
        - if there is no date at all don’t create an indicator
    - calculate *decision period length* in days as *award decision date* - *tender.bidDeadline*
    - if the result is a non-negative value in the table below search for a given country and day range, if the decision period length fits the range, create an indicator and set its value to 1
    - if the result is a non-negative value in the table below search for a given country and day range. If decision period length is out of the range, create an indicator and set its value to 0
    - if the result is a negative value don’t create an indicator

| **Country** | **Missing bidding deadline is a risk factor** | **Red flag** |
|---|---|---|
| AT | Yes | 0-56 |
| BE | No | 0-22 |
| BG | No | 0-27;120-1095 |

| | | |
|---|---|---|
| CY | No | 0-90 |
| CZ | No | 0-147 |
| DE | Yes | 0-36 |
| DK | No | 0-39;124-168 |
| EE | Yes | 0-41 |
| ES | No | 0-43 |
| FI | No | 0-65;92-127 |
| FR | No | 0-66;156-1095 |
| GR | No | 0-170 |
| HR | No | 0-26 |
| HU | No | 0-46;73-104 |
| IE | No | 0-50;87-1095 |
| IT | No | 0-200 |
| LT | No | 0-32 |
| LU | No | 0-52 |
| LV | No | 0-20;106-1095 |
| NL | No | 0-34;58- |
| NO | No | 0-70;98-229 |
| PL | Yes | 0-63 |
| PT | No | 0-63;243-1095 |
| RO | Yes | 0-56 |
| SE | No | 0-44;89-1095 |
| SI | No | 0-51;77-1095 |
| SK | No | 0-68 |
| UK | No | 0-35;165-304 |

## Use of WTO framework

Use of the WTO framework suggests good administrative capacity if the tendering process is conducted according to the WTO framework.

### Calculation

- if *tender.isCoveredByGpa* = true or if at least one *tender.lot[i].isCoveredByGPA* = true
    - Create indicator and set its value to 1
- if *tender.isCoveredByGpa* = false or for all *tender.lot[i].isCoveredByGPA* = false
    - Create indicator and set its value to 0
- Otherwise don't create an indicator

## Use of framework agreements

Framework agreement suggests good administrative capacity if the tender establishes a framework agreement.

### Calculation

- if *tender.isFrameworkAgreement* = true or if at least one *tender.lot[i].isFrameworkAgreement* = true
    - Create indicator and set its value to 1
- if *tender.isFrameworkAgreement* = false or for all *tender.lot[i].isFrameworkAgreement* = false
    - Create indicator and set its value to 0
- Otherwise don't create an indicator

## Electronic auction

Electronic auction points at good administrative capacity if the tender was conducted through an electronic auction.

### Calculation

- if *tender.isElectronicAuction* = true or if at least one *tender.lot[i].isElectronicAuction* = true
    - Create indicator and set its value to 1
- if *tender.isElectronicAuction* = false or for all *tender.lot[i].isElectronicAuction* = false
    - Create indicator and set its value to 0
- Otherwise don't create an indicator

## Call for tenders publication

Not publishing calls for tender signals a risk when no call for tenders is published prior to a contract award, decreasing the potential bidder pool.

### Calculation

- iterate over all publications
- if there is no publication of formType = PRIOR_INFORMATION_NOTICE or CONTRACT_NOTICE and the list of countries below says it’s a risk factor, create an indicator and set its value to 1
- if there is no publication of formType = PRIOR_INFORMATION_NOTICE or CONTRACT_NOTICE and the list of countries below says it’s not a risk factor, create an indicator and set its value to 0
- if there is a publication of formType = PRIOR_INFORMATION_NOTICE or CONTRACT_NOTICE, create an indicator and set its value to 0
- if decision cannot be made don’t create an indicator

| **Country** | **NO Call for Tenders publication is a risk factor** |
|---|---|
| AT | Yes |
| BE | Yes |
| BG | No |
| CY | Yes |
| CZ | Yes |
| DE | Yes |
| DK | No |
| EE | No |
| ES | No |
| FI | Yes |
| FR | Yes |
| GR | Yes |
| HR | Yes |
| HU | Yes |
| IE | Yes |
| IT | Yes |
| LT | No |
| LU | Yes |
| LV | Yes |
| NL | Yes |

| | |
|---|---|
| NO | Yes |
| PL | Yes |
| PT | Yes |
| RO | Yes |
| SE | Yes |
| SI | Yes |
| SK | Yes |
| UK | Yes |
| MT | Yes |

## Tax haven

Tax haven signals a risk when the supplier is located in a tax haven country (based on the financial secrecy index).

