PROTEUS

Scalable online machine learning for predictive analytics and real-time interactive visualization

687691

D2.3 Proteus Data Management and Ethics Plan

Lead Author: Rachel Finn, Tally Hatzakis & Kush Wadhwa, Trilateral Research

Reviewers: Ricard Martinez, Director de la Cátedra de Privacidad y Transformación Digital

&

Marcos Sacristán, Rubén Casado, Treelogic and Iván Gallego Mateos, Daniel Toilmil Martin, ArcelorMittal

Deliverable nature:	<Report (R)>
Dissemination level: (Confidentiality)	< Consortium (CO)>
Contractual delivery date:	31/5/17
Actual delivery date:	31/5/17
Version:	1.3
Total number of pages:	49
Keywords:	Data management, exploitation, research ethics, open access

Abstract

This deliverable represents the first iteration of the PROTEUS data management plan, a comprehensive document that will guide the management of data used and generated by PROTEUS through the course of the project and after the close of the project. This plan includes a consideration of the legal, ethical and policy issues relevant to the management of PROTEUS data and how the project will meet these obligations. It also outlines plans for further developing this strategy to guide partners in exploiting PROTEUS data, including providing open access to the data to allow other stakeholders to re-use it.

Executive summary

This deliverable represents the first iteration of the PROTEUS data management plan, a comprehensive document that will guide the management of data used and generated by PROTEUS through the course of the project and after the close of the project.

Data has always been an integral part of the scientific research process and is an object of scientific interest both as an object of investigation and as a potential asset for those within and outside the research process. Data management assists researchers in outlining how data needs to be protected and making informed decisions about exploiting research data or sharing them to allow others to exploit them.

Particular policies, legal frameworks and ethical guidance relevant to the management of research data frame these issues. As such, this document begins by providing foundational and contextual information related to the project, the relevant policies and legal frameworks and the initial plans for ensuring that the collection and use of data within the project conforms to the issues within this larger context. Specifically, the document examines the European Commission's Open Data Research Pilot, intellectual property rights relevant to PROTEUS, data protection standards and practices and guidance on ethical research processes.
PROTEUS will use and generate three specific data sets that fall within the following categories:

1. ArcelorMittal data provided to the project (ArcelorMittal data)
2. Derived data about the functioning of the scalable online machine learning tools, including data from the benchmarking and technical evaluation process (PROTEUS toolset data)
3. Data from the evaluation of the visualisation aspects of the tool (PROTEUS evaluation data)

The first data set raises significant intellectual property issues as it has commercial sensitivity for ArcelorMittal. The second and third data sets will be generated by the project and will raise both intellectual property issues for the consortium via the PROTEUS toolset data and data protection and ethical research issues via the PROTEUS evaluation data. Specifically, the usability evaluation exercises will be conducted with volunteers that are also ArcelorMittal employees, and thus, issues around voluntary participation, informed consent and data protection are relevant here.

The deliverable progresses by describing how the project plans to meet all of our legal, ethical and policy obligations surrounding PROTEUS data. In doing so, the project has agreed the following principles:

- Data owned by ArcelorMittal will be shared with consortium members, although consortium members can only access this data through the project coordinator
- PROTEUS partners agree to respect the commercial confidence of the data provided by AMIII
- Human participants in PROTEUS research activities are under no obligation to participate and their involvement will be strictly voluntary
- PROTEUS partners will respect data protection and ethical research principles related to the following:
    - Participant confidentiality and anonymisation
    - Informed consent
    - Data minimisation
    - Purpose limitation
    - Transparency
    - Rights of access, correction and erasure
- The coordinator will hold all commercially sensitive data and personal data and will manage access to this data for PROTEUS partners

In addition to meeting European regulations, the project will also meet legal regulations set by the Spanish government as the data and many of the research activities will be located in Spain. This includes the registration of the project coordinator as a data controller with the Spanish data protection authority and the provision of specific information required by Spanish data protection legislation. Drafts of the PROTEUS information sheet and informed consent forms are included in Annex A to demonstrate how we will meet these obligations. An independent ethical and data protection expert, Ricard Martinez, has reviewed these materials and added comments and suggestions to ensure the materials and procedures are as robust as possible (see Annex C for short CV).

Subsequent versions of this document will be produced in months 36 of the project, and these will focus on downstream issues as the project develops and matures. Chief among these will be a considered strategy for exploiting the data used by and produced within PROTEUS, including a consideration of how the project can leverage its familiarity with ArcelorMittal data to assist them in capturing opportunities associated with this data or potentially designing new value added products. It will also outline how partners plan to exploit the data produced within the project, including using it to demonstrate the added value associated with the tools developed within PROTEUS. Finally, as the project matures, later iterations of this document will evaluate whether the PROTEUS data can be made open access to allow other stakeholders to exploit the data in ways unforeseen by the project.

Document Information

IST Project Number	687691	Acronym	PROTEUS
Full Title	Scalable online machine learning for predictive analytics and real-time interactive visualization		
Project URL	http://proteus-bigdata.com		
EU Project Officer	Martina EYDNER		

Deliverable	Number	D2.3	Title	PROTEUS data management and ethics plan
Work Package	Number	WP2	Title	Industrial case: Requirements, challenges, validation and demonstration

Date of Delivery	Contractual	M18	Actual	M18
Status	version 1.3		final □	
Nature	report **X** demonstrator □ other □			
Dissemination level	public □ restricted **X**			

Authors (Partner)	Rachel Finn, Tally Hatzakis, Kush Wadhwa (Trilateral Research)			
Responsible Author	Name	Rachel Finn	E-mail	Rachel.finn@trilateralresearch.com
	Partner	Trilateral Research	Phone	+44 (0)207 559 3550

Abstract (for dissemination)	This deliverable represents the first iteration of the PROTEUS data management plan, a comprehensive document that will guide the management of data used and generated by PROTEUS through the course of the project and after the close of the project. This plan includes a consideration of the legal, ethical and policy issues relevant to the management of PROTEUS data and how the project will meet these obligations. It also outlines plans for further developing this strategy to guide partners in exploiting PROTEUS data, including providing open access to the data to allow other stakeholders to re-use it.
Keywords	Data management, exploitation, research ethics, open access

Version Log			
Issue Date	Rev. No.	Author	Change
10/5/16	0.3	Rachel Finn	Draft circulated to project partners
19/5/16	1.0	Rachel Finn	Changes from reviewers incorporated
28/4/17	1.1	Tally Hatzakis, Rachel Finn	Updates to the document given project progress over Y1.
15/5/17	1.2	Ricard Martinez	Comments on the draft
29/5/17	1.3	Tally Hatzakis, Rachel Finn & Kush Wadhwa	Revision to the document following Ricard Martinez' review

Table of Contents

Abbreviations

AEPD: Agencia Española de Protección de Datos

AMIII: ArcelorMittal

BSA: British Sociological Association

DPIA: Data Protection Impact Assessment

EC: European Commission

GDPR: General Data Protection Regulation

ICT: Information and Communication Technologies

IPR: Intellectual Property Rights

ORD Pilot: Open Research Data Pilot

1 Introduction

Data has always been an integral part of the scientific research process and is an object of scientific interest both as an object of investigation and as a potential asset for those within and outside the research process. Scientists have long been engaging in the protection of their data assets from unauthorised access and sharing and the dissemination of their data assets for use by others. However, disciplinary differences and traditions have often governed whether data was kept under lock and key or shared more widely with other scientists and stakeholders (Wessels, et al., 2014). However, recent policy advances in the field of open access have encouraged scientists and other researchers to consider managing this asset more explicitly, including assessing whether the data needs to be protected or whether it can be shared more widely to enable others to exploit already existing data for scientific and other gains. Rather than relying on tradition, all researchers are being encouraged to make an informed decision about exploiting research data or sharing them to allow others to exploit them.

In this context, research data has a specific focus and meaning. According to the European Commission, who funds this research, "research data" refers to "information, in particular facts or numbers), collected to be examined and considered and as a basis for reasoning, discussion, or calculation (EC, 2016, p. 3)." This may include data as varied as statistics, interview recordings or notes from field observations (ibid.).
This document will assist the consortium to identify:

a) what data needs to be protected, and
b) what data might be sharable outside of the consortium.

Data management plans are an important tool to enable this consideration. Data management plans encourage researchers to consider a range of different legal and ethical issues that govern whether data needs to be protected and the extent to which it can be made openly accessible or shared in other ways. Legal and ethical issues to be considered include open access policies, intellectual property rights, data protection legislation and research ethics. Data management plans enable researchers to consider these and to outline their compliance with each of these requirements. In addition, proper data management also encourages researchers to explicate how their data is being collected, processed, stored and exploited. According to Jones (2011), such planning helps researchers to check they have the necessary support for their research and enables them to make sound decisions with a clear understanding of the different options.

The PROTEUS project seeks to develop a predictive analytics system to help companies identify defects early in the manufacturing process to reduce economic and environmental waste; hence improving companies' efficiency and competitiveness, as well as environmental friendliness. The work will be based on the steel coil manufacturing process of ArcelorMittal, a large steel manufacturing company. To this end, the project is expected to utilise three types of data

a) Raw, numberic data from the company's databases that will form the basis for the tools developed by the PROTEUS project.
b) Data derived from the big data analytics tools developed by the project consortium, including data related to the technical evaluation process.
c) Evaluations, opinions and insights by company employees, who will volunteer to evaluate the utility of the tools in the operational context of ArcelorMittal.

The project will consider sensitivities in the management of each of these data types, using this data management plan. First, the project needs to consider intellectual property rights, including trade secrets, as the data upon which we are relying to build the PROTEUS toolset is owned by ArcelorMittal and is commercially sensitive. Second, the project will generate new forms of data and insights though big data analytics tools, the ownership of which is outlined in the Consortium agreement and will be revisited within this document as the data is available and the project matures. Third, PROTEUS will undertake evaluative research with company employees, in the process of which it will collect some personal data. Hence the project will be guided by the EU research ethics and data protection framework to respect participants' rights alongside sensitivities raised by doing research with employees of a specific company. Finally, the project sits within the European Commission's Horizon 2020 Open Data Research Pilot, and consequently, the project must consider the extent to which the data from this publicly funded research can be made open access.

As such, this data management plan will form a guidance document for the project. It will help the project navigate intellectual property and data protection issues and govern whether and when the consortium can provide open access to research data. The data management plan will also provide a template that will help the partners utilise and exploit the data effectively during and at the end of the project. Finally, this document will provide essential guidance on ensuring adequate protections for those participating in research, including providing personal data.

While these procedures are being co-managed by the project coordinator, Treelogic, and Trilateral Research, providing such protections is the responsibility of every partner in the PROTEUS project. This document provides guidance to all partners to help them effectively manage PROTEUS data. It will be continuously updated as the project progresses. This version is a revision of the initial version produced in M6 of the project. A final version will be produced in M36 of the project to consider developments and respond to issues raised within the research process. The UK's Data Curation Centre (2016) has the most comprehensive guidance on data management practices and the creation of a data management plan, and this document relies heavily on their advice and good practice guidance.

2 Project description

The PROTEUS project will develop a data analytics system for steel company ArcelorMittal to help the company identify defects in the steel coils they manufacture earlier in the manufacturing process. At present, these defects are discovered after manufacturing the coils, when they are inspected for quality and thickness, and the post-production discovery of defective coils has a significant economic impact. Being able to detect defective coils during the manufacturing process will enable earlier intervention, and ultimately, fewer defects and less economic and environmental waste.

