Universidad de Valladolid

ESCUELA DE INGENIERÍAS INDUSTRIALES

UNIVERSIDAD DE VALLADOLID

ESCUELA DE INGENIERIAS INDUSTRIALES

Grado en Ingeniería de Organización Industrial

RESEARCH DATA MANAGEMENT

Autor:
Araceli Amo Cubillo

Responsable de Intercambio en la UVA:
Francisco Javier Rey

Universidad de destino:
Technische Universität Dresden

Dresden, July 2018

TFG realizado en programa de intercambio

TÍTULO:	Research Data Management
ALUMNO:	Araceli Amo Cubillo
FECHA:	30/06/2018
CENTRO:	Technische Universität Dresden (Maschinenwesen)
TUTOR:	Prof. Dr.-Ing. Alexander Brosius

Resumen

Esta tesis ofrece una revisión de la actual gestión de datos en el campo de la ciencia e investigación financiada con fondos públicos. El objetivo principal es ofrecer una herramienta de soporte durante el ciclo de vida de aquellos datos que se recopilan, procesan o generan en un proyecto científico.

Primero, se expone una idea general sobre las características y técnicas principales para el manejo de grandes cantidades de datos, y se presenta el papel que desempeña Big Data en el campo de la investigación. Además, se discute la recopilación de datos, el intercambio y el almacenamiento tanto a corto como a largo plazo. También se exponen conceptos relacionados con la seguridad de los datos.

Una vez presentado el estado del arte actual sobre la gestión de datos de investigación, se propone un plan de gestión de datos específico para el experimento llevado a cabo por la TU Dresden "Macro y microestructura de herramientas de embutición profunda para conformado en seco". En esta segunda parte se proporciona un análisis de los principales elementos de una política de gestión de datos, y se describe un posible tratamiento para los datos creados en el experimento anteriormente mencionado.

Palabras clave

Gestión de datos, Big Data, investigación, Cloud Computing, selección de datos, almacenamiento y seguridad.

Abstract

This thesis offers a review of the actual data management in the public funded science and research field. The main purpose is to offer a support tool during the life cycle of those data that are collected, processed or generated in a scientific project.

First it is exposed a general view of the characteristics and main technics to handle big amount of data, and it is presented the role Big Data plays in researching. Furthermore, it is discussing data collection, sharing and data storage in the short and long-term. Issues concerning about data security are also exposed.

Once it is presented the state of the art about the research data management, it is proposed a specific data management plan for the experiment carry on by the TU Dresden “Macro- and micro-structure of deep-drawing tools for dry forming”. In this second part an analysis of the main elements of a data management policy is provided, describing a proposed treatment received by the research data collected in the mentioned experiment.

Keywords

Data management, Big Data, research, Cloud Computing, data selection, storage and security.

Universidad de Valladolid

TECHNISCHE UNIVERSITÄT DRESDEN

BACHELOR THESIS

FACULTY OF MECHANICAL SCIENCE AND ENGINEERING

RESEARCH DATA MANAGEMENT

Author:
Araceli Amo Cubillo

Tutor:
Brosius, Alexander
(Institute of Manufacturing Science and Engineering)

Dresden, July 2018

“We must all accept that science is data and that data are science, and thus provide for, and justify the need for the support of, much-improved data curation.”
(Hanson, Sugden, & Alberts,2011)

0.Index

1.Introduction

Since the improvement in the experimental and researching techniques due to the new technologies, the possible collection of digital data generated in a scientific experiment has started to grow rapidly. Today, there is tremendous amount of data generated every day in the sector of science and research. Proper processing of the data could reveal new knowledge and enable us to react to emerging opportunities and changes in a timely manner. However, the growth of the data volume in our digital world seems to out speed the advance of our computing infrastructure and our data management plans. The challenge posed by data volume is most noticeable. In science, researchers encounter computing and organizational limitation constantly due to the increasing data volume.

The new technologies not only present advantages in facts of increasing data amount, also because digital data could be shared, replicated and recombinable, it presents tremendous reuse opportunities. To enable reuse, data must be well preserved. In some cases, the effects of data loss are economic, because experiments must be re-run. In other cases, data loss represents an opportunity lost forever. The data now is rightly view as significant resources which should be safe to recover the investment made and preserve the scientific record. The researchers now need to implement data-management and data-sharing plans that address the full life cycle of data; including what happens after the project is finished.

The cost of data management, not only in proper infrastructures but also in personnel dedicating time not always are affordable to all the research projects. There are vast numbers of scientific research projects producing large amount of data which is not possible to organize, store and share.

The challenge of the research community is not to lose knowledge due to a poor data management. The only way to avoid this, is creating proper data management plans.

2. STATE OF THE ART

STATE OF THE ART

2.1 Characteristics of the actual research data management

The Research Data Management is the active process of managing the data generated in an investigation. It is carried out continuously and covers all the factors related to the management of data throughout its life cycle, beginning at the planning stage of the research and covering its execution, the formulation of results and the preservation of the data sets so that they are accurate, complete, authentic and reliable, and remain accessible and reusable over time. [1]

Data has always been the fundamental raw material for research and the advancement of knowledge. However, in the recent years, the data management has gained more importance for the research communities. There are several factors that explain this development, such as, for example, the technological advance in recent decades and the changes that it has implied in scientific practice. [1] We present some of these changes, which will allow us to better understand the situation in which the Research Data Management is located today:

- **E-Science:** The technological advance of the last decades has impacted on the volume of existing information. In the field of research, the increase in the volume of available data is mainly due to the advance of simulations thanks to the use of supercomputers, the realization of large-scale experiments and the development of sensors that can produce a large quantity of data of different type.

  Since the beginning of the 2000s it was anticipated that these advances would have as a consequence the overflow of data management capacity and the consequent development of new ways of doing science, based on networks and collaboration, and the creation of technological infrastructures to allow access to large volumes of data, their processing and visualization. It is in this context that we see the emergence of the term e-science or e-research. [2]

- **Open Data:** the meaning of this concept by Open Knowledge International is: "Open data is data that can be used, reused and redistributed freely by anyone, and that are subject, at most, to the requirement of attribution and to be shared in the same way they appear."

  Increasing collaboration between researchers and taking advantage of the potential of large and complex datasets available for research, involves the sharing of data. For this it is necessary to open the data at the most level as possible. [3]

- **Open Science:** The volume, richness and complexity of the data available for research have also been important in what is now called Open Science or Science 2.0. This science refers to a kind of research based on an open access to public scientific information, thanks to ICT tools and platforms. Also, a wider collaboration, involving no-scientific participants. Open Science is a concept which involves openness, collaboration, communication and the use of technologies for science. [4]

2.2 Big Data

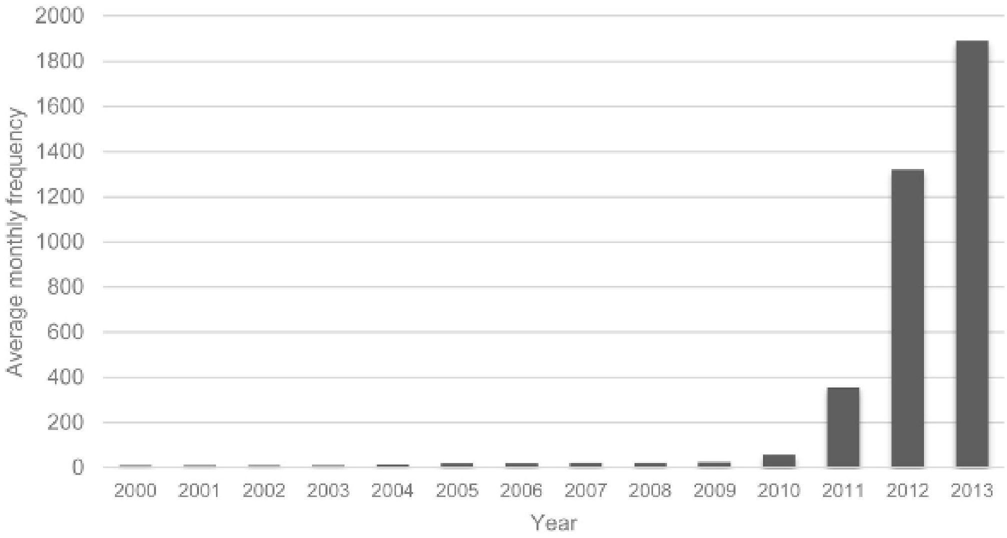
What is Big Data?

Every day, 2,5 quintillion bytes of data are created. These data come from digital pictures, videos, posts to social media sites, intelligent sensors, purchase transaction records, cell phone GPS signals, to name a few. This is known as Big Data. [6]

Big data is an evolving term that describes any voluminous amount of structured, semistructured and unstructured data that has the potential to be mined for information. Big data refers to a process that is used when traditional data mining and handling techniques cannot uncover the insights and meaning of the underlying data. Data that is unstructured or time sensitive or simply very large cannot be processed by relational database engines. This type of data requires a different processing approach called big data. [5] [6] [7]

Big Data as a concept is a nascent and has uncertain origins. Diebold (2012) argues that the term “big data…probably originated in lunch-table conversations at Silicon Graphics Inc. (SGI) in the mid-1990s”. Despite the references to the midnineties, **Fig.1** shows that the term became widespread as recently as in 2011. [8]

Fig.1 Frequency distribution of documents containing the term “big data” in ProQuest Research Library.

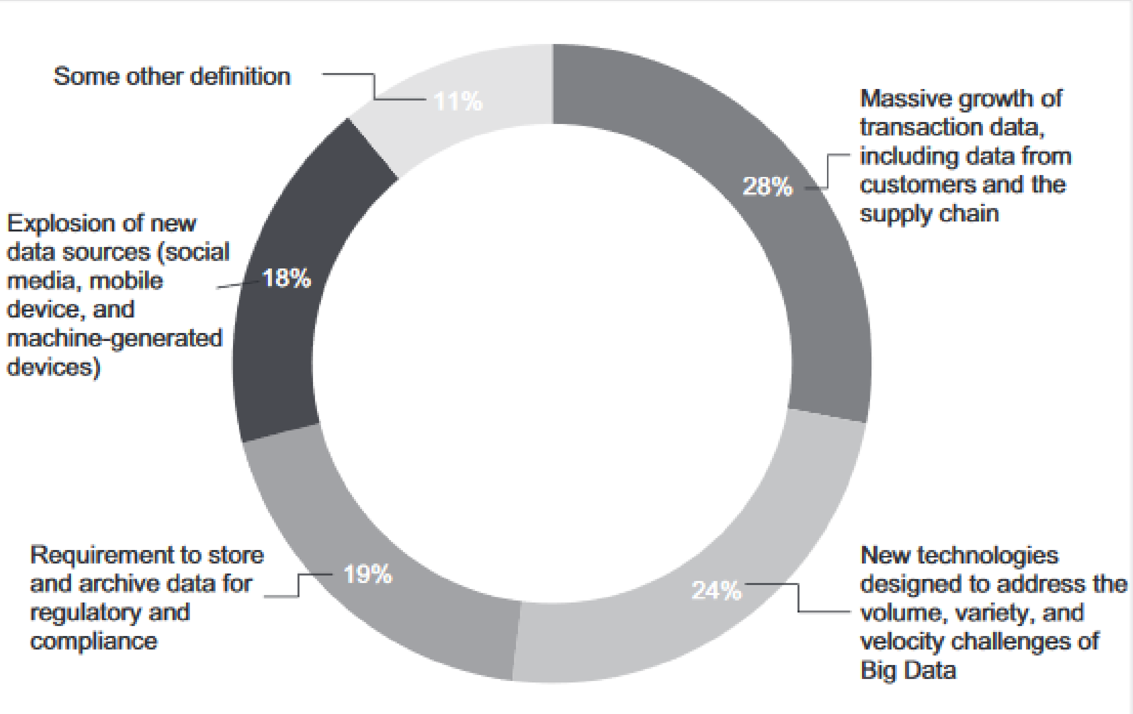


Source: A.Gandomi, M. Haider /International Journal of Information Management 35 (2015) 137-144

The current meaning can be attributed to the promotional initiatives by IBM and other leading technology companies who invested in building the niche analytics market.

