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: D2.2 Data Management Plan

Due date of deliverable: revised date 31/05/2015

Re-submission date after grant amendment:

06/02/2018 (revised version)

Author: Jan Hruby (UCAM)

Editors: Mihály Fazekas and Fiona Harrison

Organization name of lead beneficiary for this deliverable: University of Cambridge

Dissemination Level		
PU	Public	X
PP	Restricted to other programme participants (including the Commission Services)	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
Co	Confidential, only for members of the consortium (including the Commission Services)	

1. Data summary

Introduction

A central task of DIGIWHIST is to collect publicly available data about public procurement across Europe. The collection of such data and other complementary datasets is the main goal of WP2. This document is a data management plan that describes, in compliance with the guidelines on H2020 FAIR Data Management, how data will be collected, generated, maintained, secured and published.

Purpose of data collection and its relation to the objectives of the project

Procurement data

Collecting public procurement data is a fundamental part of the DIGIWHIST project. This type of data is spread across the internet on national procurement portals, municipalities' websites or NGO websites. The diversity within the sources with regard to the variety of technical sources, data completeness and data openness make the widescale collection and processing of such data, as originally envisaged in DIGIWHIST, very challenging. One of the most important goals of DIGIWHIST is to design and implement a solution that will make such data publicly available and machine readable for everyone. On top of that this data will be used by DIGIWHIST partners for scientific purposes such as computing of corruption risk indicators. This requires a well-described methodology of data collection, processing and validation.

Public sector data

To achieve all the goals of the DIGIWHIST project we also collect and create four other datasets:

- Contracting authorities register
- Public officials register
- Budget data
- Asset declaration data
-

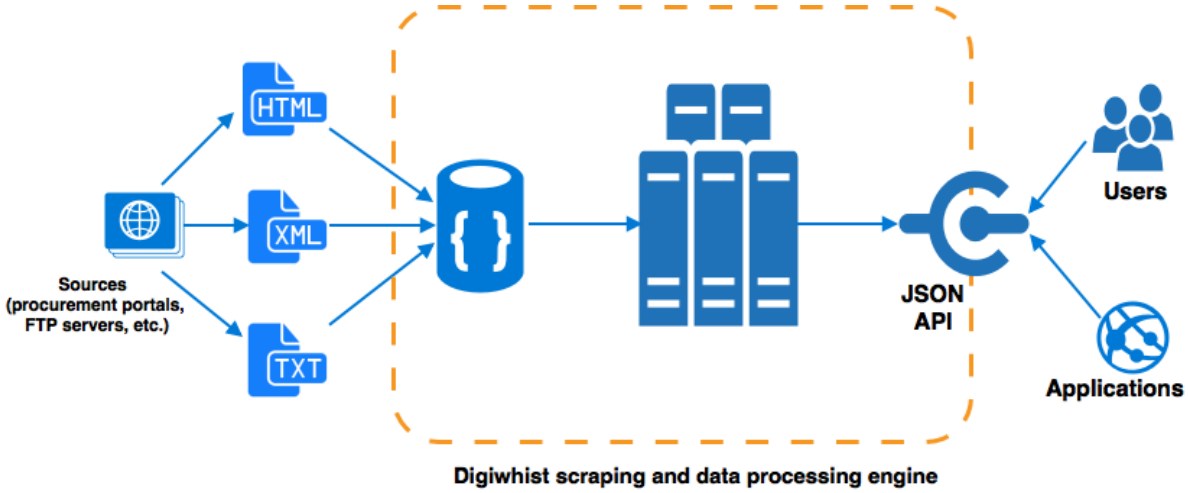
These are complementary datasets that won't be published separately as a stand-alone datasets but will be integrated into one joined database together with procurement and company data which will serve as an analytical platform for research purposes in harmony with activities described in the DoA (description of activities) particularly indicators development (WP3)

Data formats

The current state of data sharing is at a different level in each of the 35 jurisdictions covered by the project and for all data sets that are collected, meaning that it is available in a variety of formats; structured, semi-structured and unstructured. The majority of data sources provide data in a human readable format as a semi-structured HTML page. The remaining sources are either unstructured text sources or structured XML, JSON or CSV files. Some of these can be accessed programmatically via API but even structured data is mainly accessed via a direct link from HTML pages.