PROTEUS will test whether it is possible to build a predictive analytics system using ArcelorMittal production data, and which variables are relevant for identifying manufacturing defects. The predictive analytics system will rely on new scalable online machine learning techniques. These will be developed by the project and tested in the ArcelorMittal industrial environment. The project will also develop visualization methods and tools to help manufacturers identify defects and respond to alerts.

The PROTEUS project will use the large volume and variety of data collected during the manufacturing process to build the predictive platform. The project will use sensor data measuring approximately 7440 variables across the production process. These will be combined with data related to the eventual thickness and quality of the steel coil. The processing of these data are essential to build the scalable online machine learning platform that will assist ArcelorMittal in augmenting their efficiency and competitiveness by identifying defective steel coils earlier in the manufacturing process.

3 Relevant policies

As mentioned in the introduction, there are a number of legal requirements and policies relevant to the management of PROTEUS data. First, the project falls within the European Commission's Open Data Research Pilot, which means that the project is obligated to consider providing open access to the research data utilised and generated by the project. In addition, intellectual property rights are relevant, as much of the data to be utilised within the project are owned by ArcelorMittal, are vital to the company's production process and directly relate to the company's competitive advantage. Finally, data protection laws in Europe, and in Spain where the research will primarily occur, are relevant as the research involves human participants and thus collection of some personal data. Hence, research ethics associated with human participants are also of paramount importance and good practice in this area will govern how the project manages the participation of these voluntary research subjects.

3.1 EC open research data pilot

The PROTEUS project is funded by the European Commission and responded to the ICT-16-2015 call on Big data - research. As it is part of the "Introduction to Leadership in Enabling and Industrial Technologies (LEIT): Information and communication technologies (ICT)" work programme, PROTEUS falls under the auspices of the European Commission's Open Research Data Pilot (ORD Pilot).

The ORD Pilot is an initiative by the European Commission (EC) to provide greater access to scientific information. The ORD Pilot specifically refers to H2020 projects, and is "designed to improve and maximise access to and reuse of research data generated by [those] projects" (European Commission, 2016, p.7). The Commission feels that providing open access to the data generated by publicly funded research will boost innovation, prevent duplication and provide transparency to members of the public (ibid.). As such, the H2020 programme within the EC is moving in concert with other public funding bodies. Finally, a provision to provide open access to the scientific publications emanating from projects is also part of the Grant Agreement signed by projects and the EC in H2020.

In addition, the Commission also recommends the use of data management plans to guide projects in how they should manage the research data used and generated by the project. This includes identifying any intellectual property issues, privacy or data protection issues or commercial sensitivity that might require data security restrictions that will prevent unauthorised sharing (OpenAIRE, 2016).

As part of a project's participation in the ORD Pilot, they must do three things (EC, 2016):

1. Evaluate the sensitivity of their data, including issues related to intellectual property rights, privacy and data protection and commercial sensitivity. This information is used to decide how the data should be collected, stored, processed and preserved.
2. Where possible, they should deposit their research data in an appropriate repository to ensure that it is adequately preserved.
3. Take measures to enable third parties to access and re-use this data, free of charge. This may include providing information about any tools necessary to take advantage of the data, including specialised software or algorithms. The EC recommends the use of Creative Commons or other open licenses to manage the re-use of data after it has been deposited and opened.

While the EC is committed to the ethical benefits of providing open access to research data used and generated by the project, they stress that the decision to provide open access must be taken after any intellectual property, security, privacy or other legal issues have been evaluated. As such, they offer an "opt-out" if the data used or produced by the project includes issues around confidentiality, personal data protection or security issues (EC, 2016). In addition, projects may also opt out if they plan to exploit the data produced within the project themselves.

PROTEUS partners have chosen to participate in the pilot and this data management plan underpins our data management activities within the pilot to maximise access to and reuse of research data and insights that can boost innovation and derived social benefits yet respect commercial sensitivities and ethical considerations that arise as part of its work. PROTEUS will use this data management plan to identify any intellectual property, privacy or data protection and commercial sensitivity issues related to the data utilised and

generated by the project. We will use this information to evaluate whether any of the project data may be made open access, or whether it needs to be protected because of the issues related to commercial sensitivity for ArcelorMittal. The following sections describe the additional issues related to privacy and data protection and ethical research practice that will further guide our decisions about managing and providing open access to PROTEUS data.

3.2 Intellectual property rights

Using data owned by ArcelorMittal will inevitably raise intellectual property rights, as will the creation of derived data resulting from the project itself. Intellectual property rights protect works by individuals that are the result of creativity, innovation, skill and specialist effort (Korn and Oppenheim, 2011). This section examines the intellectual property rights relevant to PROTEUS, including rights around trade secrets, for ArcelorMittal data, and copyright and database rights, for project-generated data. This information will be used to govern the extent to which this data can be made openly accessible after the close of the project.

3.2.1 Trade secrets

Trade secrets are a specific form of intellectual property rights that cover commercial data. In order to qualify for protection as a trade secret, the information must be:

- secret (i.e. it is not generally known among, or readily accessible to, circles that normally deal with the kind of information in question).
- have commercial value because it is a secret.
- subject to reasonable steps by the rightful holder of the information to keep it secret (e.g., through confidentiality agreements) (World Intellectual Property Organization, no date).

Trade secrets can be protected indefinitely and they are not subject to any registration or other formal requirements. Furthermore, they must be relevant within business and provide some form of competitive advantage to the holder.

The data held by ArcelorMittal and provided to PROTEUS are subject to protection as trade secrets. They relate specifically to the core manufacturing process of the specific products. This information is not known to other organisations and forms an important part of ArcelorMittal's competitive advantage. The data is protected within ArcelorMittal, and the company has consulted with their legal team to approve the limited sharing of this data with the PROTEUS consortium.

3.2.2 Copyright and database rights

Copyright is an automatically bestowed intellectual property right that protects the author or creator of a work and determines where, how and when the work may be made publicly available as well as how it may be used. While copyright is most often associated with scientific publications in the research arena, copyright might also be relevant to data where the creation of a data set represents a significant intellectual investment (de Vries, 2012). In Europe, the 1996 Database Directive protects the intellectual investment utilised to create data sets. While factual information is not subject to intellectual property legislation, building a collection of data is protected within this framework. The Directive prevents third parties from publishing, distributing and copying research data where the owners have claimed protection. These rights can be utilised to allow researchers to dictate how their data is utilised to allow them to exploit the data themselves, and to control how the data is utilised even where open access is provided. Previous research in this area has found a potential conflict between such database rights and open access mandates (Finn, et al., 2014). Nevertheless, the push by research funders and public bodies to support the exploitation of data by researchers themselves gives significant license to researchers to enable them to choose between exploitation expectations and open access expectations or to meet both in incremental steps (ibid.). Intellectual property owners can exercise these rights through licences, and the data management plan will assist partners to identify appropriate licenses given the constraints and necessary protections identified within this document.

This legal framework will allow PROTEUS to meaningfully consider how the consortium might exploit the data produced by the project and the extent to which PROTEUS can provide open access to the derived data resulting from the project. In the project Consortium Agreement, partners have already agreed that each party

owns the results, including the data, produced by them. Furthermore, in the case of joint ownership, each party has the right to exploit the results as they see fit. The purpose of this document is to assist partners in making decisions about exploiting the research data specifically. In addition it will also support the partners in identifying data that is not subject to protection and which might be candidate for open access sharing, as well as identifying the types of licenses that would be appropriate to enable sharing whilst protecting intellectual property and the protection of personal data while also adhering to ethical research principles.

3.3 Data protection law

3.3.1 European Data Protection Directive

In addition to the EC's open research data pilot, the project is also guided by policies on the protection of personal data both in Europe and in Spain, where the research with human subjects element of the project will be situated. Rights to privacy, as contained within the European Convention of Human Rights focus on "respect for private and family life, home and communications", which are less relevant for the research carried out by the PROTEUS project.

In Europe, the 1995 Data Protection Directive (95/46/EC) controls the protection of personal data, which is being replaced by the General Data Protection Regulation (GDPR) during the lifetime of the PROTEUS project. Both of these documents set out specific principles for the protection of personal data when data is being collected or processed. The personal data collected by PROTEUS will be very low risk from the perspective of data protection; nevertheless, the following data protection principles apply.

The Data Protection Directive outlines the following principles related to the protection of personal data. Personal data must be:

a) processed fairly and lawfully (*lawfulness and fairness principles*);
b) collected for specified, explicit and legitimate purposes and not further processed in a way incompatible with those purposes (*purpose limitation principle*);
c) adequate, relevant and not excessive in relation to the purposes for which they are collected and/or further processed (*proportionality and data minimisation principles*);
d) accurate and, where necessary, kept up to data (*data quality principle*);
e) kept in a form that permits identification of data subjects for no longer than is necessary for the purposes for which the data were collected or for which they are further processed (*retention principle*) (EC, 1995).

Under the DPD, PROTEUS must meet requirements (b)-(e) in order to meet requirement (a) that the data be processed fairly and lawfully. In addition, the processing must be transparent (i.e., data subjects must be able to understand how their data will be collected and processed) and they must consent to this processing. The consent must be "freely given, specific and informed" (EC, 1995). The obligations around transparency and consent are also included within the ethical research guidance on research with human subjects, and so there is significant overlap between data protection requirements and ethical research requirements.

In addition to these principles, the DPD also outlines particular rights enjoyed by data subjects with regard to the processing of their data. These include rights of access, correction and erasure. This means that data subjects should be able to see the data that is held about them, should be able to correct inaccuracies and should be able to request that their data be removed.

However, the Data Protection Directive only applies to data that is associated with an "identified or identifiable natural person" (Art. 29 WP, 2007). Data that is effectively anonymised is not subject to the Directive.

3.3.2 European General Data Protection Regulation

The Directive has been replaced by the General Data Protection Regulation (GDPR) in May 2016 and will enter into force in May 2018. In addition to the obligations set out in the DPD, the GDPR will also include the following protections and rights relevant to PROTEUS.

