COGNITWIN

Cognitive plants through proactive self-learning hybrid digital twins

DT-SPIRE-06-2019 (870130)

Deliverable Report

Deliverable ID	D8.1	**Version**	V1
Deliverable name	Data Management Plan		
Lead beneficiary	SINTEF (SINTEF AS)		
Contributors	Arne J. Berre (SINTEF), Stein Tore Johansen (SINTEF)		
Reviewers	Peter Singstad (CYB), Sailesh Abburu (SINTEF)		
Due date	29.02.2020		
Date of final version	28.02.2020		
Dissemination level	Public		
Document approval	Frode Brakstad	28.02.2020	

The COGNITWIN project has received funding from the European Union's Horizon 2020 research and innovation programme under GA No. 870130

Executive Summary

This document is the initial Data Management Plan for the COGNITWIN project. Chapter 1 start with an introduction to the six COGNITWIN pilots and provides an overview of the type of data that is being collected within each of the pilots. Chapter 2 provides an overview of data representations and formats and data types and discussed different roles with an interest in data management and data usage. Chapter 3 introduces and explains the concept of FAIR (Findability, Accessibility, Interoperability and Re-usability) for data, as a foundation for research data management. Chapter 4 explains various aspects of data management support and corresponding roles. Chapter 5 introduces carious aspects of data security and data privacy, while chapter 6 discusses ethical concerns as related to data. The concluding chapter 7 outlines further evolution of data management in the COGNITWIN project, while the appendixes contains templates to support this further evolution.

Table of Contents

List of figures

Acronyms

DMP	Data Management Plan
GTC	Gas Treatment Centre
HF	Hydrogen Fluoride
MW	Megawatt
NmP^3	Normal cubic meter, at 1013 mbar and 273,15°K (DIN 1343)
GE	General Electric
IHEX	Internal Heat Exchanger
SGA	Smelter-Grade Alumina
IoT	Internet of Things
KPI	Key Performance Indicator
PTH	Post-Taphole

1. Introduction to COGNITWIN Data Management Plan

This document outlines COGNITWIN's data management plan (DMP), formally documenting how data will be handled both during the implementation and after the end of the project. Many DMP aspects will be considered including metadata generation, data preservation, data security and ethics, accounting for the FAIR (Findable, Accessible, Interoperable, Re-usable) data principle. This DMP will be updated over the course of COGNITWIN project whenever significant changes arise. The updates of this document will increasingly provide in-depths on COGNITWIN DMP strategies with particular interest on the aspects of findability, accessibility, interoperability and reusability of the Big Data the project produces.

The COGNITWIN project is built around 6 pilots from the Process Industry. Major elements will be to introduce robust, accurate and cost-efficient sensors using retro-fitting as well as novel new sensors as needed to achieve the planned cognitive elements in form of proactive self-learning digital twins. Although a full digitalisation of the plant is the aim, the technology demonstration will be shown in the most crucial selected parts of the industrial participants plants – i.e. the selected pilots, whereas the technology development can be transferred to the complete plant. Each pilot will use different type of data, typically collected by sensors and managed by various process control systems. In the following a short introduction to each pilot with associated data is provided. Further details about each pilot can be found in the COGNITWIN public deliverables D1.1, D2.1 and D3.1

1.1 Hydro Pilot - Aluminium Production Process

The topic of the pilot is related to Reduced energy consumption in a selected Hydro GTC (Gas Treatment Center).

The figure in the right shows how the COGNITWIN work is related to the performance of the Gas Treatment Center (top right) and the interactions with the pots (reduction cells) and the inflow of fresh alumina. The ambition is to develop a Digital Twin that allows optimal operation, acceptable emissions of HF (Hydrogen Fluoride) and which can account for the variations in alumina quality from ship load to ship load. The work will increase the overall efficiency with 10% by achieving symbiosis between the actual production (electrolysis) and the cleaning technology (Gas Treatment Centre GTC). Improved environmental impact and optimize energy consumption by maximizing the efficiency of the Gas Treatment Centre.

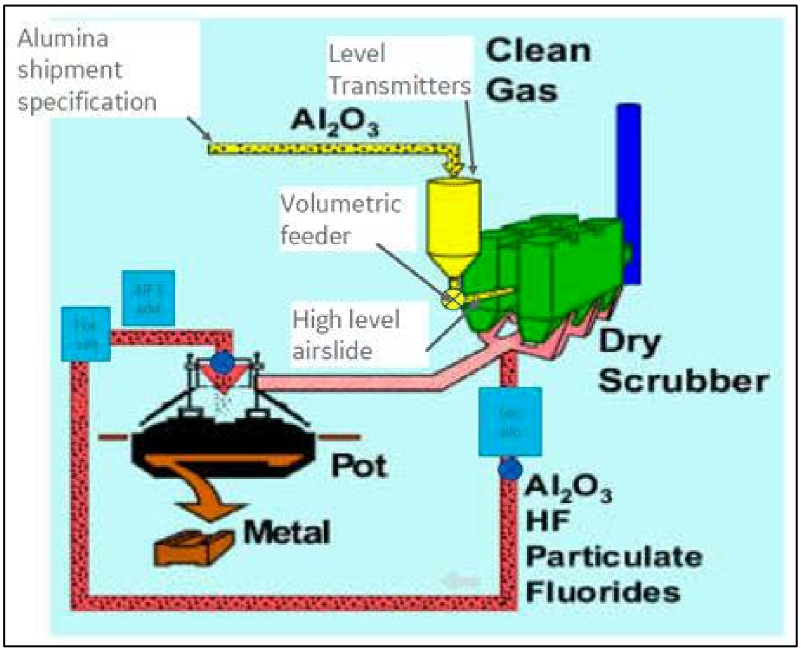

Figure 1: A schematic view of the Hall-Heroult aluminium production process.

Reduce energy consumption in GTC by 15%. Reduce suction rate overall by 10%, i.e. for the pilot in question, 1500 MWh/y saved fan work, and increased available recovered thermal energy of 13500 MWh/y. Reduced energy consumption and/or replacement by thermal energy (heating) will save CO2 emissions caused by current energy source. The goal is to balance flow distribution to different chambers within ±5% and to decrease process disturbance by preventive maintenance by 5%.

<u>**Data Management Plan – Aluminium production process**</u>

The GTC-DT App is seen as an Advisory solution, A various set of data will be collected at the pilot site (PLC, DCS, ERP, CMMS…) and sent to the Predix cloud where the to-be-developed Analytics will consume those data. The GTC-DT advisories will be available to Hydro operators & process engineers within a standard browser.

Sensors that will produce data:

1. Various Pressure Transmitters (available to some extent, gas)
2. Various Temperature Transmitters (available to some extent, gas)
3. Feeder rotation counter (available, secondary alumina, primary alumina)
4. Low cost robust gas flow measurements (desired, also gas from individual cell)
5. Online HF monitors roof emissions (available, gas)
6. Online HF monitor raw gas (desired)
7. Online HF after dry scrubber (available)
8. Individual chamber HF at-line analysis (GE-HF sniffer)
9. Ambient conditions (available, weather station)
10. Liquid flow measurement (available, IHEX gas flow estimate)

The data will be further analysed with respect to the collection needs of the pilot and further related to based on the principle's FAIR principles in the COGNITWIN Data Management Plan. Much of the data will be confidential and only accessible by those that have authorisation for this.

1.2 Elkem Pilot - Silicon Production Process

In this pilot scheme the challenge is to optimize the post tapping process in an Elkem Silicon plant. The figure in the right shows the silicon production process and the post tapping process and the focus area in this project is highlighted inside by the *green* circle (liquid metal / refining). By application of various digitalization tools and techniques to the post tapping processes (tapping into the ladle, silicon casting into molds) silicon yield can be increased, ladle lifetime can be increased, metal quality can be improved and energy consumption can be reduced. By help of new measurement

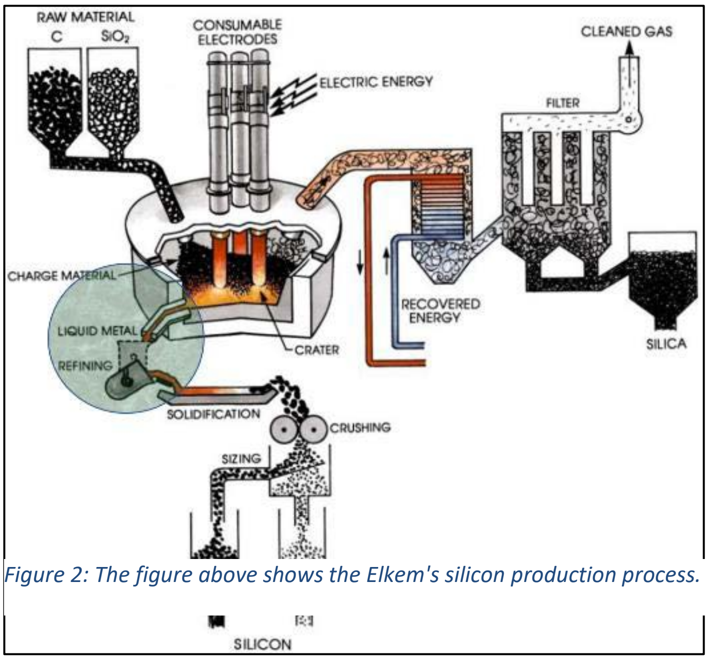

Figure 2: The figure above shows the Elkem's silicon production process.

techniques COGNITWIN will help enabling on-line estimates of the actual silicon flow and its temperatures. Application of new and old data into cognitive hybrid models will be developed to improve the product quality due to more consistent quality and lead to more profitable operation.

Data Management Plan – Silicon production process

There are several process parameters that are recorded during the post tapping process. Some of these are immediately available.

- IoT Platform Elkem IoT Platform
- Other systems Sensors, controllers, PLCs
- Simulation tools
- Data Sources (IoT) Sensor Data
- Data type Heterogeneous
- Velocity 1Hz

The data will be further analysed with respect to the collection needs of the pilot and further related to based on the principles FAIR principles in the COGNITWIN Data Management Plan. Much of the data will not be publicly available and only accessible by those that have received authorisation for this.