Despite the fact that the various sources contain data of varying quality and our options for data extraction from structured/unstructured sources are different, DIGIWHIST will endeavour to unify

data from all sources and provide it to all users in a structured machine readable format commonly used for data exchange, such as XML or JSON via API



Origin of the data and re-use of existing datasets

WP2 uses the outputs of WP1, namely D1.1 (Repository of online data sources and description of data content), as a fundamental input for data collection

Procurement portals

Most of the public procurement data is scraped from its original source which is usually a web portal where data is published in a human readable format but with no support for machine processing. The DIGIWHIST project downloads and processes such data to make it structured, links all publications together and creates outputs that aggregate more publications of each tender into one master object that is published as a final description of a tender.

Open data

Among the national portals we can find some open data portals that support machine processing of data by publishing it in a proper format (e.g XML). For example:

- ftp://ted.europa.eu/
- ftp://ftp.uzp.gov.pl/bzp/xml/
- ftp://open.iub.gov.lv/

The lack of openness is caused by not using the OCDS for publishing procurement data. Our sources use data structures which make the data less open; our task is to transform the data so it can be used in an open structure. In some cases, there are well described methodologies, such as for TED, that help to transform data into the desired format. For the remaining sources we have to use expert knowledge to map the original source data to the DIGIWHIST data template.

Company data

In order to gain powerful insights into the workings of public procurement markets, information on the participating companies must also be available. Consequently, in addition to the collection, scraping, parsing and organization of public procurement data by DIGIWHIST, company data must also be obtained. There are four relevant variable groups to be obtained: company registry

information (company name, ID, incorporation date, address, company size etc.), financial data (annual turnover, profit rate, liabilities etc.), ownership and manager information. However, in most European countries, there is no readily available and detailed company data. Although, national company registries exist, they are not always free to use and often only contain a limited set of information (e.g. no ownership or financial data is available). Furthermore, open data repositories on company characteristics do not contain enough data either (e.g. opencorporates.com), hence they can be only used for cross-checking data quality. Therefore, under the terms of project proposal, the full company data set from all 34 countries covered was purchased from a private data provider.

Public sector data

In a same way as for procurement data we use public sector data published either on national portals or on NGO portals that mainly provides data in a machine readable format. There is a big difference between quality of procurement and public sector data sources. While governments put an effort into publishing procurement data quality of public sector data like budget information or asset declarations is very poor. This data is mainly completely unstructured, not easily machine processable (scanned pdfs) and scattered across the internet.

Size of the data

Public procurement data is stored in several stages within a database

1. *Raw data*

This comprises data as it is published on its original source. In this stage we basically create a mirror of the original source so that we can access this data without needing to request it again from its original location and without any information loss. Raw data therefore contains a mixture of HTML, XML, JSON or CSV data including all the unnecessary information that accompanies the required information.

We've already collected raw data from almost all jurisdictions and therefore don't expect that the raw data size will grow dramatically although it's highly probable that after the first round of validation we will find out that some publications are missing and a crawler adjustment will be needed.

We will also collect data increments so we expect the data size will increase in the coming years. Since we have precise data for the month of May 2016 we'll be able to estimate the size of an increment in the near future.

Stage	Number of records*	Data size**	Estimated data size (GB)***
Raw	9142602	318	350

* Number of records August 2016
** Data size in GB in August 2016
*** Estimated data size in September 2017

2. Parsed data

This database contains useful information extracted from the raw documents in a structured text format. One raw document can be split into multiple parsed documents, each describing one tender. We don't parse all documents, only selected forms that contain information relevant to the project's goals; therefore, the number of parsed documents may be lower than the number of raw documents.