- Right to be forgotten – Under Art. 17 data subjects will have a right to obtain erasure from the data controller without undue delay. This means that PROTEUS research participants will have the right to have the record of their participation in the research deleted.
- Data protection by design and default – Data controllers must include appropriate provisions for anonymisation, pseudonymisation and data minimisation for all areas and stakeholders involved have been considered [IT (Chief Information-Security Officer, Legal (Legal Officer or Data Protection Officer), Business and Partners in the Project). While full anonymization is PROTEUS priority, to prove the "real or effective" participation in the project for the administrative justification of the Project and guarantee the scientific accuracy to eventual publications and the value of data in content release in open data environments (where this is agreed), the ability to re-identify user participants may be required. Hence, pseudonymization may be more appropriate in some cases. To minimize the probability of abuse or accidental misuse, the consortium will restrict the partners with access to pseudonymisation key and hence the ability to re-identify employees, and users, managers and their work council will be consulted prior to the presentation of findings to ensure that employee interests are not compromised.
- Data security – Data controllers must implement technical and organizational measures to ensure an appropriate level of security for data, including the use of pseudonymisation and encryption, ability to ensure appropriate confidentiality and resilience of systems, ability to provide access to data in a timely matter in the event of an incident and undertaking regular testing of the security of the system.
- Notification of data breaches – Authorities should be notified of any data breaches within 72 hours of their occurrence.
- Data Protection Impact Assessment (DPIA) – Data controllers should undertake a DPIA when using new technologies to collect, process or store personal data.
- Processing personal data for research purposes – Data controller and associates will ensure that data processing will respect the principles of Article 30 of GDPR, maintaining appropriate Record of processing activities, their purpose, relevant data and data handlers involved to ensure the confidentiality, integrity, availability and resilience of processing systems and services. In addition, data security safeguards (as per article 32 of GDPR) will be put in place to secure data from accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to personal data transmitted, stored or otherwise processed.
- Transparency and communication with data subjects – The consortium will ensure that all research participants will be informed prior to data collection and as part of the process of informed consent inherent in it methodology about the contact details of the data controller or third parties, if this applies, the purpose of their data processing, the recipients of processed data, forms and modalities of their communication (article 13 of GDPR) or trail of logs of not collected data (e.g. in the case of denial to provide data) as per article 14 of GDPR. Moreover, the consortium will ensure that information processing logs are kept accessible and is able to provide data subjects with clear and timely accounts of the trail of their data processing on demand as per the requirement of article 12 of GDPR.

PROTEUS will take each of these obligations into consideration when processing personal data to ensure the project complies with both current legislation and forthcoming legislation.

3.3.3 Spanish data protection law and regulatory requirements

Because the Data Protection Directive is a Directive, each European Member State was obligated to transpose the Directive via a national law. In Spain, the Data Protection Act (Law 15/1999 on the protection of personal data) implemented Directive 95/46/EC on data protection. This Act was further developed by a Regulation that was approved by Royal Decree 1720/2007 of 21 December (Data Protection Regulations). The following outlines the current specific requirements related to the protection of personal data in Spain.

Spanish data protection law requires data processing activities to be registered via the General Data Protection Registry. This can be accomplished via the Agencia Española de Protección de Datos' (AEPD, Spanish data protection authority) website. The registration must describe:

- The purpose of the data file

- The categories of personal data contained within the file
- Any intended data disclosures
- The security measures applied to the data
- Any intended international transfers (Practical Law, 2015)

Registration can be accomplished via this website:

https://www.agpd.es/portalwebAGPD/canalresponsable/inscripcion_ficheros/index-iden-idphp.php

In addition to registration, the Data Protection Act and the Data Protection Regulations introduce the following obligations for data controllers in relation to data collection and processing:

- Complying with the principles of data quality
- Informing data subjects about data processing on collection
- Obtaining data subjects' consent to process their data
- Registering personal data files
- Implementing security measures to protect personal data, including drafting a security document
- Attending to data subjects' rights of access, rectification, cancellation and opposition
- Entering into data processing agreements with data processors
- Keeping personal data confidential (ibid.)

In relation to consent, like the DPD, this must be free, unambiguous, specific and informed, and in Spain, data controllers must keep adequate records as evidence of this consent (Linklaters, 2015). In addition, informing data subjects about the data processing must include information about the following:

- The existence of a data file or data processing.
- The data controller's identity and address
- The purpose of the processing.
- The data recipients, identifying them by name and address and specifying the purpose of the data transfer.
- How the data subject can exercise his rights of access, rectification, cancellation and opposition.
- Whether answering the questions is mandatory or voluntary (unless the information can be clearly inferred from the nature of the personal data requested or the circumstances in which the data is collected).
- The consequences of providing the data or refusing to do so (unless the information can be clearly inferred from the nature of the personal data requested or the circumstances in which the data is collected). (Practical law, 2015)

Moreover, Law 37/2007 (article 8) determines the conditions to be imposed on the re-use of anonymized information in Open data contexts, stating that:

- Reusing the information of the Administrations and the public sector bodies referred to in article 2 of this Law may be subject, among others, to the following general conditions:
  (F) Where information, even if provided in a decoupled form, contains sufficient elements that would allow identification of stakeholders in the re-use process, a ban on reversing the dissociation procedure by adding new data obtained from other sources.

This rule attributes legal responsibility to those who obtain data from a public source process them, when they are personal, or perform some type of re-identification. Therefore, PROTEUS will include in Consortium Agreement, a clause regarding the prohibition or its commitment to non-reidentification either for Internal use of the data or for eventual publications.

As will be demonstrated below, many of these requirements overlap with requirements associated with ethical research practice with human volunteers. As such, PROTEUS will be able to meet each of these obligations via our informed consent mechanisms.

3.4 Ethical research guidance

Good practice in research with human volunteers has been the subject of much study in relation to social science as well as other disciplines. Disciplinary societies, such as the British Sociological Association, have published good practice guidance in relation to social science research with human volunteers, and PROTEUS will follow these guidelines very closely. The research with human volunteers that will be carried out within PROTEUS is most closely related to social science in that we are interested in how people interact with a specific technology product within an employment context. As such, PROTEUS will integrate aspects of sociology, science and technology studies and organisational studies, each of which fall under a larger interdisciplinary social science umbrella. Given this interdisciplinary social science framework, the consortium as selected the ethical research guidance produced by the British Sociological Association (BSA) to steer our research activities with human volunteers (BSA, 2002).

The BSA standards include a number of ethical research principles relevant to PROTEUS, and many of these overlap with data protection requirements. Specifically, the BSA ethical principles include an obligation to consider data protection law, human rights law and other relevant legal frameworks. Thus, data protection laws and other legal requirements are brought to the forefront of social research work.

These requirements include principles that link with data minimisation principles as well as obligations surrounding confidentiality, storage, security, sharing and rights of access, correction and erasure. With respect to data minimisation, BSA ethical standards explain that researchers should:

- consider whether specific types of data should be recorded, including sensitive data.
- keep personal information confidential,
- anonymise and pseudonymise the data as much as possible, and
- store data securely in an anticipation of threats to anonymity and confidentiality.

The standards also recognise that full anonymity cannot be absolutely guaranteed, and that participants should be given information to enable them to understand the limits of this anonymity. With respect to data storage and sharing, the guidance explains that researchers should provide information about

- how data will be stored,
- whether it will be shared with other researchers, and
- how it might possibly be used by those researchers.

In some cases, including providing open access to anonymised data, additional consent should be obtained. Finally, participants may be given rights to see copies of notes, transcripts and other research materials, and researchers should explain whether participants will be granted rights to alter or correct these. In each of these cases, the obligations surrounding research ethics and data protection are overlapping and PROTEUS will meet the requirements of both by relying upon the more stringent obligations.

In addition, research ethics and data protection also include principles of transparency and informed consent. The creation of an information sheet and informed consent form are central to the processes of both transparency and informed consent. Research ethics guidance asserts the following:

> Participation in sociological research should be based on the freely given informed consent of those studied. This implies a responsibility on the sociologist to explain in appropriate detail, and in terms meaningful to participants, what the research is about, who is undertaking and financing it, why it is being undertaken and how it is to be disseminated and used. (BSA, 2002, p. 3)

The informed consent form will accomplish the transparent provision of information about the research, the funding, the purpose of the project and how the results will be disseminated and used. It will also include information about how the data will be used by the project. Furthermore, providing this information will enable the participants to obtain informed consent, given that the information sheet is a transparency mechanism.

Finally, the BSA guidance discusses the other professional and moral responsibilities that researchers have in relation to research participants. The guidance requires researchers to consider their responsibilities to:

- Protect the interests of those involved in the research.

- Ensure, as far as possible, that the research process may be disturbing to participants and whether it may produce any unintended effects
- Ensure that participants are aware that they have a right to refuse to participate

In addition, researchers have a responsibility over the ways in which their data may be utilised and how their findings might be disseminated. These responsibilities are particularly relevant as PROTEUS research participants are employees of one of the partner organisations. Thus, their employer will have access to the research data held about them, and because research participants will be in an unequal power relationship with the researchers themselves. As such, issues around confidentiality, data security and data governance are particularly important in PROTEUS.

The following sections describe the current PROTEUS data governance considerations to enable the project to meet each of these policy, legal and ethical requirements. Specifically, the following chapters describe each of the data sets in more detail, examine our responses to the ethical and legal issues and look at issues related to data handling including access storage and sharing. Finally, the last chapters consider whether and how the consortium can exploit the project data as well as the extent to which PROTEUS data can be opened via open access provisions.

4 Data set description

This chapter outlines the specific characteristics of the three types of data that will be collected and processed by the project. The three data sets fall within the following categories:

1. ArcelorMittal data: ArcelorMittal data provided to the project
2. PROTEUS system data: Data generated by the project in the form of predictive analytics algorithms, and evaluation of their predictive accuracy against agreed KPIs.
3. PROTEUS evaluation data: User testing data of the visualisation aspects of the tool by ArcelorMittal employees

The first data set consists of real, anonymised, production data provided to the project by ArcelorMittal (AMIII), which has significant commercial sensitivity and thus is covered by intellectual property rights. This data will be used to develop and test the PROTEUS predictive analytics system. The second includes derived data about the functioning of the different versions of the system that will enable the project to identify improvements in performance, and assess the extent to which the PROTEUS system is scalable to other data sets and industrial contexts. The third data set will come from the piloting and assessing of the PROTEUS system. ArcelorMittal employees will use the visualisation dashboard to assess its utility and functionality. and their Their involvement is essential to the project, and as any human participant research, raises a number of potential ethical issues to be considered and addressed. In the sub-sections that follow we outline the methods for data collection that will be used for each of the three PROTEUS data categories, their characteristics and associated sensitivities.

4.1 Data collection and characteristics

This section provides a description of the characteristics of the different types of data that will be collected by the project. This includes information about where the data originated, the stakeholders for whom it could be useful and the uses to which it will likely be put within the project. It also includes information about the scale and volume of the data, its format and the extent to which the data might be interoperable with other data sets. While much of this information is available for the ArcelorMittal data, the details about the other two types of data that will be used within the project are less developed.

ArcelorMittal data
The ArcelorMittal data originates from the AMIII steel coil production process. Deliverable 2.1: *Scenario analysis and objectives description* gives a comprehensive overview of the steel production processes and a description of the data set relevant to PROTEUS. PROTEUS partners will use the data generated from the steel coil production process to identify variables relevant to and develop and test algorithms for detecting potential defects in the coils. As such, the data is useful to all of the consortium partners. However, because the production process has commercial sensitivity, the data that will be provided by AMIII might also be of interest to customers and competitors. Because of this sensitivity, the data provided by AMII to the project is anonymised – it has no customer data and the variables are difficult to interpret by external parties.

The ArcelorMittal data is made up of two datasets: the coil production data set and the flatness data set. These data sets comprise complex, large and high velocity data. In particular, the coil production data consists on different measures that are made available during the production process at different moments in time:
Time-series: these data are produced and available in real time during the coil production. It consists of **53 different variables** that are measured and made available in **real time** by different sensors deployed in the actual production environment. These variables present a **variety of formats**, combining both 1D and 2D information, and are produced at **different rates** (depending on the specific variable) **ranging from 50 milliseconds to around 1 second period**.
HSM: historical and final production coil data (known as HMS data). It provides more than **7,000 variables** as aggregated information per coil at the end of the production process, thus this is **available only once the production has finalised for one particular coil**.