1.3 Sidenor Pilot - Steel Production – predict the life-time of the ladle lining

In this pilot process, the project aims to predict the condition of the brick lining in the ladle and help the technicians decide if the lining in the ladle needs to be repaired or replaced completely for the next heating process. Figure 3 shows how the belt of dark gray refractory bricks is made of special material which is harder to erode during the operations. In the steel plant ladles gas injection is applied for refining and stirring. It is observed that ladle lifetime varies a lot and depends on many parameters. The COGNITWIN approach will be to develop a hybrid model that may exploit the large data that already exist. In addition COGNITWIN will apply physics based models that can handle the thermomechanical conditions in the ladle refractory, take advantage of the thermodynamic data for the steel-slag-refractory system, and account for multiphase and multicomponent mass transfer as well as the dynamic temperature variations in the system. Based one available data, new measuring techniques and physics based modeling a

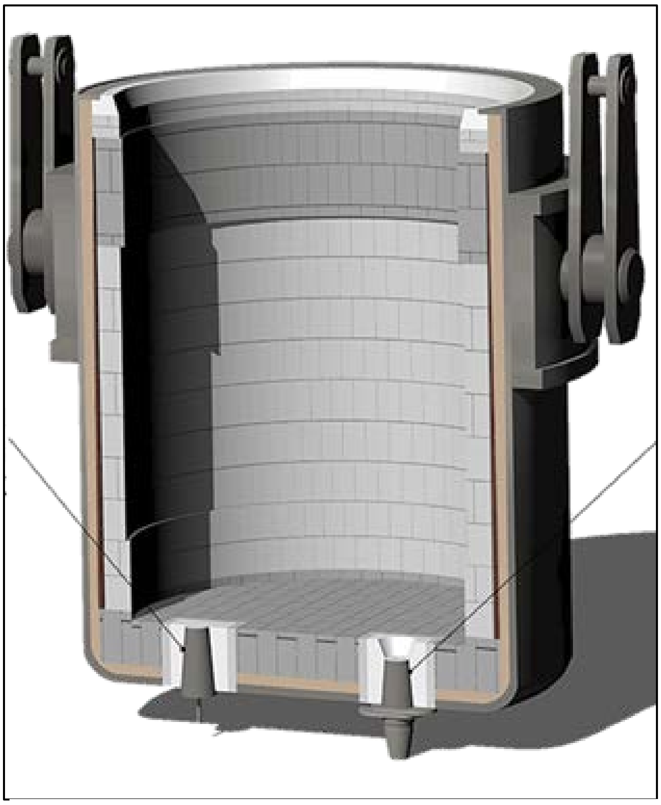

Figure 3: A steel ladle with porous bottom plugs for gas injection.

Digital Twin for the ladle operation will be developed and used to optimize the ladle lifetime and reduce operational costs.

Data Management Plan – Steel production process

Possible parameters that affect the refractory wear in the ladle as follows: Tapping temp, Slag carry over, Stirring practices, Steel grades, Secondary metallurgy needs (de-sulfurisation, degassing etc.), Types and methodology of alloys and fluxes additions of tap, Steel contact or residence time, electrical consumption KWh and taps used at ladle furnace, Thermal cycles ladle is submitted to, thermal shock, reliability of ladle, Drying curves, slag treatment and slag fluxes. The data will be further analysed with respect to the collection needs of the pilot and further related to based on the principles FAIR principles in the COGNITWIN Data Management Plan. Much of the data will not be publicly available and only accessible by those that have received authorisation for this.

1.4 Saarstahl Pilot - Tracking system for rolled bars in the rolling mill

This pilot is owned by Saarstahl AG, and where the topic is to enable tracking system for rolled bars in the rolling mill.

Figure 4: A hot, glowing bar in the Saarstahl rolling mill

Data Management Plan – Rolled bars in the rolling mil

The data provided for the tracking system will be a video stream stemming from 3 Full HD Cameras and possibly some additional video or image data. For training purposes, recorded video files will be provided in addition to the synthetic training data. The data will be further analysed with respect to the collection needs of the pilot and further related to based on the principles FAIR principles in the COGNITWIN Data Management Plan. Much of the data will not be publicly available and only accessible by those that have received authorisation for this.

1.5 Noksel Pilot - Condition monitoring by development of digital twin

This pilot is owned by Noksel. The goal is to apply a cognitive digital twin to power condition monitoring (and control) in the steel pipe manufacturing. Noksel's pilot case is the development of a digital twin for an SWP machine in steel pipe production. The digital twin will collect and analyse multiple sensors' data in real-time and enable a smart condition monitoring system for predictive maintenance. Real-time data acquisition, communication networks for monitoring, and automated recommendation generation are among the key innovative features of this pilot. Automated recommendations will also be generated.

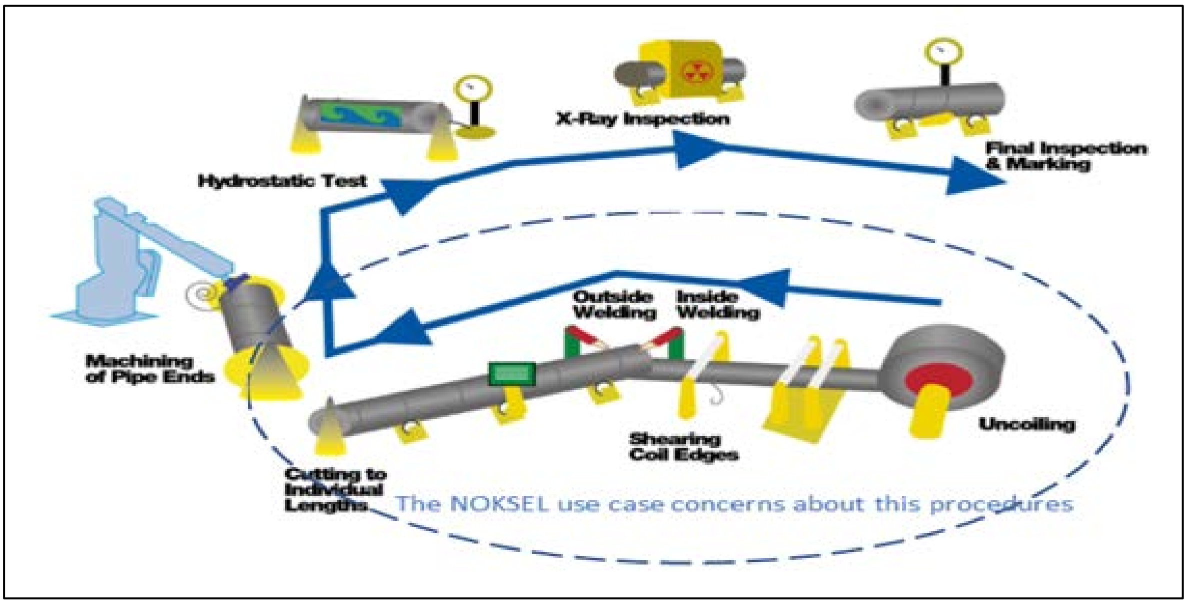

Figure 5: Noksel SWP processes for producing welded steel pipes

Smart components that use sensors to gather data about real-time status, working condition, or position will be connected to a cloud-based system that receives and processes all the data the sensors monitor. This input will be analysed against business and other contextual data through smart visualization systems. The digital twin model will allow joining physical and virtual worlds to create a new networked layer in which intelligent objects interact with each other to virtualize the steel pipe manufacturing process on the SWP machinery. The ambition is to reduce machine downtimes, decrease energy consumption, and increase total equipment performance.

Data Management Plan – production of welded steel pipes

Current tracking system provides downtimes periods and types, total working durations, effective working durations, number of produced pipes, meters of produced pipes, weight of produced pipes, which pipe produced of which labelled raw material which shows the quality of raw material. Only daily electrical consumptions have been started to be recorded since November 2019. Before then, electrical consumptions are recorded monthly in 2019. The data will be further analysed with respect to the collection needs of the pilot and further related to based on the principles FAIR principles in the COGNITWIN Data Management Plan. Much of the data will not be publicly available and only accessible by those that have received authorisation for this.

1.6 Sumitomo Pilot - Engineering Boiler operations

This pilot is owned by Sumitomo and where the objective is to minimise scaling and fouling in their customer's heat-exchanger tubes even if the fuels type, quality and its composition changes. We seek to achieve this by development cognitive digital twins.

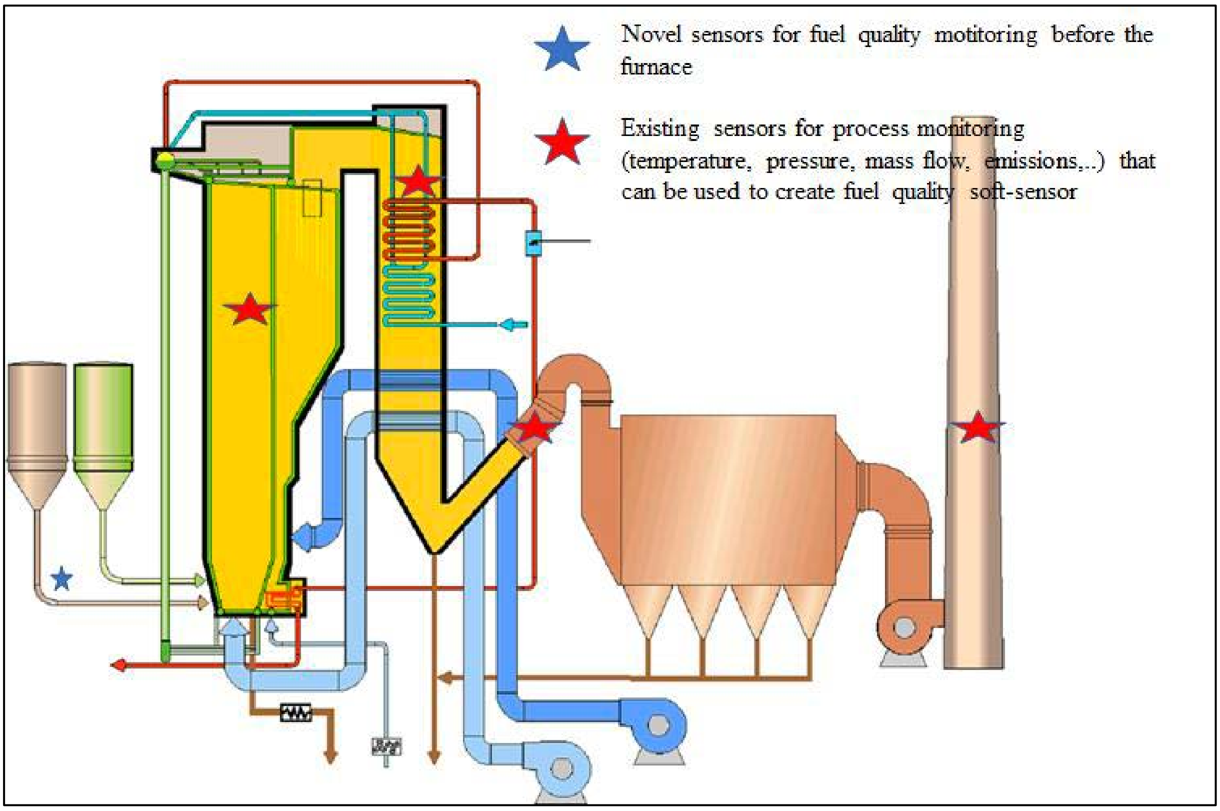
Novel sensors for fuel quality motitoring before the furnace
Existing sensors for process monitoring (temperature, pressure, mass flow, emissions,..) that can be used to create fuel quality soft-sensor

Figure 6: Overview of the boiler process in a Sumitomo made plant

The innovation and the cognitive elements here are to introduce new measuring techniques, combine measured fuel quality data, process data from the power plant and existing physics based models. The developed digital twins should predict how fuel quality changes affect the process, which enables early detection of process disturbances and overall process optimization.

Data Management Plan – Minimise heat exchanger fouling and scaling

Either private or open cloud with matching edge connectivity linked to the Sumitomo IoT platform solution, details be clarified after the final pilot plant has been selected. Process data collection from DCS historian database via OPC-DA or OPC-UA to local site server-based data storage possible. Either from DCS historian database via OPC-DA or OPC-UA or by utilizing IoT connectivity directly from DCS bus depending on the final architecture including Time stamped on-line process data. The data will be further analysed with respect to the collection needs of the pilot and further related to based on the principles FAIR principles in the COGNITWIN Data Management Plan. Much of the data will not be publicly available and only accessible by those that have received authorisation for this.

