COMPACT

CYBERSECURITY FOR PUBLIC ADMINISTRATIONS

D2.8 Data management plan (v2)

Work Package:	WP2		
Lead partner:	KUL		
Author(s):	KUL, CINI, INOV, AIT		
Due date:	31 Oct 2018		
Version number:	1.0	**Status:**	Final

Grant Agreement N°:	740712
Project Acronym:	COMPACT
Project Title:	COmpetitive Methods to protect local Public Administration from Cyber security Threats
Call identifier:	H2020-DS-2016-2017
Instrument:	IA
Thematic Priority:	Secure societies – Protecting freedom and security of Europe and its citizens
Start date of the project:	May 1st, 2017
Duration:	30 months

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

Horizon 2020 European Union Funding for Research & Innovation

Project co-funded by the European Commission under the Horizon 2020 Programme.

Revision History

Revision	Date	Who	Description
0.1	2018-09-15	KUL	Revision and adaptation of DMP v1
0.2	2018-10-12	CINI, AIT, INOV	Input from partners in relevant sections
0.3	2018-10-26	KUL	Final draft ready for review
0.4	2018-11-06	KUL	Deliverable finalised

Quality Control

Role	Date	Who	Approved/Comment
Internal reviewer	2018-10-30	BIT	Approved
Internal reviewer	2018-10-31	BOL	Approved

Disclaimer:

Table of Contents

Definitions and acronyms

COMPACT	Competitive Methods to protect local Public Administration from Cyber security Threats, financed under H2020 programme, grant agreement no. 740712
DMP	Data management plan
FAIR	Findable, Accessible, Interoperable, Reusable data
ORD	Open research data pilot

1. Introduction

This deliverable will set out the second version of the data management plan (DMP) for the COMPACT project. A DMP is a key element of good data management, which is especially important in the COMPACT context, as all Horizon 2020-funded projects from 2017 onward are required to contain a DMP.[1] Alongside open access publications, open access research data contributes to achieving open science.[2]

The DMP is applicable to data, needed to validate the results presented in scientific publications. It is part of the European Commission's Open Research Data (ORD) Pilot, which was launched as a general project requirement in 2017. According to the Commission's website, the pilot aims to 'improve and maximise access to and re-use of research data generated by Horizon 2020'. It balances between openness and protection of scientific information, commercialisation and intellectual property rights (IPRs), as well as privacy and security concerns.[3]

A previous version of this document, entitled D2.6 Data management plan (v1), was drawn up in M6 of the project (October 2017), and set out the basic principles of data management in COMPACT. The current report has been updated to reflect the advances and current status of data management in the project.

The final version of this document will be provided in M24 (April 2019).

The DMP for COMPACT is based on the European Commission's Guidelines on FAIR Data Management in Horizon 2020[4] and the COMPACT Grant Agreement.[5] It defines which data will be open by detailing the types of data generated by the project, its accessibility for verification and re-use, exploitation, as well as its curation and preservation.

In order to implement the open research data principle, the DMP sets out the following information:

- The handling of research data during and after the end of the project,
- What data will be collected, processed and/or generated,
- Which methodology and standards will be applied,
- Whether data will be shared/made open access and
- How data will be curated and preserved (including after the end of the project).

[1] European Commission, H2020 Programme Guidelines on FAIR Data Management in Horizon 2020 Version 3.0, July 2016, p. 3.
[2] http://ec.europa.eu/research/openscience/index.cfm
[3] https://ec.europa.eu/research/openscience/pdf/openaccess/background_note_open_access.pdf
[4] European Commission, H2020 Programme Guidelines on FAIR Data Management in Horizon 2020 Version 3.0, July 2016, p. 4.
[5] Grant agreement no. 740712 – COMPACT, Part B – Section 2.2.5.

Sections 2 to 6 of this document will cover the different DMP components, based on the outline suggested in the Guidelines. They are based on input from the following partners: AIT, CINI, INOV and KUL, as indicated in the relevant sections.

2. Data summary

2.1. AIT

What is the purpose of the data collection/generation and its relation to the objectives of the project?

The data collected by the human factor profiling survey is related to variables measuring knowledge about cyber security, work-related variables (i.e., time pressure, decision-making autonomy), IT-skills, organizational variables (i.e., security climate) and demographic information (i.e. gender, age, tenure). Further, we collect qualitative data based on interviews and observations when applying and evaluating the developed awareness methods Investigators Diary and Sectopia. The purpose of the data collection/data analysis is to assess the effect of psychological variables (e.g., knowledge, motivation) on cyber-secure behaviour of LPA's employees. Data is processes accumulated that no conclusions on the individuals are possible.

