**AGENT-BASED
SUPPORT TOOL FOR
THE DEVELOPMENT
OF AGRICULTURE POLICIES**

D9.2 Data Management Plan

agricore

Deliverable Number	D9.2
Lead Beneficiary	AAT
Work package	WP9
Delivery Date	M6 - February 2020
Dissemination Level	Public

www.agricore-project.eu

Document Information

Project title	Agent-based support tool for the development of agriculture policies
Project acronym	AGRICORE
Project call	H2020-RUR-04-2018-2019
Grant number	816078
Project duration	1.09.2019-31.8.2023 (48 months)

Version History

Version	Description	Organisation	Date
0.1	First template	AAT	10/02/2020
0.2	Adding datasets templates	AAT	20/02/2020
0.3	Including reused data information	AAT	26/02/2020
1.0	Final draft version	IDE	28/02/2020

Executive Summary

This document focuses on providing the AGRICORE's consortium with information on how to deal with data generated during the execution of the project AGRICORE and how it should be managed. Specifically, this document provides information about the type of data that will be generated, the standards and the accessibility to the data for verification and re-use, the curation and preservation of data procedures. At the same time, being a public document, it will allow the easy identification of the datasets generated within the project by other researchers who may be interested in reusing or validating the information produced.

This deliverable details all the data management procedures within AGRICORE, in response to article 29.3 of the grant agreement, specifically regarding open access to research data. Accordingly, it presents information for partners in the consortium as well as for external parties about the different datasets generated and used within the project. They will be categorized and technically detailed in terms of data collection, processing and generation in order to manage the data that is associated and generated by the work of the AGRICORE consortium. This topic requires a significant level of detail posing for the need of an autonomous deliverable, rather than being a part of D10.1 (Project Management Handbook) where all of the other day-to-day procedures for the running of the project are presented. The monitoring of the AGRICORE project with regard to this area will be the subject of continuous reviews within the context of WP9 at each consortium meeting.

Moreover, the current data management plan (DMP) is the first draft of this document. This means that the DMP will be a living document, continuously updated by its responsible with the information provided by all partners in the project. A final version of this document will be submitted to the European Commission by the end of the project.

This deliverable is based on the Guidelines on Data Management in Horizon 2020 document provided by the European Commission, the information provided by the UK Digital Curation Center and the DMPtool.

Abbreviations

Abbreviation	Full name
AAT	Ayesa Advanced Technologies
DCC	UK Digital Curation Centre
DMP	Data Management Plan
DPO of the project	Data Protection Officer of the project
EC	European Commission
FAIR	Findable, Accessible, Interoperable and Reusable
GDPR	General Data Protection Regulation
ID	Identifier
IDE	IDENER
IPR	Intellectual Property Rights
TBD	To Be Determined
WP	Work Package

Table of Contents

1 Introduction

This document presents the initial version of the Data Management Plan (DMP) for the AGRICORE project. This information has been prepared partially following the guidance of the UK Digital Curation Centre (DCC) (http://www.dcc.ac.uk), an internationally-recognized centre of expertise in digital curation with a focus on building capability and skills for research data management. The DCC provides expert advice and practical help to research organisations wanting to store, manage, protect and share digital research data. This DMP for AGRICORE details the public datasets that the project:

- will generate,
- whether and how it will be exploited or made accessible for verification and re-use,
- how it will be curated and preserved.

The academic papers produced within the project will be made available as open access for several years (following green or gold open access approach). In a similar way, the data generated as a result of the project's research activities will be put also online and freely available in open access repository. The DMP will include the details on the datasets generated, providing detailed information in such datasets. Specifically, this DMP details the following aspects:

- metadata generation,
- data preservation,
- data storage beyond the end of the project.

In particular, the consortium partners acknowledge their responsibilities for fulfilling and updating this DMP, which includes:

- the first version of the DMP must be operative in the first 6 months of the project,
- it will be a live document continuously updated within the project timeframe,
- data identified (generated within the project) in DMP must be shared in an online repository by the producers of the data, appropriate support will be provided to those involved in the project from both the DMP responsible and the project coordinator.