2. Data Summary

2.1 Purpose of data collection

During the lifecycle of the COGNITWIN project, data will be collected from all of the 6 pilots to be used for the various development of analytics and machine learning in the context of Cognitive and Hybrid Digital Twins. The ultimate purpose of data collection is to use the data as a source of information in the implementation of a variety of big data analytics algorithms, services and applications COGNITWIN will deploy to create a value, new business facts and insights with a particular focus on the process industry. The datasets are part of the building block input to the COGNITWIN's big data technology toolbox, that was designed to help to increase productivity for big data management and analytics.

Big Data experts provide common analytic technology support for the main common and typical Process data applications/analytics that are now emerging through the pilots in the project. Data from the past will be managed and analyzed, including many different kind of data sources: i.e., descriptive analytics and classical query/reporting (in need of variety management - and handling and analysis of all of the data from the past, including performance data, transactional data, attitudinal data, descriptive data, behavioural data, location-related data, interactional data, from many different sources). Big data from the present time will be harnessed in the process of monitoring and real-time analytics - pilot services (in need of velocity processing - and handling of real-time data from the present) - trigging alarms, actuators etc. Harnessing big data for the future time include forecasting, prediction and recommendation analytics - pilot services (in need of volume processing - and processing of large amounts of data combining knowledge from the past and present, and from models, to provide insight for the future)

2.2 Data types and formats

The COGNITWIN specific data types, formats and sources will be listed in more detail in the evolving Appendix A; below are described key features of the data used in the project.

Structured data

Structured data refers to any data that resides in a fixed field within a record or file. This includes data contained in relational databases, spreadsheets, and data in forms of events such as sensor data. Structured data first depends on creating a data model – a model of the types of business data that will be recorded and how they will be stored, processed and accessed. This includes defining what fields of data will be stored and how that data will be stored: data type (numeric, currency, alphabetic, name, date, address) and any restrictions on the data input (number of characters; restricted to certain terms etc).

Semi-structured data

Semi-structured data is a cross between structured and unstructured data. It is a type of structured data, but lacks the strict data model structure. With semi-structured data, tags or other types of markers are used to identify certain elements within the data, but the data doesn't have a rigid structure. For example, word processing software now can include metadata showing the author's name and the date created, with the bulk of the document just being unstructured text. Emails have the sender, recipient, date, time and other fixed fields added to the unstructured data of the email message content and any attachments. Photos or other graphics can be tagged with keywords such as the creator, date, location and keywords, making it possible to organize and locate graphics. XML and other markup languages are often used to manage semi-structured data. Semi-structured data is therefore a form of structured data that does not conform with the formal structure of data models associated with relational databases or other forms of data tables, but nonetheless contains tags or other markers to separate semantic elements and enforce hierarchies of records and fields within the data. Therefore, it is also known as self-describing structure. In semi-structured data, the entities belonging to the same class may have different attributes even though they are grouped together, and the attributes' order is not important. Semi-structured data are increasingly occurring since the advent of the Internet where full-text documents and databases are not the only forms of data anymore, and different applications need a medium for exchanging information. In object-oriented databases, one often finds semi-structured data.

XML and other markup languages, email, and EDI are all forms of semi-structured data. OEM (Object Exchange Model) was created prior to XML as a means of self-describing a data structure. XML has been popularized by web services that are developed utilizing SOAP principles. Some types of data described here as "semi-structured", especially XML, suffer from the impression that they are incapable of structural rigor at the same functional level as Relational Tables and Rows. Indeed, the view of XML as inherently semi-structured (previously, it was referred to as "unstructured") has handicapped its use for a widening range of data-centric applications. Even documents, normally thought of as the epitome of semi-structure, can be designed with virtually the same rigor as database schema, enforced by the XML schema and processed by both commercial and custom software programs without reducing their usability by human readers. In view of this fact, XML might be referred to as having "flexible structure" capable of human-centric flow and hierarchy as well as highly rigorous element structure and data typing. The concept of XML as "human-readable", however, can only be taken so far. Some implementations/dialects of XML, such as the XML representation of the contents of a Microsoft Word document, as implemented in Office 2007 and later versions, utilize dozens or even hundreds of different kinds of tags that reflect a particular problem domain - in Word's case, formatting at the character and paragraph and document level, definitions of styles, inclusion of citations, etc. - which are nested within each other in complex ways. Understanding even a portion of such an XML document by reading it, let alone catching errors in its structure, is impossible without a very deep prior understanding of the specific XML implementation, along with assistance by software that understands the XML schema that has been employed. Such text is not "human-understandable" any more than a book written in Swahili (which uses the Latin alphabet) would be to an American or Western European who does not know a word of that language: the tags are symbols that are meaningless to a person unfamiliar with the domain.

JSON or JavaScript Object Notation, is an open standard format that uses human-readable text to transmit data objects consisting of attribute–value pairs. It is used primarily to transmit data between a server and web application, as an alternative to XML. JSON has been popularized by web services developed utilizing REST principles. There is a new breed of databases such as MongoDB and Couchbase that store data natively in JSON format, leveraging the pros of semi-structured data architecture.

Unstructured data

Unstructured data (or unstructured information) refers to information that either does not have a pre-defined data model or is not organized in a pre-defined manner. This results in irregularities and ambiguities that make it difficult to understand using traditional programs as compared to data stored in “field” form in databases or annotated (semantically tagged) in documents. Unstructured data can't be so readily classified and fit into a neat box: photos and graphic images, videos, streaming instrument data, webpages, PDF files, PowerPoint presentations, emails, blog entries, wikis and word processing documents. In 1998, Merrill Lynch cited a rule of thumb that somewhere around 80-90% of all potentially usable business information may originate in unstructured form. This rule of thumb is not based on primary or any quantitative research, but nonetheless is accepted by some. IDC and EMC project that data will grow to 40 zettabytes by 2020, resulting in a 50-fold growth from the beginning of 2010. Computer World states that unstructured information might account for more than 70%–80% of all data in organizations.

Software that creates machine-processable structure can utilize the linguistic, auditory, and visual structure that exist in all forms of human communication. Algorithms can infer this inherent structure from text, for instance, by examining word morphology, sentence syntax, and other small- and large-scale patterns. Unstructured information can then be enriched and tagged to address ambiguities and relevancy-based techniques then used to facilitate search and discovery. Examples of "unstructured data" may include books, journals, documents, metadata, health records, audio, video, analog data, images, files, and unstructured text such as the body of an e-mail message, Web page, or word-processor document. While the main content being conveyed does not have a defined structure, it generally comes packaged in objects (e.g. in files or documents, …) that themselves have structure and are thus a mix of structured and unstructured data, but collectively this is still referred to as "unstructured data".

2.3 Big data - synergies with the European Big Data Value Reference Model

The new generation of big data management is in particular focusing on semi-structured and unstructured data, often in combination with structured data. In the BDVA reference model for big data technologies a distinction is done between 6 different big data types. The vertical dimension is according to the following Big Data types:

- Structured Data
- IoT/Time Series
- SpatioTemporal
- Media/Image
- Text/NLP
- Graph/Metadata

The BDV Reference Model[1] shown in Figure 7 has been developed by the BDVA, taking into account input from technical experts and stakeholders along the whole Big Data Value chain as well as interactions with other related PPPs. An explicit aim of the BDV Reference Model in the SRIA 4.0 document is to also include logical relationships to other areas of a digital platform such as Cloud, High Performance Computing (HPC), IoT, Networks/5G, CyberSecurity etc.

[1] http://bdva.eu/sites/default/files/BDVA_SRIA_v4_Ed1.1.pdf (page 37)

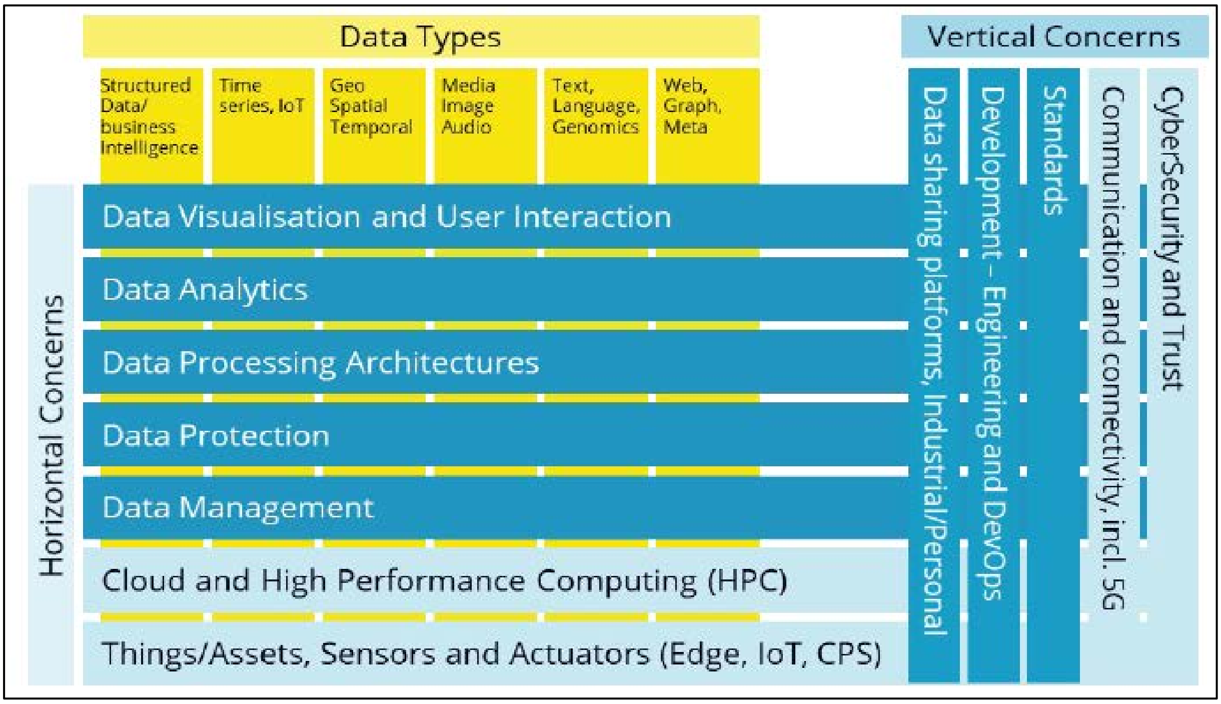
Data Types
Structured Data/ business Intelligence
Time series, IoT
Geo Spatial Temporal
Media Image Audio
Text, Language, Genomics
Web, Graph, Meta
Vertical Concerns
Data sharing platforms, Industrial/Personal
Development – Engineering and DevOps
Standards
Communication and connectivity, incl. 5G
CyberSecurity and Trust
Horizontal Concerns
Data Visualisation and User Interaction
Data Analytics
Data Processing Architectures
Data Protection
Data Management
Cloud and High Performance Computing (HPC)
Things/Assets, Sensors and Actuators (Edge, IoT, CPS)

Figure 7: BDV Reference Model as a foundation for the COGNITWIN toolbox elements

The BDV Reference Model may serve as common reference framework to locate Big Data technologies on the overall IT stack. It addresses the main concerns and aspects to be considered for Big Data Value systems.