Big data definitions have evolved rapidly, which has raised some confusion. This is evident from an online survey of 154 C-suite global executives conducted by Harris Interactive on behalf of SAP in April 2012. **Fig. 2** shows how executives differed in their understanding of big data, where some definitions focused on what it is, while others tried to answer what it does. [8]

Fig. 2 Definitions of big data based on an online survey of 154 global executives in April 2012



Source: Gandomi, A., & Haider, M. (2015): "Beyond the hype: Big data concepts, methods, and analytics.", International Journal of Information Management,35(2), 139

Three V's

Size is the first feature that comes to mind considering the question "what big data is?". However, other characteristics of big data have emerged and show that size is only one dimension, but the others are equally important.

Big data is often characterized by 3Vs: the extreme volume of data, the wide variety of data types and the velocity at which the data must be processed.

-*Volume* refers to the magnitude of data. Although big data doesn't equate to any specific volume of data, the term is often used to describe terabytes, petabytes and even exabytes of data captured over time. Definitions of big data volumes are relative and vary by factors, such as time and the type of data. What may be deemed big data today may not meet the threshold in the future because storage capacities will increase, allowing even bigger data sets to be captured. [9]

-*Variety* refers to the structural heterogeneity in a dataset. Technological advances allow firms to use various types of structured, semi-structured, and unstructured data. Structured data, which only constitutes 5% of all existing data (Cukier, 2010), refers to the tabular data found in spreadsheets or relational databases. Text, images, audio, and video are examples of unstructured data, which sometimes lack the structural organization required by machines for analysis. Spanning a continuous between fully structured an unstructured data, the format of semi-structured data does not conform to strict standards. Extensible Markup Language (XML), a textual language for exchanging data on the Web, is a typical example of semi-structured data. [8][9]

The emergence of new data management technologies and analytics, which enable organizations to leverage data in their business processes, is the innovative aspect of big data era.

-*Velocity* refers to the rate at which data are generated and the speed at which it should be analysed and acted upon. The proliferation of digital devices such as smartphones and sensors has led to an un precedent rate of data creation and is driving a growing need for real-time analytics. Traditional data management systems are not capable of handling huge data feeds instantaneously. This is where big data technologies come into play. They enable firms to create real-time intelligence from high volumes of data.

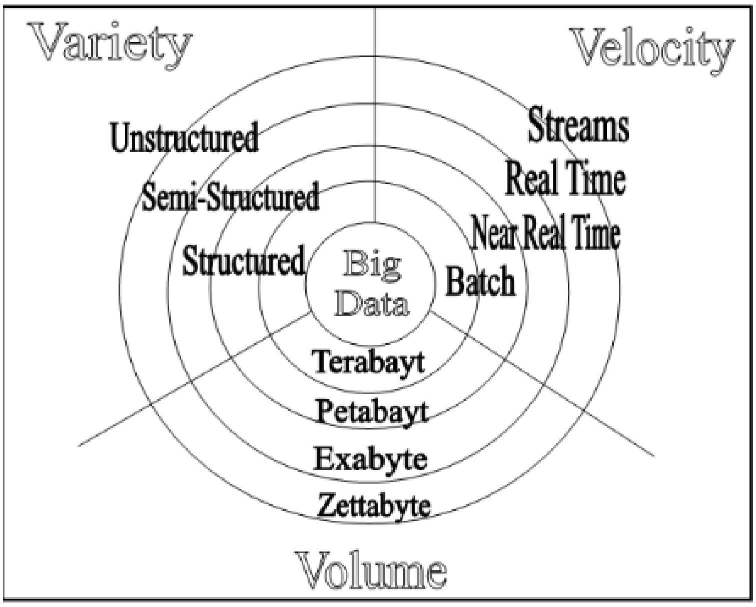


Fig. 3 The Three V's f big data

In addition to the three V's, other dimensions of big data have also been mentioned. These include *Veracity*, which represents the unreliability inherent in some sources of data. The need to deal with imprecise and uncertain data is another factor of big data, which is addressing using tools and analytics developed for management and mining of uncertain data. [8][9]

Two additional dimensions of big data are *Variability and complexity.* The first one refers to the variation in the data flow rates. Often, big data velocity is not consistent and has periodic peaks and troughs. Complexity refers to the fact that big data are generated through a myriad of sources. This imposes a critical challenge: the need to connect, match, cleanse and transform data received from different sources.

Moreover, *Value* is another attribute needed to define big data. The data received in the original form usually has a low value relative to its volume. However, a high value can be obtained by analysing large volumes of such data. [8]

Is important the fact that these dimensions are not independent of each other. As one dimension changes, the likelihood increases that another dimension will also change as a result.

2.3 What role does Big Data play in the public funded science and research field?

The ways that scientific research is practiced have shifted fundamentally in the last several decades. Thanks to Big Data, the scientific research field is able to manage with a huge amount of information that was impossible to handle before the big data revolution. Consequently, a range of opportunities in many scientific areas is opened.

The growth of data in general, and scientific research data in particular, has been driven by a number of social and technological factors. New technologies not only made it possible to gather data more quickly and cheaply, but also it is possible to store it. [12]

New technologies have also led to an increase in the amount of "born-digital" data – materials that are originally created in digital form, rather than those that are created as analogy data and subsequently digitized. For example, the widespread adoption of electronic health records means increasingly easy access to a wealth of patient data that was once difficult to utilize because it sat in hand-written paper files in clinicians' offices. Digitizing all of those records would have been an unreasonably difficult and time-consuming task, but now that patient data are gathered in electronic form as a matter of course. New types of data arising out of the internet, such as social media data, also have sometimes surprising research potential. Researchers have used social media sites like Twitter and Facebook for unexpected research purposes, like pharmacovigilance and disease surveillance.

Researchers of the 21st century often rely on large digital datasets, and sometimes they are using data that they themselves did not gather, but that they obtained from public sources for reuse. Researchers in many fields must comply with new policies from funders and journals that require them to share their data or write data management plans, tasks that most researchers have not had to undertake in the past. Whatever a researcher's particular field of study, chances are good that the ways that research is conducted have substantively changed in the last several years. As a result, researchers may find that they need new skills and knowledge to work most effectively and take advantage of the new opportunities that this age of data-driven research presents. [12]

Opportunities

There is no doubt that Big Data and especially *what we do with it* has the potential to become a driving force for innovation and value creation in the research and science field. [10] Some of the ways in which using big data can be an opportunity are the next:

Value Creation

There is no doubt that Big Data and especially *what we do with it* has the potential to become a driving force for innovation and value creation There is no doubt that Big Data and especially *what we do with it* has the potential to become a driving force for innovation and value creation. These are some of the ways in which using big data can create value.

- First, big data can unlock significant value by making information transparent and usable at much higher frequency.
- Second, as researching teams create and store more transactional data in digital form, they can collect more accurate and detailed performance information.
- Third, sophisticated analytics can substantially improve decision-making. [13] [15]

Accelerating Investigations

Many of the scientific research performed involve analysing both structured and unstructured data sources. Thanks to big data technologies, this process is more quickly and accurately. Big data technologies can capture the information and create algorithms to produce an instant roadmap into an investigation, which not only reduce budget spending but also the time spent on it. [14] [10]

Culture of sharing

The benefit of sharing data is obvious, and it could be justificated with the next four rationales:

1) to reproduce or to verify research,
   The only way to claim the veracity of an innovative research is the availability to reproduce or replicate it. The main problem is that is well known that the research activities are difficult to replicate, because the experimental conditions as well as the processing and analysis tools rarely maintain precise record of the systems or the data at each transaction. Consequently, this argument is the most problematic of sharing data. Separating data from context is a risky matter which must be taking carefully in order to not to draw the wrong conclusions.

2) to make results of publicly funded research available to the public,
   It should not be forgotten that the main goal of research is to serve the public good. With this point of view, data produced with public funds should be available for use not to be saved by researchers.

3) to enable others to ask new questions of extant data,
   This aspect refers to the ability to combine data. The greatest advantages of data sharing may be the combination of data from multiple sources. The end of a project for one researcher could be the beginning for others to ask new questions and continue the innovative process.

4) to advance the state of research and innovation,
   Thanks to the culture of sharing the use of data is maximized, the impact of findings is increased and consequently, the state of research progress faster.

A research project could affect multiple and overlapping communities of interest. Each of which may have different points of view on how to use the data. The boundaries of communities of interest are neither clear nor stable. The lack of communication between communities with the same specific proposal can lead to the duplication of efforts in the same area.

The culture of sharing aim is re-use the data between research groups to advance more effectively and make easier the work of all the scientists, because after all, the research community pursue the same objective: improve society. [16]

Challenges

Opportunities are always followed by challenges. On the one hand, Big Data bring many attractive advantages. On the other hand, we are also facing a lot of challenges when handle big data problems, difficulties lie in data capture, storage, searching, sharing, analysis, and visualization. [15]

Lack of skilled personal and sufficient resources

Big data projects require big data skills, but big data experts are hard to find. That's largely because big data is a new field, so the big data skills required for different projects are still ill-defined. Moreover, due to the technology is relatively new, the research organizations need to develop a new infrastructure which requires extensive inversions. The cost of a big data project could be unaffordable for many scientist teams, it lies in taking into account all costs of the project from acquiring new hardware, to paying a provider, to hiring additional personnel. Finding subsides or funders to support the start-up of this kind of project is usually a complicated task.

Distinction between data and knowledge

Clearly, big data provides a huge amount of information, but is scientist responsibility be able to generate the useful knowledge from it. Nate Silve (big data statistician) said '*most of the data is just noise, as most of the universe is filled with empty space*'. We will need to discover new patterns and correlations in the sea of data to offer new insights and conclusions. It is needed not only to put the right talent and technology in place but also structure workflows and incentives to optimize the use of big data. If the research field follow this way, it would take benefit from the big data opportunities but if not, it would be lost in a sea of data.

Sharing problematic

Research collaboration is not always easy. Many factors involved make the researchers unwilling to cooperate. They compete for grants, for jobs, for publications, and for students. The time and resources are limited so they need to think carefully where and when spend them. Time and money spent on preparing data which will be used for others are resources not spent in data collection, analysis, equipment, publication fees, conference travel, writing papers and proposals, or other research necessities.

Data release is costly. Even if data sharing is built into the cost of research funding, the process may increase the total cost of doing research.

Data Store well-preserved

Big data storage is concerned with storing and managing data in a scalable way, satisfying the needs of applications that require access to the data. The ideal big data storage system would allow storage of a virtually unlimited amount of data, cope both with high rates of random write and read access, flexibly and efficiently deal with a range of different data models, support both structured and unstructured data, and for privacy reasons, only work on encrypted data. Obviously, all these needs cannot be nowadays fully satisfied.