The flatness dataset: measures of the flatness of the resulting coil. It consists of **3 variables** (2 of them as 1D flatness, and an additional 2D flatness measure) following a format that is compliant with the one used in *time-series*, and synchronised with it by the spatial information (x, y). This information is measured and **available only after the coil has been fully produced, and after a certain delay** due to the production environment setup and infrastructure. As previously introduced and also in previous deliverable D2.1, this information **represents one of the main targets for the real-time learning and predictive processes** to be enabled by PROTEUS, since the capability to predict flatness in real time (without the need to wait until the process has finalised to get the actual measures) would be key to improve business operations.

The data generated by the sensors is both qualitative and quantitative, and stores approximately 7440 variables relevant to the production process. The data dates from 2010 and thus there are approximately 840,000 records for each variable. The types of measurements included are temperature, vibration intensity, tension in the rollers, speed of the plate when entering the roller and surface flatness. With respect to surface flatness, the generation rate for variables is approximately 1 per second, this requires significant computing capacity to manage this high-velocity data set. In addition, while the process data set to be shared includes historical records for each variable, these data are also generated in real time during the production process at a rate of between 32 and 500 milliseconds, depending on the specific measure. The data controller, Treelogic, will use historical data from both data sets to emulate real-time data that can be used to test the algorithms.

As noted above, this raw data has been shared in an anonymised form, so any sensitive information is excluded This analysed data set is comprised of both quantitative and qualitative data that are stored across 42 different tables. All of the tables share the same key variable, the coil identifier, which allows AMIII to join and relate information for each coil across multiple tables. The size of each of the 42 data tables is approximately 300-700MB. This data will provide some context to the initial variables that are thought to be relevant to PROTEUS.

PROTEUS predictive analytics system data
The data generated by the PROTEUS predictive analytics system and about its functioning will likely consist of algorithms that provide predictions and statistical comparisons about the predictive capacity of different algorithms and data related to the technical evaluation of its functioning using benchmarks and KPIs developed within the project. This data will be useful for the PROTEUS research and development team to optimise the PROTEUS scalable online machine-learning tool. It will also be useful for the research and development team to undertake an analysis of the impact and scalability of the PROTEUS toolset. The idea here will be to evaluate how the PROTEUS system benchmarks against other available tools like Flink. The system consists primarily of data about the functioning of the statistical tools and online machine learning libraries developed and tested within the project. Prototypes of these tools are currently in development, hence fewer details about available at this juncture. The characteristics and collection details associated with this type of data will be further developed during the course of Y2 of the project, and further details about this will be included within the next iteration of this deliverable.

PROTEUS user evaluation data
Data generated during the evaluation phase of the research involves human subjects and may include the collection of personal data. The purpose of the evaluation exercise is to assess whether this solution is fit for purpose from the point of view of industrial operators. Research participants will be presented with a series of visualisations generated by PROTEUS predictive analytics system and be requested to evaluate these visualisations with respect to their utility, reliability and validity for identifying defects. They may also be asked to consider what features they would like and how well this might be integrated into their current operations workflow. Given that they will be asked to comment on and critique current and planned workflow processes to be initiated by their employer, the research may put some participants in a potentially difficult situation. Essentially, they will be asked to indirectly critique their current workflow and work process and critique the new system. As their employer is a participant in the project and driving the project, there is a low risk that some employees may experience adverse consequences from stating their opinions. As such, the consortium plans a number of protections to ensure that research participants are not adversely affected by their participation and that their participation is fully voluntary.

Research participants will likely include employees of ArcelorMittal with, at least some, responsibility for quality control. Evaluation data will likely be collected via focus groups or workshops although the project will also consider using interviews or a survey depending on the best way to address the sensitivities raised above and any, currently unforeseen, employee concern. Any decision about the methodology, specific location and the participant pool from which volunteers will be sought will be made later in the project, in cooperation with the manager in charge of quality at a relevant location. This information is presented in greater detail in Chapter 5 and Annex A.

4.1.1 Personal data collection

The project will aim to collect as little personal data as possible, from human participants. Nevertheless, requirements around ethical research practice necessitate the collection of some personal data to manage informed consent processes and to re-connect with participants who may have questions about PROTEUS or who may wish to exercise their rights of access, correction and erasure. The personal information will likely be limited to the names and contact details of the volunteers as filled out on their informed consent sheets. After the informed consent process, the participants will be pseudonymised or anonymised as described in Chapter 5, and the link between their pseudonyms and their personal information will be held securely by the project coordinator as described in Chapter 6.

5 Ethical and legal issues

The types of data described above raise specific issues related to intellectual property, data protection and research ethics that the project will have to manage appropriately. Where relevant, Spanish law is considered alongside the European law, as Spain is the primary research location where data collection and processing activities are taking place. The following sections outline how PROTEUS will manage each of the relevant legal requirements, and describe the agreed data governance processes around access storage and sharing. Consequently, this section makes consistent reference to the material to be discussed in Chapter 6: Data Governance. This chapter addresses the management of inform consent, personal data protection, and intellectual property rights.

5.1 Informed consent

Issues related to research ethics can largely be addressed though the management of informed consent when the research is being conducted with healthy, adult volunteers. However, as the participants are employees of a partner organisation, there are some risks that they may feel pressured to participate. This risk will also be managed though the informed consent process.

Informed consent is central to ethical research practice, as adult healthy volunteers should be empowered to manage their participation and the use of their information during social science research. The consortium will provide participants adequate information about the purpose of the research, the data that will be collected, the research funders, the ways in which their data will be utilised and who will benefit from the research to ensure that participants understand the potential implications of their participation. An **information sheet** will provide this information in appropriate detail and in language that is meaningful to the participant. (See Annex A for the draft information sheet). It also sets out information about:

- how their data will be anonymised or pseudonymised
- how their data will be stored and shared with other researchers
- how participants may access the data they provided
- whether they can make corrections
- how they can request their data be removed, and
- where they can go if they have any questions, comments or complaints.

In addition, the information sheet explains any unintended effects that may result from the research. Combining each of these pieces of information will enable potential participants to evaluate whether they would like to participate in the research and whether they might experience any unintended or adverse effects.

The consortium will take steps not only to ensure that potential participants are well informed but that they also feel empowered to volunteer or decline participation as they see fit. Given that this research will be carried out with employees of one of the partner organisations, ensuring voluntary participation is paramount and will require a few additional steps. First, following good practice, the information sheet will advise participants that their participation is purely voluntary, and partner AMIII has confirmed this. In addition, non-AMIII personnel will undertake recruitment and advise participants that their participation is voluntary. Finally, during the research activity itself, those conducting the research will invite participants to re-consider their participation and to excuse themselves from the research activity. Given that the project does not involve many sensitive topics, this should be sufficient to ensure voluntary participation, but the consortium remains dedicated to addressing any further concerns raised by employees in the course of the project. Hence, the project will remain vigilant about potential conflicts and will carry out a rolling risk assessment to ensure voluntary participation. To date, an informed consent form has been drafted for future reference and appended in Annex A herein.

Moreover, given that human participants are also employees of AccelorMittal, the consortium will also consider compliance with legislation, guidelines and recommendations regarding *jurisprudence* in

accordance with the Spanish Data Protection Agency and the Working Party[1] to ensure that employee rights prevail as per the law, and guarantee that such rights cannot circumvented by the empoloyer under article 20 of the workers' statute2. To this end, the consortium will make clear to both managers and employees of AccelorMittal that the survey does not form part of the ordinary work experience or that the conditions of realization do not affect their job. It will explicitly highlight the freedom of choice of the worker, the absence of adverse consequences and the possibility to withdraw from the Project at any point by approaching the appointed Data Protection Officer, responsible for guaranteeing their rights. In addition, the consortium will research the possibility of informing and consulting the relevant Works council about the project and give them oversight rights, in accordance with articles 63, 64, and 65.2, of Spain's Workers' Statute to ensure that workers' interests are not affected.

5.2 Personal data protection

However, seeking informed consent will raise issues around the protection of personal data, as personal data, including names and contact information will be needed to record informed consent. The consortium will manage this using the following steps:

1. Participants will immediately be given a participant number linked to their name, and this will replace their name in any stored or shared data.
2. The link between a participant's name and number will be stored in a proprietary storage facility by the project coordinator
3. This information will not be shared with the project partners, and any enquiries about participants' personal information will be fed through the coordinator
4. Participants with particular identifying features or experiences may be managed by mixing these with other participants' characteristics (e.g., switching places of birth) to make each participant less identifiable. Where necessary, some identifying features may be removed from the data if it cannot be anonymised
5. Participants will be given the right to review their data and make any corrections or erasures should they have any concerns.

The project will also avoid the collection of data that is not necessary for the purposes of the research (*purpose limitation and data minimisation principles*). Each of these processes will assist in the anonymisation and pseudonymisation of any personal data, and storing this data with the coordinator will ensure that participants are adequately protected with reference to confidentiality. In addition, the information sheet will enable the project to meet requirements around transparency and provide a mechanism through which participants can exercise their rights of access, correction and erasure. The information sheet will also assist the project in meeting requirements around data retention, as the information sheet sets out how long the data will be stored and with whom it may be shared.

In addition to these, the project will also meet Spanish data protection requirements. The project coordinator will register with the AEPD as a processor of personal data. In addition, should a data breach occur, the coordinator will inform both the AEPD and research participants about the breach and provide advice on any consequences.

Thus, the overlapping requirements around ethical research practice and the protection of personal data can be met simultaneously using both the information sheet and informed consent forms for the PROTEUS

[1] SSTC 292/2000, 241/2012, 29/2013, 170/2013, 39/2016.Bărbulescu V. Romania (Application no. 61496/08)Working document on the surveillance of electronic communications in the workplace. (29 May 2002). Working Party Article 29. Opinion 15/2011 on the definition of consent. Working document on the surveillance of electronic communications in the workplace. (29 May 2002). Spanish Data Protection Agency. Data Protection in Labour Relationships.

2 Article 20. Management and control of the labour activity.(…)

3. The employer may adopt the measures he believes more convenient for supervision and control to verify the compliance by the worker of his duties and obligations, keeping in its adoption and application the consideration of his dignity and taking into account, where appropriate, the actual capacity of workers with disabilities.

research. While the amount of personal data that will be collected by the project is relatively minimal, the project will use the data protection principles to guide the collection of all data about human participants, whether personal or not, to ensure that we meet ethical research requirements. Attention to both will ensure participants receive the maximum level of protection and consideration.

5.3 Intellectual property rights

As noted above, the ArcelorMittal corporate data is subject to intellectual property protections, and the consortium will take the following specific steps to address these. First, as noted above, the data will be anonymised so that any sensitive data associated with AMIII customers is removed. Second, the data governance procedures around access, storage and sharing, discussed in Chapter 6 below, will ensure that consortium members respect AMIII's intellectual property rights. Finally, each of the partners have agreed to only use the data for the purposes of the PROTEUS project and the development and testing of PROTEUS algorithms and software. This has been agreed via the PROTEUS Consortium Agreement, a legally binding document that governs the project and partnership arrangements.