The BDV Reference Model is structured into horizontal and vertical concerns.

- **Horizontal concerns** cover specific aspects along the data processing chain, starting with data collection and ingestion, reaching up to data visualization. It should be noted, that the horizontal concerns do not imply a layered architecture. As an example, data visualization may be applied directly to collected data (data management aspect) without the need for data processing and analytics. Further data analytics might take place in the IoT area – i.e. Edge Analytics. This shows logical areas – but they might execute in different physical layers.
- **Vertical concerns** address cross-cutting issues, which may affect all the horizontal concerns. In addition, verticals may also involve non-technical aspects (e.g., standardization as technical concerns, but also non-technical ones).

Given the purpose of the BDV Reference Model to act as a reference framework to locate Big Data technologies, it is purposefully chosen to be as simple and easy to understand as possible. It thus does not have the ambition to serve as a full technical reference architecture. However, the BDV Reference Model is compatible with such reference architectures, most notably the emerging ISO JTC1 WG9 Big Data Reference Architecture – now being further developed in ISO JTC1 SC42 Artificial Intelligence.

The following technical priorities as expressed in the BDV Reference Model are elaborated in the remainder of this section:

Horizontal concerns:

- **Big Data Applications**: Solutions supporting Big Data within various domains will often consider the creation of domain specific usages and possible extensions to the various horizontal and vertical areas. This is often related to the usage of various combinations of the identified Big Data types described in the vertical concerns.

- **Data Visualisation** and User Interaction: Advanced visualization approaches for improved user experience.
- **Data Analytics**: Data analytics to improve data understanding, deep learning, and meaningfulness of data.
- **Data Processing Architectures**: Optimized and scalable architectures for analytics of both data-at-rest and data-in- motion with low latency delivering real-time analytics.
- **Data Protection**: Privacy and anonymisation mechanisms to facilitate data protection. It also has links to trust mechanisms like Blockchain technologies, smart contracts and various forms for encryption. This area is also associated with the area of CyberSecurity, Risk and Trust.
- **Data Management**: Principles and techniques for data management including both data life cycle management and usage of data lakes and data spaces, as well as underlying data storage services.
- **Cloud and High Performance Computing (HPC):** Effective Big Data processing and data management might imply effective usage of Cloud and High Performance Computing infrastructures. This area is separately elaborated further in collaboration with the Cloud and High Performance Computing (ETP4HPC) communities.
- **IoT, CPS, Edge and Fog Computing**: A main source of Big Data is sensor data from an IoT context and actuator interaction in Cyber Physical Systems. In order to meet real-time needs it will often be necessary to handle Big Data aspects at the edge of the system.

Vertical concerns:

- **Big Data Types and semantics**: The following six Big Data types have been identified – based on the fact that they often lead to the use different techniques and mechanisms in the horizontal concerns, which should be considered, for instance for data analytics and data storage: *1) Structured data; 2) Times series data; 3) GeoSpatial data, 4) Media, Image, Video and Audio data; 5) Text data, including Natural Language Processing data and Genomics representations; 6) Graph data, Network/Web data and Meta data*. In addition, it is important to support both the syntactical and semantic aspects of data for all Big Data types.
- **Standards**: Standardisation of Big Data technology areas to facilitate data integration, sharing and interoperability.
- **Communication and Connectivity**: Effective communication and connectivity mechanisms are necessary for providing support for Big Data. This area is separately elaborated further with various communication communities, such as the 5G community.
- **Cybersecurity**: Big Data often need support to maintain security and trust beyond privacy and anonymisation. The aspect of trust frequently has links to trust mechanisms such as blockchain technologies, smart contracts and various forms of encryption. The CyberSecurity area is separately elaborated further with the CyberSecurity PPP community.
- **Engineering and DevOps**: for building Big Data Value systems. This area is also elaborated further with the NESSI (Networked European Software and Service Initiative) Software and Service community.
- **Data Platforms**: Marketplaces, IDP/PDP, Ecosystems for Data Sharing and Innovation support. Data Platforms for Data Sharing include in particular Industrial Data Platforms (IDPs) and Personal Data Platforms (PDPs), but also include other data sharing platforms like Research Data Platforms (RDPs) and Urban/City Data Platforms (UDPs). These platforms include efficient usage of a number of the horizontal and vertical Big Data areas, most notably the areas for data management, data processing, data protection and CyberSecurity.
- **AI platforms**: In the context of the relationship between AI and Big Data there is an evolving refinement of the BDV Reference Model – showing how AI platforms typically include support

for Machine Learning, Analytics, visualization, processing etc. in the upper technology areas supported by data platforms – for all of the various Big Data types.

Sensor data

Within the COGNITWIN pilots, several key parameters will be monitored through sensorial platforms and sensor data will be collected along the way to support the infrastructure activities. Different types of sensor data have been already identified and namely, a) IoT data from in-situ sensors and telemetric stations and multi spectre cameras, b) imagery data from unmanned aerial sensing platforms (drones), c) imagery from hand-held or mounted optical sensors.

Internet of Things data

The IoT data are a major subgroup of sensor data involved in multiple pilot activities in the COGNITWIN infrastructure. IoT data are sent via TCP/UDP protocol in various formats (e.g. txt with time series data, json strings) and camera data. In particular, some COGNITWIN pilots will use optical (RGB), thermal or multispectral images and 3D point-clouds. The information generated by cameras is usually Image Data (JPEG or JPEG2000). A general description of the workflow is provided below.

Data acquired by an RGB sensor

The RGB sensor acquires individual pictures in **.JPG** format, together with their 'geotag' files, which are downloaded from the RPAS and processed into:

- **.LAS** files: 3D point clouds (x, y, z), which are then processed to produce Digital Models
- **.TIF** files: which are then processed into an orthorectified mosaic. In order to obtain smaller files, mosaics are usually exported to compressed **.ECW** format.

Data acquired by a thermal sensor

The Thermal sensor acquires a video file which is downloaded from the RPAS and:

- split into frames in **.TIF** format (pixels contain Digital Numbers: 0-255)
- 1 of every 10 frames is selected (with an overlap of about 80%, so as not to process an excessive amount of information)

Data acquired by the multispectral sensor

The multispectral sensor acquires individual pictures from the 6 spectral channels in **.RAW** format, which are downloaded from the RPAS and processed into:

- **.TIF** files (16 bits), which are then processed to produce a 6-bands .TIF mosaic (pixels contain Digital Numbers: 0-255)

Data from hand-held or mounted optical sensors

Images from hand-held or mounted cameras can be collected using truck-held or hand held full Range / high resolution UV-VIS-NIR-SWIR Spectroradiometer.

Machine-generated data

Machine-generated data in the COGNITWIN project are data produced by sensors and machinery. These data will serve for further analysis and optimisation of processes in the Process control systems.

2.4 Historical data

In the context of doing machine learning and predictive and prescriptive analytics it is important to be able to use historical data for training and validation purposes. Machine learning algorithms will use existing historical data as training data both for supervised and unsupervised learning. Information about datasets and the time periods concerned with historical datasets to be used for COGNITWIN can be found in the evolving Appendix A. Historical data can also serve as training complex event processing applications. In this case, historical data is injected as "happening in real-time" therefore serving as testing the complex event driven application in hand before running it in real-environment.

2.5 Expected data size and velocity

The big data "V" characteristics of Volume and Velocity is being described for each of the identified data sets in the COGNITWIN infrastructure - typically with measurements of total historical volumes and new/additional data per time unit. The COGNITWIN-specific Data Volumes and velocities (or injection rates) can be found in the evolving Appendix A.

2.6 Data beneficiaries

In this section, this document analyses the key data beneficiaries who will benefit from the use of big data in several fields as analytics, data sets, business value, sales or marketing. This section will consider both tangibles and intangibles concepts.

2.6.1 Technical Staff

Adoption rates aside, the potential benefits of utilising big data and related technologies are significant both in scale and scope and include, for example: better/more targeted marketing activities, improved business decision making, cost reduction and generation of operational efficiencies, enhanced planning and strategic decision making and increased business agility, fraud detection, waste reduction and customer retention to name but a few. Obviously, the ability of firms to realize business benefits will be dependent on company characteristics such as size, data dependency and nature of business activity.

A core concern voiced by many of those participating in big data focused studies is the ability of employers to find and attract the talent needed for both a) the successful implementation of big data solutions and b) the subsequent realisation of associated business benefits. Although 'Data Scientist' may currently be the most requested profile in big data, the recruitment of Data Scientists (in volume terms at least) appears relatively low down the wish list of recruiters. Instead, the openings most commonly arising in the big data field (as is the case for IT recruitment) are development positions~~.~~

2.6.2 Developers

The generic title of developer is normally employed together with a detailed description of the specific technical related skills required for the post and it is this description that defines the specific type of development activity undertaken. The technical skills most often cited by recruiters in adverts for big data Developers are: NoSQL (MongoDB in particular), Java, SQL, JavaScript, MySQL, Linux, Oracle, Hadoop (especially Cassandra), HTML and Spring.

2.6.3 Architects

More specifically, however, applicants for these positions are required to hold skills in a range of technical disciplines including: Oracle (in particular, BI EE), Java, SQL, Hadoop and SQL Server, whilst the main generic areas of technical Knowledge and competence required were: Data Modelling, ETL, and Enterprise Architecture, Open Source and Analytics.

2.6.4 Analysts

Particular process/methodological skills required from applicants for analyst positions were primarily in respect of: Data Modelling, ETL, Analytics and Data.

2.6.5 Administrators

In general, the technical skills most often requested by employers from big data Administrators at that time were: Linux, MySQL and Puppet, Hadoop and Oracle, whilst the process and methodological competences most often requested were in the areas of Configuration Management, Disaster Recovery, Clustering and ETL.

2.6.6 Project Managers

The specific types of Project Manager most often required by big data recruiters are Oracle Project Managers, Technical Project Managers and Business Intelligence Project Managers. Aside from Oracle (and in particular BI EE, EBS and EBS R12), which was specified in over two-thirds of all adverts for big data related Project Management posts, other technical skills often needed by applicants for this type of position were: Netezza, Business Objects and Hyperion. Process and methodological skills commonly required included ETL and Agile Software Development together with a range of more 'business focused' skills, i.e. PRINCE2 and Stakeholder Management.

2.6.7 Data Designers

The most commonly requested technical skills associated with these posts to have been Oracle (particularly BIEE) and SQL followed by Netezza, SQL Server, MySQL and UNIX. Common process and methodological skills needed were: ETL, Data Modelling, Analytics, CSS, Unit Testing, Data Integration and Data Mining, whilst more general knowledge requirements related to the need for experience and understanding of Business Intelligence, Data Warehouse, Big Data, Migration and Middleware.

2.6.8 Data Scientists

The core technical skills needed to secure a position as a Data Scientist are found to be: Hadoop, Java, NoSQL and C++. As was the case for other big data positions, adverts for Data Scientists often made reference to a need for various process and methodological skills and competences. Interestingly however, in this case, such references were found to be much more commonplace and (perhaps as would be expected) most often focused upon data and/or statistical themes, i.e. Statistics, Analytics and Mathematics.