We have only datasets from a few jurisdictions processed to the parsed stage. This comprises about 25% of all raw publications; our estimate of the final data size is therefore based on the current size and our prediction that final size will be four times bigger

Stage	Number of records*	Data size**	Estimated data size***
Parsed	2552248	13	52

* Number of records August 2016
** Data size in GB August 2016
*** Estimated data size in GB September 2017

3. Clean data

At this stage we convert the structured text information to a proper data type e.g. numbers, dates, enumeration values. This stage contains the same number of documents as a parsed stage but may contain a different number of fields from the corresponding parsed document because:

- the system can fail while cleaning some fields e.g. number is not a number; or
- the system can create a new field e.g. by mapping the national tender procedure type to enumeration value and storing both of them.

Stage	Number of records*	Data size**	Estimated data size***
Clean	1866845	11	44

* Number of records August 2016
** Data size in GB August 2016
*** Estimated data size in GB September 2017

4. Matched data

Clean data contains one document for one publication without any relation between publications describing the same tender. The matched stage connects such publications into one group. It contains the same number of records as the previous stage in a same format but adds information which connects documents.

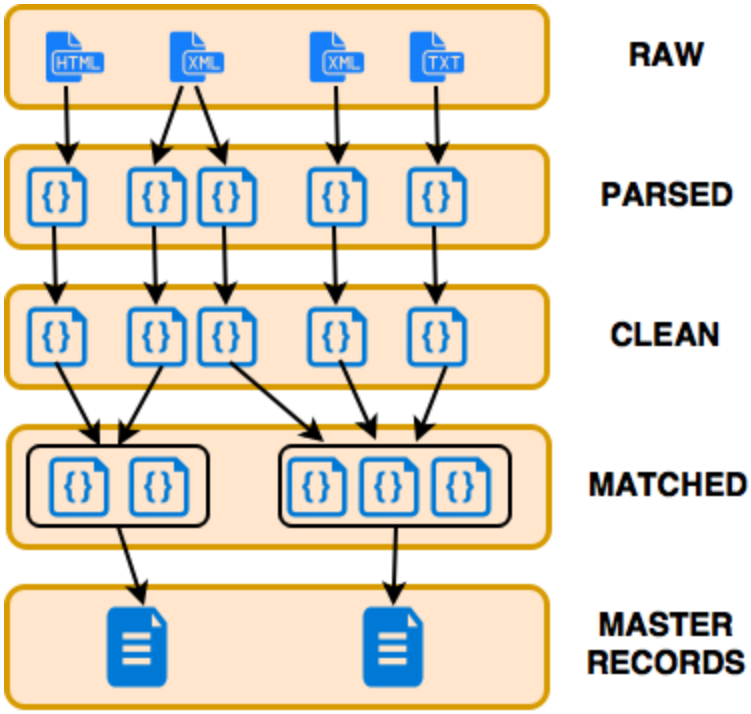
5. Master records

The mastered stage is the last stage. In this phase of data processing we aggregate data from all publications describing one tender and create one master object that is a final image of a specific

tender. This will be a final dataset that the DIGIWHIST project will publish together with some related data discussed in Chapter 2.2.

Because we are at the very beginning of matching algorithms and creating master records it's difficult to estimate how many records there will be in this stage. We can only use an expert estimate here based on the fact that:

- there are stricter rules for publishing for above threshold procurements;
- above threshold tenders will consist of a contract notice and a contract award;
- many publications will be form corrections;
- this leads to about half the number of records in comparison with matched data collection.



Company data

The company data database is an important dataset for:

- research activities;
- buyer/supplier matching algorithms;

It's currently 350GB and it comprises:

- company register - 51.288.900 records
- financial data - 67.828.500 records
- manager information - 43.954.700 records
- link - 1.489.220.000 records. This is not a final number because some data hasn't been imported yet so we can expect around 1.800.000.000

Public sector data

In comparison to other datasets the public sector data database size is negligible. Currently we have for all categories together a database of 7.5GB. Despite the fact that it is likely to grow we don't expect it will be larger than 15GB for raw data.

Data utility

Procurement data has variety of potential users. The foremost goal of the project is to create data which is usable for policy analyses and research, therefore drawing users from public institutions such as the EC or national governments and academia.