To enable reuse of the data in research and science field, data must be well preserved. The effects of data loss could be the re-run of the whole experiments, or what it is worse an opportunity lost forever. [10] [13] [14] [15]

2.4 Types of data generated

Nowadays, many applications in various domains are able to generate different types of big data. Heterogeneity is a natural property of the big data. In real-world applications, data often does not come from a single source. Big data implementations require handling data from various sources, in which data can be of different formats and models. This bring forth the challenge of data variety. The variety of data provides more information to solve problems or to provide better service. The question is how to capture the different types of data in a way that makes it possible to correlate their meanings. Typically, data can be classified into three general types—structured data, semi-structured data and unstructured data. [17]

- **Structured Data:** concerns all data which can be stored in database SQL in table with rows and columns. They have relational key and can be easily mapped into pre-designed fields. Structured data is highly organized information that uploads neatly into a relational database

  Structured data is relatively simple to enter, store, query, and analyse, but it must be strictly defined in terms of field name and type. [18] [20]

- **Unstructured Data**: may have its own internal structure but does not conform neatly into a spreadsheet or database. Such type of data becomes difficult and requires advance tools and software to access information. Today more than 80% of the data generated is unstructured.

  For Example, images and graphics, pdf files, word document, audio, video, emails, PowerPoint presentations, webpages and web contents, wikis, streaming data, location coordinates etc. [18]

- **Semi-structured Data**: Semi-structured data is information that doesn't reside in a relational database but that does have some organizational properties that make it easier to analyse. It is basically a structured data that is unorganised. Due to unorganized information, the semi-structured is difficult to retrieve, analyse and store as compared to structured data.

  Examples of semi-structured: CSV but XML and JSON documents are semi structured documents, NoSQL databases are considered as semi structured. [18] [19] [20]

The data resulted of researching is called scientific data. It is collected from data intensive experiments or applications. Examples are celestial data, high-energy physics data, genome data, health-care data…etc. This kind of data depends on the application, ranging from structured data to semi-structured data and unstructured data. Moreover, not only the original data is important in this sector, but also data provenance (recording how data are generated and transformed).

It is important to keep in mind that there is no single definition of scientific data, and that different disciplines or communities may differ in their understanding of this concept. Thus, there may be large differences between what constitutes a piece of data for researchers in the humanities, the arts or the sciences. In any of these cases, it is important to bear in mind that the data can be of a quantitative or qualitative nature, and that they can come in many formats and media, be they physical or digital. [17]

Apart from the previous classification of scientific data, there are other ways to identify and organize them:

- **According to the level of processing in the course of the investigation**

- Primary or unprocessed data: Original data that has been collected but has not yet been processed or analyzed. Some examples are sound records, observations, field notes or experiment data.

- Data processed: Data that has been digitized, translated, transcribed, cleaned, validated, and/or verified.

-Analyzed data: Models, graphs, tables, texts or others, that have been created from the primary and processed data, and that are intended to be helpful in the discovery of useful information, the presentation of conclusions and decision making.

- **According to the origin source**

- Canonical or reference data: Sets of data that can be used for validation, comparison, or search of information (for example, human genome sequences, chemical structures or spatial data portals)

- Experimental data: Data generated in scientific experiments. They are generally reproducible and can be generated by laboratory equipment.

- Models or simulations: Data generated in computers by algorithms, mathematical models, or simulations of experiments.

- Derived data: Data sets created by taking existing data and performing some type of manipulation on them.

- Observations: Data generated by recording observations of a specific and possibly unrepeatable event, in a given place and time. [21]

- **According to its form or type**

Electronic text documents, data spreadsheets, field notebooks or laboratory annotations, questionnaires or transcripts, photographs or movies, sound records, samples, artifacts, specimens, digital objects, models, algorithms, scripts, databases, metadata, metadata schemas, software configurations and pre or post software processing files.

- **According to format**

Some examples could be:
-Textual (Microsoft Wod, PDF, RTF, ODT, etc.)
-Numeric (Excel, CSV, etc.)
-Multimedia (JPEG, MPEG, WAV, etc.)
-Structured (XML, MySQL database, etc.)
-Software code (Java, C, etc.)
-Specific to software (Mesh, 3D CAD, statistical model, etc.)
-Specific to a discipline (FITS in astronomy, CIF in chemistry, etc.)
-Specific to an instrument (Olympus Confocal Microscope Data Format, Carl Zeiss Digital Microscopic Image Format)

2.5 Roles in the research data management

[22] [23] Due to the wide range of activities carried on, the long time spent, the purpose and capacities required, the data management is a responsibility shared between several actors involved in the development and communication of a research.

To achieve a successful implementation of the research data management is important all the actors involved in the project to know their own responsibilities and roles in the management process.

❖ **Researchers:** the main important role of the researchers is carry on in two moments:

During the data collection and the time of the investigation:
-Ensure that good practices are implemented while the investigation.
-Fulfill ethical and legal requirements.
-Make decisions about opening, placing, reusing and selecting data, among other aspects.

During the data reuse created by third parties:
-Follow to the reuse conditions determined by the data owners.
-Cite the origin of the data whenever necessary.

❖ **Data managers**: some of the main responsibilities of this participants are work with the researchers in the data collection and analysis, technology surveillance, and once the project is finished make decisions concerning data.

❖ **Libraries:**
- Support the management of data in activities such as: curatorship, deposit, publication, preservation, and access to data.
- Support the location and recovery of data.
- Advise on licensing, intellectual property, selection, data citations and good practices in general
- Support researchers in the creation of data management plans.

❖ **Research Institutions:** the institutions that carry out the research must ensure:
-Development and communication of policies for data management at institutional level.
-Ensure infrastructure and support services for data management (provided internally or externally)
-Educate and promote data management good practices among the parties involved.
-Control the fulfil of the requirements established for data management.

❖ **External service providers:** Some of the responsibilities that are usually assigned to external suppliers are:
-Provide storage services and data curation.
-Allow effective interoperability.
-Technological surveillance.

❖ **Information Technology Services:**
-Offer services and support in storage, use of software, authentication and access, training, and technical support
-Support in use of new technologies and data structuring
-Technological surveillance [24] [17]

2.6 Data Management

Ways to handle the data

Create Data

Before generating the data, there are some important decisions to be made in order to ensure the good organization of the information. A careful plan can help save much time and effort later. These decisions will affect the later use and access of the data in the long term. Many funders now also expect to see evidence of engagement in the data planning.

Some of the most important factors to consider are:

- Bidding for funding: data management is a crucial part of the research project life cycle from the very beginning and many funders now require a data management plan as part of the bidding process. It is important to know what funders expect, how data planning ca affect your bidding process, and who to consult when bidding for funding.
- Plan: before the start to create data is needed to make a data management plan. This will allow to make important decisions early on in the research process thus saving time and resources later on.
- Choosing formats: there are many file formats for holding data and these can differ from discipline to discipline. Choosing the formats before starting the process of generating data guarantee effective use and storage of the data in both the immediate and long term.
- Intellectual Property Rights: researchers need to maintain high ethical standards and adhere to data protection laws when obtaining data from people via questionnaires, interviews, etc. They have to ensure that any data gathered and shared is handled correctly and in accordance with the law. Links to resources will help to make it sure. [17] [23]

Organise Data

Organising the data in a consistent way will help the researchers to find and understand the results, saving valuable time both now and in the future. Good organisation reduces the chance of lost data and as a result will reduce your error rate. Suitable data documentation will also help to understand why exactly have been recorded and will allow re-use of data in the short, medium, and long-term.

Naming and organising files

A logical and consistent file naming system allows to locate and use data. Thinking about how to name and structure data should be done at the start of the project.

How to organise the files?

1) Use a logical folder structure and group files within folders so information on a topic is located in one place.
2) Structure folders hierarchically so that start with a limited number of folders with broader titles, and then create more specific folders within them.
3) Name folders after the areas of work to which they relate and not after individuals. This avoids confusion if someone leaves the project.
4) Follow existing folder naming procedures in the team or department if necessary.
5) Be consistent with the folder naming scheme.

What to consider when creating a file name?

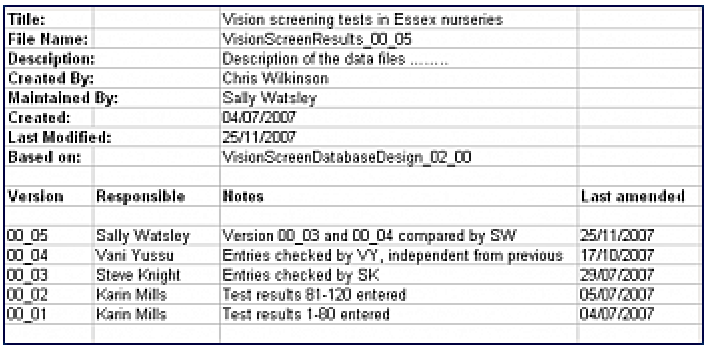
1) Decide on a file naming convention at the start of the project.
2) Be brief but meaningful – file names should be easy to read but equally it is needed to be able to understand them.
3) Avoid very long file names.
4) Avoid using spaces and specialist characters – these might be processed differently by various software, causing difficulties in the future. It is often recommended use underscores rather than fullstops or spaces for these reasons.
5) File names should outlast the data file creator who originally named the file.
6) The filename should include as much descriptive information that will assist identification independent of where it is stored.
7) Agree on a logical use of dates so that they display chronologically i.e. YYYY-MM-DD.
8) Confirm which element of the naming system will be written first, so that files on the same theme are listed together and can therefore be found easily.
9) Specify the amount of digits that will be used in numbering so that files are listed numerically

How to organise the files to know the most recent version?

In the researching field is quite often the use of many laptops and working in many locations. Many researchers use to be involved in the project and all of them might make changes. It is needed to know if all the different versions are going to be stored or will the files need to be synchronised.

In which case, it is good practice to record file status/versions using a 'revision' numbering system. Major changes to a file can be indicated by whole numbers, for example, v01 followed by v02 and so on. Minor changes can be indicated by increasing the decimal figure, for example, v01_01 indicates a minor change has been made to the first version, and v03_01 a minor change has been made to the third version.

Also, include a 'version control table' for each important document, noting changes and their dates alongside the appropriate version number of the document.



Title:		Vision screening tests in Essex nurseries	
File Name:		VisionScreenResults_00_05	
Description:		Description of the data files	
Created By:		Chris Wilkinson	
Maintained By:		Sally Watsley	
Created:		04/07/2007	
Last Modified:		25/11/2007	
Based on:		VisionScreenDatabaseDesign_02_00	
Version	**Responsible**	**Notes**	**Last amended**
00_05	Sally Watsley	Version 00_03 and 00_04 compared by SW	25/11/2007
00_04	Vani Yussu	Entries checked by VY, independent from previous	17/10/2007
00_03	Steve Knight	Entries checked by SK	29/07/2007
00_02	Karin Mills	Test results 81-120 entered	05/07/2007
00_01	Karin Mills	Test results 1-80 entered	04/07/2007

Lastly, agree who will complete any final files and mark them as 'final'. [17] [23]

Documentation and metadata

Digital data are machine-readable. However, the task of interpreting data falls to humans. For this to work we must have sufficient contextual information. Thus, the importance of documenting the data during the collection and analysis stage of the research cannot be underestimated.

Whilst collecting data the scientist might be able to remember what all the classification systems mean, but the chances are slim that this will still be the case a few months or a year's time. Sufficient documentation, explaining the codes and classifications used, will eliminate this possibility.

Others may also want to examine the data for many reasons, such as: understanding your findings, verifying your results, reviewing your submitted publication, replicating your results, designing a similar study or archiving your data for access and re-use. Good documentation will ensure all these reasons are possible regardless of what system or software they might be using. [17] [23]

When and how include documentation/metadata?

Document the data should be done from the very beginning of the research project. Information can then be added as the project progresses. Include procedures for documentation in the data plan. Documentation can be added to the data at various levels: [24]

- Variable level documentation: this documentation can be included within the data or document itself, as a header or at a specified location within a file. Examples of variable level documentation can include information on labels, codes, classifications, missing values, derivations and aggregations.

- File or database level documentation: this type of documentation explains how all the files that make up the dataset relate to each other, what format they are in or whether particular files are intended replace other files, etc. A readme.txt file is an established way of accounting for all the files and folders in a project.

