DANUBIUS-PP

Preparatory Phase for the pan-European Research Infrastructure DANUBIUS–RI "The International Centre for advanced studies on river-sea systems"

Set of techniques and tools for data processing

Deliverable 8.10

European Commission

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 739562

Project Full title	Preparatory Phase for the pan-European Research Infrastructure DANUBIUS–RI "The International Centre for advanced studies on river-sea systems"
Project Acronym	DANUBIUS-PP
Grant Agreement No.	739562
Coordinator	Dr. Adrian Stanica
Project start date and duration	1st December 2016, 36 months
Project website	www.danubius-pp.eu

Deliverable No.	8.10	**Deliverable Date**	M32
Work Package No.	WP8		
Work Package Title	ICT e-Infrastructure and digital data cloud storage		
Responsible	22 - POS		
Authors & Institutes Acronyms			
Status:	**Final (F)**		
	Draft (D)	●	
	Revised draft (RV)		
Dissemination level:	**Public (PU)**	-	
	Restricted to other program partic-ipants (PP)	-	
	Restricted to a group specified by the consortium (RE)	-	

	Confidential, only for members of the consortium (CO)	●

Executive summary / abstract

What is the focus of this Deliverable?

The main objective of this document is to introduce beneficiaries and other readers to the different techniques and tools for data processing. First, a generalised process for pre-processing and data preparation will be explained; then some useful metrics, transformations and statistics in river-sea systems will be shown in the context of Danubius PP.

Then, it is presented the concept of *FAIR* data, explaining the four principles and meaning and use of FAIR digital objects. Related with this, document covers the semantic web (and ontology) concepts. To finalise the part, a summary of the CKAN platform. An important part to be reviewed in this deliverable is R&D projects indicators: metrics and KPIs (and relationships between them).

There is a compendium of the tools for data processing, including: spreadsheet software, database managers, programming languages, applications to visualise and interact with collected data, and existing European platforms to achieve this purposes[1] (in this case it is a brief description, for supplementary information, please check D8.9 *"Set of computational tools oriented towards HPC and Cloud computing").*

What is next in the process to deliver the DANUBIUS-PP results?

As said before, this document relies in part on deliverable D8.9; some of the tools presented in this deliverable were already named in the previous one. It will serve as support for the D8.11 *(Set of modelling and benchmark tools)* and will also facilitate the implementation of D8.15 *(Implementation of application interfaces - final version).*

[1] European Union in the area of 2020 horizon and the most used by data analysts and Big Data specialists worldwide.

What are the Deliverable contents?

This document is divided into the following sections:

- This executive summary, and the introduction to put this document into context. Useful acronyms table.
- Data Pre-processing: the orderly process used to prepare the data for subsequent modelling. Depending on the nature of the data, it will proceed with some methods or others; there are some steps that are obligatory for all the sets.
- The FAIR data concept, explaining every principle, digital objects. Introducing concepts of semantic web and ontology (vocabulary).
- Commonly used indicators for projects: metrics and KPIs.
- Brief description of the different tools used to manipulate data (spreadsheet, database, programming languages, visualization and interactivity, and existing European Platforms from H-2020 program).
- Conclusions, with special emphasis in the FAIR (Findable, Accessible, Interoperable and Reusable) data principles, promoted in the EC's Horizon 2020.
- Bibliographic references.

Contents

List of figures and tables

Acronyms

ADCP	Acoustic Doppler Current Profilers
API	Application Programming Interface
CKAN	Comprehensive Knowledge Archive Network
CSV	Comma Separated Values
DMP	Data Management Plans
EC	European Commission
EGI	European Grid Infrastructure
ENVRI	ENVironmental Research Infrastructure
EOSC	European Open Science Cloud
ERIC	European Open Science Cloud
FAIR	Findable, Accessible, Interoperable and Reusable
FOSS	Free and Open Source Software
GIS	Geographic Information System
GNU	GNU Not Unix
HTML	HyperText Markup Language
IaaS	Internet as a Service
ICA	Independent Component Analysis
INDIGO	INtegrating Distributed data Infrastructures for Global explOitation
ITSM	IT Service Management
KPI	Key Performance Indicator
LDA	Linear Discriminant Analysis
MAR	Missing At Random
MCAR	Missing Completely at Random
MNAR	Missing Not At Random
NA	Not Available
NS	NameSpace
ORCID	Open Researcher and Contributor ID
ORCID	Open Researcher and Contributor ID
ORM	Object Relational Mapper
PCA	Principal Component Analysis
PDF	Portable Document Format
PID	Persistent IDentifier
PP	Preparatory Phase
PRACE	Partnership for Advanced Computing in Europe
PROV	Data PROVenance

RAID	Risks, Assumptions, Issues and Dependencies
RDA	Research Data Alliance
RDF	Resource Description Framework
RI	Research Infrastructure
RRID	Research Resource Identifiers
RS	River-Sea
SaaS	Software as a Service
SDK	Software Development Kit
SQL	Structured Query Language
SWF	Scientific WorkFlow
SWFS	Scientific WorkFlow System
URL	Universal Resource Locator
USGS	U.S. Geological Survey
VMT	Velocity Mapping Toolbox
W3C	World Wide Web Consortium
XML	eXtensible Markup Language

1 Introduction

DANUBIUS-RI will be a pan-European distributed research infrastructure dedicated to interdisciplinary studies of large river–sea (RS) systems. It will enable and support research addressing the conflicts between society's demands, environmental change and environmental protection in river–sea systems worldwide. DANUBIUS-RI will be a distributed RI that will collaborate with, and enhance, existing European research infrastructure and programmes.

Key elements of this preparatory phase are to determine the necessary computing, storage and communication infrastructure and simulation tools for the distributed RI and to provide the tools and techniques for the development of all computing elements.

The main objective of the current WP8 is to provide necessary computing, storage, communication infrastructure (between the Data Centre, the Hub, Nodes, Supersites and users) and simulation tools. A second goal is to provide the computing elements which include HPC and Cloud Computing, application interfaces and virtual Research Environments.

In this report basic user requirements will be addressed by developing procedures for data aggregation, reformatting, transformation, inter/extrapolation and averaging, amongst others. These techniques will allow data flagging and, extracting trends and probability distribution functions, thus facilitating data comparison at different scales and providing added value to already existing datasets.

2 Data pre-processing

For a dataset, there are a sequence of steps that must be performed before carrying out any other type of operation with them. These steps or methodology are called data pre-processing. Data pre-processing consists of different stages:

2.1 Data cleansing[2]

The first step in data pre-processing is auditing the data to find the types of anomalies contained within it. The data is audited using statistical methods and parsing the data to detect syntactical anomalies. The instance analysis of individual attributes (data profiling) and the whole data collection (data mining) derives information such as minimal and maximal length, value range, frequency of values, variance, uniqueness, occurrence of null values, typical string patterns as well as patterns specific in the complete data collection (functional dependencies and association rules).

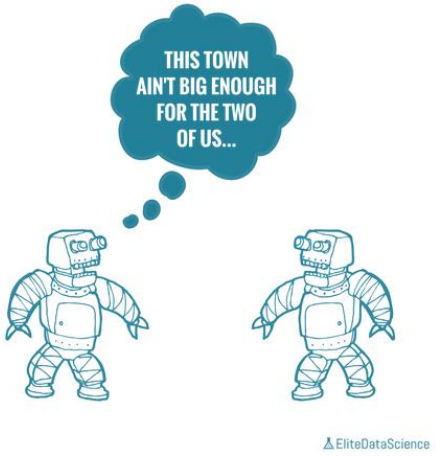


Figure 1. Duplicate observations most frequently arise during data collection.

The results of auditing the data support the specification of integrity constraints and domain formats. Integrity constraints are depending on the application domain and are specified by domain expert. Each constraint is checked to identify possible violating tuples. For one-time data cleansing only those constraints that are violated within the given data collection must be further regarded within the cleansing process. Auditing data also includes the search for characteristics in data that can later be used for the correction of anomalies.

[2] Also known as “Data cleaning” or “Data scrubbing”.

As a result of the first step should be an indication for each of the possible anomalies to whether it occurs within the data collection and with which kind of characteristics. For each of these occurrences a function, called tuple *partitioner*, for detecting all its instances in the collection should be available or directly inferable.

2.2 Handling of missing values (NA)

If it is noted that there are many tuples that have no recorded value for several attributes, then the missing values can be filled in for the attribute by various methods described below. There are some reasons why a data goes missing:

- **Missing at Random (MAR):** Missing at random means that the propensity for a data point to be missing is not related to the missing data, but it is related to some of the observed data
- **Missing Completely at Random (MCAR):** The fact that a certain value is missing has nothing to do with its hypothetical value and with the values of other variables.
- **Missing not at Random (MNAR):** Two possible reasons are that the missing value depends on the hypothetical value (e.g. People with high salaries generally do not want to reveal their incomes in surveys) or missing value is dependent on some other variable's value (e.g. Let's assume that females generally don't want to reveal their ages! Here the missing value in age variable is impacted by gender variable)

In the first two cases, it is safe to remove the data with missing values depending upon their occurrences, while in the third case (MNAR) removing observations with missing values can produce a bias in the model. So, it must be careful before removing observations. Note that imputation does not necessarily give better results. In Figure 2 (below) a schema of how handling missing values is shown.

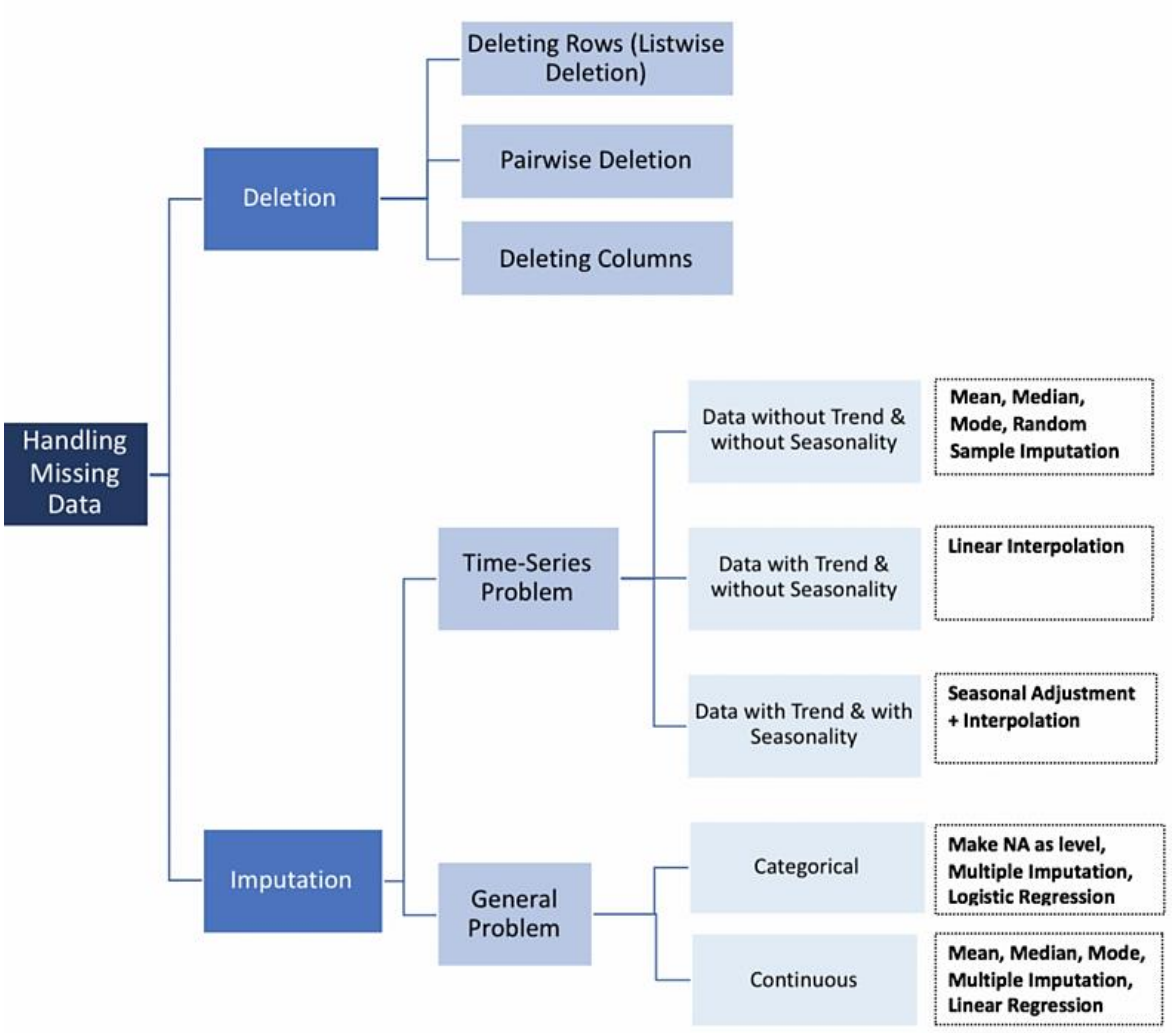


Figure 2. How to proceed with Missing Data

2.3 Range Normalization

The goal of normalization is to transform features to be on a similar scale. This improves the performance and training stability of the model. Two common normalization techniques may be useful:

- **Lineal Normalization:** For every feature, the minimum value of that feature gets transformed into a 0, the maximum value gets transformed into a 1, and every other value gets transformed into a decimal between 0 and 1.

$$x' = \frac{x - x_{min}}{x_{max} - x_{min}}$$

  By viewing the formula, this method is also known as “min-max” normalization.

- **Z-Score:** is a variation of scaling that represents the number of standard deviations away from the mean. You would use z-score to ensure your feature distributions have mean = μ = 0 and std = σ = 1.

$$x' = \frac{x - \mu}{\sigma}$$

  It is useful when the actual minimum and maximum of an attribute to be normalized are unknown.
- By **decimal scaling**. This normalizes by moving the decimal point of values of attribute.

2.4 Identification of out-of-range values (outliers)

In Data Science, an *outlier* is an observation point that is distant from other observations. An outlier may be due to variability in the measurement or it may indicate experimental error.

The decision to consider or discard an outlier needs to be taken at the time of building the model. Outliers can drastically bias/change the fit estimates and predictions. It is left to the best judgement of the analyst to decide whether treating outliers is necessary and how to go about it.