At a deeper level of detail, the aim of the DMP is that its execution and update will lead to:

- a better understanding of the data produced as output from the project,
- clarity on how the data is actually used within the project and outside of it,
- continuity in the work of the consortium in the event of staff leaving or entering the project during its lifecycle or equivalently staff changing role within the project during its lifecycle. This includes such areas as:
    - o avoiding duplication of effort i.e., re-collecting or re-working data,
    - o enabling validation of results,
    - o contributing to collaboration through data sharing,
    - o increasing visibility of output and thereby leading to greater impact. In particular, enabling other researchers to cite and use the datasets generated within the project.

AGRICORE embraces the open-access approach incentivised by the EC. This means that the project aims to improve and maximise access to and re-use of research data generated by it and considers the need to balance openness and protection of scientific information, commercialisation and Intellectual Property Rights (IPR), privacy concerns, security as well as data management and preservations. Within the AGRICORE project, most of the deliverables produced are done as public documents. A similar approach will be undertaken with the generated data.

2   Data Management Plan

The DMP aims to regulate all the processes related to the data life cycle.

At this stage of the project, the consortium is working in the first stage of the data collection and description. Figure 1 shows the whole process developed in the data lifecycle, composed of mainly 4 steps:

- Data storage. This step is composed of the next 4 internal procedures: collection, description/characterization, analysis and re-collection.
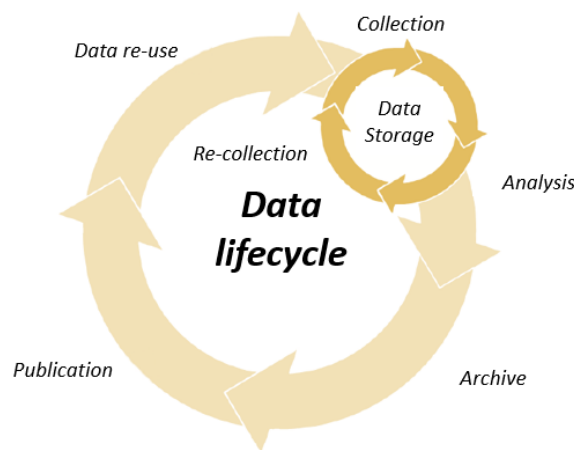
- Archive

- Publication

- Data re-use

Figure 1: Data Life cycle

In addition, the usage of the data in the consortium will be also considered and monitored.

This initial DMP provides templates to characterise the datasets generated or employed within the project (which consists mainly of primary, derived and simulation data). At the moment, there is no identified need for collecting personal data within the project. Nonetheless, the participatory research activities (still under definition) may involve surveying such information. Should it be the case, any data collection process will be done considering the anonymization of the data, target that will be assessed and guided by the recommendations of the DPO of the project.

Moreover, as part of the project work plan, the partners participating in AGRICORE will generate an Index tool of the data sources available at EU level that can be used for conducting agricultural research. As part of this work, the consortium is identifying and characterising these databases. The information contained in such data sources will not be stored in any repository but rather characterized in the Confluence platform first and later in the Index tool (which is the scope of Task 1.8). The characterisation and mapping of this information are separated from the other project activities, meaning that the data sources analysis will go further than the specific needs for executing the rest of the project's tasks. This approach aims to produce a more extensive mapping of data sources, promoting and facilitating future research on related topics.

3 Data summary

This section includes in the initial DMP the types and formats of data generated and/or collected.

In order to define an effective DMP, the data should be categorized. In this way, the data can be categorized according to

- The source
- Format
- Stability
- Volume

Regarding the source, the data can be categorized into 4 types: observational, experimental, simulation and derived/compiled. The AGRICORE project will generate experimental, simulation and derived datasets. Nevertheless, the project is still investigating whether observational data will be needed and will be generated by involving stakeholders in participatory research activities.

- Observational data may be needed in the context of participatory research with stakeholders. According to the gaps existing in the identified datasets, further information might be needed to execute the activities of the AGRICORE project.
- Experimental data. No experimental datasets are expected to be generated within the project.
- Simulation data are generated by the models produced within the project. These data will be reproducible if the model and inputs are preserved. Moreover, these data will be reproducible by the platform to be developed in the project.
- Derived data are generated by existing datasets and are reproducible. These data are characterized in the context of WP1, where available public data sources are explored to provide useful information for the impact assessment of agricultural policies at European and national levels.