Various studies such as PwC(2013)[1] identify lack of reliable data (especially in terms of unified structure and centralisation) as a major drawback for similar policy analyses. Additionally data including various red flags might be highly beneficial to anti-fraud agencies such as OLAF and various NGOs focusing on anti-corruption activities.

2. FAIR data

2.1 Making data findable, including provisions for metadata

Data discoverability

We are active members of the Open Contracting community which is dedicated to the publication of public procurement data. We plan to follow the standard defined by the OCDS and, together with other OCDS publishers, our outputs will be linked from http://www.open-contracting.org/why-open-contracting/worldwide/ which is the central directory of similar datasets.

Naming conventions

Although we designed our own data template for recording public procurement data our outputs will be published in the format of the Open Contracting Data Standard (OCDS) (http://standard.open-contracting.org) that is currently the only widely used standard for publishing this type of data.

Keywords

Individual structures, fields and enumeration values follow the OCDS which makes our data easy searchable for everyone.

Versioning

Each data release will be versioned in accordance with the OCDS for package releases[2].

[1] https://ec.europa.eu/anti-fraud/sites/antifraud/files/docs/body/identifying_reducing_corruption_in_public_procurement_en.pdf
[2] http://standard.open-contracting.org/latest/en/schema/release_package/

Metadata

To make our data more open and findable we will publish metadata based on the OCDS[3] like URL, published date, publisher etc. When we decide to publish metadata that is not described in the OCDS we will do it in such a way that it only extends it and remains compatible.

2.2. Making data openly accessible

Processed and produced datasets

The main goal of WP2 is to publish a data collection that will best reflect each tender based on the raw data detected and obtained by our software together with additional tender related information such as indicators that are outputs of WP3. Some secondary data collected and/or processed by our software won't be published as a separate dataset because it's either under the protection of contract or it's not a goal of DIGIWHIST project to publish such datasets and we use it only for purposes of making our final data better and more accurate.

Public datasets

- Public procurement data: DIGIWHIST will publish all tender information it detects in the described format together with a methodology of how the data was collected and created since it will be an aggregation of more public data sources.
- Indicators developed within the scope of WP3 as deliverable D3.6 (*Indicators implemented in database*) will be a part of the public procurement data. They will be published as tender-related information
- Public sector data collected within WP2 will be published, in compliance with the Grant Agreement, as tender/buyer related information or aggregate statistics

Non-public datasets

- Company data is a dataset that the DIGIWHIST project bought in compliance with the Grant Agreement. Its usage is defined by a contract with the supplier (Bureau van Dijk). This prevents DIGIWHIST from making it public but it enables DIGIWHIST partners to use it for scientific research.
- Tender-related data of a speculative nature. Within the complex process of data cleaning and merging we obtain some variables with some informative value, yet with a high risk of being erroneous. These will be valuable for research purposes yet their publication might bring serious legal and misinterpretation risks (for example through publishing the wrong supplier). Even with rigorous disclaimer release of such data might in fact reduce understandability and usability of the data to journalists, researchers etc.

Data access

All data published within WP2 will be accessible through an API designed to be easily usable and machine readable. It will use SSL for authentication and encryption of communication between API and end user so that delivered data can't be modified by anyone during its transmission from source system to its destination and the receiving party is thus certain of the source.

[3] http://standard.open-contracting.org/latest/en/schema/reference/?highlight=metadata

Access methods

The DIGIWHIST API will use a standard HTTP protocol which means there is no need for special software to access the data. All popular programming languages implement functions or libraries that enable developers to communicate via HTTP protocol. On top of that anyone can access the data via a web browser such as Internet Explorer, Chrome or Firefox.

Documentation

As well as the data and the API itself, the documentation of the API is under development and this will be released together with the data by the end of a project. This documentation will describe all API endpoints and methods in detail and will be the only document needed to successfully connect to the described data source.

Restrictions

There are no explicit restrictions on re-use of published data therefore there is also no need for a data access committee. The software as well as the data documentation will be released as D2.8 (*Methods paper describing database content, data collection, cleaning, and linking*)

Software license

All software products developed within the framework of WP2 will be published as open source under the MIT[4] licence. This licence grants permission, free of charge, to any person obtaining a copy of the software and associated documentation files (the "Software"), to deal in the Software without restriction, including without limitation on the rights to re-use, copy, modify, merge, publish, distribute, sublicense, and/or sell copies of the Software, and to permit persons to whom the Software is furnished to do so.