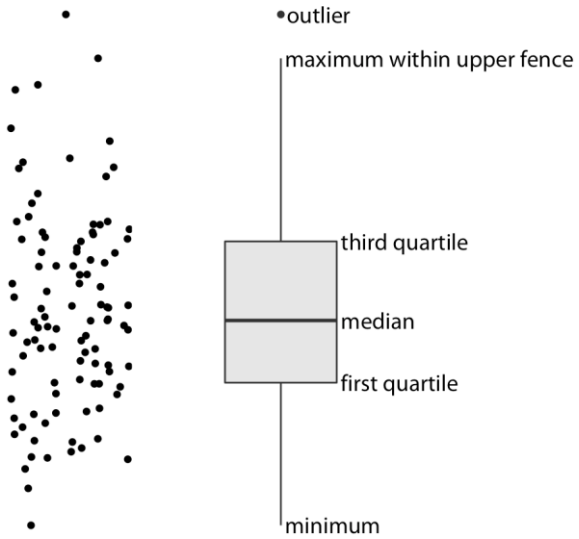


Figure 3. Outliers detection using boxplots

In descriptive statistics, a box plot is a method for graphically depicting groups of numerical data through their quartiles. Box plots may also have lines extending vertically from the boxes (whiskers) indicating variability outside the upper and lower quartiles, hence the terms box-and-whisker plot and box-and-whisker diagram (see previous figure). Outliers may be plotted as individual points.

2.5 Discretization

The purpose of discretization is finding a set of cut points to partition the range into a small number of intervals that have good class coherence, which is usually measured by an evaluation function. In addition to the maximization of interdependence between class labels and attribute values, an ideal discretization method should have a secondary goal to minimize the number of intervals without significant loss of class attribute mutual dependence.

The term “cut-point” refers to a real value within the range of continuous values that divides the range into two intervals, one interval is less than or equal to the cut-point and the other interval is greater than the cut-point. For example, a continuous interval [a, b] is partitioned into [a, c] and (c, b], where the value ***c*** is a cut-point (also known as split-point). The term *“arity”* in the discretization context means the number of intervals or partitions. Before discretization of a continuous feature, arity can be set to k—the number of partitions in the continuous features. The maximum number of cut-points is k − 1. Discretization process reduces the arity but there is a trade-off between arity and its effect on the accuracy.

A typical discretization process broadly consists of four steps: **(1) sorting** the continuous values of the feature to be discretized, **(2) evaluating a cut-point** for splitting or adjacent intervals for merging, **(3)** according to some criterion, **splitting or merging intervals** of continuous value, and **(4)** finally **stopping** at some point.

Data discretization include following techniques:

- **Binning**. This is a top-down unsupervised splitting technique based on a specified number of bins.
- **Histogram analysis**. Histogram partitions the values of an attribute into disjoint ranges called buckets or bins. It is also an unsupervised method.
- **Clustering** algorithm can be applied to discretize a numerical attribute by partitioning the values of that attribute into clusters or groups.
- **Decision tree analysis**. It employs a top-down splitting approach; it is a supervised method. To discretize a numeric attribute, the method selects the value of the attribute that has minimum entropy as a split-point, and recursively partitions the resulting intervals to arrive at a hierarchical discretization.
- **Correlation analysis**. Bottom-up approach by finding the best neighbouring intervals and then merging them to form larger intervals, recursively. It is supervised method.

2.6 Dimensionality Reduction

There are many techniques for dimensionality reduction. The objective of dimensionality reduction techniques is to appropriately select the k dimensions (and also the number k) that would retain the important characteristics of the original object. For example, when performing dimensionality reduction on an image, using a wavelet technique, then the desirable outcome is for the difference between the original and final images to be almost imperceptible.

When performing dimensionality reduction not on a single object, but on a dataset, an additional requirement is for the method to preserve the relationship between the objects in the original space. This is particularly important for reasons of classification and visualization in the new space.

There exist two important categories of dimensionality reduction techniques:

- **Feature selection techniques**, where only the most important or descriptive features/dimensions are retained, and the remaining are discarded.
- **Feature extraction methodologies**, which project the existing features onto different dimensions or axes. The aim here is again, to find these new data axes that retain the dataset structure and its variance as closely as possible. Examples: PCA, LDA, ICA, …

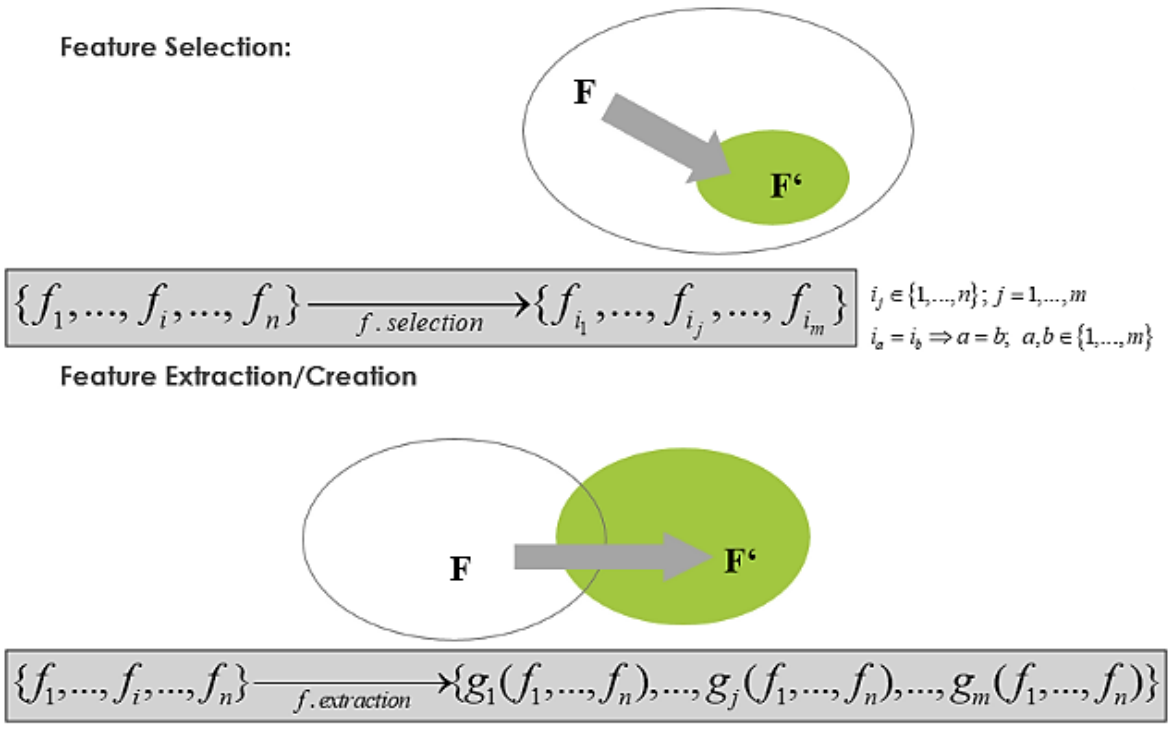


Figure 4. Feature selection vs feature extraction

3 FAIR Data

The FAIR acronym stands for data **F**indable, **A**ccessible, **I**nteroperable and **R**eusable.

The last thirty years have witnessed a revolution in digital technology. The rate and volume at which research data are created and the potential to make outputs readily available for analysis and reuse has increased exponentially. It has long been recognised that it is not enough simply to post data and other research-related materials onto the web and hope that the motivation and skill of the potential user would be enough to enable reuse. There is a need for various things, including contextual and supporting information *(metadata)*, to allow those data to be discovered, understood and used. From the G8 Science Ministers Statement (June-2013):

"Open scientific research data should be easily discoverable, accessible, assessable, intelligible, useable, and wherever possible interoperable to specific quality standards".

3.1 The four principles

- **Findable**: discoverable with metadata, identifiable and locatable by means of a standard identification mechanism. The way to achieve this is by ensuring: data has a persistent identifier (PID), it has rich metadata, and it is searchable and discoverable online.
- **Accessible**: always available and obtainable. Data should be retrievable online using standardised protocols. This *A* does not necessarily mean 'Open' or 'Free', but rather, gives the exact conditions under which the data are accessible. More information at https://www.dtls.nl/fair-data/fair-principles-explained/
- **Interoperable**: both syntactically *parseable* and semantically understandable, allowing data exchange and reuse between researchers, institutions, organisations or countries. Using common formats and standards, and controlled vocabulary.
- **Reusable**: sufficiently described and shared with the least restrictive licences, allowing the widest reuse possible and the least cumbersome integration with other data sources. To make this possible, ensure it is well-documented and it has a clear license and provenance information.

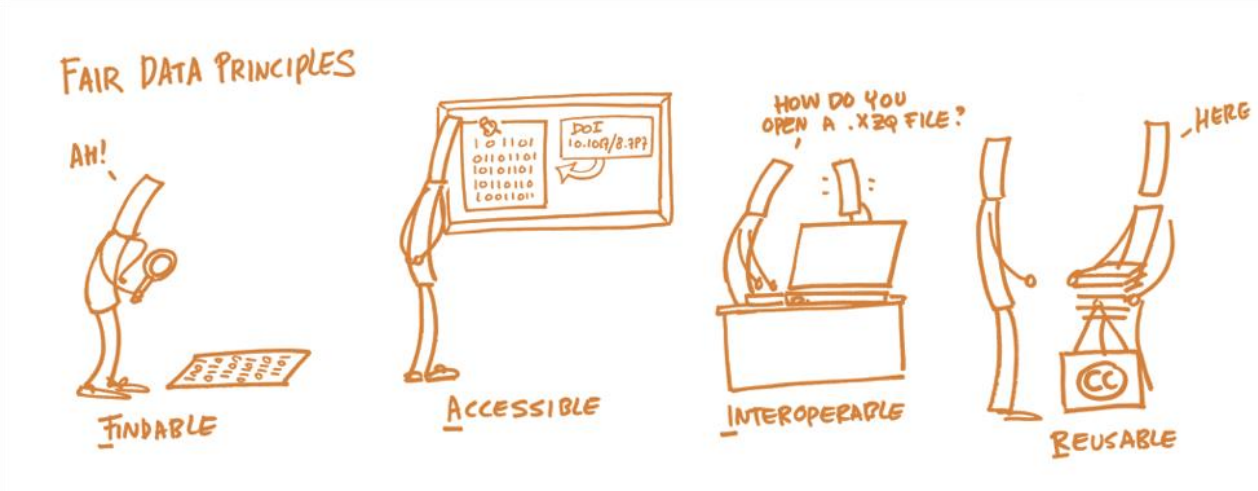


Figure 5. FAIR data principles (image from https://book.fosteropenscience.eu).

It is important to remind that FAIR is a set of **principles**; not a standard. In document https://doi.org/10.1038/sdata.2016.18 we could find them detailed:

FINDABLE

- F1. (meta)data are assigned a globally unique and persistent identifier.
- F2. data are described with rich metadata (defined by R1 below).
- F3. metadata clearly and explicitly include the identifier of the data it describes.
- F4. (meta)data are registered or indexed in a searchable resource.

ACCESSIBLE

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

INTEROPERABLE

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data uses vocabularies that follow FAIR principles.

- I3. (meta)data include qualified references to other (meta)data.

REUSABLE (REPRODUCIBLE)

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

Consider that FAIR is **not** a standard, **not equal to** "open" or "free", and FAIR **is not guarantee** for successful open science.

3.2 Key aspects to make your data management FAIR

There are some interesting aspects of FAIR principles to consider.

3.2.1 Semantic Web & Ontology

In addition to the classic "Web of documents" the W3C is helping to build a technology stack to support a "Web of data," the sort of data you find in databases. The term **Semantic Web** refers to W3C's vision of the Web of linked data. Semantic Web technologies enable people to create data stores on the Web, build vocabularies, and write rules for handling data.

On the Semantic Web, *vocabularies* define the concepts and relationships (also referred to as "terms") used to describe and represent an area of concern. In practice, vocabularies can be very complex (with several thousands of terms) or very simple (describing one or two concepts only). There is no clear division between what is referred to as **vocabularies** and **ontologies**. The trend is to use the word "ontology" for more complex, and possibly quite formal collection of terms, whereas "vocabulary" is used when such strict formalism is not necessarily used or only in a very loose sense

The semantic web principles are implemented in the layers of Web technologies and standards (see Figure 6).

- The **Unicode and URI** layers make sure that we use international characters sets and provide means for identifying the objects in Semantic Web.
- The **XML layer with namespace and schema** definitions make sure we can integrate the Semantic Web definitions with the other XML based standards.

- With **RDF and RDFSchema** (RDFS) it is possible to make statements about objects with URI's and define vocabularies that can be referred to by URI's. Here user can give types to resources and links.
- The **Ontology** layer supports the evolution of vocabularies as it can define relations between the different concepts.
- The **Logic** layer enables the writing of rules while the **Proof** layer executes the rules and evaluates together with the **Trust** layer mechanism for applications whether to trust the given proof or not. With the **Digital Signature** layer for detecting alterations to documents, these are the layers that are currently being standardized in W3C working groups.

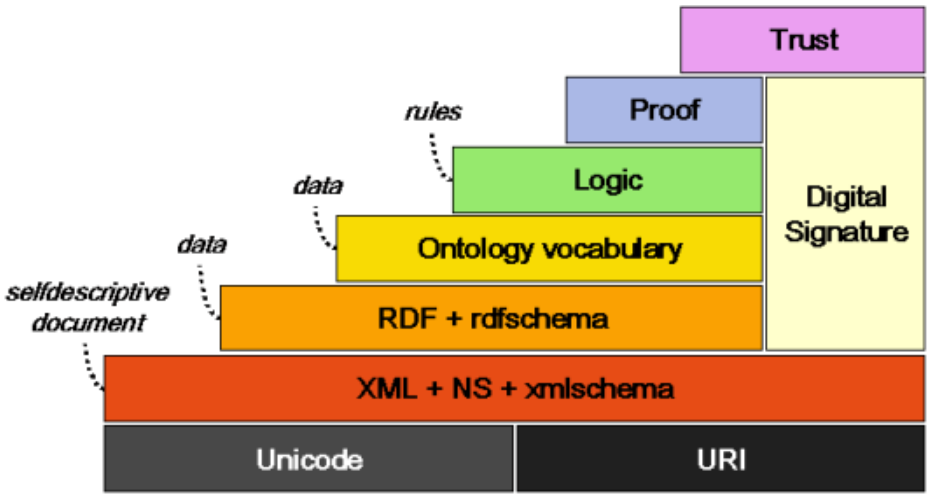


Figure 6. Semantic web layers (T. Berners-Lee)

To make the Web of Data a reality, it is important to have the huge amount of data on the Web available in a standard format, reachable and manageable by Semantic Web tools. Furthermore, not only does the Semantic Web need access to data, but relationships among data should be made available, too, to create a *Web of Data*.