2.6.9 Research and education

Researchers, scientists and academics are one of the largest groups for data reuse. COGNITWIN data published as open data will be used for further research and for educational purposes (e.g. thesis).

2.6.10 Policy making bodies

The COGNITWIN data and results might serve as a basis for decision making bodies, especially for policy evaluation and feedback on policy implementation on Environmental issues and sustainability related to Process Industry. This includes mainly the European Commission, national and regional public authorities.

3. FAIR Data

The FAIR principle ensures that data can be discovered through catalogues or search engines, is accessible through open interfaces, is compliant to standards to interoperable processing of that data, and therefore can be easily being reused.

3.1 Data findability

3.1.1 Data discoverability and metadata provision

Metadata is, as its name implies, data about data. It describes the properties of a dataset. Metadata can cover various types of information. Descriptive metadata includes elements such as the title, abstract, author and keywords, and is mostly used to discover and identify a dataset. Another type is administrative metadata with elements such as the license, intellectual property rights, when and how the dataset was created, who has access to it, etc. The datasets on the COGNITWIN Infrastructure are either added locally, by a user, harvested from existing data portals, or fetched from operational systems or IoT ecosystems. In COGNITWIN, the definition of a set of metadata elements is necessary in order to allow identification of the vast amount information resources managed for which metadata is created, its classification and identification of its geographic location and temporal reference, quality and validity, conformity with implementing rules on the interoperability of spatial data sets and services, constraints related to access and use, and organization responsible for the resource. In addition, metadata elements related to the metadata record itself are also necessary to monitor that the metadata created are kept up to date, and for identifying the organization responsible for the creation and maintenance of the metadata. An analysis will be done to identify relevant metadata schemas that can be applied.

3.1.2 Data identification, naming mechanisms and search keyword approaches

For data identification, naming and search keywords appropriate meta data registers will be used, managing unique identifiers for relevant assets. Registers provide a means to assign identifiers to items and their labels, definitions and descriptions (in different languages). A registry is a service giving access to semantic assets (e.g. application schemas, meta/data codelists, themes), and assigning to each of them a persistent URI. As such a metdata service can be considered also as a metadata directory/catalogue as well as a registry.

3.1.3 Data lineage

Data lineage refers to the sources of information, such as entities and processes, involved in producing or delivering an artifact. Data lineage records the derivation history of a data product. The history could include the algorithms used, the process steps taken, the computing environment run, data sources input to the processes, the organization/person responsible for the product, etc. Provenance provides important information to data users for them to determine the usability and reliability of the product. In the science domain, the data provenance is especially important since scientists need to use the information to determine the scientific validity of a data product and to decide if such a product can be used as the basis for further scientific analysis.

The provenance of information is crucial to making determinations about whether information is trusted, how to integrate diverse information sources, and how to give credit to originators when reusing information . In an open and inclusive environment such as the Web, users find information that is often contradictory or questionable. Reasoners in the Semantic Web will need explicit representations of provenance information in order to make trust judgments about the information they use. With the arrival of massive amounts of Semantic Web data (eg, via the Linked Open Data community) information about the origin of that data, ie, provenance, becomes an important factor in developing new Semantic Web applications. Therefore, a crucial enabler of the Semantic Web deployment is the explicit representation of provenance information that is accessible to machines, not just to humans. Data provenance as the information about how data was derived. Both are critical to the ability to interpret a particular data item. Provenance is often conflated with metadata and trust. Metadata is used to represent properties of objects. Many of those properties have to do with provenance, so the two are often equated. Trust is derived from provenance information, and typically is a subjective judgment that depends on context and use[2].

W3C PROV Family of Documents defines a model, corresponding serializations and other supporting definitions to enable the interoperable interchange of provenance information in heterogeneous environments such as the Web[3]. Current standards include[4]:

PROV-DM: The PROV Data Model - PROV-DM is a core data model for provenance for building representations of the entities, people and processes involved in producing a piece of data or thing in the world. PROV-DM is domain-agnostic, but with well-defined extensibility points allowing further domain-specific and application-specific extensions to be defined. It is accompanied by PROV-ASN, a technology-independent abstract syntax notation, which allows serializations of PROV-DM instances to be created for human consumption, which facilitates its mapping to concrete syntax, and which is used as the basis for a formal semantics[5].

PROV-O: The PROV Ontology - This specification defines the PROV Ontology as the normative representation of the PROV Data Model using the Web Ontology Language (OWL2). This document is part of a set of specifications being created to address the issue of provenance interchange in Web applications[6].

Constraints of the PROV Data Model - PROV-DM, the PROV data model, is a data model for provenance that describes the entities, people and activities involved in producing a piece of data or thing. PROV-DM is structured in six components, dealing with: (1) entities and activities, and the time at which they were created, used, or ended; (2) agents bearing responsibility for entities that were generated and activities that happened; (3) derivations of entities from entities; (4) properties to link entities that refer to a same thing; (5) collections forming a logical structure for its members; (6) a simple annotation mechanism[7].

PROV-N: The Provenance Notation - PROV-DM, the PROV data model, is a data model for provenance that describes the entities, people and activities involved in producing a piece of data or thing. PROV-DM is structured in six components, dealing with: (1) entities and activities, and the time at which they were created, used, or ended; (2) agents bearing responsibility for entities that were generated and activities that happened; (3) derivations of entities from entities; (4) properties to link entities that refer to the same thing; (5) collections forming a logical structure for its members; (6) a simple annotation mechanism. Figure 8 is a generic data lifecycle in the context of a data processing

[2] https://www.w3.org/2005/Incubator/prov/wiki/What_Is_Provenance
[3] https://www.w3.org/TR/prov-overview/
[4] https://www.w3.org/standards/techs/provenance#w3c_all
[5] https://www.w3.org/TR/2013/REC-prov-dm-20130430/
[6] https://www.w3.org/TR/2013/REC-prov-o-20130430/
[7] https://www.w3.org/TR/2013/REC-prov-constraints-20130430/

environment where data are first discovered by the user with the help of metadata and provenance catalogues[8].

During the data processing phase, data replica information may be entered in replica catalogues (which contain metadata about the data location), data may be transferred between storage and execution sites, and software components may be staged to the execution sites as well. While data are being processed, provenance information can be automatically captured and then stored in a provenance store. The resulting derived data products (both intermediate and final) can also be stored in an archive, with metadata about them stored in a metadata catalogue and location information stored in a replica catalogue.

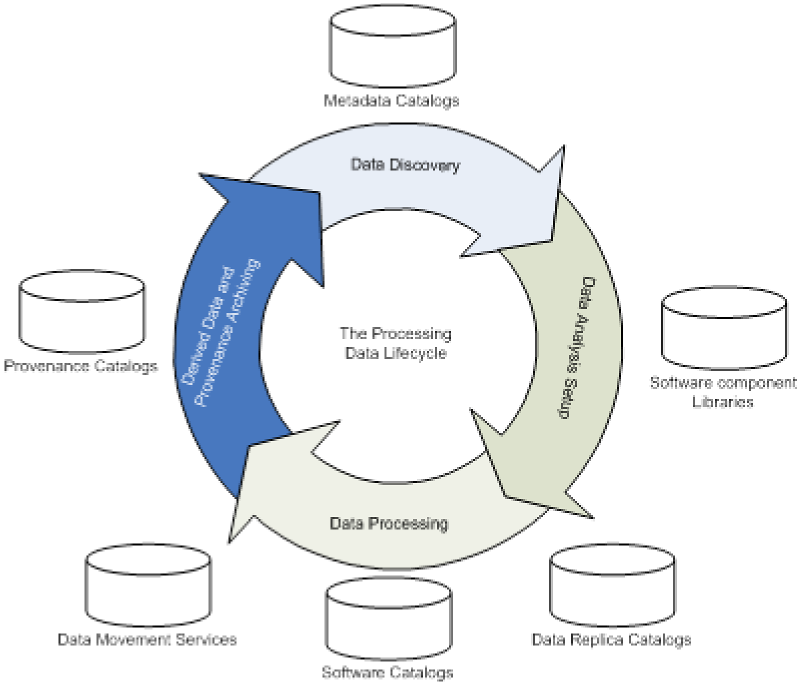

Figure 8: The processing data lifecycle

Data Provenance is also addressed in W3C DCAT Metadata model[9]. dcat:CatalogRecord describes a dataset entry in the catalog. It is used to capture provenance information about dataset entries in a catalog. This class is optional and not all catalogs will use it. It exists for catalogs where a distinction is made between metadata about a dataset and metadata about the dataset's entry in the catalog. For example, the publication date property of the dataset reflects the date when the information was originally made available by the publishing agency, while the publication date of the catalog record is the date when the dataset was added to the catalog. In cases where both dates differ, or where only the latter is known, the publication date should only be specified for the catalog record. W3C PROV Ontology [prov-o] allows describing further provenance information such as the details of the process and the agent involved in a particular change to a dataset. Detailed specification of data provenance is also additional requirements for DCAT – AP specification effort[10].

[8] https://www.w3.org/TR/2013/REC-prov-n-20130430/
[9] http://arxiv. org/ftp/arxiv/papers/1005/1005.2643.pdf
[10] http://adsabs.harvard.edu/abs/2014AGUFMIN34B..05D

3.2 Data accessibility

Through COGNITWIN experiments with a large number of tools and technologies a common data access pattern shall be developed. Ideally, this pattern is based on internationally adopted standards, relevant for process industry data and related to state of the practice in the SPIRE community.

3.2.1 Open data and closed data

Everyone from citizens to civil servants, researchers and entrepreneurs can benefit from open data. In this respect, the aim is to make effective use of Open Data. This data is already available in public domains and is not within the control of the COGNITWIN project.

All data rests on a scale between closed and open because there are variances in how information is shared between the two points in the continuum. Closed data might be shared with specific individuals within a corporate setting. Open data may require attribution to the contributing source, but still be completely available to the end user.

Generally, open data differs from closed data in three key ways[11]:

1. Open data is accessible, usually via a data warehouse on the internet.
2. It is available in a readable format.
3. It's licensed as open source, which allows anyone to use the data or share it for non-commercial or commercial gain.

Closed data restricts access to the information in several potential ways:

1. It is only available to certain individuals within an organization.
2. The data is patented or proprietary.
3. The data is semi-restricted to certain groups.
4. Data that is open to the public through a licensure fee or other prerequisite.
5. Data that is difficult to access, such as paper records that haven't been digitized.

Within the COGNITWIN project we mostly will have closed data only available for relevant groups within the project, but still relevant for a data management plan.

3.2.2 Data access mechanisms, software and tools

Data access is the process of entering a database to store or retrieve data. Data Access Tools are end user oriented tools that allow users to build structured query language (SQL) queries by pointing and clicking on the list of table and fields in the data warehouse.

Thorough computing history, there have been different methods and languages already that were used for data access and these varied depending on the type of data warehouse. The data warehouse contains a rich repository of data pertaining to organizational business rules, policies, events and histories and these warehouses store data in different and incompatible formats so several data access tools have been developed to overcome problems of data incompatibilities.

Recent advancement in information technology has brought about new and innovative software applications that have more standardized languages, format, and methods to serve as interface among different data formats. Some of these more popular standards include SQL, OBDC, ADO.NET, JDBC, XML, XPath, XQuery and Web Services.