- Project level documentation: explains the aims of the study, what the research questions/hypotheses are, what methodologies are being used, what instruments and measures are being used, etc. This information is contained in separate files accompanying the data in order to provide context, explanation, or instructions on data use or reuse. Examples of project level documentation include: working papers, laboratory books, questionnaires, interview guides, final project reports and publications.

- Metadata: this is structured information that has the potential for machine-to-machine interoperability. It can be used to identify and locate the data via a web browser or web-based catalogue. Usually structured according to an international standard, metadata descriptors are crucial if the scientists intend to share the data online. Researchers usually create metadata when completing a data centre's deposit form or metadata editor. Fields include: title, description, abstract, creator, geographic location, keywords. [25] [26]

2.7 Ways to share data

Among different researchers in different locations working together on the same research project.

Cloud Computing

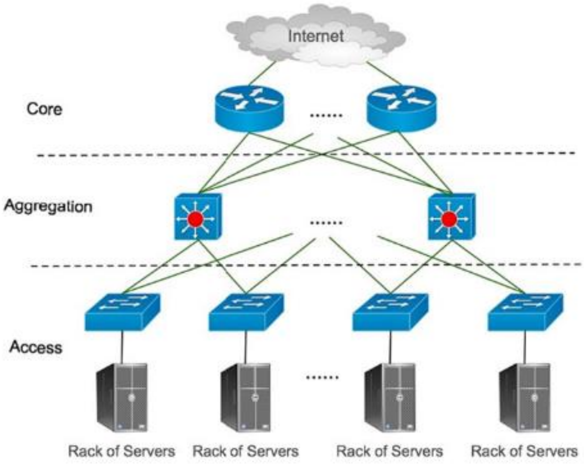
The main idea of cloud computing was already envisioned by John McCarthy in the 1960s, referring to computing facilities will be provided to the general public like a utility. The term “cloud” has also been used in various contexts. However, it was after Google’s CEO Eric Schmidt used the word to describe the business model of providing services across the Internet in 2006, that the term really started to gain popularity. The definition of cloud computing provided by The National Institute of Standards and Technology (NIST) is the next:

NIST definition of cloud computing: *“Cloud computing is a model for enabling convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction.”*

How cloud computing works?

Cloud Computing uses a network layer to connect users peripheral point devices, such as computers, smartphones, and portable accessories, to resources centralized in the data center. This data centre is home to the computation power and storage, is central to cloud computing and contains thousands of devices like serves, switches and routers. Proper planning of this network architecture is critical, as it will heavily influence applications performance and throughput in such a distributed computing environment. Currently, a layered approach is the basic foundation of the network architecture design, which has been tested in some of the largest deployed data centers. The basic layers of a data center consist of the core, aggregation, and access layers. [27] [28]

Fig. 6 Basic layered design of data center network infrastructure.



Source: Qi Zhang, Lu Cheng, Raouf Boutaba (2010): “Cloud computing: state-of-the-art and research challenges”

Types of Clouds

Public clouds

A cloud in which service providers offer their resources as services to the general public. Public clouds offer several key benefits to service providers, including no initial capital investment on infrastructure and shifting of risks to infrastructure providers. However, public clouds lack fine-grained control over data, network and security settings, which hampers their effectiveness in many business scenarios.

Private clouds

Also known as internal clouds, private clouds are designed for exclusive use by a single organization. A private cloud may be built and managed by the organization or by external providers. A private cloud offers the highest degree of control over performance, reliability and security. However, they are often criticized for being similar to traditional proprietary server farms and do not provide benefits such as no up-front capital costs.

Hybrid clouds

A hybrid cloud is a combination of public and private cloud models that tries to address the limitations of each approach. In a hybrid cloud, part of the service infrastructure runs in private clouds while the remaining part runs in public clouds. Hybrid clouds offer more flexibility than both public and private clouds. Specifically, they provide tighter control and security over application data compared to public clouds, while still facilitating on-demand service expansion and contraction. On the down side, designing a hybrid cloud requires carefully determining the best split between public and private cloud components.

Virtual Private Cloud

An alternative solution to addressing the limitations of both public and private clouds is called Virtual Private Cloud (VPC). A VPC is essentially a platform running on top of public clouds. The main difference is that a VPC leverages virtual private network (VPN) technology that allows service providers to design their own topology and security settings such as firewall rules. VPC is essentially a more complete design since it not only virtualizes servers and applications, but also the underlying communication network as well. Additionally, for most companies, VPC provides seamless transition from a proprietary service infrastructure to a cloud-based infrastructure, owing to the virtualized network layer. [27]

Advantages of Cloud Computing

- **Outsourcing Hardware:** when a company provides its own software, it must deal with the servers. These servers require an exclusive power supply and replacement parts. Moreover, it is also necessary to configure and supervise them in case of performance problems and on-call experts could be needed to solve them. With Cloud Computing all these problems and the variable costs disappeared. The cloud computing provider is responsible for dealing with those concerns. Is their business to ensure an efficient, correct, and interrupted process in exchange of fixed and reasonable cost.

- **Security:** Cloud Computing is extremely secure, most of the times better than the levels obtained with the traditional computing systems. Cloud computing providers work with a much larger budget. As they need to guarantee the safety of all customers, modern security technologies are implemented. Each company benefits from the large group, meaning a higher level of security for all. With a stronger infrastructure, careful monitoring and application of security protocols, cloud computing can offer small and medium enterprises the same protection as organizations with more demanding requirements. [28]

- **Data Recovery:** Cloud Computing providers manage recovery problems more quickly than recoveries that are not in the cloud.

- **Automatic software updates:** Cloud Computing providers are responsible for server maintenance, including security updates.

- **Simplicity:** generally, the final user does not need to know what happens in the cloud. The user only needs to start the session and work on the task.

- **Flexibility:** since resources can be allocated or deallocated on-demand, service providers are empowered to manage their resource consumption according to their own needs. Furthermore, the automated resource management feature yields high agility that enables service providers to respond quickly to rapid changes in service demand such as the flash crowd effect.

- **Geo-distribution and ubiquitous network access:** clouds are generally accessible through the Internet and use the Internet as a service delivery network. Hence any device with Internet connectivity, be it a mobile phone, a PDA or a laptop, is able to access cloud services. Additionally, to achieve high network performance and localization, many of today's clouds consist of data centers located at many locations around the globe. A service provider can easily leverage geo-diversity to achieve maximum service utility.

- **Collaboration:** all the employees can share applications and documents at the same time. [28] [29]

Categories of cloud services

SaaS

Refers to providing on-demand applications over the Internet. In this type of service, the user access through the browser without taking into account the software. The design, maintenance as well as the updates and backup copies are responsibilities of the provider. If the service fails it is the provider's business to make it work again.

This is one of the most popular applications of the Cloud Computing: it is estimated that around 59% of the total workloads in the cloud will be SaaS until 2018. Some of the most used Saas examples are Google, Docs, Salesforce, Dropbox or Gmail. [27] [28] [29]

Saas	

PaaS

PaaS refers to providing platform layer resources, including operating system support and software development frameworks. This type takes the advantages of Cloud Computing while allows the user the freedom to develop custom software applications. Users can access Paas in the same way that is done with SaaS.

The provider continues being the responsible for the maintenance of the operational system, the network, the servers, and security. For developers who ignore the infrastructure to be used and just want to focus about the software, this is the best alternative. Some examples could be Google App Engine, Heroku or Microsoft Windows Azure. [27] [28] [29]

Paas	App Engine HEROKU Windows Azure

IaaS

IaaS refers to on-demand provisioning of infrastructural resources. The cloud provider hosts the infrastructure components, including servers, storage, networking hardware, as well as a range of services to accompany those infrastructure company. The main difference is that IaaS users have more control in comparison with Paas and Saas. The user is responsible to scale the applications according to the needs, in addition to preparing all the machines. This service enabling IaaS servers to implement greater levels of automation and orchestration for important infrastructure tasks. For example, a user can implement policies to drive load balancing to maintain application availability and performance. Examples of IaaS providers include Amazon EC2, GoGrid, Flexiscale or Google Cloud Platform. [27] [28] [29]

Iaas	amazon web services™ EC2 GOGRID FlexiScale™ UTILITY COMPUTING ON DEMAND Google Cloud Platform

Hadoop

The history of Hadoop is necessarily linked to Google. In fact, it could be said that Hadoop was born at the moment in which Google urgently needs a solution that allows it continues processing data at the rhythm needed. Google is unable to index the web to the level demanded by the market and therefore it decides to find a solution, which is based on a distributed file system, creating the slogan: "Divide and conquer".

This solution, which will later be called Hadoop, is based on a large number of small computers, each of which is responsible for processing a portion of information. The innovation is that, although each of them works independently and autonomously, all act together, as if they were a single computer of incredible dimensions.

In 2006, Google publishes all the details about its new discovery, sharing its knowledge and experience with all users. From this moment, the group of beneficiaries investigates the possibilities of the idea and develop an implementation that they call Hadoop. At this moment, Yahoo takes over, promoting its expansion, in order to reach large and iconic companies in the computer world, such as Facebook. [30]

What is Hadoop?

Hadoop is an open source system used to store, process, and analyse large volumes of data. The basic components of Hadoop are:

- HDFS: It consists of a distributed file system, which instead of store the data file in a single machine, it distributes the information to different devices.

- MAPREDUCE: It is a working framework that makes it possible to isolate the programmer from all the tasks inherent to parallel programming. It means, it allows a program which has been written in common programming languages, to be run I a Hadoop cluster. The great advantage is that it makes it possible to choose and use the language and the most appropriate tools for the specific task that will be carried out. [31]

Hadoop Function

Hadoop is a system that can be implemented on hardware at a relatively low cost, being at the same time totally free for software. Thanks to this, the users are able to store, manage and analyse all the information that before could not be processing due to technology limits or economic barriers.

Any organization that uses Hadoop can obtain new information, at the same time that it discovers and applies any other type of analysis to its data, such as a linear regression on millions of data archives of its history.

It is precisely because of that its use is expanding so much among companies that benefit from:

- The relatively low cost involved.
- The rapid return of the investment it provides
- The possibility of facing new challenges and solving problems that they could not assume before, or that were left unanswered.

One of its applications that we are most interested in is the low-cost data storage and archiving. The affordable cost of hardware makes Hadoop useful for storing and combining data such as transactional data, social networks, sensors, machines, scientists, etc. Low cost storage allows you to keep information that is not currently considered critical but that you may need to analyze later. [30] [32]

With the public and other researchers that are not directly connected to the project.

After a project is concluded and the final report is written the aim of the research world is widespread the new knowledge acquired. This refers not only to other researchers who can take the information to continue and improve the research process but also to all the public who found the new evidences useful and interesting. [33] The main ways to share the final work with all the people not involved in it are:

Archives or Repositories

Archives or repositories are organisations whose purpose is to preserve material for future use. A large number of data archives exist: these include both national (and international) services focused on data from one specific discipline or area, and institutional repositories maintained by a single university.

A growing number of funders expect data with acknowledged long-term value to be preserved and remain accessible and usable for future research. Some funders have set up specialised archives (sometimes known as data centres) to curate, disseminate and preserve data created as part of their funded programmes. In these cases, researchers are expected to deposit their data in the designated place.

A big advantage of depositing your data in an archive or data centre is that it will be preserved in the long term. [34] [35]

Submit data to a journal

A condition of publication in some journals is that authors are required to make data promptly available to others without undue restrictions. Datasets must often be made freely available to readers from the date of publication and must be provided to editors and peer-reviewers at submission, for the purposes of evaluating the manuscript. Some publishers are already creating persistent links from articles to relevant datasets. [34]

Informal Sharing

Project websites can offer easy immediate storage and dissemination, but will offer less sustainability, and it is difficult to control who uses your data and how they use it unless administrative procedures are in place.