The source of such extra information can be defined via vocabularies or rule sets. Both approaches draw upon knowledge representation techniques. In general, **ontologies** concentrate on classification methods, putting an emphasis on defining *classes*, *subclasses*, on how individual resources can be associated to such classes, and characterizing the relationships among classes and their instances. **Rules**, on the other hand, concentrate on defining a general mechanism on discovering and generating new relationships based on existing ones.

Usually, chose of a specific ontology (and its evaluation) is hard:

- Large **number and variety** of ontologies (versions, platforms, formats, etc.), different **complexity** level (from terminology to ontologies).
- **Automation** of the selection process?
- Diversity of **user requirements** and expectations
- What are the **risks of a bad choice**?

3.2.2 Metadata

From the beginning of FAIR Data section appears the concept of *metadata*. One definition could be:

> ***Is data about data, information about information... More comprehensive definitions address metadata as structured data supporting functions associated with an object, an object being any "entity, form, or mode"***

Examples of metadata:

- Metadata elements – schemas: *title, author, subject, date, type, coordinates…*
- Metadata values – schemas: *"physics", "2004-01-23"*
- *Digital format, terms and conditions, location & PID*.

Metadata is a part of the global conversation and is already recognized as a necessity for **interoperability**, **discovery** and **contextualisation**. Given that metadata is essential for data science endeavours, different communities of practice keep sharing their knowledge regarding metadata development, harmonisation and adoption to help perform better throughout the research process.

The *Research Data Alliance (RDA)* has defined several metadata principles:

1. The only difference between metadata and data is **mode of use**.
2. Metadata is not just for data, it is also for users, software services, computing resources.
3. Metadata is not just for **description** and **discovery**; it is also for **contextualisation** (relevance, quality, restrictions, rights, costs) and for coupling users, software and computing resources to data,
4. Metadata must be **machine-understandable** as well as human understandable for autonomicity (formalism).
5. **Management** (meta)data is also relevant (research proposal, funding, project information, research outputs, outcomes, impact…)

3.2.2.1 *Provenance*

Data provenance (PROV) is information about entities, activities, and people involved in producing (influencing or delivering) a piece of data. Concept comes from French *provenir* = to come from, and was originally used to keep track of the chain of ownership of cultural artefacts.

Provenance is often conflated with metadata. They are related, but not the same: Provenance is a kind of metadata, **but** not all metadata is provenance. For example, the *title of a book* is metadata, but it is not part of its provenance; the *date of creation, author, publisher or the license of a book* are part of its PROV. Picking up the figure Semantic web layers (T. Berners-Lee) (p. 20), provenance covers up to three layers:

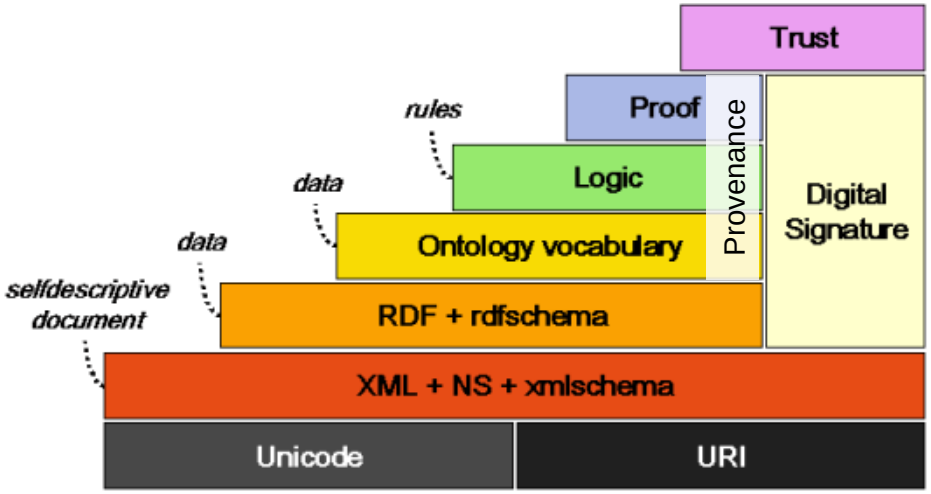


Figure 7. The Provenance concept in the Semantic web layers

3.2.3 PID (Persistent Identifier)

At a basic level, a PID is a **reference** to a person, place, or thing, which can be used to uniquely identify them, in perpetuity. PIDs can be

- **Internal** (for using within a single organisation).
- **Proprietary** (within a single system).
- **Open** (fully interoperable in any system).

Most commonly used PID for researching is *ORCID (Open Researcher and Contributor ID)*, probably the best, since they are the easiest to work with for everyone. Internal PIDs can also be added to ORCID records and shared in our own data files, but we only allow them to be categorized as *Non-standard ID from work data source*.

As a benefit of membership in ORCID, organizations can ask them to support additional PID types in the ORCID Registry[3]. For example,[4]: *asin* for "Amazon Standard Identification Number"; *doi* = "Digital Object Identifier"; of *isbn* "International Standard Book Number".

Other "desirable" features should have PIDs:

- **Resolvable**. These are either URLs (links), or can be transformed into URLs, which resolve directly to a document or a human-readable landing page using well-known rules.
- **FAIR**. PIDs can also be used to discover open, interoperable, well-defined metadata containing provenance information in a predictable manner.

3.2.4 Scientific Workflow (SWF)

SWFs allow users to easily express multi-step computational tasks. Typical phases can be data accessing, scheduling, generation, transformation, aggregation, analysis, visualization design, testing, share, deployment, execute, or reuse other SWFs. (Please review section 2 Data preprocessing).

There are defined several requirements for the Scientific Workflow Systems (SWFS), like:

- **Design tool**, especially for non-expert users.
- **Ease of use**: simple user interface having more complex features hidden in background.
- **Reusable** generic features.
- **Extensibility** for the expert user. Almost a visual programming interface.
- **Registration and publication** of data products and "process products" (workflows); provenance.
- **Error detection** and **recovery** from failure.
- **Logging information** for each workflow.
- Allow data-intensive and compute-intensive **tasks.**

[3] Request to add a new identifier type https://orcid.org/content/identifier-requests
[4] More identifiers at https://pub.orcid.org/v2.0/identifiers

- **Data management/integration.**
- Allow **status checks** and on the fly updates.
- **Semantics** and **metadata**-based dataset access.
- Certification, trust and security.

Researchers can find some SWFSs tools:

- **Taverna Workflow System**. Taverna enables a scientist who has a limited background in computing, limited technical resources and support, to construct highly complex analyses over data and computational resources that are both public and private.

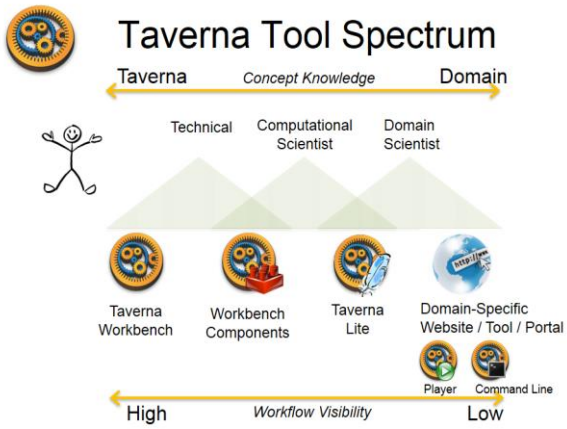


Figure 8. The Taverna tool spectrum

- **Galaxy.** It has produced numerous open source software offerings to help you build your science analysis infrastructure. This software covers the gamut from helping you integrate new software into our platform, to a production-ready engine to run those programs in complex *MapReduce* workflows.
- **Kepler** Project is dedicated to furthering and supporting the capabilities, use, and awareness of the free and open source, scientific workflow application, *Kepler*. It is designed to help scientists, analysts, and computer programmers create, execute, and share models and analyses across a broad range of scientific and engineering disciplines.

3.3 FAIR Digital Objects

Implementing FAIR requires a model for FAIR Digital Objects. These, by definition, have a PID linked to different types of essential metadata including provenance and licencing. See Figure 9 below.

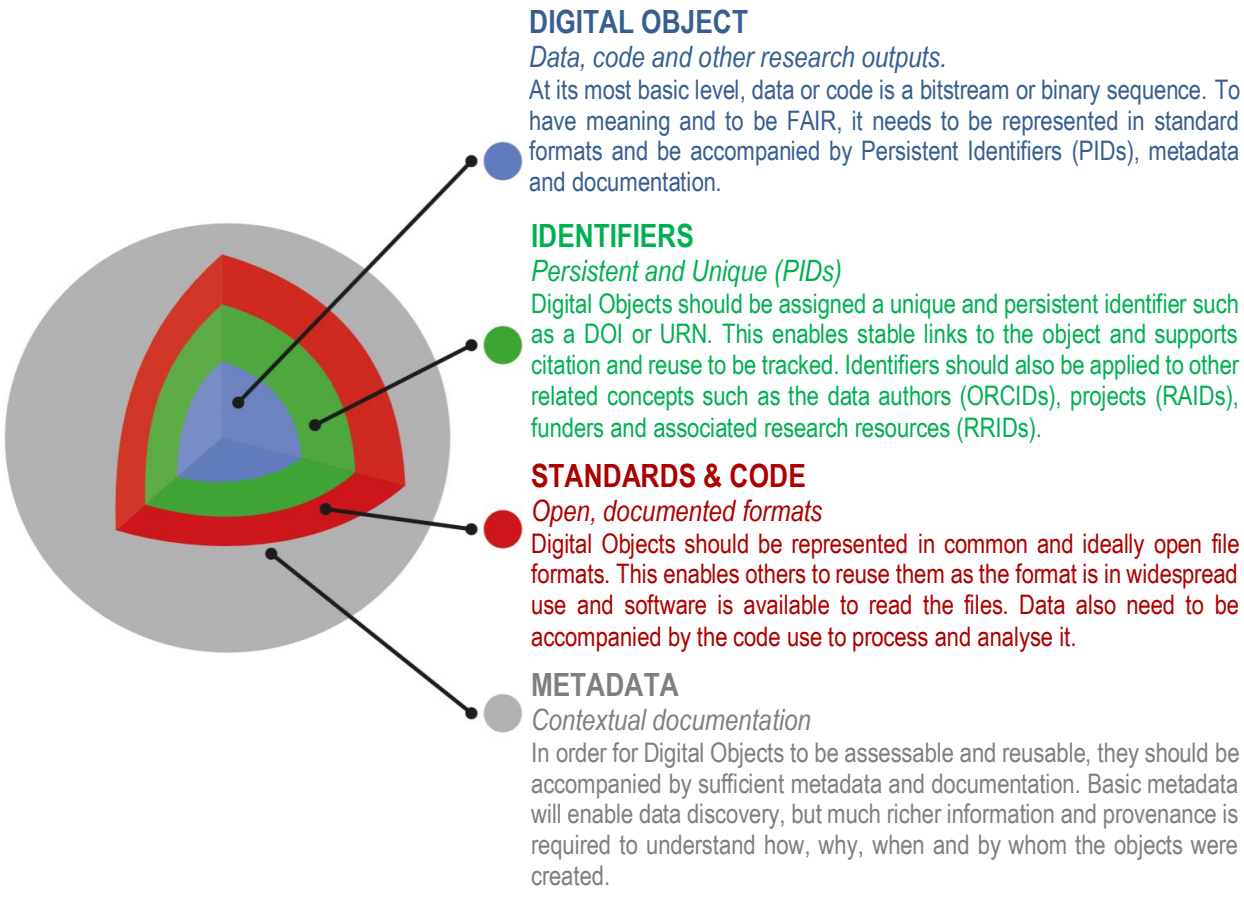


Figure 9. A layer model for FAIR Digital Objects.

3.3.1 Data life cycle

Figure 10 (top) shows the researching data cycle according to ENVRI reference model. Below, the same diagram with the provenance concept.

Figure 10. (Top) Research data life cycle; (Bottom) Data cycle with provenance.

3.4 CKAN

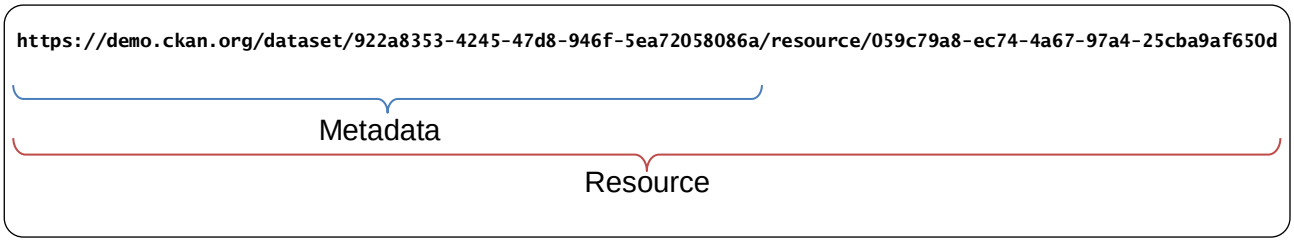
CKAN is Comprehensive Knowledge Archive Network. A web-based open-source management system for the storage and distribution of open data, and a powerful data catalogue system that is mainly used by public institutions to share their data.

It is built with Python on the backend and JavaScript on the frontend and uses The Pylons web framework and SQLAlchemy as its ORM. Its database engine is PostgreSQL and its search is powered by *Solr*. It has a modular architecture that allows extensions to be developed to provide additional features such as harvesting or data upload. CKAN provides a streamlined way to make your data discoverable and presentable. Each dataset is given its own page for the listing of data resources and a rich collection of metadata, making it a valuable and easily searchable data catalogue.

A use case of platform developed with CKAN is *Depositar*. It is a public platform for storing, preserving, managing, and exploring research data. A dataset (Figure 11) contains two things:

- Information or **metadata** about the data. For example, the title and publisher, date, what formats it is available in, what license it is released under, etc.
- A number of **resources**, which hold the data itself. CKAN does not mind what format the data is in. A resource can be a CSV or Excel spreadsheet, XML file, PDF document, image file, linked data in RDF format, etc. CKAN can store the resource internally, or store it simply as a link, the resource itself being elsewhere on the web. For example, different resources might contain the data for different years, or they might contain the same data in different formats.