3.2.3 Big data warehouse architectures and database management systems

[11] www.opendatasoft.com

Depending on the project needs, there are different possibilities to store data:

Relational Database

This is a digital database whose organization is based on the relational model of data. The various software systems used to maintain relational databases are known as a relational database management system (RDBMS). Virtually all relational database systems use SQL (Structured Query Language) as the language for querying and maintaining the database. A relational database has the important advantage of being easy to extend. After the original database creation, a new data category can be added without requiring that all existing applications be modified. This model organizes data into one or more tables (or "relations") of columns and rows, with a unique key identifying each row. Rows are also called records or tuples. Generally, each table/relation represents one "entity type" (such as customer or product). The rows represent instances of that type of entity and the columns representing values attributed to that instance. The definition of a relational database results in a table of metadata or formal descriptions of the tables, columns, domains, and constraints.

When creating a relational database, the domain of possible values can be defined in a data column and further constraints that may apply to that data value can be described. For example, a domain of possible customers could allow up to ten possible customer names but be constrained in one table to allowing only three of these customer names to be specifiable. An example of a relational database management system is the Microsoft SQL Server, developed by Microsoft. As a database server, it is a software product with the primary function of storing and retrieving data as requested by other software applications—which may run either on the same computer or on another computer across a network (including the Internet). Microsoft makes SQL Server available in multiple editions, with different feature sets and targeting different users.

PostgreSQL – for specific domains: PostgreSQL, often simply Postgres, is an object-relational database management system (ORDBMS) with an emphasis on extensibility and standards compliance. As a database server, its primary functions are to store data securely and return that data in response to requests from other software applications. It can handle workloads ranging from small single-machine applications to large Internet-facing applications (or for data warehousing) with many concurrent users; on macOS Server, PostgreSQL is the default database. It is also available for Microsoft Windows and Linux. PostgreSQL is developed by the PostgreSQL Global Development Group, a diverse group of many companies and individual contributors. It is free and open-source, released under the terms of the PostgreSQL License, a permissive software license. Furthermore, it is ACID-compliant and transactional. PostgreSQL has updatable views and materialized views, triggers, foreign keys; supports functions and stored procedures, and other expandability.

Big Data storage solutions

A NoSQL (originally referring to "non-SQL", "non-relational" or "not only SQL") database provides a mechanism for storage and retrieval of data which is modeled in means other than the tabular relations used in relational databases. Such databases have existed since the late 1960s, but did not obtain the "NoSQL" moniker until a surge of popularity in the early twenty-first century, triggered by the needs of Web 2.0 companies such as Facebook, Google, and Amazon.com. NoSQL databases are increasingly used in big data and real-time web applications. NoSQL systems are also sometimes called "Not only SQL" to emphasize that they may support SQL-like query languages.

Motivations for this approach include: simplicity of design, simpler "horizontal" scaling to clusters of machines (which is a problem for relational databases), and finer control over availability. The data structures used by NoSQL databases (e.g. key-value, wide column, graph, or document) are different from those used by default in relational databases, making some operations faster in NoSQL. The particular suitability of a given NoSQL database depends on the problem it must solve. Sometimes the data structures used by NoSQL databases are also viewed as "more flexible" than relational database tables.

MongoDB: MongoDB (from humongous) is a free and open-source cross-platform document-oriented database program. Classified as a NoSQL database program, MongoDB uses JSON-like documents with schemas. MongoDB is developed by MongoDB Inc. and is free and open-source, published under a combination of the GNU Affero General Public License and the Apache License.

MongoDB supports field, range queries, regular expression searches. Queries can return specific fields of documents and also include user-defined JavaScript functions. Queries can also be configured to return a random sample of results of a given size. MongoDB can be used as a file system with load balancing and data replication features over multiple machines for storing files. This function, called Grid File System, is included with MongoDB drivers. MongoDB exposes functions for file manipulation and content to developers. GridFS is used in plugins for NGINX and lighttpd. GridFS divides a file into parts, or chunks, and stores each of those chunks as a separate document.

3.3 Data interoperability

Data can be made available in many different formats implementing different information models. The heterogeneity of these models reduces the level of interoperability that can be achieved. In principle, the combination of a standardized data access interface, a standardized transport protocol, and a standardized data model ensure seamless integration of data across platforms, tools, domains, or communities. When the amount of data grows, mechanisms have to be explored to ensure interoperability while handling large volumes of data. Currently, the amount of data can still be handled using OGC models and data exchange services. We will need to review this element during the course of the infrastructure evolution. For now, data interoperability is envisioned to be ensured through compliance with internationally adopted standards.

Eventually, interoperability requires different phenotypes when being applied in various “disciplinary” settings. The following figure illustrates that concept (source: Wyborn 2017).

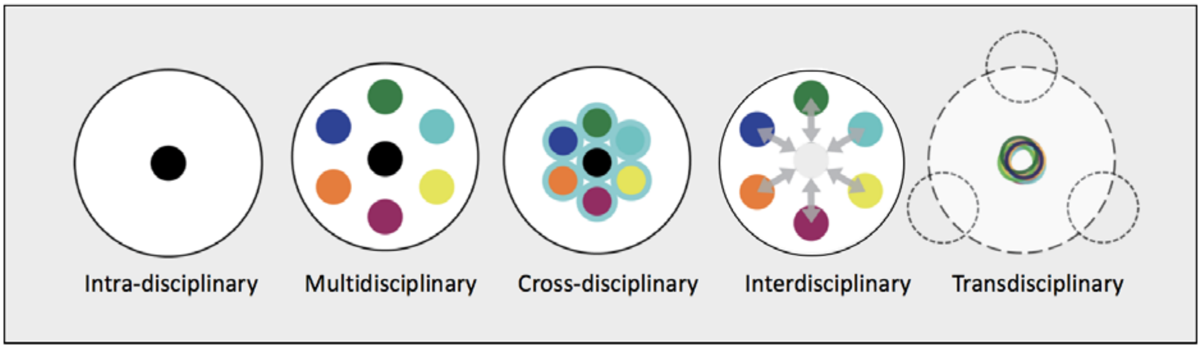

Figure 9: The "disciplinary data integration platform: where do you sit? (source: Wyborn)

The intra-disciplinary type remains within a single discipline. The level of standardization needs to cover the discipline needs, but little attention is usually paid to cross-discipline standards. The multi-disciplinary situation has many people from different domains working together, but eventually they all remain within their silos and data exchange is limited to the bare minimum.

The cross-disciplinary setting is what we are experiencing at the beginning of COGNITWIN. All disciplines are interfacing and reformatting their data to make it fit. The model works as long as data exchange is minor, but does not scale, as it requires bilateral agreements between various parties. The interdisciplinary approach is targeted in COGNITWIN. The goal here is to adhere to a minimum set of standards. Ideally, the specific characteristics are standardized between all partners upfront. This model adds minimum overhead to all parties, as a single mapping needs to be implemented per party (or, even better, the new model is used natively from now on). The transdisciplinary approach starts with data already provided as linked data with links across the various disciplines, well-defined vocabularies, and a set of mapping rules to ensure usability of data generated in arbitrary disciplines.

3.3.1 Interoperability mechanisms

Key to interoperable data exchange are standardized interfaces. Currently, the amount of data processing and exchange tools is large. We expect a consolidation of the number of tools during the first year of the project. We will revise the requirements set by the various pilots and the data sets made available regularly to ensure that proper recommendations can be given at any time.

Inter-discipline interoperability and ontologies

A key element to interoperability within and across disciplines are shared semantics, but the Semantic Web is still in its infancy and it is not clear to which extent it will become widely accepted within data intensive communities in the near future. It requires graph-structures for data and/or metadata, well defined vocabularies and ontologies, and lacks both the necessary tools to get COGNITWIN data operational within reasonable amounts of time. Therefore, at this stage it is mainly recommended to observe the topic of vocabularies and ontologies, but concentrate on initial base-vocabularies and their governance to ensure that at least base parameters are well defined.

Promoting data reuse

The reuse of data is a key component in FAIR. It ensures that data can be reused for purposes other than it was initially created for. This reuse improves the cost-balance of the initial data production and allows cross-fertilization across communities. COGNITWIN will advertise all the data produced to ensure that they are known to wider audience. In combination with standardized models and interfaces as described above and complemented with metadata and a catalog system that allows proper discovery, COGNITWIN can serve as valuable input outside of the project.

At this stage, it is not clear what licensing models need to be applied for the various data products produced in COGNITWIN. Generally, the focus shall be on public domain attribution and open licenses that maximize reusability in other contexts. All data products produced by COGNITWIN will be reviewed for FAIR principles once a year by the data producing organization. on the other hand, COGNITWIN is open to any third-party data and process provisioning. Data quality is a key component for data reuse. Without proper quality parameters, data cannot be integrated in external processes, as the level of uncertainty of the remote processes becomes undefined. COGNITWIN will review its data products for quality information provided as part of the metadata.

3.4 Data management support

3.4.1 FAIR data costs

The COGNITWIN consortium will handle both open data and data with restricted access. These data will be used by the project and the project pilots to demonstrate the power of big data. These data will be published through the COGNITWIN infrastructure. The current list of datasets and their details are to be further detailed. All data are either open data or data with restricted access provided for free to the consortium partners for project purposes. COGNITWIN does not foresee to purchase any data. The consortium has the knowledge and tools to make data FAIR, i.e. findable, accessible, interoperable and reusable. To make data FAIR is one of the project objectives and appropriate resources were allocated by each partner to cover costs for data harmonisation, integration and publication.

The COGNITWIN infrastructure has allocated appropriate resources to the sustainability of the infrastructure results. This includes the sustainability of FAIR data that are in the scope of the infrastructure. To satisfy the dataset reusability requirement, COGNITWIN anticipated several strategies for data storage and preservation. Dataset storage and preservation plan will include but not limited to disk drives, solid-state drives, in-memory functions and off-premises storage. Insofar as security concerns are not an issue, COGNITWIN partners will be encouraged to store data in the publicly available certified data repositories.

3.4.2 Big data managers

Managing Big Data also includes a specific structure or role-system, which means in fact types of people how manage or use Big Data in a specific way. Following chapter will describe the team structures for Big Data Management in COGNITWIN.

COGNITWIN will employ a two-layer approach for the management of the data used. On the first layer, the management of data provided in any of the participating institutions is done locally. On the second layer, data used in the context of COGNITWIN and needed in the context of data exchange or integration across organizations will be subject to the methodologies described within this document. These are enforced by the roles described below.

3.4.3 Project manager

COGNITWIN includes a diverse group of talented professionals, which have to be led. Beside the complex pilot-driven management structure.

3.4.4 Business Analysts

Business analysts are business-oriented domain experts, which are comfortable with data handling. They have deep insights in business requirements and logics and make sure that big data applications and platforms are capable to them. Business analysts are the connection between "non-technical" business user and technical developers. This includes techno-economic analysis as well as advanced visualisation services.

3.4.5 Data Scientists

Data scientists represent the data experts and analysis within the COGNITWIN consortium. They are able to turn raw data into purified insights and value with data science methods, techniques and tools. They have strong programming skills and can handle big data as well as linked data (incl. metadata). Furthermore, they are able to identify datasets for different requirements and develop solutions with regard to common standards. They are also able to visualise eloquently the results and findings.