When describing the format categorization, the data can be generated in different forms: text, numeric, audiovisual, models, computer code, discipline-specific and instrument-specific. The AGRICORE project will generate data in all these forms. It is important to notice that the project will be gathering mainly data in the text and numeric formats of categorisations to populate the data warehouse needed and employed in the project.

Regarding the stability of the data, the data can be fixed or change over the course of the project and this will affect the way it is managed. The common categories of a dataset are:

- Fixed datasets never change after being generated or collected.
- Growing datasets: new data may be added while old data are never changed or deleted.
- Revisable data: new data may be added while the old data may be changed or deleted.

In this context, partners in the AGRICORE project plan to use all of the three types listed above throughout the execution of the project, requiring to keep track of data versions.

Finally, it is important to define **the volume of data** to be prepared. In general, the AGRICORE project is expected to handle a high volume of data, including both the datasets used and the generated ones.

3.1 Produced datasets

At this initial stage of the project, no specific details about the datasets to be produced have been defined yet. Nonetheless, this section provides the template that all users should be filling for each of the produced datasets. As part of the continuous updating process of this document, the newly identified datasets will be included both in this document and in the internal portal of the project (Confluence).

The template for the datasets produced within the project is the next one:

Partner: (Partner name)	
Identifier	The identifier will have the following format (where internalID is an incremental ID starting from 1 for each dataset produced in the project): AGRICORE_shortDescription_MainAuthor_internalID
Dataset description	A detailed description of the dataset
Purpose of the data	Explanation of the purpose of this dataset
Type of data	Numeric, text, etc.
Form of the data	The way the information is provided and generated
Format of the data	The format in which the data are released
Origin of the data	The source generating this data. It should also include details on reused data
Dataset stability	Details on the stability (as defined in the previous section) of the data
Size of data	Estimation (or measurement when already produced) of the data size

3.2 Reused data sources

As indicated in the previous section, the AGRICORE project includes dedicated activities to map and characterise available data sources useful for agricultural modelling research. The findings of such activities will be released as part of the Index tool generated within the project. Nonetheless, a list of the data sources identified will be included in this document. At the current stage of the project (M06 -February 2020), the consortium has already identified and initially characterised the following datasets:

Region covered	Main topic covered	Dataset name
EU level	Climate-Related	European Climate Assessment & Dataset NASA Prediction Of Worldwide Energy Resources (POWER) Modern-Era Retrospective analysis for Research and Applications, Version 2 (MERRA-2) ADMS AgMERRA Climate Forcing Dataset for Agricultural Modeling AgCFSR Climate Forcing Dataset for Agricultural Modeling CRU TS (Climatic Research Unit Time-series) dataset Global Summary of the Day (GSOD) WorldClim Version2

	Soil / Land / Quality / Biodiversity-related	GIOŚ LUCAS - Land Cover and Land Use Landscape European Soil Database v2 Raster Library MIRCA2000 SoilGrids SoilHydroGrids
	General Agriculture data	FADN FSS (Farm Structure Survey) Eurostat OECD FaoStat IACS - Integrated Administration and Control System Feedipedia Feedprint AFFRIS - Aquaculture Feed and Fertilizer Resources Information System
	Geo-referenced datasets	CORINE Land Cover LUCAS Earthstat Gridded Livestock of the World WorldClim - Global Climate Data MOD16A3 v006 MODIS/Terra Net Evapotranspiration Yearly L4 Global 500 m SIN Grid Rainfall Erosivity Database on the European Scale (REDES) RUSLE High-Resolution Layer: Imperviousness Density (IMD) 2015
National and regional datasets	Farm characterization datasets	(Italian) RICA(Spanish) SSAO2016 Survey on the Structure of Agricultural Holdings (Polish) ARMA - RDP 2014-20 (PL) Implementation reports (Polish) Statistics Poland - Agricultural and horticultural crops (Polish) Statistics Poland - Animal production, Farm animals (Polish) Statistics Poland - LOCAL DATA BANK (Polish) ARMA - RDP 2014-20 (PL) Implementation reports (Greece) ELSTAT - Annual Agricultural Statistical Survey (Greece) ELSTAT - Crops Survey (Greece) ELSTAT - Farm Structure Survey (FSS) (Greece) ELSTAT - Livestock Surveys (Greece) Hellenic Statistical Authority on Agriculture and Livestock (Spanish) Statistics on Means of Production: Use of Fertilizers (Spanish) Statistics on Means of Production: Commercialization of Phytosanitary Products (Spanish) Statistics on Means of Production: Registration of New Machinery (Spanish) International Penn TablesSSAO2016 (Spanish) Agrifood Foreign Trade Statistics (Spanish) BDSICE. Cost and price index (Spanish) BDSICE. National production and demand indicators (Spanish) BDSICE. Price and costs. Agricultural wage index (Spanish) BDSICE. Price and costs. Salary increases in agreement and salary increases registered in agriculture (Spanish) Climate data obtained by the Agroclimatic Stations Net (Spanish) Evolution of provincial agricultural macromagnitudes 2005-2014 (Spanish) FEGA (Spanish) MARM. Household consumption database (Spanish) Monthly production, movement and stock data (AICA) (Spanish) Multi-territorial Information System of Andalusia (SIMA) (Spanish) National Agrarian Accounting Network RECAN