What types and formats of data will the project generate/collect?

Data is collected in the form of numbers (i.e., from rating scales), statements (i.e., open questions) and transcribed data (i.e., from observations).

Will you re-use any existing data and how?

Next to project purposes, data will be used for scientific publications. Therefore, data will be anonymized and analysed cumulatively.

What is the origin of the data?

Data is empirical data collected by applying research methods like questionnaires, interviews, or observations.

What is the expected size of the data?

The expected size of the data depends on the number of participants. The expected size of the data for the remaining data collection activities cannot be predicted at this stage of the project. We expect empirical data from the human factor profiling survey from about 50 employees, and 10 people using the Investigator's Diary and Sectopia.

To whom might it be useful ('data utility')?

Within the COMPACT project, data will be useful to analyse existent behaviour and attitudes of LPA employees (from the respective end-user partners). Based on this, adequate awareness method will be chosen to train and increase cyber-secure behaviour. Outside of the COMPACT project, data may be useful as a basis for creating more general insights on cyber secure behaviour and attitudes of LPA employees. This may be done by enriching existing data with further data of other LPA employees.

2.2. CINI

What is the purpose of the data collection/generation and its relation to the objectives of the project?

Within COMPACT Project, CINI is in charge of developing an advanced Security Information and Event Management (SIEM) system endowing LPAs' organisation with real-time monitoring capabilities. SIEM services receive log files, which represent records of the events occurring within an organization's systems and networks when a user attempts to authenticate into the system or a system event occurs (such as starting a service or shutting down the system, etc.). The content/records of these log files related to computer security information are then analysed for investigating malicious activities. A particular alarm or event is generated in relation to the particular detected attack.

What types and formats of data will the project generate/collect?

SIEM systems come with a number of adapters for receiving data/events from a wide variety of sources, such as Operating System (OS) log files (in proprietary or open formats) or Commercial Off The Shelf (COTS) products for logical and physical security monitoring, including: Windows registries, Wireshark, Nessus, Nikto, Snort, Ossec, Argus, Cain & Abel, OpenNMS, Nagios, CENTEROS, Ganglia, Milestone, openHAB, IDenticard, FieldAware, and CIMPLICITY. In terms of data generated, the data format is JSON and it is stored in an Elasticsearch analytics engine.

Will you re-use any existing data and how?

Currently no existing LPA data is used in order to develop/test the SIEM. However, if needed, we foresee to use existing anonymized data present within the archives of some of the LPAs involved in the project.

What is the origin of the data?

Data collected and analysed by the SIEM system will originate from testing activities carried out at pilot sites (LPAs involved within the project) and will be relied on the collection and analysis of log files of the LPAs participating in the COMPACT project.

What is the expected size of the data?

At this stage of the project, it is not possible to predict the size of the data that will be processed and stored, because it highly varies on the number, type and frequency of monitored data sources.

To whom might it be useful ('data utility')?

Data collected by the SIEM system can be useful to the other technical partners in charge of developing COMPACT's tools and services, such as risk assessment tool or personalization of training courses for LPAs' employees, etc.. They could be useful to other research groups working on similar research, as well as for testing alternative SIEM solutions.

What is the purpose of the data collection/generation and its relation to the objectives of the project?

During the COMPACT project, INOV will collect data to test and demonstrate its Business process intrusion detection system (BP-IDS). This data collection is related with the project objective "SO3: Lower the entry barrier to timely detection and reaction to cyber-threats", and may occur during the tasks: "Task 4.3 Threat intelligence and monitoring Component"; "Task 4.5 Integration of solutions in a unified platform"; "Task 5.1 Validation and Demonstration scenarios"; "Task 5.2 Trials Setup"; and "Task 5.3 Pilot execution and demonstration".

What types and formats of data will the project generate/collect?

It is not definitive in this phase, since the data to collect is still being defined. However, it is foreseen that INOV technology will at least collect three types of information: documentation, operative data and statistical data. The documentation will be collected before the trials, and composed of data produced by CMA that explains the business processes employed in their municipality. The operative data will be collected during the trials by monitoring the interactions between the IT systems present in the CMA IT infrastructure and its database, but it will only be processed within CAM IT infrastructure. While the statistical data will be collected by INOV after the trials to evaluate the performance of BP-IDS during the trials.

Will you re-use any existing data and how?