Example:



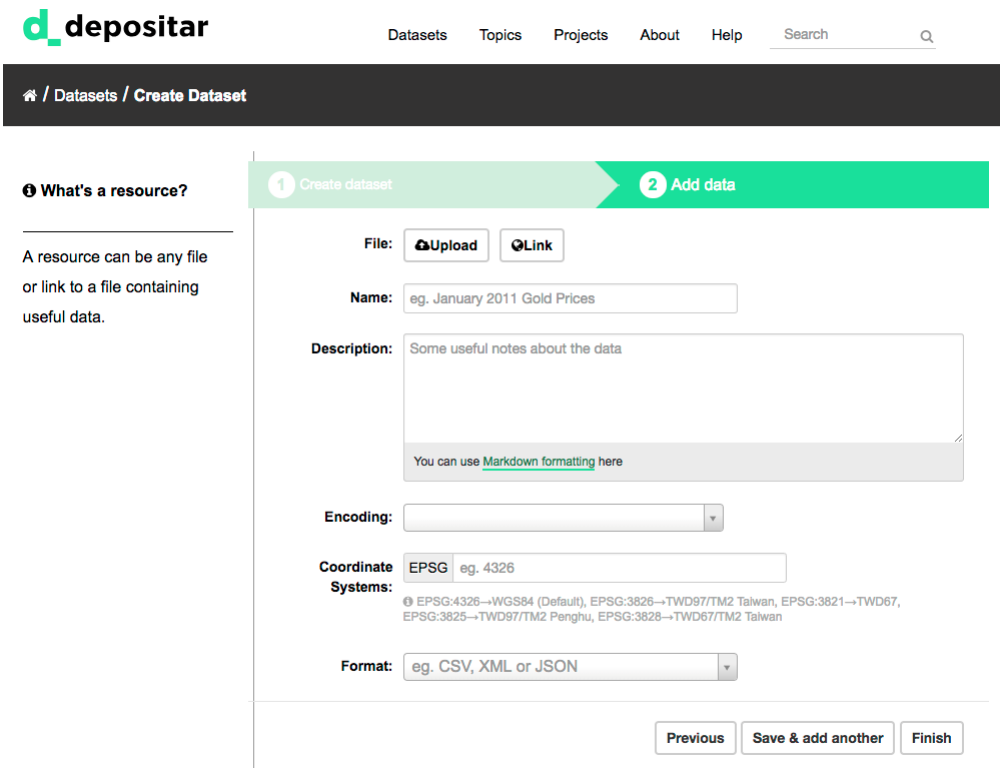


Figure 11. Screenshot for adding data in Depositar website

To find datasets in CKAN, user types any combination of search words (e.g. “health”, “transport”, etc) in the search box on any page. CKAN displays the first page of results for your search (see below figure). Options:

- View more pages of results.
- Repeat the search, altering some terms.
- Restrict the search to datasets with particular tags, data formats, etc using the filters in the left-hand column.

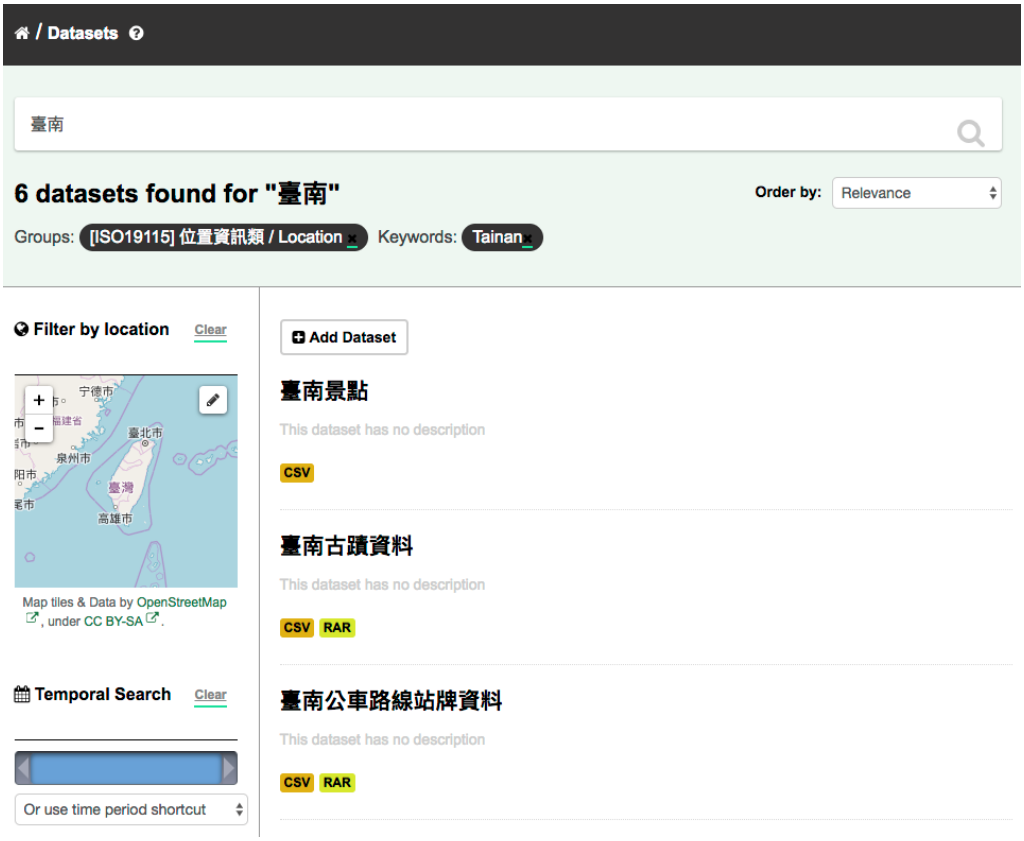


Figure 12. Screenshot of an example of searching in Depositar

4 Indicators

We can distinguish two common indicators to quantify the feasibility or status of a project (originally are design for business): **KPIs** *(Key Performance Indicator)* and **metrics**. They are similar, but not equal.

Both are as a quantifiable measurement of a strategic or tactical activity. *KPIs* are a quantifiable or measurable value that reflects a business goal or objective *(strategic)*. A *metric* is also a quantifiable or measurable value, but it reflects how successful the activities taking place are *(tactical)* to support the accomplishment of the KPI. Starting with metrics…

4.1 Metric

In its simplest form, a metric is a measurement that is recorded to track some aspect of the business activity and quantity the success or failure of the performance of that activity. A metric is simply a

number. For example, digital analytics tools offer a lot of numbers; web analytics reports typically include a lot of values.

4.1.1 RS systems metrics example

In the context of the *Danubius-PP* project, we can mention the US Department of the Interior (DOI) report titled *"Recommendations for assessing the effects of the DOI Hurricane Sandy Mitigation and Resilience Program on ecological system and infrastructure resilience in the Northeast coastal region"*. The document includes recommended ecological performance metrics for an assessment of regional changes in resilience along the Northeast coast and how these projects may increase resilience of ecological systems since Hurricane Sandy. It covers four main assessment components: **ecological performance metrics, socio-economic performance metrics, data management metrics,** and **filling of baseline data and information gaps**.

Inside Appendix C1 of mentioned report, we find the Ecological performance metrics (there are more zones than showed here):

- Riverine and Riparian zone. Categories:
    - **Biotic:** fish species health/recruitment stressors; fish migration rates and patterns; fish assemblage/abundance pre and post project; invasive species extent, mobility; biomass diversity, macro invertebrates pre-post; riparian plant community pre-post; biologic assimilation of contaminants; riparian and channel habitat measurements; habitat availability (stream miles made accessible to aquatic species upstream, pre and post).
    - **Abiotic:** river flow and depth; flooding extent and depth to create a volumetric measurement of stormwater retention capacity; flow rates across obstructions.; inundation area pre-post engineered change; sediment composition and contaminants (pre and post project); modelled potential for changes in flood regime up/down stream; water quality (temperature, salinity, pH, dissolved oxygen, turbidity, nutrients, contaminants); river biogeochemistry (flow-weighted water quality parameters); observed water levels (surface and ground); erosion rate and changes to sediment transport processes of system pre-post project.
    - **Structural/Engineering:** river hydrology (geomorphic mapping pre-post constriction removal); minimum change in connectivity needed to allow fish passage; elevation

change across obstructions; river hydrology change (percentage of flood-risk reduction, riparian buffer dimensions, position, pre-post project); number of barriers removed or remediated.

- Estuaries and Ponds.
    - **Biotic:** submerged aquatic vegetation biomass, species, density, extent, health; invasive species; invertebrate fauna; migratory shorebirds; species composition and abundance for fish, plankton, and the benthic animal community.
    - **Abiotic:** water depths- transects shore to vegetation perimeter; water flow patterns (circulation within an estuary, and ground and surface water inputs); inundation extent, rates and frequency; water levels, flows and wave heights; water quality (temperature, salinity, pH, dissolved oxygen, turbidity, nutrients, chlorophyll, contaminants).
    - **Structural/Engineering:** physiographic characteristics (depth, fetch, substrate, perimeter characteristics, flow patterns).
- Beach System (beach, barrier island and dune).
    - **Biotic:** fish and wildlife population/recruitment/overwintering/stopover weight/health relative to other mitigating factors; intra-faunal abundance and diversity (invertebrates); vegetation cover of dunes pre and post events.
    - **Abiotic:** surge, wave and tide hydrodynamic network; pre-post storm wave height, inundation level; water flow velocity and current dynamics; water levels (including back bays); pre and post storm rates of erosion; volumes of material in flood and ebb shoals; change in near shore sediment character and movement; water quality (ex. Temperature, salinity, pH, dissolved oxygen, turbidity, nutrients, contaminants).
    - **Structural/Engineering:** dune characterization (height, width, length, texture, substrate); beach width, elevation, volume, shoreline position; breach morphology; shoreline position and topography.

4.2 KPI

Like metrics, KPIs must be very well defined and are also quantifiable. The difference is KPIs define a set of values against which the metrics are measured.

According to R. Hatheway[5], *"How they relate to each other is extremely simple: metrics support KPIs. KPIs in turn support the overall business strategic goals and objectives."*

They can be categorized according to:

- **Qualitative vs Quantitative**. Facts without distortion from personal feelings, prejudices, or interpretations presented with a specific value.
- **Lagging and leading** indicators. First one measures an output, a result; the leading indicator is a predictor, though imperfect of the future state of related lagging indicators.

KPIs should follow the **SMART** criteria. This means the measure has a **Specific** purpose for the business, it is **Measurable** to really get a value of the KPI, the defined norms have to be **Achievable**, the improvement of a KPI has to be **Relevant** to the success of the organization, and finally it must be **Time** phased, which means the value or outcomes are shown for a predefined and relevant period.

Figure 13. The SMART objectives for KPIs

In R&D scope, given the variety of disciplines, processes, and study designs associated with this field, it is important to have a holistic KPI program that considers different types of research and organizational goals. The traditional budgeting approach to measuring innovations does not work. A typical innovation is not a production line that converts ideas into commercial products; it involves many stakeholders, and the measurement efforts should take them into account. In Figure 14, reader can find proposed innovation funnel.

[5] Richard Hatheway is Full Stack Marketing Executive / Business Consultant. https://www.linkedin.com/pulse/real-difference-between-metrics-kpis-richard-hatheway/

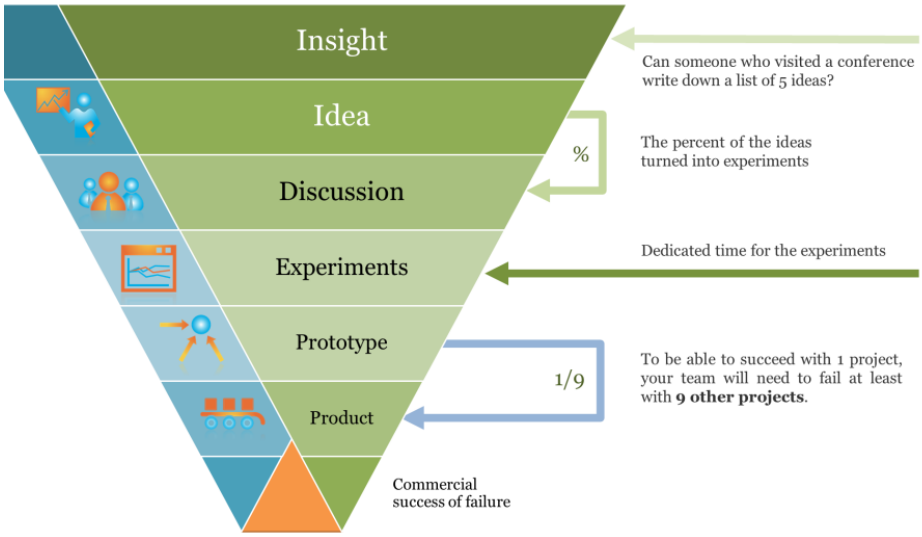


Figure 14. How to measure innovations with KPIs (image from https://bscdesigner.com)

One case of use of KPIs in a real project is *Hanson UK*. It is a leading supplier of heavy building materials to the construction industry.

In the web site[6], Hanson UK has a section about **water**. The use of mains water fell to its lowest recorded level of 17.5 litres per tonne of product, while absolute mains water uses also fell. We have also installed smart meters at 25 of our biggest water-using sites, allowing better management. They developed biodiversity action plans (BAPs) in place for all our quarries and they are all published on our website along with several geodiversity action plans (GAPs). They also own an indicator looking at quarries with high biodiversity value. Viewing document *"Summary of KPI performance against 2020 targets"*, some KPIs can be found related with:

- **People and communities**, including health and safety; stakeholder performance, environmental incidents and emissions; employment and skills; and local community.
- **Carbon and energy** comprising energy efficiency, waste as fuel, CO_2 emissions from production; and CO_2 emissions from transport.
- **Waste and raw materials**: waste minimisation; materials efficiency and recycling; product quality and performance.
- **Water and biodiversity**. Containing these targets for 2020:

[6] https://www.hanson-sustainability.co.uk/en/water/water

- *Water*: reduce mains water consumption by 25 per cent per tonne across the business by 2020 based on 2010. Reduce the sum of mains and abstracted water for concrete by 10 per cent per tonne by 2020 (based on 2010).
- *Biodiversity and site stewardship*: All quarries to implement published biodiversity action plans.

4.3 Metrics and KPIs for data processing/computing

This section centres into metrics (and KPIs) to measure the quality of data handling and level of computation of the servers. Among all existing literature, we remark some of them.