		(Spanish) National Agricultural Economic Statistics: Agricultural rates and prices paid (Spanish) National Agricultural Economic Statistics: Agricultural rates and prices received (Spanish) National Agricultural Economic Statistics: Agricultural rates and salaries (Spanish) National Agricultural Economic Statistics: Average land prices for agricultural use (Spanish) National Agricultural Economic Statistics: Short-term prices of agricultural products (Spanish) National Statistics Institute: Agricultural Census (2009) (Spanish) National Statistics Institute: Survey on Production Methods in Agricultural Operations (2009) (Spanish) Olive. Data obtained from the monitoring of pests and diseases in the biological control stations (Spanish) Organic Farming in Spain (Spanish) Rice. Data obtained from the monitoring of pests and diseases in the biological control stations (Spanish) SIGGAN
	Land Quality and Characteristics Datasets	(Polish) Database on mineral nitrogen content in Poland (3 depths, 2 times per year) (Spanish) SIGPAC (Spanish) ESYRCE (Spanish Survey on Crop Surfaces and Yields) (Spanish) National Soil Erosion Inventory (INES) (Spanish) National Hydrological Report
	Climate-Related Datasets	(Polish) Climate-related dataset for Poland (Polish) ADMS (Polish) Data set on mineral nitrogen content in Poland (3 depths, 2 times per year) (Greece) Climate-related dataset

4 FAIR Data

This section aims at presenting how the AGRICORE consortium should save the information in order to make the data Findable, Accessible, Interoperable and Reusable (FAIR).

4.1 File formats

The file format to be employed is a primary factor to generate FAIR data. The file format should be accessible in the future so that the selected formats should be non-proprietary, open, with documented standards, in common usage by the research community, using standard character encodings and uncompressed.

The following table includes the recommended and acceptable formats that AGRICORE consortium adopts following the UK Data Service guidance on recommended formats:

Type of data	Recommended formats	Acceptable formats
Tabular data with extensive metadata (Variable labels, code labels, and defined missing values)	• SPSS portable format (.por) • Delimited test and command ('setup') file (SPSS, Stata, SAS, etc.) • Structured text or mark-up file of metadata information, e.g. DDI XML file	• Proprietary formats of statistical packages: SPSS (.sav), Stata (.dta), MS Access (.mdb/accdb) • Matlab formatted datafiles (.MAT), • R data files (.rdata or .rda; R workspace) • GAMS datafiles GAMS Data eXchange (GDX), • GIS datafiles Esri Shapefile .SHP, .DBF, .SHX, JSON-sat • NLOGIT/LIMDEP data files .lpj .sav • SAS datafiles (.sas) • ODBC datasource • GRETL datafiles • GAUSS datafiles • Eviews datafiles.
Tabular data with minimal metadata (column headings, variable names)	• Comma-separated values (.csv) • Tab-delimited file (.tab) • Delimited text with SQL data definition statements	• Delimited text (.txt) with characters not present in data used as delimiters • Widely-used formats: MS Excel (.xls/.xlsx), MS Access (.mdb/.accdb), dBase (.dbf), OpenDocument Spreadsheet (.ods)
Textual data	• Rich Text Format (.rtf)	• Hypertext Mark-up Language (.html)