It is not definitive in this phase, since the datasets are still being defined. However, from the three datasets that have been selected only the documentation will be reused based on what was provided by CMA.

What is the origin of the data?

It has not been completely defined, the datasets need to be specified first to respond to this question. However, at this stage the three datasets chosen will be collected from the Amadora municipality. Specifically, the documentation will be based on the Amadora archive documents, the operational data based on the data produced on this LPA's computers, and the statistical data will be produced created based on the deployments of INOV's tool in the Amadora's infrastructure.

What is the expected size of the data?

It is difficult to estimate the size of the data at this stage, because the size of the data highly varies on the network protocols or the files monitored.

To whom might it be useful ('data utility')?

The principal benefactor of this dataset will be CMA that will use the tools developed to monitor threats against their infrastructure. INOV will use it for adapting BP-IDS for LPAs. Besides INOV this dataset might be useful to technical partners in the COMPACT project, that require live data to adapt their technical solutions to LPA environments. Outside the COMPACT's consortium the data may be useful to technical partners that develop tools for local public administrations and want to test their tools in a realistic environment. Data that can be useful by third parties may contain personal data and will only be processed within CMA IT infrastructure, INOV will not collect this data or process it in anyway. The statistical data collected by INOV, will be evaluated, and if considered of interest to the research community, it will be made available (if proper authorized by CMA).

3. FAIR data

Under Horizon 2020's principle of open access to data, research data must be FAIR: findable, accessible, interoperable and reusable. This will contribute to the use of data in future research.[6]

In order to be **Findable**:

- F1. (meta)data are assigned a globally unique and eternally persistent identifier.
- F2. data are described with rich metadata.
- F3. (meta)data are registered or indexed in a searchable resource.

[6] Mark D. Wilkinson et al., The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data* **3**, Article number: 160018 (2016).

- F4. metadata specify the data identifier.

In order to be **Accessible**:

- A.1. (meta)data are retrievable by their identifier using a standardized communications protocol.
- A1.1. the protocol is open, free, and universally implementable.
- A1.2. the protocol allows for an authentication and authorization procedure, where necessary.
- A2. metadata are accessible, even when the data are no longer available.

In order to be **Interoperable**:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles.
- I3. (meta)data include qualified references to other (meta)data.

In order to be **Re-usable**:

- R1. meta(data) have a plurality of accurate and relevant attributes.
- R1.1. (meta)data are released with a clear and accessible data usage license.
- R1.2. (meta)data are associated with their provenance.
- R1.3. (meta)data meet domain-relevant community standards.

Answering the following questions will contribute towards compliance with the FAIR data standards. The answers are provided in a comprehensive manner, not on a yes/no basis.

3.1. Making data findable, including provisions for metadata

3.1.1. AIT

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

The participant ID is a unique and persistent identifier which is linked to the data set of one person. In terms of the HFP, this information will not be made available for the LPAs – just the aggregated data. Interview and observed data will be anonymised. Meta data is not (yet) linked to a standard identification mechanism.

What naming conventions do you follow?

With regard to participant IDs, these are numbered sequentially.

Will search keywords be provided that optimize possibilities for re-use?

No

Do you provide clear version numbers?

Regarding data collection and data analysis, version numbers of the according files are provided.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Metadata standards do not exist in our discipline. The metadata created within the project are mainly aggregated statistical data (e.g., means, standard deviations) as well as aggregated qualitative data (e.g., interpreted data from sources like interviews or observations).

3.1.2. CINI

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

Not defined yet. However, if the Zenodo Repository will be adopted for data storing and sharing, the persistent identification through DOIs for sharing research results will be adopted.

What naming conventions do you follow?

We refer to the "Glossary of Key Information Security Terms" provided by NIST[7] or, in turn, to the SANS Glossary of Security Terms[8].

Will search keywords be provided that optimize possibilities for re-use?

At this stage we have not yet planned to provide keywords for optimizing re-use.

Do you provide clear version numbers?

Not defined yet, it is possible that the gathered data will be not modified after their collection, so one version for each session will be provided.

[7] R. Kissel. Glossary of key information security terms. NIST Interagency Reports NIST IR 7298 Revision 1, National Institute of Standards and Technology, February 2011.
[8] http://www.sans.org/security-resources/glossary-of-terms/.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Not defined yet

3.1.3. INOV

Are the data produced and/or used in the project discoverable with metadata, identifiable and locatable by means of a standard identification mechanism (e.g. persistent and unique identifiers such as Digital Object Identifiers)?