4.3.1 KPIs for Big Data

Big data is used for a wide range of predictive and behaviour analysis. Organizations apply big data to reduce costs, understand customer needs better, and to mitigate risks. The web article *"KPIs for Big Data Initiatives"* by A. Savkin[7] distinguishes three levels to tackle the problem of measurement:

1. **The 3-V metrics** can be easily quantified:
   - **Volume** of data is a measure by itself.
   - **Variety** can be quantified as the number of different types of data sources.
   - **Velocity** is defined by the volume of data generated/analysed per time period.

   Sometimes it added a fourth "V" for **Veracity**. It might be more difficult to quantify. You will need to define what your team qualifies as accurate data and that depends on the context.

[7] Available at https://bscdesigner.com/kpis-for-big-data.htm

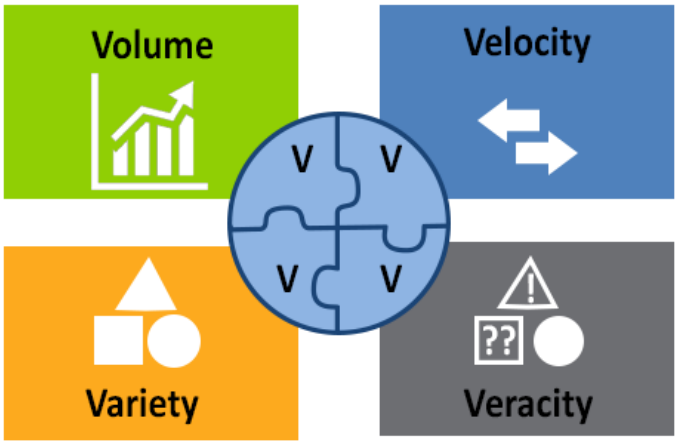


Figure 15. The "4-Vs" of Big Data metrics (image from https://www.zarantech.com/blog/the-4-vs-of-big-data/)

2. **Big data process metrics** (time-related):
    - **Frequency of data collection**.
    - **Time needed for data to be available** for analysis.
    - **Time needed for data to be reported** in a form of KPIs.

    In this group, we could add: **Query to report conversion rate** (%); and **Data capturing capabilities**.

3. **Lagging KPIs, to validate big data success**
    - What lessons do we learnt from big data? What cost saving was achieved after implementation of those ideas?
    - How did the customer retention rate change due to delivering a tailor-made experience? How is customer lifetime value changing?
    - Does big data help customer service to be more effective? How did the first-call resolution rate change?
    - How did the hiring processes change after starting to use big data? How did the time to performance HR metric change?
4. **Leading KPIs: ensuring big data success:**
    - Funds invested in big data initiatives.
    - Time spend on big data initiatives.

4.3.2 Metrics for Sustainable Data Centres

This article is from V. Dinesh Reddy, Brian Setz, G. Subrahmanya V. R. K. Rao, G. R. Gangadharan, and Marco Aiello. They define up to 9 metrics categories or dimensions. A summary is presented in table on next page.

Dim.	Use	Metrics	Issues and Challenges
Energy Efficiency	A series of indicators relevant to quantitative measure of energy efficiency of datacentre and its components. Some metrics are used to know how efficiently a data centre transfers power from the source to the IT equipment and some metrics define IT load versus overhead	APC: Adaptability Power Curve CADE: Corporate Average Data Centre Efficiency CPE: Compute Power Efficiency DCA: DCAdapt DCcE: Data Centre Compute Efficiency DCeP: Data Centre Energy Productivity DCiE: Data Centre Infrastructure Efficiency DCLD: Data Centre Lighting Density DCPD: Data Centre Power Density DCPE: Data Centre Performance Efficiency DC-FVER: Data Centre Fixed to Variable Energy Ratio DH-UE: Deployed Hardware Utilization Efficiency DH-UR: Deployed Hardware Utilization Ratio DPPE: Data Centre Performance Per Energy DWPE: Data centre Workload Power Efficiency EES: Energy ExpenseS EWR: Energy Wasted Ratio GEC: Green Energy Coefficient H-POM: IT Hardware Power Overhead Multiplier ITEE: IT Equipment Energy ITEU: IT Equipment Utilization OSWE: Operating System Workload Efficiency PDE: Power Density Efficiency	Energy consumption data disaggregated by data centre subcomponents may not be available. It is hard to know the number of operating systems and virtual machines running in a data centre.

Dim.	Use	Metrics	Issues and Challenges
		PEsavings: Primary Energy Savings PUE1-4: Power Usage Effectiveness Level 1-4 PUEscalability: Power Usage Effectiveness Scalability pPUE: Partial Power Usage Effectiveness PpW: Performance per Watt ScE: Server Compute Efficiency SI-POM: Site Infrastructure Power Overhead Multiplier SPUE: Server Power Usage Efficiency SWaP: Space, Watts and Performance TUE: Total-Power Usage Effectiveness	
Cooling	These metrics characterize the efficiency of the HVAC systems and how well these serve the cooling demand	AEUF: Air Economizer Utilization Factor CoP: Coefficient of Performance Ensemble DCCSE: Data Centre Cooling System Efficiency DCSSF: Data centre Cooling System Sizing Factor EER: Energy Efficiency Ratio HSE: HVAC System Effectiveness RI: Recirculation Index WEUF: Water Economizer Utilization Factor	It is challenging to determine whether there is adequate under–floor cooling in a consistently advancing environment, where heat densities change within rack and one rack to the next. Data centre cooling system must balance ambient environment with supplemental cooling to optimize efficiency.

Dim.	Use	Metrics	Issues and Challenges
Greenness	explain the carbon foot- print of the data centres and IT equipment. Also, we can assess how much green energy used, how much energy is exported for reuse and how efficiently a data centre is using water	CO2 Savings: Ratio CUE: Carbon Usage Effectiveness EDE: Electronics Disposal Efficiency ERE: Energy Reuse Effectiveness ERF: Energy Reuse Factor GEC: Green Energy Coefficient GUF: Grid Utilization Factor MRR: Material Recycling Ratio Omega: Water Usage Energy / v TCE: Technology Carbon Efficiency TGI: The Green Index WUE: Water Usage Effectiveness	Some of these metrics requires seasonal benchmarking to capture region and season changes.

Dim.	Use	Metrics	Issues and Challenges
Performance	These metrics measure the productivity of data centre, effectiveness in delivering service and agility in responding dynamically to change.	ACE: Availability, Capacity, and Efficiency Performance Score CPU: Central Processing Unit Usage DCP: Data Centre Productivity DEEPI: Data Centre Energy Efficiency and Productivity Index DR: Dynamic Range EP: Energy Proportionality FpW: Flops per Watt IPR: Idle-to-peak Power Ratio LD: Linear Deviation LDR: Linear Deviation Ratio PG: Proportionality Gap SWaP: Space, Watts and Performance UDC: Data Centre Utilization Userver: Server Utilization UCF: Uninterruptible Power Supply Crest Factor UPEE: Uninterruptible Power Supply Energy Efficiency UPF: Uninterruptible Power Supply Power Factor UPFC: Uninterruptible Power Supply Power Factor Corrected USF: Uninterruptible Power Supply Surge Factor	“useful computing work” is not defined uniquely. Correct base scores may be challenging without the right tools

Dim.	Use	Metrics	Issues and Challenges
Thermal & Air Management	These metrics help us to take care of efficient air flow, temperature issues and aisle pressure management.	-: Airflow Efficiency BPR: Bypass Ratio BR: Balance Ratio CI: Capture Index DC: Data Centre Temperature DP: Dew Point HF: Heat Flux IoT: Imbalance of Temperature -: Mahalanobis Generalized Distance (D2) M: Mass Flow Mc, Mn, Mbp, Mr, Ms RCI: Rack Cooling Index -: Relative Humidity RHI: Return Heat Index RR: Recirculation Ratio RTI: Return Temperature Index SHI: Supply Heat Index -: b-index	It is difficult to make proper aisle arrangement. For efficient airflow, we must address bypass & re-circulation air flow

Dim.	Use	Metrics	Issues and Challenges
Network	These metrics give the data centre network energy efficiency, utilization and traffic demands	BJC: Bits per Joule Capacity CNEE: Communication Network Energy Efficiency DS: Diameter Stretch ECR-VL: Energy Consumption Rating Variable Load NPUE: Network Power Usage Effectiveness -: Network Traffic per Kilowatt-Hour PS: Path Stretch RSmax: Maximum Relative Size TEER: Telecommunications Energy Efficiency Ratio Unetwork: Network Utilization	Measuring variable energy vary from one to another operator. Useful work is not defined properly
Storage	Using these metrics storage operations and performance can be monitored. We get better visibility into how proficiently our capacity is being utilized to store client information	-: Capacity LSP: Low-cost Storage Percentage -: Memory Usage OSE: Overall Storage Efficiency RT: Response Time SU: Slot Utilization -: Throughput Ustorage: Storage Usage	Measuring customer stored data and its criticality is difficult due to data duplication and the users view differ from the storage frame view

Dim.	Use	Metrics	Issues and Challenges
Security	These metrics are useful for protecting servers from attacks and continuously monitor physical and virtual servers and clouds. Further these metrics cover some basic measurements of firewall performance in a data centre	ACPR: Average Comparisons Per Rule AS: Accessibility Surface ATR: Application Transaction Rate CC: Concurrent Connections CER: Connection Establishment Rate CTR: Connection Tear down Rate DeD: Defense Depth DeP: Detection Performance DTE: Data Transmission Exposure FC: Firewall Complexity -: HTTP Transfer Rate IAS: Interface Accessibility Surface IPFH: IP Fragmentation Handling -: IP throughput ITH: Illegal Traffic Handling -: Latency RA: Rule Area RC: Reachability Count RCD: Rogue Change Days T: Vulnerability Exposure	These metrics are highly dependent on internal governance, compliance standards and SLA

Dim.	Use	Metrics	Issues and Challenges
Financial Impact	These metrics calculate total cost of ownership, financial impact of data centre outages, return on investments on management tools and technologies for sustainable data centre	A: Availability BVCI: Business Value of Converged Infrastructure CapEx: Capital Expenditure CCr: Carbon Credit MTBF: Mean Time Between Failures MTTF: Mean Time To Failure MTTR: Mean Time To Repair OpEx: Operational Expenditure ROI: Return On Investment TCO: Total Cost of Ownership - : Reliability	Confidentiality concerns associated with revealing costs for a particular facility. Carbon Credit may vary based on the country policies

5 Tools

5.1 Spreadsheets

A spreadsheet is the computerized analogue of paper accounting worksheets. It enables the user to tabulate and collate data. This data can then be used to make calculations, show graphical representations or analysis. A spreadsheet comprises of a grid of “grid” arranged in rows and columns and information can be inserted into each cell. What makes a spreadsheet software program unique is its ability to calculate values using mathematical formulas and the data in cells.

Some computer software to manage spreadsheets:

- Microsoft Excel (part of Microsoft Office).
- Calc (part of OpenOffice suite).
- Google (spread)Sheets (online)
- iWork Numbers (Apple Office suite).
- Lotus Symphony Spreadsheets (IBM).

5.1.1 Microsoft Excel

Spreadsheet software developed for Windows, MacOS, Android and iOS. It features calculation, graphing tools, pivot tables, and a macro programming language called *Visual Basic for Applications (VBA)*.

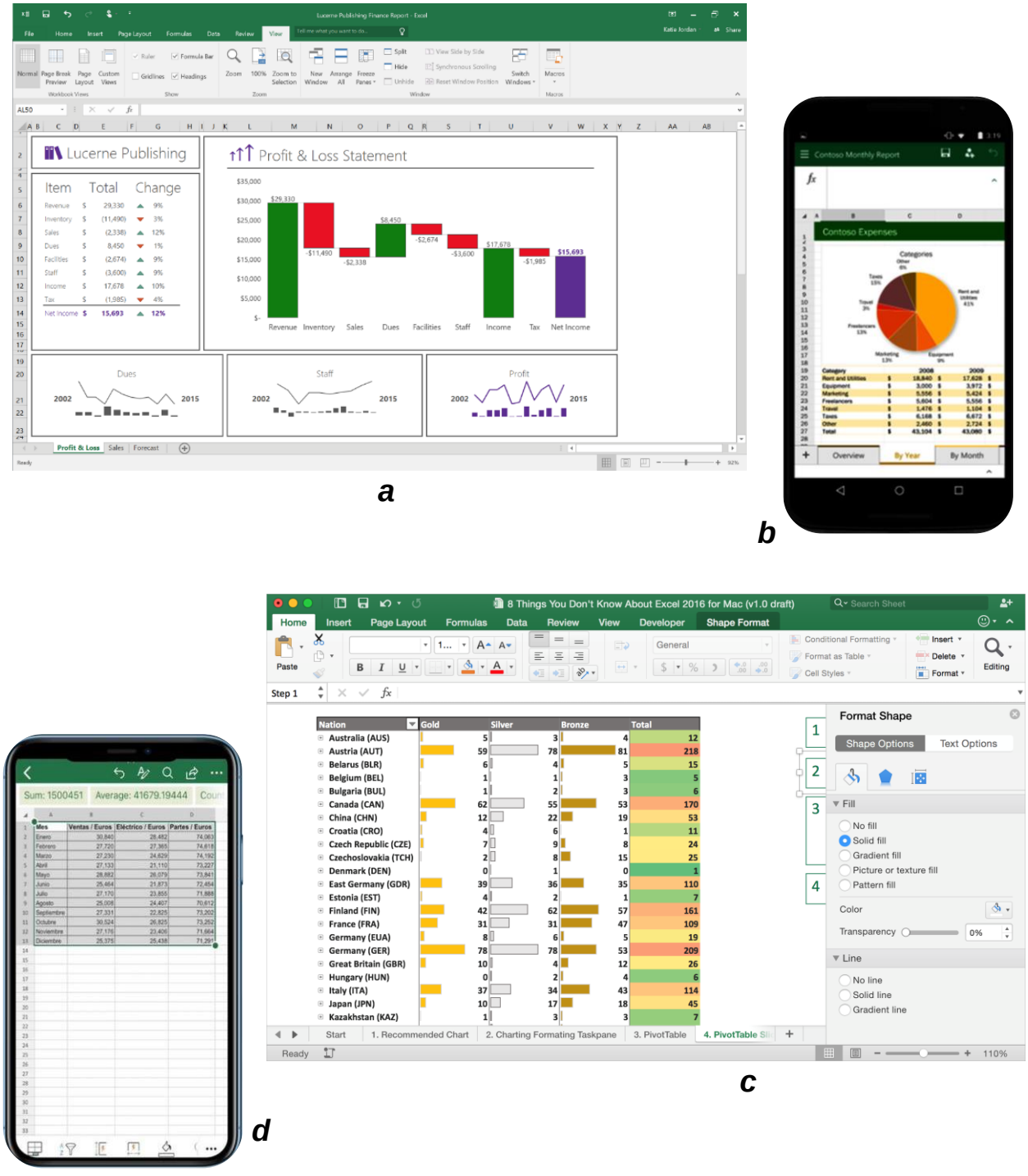


Figure 16. Screenshots of MS Excel in different environments: (a) Windows, (b) Android, (c) MacOS, (d) iOS.