Data, documentation and code repository

Source codes and documentation

All source codes and their documentation will be stored in a public GitHub repository. This is a well-known repository in an open source community and it is considered to be a best practice to share codes in this way because it creates a centralised point for third parties to obtain and re-use the code which has been created.

The choice was thus motivated by current best practice, which is based on the following GitHub features:

- security of a repository;
- connectivity reliability;
- data durability;
- graphical user interface;
- easy re-use enabled by major third-party software development tools.

[4] https://opensource.org/licenses/MIT

Published data

Due to the fact that the data size may grow and the nature of a data may change every day we decided that it's more appropriate to implement a custom solution for data publication that allows users to access updated data on demand and in small portions instead of always having to download the whole dataset from a static repository, even if a significant part of a content hadn't changed. This solution is referred to as an API in this document and is based on standard technologies like HTTP protocol or JSON data format. A proper backup and recovery plan needs to be implemented in the production phase to avoid potential system failures or data loss.

2.3. Making data interoperable

It is a top priority goal for DIGIWHIST to make data from various sources using various national code tables and enumerations interoperable and easily understandable.

Standard vocabularies

To make data easily readable and processable we follow open contracting data standard structures and enumerations. This should make data completely clear for everyone who wants to use it. The importance of standard vocabularies like common procurement vocabulary (CPV) or NUTS code arises when we take into consideration that we publish data in many different languages. For users it's almost impossible to understand all of those languages but standard vocabularies help to make basic information like the subject of a tender or location of works understandable.

Mapping

Where various national values for different fields are used (e.g. tender procedure type) we put extensive effort into mapping national values to standard vocabularies. We do such mapping for fundamental data like:

- lot status;
- tender size;
- procedure type; and
- other fields that have enumeration values in OCDS.

2.4. Increase data re-use

Data licence

The licensing of the data produced is still an open issue given the legal differences across all jurisdictions and the differences in rights granted by official data providers. Even though there are licenses designed for open data (e.g. ODbL[5]) any licensing can be complicated and we will have to proceed country by country. For example, copyright law in the Czech Republic explicitly excludes "data in public registries whose distribution is in public interest" from any possibility of licensing or protection.

[5] http://opendatacommons.org/licenses/odbl/

Data availability

D2.6 is due in month 31 of DIGIWHIST. This means that a final linked database and related algorithms will be published by the end of September 2017. In compliance with the Grant Agreement, data will be available for at least three years after the end of a project.

Data reusability

Data published by DIGIWHIST will be accompanied by an OCDS version that it is compatible with. This is especially important because data standards are evolving all the time and it is expected that some changes in OCDS will occur during the implementation phase which ends in month 31 of the project or during following years. Adding a compatible OCDS makes implementation of the data processing software much easier.

Quality assurance

There are several consortium members contributing to the quality assurance process. This is led by Datlab which validates data at several levels.

1. **Consistency** - examining the integrity of the data, its structural consistency with the designed model and suggesting further changes to the model. Responsible organisation: Datlab
2. **Completeness** - ensuring that all the relevant data (at the form level) has been obtained from the source. Responsible organisation: Datlab
3. **Correctness** - ensuring that the raw data obtained is consistent with the source, i.e. containing the same values, codelists match national legislation etc. Responsible organisations Datlab + UCAM domain experts
4. **Data availability** - evaluating the quality of the processed data in terms of availability of variables (in contrast to Correctness this is not looking for the errors in our software anymore, but assessing the quality of the data, which possibly carries many imperfections from the source systems). Responsible organisation: Datlab.

The outputs of this process, most importantly the Data availability step, will be described in detail together with validation results in D2.7 which will be released together with the final database.