Not defined yet, the data collected so far does not contain any metadata. Thus additional measures to create the metadata are necessary to make data discoverable.

What naming conventions do you follow?

Not defined yet, the data collected so far does not follow naming conventions. Thus, additional measures to convert data are necessary to make data compatible with the naming conventions chosen.

Will search keywords be provided that optimize possibilities for re-use?

Not defined yet, the data collected is not structured. Thus, additional measures to convert data are necessary to make data searchable through keywords.

Do you provide clear version numbers?

Not defined yet, it is probable that most data collected will not be altered and that only one version will be created per data.

What metadata will be created? In case metadata standards do not exist in your discipline, please outline what type of metadata will be created and how.

Not defined yet, but due to the nature of the data collected creating metadata will be very unlikely.

3.2. Making data openly accessible

3.2.1. AIT

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under

restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.
Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out. There has been no opt-out in the COMPACT project yet.

Data collected within the project will be pseudonymised. After data analysis, aggregated data may be openly available (e.g., in the form of publications or project reports). Overall, data cannot be linked to a specific person.

How will the data be made accessible (e.g. by deposition in a repository)?

Only anonymised data will be made accessible if an assigned scientific article is published, and this one is part of an open access data strategy.

What methods or software tools are needed to access the data?

Data are recorded as numeric values or strings. Data will be provided as csv or Excel files.

Is documentation about the software needed to access the data included?

No, this is not needed.

Is it possible to include the relevant software (e.g. in open source code)?

n.a.

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

Not defined yet.

Have you explored appropriate arrangements with the identified repository?

n.a.

If there are restrictions on use, how will access be provided?

n.a.

Is there a need for a data access committee?

n.a.

Are there well described conditions for access (i.e. a machine-readable license)?

n.a.

How will the identity of the person accessing the data be ascertained?

n.a.

3.2.2. CINI

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.
Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out. There has been no opt-out in the COMPACT project yet.

Data produced and/or used in the project will be made openly available by default only after a pseudonymisation and/or anonymization process in order to prevent that the data can be attributed to a specific person. This is the case for example of the data representing the USB usage of the CDA.

How will the data be made accessible (e.g. by deposition in a repository)?

Data will be made accessible through a research data repository. The consortium will take measures to enable third parties to access, mine, exploit, reproduce, and disseminate the data free of charge.

What methods or software tools are needed to access the data?

The best candidate tool for data sharing – at the time of this writing – is ZENODO, an OpenAIRE/CERN compliant repository. Zenodo builds and operates a simple and innovative service that enables researchers, scientists, EU projects and institutions to share, preserve and showcase multidisciplinary research results (data and publications), that are not part of the existing institutional or subject-based repositories of the research communities. Zenodo enables researchers, scientists, EU projects and institutions to: • easily share the long tail of small research results in a wide variety of formats, including text, spreadsheets, audio, video, and images across all fields of science. • display the research results and receive credit by making the research results citable and integrating them into existing reporting lines to funding agencies like the European Commission.

- easily access and reuse shared research results.

Is documentation about the software needed to access the data included?

Yes, it is.

Is it possible to include the relevant software (e.g. in open source code)?

It is possible, but not decided yet if open source code will be included.

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

The consortium plans to deposit data in an OpenAIRE compliant research data repository.

Have you explored appropriate arrangements with the identified repository?

If the consortium choice will be Zenodo, this information is available here: http://about.zenodo.org/

If there are restrictions on use, how will access be provided?

If the consortium choice will be Zenodo, this information is available here: http://about.zenodo.org/

Is there a need for a data access committee?

Not defined yet

Are there well described conditions for access (i.e. a machine-readable license)?

If the consortium choice will be Zenodo, this information is available here: http://about.zenodo.org/

How will the identity of the person accessing the data be ascertained?

If the consortium choice will be Zenodo, this information is available here: http://about.zenodo.org.

3.2.3. INOV

Which data produced and/or used in the project will be made openly available as the default? If certain datasets cannot be shared (or need to be shared under

restrictions), explain why, clearly separating legal and contractual reasons from voluntary restrictions.
Note that in multi-beneficiary projects it is also possible for specific beneficiaries to keep their data closed if relevant provisions are made in the consortium agreement and are in line with the reasons for opting out. There has been no opt-out in the COMPACT project yet.

Not defined yet. However since making the data public reveals CMA's business secrets (like detailed business processes specification, operative data with personal data or sensitive information), data should only be available to a restricted number of personnel to avoid any risk to CMA.