The Consortium Agreement and this document will also guide the intellectual property rights claimed by the consortium with respect to PROTEUS toolset data. The consortium will agree a license that adequately describes how the data will be used and shared within the consortium and at the close of the project. Underpinning this will be the agreement, contained within the Consortium Agreement, that each partner owns the intellectual property, including data, which they create. Nevertheless, the Consortium Agreement also provides for joint or multiple ownership, and in these cases, relevant partners will agree on the license to be used. Consideration of these intellectual property rights will also govern the extent to which PROTEUS toolset data can be made openly accessible at the close of the project. If this option is selected, partners will agree an open license to manage the use of this data, and will likely select a license such as CC-BY – a creative commons license that requires users to attribute the data to those who originally created it. The outcome of these discussions will feed into the PROTEUS intellectual property rights and innovation committee that will undertake the final decision regarding licensing. This issue will be re-visited in the next iteration of this plan.

6 Data governance

This section outlines the rules for the governance of PROTEUS data. The information here relies upon information provided by consortium partners, as well as governance mechanisms agreed in the Consortium Agreement. Each of the sub-chapters that follow outline the agreed rules about how data will be accessed, stored and shared.

6.1 Access

With respect to the ArcelorMittal data provided to PROTEUS, the partners have agreed that the project coordinator will manage access to the data. AMIII have agreed to allow access to the required historical and simulated real-time data in an anonymised format, provided that data is only accessed by consortium partners and only for project activities. Partners have agreed not to seek access to raw, non anonymised, data. Anonymised AMIII data will be provided directly to the coordinator, who will store the data in their existing ICT infrastructure and enable futher access by consortium partners. Any partner with a user password will be able to access the data to develop or test their algorithm or software. Data generated by the final testing of the PROTEUS solution will be stored within AMIII facilities. Partners will maintain access only rights during testing, but will not be able to store such data. Unauthorised access to this data will be prevented via Treelogic's existing data and information security mechanisms, which meet existing information security standards.

Access to PROTEUS predictive analytics system and PROTEUS evaluation data will be restricted to the consortium during the course of the project. Near the end of the project, the consortium will begin to consider the extent to which this data can be made open access, as well as the optimal licensing framework that should be used to govern the use of this data. This decision is dependent on the eventual characteristics of this data, which will become clearer as the project itself develops. As a starting point, the algorithms and system evaluation data will be held by the partners participating in system development and evaluation, and shared among the consortium for the purposes of carrying out the work described in the Description of Action. Access to the PROTEUS system and evaluation data after the close of the project will be initially considered in the last iteration of this document in M36, where the commercial implication for ArcelorMittal and for partners are clearly and comprehensivley understood.

6.2 Storage and processing

With respect to storage, each of the three types of data will be stored and backed-up slightly differently. Treelogic will store the ArcelorMittal provided data in their existing infrastructure. All of this data will be backed-up in the cluster, automatically, so that the data can be recovered in the event of an incident. The security of this data will be maintained via the security policies and mechanisms that Treelogic already has in place for protecting their sensitive commercial data. These follow Treelogic's existing data and information security standards which comply with the ISO/IEC 27001:2013 for which the company has been certified by IQNet and AENOR (See Annex B - Treelogic ISO/IEC 27001:2013 Certification). In addition, for personal data relating to user evaluation that are kept and processed in the Cloud security standard ISO / IEC 27017: 2015, Information technology, Security techniques, Code of practice for information security controls based on ISO / IEC 27002 for cloud services. Information is stored in the cluster using the Hadoop Distributed File Systems (HDFS). Concretely, Treelogic uses the .20.20x distributions of Hadoop which focus on security issues by utilizing the following:

- Mutual Authentication with Kerberos RPC (SASL/GSSAPI) on RPC connections**:** SASL/GSSAPI was used to implement Kerberos and mutually authenticate users, their processes, and Hadoop services on RPC connections.
- "Pluggable" Authentication for HTTP Web Consoles: meaning that implementers of web applications and web consoles could implement their own authentication mechanism for HTTP connections. This could include (but was not limited to) HTTP SPNEGO authentication.

- *Enforcement of HDFS file permissions:* Access control to files in HDFS could be enforced by the NameNode based on file permissions - Access Control Lists (ACLs) of users and groups.
- Delegation Tokens for Subsequent Authentication checks*:* these were used between the various clients and services after their initial authentication in order to reduce the performance overhead and load on the Kerberos KDC after the initial user authentication. Specifically, *delegation tokens* are used in communication with the NameNode for subsequent authenticated access without using the Kerberos Servers.
- Block Access Tokens for Access Control to Data Block**:** when access to data blocks were needed, the NameNode would make an access control decision based on HDFS file permissions and would issue *Block access tokens (using HMAC-SHA1)* that could be sent to the DataNode for block access requests. Because DataNodes have no concept of files or permissions, this was necessary to make the connection between the HDFS permissions and access to the blocks of data.
- Job Tokens to Enforce Task Authorization*:* Job tokens are created by the JobTracker and passed onto TaskTrackers, ensuring that Tasks could only do work on the jobs that they are assigned. Tasks could also be configured to run as the user submitting the job, making access control checks simpler.
- From "Pluggable Authentication" to HTTP SPNEGO Authentication: Although the 2009 security design of Hadoop focused on pluggable authentication, the Hadoop developer community decided that it would be better to use Kerberos consistently, since Kerberos authentication was already being used for RPC connections (users, applications, and Hadoop services). Now, Hadoop web consoles are configured to use HTTP SPNEGO Authentication, an implementation of Kerberos for web consoles. This provided some much-needed consistency.
- Network Encryption*:* Connections utilizing SASL can be configured to use a Quality of Protection (QoP) of confidential, enforcing encryption at the network level – this includes connections using Kerberos RPC and subsequent authentication using delegation tokens. Web consoles and MapReduce shuffle operations can be encrypted by configuring them to use SSL. Finally, HDFS File Transfer can also be configured for encryption

Treelogic will use an Apache Kafka end point to provide test data to the partners. The 0.9.x release used by the PROTEUS Project includes a number of features that, whether used separately or together, will increase security in a Kafka cluster. This includes the following security measures:

- Authentication of connections to brokers from clients (producers and consumers), other brokers and tools, using either SSL or SASL (Kerberos)
- Authentication of connections from brokers to ZooKeeper
- Encryption of data transferred between brokers and clients, between brokers or between brokers and tools using SSL (However, there is a performance degradation when SSL is enabled, and the magnitude of this degradation depends on the CPI type and the JVM implementation utilized.
- Authorisation of read/write operations by clients
- Authorisation is pluggable and integration with external authorisation services is supported (Apache Kafka, 2016)

PROTEUS toolset data and anonymised evaluation data will be stored by individual partners and in the consortium's file repository that is managed by Treelogic. The sharing of this data within the consortium will create back-ups should an incident occur, however, like the AMIII data, storing this data within Treelogic's file repository would trigger automated back-ups and enable recovery. Finally, Treelogic will store the personal data associated with the informed consent within a separate, but equally secure, storage space that is not accessible to project partners to protect the personal data of those participating in the project. Treelogic's existing data and information security protocols and tools will also protect this data.

6.3 Sharing

At present, the data will not be shared outside the consortium. However, this will be reconsidered as the specifics of the PROTEUS system and the PROTEUS evaluation data develop and the implication for ArcelorMittal and consortium partners is clearly understood.

7 Standards and metadata

This section considers the standards that PROTEUS will use to represent data generated by the project, and the additional standards around data security, etc. that might be useful to govern the data used within and generated by the project. In addition, this section also includes a consideration and selection of the metadata that will be most effective in describing the PROTEUS data set.

This includes a consideration and evaluation of existing standards and our reasoning for selecting specific standards. At present, the consortium has agreed to avoid proprietary data formats as far as possible, as these will make it difficult for both the consortium and any external stakeholders to utilise the PROTEUS data after the close of the project. This is for three specific reasons – first, because proprietary programmes evolve and data formats may become defunct, thus maintaining proprietary data formats represent a significant needed investment to keep the data relevant and accessible. Second, because access to the data would be restricted to those who have access to the appropriate analysis tools if a proprietary data format was utilised. Third, because it would be difficult to combine PROTEUS data with other data as proprietary data formats often raise interoperability issues. Thus, when designing PROTEUS data generation and collection methodologies, the project will consider each of these issues.

Given the proprietary nature and commercial sensitivity of utilised data and the requirements of the project for a realistic simulation environment on which the PROTEUS prototype can be test, a concerted effort between ArcelorMittal and Treelogic is being undertaken to jointly curate the data and ensure that there is a clear and accurate understanding of what the data categories represent amongst data partners.

In addition to these, PROTEUS will also consider standards around other issues that could govern the storage and representation of PROTEUS data. This includes data security standards such as ISO 27001. As the PROTEUS data management plan develops alongside the project, partners will consider each of these relevant standards and make an informed selection for PROTEUS system and evaluation data.

Finally, the project will also consider and select effective metadata for describing the PROTEUS system and evaluation data. Effective metadata will assist project partners and potential additional data users by providing “clear and detailed data descriptions and annotation”, accessibly written accompanying documentation and any contextual information that is relevant when the data is re-used (UK Data Service, 2016). This consideration of metadata is linked to the next section on data exploitation, in that the metadata provided should consider the uses to which the data can be put in order to provide sufficient and relevant information to potential users.

8 Data exploitation

This section will manage the exploitation of PROTEUS data by project partners and external stakeholders for additional commercial or research purposes. It will identify the extent to which the intellectual property issues discussed above can facilitate such exploitation and define which partners have rights to exploit PROTEUS data.

As the project develops, the PROTEUS consortium will use this document to outline potential ways in which the project partners could exploit the data used and generated by PROTEUS. This includes two potential streams of activity. First, the project will consider how the data generated within PROTEUS, via the PROTEUS system development and evaluation, could be further exploited. One suggested avenue for this is to use the data to demonstrate the value of PROTEUS to enable its exploitation by additional data scientists and potential customers within industry or other fields. This is expected to generate value for the project partners, both in terms of reputation enhancement and in terms of potential commercialisation of the tools and/or components of the tools. As part of this discussion, partners will evaluate the different licenses that could be used to manage the utilisation of PROTEUS data.

The second stream of exploitation activity will include a consideration of any additional added value that PROTEUS partners may be able to provide to ArcelorMittal, given partners' familiarity with the AMIII production process and data set developed during the course of the project. This includes a consideration of how AMIII data might be better curated to enable the company to identify additional insights, efficiencies or interactions. It may also include the development of new tools and services to optimise the production process or make use of the data in unforeseen ways. For example, this manufacturing data could be combined with earth sciences data to better manage environmental impacts. The data management plan will use this PROTEUS collaboration as an opportunity to assist ArcelorMittal to transform their data into insights and actionable intelligence to benefit additional business areas within the company.

The final version will include a more comprehensive consideration of each of these issues in concert with Deliverable 6.1: *PROTEUS business plan* and D6.2: *PROTEUS evaluation and impact assessment*.

9 Long-term archiving and preservation (including open access)

PROTEUS partners will use this section of the data management plan to outline a strategy for long term preservation of PROTEUS data beyond the close of the project. A consideration of these issues needs to take place alongside the planning of the research process for generating PROTEUS toolset and evaluation data, and this section will be updated to reflect these developments. As a guideline, this section will describe the processes and procedures that will be put into place to guide the long-term preservation of the data. This includes an indication of how long the data might be preserved, its approximate volume and characteristics as well as information about how the veracity of the data will be ensured. The project will evaluate where the data should be stored, including evaluating different repositories, and who might be able to access it. Central to this chapter will be an evaluation of whether the data can be made openly accessible in line with PROTEUS' participation in the EC open research data pilot. This chapter will also consider the costs associated with preparing the data and arranging its preservation, and deliver a strategy for how these costs are going to be covered.