3. Allocation of resources

Costs for making data FAIR and its coverage

Making data FAIR is significant part of the project. Almost the whole of WP2 entails re-creating data from original sources and making it FAIR. Thus, in some sense, at least 36% of overall project costs (the WP2 share of the work) is dedicated to this. Since other work packages such as WP1 also contribute to that goal, we can conclude that overall considerable resources and time are dedicated to publishing data in accordance with FAIR principles. The costs of achieving this are built into the project budget. Some of the activities (deliverables) which are crucial for this include:

- Legal and regulatory mapping (D 1.1)
- Implement data templates compatible with OCDS (D2.3)
- Raw (D2.4), Cleaned and structured databases (D2.5), Final linked database (D2.6)

- Data validation (D2.7)
- Methods paper describing database content, data collection, cleaning, and linking (D2.8)

Responsibilities for data management

Until the end of project the UCAM team is responsible for making the data public, documented and secure. After that OKFN will take over the sustainability phase, ensuring the availability of published resources at least for five years after the project end.

The current distribution of labour requires several steps of complex data gathering and processing which is designed and coordinated by the UCAM IT team. Other consortium members take responsibility as part of that process for particular actions:

1. Source annotation (UCAM domain experts)
2. Parsing and processing of the data from sources (UCAM IT)
3. Validation and bug reporting (Datlab, UCAM domain experts)
4. Data release and provision to other partners (UCAM IT)

The process further involves many decisions which will affect the final quality and scope of the data. This includes, for example, the prioritization of countries, sources (especially if multiple sources are available in given countries) and individual variables in order to deliver the most comprehensive dataset with the resources available. Such decisions are made following discussion amongst consortium members to reflect both future usability of data and the practical costs of gathering it.

Are the resources for long term preservation discussed?

As explained earlier, the current infrastructure is run and further developed by the UCAM IT team. The choice of storage now, as well as in the future, is primarily made by balancing the costs, ease of processing and potential re-use.

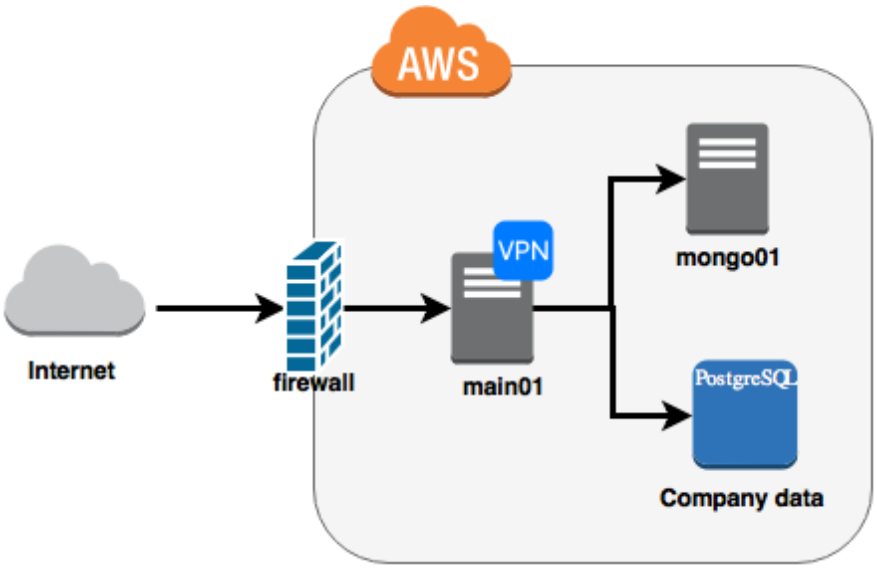
Thus far we have designed (D2.1) and implemented the whole architecture using the AWS IAAS (Infrastructure as a Service) provider and we will run it until September 2017. Thereafter OKFN will be responsible for ensuring that the data gathered during the implementation phase is available until the end of the sustainability phase. One of the key upcoming decisions is to agree with OKFN what kind of architecture they will use to do this; alternatives include using the existing architecture, or they could run the database and software using their own servers or they could buy some servers from a hosting company. The chosen solution will reflect above mentioned principles in order to facilitate one of key project goals - making the data available for further re-use.