How will the data be made accessible (e.g. by deposition in a repository)?

Not defined yet. Data is stored and only accessed in computers located on CMA headquarters. INOV will only have access to the information by deploying the tool in closed premises.

What methods or software tools are needed to access the data?

To access the data INOV will physically access the information by visiting the Amadora's headquarters, or remotely by accessing Amadora VPN. Either way, the information collected will remain in CMA's premises.

Is documentation about the software needed to access the data included?

Not defined yet, but most likely it will not be available since it may undermine CMA's security methodology.

Is it possible to include the relevant software (e.g. in open source code)?

Not defined yet

Where will the data and associated metadata, documentation and code be deposited? Preference should be given to certified repositories which support open access where possible.

Not defined yet

Have you explored appropriate arrangements with the identified repository?

Not defined yet

If there are restrictions on use, how will access be provided?

Not defined yet

Is there a need for a data access committee?

Not defined yet

Are there well described conditions for access (i.e. a machine-readable license)?

Not defined yet

How will the identity of the person accessing the data be ascertained?

Not defined yet

3.3. Making data interoperable

3.3.1. AIT

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

As the produced data are based on specific methods and measures, they are interoperable with further data stemming from the same methods or measures. Further, the quantitative and qualitative data may be used for complementing other research in this area.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

Not defined yet.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

Not defined yet.

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

Not defined yet.

3.3.2. CINI

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

Yes

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

The data format used to represented the data is JSON, a lightweight data-interchange format supported by all modern programming languages support in one form or another.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

Not defined yet

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

Not defined yet

3.3.3. INOV

Are the data produced in the project interoperable, that is allowing data exchange and re-use between researchers, institutions, organisations, countries, etc. (i.e. adhering to standards for formats, as much as possible compliant with available (open) software applications, and in particular facilitating re-combinations with different datasets from different origins)?

Not defined yet, data produced will be produced according to CMA standard formats and most likely specific to their environment. Thus additional measures may be necessary to make data interoperable.

What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?

Not defined yet.

Will you be using standard vocabularies for all data types present in your data set, to allow inter-disciplinary interoperability?

Not defined yet

In case it is unavoidable that you use uncommon or generate project specific ontologies or vocabularies, will you provide mappings to more commonly used ontologies?

Not defined yet

3.4. Increase data re-use (through clarifying licenses)

3.4.1. AIT

How will the data be licensed to permit the widest re-use possible?

n.a.

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

n.a.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

n.a.

How long is it intended that the data remains re-usable?

Not defined yet.

Are data quality assurance processes described?

Not described yet.

3.4.2. CINI

How will the data be licensed to permit the widest re-use possible?

If the consortium choice will be Zenodo, this information is available here: ://about.zenodo.org/

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

At this stage of the project it is not possible to predict the date for making re-use available.

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

Anonymised data will be usable by third parties.

How long is it intended that the data remains re-usable?

We intend to store data and make it re-usable for an appropriate period of time according to the key guidelines established by the following regulations: Directive (EU) 2016/1148 of the European Parliament and of the Council of 6 July 2016 concerning measures for a high common level of security of network and information systems across the Union (NIS Directive) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)

Are data quality assurance processes described?

Not defined yet

3.4.3. INOV

How will the data be licensed to permit the widest re-use possible?

Not defined yet

When will the data be made available for re-use? If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

Not defined yet

Are the data produced and/or used in the project useable by third parties, in particular after the end of the project? If the re-use of some data is restricted, explain why.

Not foreseeable at this stage

How long is it intended that the data remains re-usable?

For the duration of the project.

Are data quality assurance processes described?

Not defined yet

4. Allocation of resources – the whole consortium

According to the Horizon 2020 rules, costs related to open access to research data are eligible for reimbursement during the duration of the project under the conditions defined in the COMPACT Grant Agreement, in particular Articles 6 and 6.2.D.3.[9] These are direct costs, related to subcontracting of project tasks, such as subcontracting the open access to data.

What are the costs for making data FAIR in your project?

According to the budget, AIT has planned to address 10000 EUR for making data FAIR in the project.

[9] See the Annotated Model Grant Agreement, available at: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf#page=83

How will these be covered? Note that costs related to open access to research data are eligible as part of the Horizon 2020 grant (if compliant with the Grant Agreement conditions).

Not defined yet

Who will be responsible for data management in your project?

A specific role has been foreseen in the project, the Data Controller (DC). Salvatore D'Antonio from CINI has been appointed as Data Controller and will be responsible for data management.