Informal peer-to-peer sharing makes it difficult to know which data can be obtained where: it requires people to have the right contacts, makes managing data access a burden, and does not ensure availability of the data in the long-term. [34]

2.8 How is safe data storage ensured?

Looking after research data for the longer-term and protecting them from unwanted loss requires having good strategies in place for securely storing, backing-up, transmitting, and disposing of data.

Physical security, network security and security of computer systems and files all need to be considered to ensure security of data and prevent unauthorised access, changes to data, disclosure, or destruction of data. Data security may be needed to protect intellectual property rights, commercial interests, or to keep personal or sensitive information safe. [36] [37]

Physical data security requires:

- controlling access to rooms and buildings where data, computers or media are held.

- logging the removal of, and access to, media or hardcopy material in store rooms.

- transporting sensitive data only under exceptional circumstances, even for repair purposes, e.g. giving a failed hard drive containing sensitive data to a computer manufacturer may cause a breach of security.

Network security means:

- not storing confidential data such as those containing personal information on servers or computers connected to an external network, particularly servers that host internet services.

- firewall protection and security-related upgrades and patches to operating systems to avoid viruses and malicious code. [38]

Security of computer systems and files may include:

- locking computer systems with a password and installing a firewall system.

- protecting servers by power surge protection systems through line-interactive uninterruptible power supply (UPS) systems.

- implementing password protection of, and controlled access to, data files, e.g. no access, read only, read and write or administrator-only permission.

- controlling access to restricted materials with encryption.

- imposing non-disclosure agreements for managers or users of confidential data.

- not sending personal or confidential data via email or through File Transfer Protocol (FTP), but rather transmit as encrypted data.

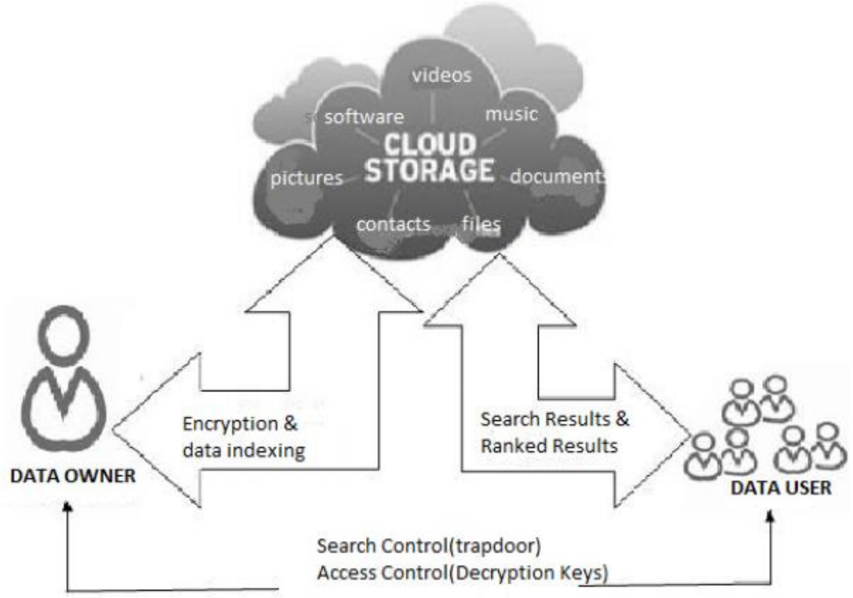
- destroying data in a consistent manner when needed. [39]

Data transmission and encryption

(Nar16) Transmitting data between locations or within research teams can be challenging for data management infrastructure. In the world of Cloud Computing the users are unwilling to place their confidential or sensitive data due to once the data are placed in cloud datacentre, the cloud client lost their direct control over their data.

To ensure the privacy of the data it must be encrypted to an appropriate standard. Encryption ensure the data during transmission. Encryption requires the creation of a public and private key pair and a passphrase. The private PGP key and passphrase are used to digitally sign each encrypted file, and thus allow the recipient to validate the sender's identity. The recipient's public PGP key is installed by the sender in order to encrypt files so that only the authorised recipient can decrypt them. [38] [40]

Fig.7 Cloud data storage model



Source: Naresh vurukonda, B.Thirumala Rao (2016): "A Study on Data Storage Security Issues in Cloud Computing"

Data storage challenge and issues

The only encryption does not give full security over the stored data, consequently the cloud data storage presents the next main challenges and issues:

- Data Storage Issues:
  - Data privacy and integrity: Because of simplicity cloud users are increasing exponentially and applications are hosted in cloud is very high. These situations lead to greater security threats to cloud clients. If any attack is successful on data entity will leads to data breach and takes an unauthorized. access to data of all cloud users.
  - Data recovery and vulnerability: the cloud ensures dynamic and on-demand resource provisioning to the users. The resource allocated to a particular user may be assigned to the other user at some later point of time. In case of memory and storage resources, a malicious user can employ data recovery techniques to obtain the data of previous users. [38]

  - Improper Media Sanitization
  - Data Backup: the data backup is important when accidental or intentional disasters, but the backup by themselves could be attacked by malicious activities.

- Identity Management and Access Control
  It is important to maintain track record for user identity for avoiding unauthorized access to the stored data. The cloud has to maintain quickly updating and managing identity management for joining and leaving users over cloud resources.
  - Malicious Insider
  - Outside Intruder

- Contractual and Legal Issues
  - Service Level Agreements
  - Legal Issues [41]

Data Storage issues solutions

1) SecCloud: it provides a storage security protocol for cloud customer's data and it not only secures the stored data but also provides security on computational data. The SecCloud protocol uses encryption for storing data in secure mode. By receiving encrypted data, the cloud decrypts the data, verifies the digital signature, and stores the original data in specified location in cloud.
2) The File Assured Deletion (FADE): it is a light weight protocol which provides a dey management with data integrity and privacy. It uses both asymmetric and symmetric key encryption of data. A group of key managers are used by FADE protocol, those acts as a trusted third party.
3) Time PRE: in this scheme the time period is associated with every user and by expiration the revocation automatically by Cloud Service Provider. This time based encryption scheme allows users to share keys in prior with CSP and CSP generate re-encryption keys by taking request from user. This scheme ensures the privacy and availability of data among the group peoples but doesn't concentrate on data integrity. [38]

Proposed Scheme	Services	Privacy	Integrity	Availability	Confidentiality
SecCloud, for securing cloud data	Encryption Bilinear pairing Signature verification Trusted third party	✓	✓	✗	✓
FADE, a protocol for data privacy and integrity	Encryption Trusted third party Assured deletion Threshold secret sharing	✓	✓	✗	✓
TimePRE, a scheme for secure data sharing in cloud	Proxy re-encryption Attribute based Encryption	✓	✗	✗	✓

Identity Management and Access control solutions

1) SPICE (Simple Privacy preserving Identity Management for Cloud Environments): ensures group signature for providing the unidentified authentication, access control, accountability, unlink ability, and user centric authorization. After user registration with trusted third party they obtain unique credentials for all the services. By using the credentials, user generates authentication certificate.
2) RB_MTAC (Role Based Multi-Tenancy Access Control): This requires user registration and obtains single credential that should be unique. The user has to choose the password while registration portal. By using these credentials, the user can enter into the cloud environment by passing through identity module that uniquely identifies the user and after that it will be redirected to role assignment module that establish a connection to the RB_MTAC database and assigns the roles to registered user based on enrolled information.
3) Identity Management framework: Here the Costumer service Provider acts as a host, while the authorized used acts as service owner. The authorization manager handles the service management and service requesting users also managed by authorization manager. This scheme ensures the identity management and access control across multiple Cloud providers with the help of authorization management. [38] [42]

Proposed Scheme	Services	Integrity	Availability	Confidentiality
SPICE, identity management framework	Anonymous and delegatable Authentication Access control Accountability	✓	✓	☒
RB_MTAC	Access control	✓	✓	☒
Identity management framework	Identity management Authentication Access control	✓	☒	✓

Data disposal

Having a strategy for reliably erasing data files is a critical component of managing data securely and is relevant at various stages in the data cycle. During research, copies of data files no longer needed can be destroyed. At the conclusion of research, data files which are not to be preserved need to be disposed of securely. Simple deleting does not erase a file on most systems, but only removes a reference to the file. It takes little effort to restore files deleted in this way. Files need to be overwritten to ensure they are effectively scrambled.

Software is available for the secure erasing of files from hard discs, meeting recognised standards of overwriting to adequately scramble sensitive files. Example software is BC Wipe, Wipe File, DeleteOnClick and Eraser for Windows platforms. Mac users can use the standard ‘secure empty trash’ option; an alternative is Permanent Eraser software.

Flash-based solid-state discs, such as memory sticks, are constructed differently to hard drives and techniques for securely erasing files on hard drives cannot be relied on to work for solid state discs as well. Physical destruction is advised as the only certain way to erase files. [38] [43]

USE-CASE

DATA MANAGEMENT PLAN

3.1 Use-Case Aim

The aim of the next section is to propose a formal document that should be prepared by the researcher or research group, which is developed at the beginning of a research project. It describes all aspects of data management, in terms of, what can be done with the data during and after the research project. It pretends to be a vital step in the research process that cannot be skipped. This instrument would help ensure that research data is accurate, complete, reliable and safe both during and after completing an investigation.

The purpose is to provide an analysis of the main elements of a data management policy, and describes the treatment received by the research data collected or generated in the course of a research project. Its purpose is to serve as a support tool during the life cycle of those data that are collected, processed or generated in a project.

Thanks to analyze a real use-case it is showed the main areas needed to be considered and the main recommended actions to allow the correct and effective management of research data. [23]

3.2 Use-case description

"Due to the economic and ecological requirements in the use of lubricants, the research group of the Technische Universität Dresden made up of Prof. Alexander Brosius (Project manager) and M.Sc. Ali Mousavi (Project agent) carry out the project named "Macro- and micro-structure of deep-drawing tools for dry forming". The goal of this experiment is to achieve a lubricant-free forming process in practice. Therefore, the deep drawing with the aim of forming drying is investigated.

The omission of the positive tribological functions of the waiver lubricants should be compensated in this project by a specific process and tool development. For this purpose, an integrated approach is chosen which includes the combination of macro and micro-structure with a corresponding coating of the tools. By combining the effects thereby obtained, the friction between the sheet and the tool is significantly reduced. [44] Thus, a lubricant used is unnecessary. The relevant cause-effect relationships are using analytical approaches, analyzed FE calculations and experimental investigations and thus made available for industrial application. As a result of using a macro structured thermoforming mold following four positive and stabilizing effects can be achieved:

- Reducing the contact area by 80%,
- Reducing the tendency to form wrinkles,
- Reducing the clamping force by 90%
- Material flow control.

The conducted experiments show that a deep-drawing process by macro- and micro-structured tools can be successfully carried out at constant process window without the use of a lubricant.

In the future, the method for minimizing frictional forces and wear for stamping complex shape elements can continue to develop and improve, so that a significant increase in the previously usable process window can be achieved.” [44]

3.3 Data Characterization

The described experiment has been carried out for five years. In this time, a long amount of data has been created. The first step in a Data Management Plan is identify the data. It is needed a detail description of the data obtained. Depending on the formats, standards and methodology used, the storage and sharing of the data would change.