10 Conclusion

This deliverable represents the first iteration of the PROTEUS data management plan. It provides foundational and contextual information related to the project, the relevant policies and legal frameworks and the initial plans for ensuring that the collection and use of data within the project conforms to the issues within this larger context. Specifically, the document examines the European Commission's Open Data Research Pilot, intellectual property issues relevant to PROTEUS, data protection standards and practices and guidance on ethical research processes. The deliverable progresses by describing the data that will be used and generated by PROTEUS and outlining how the project plans to meet all of our legal, ethical and policy obligations surrounding PROTEUS data. In doing so, the project has agreed the following principles:

- Data owned by ArcelorMittal will be shared with consortium members, although consortium members can only access this data through the project coordinator
- PROTEUS partners agree to respect the commercial confidence of the data provided by AMIII
- Human participants in PROTEUS research activities are under no obligation to participate and their involvement will be strictly voluntary
- PROTEUS partners will respect data protection and ethical research principles related to the following:
    - Participant confidentiality and anonymisation
    - Informed consent
    - Data minimisation
    - Purpose limitation
    - Transparency
    - Rights of access, correction and erasure
- The coordinator will hold all commercially sensitive data and personal data and will manage access to this data for PROTEUS partners

In addition to meeting European regulations, the project will also meet legal regulations set by the Spanish government as the data and many of the research activities will be located in Spain. This includes the registration of the project coordinator as a data controller with the Spanish data protection authority and the provision of specific information required by Spanish data protection legislation. Drafts of the PROTEUS information sheet and informed consent forms are included below to demonstrate how we will meet these obligations.

Future iterations of this document will leverage the development of the project and the research process, in general, to outline more specific information about the data characteristics and governance of the data generated by the project. It will also include more information about how the project intends to exploit, store and share the data moving forward. These issues will be outlined in detail in the third iteration of the document, which will be submitted in M36 of the project.

References

[1] Wessels, Bridgette, Rachel L. Finn, Peter Linde, Paolo Mazzetti, Stefano Nativi, Susan Riley, Rod Smallwood, Mark J. Taylor, Victoria Tsoukala, Kush Wadhwa and Sally Wyatt, "Issues in the development of open access to research data", *Prometheus: Critical Studies in Innovation,* Vol. 32, Issue 1, 2014, pp. 49-66.

[2] European Commission, *Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020*, Version 2.1, 15 February 2016. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

[3] Jones, Sarah. (2011). 'How to Develop a Data Management and Sharing Plan'. DCC How-to Guides. Edinburgh: Digital Curation Centre. Available online: http://www.dcc.ac.uk/resources/how-guides

[4] Data Curation Centre, "Data management plans", 2016. http://www.dcc.ac.uk/resources/data-management-plans

[5] OpenAIRE, "What is the Open Research Data Pilot?", 22 Feb 2016. https://www.openaire.eu/opendatapilot

[6] Korn, Naomi, and Charles Oppenheim, *Licensing Open Data: A Practical Guide*, June 2011 version 2.0. http://discovery.ac.uk/files/pdf/Licensing_Open_Data_A_Practical_Guide.pdf

[7] World Intellectual Property Organization, "How are trade secrets protected?", no date. http://www.wipo.int/sme/en/ip_business/trade_secrets/protection.htm

[8] Marc de Vries, *Open Data and Liability,* European Public Sector Information Platform Topic Report No. 2012 / 13, December 2012.

[9] Finn, Rachel, Kush Wadhwa, Mark Taylor, Thordis Sveinsdottir, Merel Noorman and Jeroen Sondervan, *Legal and ethical issues in open access and data dissemination and preservation*, RECODE project Deliverable 3.1, 30 April 2014.

[10] European Parliament and the Council, Directive 95/46/EC of 24.10.1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, OJ L 281, 23.11.1995.

[11] Article 29 Data Protection Working Party, Opinion 4/2007 on the concept of personal data, Brussels, 20 June 2007, http://ec.europa.eu/justice/policies/privacy/docs/wpdocs/2007/wp136_en.pdf ("A29WP Opinion 4/2007").

[12] Practical Law, Data protection in Spain: overview, 1 Dec 2015. http://uk.practicallaw.com/1-520-8264#

[13] Linklaters, "Data Protected: Spain", July 2015. https://clientsites.linklaters.com/Clients/dataprotected/Pages/Spain.aspx

[14] British Sociological Association, *Statement of ethical practice for the British Sociological Association*, British Sociological Association, Durham UK, March 2002.

[15] Apache Kafka, "7.1 Security Overview", 2016. https://kafka.apache.org/090/security.html

[16] UK Data Service, "Data management planning for ESRC researchers", 2016. https://www.ukdataservice.ac.uk/manage-data/plan/dmp-esrc

Annex A – PROTEUS informed consent and information sheets

PROTEUS Information sheet

PROTEUS
Scalable online machine learning for predictive analytics and real-time interactive visualization

The PROTEUS research project[3] conducted by NAME of the INSTITUTION is funded by the European Union[4]. The purpose of the research is to evaluate a new system for identifying defective steel coils sooner in the production process. The evaluation exercise is based on the tool developed within the project that will provide alerts and information about the quality of the steel coils. This research will help ArcelorMittal and other large companies make better use of the data they have about the manufacturing process and cut down on waste and defective products. The research will also help companies developing new software and analysis tools by providing a real world data challenge. I will contribute to this software by evaluating its usefulness and user-friendliness. My role is to help the researchers test how easy the system is to use from operatives' point of view, and tell them whether I understand the alerts and information I see, so they can assess and where necessary refine the design of the system to make it useful and user-friendly.

This research will involve a focus group / workshop / consultation lasting 1-2 hours where I will be invited to give feedback about the tool, discuss my understanding of what I see and provide feedback about how this might be integrated into the manufacturing process. Participation to this project is ***not*** an employment obligation and it is entirely in my discretion to participate. Participating in this research is entirely voluntarily and I am free to leave at any time, without any impacts on my employment. I am also aware that I am free to refuse to answer any questions that I feel are commercially or institutionally sensitive or relate to topics that I do not wish to discuss. I understand that I have the right to ask questions and receive understandable answers before making any decision.

I understand that I will only be asked to provide professional, not personal, information, and the record of my involvement in the research will be kept confidential. I have been informed that everything I say will be anonymous. The event data will be recorded via paper notes/tape recorder and I understand that I can request a copy of the notes/transcript to review, if I wish, as well as a log of how my data has been processed, stored, interpreted and communicated between parties. I understand that I am also allowed to delete or make any changes to the notes/transcript if I feel the information I provided could be improved or clarified. I understand that a record about my personal involvement in the research will be kept in a secure location at Treelogic, solely for the purpose of double checking the accuracy of data and for proving to funding and other authorities that the evaluation exercise was indeed carried out. I understand that the data I provide will be anonymised and the record of my participation will be kept in a file separate from the research data to ensure my anonymity. All other parties and researchers within PROTEUS will work with information in an anonymised format. I understand that the data will be kept for one year after the PROTEUS project ends, but my personal data will be destroyed when the project ends.

[3] http://proteus-bigdata.com/

[4] Grant agreement number 687691

I understand that this research conforms to European Commission guidelines and that it has been approved by the Ethics Committee in the Research Executive Agency managing Horizon 2020 projects. Finally, I have been given the contact details of the research team and I have been informed that I am free to contact Marcos Sacristan (Project Coordinator and acting Data Protection Officer) about any queries relating to my data or the project itself.

Marcos Sacristan
Treelogic
T: +34 910 05 90 88
marcos.sacristan@treelogic.com

PROTEUS Informed consent sheet

PROTEUS

Scalable online machine learning for predictive analytics and real-time interactive visualization

Lead researchers: Insert name of person/institution conducting the research activity

Participant Identification Number for this project: **Please initial box**

1. I confirm that I have read and understand the information sheet/letter (delete as applicable) dated *[insert date]* explaining the above research project and I have had the opportunity to ask questions about the project. ☐

2. I understand that my participation is purely voluntary and that I am free to withd at any time without giving any reason and without there being any negative consequences. In addition, should I not wish to answer any particular question or questions, I am free to decline and can contact *Project Coordinator Marcos Sacristan* via telephone or e-mail at +34 910 05 90 88 or marcos.sacristan@treelogic.com ☐

3. I understand that my responses will be kept strictly confidential. ☐

4. I give permission for members of the research team to have access to my anonymised responses. I understand that my name will not be linked with the research materials, and I will not be identified or identifiable in the report or reports that result from the research. ☐

5. I agree to take part in the above research project. ☐

_____	_____	_____
Name of Participant (*or legal representative*)	Date	Signature
_____	_____	_____
Name of person taking consent (*if different from lead researcher*)	Date	Signature
_____	_____	_____
Lead Researcher	Date	Signature

To be signed and dated in presence of the participant

Copies:

Once this has been signed by all parties the participant should receive a copy of the signed and dated

participant consent form, the letter/pre-written script/information sheet and any other written information provided to the participants. A copy of the signed and dated consent form should be placed in the project's main record (e.g. a site file), which must be kept in a secure location.

Annex B- Treelogic ISO/IEC 27001:2013 Certification

CERTIFICATE

IQNet and
AENOR
hereby certify that the organization

TREELOGIC TELEMÁTICA Y LÓGICA RACIONAL PARA LA EMPRESA EUROPEA, S.L.

PQ PARQUE TECNOLÓGICO DE ASTURIAS, 30.
33428 - LLANERA
(ASTURIAS)

AV MANOTERAS, 38 OFICINA D 614.
28050 - MADRID

for the following field of activities

The Information systems that supports the processes: management, commercial and Information Tecnology consulting (design, development, implementation, integration and maintenance of web and communication solutions, R&D projects, Smart Mobility). according to the current applicability statement.

has implemented and maintains a

Information Security Management System

which fulfills the requirements of the following standard

ISO/IEC 27001:2013

First issued on: 2017-04-20 Validity date: 2020-04-20

Registration Number: **ES-SI-0015/2017**

AENOR

Michael Drechsel
President of IQNet

Avelino BRITO
General Manager

Original Electronic Certificate

IQNet Partners*:
AENOR *Spain* AFNOR Certification *France* AIB-Vincotte International *Belgium* ANCE *Mexico* APCER *Portugal* CCC *Cyprus*
CISQ *Italy* CQC *China* CQM *China* CQS *Czech Republic* Cro Cert *Croatia* DQS Holding GmbH *Germany*
FCAV *Brazil* FONDONORMA *Venezuela* ICONTEC *Colombia* IMNC *Mexico* Inspecta Certification *Finland* IRAM *Argentina*
JQA *Japan* KFQ *Korea* MIRTEC *Greece* MSZT *Hungary* Nemko AS *Norway* NSAI *Ireland* PCBC *Poland*
Quality Austria *Austria* RR *Russia* SII *Israel* SIQ *Slovenia* SIRIM QAS International *Malaysia*
SQS *Switzerland* SRAC *Romania* TEST St Petersburg *Russia* TSE *Turkey* YUQS *Serbia*
IQNet is represented in the USA by: AFNOR Certification, CISQ, DQS Holding GmbH and NSAI Inc.
* The list of IQNet partners is valid at the time of issue of this certificate. Updated information is available under www.iqnet-certification.com