Microsoft developed a competing spreadsheet (Lotus 1-2-3 suite dominated the mid-80s market), and the first version of Excel was released in 1985 for Apple Inc.'s Macintosh computer. Featuring strong graphics and fast processing, the new application quickly became popular.

It has a battery of supplied functions to answer statistical, engineering and financial needs. In addition, it can display data as line graphs, histograms and charts, and with a very limited three-dimensional graphical display. It allows sectioning of data to view its dependencies on various factors for different perspectives. In a more elaborate realization, an Excel application can automatically poll external databases and measuring instruments using an update schedule, analyse the results, make a *Word* report or *PowerPoint* slide show, even e-mail these presentations.

Some remarkable features include:

- **Macro programming.** The Windows version of Excel supports programming through Microsoft's Visual Basic for Applications, which is a dialect of Visual Basic. Programmers may write code directly using the Visual Basic Editor, which includes a window for writing, debugging, and code module organization environment. The user can implement numerical methods as well as automating tasks such as formatting or data organization in VBA and guide the calculation using any desired intermediate results reported back to the spreadsheet.

  A useful (and common) way to generate VBA code is by using the Macro recording function. It records actions of the user and generates the application code in the form of a macro. Then, the user can modify and edit this "base" code to add graphical user prompts.

  VBA code interacts with the spreadsheet through the *Excel Object Model*, a vocabulary identifying spreadsheet objects. Each object has its own properties and methods that we can use to make decisions and take actions with our code.

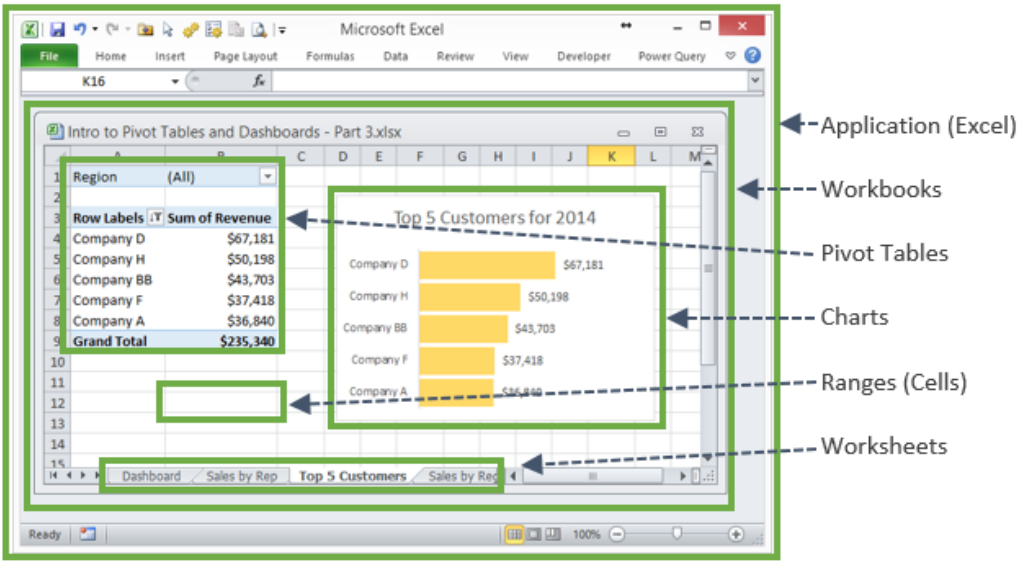


Figure 17. Type of objects in an Excel spreadsheet: "everything is an object".

- **Charts.** Excel supports charts, graphs, or histograms generated from specified groups of cells. The generated graphic component can either be embedded within the current sheet or added as a separate object. These displays are dynamically updated if the content of cells change.
- **Add-ins.** They are like packages that add new functions to Excel. Some of the are provided with the software: *Analysis ToolPak* (statistical and engineering tools), *Euro Currency Tools*, or *Solver Add-In* (optimization and equation solving).

5.1.2 Open spreadsheet software

Inside this title we can find several organisations/enterprises that develop office suites[8]. For example, first releases were from *StarOffice*, and nowadays (summer-2019) we find:

- **NeoOffice** is an office suite for MacOS developed by *Planamesa Inc*. It is a commercial fork of the free/open source OpenOffice.org that implements most of the features of OpenOffice.org, including a word processor, spreadsheet, presentation program, and graphics program,
- **LibreOffice** is a free and open-source office suite, a project of The Document Foundation. It was forked from OpenOffice.org in 2010, which was an open-sourced version of the earlier *StarOffice*. It comprises programs for word processing, the creation and editing of spreadsheets, slideshows, diagrams and drawings, working with databases, and composing mathematical formulae.
- **Apache OpenOffice** is an open-source office productivity software suite. It is one of the successor projects of *OpenOffice* and the designated successor of IBM Lotus Symphony. It is a close cousin of LibreOffice and NeoOffice. It contains a word processor (Writer), a spreadsheet (Calc), a presentation application (Impress), a drawing application (Draw), a formula editor (Math), and a database management application (Base).

This part covers *Apache OpenOffice Calc*. Newcomers will find **Calc** intuitive and easy to learn; professional data miners and number crunchers will appreciate the comprehensive range of advanced functions.

[8] Differences between *free software, Open source*, and *Freeware*: https://dzone.com/articles/free-software-vs-open-source-vs-freeware-whats-the

- *DataPilot* is an advanced technology that makes it easy to pull in raw data from corporate databases; cross-tabulate, summarize, and convert it into meaningful information.
- *Natural language formulas* let you create formulas using words (e.g. "sales - costs").
- *Natural language* formulas let you create formulas using words (e.g. "sales - costs").
- *Wizards* guides you through choosing and using a comprehensive range of advanced spreadsheet functions or download templates.
- *Scenario Manager* allows "what if ..." analysis at the touch of a button - e.g. compare profitability for high / medium / low sales forecasts.
- Calc's *solver* component allows solving optimization problems where the optimum value of a spreadsheet cell must be calculated based on constraints provided in other cells.

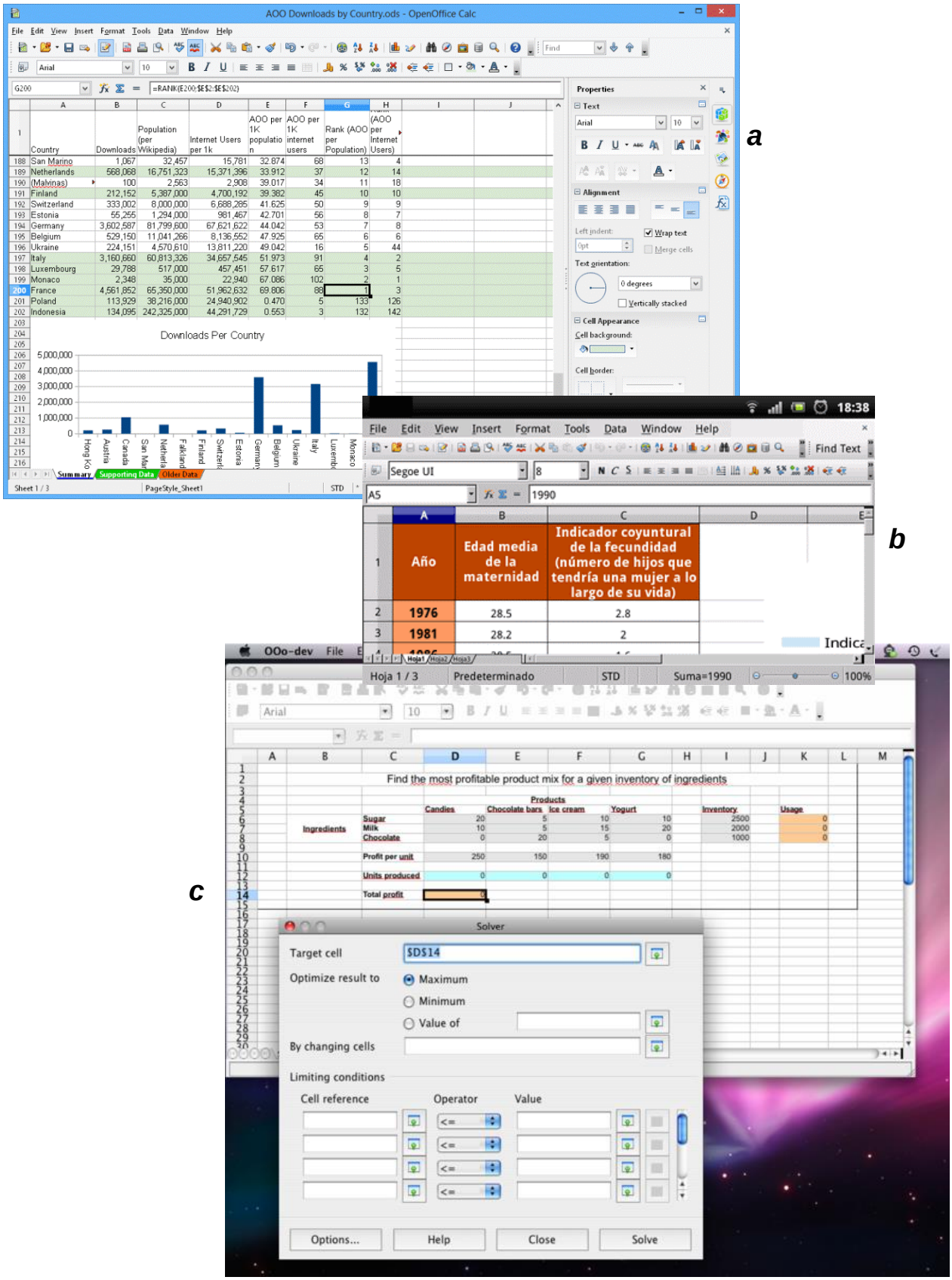

a
b
c
AOO Downloads by Country.ods - OpenOffice Calc
Downloads Per Country
Summary
Supporting Data
Older Data
Año
Edad media de la maternidad
Indicador coyuntural de la fecundidad (número de hijos que tendría una mujer a lo largo de su vida)
1976
28.5
2.8
1981
28.2
2
Find the most profitable product mix for a given inventory of ingredients
Products
Candies
Chocolate bars
Ice cream
Yogurt
Inventory
Usage
Ingredients
Sugar
Milk
Chocolate
Profit per unit
Units produced
Total profit
Solver
Target cell
D14
Optimize result to
Maximum
Minimum
Value of
By changing cells
Limiting conditions
Cell reference
Operator
Value
Options...
Help
Close
Solve

Figure 18. Screenshots of OpenOffice CALA in different environments: (a) Windows, (b) Android, (c) MacOS.

5.1.3 Google Sheets

It is the spreadsheet part of the web-based software office suite offered by Google. Google Sheets is available as a web application, mobile app for Android, iOS, Windows, BlackBerry, and as a desktop application on Google's ChromeOS. The Ajax-based program is compatible with Microsoft Excel and CSV files. Spreadsheets can also be saved as HTML.

Major features:

- **Collaboration and revision history**. Documents can be shared, opened, and edited by multiple users simultaneously and users are able to see *character-by-character* changes as other collaborators make edits. Changes are automatically saved to Google's servers, and a revision history is automatically kept so past edits may be viewed and reverted to.
- **Explore**. This function uses machine learning enabling users to type something wanted to know in natural language, such as *How many units were sold on two-years-ago Black Friday?* Google will provide answers and even automatically generate formulas.
- **Add-ons**. Like Excel, Google Sheets can incorporate software packages to enhance spreadsheets (with new functionalities, styles…).

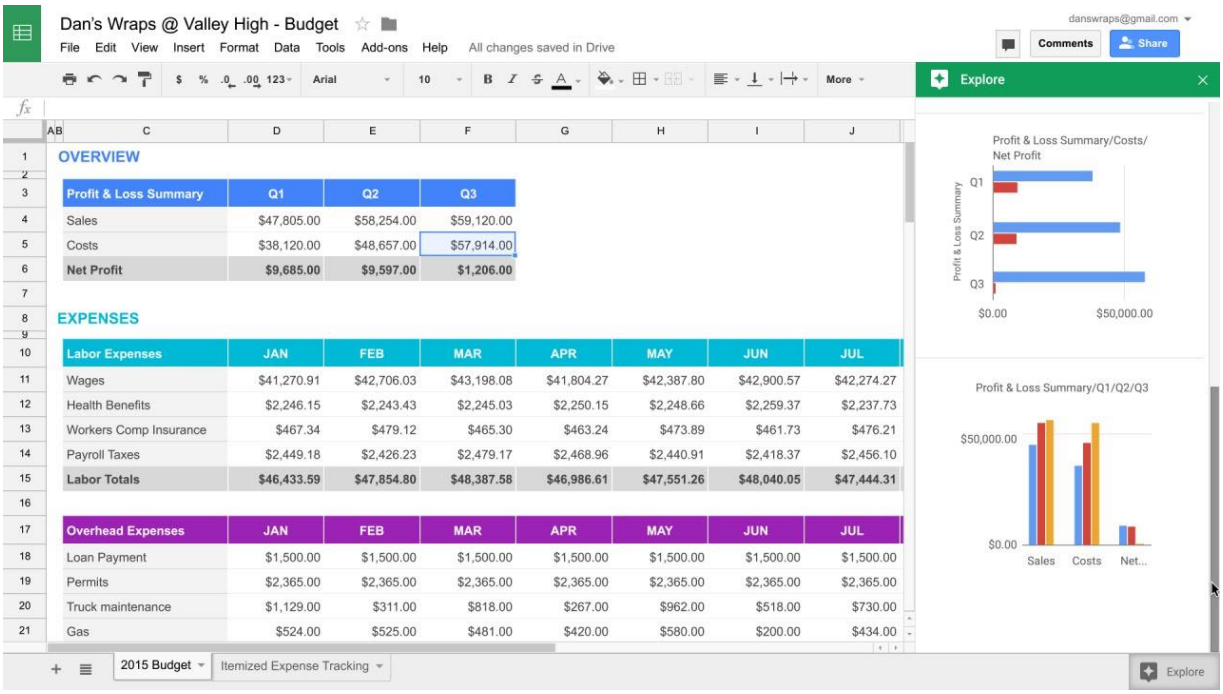


Figure 19. Screenshots of Google Sheet with the Explore tool.