In this specific experiment we find a wide range of data:

Identify Data			
Type	Sources	Volume	Format
Experimental Results	○ Machin / Sensors ○ Analysis and evaluation with software	1 GB	.txt .ASCII .CSV
Analytical calculations	○ Personal calculations ○ Software calculations	1.5 GB	Excel tables
FEM	○ Simufact ○ MSC Marc	900 GB	.sfp .mud .dat
Constructions	○ SolidWorks	120 GB	.SLD
Literatures	○ Books ○ Papers (Journal / Conference) ○ Dissertations ○ Magazine	20 GB	.pdf .doc .docx .ppt
Publications	○ Own publications	1 GB	.pdf .doc .docx .ppt
Pictures	○ Own pictures ○ Own painted pictures ○ Online pictures ○ Pictures from simulations	15 GB	.jpg .png .gif .tif .ppt

3.4 Organization and Documentation

In this section we intend to define a set of rules that establish the way in which identifiers must be assigned to digital objects. The use of consistent practices when assigning names will help to facilitate access to files. The main purpose of this section is to allow the researcher to differentiate the files and understand the content and status of them. In this way, we will avoid wasting time in the future trying to remember the content of each file.

File Name

To facilitate access to data sets and their versions, practices for structuring folders will be established. First, we will organize the data by type and then by research activity. Inside this folder we will try not to create structures of more than three or four levels. Moreover, we will avoid store more than ten items in each folder.

To define the name of the files we will follow these indications:
-Not to use special characters.
-Use only alphanumeric characters, except for dashes (-) or low bars (_). No points either gaps should be used.
-Only use lowercase.
- Ensure that the name is self-explanatory, containing sufficient descriptive information regardless of its location. Probably the files could change the storage location so this is important due to the new location could not to follow the same files structure.
-If we decide to include dates follow the standard ISO 8601 (AAAA_MM_DD or AAAAMMDD)
-Precede zeros when we include numbers (for examples, instead of use “1”, use, “001”)
-Include version number (for instances V01, V02, etc.). In the definitive version we could change the number for “final”.

Version Control

In the development of the research experiment, the information can be handled simultaneously by different researchers in different locations. For this reason, is quite important have a version control. It allows the researchers to know which version of a certain file it is being used. As well is important to communicate the changes realised.

There are different ways to register the versions, such as registering dates, version number or document status in the file name. You can also include change control tables or use specific software.

Proposed control table

Title:		Analysis test in Machine
File Name:		ExperimentalResults_00_05
Description		Result data of Vision screen Test carried out in …
Created by:		M.Sc. Ali Mousavi
Maintened by:		Prof. Alexander Brosius
Created:		04/07/2015
Last Modified:		20/08/2017
Based on:		VisionScreenDatabaseDesign_02_00

Version	Responsible	Notes	Last amended
00_05	Brosius	Version 00_03 and 00_04 compared	20/08/2017
00_04	Ali Mousavi	Entries checked by	15/06/2017
00_03	Brosius	Entries checked by	04/03/2016
00_02	Ali Mousavi	Test results MMtool	05/01/2016
00_01	Ali Mousavi	Test results XXtool	04/07/2015

3.5 Data Selection

In a kind of experiment as the use-case studied, the amount of data obtained require a selection process. This is due to the unlimited storage capacity of the institution. The selection process aims to make further choices about what to keep for the long-term, selecting what data to make available or to dispose of. The best time to do this data appraisal is well before the end of the project, or periodically if it's a longitudinal or reference data collection. [45] [46]

In this section it is proposed a guide to help make what may be quite difficult choices around what data to keep in order to meet the experiment purposes and satisfy the institution.

1. **Identify purposes that the data could fulfil**
   In this first step, you have to wonder if the data obtained could be useful for a future research. The reuse of the data is a good way to safe money and time for an institution. Consequently, is quite important to analyse carefully if the data could serve even beyond the research context in which it was created or collected.

   Some of the reasons that could justify retaining data for long-term are: enable others to verify the research with the willing to continue with the research process, increase the academic reputation allowing citations, to contribute the culture of sharing between other institutions and research groups and, finally, for a future private use.

2. **Identify data that must be kept**
   In some contexts, it could be compulsory requisites about what to keep and under which access conditions. It could be due to institutional policies, regulations, other legal or contractual reasons, and issues related with personal data. We must know all the data which we must preserve to avoid legal problems.

3. **Identify data that should be kept**
   This third step is the hardest one due to not always is easy to the research to guess which data would be useful. Moreover, we are handling a huge amount of data what means spending a lot of time to analysis it in order to decide what to keep.

   We propose here criteria to make easier the choice: a group of data should be kept if at least in two of the next sections, one question has an affirmative answer.

I. Is the data good enough?
   a. Is there enough information about what the data is, how and why it was collected and how it as been processed, to assess its quality and usefulness for the aims you identified?
   b. Is the data quality good enough in terms of completeness, sample size, accuracy, validity, reliability, representativeness or any other criteria relevant in the domain?

II. Is there likely to be demand?
   a. Are there users waiting for this data?
   b. Will making the data available be likely to significantly enhance a group or project's reputation?
   c. Could the data have social appeal in other contexts or previous researchers?

III. How difficult is it to replicate?
   a. Would reproducing the data be difficult or impossible?

IV. Is it the only copy?
   a. Is this the only and most complete copy of the data?

4. **Weigh up the costs**

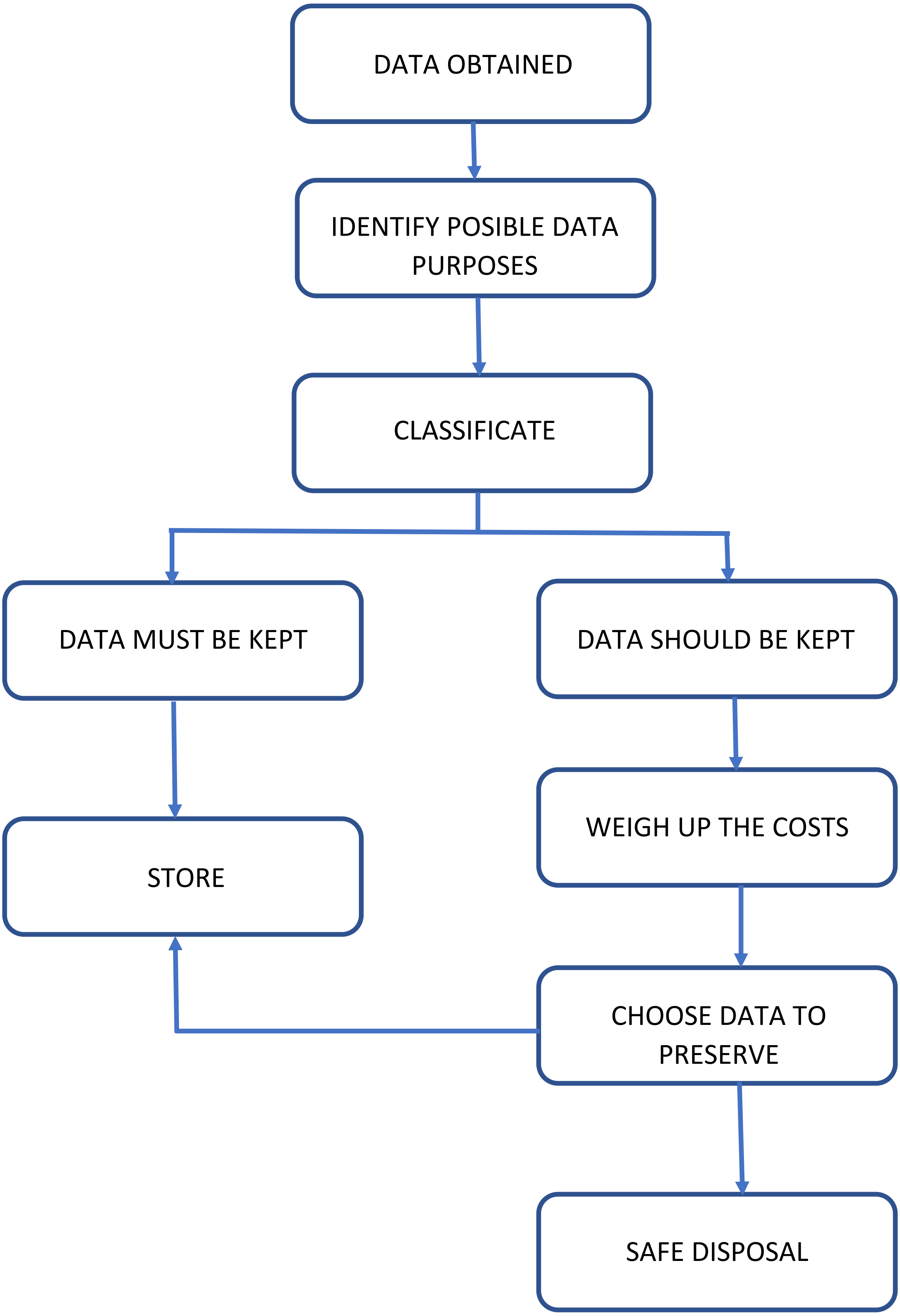
   It is quite important to consider the costs involved in the process of data selection. First, we have to be sure that there is enough funding to develop the data management plan. Furthermore, the storage and long-term curation charges must be paid.

   If the research group does not have the enough funding, the selection process must be simplified and just keep the most relevant information.

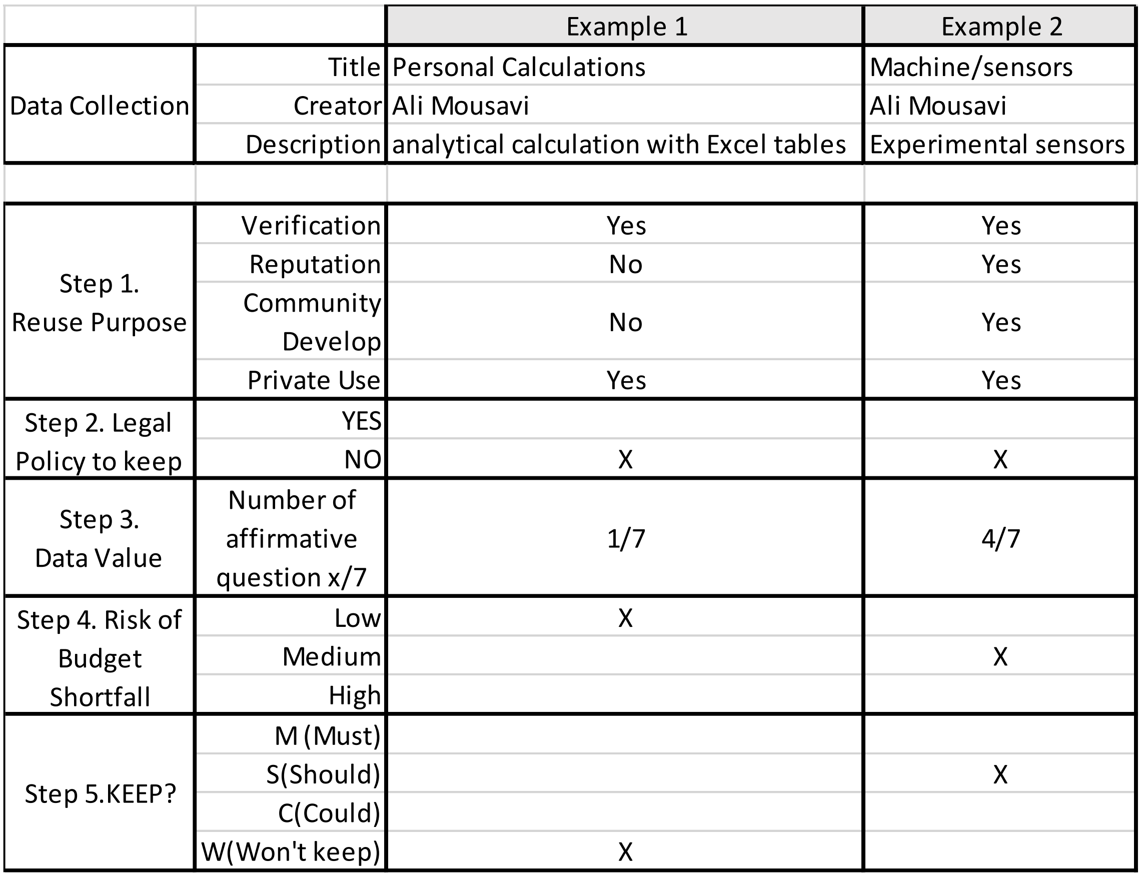
5. **Complete the data appraisal**

   Finally, after considering all the last factors, a choice about which data preserve must be made. This step consists in sum up all the conclusions of the previous sections to register to final decision. [45] [46]

Data selection process



Proposal table to Selection Data



		Example 1	Example 2
Data Collection	Title	Personal Calculations	Machine/sensors
	Creator	Ali Mousavi	Ali Mousavi
	Description	analytical calculation with Excel tables	Experimental sensors
Step 1. Reuse Purpose	Verification	Yes	Yes
	Reputation	No	Yes
	Community Develop	No	Yes
	Private Use	Yes	Yes
Step 2. Legal Policy to keep	YES		
	NO	X	X
Step 3. Data Value	Number of affirmative question x/7	1/7	4/7
Step 4. Risk of Budget Shortfall	Low	X	
	Medium		X
	High		
Step 5.KEEP?	M (Must)		
	S(Should)		X
	C(Could)		
	W(Won't keep)	X	

3.6 Storage and Security

Wrong decisions regarding storage and security can result in unauthorized access to a data set, its corruption and even its total loss. To avoid these events, it is important to select the proper storing data method as well as establish the necessary controls to keep the data safe.