	• Plain text, ASCII (.txt) • eXtensible Mark-up Language (.xml) text according to an appropriate Document Type Definition (DTD) or schema	• Widely-used formats: MS Word (.doc/.docx) • Some software-specific formats: NUD*IST, NVivo and ATLAS.ti
Documentation and scripts	• Rich Text Format (.rtf) • PDF/UA, PDF/A or PDF (.pdf) • XHTML or HTML (.xhtml, .htm) • OpenDocument Text (.odt)	• Plain text (.txt) • Widely-used formats: MS Word (.doc/.docx), MS Excel (.xls/.xlsx) • XML marked-up text (.xml) according to an appropriate DTD or schema, e.g. XHMTL 1.0
Audio data	• Free Lossless Audio Codec (FLAC) (.flac)	• MPEG-1 Audio Layer 3 (.mp3) if original created in this format • Audio Interchange File Format (.aif) • Waveform Audio Format (.wav)
Image data	• TIFF 6.0 uncompressed (.tif)	• TIFF other versions (.tif, .tiff) • RAW image format (.raw) • Photoshop files (.psd) • BMP (.bmp) • PNG (.png) • Adobe Portable Document Format (PDF/A, PDF) (.pdf) • JPEG (.jpeg, .jpg, .jp2) if original created in this format • GIF (.gif)
Video data	• MPEG-4 (.mp4) • OGG video (.ogv, .ogg) • Motion JPEG 2000 (.mj2)	• AVCHD video (.avchd)

Tabular data and Textual data are going to be commonly handled in implementing the activities planned in the AGRICORE project. Hence, the next recommendations are given to the consortium be make datasets easier to understand and to export:

• Do not put more than one table on a worksheet
• Create charts on new sheets. Do not embed them in the worksheet with the data
• Include a header row with a clear title for each column
• Upon manipulating Excel worksheets, be careful when deleting the content of row or column cells. The simple deletion of the content of the cells may yield missing values when importing the data in other software packages (i.e., Stata) using selected procedures. Please take care of deleting the whole rows and/or columns
• Be mindful of the possibility that data are not written following the international notation employing the comma (,) as the separator of thousands and the dot (.) as the separator of decimals

- Relying on a data conversion programme like Stat/Transfer (https://stattransfer.com/) may facilitate data management

4.2 Metadata: Data document

Clear and detailed documentation is essential for data to be understood, interpreted and used. The data document describes the content, formats and internal relationships of the data. The following template describes the documentation required within the project to document the datasets used and generated. This information will be indexed in the internal project's platform (Confluence). If during the elaboration of the data formats any field is not determined yet, the label TBD should be included. In case that any field is provisional, the * should be included at the beginning of the text.

At this stage of the project, most of the information which is planned to be gathered is coming from public sources, which are generally available online. Confluence will collect the analysis of all the data sources in order to consolidate the information providing a "Master Table" to feed the Data Index that will be created as a platform in Task 1.8 of the project. In addition to the "Master Table", a detailed database characterization will be developed aiming to provide the information required to use the data effectively within the project. Even if the consortium is aware of the heterogeneity of the existing data provided by different sources, the consortium will make its best effort to gather the following information to ensure these domains are covered

- Dataset information
- Content description
- Technical description
- Access

Dataset information	
Name	Name of the dataset
Data contained	Describing the data collected and the agroupation of data
Dataset structure	Detailing tables structures and main indications to help understanding the reading of the dataset
Dataset format	How can this data be available: format to be used
Data source access	How can this data can be accessed: Web, download, individual access...
Source	Source of the dataset
Generation process	Process followed to gather/produce the dataset
Author	Author of the data
Maintainer	Institution/company in charge of maintaining
Last actualization	Date
Update frequency	Period of time
Periods covered	Time when the dataset started to gather information
Release date (past and expected future ones)	Figure detailed in days
Additional information	Relevant information to be included for a better understanding of the data source

Rights	Any known IPR, licences or restrictions on the use of the data (if applicable)
Access information	Where and how to access the data (if applicable). Publicly available/Access request required
Embargo	Description of the embargo that will be held on the research data (i. e., previous publication on paid-for journals and later published openly) and the period that this embargo will entail (If applicable).
Publication information	Sharing policies among the following options: Green Open Access; Gold Open Access; Non-shared; Published in paid-for journals; Other (Describe)

In addition, the next template will also be filled in by each partner characterising an existing or produced dataset for the purpose of promoting FAIR principles:

Dataset identifier
What naming conventions do you follow?
Will keywords for searching the dataset identifier be provided to optimize the opportunity for re-use?
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. (If applicable)
Will this dataset generated and/or used in the project be made openly available as the default? If positive, How?
What methods or software tools are needed to access the data?
If there are restrictions on use, how will access be provided?
What data and metadata vocabularies, standards or methodologies will you follow to make your data interoperable?
How will the data be licensed to allow for the widest re-use possible?
When will the data be made available for reuse?
Are the data produced and/or used in the project useable by third parties, in particular after the end of the project?
How long is it intended that the data remains reusable?

