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

EU Grant Agreement number: 645852

Project acronym: DIGIWHIST

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

Work Package: 2 - Data Collection and Cleaning

Title of deliverable: D2.6 Final Linked Database and related alghorithms

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,
Tomáš Mrázek, Marek Mikeš, (UCAM),
Michal Říha 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)	

All rights reserved. This document has been published thanks to the support of the European Union's Horizon 2020 research and innovation Programme under grant agreement No 645852.

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

Introduction

The purpose of deliverable D2.6 is to publish source codes of the whole DIGIWHIST data processing system and final DIGIWHIST database which is the result of processing:

- 25 public procurement data sources
    - TED + TED archive
    - Current procurement portal + archive for CZ, UK, HU
    - One source for SK, PL, ES, NL, FR, LV, PT, EE, GE, SI, IE, NO, CH, LT, HR, BG, RO
- 4 public officials data sources
    - http://everypolitician.org/
    - http://www.politicaldatayearbook.com/
    - http://rulers.org/
    - https://www.cia.gov/library/publications/world-leaders-1/index.html
- company database
- 3 budget data sources
    - UK, ES, CZ

The key component of the whole process is public procurement data crawling, structuring, formatting, linking and merging of linked records, covering 35 jurisdictions. It also includes integration with the above mentioned databases like company database, public officials database and budget database. This integration is represented in our final database by several tender related indicators like Tax haven indicator, Political connections indicator or Publication rate indicator.

Methodologically the process is described in other deliverables of WP2 of the DIGIWHIST project.

Data

Data are available for bulk download as archives in two data standards: the original DIGIWHIST data standard (DDS) and also the OCDS (Open Contracting Data Standard). Each archive contains one file that consists of all tender records for one country. Each line in a file represents one tender record and is in a valid JSON format. Each file might contain data for a country from its procurement portal and from TED. This means there can be duplicated tenders, each one based on the data published on a different source.

- If the tender is based on TED data the field *createdby* in DDS has a value (name of a programme which created the record)
    - *eu.digiwhist.worker.eu.master.TedTenderMaster*
- If the tender is based on data from a national procurement portal then the field *createdby* has a source specific value
    - *eu.digiwhist.worker.<country ISO2 code>.master.<programme name>*

The structure of the DDS data is described in the Apiary public project that can be found at http://docs.digiwhist.apiary.io . This documentation describes an API that is not public and serves only for internal DIGIWHIST project purposes (for example opentender.eu portal), but the structure of the exported data is identical to the one described there because it was used to export data.

Data in OCDS format are valid against OCDS 1.1 schema and make use of the Lots, Bids and Requirements extensions. - http://standard.open-contracting.org/latest/en/
Speaking in OCDS terms each file consists of multiple JSON documents (one document per line). Each document represents one release package containing one compiled release.

The following table contains links to archives containing data for a particular jurisdiction in the DIGIWHIST data standard.

Poland	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/PL_data.json.tar.gz
France	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/FR_data.json.tar.gz
Portugal	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/PT_data.json.tar.gz
Spain	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/ES_data.json.tar.gz
Czech Republic	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/CZ_data.json.tar.gz
Germany	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/DE_data.json.tar.gz
Hungary	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/HU_data.json.tar.gz
Norway	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/NO_data.json.tar.gz
Georgia	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/GE_data.json.tar.gz
Estonia	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/EE_data.json.tar.gz
United Kingdom	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/UK_data.json.tar.gz
Latvia	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/LV_data.json.tar.gz
Slovakia	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/SK_data.json.tar.gz
Netherlands	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/NL_data.json.tar.gz
Italy	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/IT_data.json.tar.gz
Sweden	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/SE_data.json.tar.gz
Ireland	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/IE_data.json.tar.gz
Belgium	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/BE_data.json.tar.gz
Romania	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/RO_data.json.tar.gz
Bulgaria	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/BG_data.json.tar.gz
Finland	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/FI_data.json.tar.gz
Austria	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/AT_data.json.tar.gz
Switzerland	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/CH_data.json.tar.gz
Denmark	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/DK_data.json.tar.gz
Greece	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/GR_data.json.tar.gz
Lithuania	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/LT_data.json.tar.gz
Slovenia	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/SI_data.json.tar.gz
Croatia	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/HR_data.json.tar.gz
Luxembourg	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/LU_data.json.tar.gz
Cyprus	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/CY_data.json.tar.gz
Malta	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/MT_data.json.tar.gz
Iceland	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/IS_data.json.tar.gz
Serbia	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/RS_data.json.tar.gz
Armenia	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/AM_data.json.tar.gz

The following table contains links to archives containing data for a particular jurisdiction in OCDS.

Poland	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/PL_ocds_data.json.tar.gz
France	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/FR_ocds_data.json.tar.gz
Portugal	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/PT_ocds_data.json.tar.gz
Spain	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/ES_ocds_data.json.tar.gz
Czech Republic	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/CZ_ocds_data.json.tar.gz
Germany	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/DE_ocds_data.json.tar.gz
Hungary	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/HU_ocds_data.json.tar.gz
Norway	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/NO_ocds_data.json.tar.gz
Georgia	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/GE_ocds_data.json.tar.gz
Estonia	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/EE_ocds_data.json.tar.gz
United Kingdom	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/UK_ocds_data.json.tar.gz
Latvia	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/LV_ocds_data.json.tar.gz
Slovakia	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/SK_ocds_data.json.tar.gz
Netherlands	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/NL_ocds_data.json.tar.gz
Italy	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/IT_ocds_data.json.tar.gz
Sweden	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/SE_ocds_data.json.tar.gz
Ireland	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/IE_ocds_data.json.tar.gz
Belgium	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/BE_ocds_data.json.tar.gz
Romania	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/RO_ocds_data.json.tar.gz
Bulgaria	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/BG_ocds_data.json.tar.gz
Finland	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/FI_ocds_data.json.tar.gz
Austria	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/AT_ocds_data.json.tar.gz
Switzerland	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/CH_ocds_data.json.tar.gz
Denmark	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/DK_ocds_data.json.tar.gz
Greece	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/GR_ocds_data.json.tar.gz
Lithuania	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/LT_ocds_data.json.tar.gz
Slovenia	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/SI_ocds_data.json.tar.gz
Croatia	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/HR_ocds_data.json.tar.gz
Luxembourg	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/LU_ocds_data.json.tar.gz
Cyprus	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/CY_ocds_data.json.tar.gz
Malta	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/MT_ocds_data.json.tar.gz
Iceland	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/IS_ocds_data.json.tar.gz
Serbia	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/RS_ocds_data.json.tar.gz
Armenia	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/AM_ocds_data.json.tar.gz

Source codes

- All source codes, including DB creation scripts, are publicly available on a GitHub
    - https://github.com/digiwhist/backend
- The technological stack needed to properly run the whole system is
    - Postgresql 9.4 and higher
    - RabbitMQ 3.6
    - Java 8
    - Maven
- Company DB is not available because it was purchased only for the purposes of the DIGIWHIST project and is not licensed for further data publication. Indicators derived using company data are part of the DIGIWHIST data for example *Tax haven* or *New company* indicators.