Storage

One of the aspects that will need to be considered from the moment the research is planned is the way in which the data that will be generated or collected will be stored. It exists a wide range of storage methods already explained. It is important to keep in mind that the technique used may vary depending on the needs of the researchers and that the options used throughout the development of the research will not necessarily be appropriate for storing and giving access to data once it has finished. It means, that the methods would change throughout the progress of the study. On the other hand, the storage methods that can be used are not exclusive, so we can choose various and complement it each other's.

In the first steps of the research methods related with the personal work data are needed, these are: laptops, USB, CDs, HDD, network units inside the institution… These alternatives are recommended only to be used in the course of the investigation, since the necessary characteristics to guarantee access and preservation of data in the long term are not included. They don´t allow the disposition of other people beyond the group of researchers. These storage devices are affordable, but have disadvantages, such as their rapid degradation over time, the speed with which they can become obsolete, relatively error rates frequent, limited size or security risks to which they may be subject due to their portability. It is recommended that its use be limited to the storage of copies (never master data files) and should always be supported by more secure media.

On the other hand, cloud computing is really useful throughout the research, but the access should be limited to the people involved in the project. For the specific use-case experiment studied it is proposed to use the cloud computer category *Saas,* because the research group is not interested in developing custom software applications so could be better not to take into account the software.

Once the project is finished, it should migrate to an archive or repository where it can be preserved in the long-term, and in addition, to allow the sharing with the research community.

Security

The more security methods you use on your data, the less likely it is an unauthorised access or unwanted destruction to happened . The proposal ways to keep the security of the data in this experiment are:

1) Encryption: it is needed to convert the data to a non-recognizable or readable form or code. On this way, we prevent third parties to access or steal the data. Encryption could be use in various levels like devices, laptops or information transmitted by the network. Depending on the value of the data may not be necessary to encrypt all the data collected. It is proposed to carry out the encryption process in advanced stage of the project when the results and conclusions obtained are more valuable.
2) Use and constant update of Antivirus
3) Use of proper Passwords
4) Safe data disposal
5) Periodical Backup

3.7 Sharing and long-term preservation

Once the project is finished, the group would have to decide if it is convenient to share the investigation. If the answer is positive, more question would have need to be done in terms of which data should be released: the whole project, just the results and conclusions, the data obtained by experimentation…

Sharing the experiment	
Advantages	**Disadvantages**
It allows the results verification by other scientists and may lead new researches.	Competitive feeling between institutions. Unwilling to cooperate.
It increases the collaboration. This time is your own institution who release the information but next time you could be the one who obtain benefits.	Financial value associated with intellectual property.
It increases the impact and visibility of the investigation.	Regulations in terms of data property.
The institutional and research group reputation increase.	

In the situation of the experiment we can conclude that sharing would be the best choice. The main people in charge of the project should be responsible for deciding the depth in which they want the data to be release.

Preservation

The characteristics of the experiment carry out like,

- Large volume of data: for five years the research has created around 1058,5 GB.
- The variety of formats and types of data: unfortunately, the research not only consists in one kind of data, but it is a wide range of data types which make the management and preservation more difficult.
- Related with the above, the dependence of specific types of hardware and software to access and use the data could be a problem, due to the quick speed with which it become obsolete.

, make the long-term preservation a challenge. Consequently, a good planification of preservation is needed.

The most proper way in this specific use-case is deposit the data in repositories created for that purpose. Although the data can be shared informally (for example, when requested by e-mail), this will in no way ensure the availability of data in the medium and long term and introduces unnecessary barriers both for those who share the data and for those who they want to access them. To choose which repository is better it is good to take into consideration that there are two options:

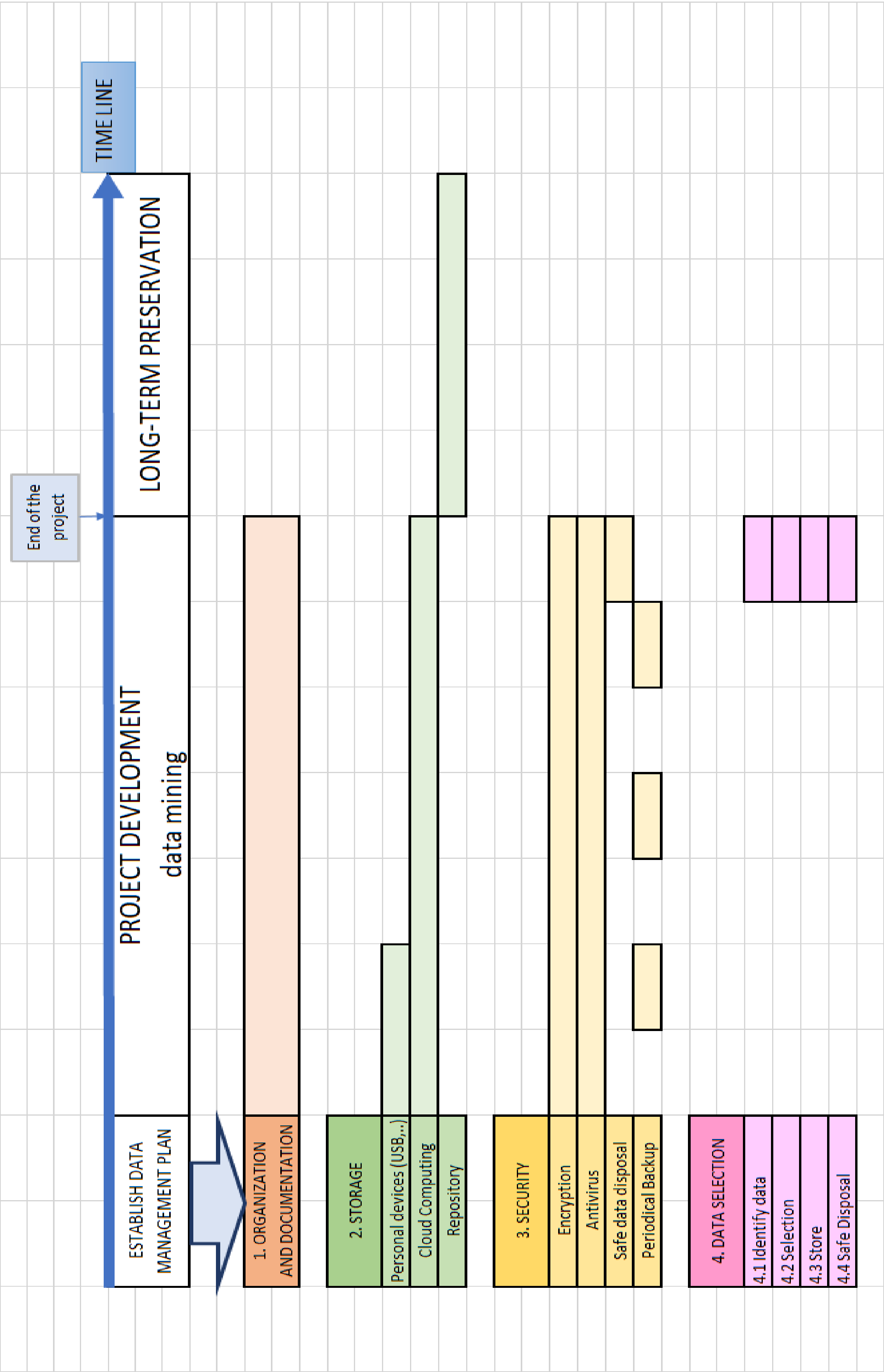
- **Institutional repositories:** generally, responsibility of a university or research center. It allows their own researchers to deposit and assign licenses. In this way, the data can be consulted and reuse by others.

- **National repositories:** They are usually associated with research funding agencies or national agencies linked to science and technology. In these cases, they can offer storage services, licensing and access to data to different communities of researchers, without necessarily having to be affiliated with a specific institution.

The choice of which repository is more convenient depends on the scope of the project. If the responsible people consider that the interest of the research does not go beyond the circle of the own institutional researchers, the institutional repository would be enough. But in the opposite way, if it is rather to give a greater scope to the project allowing a larger community to access it, the national repository will be more appropriate.

3.8 Sum up

Data Management Plan



4. Conclusions

Nowadays, most of the research groups find the same problematic handling data which is collected in scientific experiments. The large amount generated entails the necessity of a more accurate and specific curation of data. The wide range of possible types of data created, the different sources where it came from, the common fact that more than one person is working at the same time in the data mining of a project, make the organization of data much more complex than it used to be some years ago. To allow the management of this data-rainfall the research institutions must to invest time, personnel and infrastructures. As this is relatively a new problematic, most of the groups are not sufficiently aware or sensitive to the seriousness of it. Consequently, a huge risk is being taken that can result not only in the waste of money and time but also in the definitive loss of useful knowledge for the society.

To only way to solve the problem is to establish a proper Data Management Plan. It must be developed at the beginning of a research project and all the personnel involved must commit to follow it. First, the researchers must describe and characterize the data collected, in terms of types, sources, volume and format. Then the Data Management Plan should provide a proper guide to organize and document all the data. The main purpose is to allow the researcher to differentiate the files and understand the content and status of them. The principal factor to consider are the files names and version control.

At some point in the develop of the project, the research group would need to carry out a process of data selection. The limit of storage does not allow to keep all the data created. It is a delicate and complex process making the decision about what to keep and what to delete. If the wrong decision is made, in the future deleted data may be required which means that the experiments would have to be carried out again. Now, it does not exist an accurate method to choose which data preserve to the long-term. Statistical methods or selection algorithms could be used in the future to solve this problem. Until then, a continuous organizational process is proposed that aims to facilitate the decision through a table.

A Data Management Plan also consider storage and security. The first one is as well a problematic issue. It must allow the researchers to store the data while the project is carry out and allow sharing the information among different researchers in different locations who are working together on the same research project. Cloud Computing is proposed as the best way to permit this. On the other hand, a proper storage method must ensure the preservation of the information in the long-term. For this, the best choice is migrating the information to repositories or archives where the material can be not only preserve but also accessible to all the research community allowing the reuse. About security, the Data Management Plan must provide good strategies to protect the data from unwanted loss. Some of them are physical security, encryption, antivirus, proper passwords, safe data disposal and periodical backup, among others.