Annex C – Ricard Martinez Review report

PROTEUS

Scalable online machine learning for predictive analytics and real-time interactive visualization

687691

D2.3 Annex: Proteus Data Management and Ethics Plan data Protection Review

Lead Author: Ricard Martínez

Deliverable nature:	<Report (R)>
Dissemination level: (Confidentiality)	
Contractual delivery date:	31/05/2017
Actual delivery date:	15/05/2017
Version:	First Draft
Total number of pages:	16
Keywords:	Data management, exploitation, research ethics, open access

1. Executive summary

This deliverable contains the analysis of the first iteration of the PROTEUS data management plan. According with the Document D2.2 [Proteus Data Management and Ethics Plan] PROTEUS will use and generate three specific data sets that fall within the following categories:

4. ArcelorMittal data provided to the project (ArcelorMittal data)
5. Derived data about the functioning of the scalable online machine learning tools, including data from the benchmarking and technical evaluation process (PROTEUS toolset data)
6. Data from the evaluation of the visualisation aspects of the tool (PROTEUS evaluation data)

As the document states the project has agreed the following principles:

- Data owned by ArcelorMittal will be shared with consortium members, although consortium members can only access this data through the project coordinator
- PROTEUS partners agree to respect the commercial confidence of the data provided by AMIII
- Human participants in PROTEUS research activities are under no obligation to participate and their involvement will be strictly voluntary
- PROTEUS partners will respect data protection and ethical research principles related to the following:
    - Participant confidentiality and anonymization
    - Informed consent
    - Data minimisation
    - Purpose limitation
    - Transparency
    - Rights of access, correction and erasure
- The coordinator will hold all commercially sensitive data and personal data and will manage access to this data for PROTEUS partners

In addition to meeting European regulations, the project will also meet legal regulations set by the Spanish government since data and many of the research activities will be located in Spain. This includes the registration of the project coordinator as a data controller with the Spanish data protection authority and the provision of specific information required by Spanish data protection legislation.
The object of this report consists on the reviewing of the document.

2. Introduction

This report refers exclusively to the compliance of the legal framework in the protection of personal data. This therefore excludes any opinion regarding intellectual property and other legal aspects concerning the PROTEUS Project.

For this purpose, the following references have been taken into account

A.-Legal Framework.

▪ 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) (Text with EEA relevance).

▪ Organic Law 15/1999, 13 December, of protection of Personal data. (Consolidated text. Last updated: 5 march 2011).

▪ Royal Decree 1720 / 2007, of 21 December, approving the regulations implementing the organic law 15/1999, 13 December, of protection of personal data. (Consolidated text. Last updated: 8 march 2012).

▪ Royal Legislative Decree 2/2015 of 23 October approving the consolidated text of the Statute of Workers updates and derogates Royal Legislative Decree 1/1995 of 24 March approving the consolidated text of the Statute of Workers.

B.- Case Law

● *Court of Justice of the European Union.*

▪ Volker und Markus Schecke (C-92/09).

▪ Hartmut Eifert (C-93/09) y Land Hessen.

▪ Google Spain SL and Google Inc. v Agencia Española de Protección de Datos (AEPD) and Mario Costeja González (C-131/12).

▪ Digital Rights Ireland (C-293/12).

▪ Commission v Hungary (C-288/12).

● Spanish Constitutional Court.

▪ STC 292/2000

▪ STC 241/2012.

▪ STC 29/2013.

▪ STC 170/2013.

▪ STC 39/2016.

● European Court of Human Rights.

▪ Case of Bărbulescu V. Romania (Application no. 61496/08)

C.-Reports and Guidelines

▪ Working document on the surveillance of electronic communications in the workplace. (29 May 2002).

▪ Working Party Article 29. Opinion 15/2011 on the definition of consent.

▪ Working Party Article 29. Opinion 05/2014 on Anonymization Techniques.

▪ Spanish Data Protection Agency. Data Protection in Labour Relationships.

▪ Spanish Data Protection Agency. Guidance and guarantees in the procedures of anonymization of personal data.

▪ Spanish Data Protection Agency. Code of good practices in data protection for Big Data projects.

3. General Issues

3.1 The Right to data protection a European Fundamental right.

An appropriate approach to data protection requires that the Court of Justice of the European Union has consolidated in its various judgments the fundamental right to data protection. In this regard, the consequences of three recent judgments (Costeja, Digital Right Ireland and European Commission V. Hungary) and the judgment of the European Court of Human Rights in the Bărbulescu case are briefly outlined.

The following conclusions can be drawn from these judgments relevant to the PROTEUS Project:

▪ Any processing of personal data affects a fundamental right. It therefore receives the highest consideration in the Constitutional Orders of Member State.

▪ The fundamental right is based on respect for the principles of transparency and consent, and the guarantee of rights of access, rectification, cancellation and opposition to the processing.

▪ Data protection also requires guaranteeing what has been inserted in the doctrine in the so-called fair information principles or fairness principles: purpose limitation principle, proportionality and data de minimisation principles, retention principle and security principle.

▪ The identifiability in the internet of the search engines can have an intense repercussion on the data released in the networks.

▪ The publication of information in open data environments must respond to the principle of minimizing the impact on data protection.

▪ Data Protection Authorities play a fundamental role as guarantors of the fundamental right to data protection. A very direct consequence of this is that their guidelines, reports or recommendations must be considered as Soft Law.

▪ In the context of labour relations, strict compliance with data protection principles must be ensured. In this sense the duty of transparency is fundamental to eliminate in its case the expectation of privacy.

On the other hand, it should be taken into account that consent should not be required in those cases where the processing is part of the labour relationship.

3.2 A risk analysis under General Data Protection Regulation (GDPR).

The General Data Protection Regulation provides a new approach based on accountability and risk analysis as well as new obligations that will be fully implemented as from May 25, 2018.

Nevertheless it is necessary to point out that the Spanish Agency of Data Protection has distinguished in its website between ‘entry into force’ and ‘full implementation’ pointing out that the Regulation shall enter into force on the twentieth day following its publication in the Official Journal of the European Union (Article 99).

The Deliverable subject of this report includes different references to elements of the GDPR in section "3.3.2 3.3.2 European General Data Protection Regulation". In addition an explicit commitment is included: "PROTEUS will take each of these obligations into consideration when processing personal data to ensure the project complies with both current legislation and forthcoming legislation".

However, there are at least three recommendations that this expert believes should be followed. The Project should consider the inclusion of specific policies of GDPR compliance in the following areas:

▪ *Protection of data from the design*

In this matter the Deliverable suggests that this principle has been adequately considered. This is demonstrated by the design of consent processes, or references to pseudonymization.

However, it would be convenient to document:

a. If all areas or stakeholders involved have been considered [IT (Chief Information-Security Officer, Legal (Legal Officer or Data Protection Officer), Business and Partners in the Project).

b. If any specific methodology has been followed, such as the Privacy by Design Guidelines of the Ontario[5] Data Protection Authority or Guidelines issued in the matter by the above mentioned Spanish Data Protection Agency, or other authorities[6].

c. If the opening of a Record of processing activities (article 30 GDPR) has been considered.

d. Whether the information will be extended to the data in accordance with the duty of Transparency provided for in Articles 12 to 14 of the GDPR.

▪ From the reading of the document it is deduced that PROTEUS complies with the risk-based approach. However, there is no specific section where the methodology developed is confirmed or where the identified risks and the measures necessary to be eliminated or mitigated are listed.

[5] See for example Privacy by Design (2011), Privacy by Design Solutions for Biometric One-to-Many Identification Systems (2014) Big Data and Innovation – Setting the Record Straight: De-Identification Does Work (2014). Available at https://www.ipc.on.ca/about-us/guidance-documents/?yr=&topic=privacy.

[6] ICO. Big data, artificial intelligence, machine learning and data protection. Available at https://ico.org.uk/media/for-organisations/documents/2013559/big-data-ai-ml-and-data-protection.pdf.

7. Ethical and Legal Issues

Following the structure of Deliverable, sections 4 and 5 in relation to Annex A, this section will discuss the following issues:

- Excluded processes
- Informed consent. The processing of personal data in the framework of labour relations.
- The impact of pseudonymization and anonymization on PROTEUS and its impact on the project.
- Some considerations regarding safety

4.1 Excluded processes.

According to the documentation the processes in the datasets Arcelor Mittal data and PROTEUS predictive analytics system data do not include personal data. The expert has not had access of such data. However from the descriptions contained in the Deliverable, the absence of data is clearly deduced. The data obtained refer exclusively to technical data related to industrial processes and to results obtained through Big Data Analytics Tools.

4.2 The processing of personal data in the framework of labour relations.

The fact that the data subject to processing belong to Arcelor Mittal workers implies taking into account a series of specific forecasts. In this respect the Deliverable states:

> 5.1 Informed Consent
> (…)
> The consortium will take steps not only to ensure that potential participants are well informed but that they also feel empowered to volunteer or decline participation as they see fit. Given that this research will be carried out with employees of one of the partner's organisations, ensuring voluntary participation is paramount and will require a few additional steps. First, following good practice, the information sheet will advise participants that their participation is purely voluntary, and partner AMIII has confirmed this. In addition, non-AMIII personnel will undertake recruitment and advise participants that their participation is voluntary. Finally, during the research activity itself, those conducting the research will invite participants to re-consider their participation and to excuse themselves from the research activity. Given that the project does not involve many sensitive topics, this should be sufficient to ensure voluntary participation, but the consortium remains dedicated to addressing any further concerns raised by employees in the course of the project. Hence, the project will remain vigilant about potential conflicts and will carry out a rolling risk assessment to ensure voluntary participation. To date, an informed consent form has been drafted for future reference and appended in Annex A herein

It is considered that the regulatory compliance policies foreseen here are correct although they can be supplemented by some issues consolidated in the legislation, jurisprudence and guidelines and recommendations of the Spanish Data Protection Agency and the Working Party[7].

[7] SSTC 292/2000, 241/2012, 29/2013, 170/2013, 39/2016.

Bărbulescu V. Romania (Application no. 61496/08)

Working document on the surveillance of electronic communications in the workplace. (29 May 2002).

Working Party Article 29. Opinion 15/2011 on the definition of consent. Working document on the surveillance of electronic communications in the workplace. (29 May 2002).

Spanish Data Protection Agency. Data Protection in Labour Relationships.

The jurisprudence of the Constitutional Court and the Supreme Court in Spain, as well as of the European Court of Human Rights, has repeatedly pointed out that within the framework of the labour relationship, the worker's right to privacy, prevails[8]. This right may be limited in different cases, either by the nature of the work performed or because the employer exercises the control for which he/she is empowered by article 20 of the workers' statute[9]. To this concern there are two basic principles:

▪ When the processing of a worker's personal data in the context of labour relations is directly related to the job benefit, it is not feasible to base such processing on consent.

In this sense, for a better understanding of the voluntary nature of the processing, elements such as making it clear that the survey does not form part of the ordinary work experience or that the conditions of realization do not affect the job, could be highlighted. The specific references to the freedom of choice of the worker, the absence of adverse consequences and the possible withdrawal of the Project when it is decided should be noted as an extremely positive element of the Deliverable and Annex. And also, to the existence of a contact person who, as Data Protection Officer, plays a role of guarantor of their rights.

