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 Grant Agreement: 28/02/2017

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

Actual submission date: 28/09/2017

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)	

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. 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* 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 JSON 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.

The following table contains links to archives containing data for a particular jurisdiction. Each archive is encrypted by the standard encryption software GnuPG. To decrypt archives on Windows,

the installation of software like gpg4win (https://www.gpg4win.org/download.html) is required. Linux distributions and Mac operating systems usually have this software built-in. After the installation of gpg4win, the decrypt function is available in a context menu. Decryption also requires a password which we provide to EC members on demand by emailing digiwhist.aws@gmail.com .

In compliance with the Description of the Action , the public release of the final database will only take place once the validation tests are positive and the consortium is confident that the highest possible data quality was achieved. Based on results of validation process this is expected to happen at the end of November 2017. We will widely advertise the public release on social media and at our dissemination workshops and other events.

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

Malta	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/MT_data.json.tar.gz.gpg
Iceland	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/IS_data.json.tar.gz.gpg
Serbia	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/RS_data.json.tar.gz.gpg
Armenia	https://s3.eu-central-1.amazonaws.com/digiwhist-data/D2_6/tender/AM_data.json.tar.gz.gpg

Source codes

- All source codes including DB creation scripts are publicly available on a GitHub
    - https://github.com/digiwhist/backend
- RabbitMQ messaging system is needed to properly run the whole infrastructure
- 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 company tax haven registration.