### Calculation

- for each bidder of each winning bid
    - calculate *winner country* as *bidder.address.country*
    - calculate *award publication year* as a year value from *tender.publications[i].publicationDate* where *tender.publications[i].formType* = CONTRACT_AWARD
        - if there are multiple CONTRACT_AWARD publications use the oldest
    - Search FSI(Financial Secrecy Index) table for combination of *winner country* (column A) and *award publication year* (column D)
        - https://github.com/digiwhist/wp2_documents/blob/master/FSI_scores.xls
        - if this combination determines *Yes* value, create an indicator and set its value to 1
        - if this combination determines *No* value, create an indicator and set its value to 0
        - if a decision cannot be made don’t create an indicator
- create an indicator just once even when there are more bidders fulfilling such condition
- store the relevant bidder ID in the indicator metadata

## English as foreign language

English as a foreign language suggests good administrative capacity if the bids can be submitted in English as a foreign language.

### Calculation

- if English is in *tender.eligibleBidLanguages* and *tender.country* is not the UK or IE
    - create an indicator and set its value to 1
- if English is in *tender.eligibleBidLanguages* and *tender.country* is the UK or IE
    - Don't create an indicator
- if English is not in *tender.eligibleBidLanguages* and *tender.country* is not the UK or IE
    - create an indicator and set its value to 0

## Procedure type

Non-open procedures signal a risk of using procedures types which are less open for competition and more readily used for directly contracting connected companies (e.g. negotiated without publication).

### Calculation

Use *tender.country* and *tender.procedureType* values to search for this combination in a non-open procedures matrix

- if procedure type is missing and a column *missing* contains the value *Yes* for a particular country
    - Create an indicator and set its value to 1
- if the combination says *Yes*
    - Create an indicator and set its value to 1
- If the combination say *No*
    - Create an indicator and set its value to 0
- if a decision cannot be made don't create an indicator

| DIGIWHIST enumeration | NEGOTIATED_WITH_PUBLICATION | RESTRICTED | OUTRIGHT_AWARD | COMPETITIVE_DIALOG | NEGOTIATED_WITH_PUBLICATION | NEGOTIATED_WITHOUT_PUBLICATION | OPEN | RESTRICTED | |
|---|---|---|---|---|---|---|---|---|---|
| Country | Accelerated negotiated | Accelerated restricted | Award without publication | Competitive dialogue | Negotiated with competition | Negotiated without competition | Open | Restricted | Missing/error |
| AT | Yes | No | Yes | Yes | No | Yes | No | No | No |
| BE | Yes | No | Yes | No | Yes | Yes | No | No | No |
| BG | No | No | Yes | No | Yes | Yes | No | Yes | No |
| CY | No | No | No | No | No | No | No | No | No |
| CZ | Yes | No | Yes | Yes | No | Yes | No | No | No |
| DE | Yes | Yes | Yes | No | Yes | Yes | No | No | No |
| DK | No | No | Yes | Yes | Yes | Yes | No | No | No |
| EE | Yes | No | Yes | Yes | No | Yes | No | No | Yes |
| ES | Yes | No | Yes | No | Yes | Yes | No | No | No |
| FI | No | Yes | Yes | No | Yes | Yes | No | No | No |
| FR | Yes | Yes | No | No | Yes | Yes | No | No | No |
| GR | No | No | No | No | No | No | No | No | No |
| HR | No | No | No | No | No | No | No | Yes | No |
| HU | Yes | Yes | No | No | Yes | Yes | No | No | No |
| IE | No | No | No | Yes | Yes | No | No | No | No |
| IT | Yes | No | Yes | No | No | Yes | No | Yes | No |
| LT | No | No | No | No | Yes | Yes | No | No | No |
| LU | No | No | No | No | No | No | No | No | No |
| LV | No | No | Yes | No | Yes | Yes | No | No | No |
| NL | Yes | No | Yes | No | No | Yes | No | No | No |
| NO | Yes | No | Yes | No | Yes | Yes | No | No | No |
| PL | Yes | No | Yes | No | Yes | Yes | No | Yes | No |
| PT | No | No | Yes | No | No | Yes | No | No | Yes |
| RO | Yes | Yes | No | Yes | No | Yes | No | No | No |
| SE | No | No | No | No | No | Yes | No | No | No |
| SI | Yes | No | Yes | Yes | No | Yes | No | No | No |
| SK | Yes | Yes | No | No | Yes | Yes | No | No | No |
| UK | No | Yes | Yes | No | Yes | Yes | No | No | No |