One of the most important parts of COGNITWIN is making sense and value of data in different bio-economic sectors. In order to do so, methods, techniques and tools of machine learning are necessary to handle the huge amount of data.

3.4.6 Data Engineer / Architect

Data Engineers or Architects are data professionals who prepare the big data to be ready for analysis. This includes data discovery, data integration, data processing (and pre-processing) extraction and exchange as well as the quality control. Furthermore, they focus on design and architecture.

3.4.7 Platform architects

The data platform and its architecture is one of the most important part of COGNITWIN. In order to ensure a valid platform design, systems integration and platform development, high experienced platform architects are needed. IT/Operation manager

Some of the realized pilots will be very processing intensive, which requires a very good infrastructure. In order to provide and manage this infrastructure specific operation manager are needed.

3.4.8 Consultant

Big Data Consultant are responsible for support, guidance and help within all design and implementation phases. That includes high knowledge and practice in design big data solutions as well as develop data pipelines that leverage structured and unstructured data from multiple sources.

3.4.9 Business User

Business users are direct (business) beneficiaries of the developed COGNITWIN solutions. Further, they are important to specify detailed domain requirements and implement the solutions.

3.4.10 Pilot experts

In order to specify and prioritize requirements as well as manage the different pilots, finding synergies and connecting the different experts into the pilot, domain experts are needed.

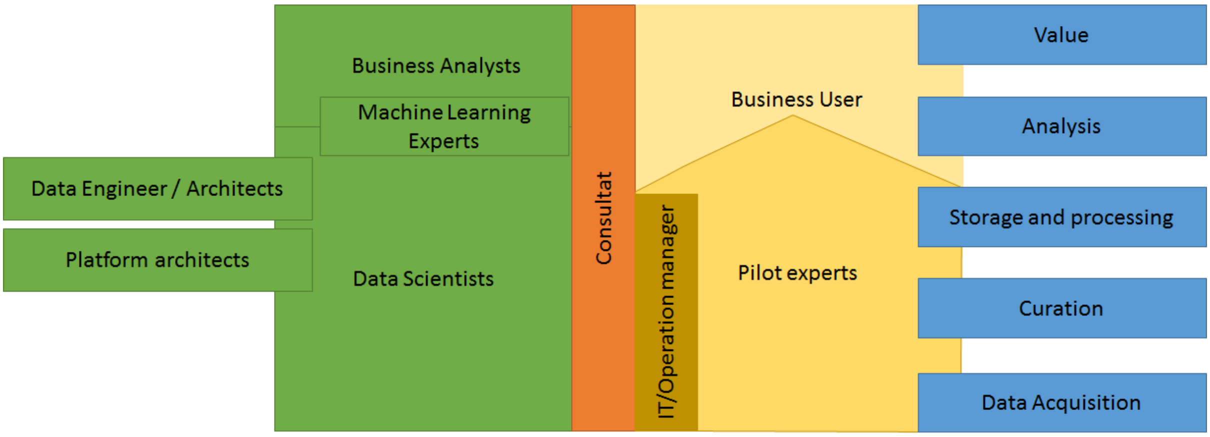

Figure 10: COGNITWIN's data managers

4. Data security

4.1.1 Introduction

In order to be able to address data security properly, one has to identify the various phases of data lifecycle, from their creation, to their use, sharing, archive and deletion. Handling project data securely throughout their lifecycle lays the foundations of a sensitive data protection strategy. In this context, the project consortium will determine specific security controls to apply in each phase, evaluating during the course of the project their level of compliance. Those data lifecycle phases are featured in the image below and are summarized as follows:

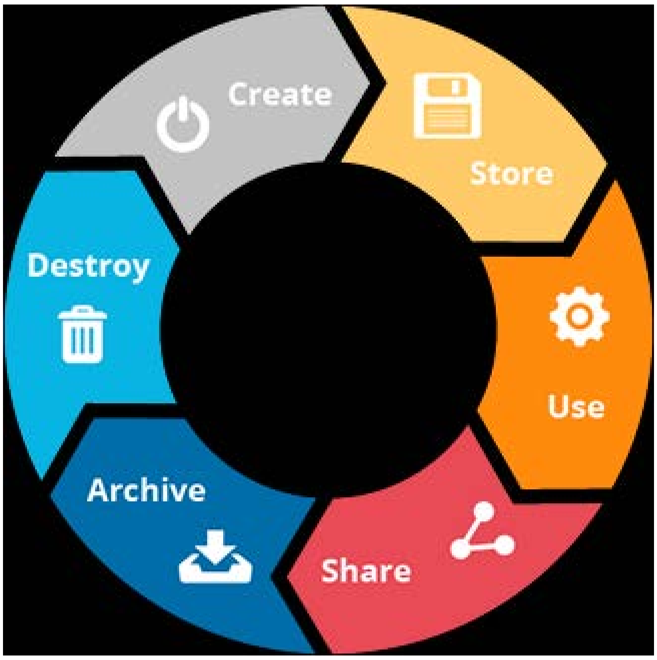

Figure 11: Data lifecycle

1. *Phase 1: Create*

This first phase includes the creation of structured or unstructured (raw) data. For the needs of the COGNITWIN project, those sensitive data are classified in the following categories: a) **Enterprise Data** (commercially sensitive data), b) **Personal Data** (personal sensitive data) and c) **other data** that are not applicable in one of the previous categories. Especially for the enterprise data, upon the creation phase already, security classification occurs based on an enterprise data security policy.

2. *Phase 2: Store*

Once data is created and included in a file, then it is stored somewhere. What needs to be ensured is that stored data is protected and the necessary data security controls have been implemented, so as to secure and minimize risk of information leak, ensuring efficient data privacy. More information about this phase is found in sections 5.2 about **data recovery** and 5.3 about **secure storage**.

3. *Phase 3: Use*

During this phase when data is viewed, processed, modified and saved, security controls are directly applied to data, with a focus on monitoring user activity and applying security controls to ensure data leak prevention.

4. *Phase 4: Share*

Data is constantly being shared between employees, customers and partners, necessitating a strategy that continuously monitors **data stores** and users. Data move among a variety of public and private storage locations, applications and operating environments, and are accessed by various data owners from different devices and platforms. That can happen at any stage of the data security lifecycle, which is why it's important to apply the right security controls at the right time.

5. *Phase 5: Archive*

In the case of data leaving active use but still needed to be available, they should be securely archived in appropriate storages, normally of low cost and performance, sometimes offline. This may cover also version control where older versions of original (raw) data files and data source processing programs are maintained in archive storages, per case. These backups are then stored and can be brought back online within a reasonable timeframe that will ensure that there is no detrimental effect of the data being lost or corrupted.

6. *Phase 6: Destroy*

In the case of data no longer needed, this data should be deleted securely so as to avoid any data leakage.

4.2 Data recovery

Data recovery strategy (also called disaster recovery plan) is not only a plan, but also ongoing process of minimizing a risk of data loss that can be a consequence of different random events.

Since COGNITWIN is a project dealing with Big Data scenarios, the context of data recovery is focused mostly on management procedures of data centers that are able to store and process significant amount of data. The disasters that can occur can be classified into two categories:

- Natural disasters (floods, hurricanes, tornadoes or earthquakes): because they cannot be avoided it is possible to minimize their effects on IT infrastructure (distributed backups)
- Man-made disasters (infrastructure failure, software bugs, hackers attacks): besides minimizing the effect it is possible to prevent them in different ways (regular software updates, good, active protection mechanisms, regular testing procedures)

The most important elements of Data recovery plan are:

- Backup management: well-designed automatic procedures for regular storing copies of datasets on separate machines or even geographically distributed places
- Replication of data to an off-site location, which overcomes the need to restore the data (only the systems then need to be restored or synchronized), often making use of storage area network (SAN) technology

- Private Cloud solutions that replicate the management data (VMs, Templates and disks) into the storage domains that are part of the private cloud setup.
- Hybrid Cloud solutions that replicate both on-site and to off-site data centers. These solutions provide the ability to instantly fail-over to local on-site hardware, but in the event of a physical disaster, servers can be brought up in the cloud data centers as well.
- The use of high availability systems which keep both the data and system replicated off-site, enabling continuous access to systems and data, even after a disaster (often associated with cloud storage)

The pilots in the project have various existing approaches for their data management which will be related to.

4.3 Privacy and sensitive data management

4.3.1 Introduction

With regards to privacy and sensitive data management, it is confirmed that these activities will be rigorously implemented in compliance to the privacy and data collection rules and regulations as they are applied nationally and in the EU, as well as with the national rules. The next sections include more specific information regarding those activities, rules and measures based on the classification of data made in the introduction of this section (5.1).

4.3.2 Enterprise Data (commercial sensitive data)

This category of data includes the (raw) data coming from specific sensor nodes and other similar data management systems and sources from the various project partners in each pilot case. They also include data about technologies and other assets protected by IPR and are considered to be highly-commercially sensitive, belonging to the providers that provides them for the various research and pilot activities within COGNITWIN project. Therefore, access to those data will be controlled and exchanges normally take place between specific end users and partners involved in their use and management within each pilot case for COGNITWIN related activities.

Following also project agreements, each partner who provides or otherwise makes available to any other project partner shared information represents that: (i) it has the authority to disclose this shared information, (ii) where legally required and relevant, it has obtained appropriate informed consents from all individuals involved, or from any other applicable institution, all in compliance with applicable regulations; and (iii) there is no restriction in place that would prevent any such other project partner from using this shared information for the purpose of COGNITWIN project and the exploitation thereof.

The abovementioned rules are also applied to any new data stemming from the project activities. This data will be also anonymised and protected and only based on the above rules our partners will be able to make data available to external industry stakeholders to utilise them for their own purposes. Related publications will be released and disseminated through the project dissemination and exploitation channels to make these parties aware of the project as well as appropriate access to any data (see Appendix A for COGNITWIN specific data). On a technical level, data are protected by IPRs are often accessed as a service, with specific access rights given under specific terms. Alternatively, they are shared encrypted or similarly protected with the keys provided under specific terms.

4.3.3 Personal Data

According to the Grant Agreement, it has been agreed by all partners that any Background, Results, Confidential Information and/or any and all data and/or information that is provided, disclosed or otherwise made available between the Parties **shall not include personal data**. Accordingly, each Party agreed that it will take all necessary steps to ensure that all Personal Data is removed from the Shared Information, made illegible, or otherwise made inaccessible (i.e. de-identify) to the other Parties prior to providing the Shared Information.

Therefore, no personal sensitive data are included in data exchanged between partners within COGNITWIN. Data created within project activities, e.g. some pilot activities, could initially involve personal and/or sensitive data from human participants, like location and id, COGNITWIN will apply specific security measures for their informed consent and data protection in line with the legislation and regulations in force in the countries where the research will be carried out, with most relevant rules to the project being the following:

- The Charter of Fundamental Rights of the EU, specifically the article concerning the protection of personal data
- Council Directive 83/570/EEC of 26 October 1983 amending Directives 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.