4.3 Data sharing: Publication

The AGRICORE consortium fully embraces the H2020 requirement for Open Access publishing, following the guidelines presented by the EC. The project will ensure '**green**' and '**gold**' publishing, i. e., self-archiving in one or more repositories. The associated costs of publications will be claimed as part of the H2020 grant.

The project will make its datasets of results publicly available through the following repositories:

- The project website: https://agricore-project.eu/

- The European data sources Index Module developed in Task 1.8

- Public software repositories such as GitLab and GitHub

- The dissemination of the software tools will not be limited to the distribution of the source code but will also include an extensive plan for promoting the adoption of the proposed technologies by other researchers and developers. This will be done in the frame of the planned dissemination and communication activities and especially during the clustering and workshop activities.

In any case, the partners will not publish their results when the publication of the research data violates any confidential restriction or IPR of any of the partners.

5   Allocation of resources

5.1   Responsibility and roles

IDE is the project coordinator and will be in charge of the management of the Confluence site.

Ayesa Advanced Technologies (AAT) will be responsible for maintaining and updating the DMP. Therefore, AAT will be in charge of continuously updating the DMP to produce a final version by the end of the project by providing clear guidelines on the management of personal data and compliance with the General Data Protection Regulation (GDPR). The DPO of the project provided by AAT has an email channel to efficiently manage everything related to data for the duration of the project. The email at which all concerns about data management can be addressed to is dpo_agricore@ayesa.com

All partners will be responsible for updating and contributing to the Confluence site of the project. Partners will be also requested to deliver periodically:

- Pre-print manuscripts of any accepted publication
- Slides and posters shown in conferences
- Raw data supporting the figures is papers and deliverables
- PhD dissertations generated in the frame of the project

In case new personnel are assigned to a relevant role, responsibilities with respect to the DMP are also taken over. For details on the management roles and structure of AGRICORE see D10.1. In case the contact person for a certain dataset is leaving the project, the institution of the original contact person will take over the responsibility and will assign a new contact person.

6 Ethical aspects

In this section, the ethical aspects and legal compliances about DMP are presented. None of the data generated during the execution of AGRICORE project will contain information on individual or companies, so every data will be anonymized. The technical data will not represent or use any human being or corporate entity. Moreover, the AGRICORE consortium will do its best to adapt data management to the European GDPR.

With respect to the IPR, the whole project is using open-access data and will be ensuring data generated will be released as open access.

Only when the achievement of the main objective of a partner is jeopardised by making specific parts of the research data openly accessible, the consortium will evaluate the reasons for not giving access in order to protect the partner's commercial activity.

6.1 Data, samples and equipment exchange between EU and non-EU parties

The consortium will keep an account of the scientific samples and data exchanged with non-EU parties, which will be stored on the Confluence platform on the page "Data exchanged between EU and non-EU members". This will be available to the EC services at any time. The potential exchange between non-EU parties will be closely evaluated before proceeding, requesting permission to the concerned entities. Nonetheless, private data (not already publicly available) will not be shared with non-EU partners unless strictly necessary and always after receiving the required authorisations.

All the ethics issues will be developed in WP11 and addressed and reported in D11.1, D11.2 and D11.3

7   Other issues for data management

At the moment, the AGRICORE project does not make use of other information requiring a special treatment different than the one already explained within this document, and no other issues have been found in this initial phase of writing the DMP. If any other concerns are raised, the DPO of the project will provide the best recommendation to address the situation.

In preparing this report, the following deliverables have been taken into consideration:

Deliverable Number	Deliverable Title	Lead beneficiary	Type	Dissemination Level	Due date
D10.1	Project management handbook	IDE	Report	Confidential	1