## Number of key missing fields in form

Number of key missing fields is the ratio of fields with missing values in the call for tender and contract award announcement related to a given tender (based on the most recently published versions).

### Calculation

For the most recently published publication of type CONTRACT_NOTICE and most recently published publication of type CONTRACT_AWARD

- calculate a ratio of key missing fields as a number of fields that has an empty value set divided by a number of tested fields
- Value of an indicator is not 1/0 in this case but the ratio itself
- Tested field are
    - tender.addressOfImplementation.nuts
        - if form type is CONTRACT_AWARD
    - tender.awardCriteria.name
        - if tender.selectionMethod=MEAT and form type is CONTRACT_NOTICE
    - tender.awardCriteria.weight
        - if tender.selectionMethod=MEAT and form type is CONTRACT_NOTICE
    - tender.lot.bid.bidder.name
        - if tender.lot.bid.isWinning=TRUE and form type is CONTRACT_AWARD
    - tender.lot.bid.price.netAmount
        - if tender.lot.bid.isWinning=TRUE and form type is CONTRACT_AWARD
    - Only one of these fields will be given so it’s only missing if all is missing in publication of form type CONTRACT_NOTICE
        - tender.lot.estimatedStartDate
        - tender.lot.estimatedCompletionDate
        - tender.lot.estimatedDurationInMonths
        - tender.lot.estimatedDurationInDays
    - tender.eligibleBidLanguages
        - if form type is CONTRACT_NOTICE
    - tender.selectionMethod or lot.selectionMethod
        - if form type is CONTRACT_NOTICE
    - tender.funding.isEuFund
        - if form type is CONTRACT_AWARD
    - tender.CPVs.code
        - if form type is CONTRACT_NOTICE
    - tender.lot.bid.isSubcontracted
        - if form type is CONTRACT_AWARD

## Discrepancies between call for tender and award

Discrepancies between call for tender and contract award notices is the ratio of fields with different values in the call for tender and contract award announcements related to a given tender (based on most recently published versions).

### Calculation

For the most recently published publication of type CONTRACT_NOTICE and most recently published publication of type CONTRACT_AWARD compare corresponding values from contract notice and contract award

- Calculate a ratio as a number of fields with different values divided by number of compared fields
- Value of an indicator is not 1/0 in this case but the ratio itself
- If there is no comparable variable then no indicator is created
- Compared fields are
    - tender.buyer.address.street
    - tender.buyer.address.postcode
    - tender.addressOfImplementation.nuts
    - tender.awardCriteria.name
    - tender.awardCriteria.weight
    - tender.lot.lotNumber
    - tender.title
    - tender.isCoveredByGPA
    - tender.selectionMethod
    - tender.isElectronicAuction
    - tender.funding.isEuFund
    - tender.isFrameworkAgreement
    - tender.isDPS

## Political connections of suppliers

Political connection of the supplier captures the risk of at least one owner or manager holding a political office.

### Calculation

- If a link from a winner of a tender to a record from the public officials database described in chapter Company data exists, create an indicator and set its value to 1
- If there is no link don’t create an indicator

## Publication rate

Publication rate is the ratio of advertised public procurement spending over the total public procurement spending of a given contracting authority.

### Calculation

Percent of advertised public procurement spending over total public procurement spending calculated as

- For each tender published by a given contracting authority which is not not published on behalf of some other contracting authotiry
    - Take a *tender.finalPrice* or *bid.price* of winning bid if final price is not in a data
    - Add this value to a total sum of advertised public procurement spendings
- Divide advertised public procurement spending by a budgeted amount of the same contracting authority (public procurement spending estimate from budget data is calculated following OECD/Eurostat methodology)
- This indicator is evaluated on a yearly basis

## Description length

Product description length signals the risk of product description being tailored to one company, when the description is excessively long.