5.2 Databases

When it comes to choosing a database, one of the biggest decisions is picking a relational *(Structured Query Language, SQL)* or non-relational (NoSQL) data structure. While both are viable options, there are certain key differences between the two that users must keep in mind when making a decision. SQL databases are table-based, while NoSQL databases are either document-based, key-value pairs, graph databases or wide-column stores.

Figure 20. Comparison between SQL and NoSQL databases.

- **SQL** is a strong choice for any project that will benefit from its pre-defined structure and set schemas. For example, applications that require multi-row transactions - like accounting systems or systems that monitor inventory - or that run on legacy systems will thrive with the SQL structure.
- **NoSQL**, on the other hand, is a good choice for businesses that have rapid growth or databases with no clear schema definitions. More specifically, if you cannot define a schema for your database, if you find yourself de-normalizing data schemas, or if your schema continues to change - as is often the case with mobile apps, real-time analytics, content management systems, etc. - NoSQL can be a strong choice.

Figure 21. Examples of SQL engineers (left side) and NoSQL engineers (right).

5.3 Programming Languages & Related Software

Commonly used languages for bio and geographic data processing, and associated software.

5.3.1 R

R is a language and environment for statistical computing and graphics. It is a GNU project which is similar to the S language and environment which was developed at Bell Laboratories (formerly AT&T, now Lucent Technologies) by John Chambers and colleagues. R can be considered as a different

implementation of S. There are some important differences, but much code written for S runs unaltered under R. It compiles and runs on a wide variety of UNIX platforms, Windows and MacOS.

R provides a wide variety of statistical (linear and nonlinear modelling, classical statistical tests, time-series analysis, classification, clustering, …) and graphical techniques, and is highly extensible. The S language is often the vehicle of choice for research in statistical methodology, and R provides an Open Source route to participation in that activity.

One of R's strengths is the ease with which well-designed publication-quality plots can be produced, including mathematical symbols and formulae where needed. Great care has been taken over the defaults for the minor design choices in graphics, but the user retains full control.

An example of project using the R language oriented to Hydrological Data and Modelling is located in the URL https://cran.r-project.org/web/views/Hydrology.html. It is a part of the **CRAN** (Comprehensive R Archive Network), and divided in several sections:

- **Data Retrieval**. Hydrological data sources (surface water/groundwater quantity and quality), and Meteorological data (precipitation, radiation, temperature, etc - including both measurements and reanalysis).
- **Data Analysis**. Data tidying (gap-filling, data organization, QA/QC, etc); Hydrograph analysis (functions for working with streamflow data, e.g. flow statistics, trends, biological indices, etc.); Meteorology (functions for working with meteorological and climate data); Spatial data processing; etc.
- **Modelling**. Process-based modelling (scripts for preparing inputs/outputs and running process-based models); and Statistical modelling (hydrology-related statistical models).

All developed code is organized into CRAN packages (90+).

5.3.2 Python

In technical terms, Python is an object-oriented, high-level programming language with integrated dynamic semantics primarily for web and app development. It is extremely attractive in the field of Rapid Application Development because it offers dynamic typing and dynamic binding options.

Developers can read and translate Python code much easier than other languages. In turn, this reduces the cost of program maintenance and development because it allows teams to work collaboratively without significant language and experience barriers.

The usefulness of Python for data manipulation stems primarily from the large and active ecosystem of third-party packages:

- **NumPy** for manipulation of homogeneous array-based data.
- **Pandas** for manipulation of heterogeneous and labelled data.
- **SciPy** for common scientific computing tasks.
- **Matplotlib** for publication-quality visualizations.
- **Scikit-Learn** for machine learning.
- **TensorFlow** for Deep Learning.
- …

5.3.3 Julia

Julia is a high-level dynamic programming language designed to address the needs of high-performance numerical analysis, and scientific computing is rapidly gaining popularity amongst the data scientists. It is a newer language, also capable of general-purpose programming as well and hasn't been around as long as R or Python.

Due to its faster execution, Julia has become a perfect choice for dealing with complex projects containing high volume data sets. For many basic benchmarks run 30 times quicker than Python and regularly run somewhat quicker than C code. If you like Python's syntax while you have a massive amount of data, then Julia is the next programming language to learn.

A joint effort between Jupyter and Julia communities, it gives a fantastic program based graphical notebook interface to Julia. People, who are searching for the best performance parallel computing language focused on numerical computing, Julia is a perfect language for them.

5.3.4 IBM SPSS

It is the acronym for *Statistical Package for the Social Sciences*. Although originally was designed for social science, nowadays its market includes health sciences, marketing, data miners, etc. It offers advanced statistical analysis, a vast library of machine-learning algorithms, text analysis, open-source extensibility, integration with big data and seamless deployment into applications. Base software provides:

- **Descriptive statistics**: Cross tabulation, Frequencies, Descriptive, Explore, Descriptive Ratio Statistics.
- **Bivariate statistics**: Means, t-test, ANOVA, correlation (bivariate, partial, distances), Non-parametric tests, Bayesian.
- **Prediction for numerical outcomes**: Linear regression
- **Prediction for identifying groups**: Factor analysis, cluster analysis (two-step, K-means, hierarchical), discriminant.
- **Geo spatial analysis, simulation**.

SPSS competes with other (both licenced and open) software, like R language, so IBM developed a free packet called *PSPP*.

As an illustration, mention the publication *"Prediction Of Future Surface Water Quality In Tiruchirappalli District Using SPSS Software"*. In this study, the water quality index is calculated and predicted using a regression model. Climate change has and will continue to have a profound impact on the water sector through the hydrologic cycle, water availability, water demand, and water allocation at the global, regional, basin, and local levels.

5.3.5 MATLAB

MATLAB stands for *MATrix LABoratory*. It allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of user interfaces, and interfacing with programs written in other languages (C, C++, C#, Java, Python…).

Regarding the data science, MATLAB makes data science easy with tools to access and pre-process data, build machine learning and predictive models, and deploy models to enterprise IT systems.

- Access data stored in flat files, databases, data historians, and cloud storage, or connect to live sources such as data acquisition hardware and financial data feeds.
- Manage and clean data using datatypes and pre-processing capabilities for programmatic and interactive data preparation, including apps for ground-truth labelling.
- Explore a wide variety of modelling approaches using machine learning and deep learning apps.

One project developed in MATLAB is analysing and visualizing flows in rivers and lakes. With the aid of Velocity Mapping Toolbox (VMT), the U.S. Geological Survey (USGS) rapidly process all the raw data recorded by ADCPs (acoustic Doppler current profilers). The evolution of VMT included adding more visualization tools for ADCP data. For example, researchers can compare flow velocity data at different depths and strata, map primary and secondary circulation patterns, and plot depth-averaged velocities on aerial maps (see Figure 22).

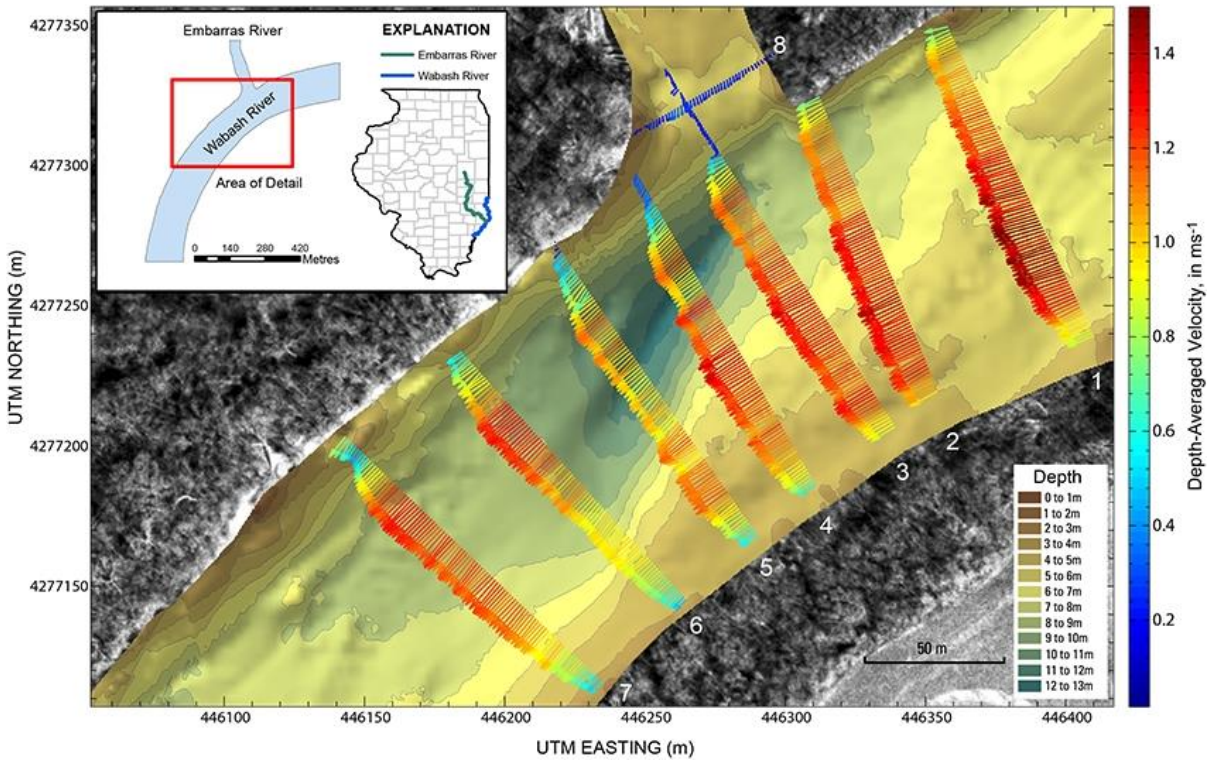


Figure 22. Depth-averaged velocities plotted using VMT on an aerial view of the confluence of the Wabash and Embarras Rivers (Illinois) with ADCP-derived bathymetry.

5.4 Visualization and interactivity

A wide variety of tools are available for interacting and visualizing data models. The following is a brief summary of the most commonly used in the field of numerical models, data science and GIS

5.4.1 Jupyter Notebook

The Jupyter Notebook is an open-source web application that allows you to create and share documents that contain live code, equations, visualizations and narrative text. Uses include: data cleaning and transformation, numerical simulation, statistical modelling, data visualization, machine learning, and much more.

Originally developed for data science applications written in **Ju**lia, **Pyt**hon and **R**, Jupyter Notebook is useful in all kinds of ways for all kinds of projects:

- **Data visualizations.** Most people have their first exposure to Jupyter Notebook by way of a data visualization, a shared notebook that includes a rendering of some data set as a graphic. Jupyter Notebook lets you author visualizations, but also share them and allow interactive changes to the shared code and data set.
- **Code sharing.** Cloud services like GitHub and Pastebin provide ways to share code, but they're largely non-interactive. With a Jupyter Notebook, you can view code, execute it, and display the results directly in your web browser.
- **Live interactions with code.** Jupyter Notebook code is not static; it can be edited and re-run incrementally in real time, with feedback provided directly in the browser. Notebooks can also embed user controls (e.g., sliders or text input fields) that can be used as input sources for code.
- **Documenting code samples.** If you have a piece of code and you want to explain line-by-line how it works, with live feedback all along the way, you could embed it in a Jupyter Notebook. Best of all, the code will remain fully functional—you can add interactivity along with the explanation, showing and telling at the same time.

5.4.2 ArcGIS

ArcGIS is the most used geographic information system (GIS) for working with maps and geographic information. It is used for creating and using maps, compiling geographic data, analysing mapped information, sharing and discovering geographic information, using maps and geographic information in a range of applications, and managing geographic information in a database. The software is provided by *Esri*.

The system provides an infrastructure for making maps and geographic information available throughout an organization, across a community, and openly on the Web.

ArcGIS for Desktop consists of several integrated applications, some of them are:

- **ArcCatalog** is the data management application, used to browse datasets and files on one's computer, database, or other sources. In addition to showing what data is available, ArcCatalog also allows users to preview the data on a map. ArcCatalog also provides the ability to view and manage metadata for spatial datasets.
- **ArcMap** is the application used to view, edit and query geospatial data, and create maps. The ArcMap interface has two main sections, including a table of contents on the left and the data frame(s) which display the map. Items in the table of contents correspond with layers on the map Integrated Platforms.
- **ArcToolbox** contains geoprocessing, data conversion, and analysis tools, along with much of the functionality in ArcInfo. It is also possible to use batch processing with ArcToolbox, for frequently repeated tasks.

5.4.3 GRASS GIS

GRASS stands for Geographic Resources Analysis Support System. It is an open-source alternative to ArcGIS, and used for data management, image processing, graphics production, spatial modelling, and visualization of many types of data.

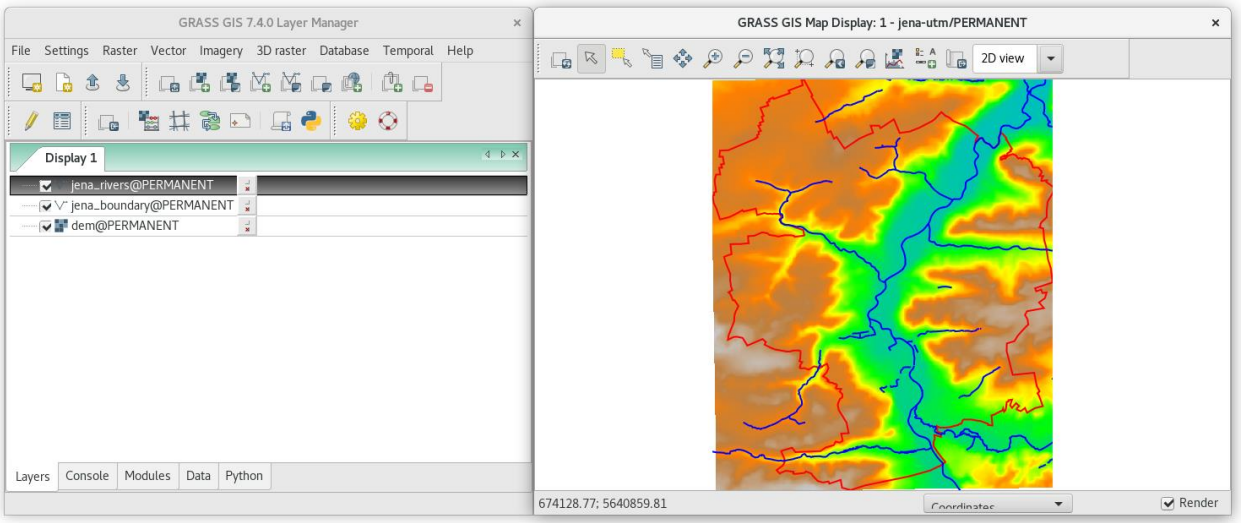


Figure 23. Screenshot of Jena city boundary and rivers in GRASS.