Are the resources for long term preservation discussed (costs and potential value, who decides and how what data will be kept and for how long)?

Not defined yet

5. Data security

5.1.1. AIT

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

Data is stored only internally in our facilities which provide state of the art IT security. State of the art IT security measures and company policies mitigate most of the risk of illegitimate access. Firewalls (to prevent illegitimate access from outside) and a rights-based-file system (to prevent illegitimate access from inside) are the countermeasures against this risk.

Is the data safely stored in certified repositories for long term preservation and curation?

It is not definitive in this phase. Potentially, collected data are stored in a repository.

5.1.2. CINI

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

Regarding security, all the data collected will be stored on a database only accessible to authenticated users on the partner premises. Regarding the data recovery, database backups will be stored on premises and only accessible to CINI. Sensitive data will never be transferred outside the LPA premises except in an anonymised form.

Is the data safely stored in certified repositories for long term preservation and curation?

It is not definitive in this phase.

5.1.3. INOV

What provisions are in place for data security (including data recovery as well as secure storage and transfer of sensitive data)?

Regarding security, all the data collected will be stored on a database only accessible to authenticated users on the partner premises. Regarding the data recovery, database backups will be stored on premises and only accessible to INOV.

Is the data safely stored in certified repositories for long term preservation and curation?

It is not definitive in this phase, but it is not expected to store the collected data in a repository.

6. Ethical and legal aspects

Are there any ethical or legal issues that can have an impact on data sharing? These can also be discussed in the context of the ethics review. If relevant, include references to ethics deliverables and ethics chapter in the Description of the Action (DoA).

Addressing legal and ethics challenges is an important part of the COMPACT work plan. As already indicated in the Section 5 of the Description of the Action, special attention has been paid to these issues since the very beginning of the project. A legal partner (KUL) forms part of the consortium, handling guidance and providing relevant expertise. Internal ethics controls in COMPACT include setting up an internal ethics committee and defining checklists for project compliance, as per Task 1.4. The consortium has also appointed an ethics and privacy manager.

Moreover, a dedicated work package (WP8) deals specifically with ethics requirements, such as notifications to competent data protection authorities (POPD - Requirement No. 4), authorisation for use of non-public data (POPD - Requirement No. 5) and details on preventing the misuse of research findings (M - Requirement No. 6). All these requirements have been duly met by relevant partners.

SELP, or security, ethics, legal and privacy, is one of the building blocks of setting up COMPACT products. Specific tasks have been allocated to deal with SELP aspects of COMPACT (T2.5, T3.4), especially research ethics, privacy rights and data protection regime under the GDPR.

Nevertheless, a DMP is on principle not part of general GDPR compliance due to the latter's slightly different scope of application. Namely, the GDPR applies to processing of personal data. Personal data are defined in Art. 4(1) of the GDPR as any information relating to an identified or identifiable natural person ('data subject'). Open research data, on the other hand, can be any kind of data resulting from research, whether personal, pseudonymised (which are still personal data), anonymised formerly personal data, but also data from (chemical) lab trials, industrial data or any kind of data that have no connection to an individual person, thus falling wholly outside the scope of the GDPR.

However, should personal data be used as part of the DMP, they will be anonymised, or if that is not possible, they will be pseudonymised according to the current state of the art, and the additional information necessary for re-identifying the individual will be kept separately (according to Art. 4(5) of the GDPR). Pseudonymiation is a permissible measure for data protection in research, according to Art. 89; nevertheless, if possible, further identification should not be possible. Moreover, for personal data processed in the context of COMPACT trials, relevant controller-processor agreements for data sharing will be concluded between end users and technology partners.

Is informed consent for data sharing and long term preservation included in questionnaires dealing with personal data?

Informed consent forms & information sheets – updated to reflect the GDPR requirements

The internal ethics committee has provided COMPACT partners with informed consent forms and information sheets, updated to reflect the new GDPR requirements. The information requirements are laid down in Art. 13 and 14. Accordingly, the information sheets give research participants information about, inter alia:

- Purposes of data collection, data processing and data analysis
- Types of personal data processed
- Transfer of their personal data between their employer/LPA and the relevant technical partner(s), involved in the trials
- The rights they have as data subjects, and information on how to exercise them
- The period for which the data will be stored

The participants will receive the information sheet together with informed consent forms before they start trials. They have the right to withdraw from research at any time without any adverse consequences.

7. Other

Do you make use of other national/funder/sectorial/departmental procedures for data management? If yes, which ones?

Not defined yet.