In any case it is advisable to describe in as much detail as possible the conditions under which the processing will be carried out in order to establish whether it is in fact a voluntary activity or if it should be considered part of the contractual relationship. In this sense, the statement in Annex A is sufficiently clear, although it could be supplemented by a brief description of the project or a document describing the project in general and the voluntary nature of the participation. The textual structure of the informed consent of Annex A would probably be more understandable or friendly if, instead of introducing the explanatory details following phrases like “I consent”, “I understand”, there would be a brief introduction of the Project and the consequences of participation of the worker.

▪ The duty of transparency plays a key role in the

> ■ INFORMATION ON THE PROCESSING OF PERSONAL DATA. MODES:
> The duty of disclosure of Article 5 of the LOPD is an essential part of the right to data protection. The essential nature of this data applies both to the collection of personal data for processing requiring consent, and to cases in which consent is not required.
> Article 5 of Organic Law 15/1999 establishes a duty generally imposed on data controllers, so that in principle affected parties must be notified of the processing of their data, both in cases where processing has the consent of the data subject, and for cases in which processing has been enabled due to other causes admissible under Article 6 of this Law.” (Report 60/2004)
> In the first case, this is because the consent, in addition to being given freely, specifically and in advance, must also be communicated and therefore the failure to inform invalidates the declaration of consent by the data subject or interested party. In addition, the content of the information defined under Article 5 of

[8] Thus the Statute of Workers establishes:

Article 18. Inviolability of the worker.

Records may only be made on the worker, in their lockers and private effects, when they are necessary for the protection of the corporate assets and of the other workers of the company, in the workplace and in working hours. In doing so, dignity and privacy of the worker will be respected and a legal representative of the workers or, in the absence of the work center, of another worker of the company, will be available whenever possible.

[9] Article 20. Management and control of the labour activity.

(…)

3. The employer may adopt the measures he believes more convenient for supervision and control to verify the compliance by the worker of his duties and obligations, keeping in its adoption and application the consideration of his dignity and taking into account, where appropriate, the actual capacity of workers with disabilities.

> the LOPD constitutes a guarantee of data subjects' rights, since it enables them to know in respect of whom these rights may be exercised. (…)
> In the area of personnel management, organisations must be particularly careful, both in the procedure they choose and the time at which they carry out the capture of personal data.

As regards compliance with the principle of transparency, it should be emphasized that the GDPR leaves a broad field to national legislation.

> Article 88
> Processing in the context of employment
> 1. Member States may, by law or by collective agreements, provide for more specific rules to ensure the protection of the rights and freedoms in respect of the processing of employees' personal data in the employment context, in particular for the purposes of the recruitment, the performance of the contract of employment, including discharge of obligations laid down by law or by collective agreements, management, planning and organisation of work, equality and diversity in the workplace, health and safety at work, protection of employer's or customer's property and for the purposes of the exercise and enjoyment, on an individual or collective basis, of rights and benefits related to employment, and for the purpose of the termination of the employment relationship.
> 2. Those rules shall include suitable and specific measures to safeguard the data subject's human dignity, legitimate interests and fundamental rights, with particular regard to the transparency of processing, the transfer of personal data within a group of undertakings, or a group of enterprises engaged in a joint economic activity and monitoring systems at the work place.

Precisely for this reason the provisions on data protection must be supplemented by those of the Workers' Statute. Article 63 of this regulation refers to the works council as "representative and associated of all workers in the company or work center to defend their interests, being constituted in each work center which census is of fifty or more workers. Article 64 gives this Committee information and consultation rights. The works council will have the right to be informed and consulted by the employer on those issues that may affect the workers. Therefore, although the rule does not specifically foresee a duty of information in this regard, since it is an issue related to fundamental rights, it is advisable to inform the Works council about the Project if it exists. It should be noted that in accordance with article 65.2 of the statute, the members of the committee may be required to «observe the duty of discretion with respect to information that, in the legitimate and objective interest of the company or the work center, has been expressly communicated on a reserved basis».

4.3 Pseudonymization and anonymization.

Throughout the document, policies of pseudonymization and/or anonymization are proposed. In this regard, it should be noted that both the Working Party and the Spanish Data Protection Agency are based on a strict interpretation based on the "26th Whereas" of Directive 95/46 /EC of the European Parliament and of the Council of 24 October 1995 On the protection of individuals with regard to the processing of personal data and on the free movement of such data

> (26) Whereas the principles of protection must apply to any information concerning an identified or identifiable person; whereas, to determine whether a person is identifiable, account should be taken of all the means likely **reasonably to be used either by the controller or by any other person to identify** the said person; whereas the principles of protection shall not apply to data rendered anonymous in such a way that the data subject is no longer identifiable; whereas codes of conduct within the meaning of Article 27 may be a useful instrument for providing guidance as to the ways in which data may be rendered anonymous and retained in a form in which identification of the data subject is no longer possible;

This implies that the conditions of anonymization require that "no third party" can reidentify. In the Project, pseudonymization is designated as the default policy instead of anonymization. For this, the identity of the participants and the responses or evaluations derived from the interviews and from viewing the results are separated into two data sets.

In health research, and in general when some risk can be derived, or a right must be guaranteed for the participants, research methodologies advise or require re-identification. In the case of PROTEUS, it appears that there is probably a need to prove the "real or effective" participation in the investigation for the administrative justification of the Project, in order to guarantee scientific accuracy to eventual publications, or to guarantee the value of data in content release in open data environments. If these were reasons for pseudonymization, it would be appropriate to underline them explicitly and explain why a complete anonymization is not done.

On the other hand, section 5.2 describes how a simple method for anonymization/pseudonymization has been used by assigning a numerical value to each participant. In addition, special cases are considered:

> 4. Participants with particular identifying features or experiences may be managed by mixing these with other participants' characteristics (e.g., switching places of birth) to make each participant less identifiable. Where necessary, some identifying features may be removed from the data if it cannot be anonymised.

The methodology is fully shared although the example "switching places of birth" is not adequately understood. Probably, from the point of view of third parties there are more sensitive elements that allow reidentification as the unique characteristics of the workplace. Also, ignoring if the questionnaires or battery of questions to be used are different depending on the job profile of the worker, an opinion cannot be given on whether there could be some type of reidentificación from the answers.

On the other hand, if we take into account the Spanish sector legislation there are two laws from which we can extract complementary ideas useful for the project. We refer to:

- Law 37/2007, of November 16, on reuse of public sector information.
- Law 19/2013, of December 9, on transparency, access to public information and good governance.

Law 37/2007 allows conditions to be imposed on the re-use of anonymized information in Open data contexts:

> Article 8. Conditions for reuse.
> Reusing the information of the Administrations and the public sector bodies referred to in article 2 of this Law may be subject, among others, to the following general conditions:
> (F) Where information, even if provided in a decoupled form, contains sufficient elements that would allow identification of stakeholders in the re-use process, a ban on reversing the dissociation procedure by adding new data obtained from other sources.

Both this rule and the Transparency Law attribute legal responsibility to those who obtain data from a public source process them, when they are personal, or perform some type of reidentification.

Therefore, it would be advisable to include in the regulatory framework among the parties in the consortium and in the terms and conditions of use of the open data spaces a clause adding a prohibition or express commitment of non-reidentification both for the Internal use of the data and in eventual publications.

4.4 Some considerations about safety.

According to the description of the data in the Project, the applicable security level will be the basic one provided in Article 81 of Royal Decree 1720/2007 of 21 December, which approves the Regulation for the development of Organic Law 15/1999, of 13 December, on the protection of personal data. As stated in the document, the organization applies the ISO / IEC 27001: 2013 standard capable of meeting the high standard of the same regulation.

On the other hand, security information regarding encryption of communications is completed. It should be recalled that the Working Party, and now Articles 32 and 34 GDPR consider encrypting data itself as a high quality measure. In this sense, article 34 excludes the duty of notification of security incidents to the data subject when the data had been encrypted.

If personal data are kept in local systems there is no consideration to be made. However it should be remembered that for the processing in Cloud there is the security standard ISO / IEC 27017: 2015, Information technology, Security techniques, Code of practice for information security controls based on ISO / IEC 27002 for cloud services.

5.1 RECOMMENDATIONS

Taking into account the above mentioned considerations, the following recommendations should be followed:

▪ To document, if applicable, the methodologies of data protection from design and by default and give an impact analysis on data protection, and their results.

▪ Consider opening an internal Registry of the processes.

▪ Verify if the information to data subjects complies with the provisions of art. on transparency provided for in the GDPR.

▪ Describe in as much detail as possible the conditions under which the processing will be carried out in order to establish whether it is in fact a voluntary activity or if it should be considered part of the contractual relationship.

▪ Inform about the processing of personal data in the framework of the Project to the works council

▪ Explicitly document the reasons for the pseudonymization and explain in such a case why a complete anonymization is not performed.

▪ Include in the regulatory framework between the parties in the consortium and in the terms and conditions of use of the open data spaces that a clause could be established to add a prohibition or express commitment of non-reidentification both for the internal use of the data and for any publications.

▪ It should be remembered that for the processing in Cloud there is the security standard ISO / IEC 27017: 2015, Information technology, Security techniques, Code of practice for information security controls based on ISO / IEC 27002 for cloud services.

Annex D - Ricard Martinez CV

Data Protection Officer of the University of Valencia and Director of the Microsoft- University of Valencia Privacy & Digital Transformation Institutional Chair.

Former Head of Area of Transparency and Open Government (Diputación de Valencia). Former President of the Spanish Privacy Professionals Association (May 2016). Data Protection Officer (University of Valencia-November 2015). Head of the Studies Area of the Spanish Data Protection Authority (2007-2011). Assistant Professor in Constitutional Law, and law and Information Technologies at Open Catalonia University (UOC). Collaborator in various master´s degrees at U. Valencia, U. Polytechnic of Valencia, U. Carlos IIII. Director of the Official Master of Data protection Law at the International University of la Rioja. Part Time Researcher at U. Jaume I (General Data Protection Regulation Impact (DER2015-63635-R)). Ph. D. in Law (with distinction), University of Valencia. Dissertation about Data protection Right, Technology and Freedoms. Member of the Academic Council of Fide. Publications.
https://dialnet.unirioja.es/servlet/autor?codigo=173672

Annex E – List of changes after Ricard Martinez review

Following the review and recommendations of our dedicated Spanish data expert, Richard Martinez, the following amendments have taken place in D2.3.

	In response to the following recommendations from Richard Martinez' report:		Changes in the following sections of the final D2.3 report were made:
p.7	Section 3.2	p.14	**Section 3.3.2** "Data protection by design and defaut" was amended to include all stakeholders and clarify the rationale for pseudonyisation "Processing personal data for research purposes" was amended to include requirement of article 30, 32 of GDPR. A new bullet point on Transparency and communication with data subjects was added to include GDPR requirements outlined in Articles 12-14 of GDPR
P11-12	Section 4.3	p.15	**Section 3.3.3** Reference to Law 37/2007 (article 8) regarding re-identification was introduced at the end of the section
p.9		p.22	**Section 5.1** Reference to principles of jurisprudence as per the Spanish law and prior consultation with the Works council about the project was introduced at the end of the section.
p.12	Section 4.4	p.24	**Section 6.2** Reference to ISO standards with respect to cloud security processing of personal data was added.
		Annex A	Changes in the information sheet were made to make more explicit and prominent recommendation regarding jurisprudence and transparency and explicitly declare a Data Protection Officer