Regarding the procedure that is required in order to be able to participate in any COGNITWIN activities, we foresee that all potential participants will have to read and sign an informed consent form before starting the participation. This form aims to fully inform the participants about the study procedure and goals in order to guarantee that they have basic information in order to make the decision about whether to participate or not in the project activity. It shall include a summary and schedule of the study, the objectives and descriptions of the COGNITWIN system and its components. All participants have the right to receive a copy of the documents of this form. Participants will receive a generic user ID to identify them in the system and to anonymise their identities. No full names will be stored anywhere electronically. All gathered personal data shall be password protected and encrypted. Users' personal data will be safeguarded from other people not involved in the project. No adults unable to give informed consent will be involved.

It should be stated that the protection of the privacy of participants is a responsibility of all persons involved in research with human participants. Privacy means, that the participant can control the access to personal information and is able to decide who has access to the collected data in the future. Due to the principle of autonomy, the participants will be asked for their agreement before private and personal information is collected. It will be ensured that all persons involved in the project activities understand and respect the requirement for confidentiality. The participants will be informed about the confidentiality policy that is used in this research project.

4.4 General privacy concerns

Other privacy concerns will be addressed as following:

- External experts: Any external experts that will be involved in the project shall be required to sign an appropriate non-disclosure agreement prior to participating in any project related meeting, decision or activity.

- Publications: Hints to or identifiable personal information of any participant in (scientific) publications should be omitted. It is avoided to reveal the identity of participants in research deliberately or inadvertently, without the expressed permission of the participants.
- Dissemination: Dissemination of data between partners. This relates to access to data, data formats, and methods of archiving (electronic and paper), including data handling, data analyses, and research communications. Access to private information will be granted only to COGNITWIN partners for purposes of evaluation of the system and only in an anonymised form, i.e. any personally identifiable information such as name, phone number, location, address, etc. will be omitted.
- Protection: The lead project partner of every pilot case is responsible for the protection of the participants' privacy throughout the whole project, including procedures such as communications, data exchange, presentation of findings, etc.
- Control: The responsible project partners are not allowed to circulate information without anonymisation. This means that only relevant attributes, i.e. gender, age, etc. are retained.
- Information: As already mentioned above, the protection of the confidentiality implies informing the participants about what may be done with their data (i.e. data sharing). Individuals that participate in any study must have the right to request and obtain free of charge information on his/her personal data subjected to processing, on the origin of such data and on their communication or intended communication.

4.5 Ethical issues

In line with the Consortium's commitment in the COGNITWIN proposal, the ethics and responsibility work in the project is guided by the principles of responsible research and innovation in the information society (http://renevonschomberg.wordpress.com/implementing-responsible-research-andinnovation/), by the guidelines of European Group on Ethics (http://ec.europa.eu/bepa/european-groupethics). Since the research activities do not include any human trial, animal intervention or acquisition of tissues thereof, there are no ethical concerns.

Related to the area of AI and Machine Learning there is currently an international effort both in EU and in ISO SC42 on Ethics and AI. EU har provided Ethics guidelines for trustworthy AI[12] - and the project will ensure relationship to these guidelines, whenever relevant. The Partners agreed that any Background, Results, Confidential Information and/or any and all data and/or information that is provided, disclosed or otherwise made available between the Partners during the implementation of the Action and/or for any Exploitation activities ("Shared Information"), shall not include personal data as defined by Article 2, Section (a) of the Data Protection Directive (95/46/EEC) (hereinafter referred to as "**Personal Data**"). Accordingly each Partner agrees that it will take all necessary steps to ensure that all **Personal Data** is removed from the Shared Information, made illegible, or otherwise made inaccessible (i.e. de-identify) to any other Party prior to providing the Shared Information to such other Party.

Each Partner who provides or otherwise make available to any other Partner Shared Information ("Contributor") represents that: (i) it has the authority to disclose the Shared Information, if any, which it provides to the Partner; (ii) where legally required and relevant, it has obtained appropriate informed consents from all the individuals involved, or from any other applicable institution, all in

[12] https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

compliance with applicable regulations; and (iii) there is no restriction in place that would prevent any such other Partner from using the Shared Information for the purpose of the COGNITWIN Action and the exploitation thereof.

Any Advisory Board member or external expert shall be required to sign an appropriate non-disclosure agreement prior to participating in any project related meeting, decision or activity.

4.6 Conclusions

As COGNITWIN participates in the European set of data-oriented projects a Data Management Plan (DMP) is required and as a consequence, COGNITWIN project's datasets will be as open as possible and as closed as necessary, focusing on sound big data management for the sake of best research practice. The underlying motivation is to ensure that scientific articles and statements can be accompanied by associated data, and also to encourage a wider sharing of industrial data in Europe. European activities in the area of Open Science[13], Research Data Sharing[14] and Industrial Data sharing[15] is supporting this. The data management life cycle for the data to be collected, processed and/or generated by COGNITWIN project was described, accounting also for the necessity to make research data findable, accessible, interoperable and re-usable, without compromising the security and ethics requirements.

As a part of the project implementation, COGNITWIN's partners will be encouraged to adhere to sound data management to ensure that data are well-managed, archived and preserved. This is the first version of COGNITWIN DMP; it will be updated over the course of the project as warranted by significant changes arising during the evolution of the project implementation. The scheduled advanced releases of this document will particularly include information on the repositories where the data will be preserved, the security measures, and several other FAIR aspects.

[13] https://en.wikipedia.org/wiki/Open_science

[14] https://www.rd-alliance.org/

[15] https://ec.europa.eu/digital-single-market/en/policies/building-european-data-economy

Appendix A COGNITWIN Datasets

The following provides a general template for description data set (both open and with various protection/access levels) and a characterisation of each data set – which will be input for the further detailed meta data about the COGNITIWIN data sets:

Name	<Name of the data set>
Identifier	< e.g. D07.01 : D = "Data" , 07= partner number, 01 = sequential number 01, 02,...>
Owner	<Provider of the dataset/data model >
Description	<Describe the level of dataset (e.g. Landsat 8 archive) and not other more detailed levels (e.g. image from 2017/01/26)>
Classification(s)	<Keywords, e.g. EO data>
Date	<Date (range) when the resource will become or did become available >
Area coverage	<Geographical area of the dataset>
Time coverage	<Period of time that the dataset describes>
Format	<e.g. SENTINEL-SAFE format, Excel ...>
Licence	<Information about rights held in and over the dataset, including specific license name>
Related datasets	<Link to the related datasets, identifiers of the descriptions>
Data set size	<Indicative data size>
Frequency of update	<e.g. daily, yearly..>
Access interfaces	<e.g. SQL, REST>
Contact point	<Email of the contact person>

5. Data Management Plan - TEMPLATE

5.1 Standard elements of a data management plan

During the evolution of the COGNITWIN Data Management Plan we will ensure the answers to the questions below for all of the relevant data sets for the various pilots.

5.1.1 ADMIN DETAILS

Project Name:

Principal Investigator / Researcher:

5.1.2 DATA COLLECTION

1. What data will you collect or create?

2. What type, format and volume of data?
3. Do your chosen formats and software enable sharing and long-term access to the data?
4. Are there any existing data that you can reuse?
5. Are there any existing data or methods that you can reuse?
6. Do you need to pay to reuse existing data?
7. Are there any restrictions on the reuse of third-party data? Can the data that you create - which may be derived from third-party data - be shared?
8. Do you have sufficient storage?
9. Do you need to include costs for additional managed storage? Will the scale of the data pose challenges when sharing or transferring data between sites?
10. What types of data will you create?
11. Which types of data will have long-term value?
12. What format will your data be in?
13. Why have you chosen to use particular formats?
14. Do the chosen formats and software enable sharing and long-term validity of data?
15. How will the data be collected or created?

5.1.3 DOCUMENTATION AND METADATA

1. What documentation and metadata will accompany the data?
2. What information is needed for the data to be to be read and interpreted in the future? How will you capture / create this documentation and metadata?
3. What metadata standards will you use and why?
4. How will you capture / create the metadata?
5. Can any of this information be created automatically?
6. What metadata, documentation or other supporting material should accompany the data for it to be interpreted correctly?
7. What information needs to be retained to enable the data to be read and interpreted in the future?

5.1.4 ETHICS AND LEGAL COMPLIANCE

1. How will you manage any ethical issues?
2. Have you gained consent for data preservation and sharing?
3. How will you protect the identity of participants if required? e.g. via anonymisation
4. How will sensitive data be handled to ensure it is stored and transferred securely?
5. How will you manage copyright and Intellectual Property Rights (IPR) issues?
6. Who owns the data?
7. How will the data be licensed for reuse?
8. Are there any restrictions on the reuse of third-party data?
9. Will data sharing be postponed / restricted e.g. to publish or seek patents?

5.1.5 STORAGE AND BACKUP

1. How will the data be stored and backed up during the research?
2. Do you have sufficient storage or will you need to include charges for additional services?
3. How will the data be backed up?
4. Who will be responsible for backup and recovery?

5. How will the data be recovered in the event of an incident?
6. Where will the data be stored?
7. How will the data be backed up? i.e. how often, to where, how many copies, is this automated…
8. Who will be responsible for storage and backup?
9. Do you have access to enough storage or will you need to include charges for additional services?
10. How will you manage access and security?

5.1.6 SELECTION AND PRESERVATION

1. Which data are of long-term value and should be retained, shared, and/or preserved?
2. What data must be retained/destroyed for contractual, legal, or regulatory purposes? How will you decide what other data to keep? What are the foreseeable research uses for the data? How long will the data be retained and preserved?
3. Which data are of long-term value and should be shared and/or preserved? How will you decide what to keep?
4. What is the long-term preservation plan for the dataset?
5. Where e.g. in which repository or archive will the data be held? What costs if any will your selected data repository or archive charge? Have you costed in time and effort to prepare the data for sharing / preservation?
6. What is the long-term preservation plan for the dataset? e.g. deposit in a data repository Will additional resources be needed to prepare data for deposit or meet charges from data repositories?
7. Where (i.e. in which repository) will the data be deposited?

5.1.7 DATA SHARING

1. How will you share the data?
2. How will potential users find out about your data?
3. With whom will you share the data, and under what conditions?
4. Will you share data via a repository, handle requests directly or use another mechanism?
5. When will you make the data available?
6. Will you pursue getting a persistent identifier for your data?
7. How will you make the data available to others?
8. Are any restrictions on data sharing required?
9. What action will you take to overcome or minimise restrictions?
10. For how long do you need exclusive use of the data and why? Will a data sharing agreement (or equivalent) be required?
11. Are any restrictions on data sharing required? e.g. limits on who can use the data, when and for what purpose.
12. What restrictions are needed and why?
13. What action will you take to overcome or minimise restrictions?

5.1.8 RESPONSIBILITIES AND RESOURCES

1. Who will be responsible for data management?
2. Who is responsible for implementing the DMP, and ensuring it is reviewed and revised?

3. Who will be responsible for each data management activity?
4. How will responsibilities be split across partner sites in collaborative research projects?
5. Will data ownership and responsibilities for RDM be part of any consortium agreement or contract agreed between partners?
6. How are responsibilities split across partner sites in collaborative research projects?
7. What resources will you require to deliver your plan?
8. Is additional specialist expertise (or training for existing staff) required?
9. Do you require hardware or software which is additional or exceptional to existing institutional provision?
10. Will charges be applied by data repositories?
11. What additional resources are needed to deliver your plan?
12. Do you have sufficient storage and equipment or do you need to cost in more?
13. Have you costed in time and effort to prepare the data for sharing / preservation?