To sum up, the data management plays an important role in the public funded science and research field. A proper organization plan allows the researchers to get the most out of their investigations and facilitate the expansion of the new knowledge to all the community.

Annex I: Experimental Design

The purpose of this section is to propose the application of experimental design techniques for the study of the explained experiment. The design of experiments is a tool of fundamental importance in the field of engineering to improve the performance of a process. Using this statistical process, it could be possible set some requirements to separated which data to keep an which to remove. [47]

The design of experiments or statistical design of experiments is a discipline based on statistical principles and built through years of experience in science and engineering. It involves the process of designing and planning the experiment, so that appropriate data can be collected and later analyzed by statistical methods to reach valid and objective conclusions.

Some of the benefits that may result from the application of the experimental design are:

- Improved process performance
- Reduced variability and closer compliance with the projected requirements
- Reduction of process time
- Reduction of global costs

Among the applications of experimental design in engineering design we can find:

- The evaluation and comparison of basic design configurations.
- The evaluation of possible materials.
- The selection of design parameters so that the product has a good performance in a wide variety of conditions.
- Determination of the key parameters of the product design that affect its performance

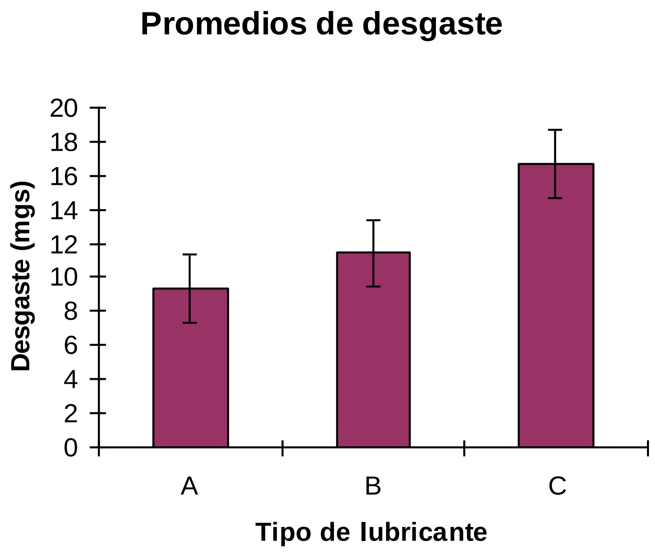
The proposed idea is to use this statistical technique to decide which experiments among all the test carried out during the project are significant enough to make it worthwhile to keep them.

Just to exemplify how this technique could be applied to the specific experiment, a simplified case is proposed:

➢ The next data refers to the loss of friction due to the use of three different tools. The tool C has been used to the process but now two new possibilities are evaluated.

Tool	Friction	Tool	Friction	Tool	Friction
A	12.2	B	10.9	C	12.7
A	11.8	B	5.7	C	19.9
A	13.1	B	13.5	C	13.6
A	11	B	9.4	C	11.7
A	3.9	B	11.4	C	18.3
A	4.1	B	15.7	C	14.3
A	10.3	B	10.8	C	22.8
A	8.4	B	14	C	20.4

By means of a bar chart with its standard error, we can determine the friction and the variability of the observations.



As can be seen in the graph, the dispersion patterns within each tool are very similar (due to the similarity in the standard error bars). It can also be observed that the tool with the highest value in the answer is C, followed in decreasing order by tool B and the tool with the lowest average response in tool A.

Statement of the hypothesis:
Ho: The friction averages under the three tools used are the same.
H1: At least one average friction associated with a tool is different.

The statistical model for this data set corresponds to a completely random design.

$$Y_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

Y_{ij}: it corresponds to the friction value in the repetition j of the treatment i.

μ: the general average of friction.

τi : the effect of the tool i.

εij : the experimental error in the repetition j of the tool i.

Analysis of variance to test the hypothesis:

TABLE ANALYSIS VARIANCE

Sources of variation	*SS*	*DF*	*MS*	*F*	*Probability*	*Critical Value for F*
Among groups	230.585833	2	115.292917	8.74681521	0.00172472	3.46680011
Inside the groups	276.80375	21	13.181131			
Total	507.389583	23				

R^2=0,4544

The variation in the data is due in 45% to the change of the tool and 55% factors which are not being considering in the investigation.

P_Value=0,0017

As this value is less than the significance of the test (0.05) Ho is rejected. This means that the friction media in the three tools are not equal what implies that the tool has an impact on the level of friction. Then, it could be possible to decide about which tool is the most convenient. In this case, if a decrease in the friction is desired we should chose the tool A.

This is just an example of the many applications experimental design has. Working with these techniques it would be possible to establish requirements that will provide us a statistical decision on which data to maintain and which to be removed. [47]

Bibliography

[1] Jim Gray, David T. Liu, Maria Nieto-Santisteban & Alex Szalay, David J. DeWitt, Gerd Heber (2005): “Scientific Data Management in the Coming Decade.”

[2] The Research Council of Norway (2005): “eScience – Infrastructure, Theory and Application (eVITA)”

[3] Molloy JC (2011) The Open Knowledge Foundation: Open Data Means Better Science. PLoS Biol 9(12): e1001195.

[4] Paul A. David; Understanding the emergence of ‘open science’ institutions: functionalist economics in historical context, *Industrial and Corporate Change*, Volume 13, Issue 4, 1 August 2004, Pages 571–589

[5] Saint John Walker (2014) Big Data: “A Revolution That Will Transform How We Live, Work, and Think, International Journal of Advertising”, 33:1, 181-183

[6] Trevor J Barnes (2013): “Big data, little history”

[7] JOUR Lynch, Clifford (2008): “How do your data grow?”

[8] Gandomi, A., & Haider, M. (2015): “Beyond the hype: Big data concepts, methods, and analytics.”, International Journal of Information Management,35(2), 137–144

[9] Seref Sagiroglu, Duygu Sinanc (2013): “Big Data: a review”

[10] Min Chen, Shiwen Mao, Yin Zhang, Victor C. M. Leung (2014): “Big Data related technologies, challenges and prospects.” pp 33-49

[11] James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers (2011): “Big data: The next frontier for innovation, competition, and productivity.”

[12] University of Sussex (2018): “Research data management”

[13] Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, Vishanth Weerakkody (2015): “Critical analysis of Big Data challenges and analytical methods”, Journal of Business Research, Volume 70, 2017, pp. 263-286

[14] Chen, C. L. P., & Zhang, C. Y. (2014): “Data-intensive applications, challenges, techniques and technologies: a survey on big data”, Information Sciences

[15] Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money (2013): “Big Data: Issues and Challenges Moving Forward”, System Sciences

[16] Christine L. Borgman (2010): “ADVANCES IN INFORMATION SCIENCE; The Conundrum of Sharing Research Data”

[17] Veerle Van den Eynden, Louise Corti, Matthew Woollard, Libby Bishop and Laurence Horton (2011): “Managing and sharing data”

[18] Guoliang Li, Beng Chin Ooi, Jianhua Feng, Jianyong Wang, Lizhu Zhou (2008): “EASE: an effective 3-in-1 keyword search method for unstructured, semi-structured and structured data.”

[19] Abiteboul, S. (1997, January). Querying semi-structured data. In *International Conference on Database Theory* (pp. 1-18). Springer, Berlin, Heidelberg.

[20] Chen, Y., Wang, W., Liu, Z., & Lin, X. (2009, June). Keyword search on structured and semi-structured data. In *Proceedings of the 2009 ACM SIGMOD International Conference on Management of data* (pp. 1005-1010). ACM.

[21] https://libguides.macalester.edu/c.php?g=527786&p=3608643

[22] Jinchuan CHEN, Yueguo CHEN, Xiaoyong DU, Cuiping LI, Jiaheng LU, Suyun ZHAO, Xuan ZHOU (2013): "Big data challenge: a data management perspective"

[23] Biblioteca CEPAL (2015-2017): "Gestión de datos de investigación", Proyecto LEARN-Leaders Activating Research Networks.

[24] McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big data: the management revolution. *Harvard business review*, *90*(10), 60-68.

[25] Nathan, D., & Austin, P. K. (2004). Reconceiving metadata: language documentation through thick and thin. *Language documentation and description*, *2*, 179-187.

[26] Greenberg, J. (2003). Metadata generation: Processes, people and tools. Bulletin of the American Society for Information Science and Technology, 29(2), 16-19.

[27] Qi Zhang, Lu Cheng, Raouf Boutaba (2010): "Cloud computing: state-of-the-art and research challenges"

[28] Voorsluys, W., Broberg, J., & Buyya, R. (2011). Introduction to cloud computing. *Cloud computing: Principles and paradigms*, 1-41.

[29] Wu, J., Ping, L., Ge, X., Wang, Y., & Fu, J. (2010, June). Cloud storage as the infrastructure of cloud computing. In *Intelligent Computing and Cognitive Informatics (ICICCI), 2010 International Conference on* (pp. 380-383). IEEE.

[30] Zikopoulos, P., & Eaton, C. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data*. McGraw-Hill Osborne Media.

[31] White, T. (2012). *Hadoop: The definitive guide*. " O'Reilly Media, Inc.".

[32] Hadoop, A. (2011). Apache hadoop.

[33] Kowalczyk, S., & Shankar, K. (2011). Data sharing in the sciences. *Annual review of information science and technology*, *45*(1), 247-294.

[34] Schofield, P. N., Bubela, T., Weaver, T., Portilla, L., Brown, S. D., Hancock, J. M., ... & Rosenthal, N. (2009). Post-publication sharing of data and tools. *Nature*, *461*(7261), 171.

[35] Kim, Y., & Adler, M. (2015). Social scientists' data sharing behaviors: Investigating the roles of individual motivations, institutional pressures, and data repositories. *International Journal of Information Management*, *35*(4), 408-418.

[36] https://www.youtube.com/watch?v=vlCO7J4KCz8

[37] Kaufman, L. M. (2009). Data security in the world of cloud computing. IEEE Security & Privacy, 7(4).

[38] Naresh vurukonda, B.Thirumala Rao (2016): “A Study on Data Storage Security Issues in Cloud Computing”

[39] Chen, D., & Zhao, H. (2012, March). Data security and privacy protection issues in cloud computing. In *Computer Science and Electronics Engineering (ICCSEE), 2012 International Conference on* (Vol. 1, pp. 647-651). IEEE.

[40] Robling Denning, D. E. (1982). *Cryptography and data security*. Addison-Wesley Longman Publishing Co., Inc..

[41] Katal, A., Wazid, M., & Goudar, R. H. (2013, August). Big data: issues, challenges, tools and good practices. In *Contemporary Computing (IC3), 2013 Sixth International Conference on* (pp. 404-409). IEEE.

[42] Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013, January). Big data: Issues and challenges moving forward. In *System sciences (HICSS), 2013 46th Hawaii international conference on* (pp. 995-1004). IEEE.

[43] Jones, A. (2009). Lessons not learned on data disposal. *Digital Investigation*, *6*(1-2), 3-7.

[44] Description of the experiment: http://www.trockenumformen.de/schwerpunktprogramm/projekte/makro-und-mikrosturkturierung/

[45] DCC (2014): 'Five steps to decide what data to keep: a checklist for appraising research data v.1'. Edinburgh: Digital Curation Centre.

[46] Whyte, A. & Wilson, A. (2010). "How to Appraise and Select Research Data for Curation". DCC How-to Guides. Edinburgh: Digital Curation Centre.

[47] Douglas C Montgomery (2017): “Design and analysis of experiments”