4. Data security

Access

We apply different security mechanisms on different levels to ensure the security of the production infrastructure:

1. Access to the production infrastructure is granted to approved personnel only.
2. The production environment is secured by firewall.
3. There is only one entry point to the infrastructure. There is no direct access vector to servers/services(PostgreSQL, RabbitMQ etc.)
4. All communication with the production environment (API, server access) is possible only via channels with a strong cryptography enabled OpenVPN, SSH)



OpenVPN

There is an OpenVPN server installed on the infrastructure entry point. The clients have first to connect to the OpenVPN with a proper certificate (certificates are user specific). OpenVpn is configured to disable “visibility” of connected clients between each other.
Once the OpenVPN connection is successfully established, the user can continue with SSH access to the rest of infrastructure.

SSH

Connection to servers is possible only via SSH. The SSH is configured to disallow password authentication, only public/private key authentication is possible. Keys are not shared amongst users.

Administrator access

Client trying to access the infrastructure have to:

1. Connect to OpenVPN with a proper certificate
2. Connect to the entry point server via SSH

Service access

To connect to one of the services(MongoDB, RabbitMQ) from the client directly:

1. Connect to OpenVPN with a proper certificate
2. Connect to the entry point server via SSH
3. Create an SSH tunnel to target service/port.

Backup

The data is backed up on a daily basis with 30 day retention period. The backups are stored as encrypted snapshots to Amazon S3 infrastructure into geographically different locations.

Availability and recovery strategy

The current state of the project does not require 24/7 availability setup. In case of service failure, we are able to restore whole production environment in a short period.

Encryption

The data is stored on encrypted storage devices. Database storages as well as backups, logs etc. are placed only on encrypted volumes.

Software patches

Software patches are applied on a regular basis. Critical path updates and security patches bulletins are reviewed and fixes are applied within hours when necessary.

5. Ethical aspects

The EC original Ethics Check and RP1 Ethics Check have both raised a number of concerns around the impact of data sharing :

Data Protection & Privacy

Detailed information was sought in relation to the procedures that will be implemented for data collection, storage, protection, retention and destruction along with confirmation that they comply with national and EU legislation**.** A lengthy description of data security was provided which covers the service provider for data storage, encryption, backup, secure access, network configuration, user accounts, software patches, log audit, recovery strategies, data destruction and passing data to third parties. The full response, which has been accepted by the EC, can be found in our Consortium Ethics Check Response RP1.

Personal information

Detailed information was also sought on the type of personal information that is to be collected from interviewees/informants as well as the privacy/confidentiality issues related the personal data. We provided a detailed explanation of how the data will be accessible through password login and will be kept in encrypted files which are backed up daily. Data will only be kept for the length of the project. All participants will be made aware of how their data will be used and will sign consent forms. For the purposes of analysis, informant's personal data will be anonymised so they cannot be identified. We will not publish any information which would allow the identification of interviewees/informants. The full response, which has been accepted by the EC, can be found in our Consortium Ethics Check Response RP1.

Protection of whistleblowers

Whistle-blowers may face severe professional and physical reprisals if their identities were wrongfully disclosed. Our national portals will not themselves provide the whistleblowing function, but will link to a national partner's website that provides such a channel so no personal data will be

transmitted through and stored on DIGIWHIST servers. All the national partners will be experienced in running such portals and will be thoroughly vetted in advance. Each will sign a Memorandum of Understanding requiring them to comply with EU and national whistle-blowing and data protection legislation. We will only enable the whistleblower function in countries where we can identify partners that are capable of implementing the required national and international standards.

The management of the potential discovery of illegal activities, in particular corruption

We have agreed with the EC that we will develop a set of guidelines, including for interviewers, on how to manage such situations based on the best practice required by the University of Cambridge and with input from all consortium partner institutions.

The stigmatization of organizations and/or individuals because of false alarms caused by the developed indicators and systems

The possible stigmatization of individuals has been addressed satisfactorily as we will not share any individual data at all – neither for private nor public persons – so the issue will not arise.

The possible stigmatization of organisations has yet to be resolved and is still being discussed by the Consortium with the EC (as at September 2016). This is an ongoing “conversation”.