### Calculation

Calculation of this indicator is based on aggregated information from all tenders, therefore, it's not a part of the published database. This indicator will be present on opentender.eu portal as a result of an aggregation function and will be calculated as

- 0=description length is not in the top 5% of the market
- 1=description length is in the top 5% of the market
- See D3.6 for a detailed formula

## Eligibility criteria length

Eligibility criteria length signals the risk of criteria being tailored to one company with the criteria description being excessively long.

### Calculation

Calculation of this indicator is based on aggregated information from all tenders, therefore, it's not a part of the published database. This indicator will be present on opentender.eu portal as a result of an aggregation function and will be calculated as

- 0=eligibility criteria length is not in the top 5% of the market
- 1=eligibility criteria length is in the top 5% of the market
- See D3.6 for a detailed formula

## Evaluation criteria

Evaluation criteria signal the risk of using non-quantitative criteria to assess bidders which are subjective and can easily be manipulated.

### Calculation

Calculation of this indicator is based on aggregated information from all tenders, therefore, it's not a part of the published database. This indicator will be present on opentender.eu portal as a result of an aggregation function and will be calculated as

- 0=non-price related evaluation criteria are not in the top 5% of the market distribution and lowest price criteria are also unrelated to corruption risks
- 1=non-price related evaluation criteria are in the top 5% of the market distribution OR the lowest price criteria are unrelated to corruption risks
- See D3.6 for a detailed formula

### Winner contract share

Winner contract share is the share of contract value won by a given company from a given buyer in a given year.

#### Calculation

Calculation of this indicator is based on aggregated information from all tenders, therefore, it's not a part of published database. This indicator will be present on opentender.eu portal as a result of an aggregation function and will be calculated as

- The total share of contracts won by the winner company from the contracting authority per year
- See D3.6 for a detailed formula

## OCDS conversion

All processes described in this paper work with a DIGIWHIST data  standard. To make the data as understandable and interoperable as possible it is also published in OCDS data format. The conversion from a DIGIWHIST data standard to OCDS is described using official field-level mapping template published by OCP. This mapping can be found online - https://github.com/digiwhist/wp2_documents/blob/master/digiwhist_ocds.xlsx

# Annex 1 - Future data releases

In order to further fine-tune the database following user feedback from stakeholder workshops and various dissemination events, the consortium will release further versions of the database containing higher quality data. On top of these refinements, a number of key extensions were also identified which go beyond the mere fulfillment of the Grant Agreement by adding high public value to the core database. These extensions will also be part of later data releases. This chapter shortly describes some of these key improvements which are planned for implementation during autumn 2017 up until February 2018. Updated version of this document can be found online on www.

## Data cleaning

### Crazy values elimination

Some source data contains obviously crazy values like septilions or dates from the Middle Ages. These should be removed from a final dataset.

### Completion of price object

If the price object does not contain VAT, a standard VAT rate for given country from https://ec.europa.eu/taxation_customs/sites/taxation/files/resources/documents/taxation/vat/how_vat_works/rates/vat_rates_en.xls (tab Evolution of VAT rates) is taken

### Postcode to NUTS conversion

Conversion files from postcode to NUTS code are available for European countries on http://ec.europa.eu/eurostat/tercet/flatfiles.do . Final data will be enriched by NUTS codes where a postcode is available.

## Mastering matched data

### Address rule

This rule will change to take an age of address into consideration so that a complete but very old address is not published for a body.

### Size

A new rule for a size calculation will be introduced. Currently the size is published only if the information is present in a source. The information can be also inferred from the price of whole tender, supply type, contract notice or contract award publication date and buyer type. Basic logic is that for some combination of parameters, price thresholds are declared by the EC every two years. Based on those thresholds a decision can be made whether the tender is above or below threshold.

## Contract updates

A new form F20 - modification notice[10] was introduced in directive 2014/23/EU. This form allows information in previously published notices to be updated by specifying a section of the original notice, original value and replacement value. These updates will be applied to master tenders.

[10] http://simap.ted.europa.eu/documents/10184/99173/EN_F20.pdf