It contains over 350 modules to render maps and images on monitor and paper; manipulate raster, and vector data including vector networks; process multispectral image data; and create, manage, and store spatial data. GRASS GIS offers both an intuitive graphical user interface as well as command line syntax for ease of operations. GRASS GIS can interface with printers, plotters, digitizers, and databases to develop new data as well as manage existing data.

5.4.4 QGIS

QGIS is a professional GIS application that is built on top of and proud to be itself Free and Open Source Software (FOSS). QGIS offers many common GIS functions provided by core features and plugins. General categories include:

- **Data viewing**. User can view combinations of vector and raster data (in 2D or 3D) in different formats and projections without conversion to an internal or common format.
- **Explore data and compose maps**. Compose maps and interactively explore spatial data with a friendly GUI.
- **Create, edit, manage and export** vector and raster layers in several formats.
- **Analyse data**. Performing spatial data analysis on spatial databases and other OGR-supported formats. QGIS currently offers vector analysis, sampling, geoprocessing, geometry and database management tools. You can also use the integrated GRASS GIS tools.
- **Publish maps on the Internet**. QGIS can be used as a WMS, WMTS, WMS-C or WFS and WFS-T client, and as a WMS, WCS or WFS server.
- **Extend QGIS functionality** through plugins.

Figure 24. QGIS map capture of Natural Earth project.

5.4.5 Microsoft Power BI

Power BI is a business analytics service that delivers insights to enable fast, informed decisions.

There are three main targets for Power BI:

- **Analysts**. Connect to and transform data with advanced data preparation capabilities; create interactive data visualizations and uncover important insights; and publish dashboards and share insights.
- **IT**. Reduce your training and support costs by taking advantage of these (Microsoft) familiar tools as part of your enterprise BI deployment. Lower implementation costs and simplify management. Centrally control how data is accessed and used—even on mobile devices. Set

and monitor policies, detect anomalies, and act. Pixel-perfect paginated reporting, enterprise-scale modelling and self-service BI in one modern platform.

- **Developers**. Business intelligence and analytics from Microsoft Power BI can be customized, extended, and embedded in applications using our comprehensive set of APIs and fully documented SDK libraries.

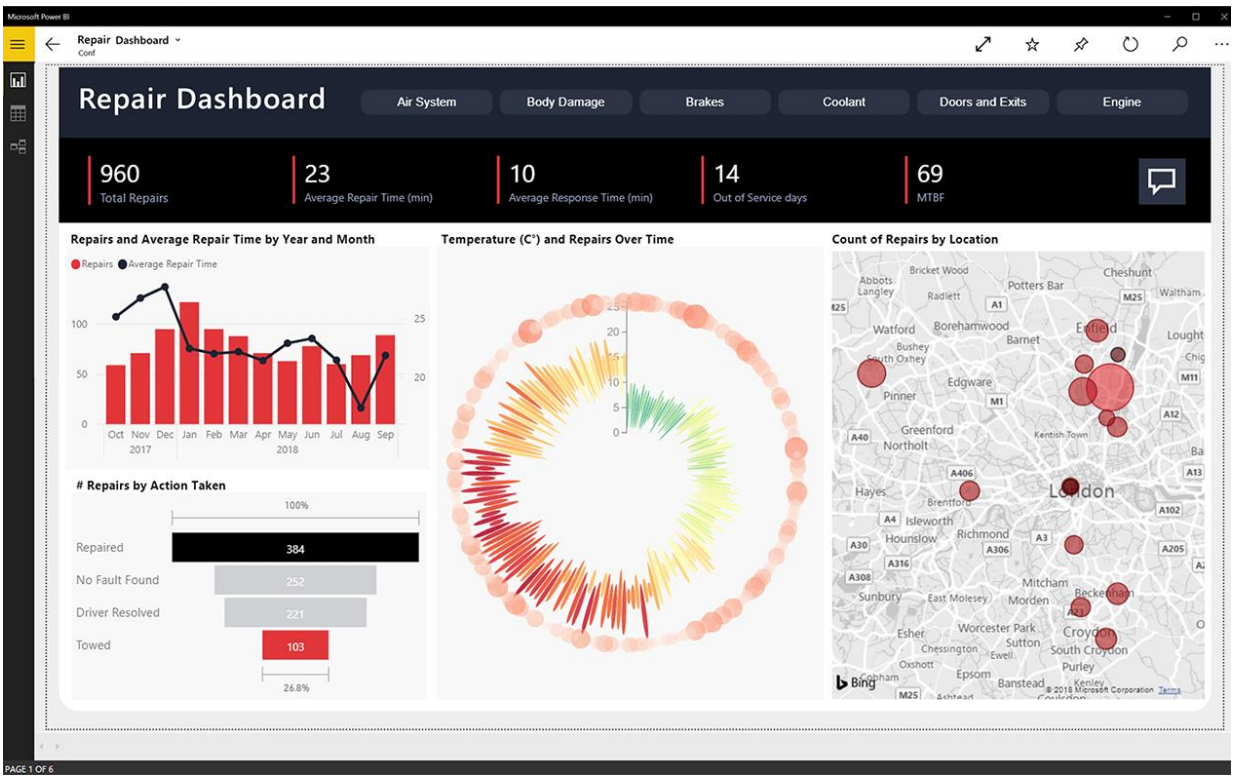


Figure 25. Screenshot of a Power BI dashboard

Key components of the Power BI ecosystem comprise: **Desktop**, **Service**, **Mobile Apps**, **Gateway**, **Embedded**, **Report Server**, and **Visuals Marketplace**.

Taking up the previous section ArcG, the ESRI web page offers the maps for Power BI.

5.5 Integrated Platforms

This section will review most interesting organisms/infrastructures for researching. It is a summary; detailed information can be found in **D8.9** (*Set of computational tools oriented towards HPC and Cloud computing).*

5.5.1 GÉANT

The GÉANT network continues to set the standard for speed, service availability, security and reach, delivering the high performance that more than 50 million users rely on. A separate ultra-high-speed internet, just for research and education. A summary of the services from GÉANT:

- **Internet as a Service (IaaS).** Amazon Web Services (AWS), Microsoft Azure, T-Systems Cloud Service
- **File storage, Synchronization and Collaboration**. Nextcloud, Owncloud, Dropbox.
- **Real-Time Communications**. MVC, Kinly.
- **Connection**. Microsoft Azure ExpressRoutes, Data Egress Charge Waiver for Amazon Web Services
- **Education and e-Learning**. Edu Zone.
- **Software as a Service (SaaS)**. Microsoft Office 365.

Figure 26. Some logos of the services included in GÉANT partnership

5.5.2 EGI (European Grid Infrastructure)

EGI is a series of efforts to provide access to high-throughput computing resources across Europe using grid computing techniques.

The latest project (2018~)by EGI is **EOSC-Hub**, to bring together an extensive group of national and international service providers that will create the *Hub*: a central contact point for European researchers and innovators to discover, access, use and reuse a broad spectrum of resources for advanced data-driven research.

About the enabled services by EGI we could find *(some of them are in beta phase)*:

- **Cloud computing**. It enables to deploy on-demand IT services via standards-based interface onto federate academic and commercial clouds from multiple provider.
- **Cloud Container Compute** gives you the ability to deploy and scale *Docker containers* on-demand.
- **High-Throughput Compute**. Run computational jobs at scale on the EGI infrastructure. It allows you to analyse large datasets and execute thousands of parallel computing tasks.
- **Workload Manager**. Manage and distribute your computing tasks in an efficient way while maximising the usage of computational resources.
- **Online Storage** allows to store data in a reliable and high-quality environment and share it across distributed teams.
- **Archive Storage** allows to store large amounts of data in a secure environment freeing up your usual online storage resources.
- **Data Transfer** allows to move any type of data files asynchronously from one place to another.
- **Check-in** provides a reliable and interoperable AAI solution that can be used as a service for third parties.
- **Applications on Demand** gives access to online applications and application-hosting frameworks for compute-intensive data analysis
- **Notebooks** is a browser-based tool for interactive analysis of data using EGI storage and compute services, by using Jupyter Notebook technology.
- **FitSM Training**. The user will learn, through lessons and examples, the fundamentals of IT service management and how to implement FitSM[9] in your organisation.

[9] FitSM is the name for a family of lightweight standards for IT service management (ITSM).

- **ISO 27001 Training**. It is a standard designed to help organisations keep information assets secure.
- **Training Infrastructure**. It is useful to organise onsite tutorials or workshops and online training courses or as a platform for self-paced learning.

5.5.3 PRACE (Partnership for Advanced Computing in Europe)

The aim of PRACE is to enable high-impact scientific discovery and engineering research and development across all disciplines to enhance European competitiveness for the benefit of society.

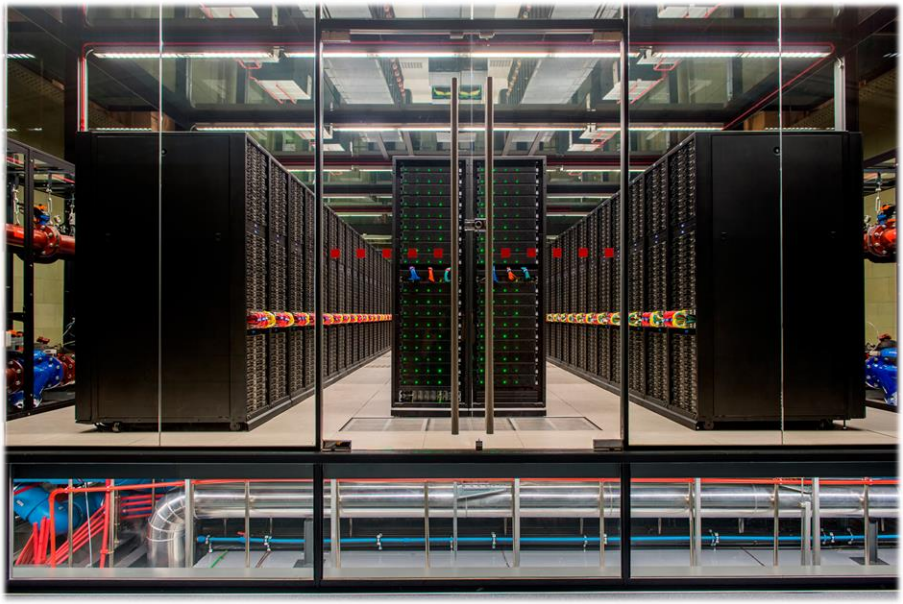


Figure 27. Image of the MareNostrum 4 supercomputer (Spain). This is one of the five hosting members in the PRACE research infrastructure.

5.5.4 HELIX NEBULA

The Helix Nebula Initiative is a partnership between industry, space and science to establish a dynamic ecosystem, benefiting from open cloud services for the seamless integration of science into a business environment.

5.5.5 EUDAT

EUDAT's vision is *Data is shared and preserved across borders and disciplines*. One of EUDAT's main ambitions is to bridge the gap between research infrastructures and e-Infrastructures through an active engagement strategy.

Figure 28. Summary of the EUDAT Services.

5.5.6 INDIGO Data Cloud

INDIGO (INtegrating Distributed data Infrastructures for Global explOitation) is a H2020 project that aims at developing a data and computing platform targeting scientific communities, deployable on multiple hardware and provisioned over hybrid (private or public) e-infrastructures.

5.5.7 LifeWatch

LifeWatch ERIC[10] is a European consortium providing e-Science research facilities to scientists seeking to increase our knowledge and deepen our understanding of Biodiversity organisation and Ecosystem functions and services in order to support civil society in addressing key planetary challenges.

Services provided by this infrastructure are categorized in:

- **Virtual Labs**, also called *VRE* (Virtual Research Environment). Resources and tools to support advanced scientific work in the hot biodiversity research topics.
- **Data portals**, to access environment catalogues.
- **"Actual" Services**. User can find a variety of web services. For example: Alien Species Thesaurus, Genetic Services, or Zooplankton Traits Thesaurus.

[10] European Research Infrastructure Consortium.

6 Conclusions

The present document is divided into four important blocks: data pre-processing operations; the concept of FAIR data; indicators to measure the quality of computing and handling data; and a variety of tools to do this data management.

Going deeper in the FAIR concept, and as discussed in the FAIR Data section (p. 17), there is an urgent need to improve the infrastructure supporting the reuse of scholarly data[11]. A diverse set of stakeholders —representing academia, industry, funding agencies, and scholarly publishers— have come together to design and jointly endorse a concise and measurable set of principles: the FAIR Data Principles. The outcomes from good data management and stewardship are high quality digital publications that facilitate and simplify this ongoing process of discovery, evaluation, and reuse in downstream studies.

From the European Union's Horizon 2020 exist several projects and initiatives to achieve this objective:

- **EUDAT** through its service *B2SHARE* is collaborating in the new guidelines on Data Management Plans (DMP).
- **OpenAIRE** infrastructure supports projects and researchers in complying to the EC's open research data policy and therefore developed supporting material to aid projects to comply with the new guidelines on FAIR data management.
- **GO FAIR** initiative will contribute to and coordinate the coherent development of the Internet of FAIR Data & Services through community-led initiatives. It follows a bottom-up open implementation strategy for the *European Open Science Cloud (EOSC).*
- **FAIRsFAIR** (Fostering Fair Data Practices in Europe) aims to supply practical solutions for the use of the FAIR data principles throughout the research data life cycle. Emphasis is on fostering FAIR data culture and the uptake of good practices in making data FAIR. Using FAIR data with *LifeWatch* (and others).
- **LifeWatch ERIC** is designed to tackle the constraints affecting research activities and the pressing need for increasingly diverse data and larger and more advanced models, open

[11] *"Percentage of time spent finding and organising data according to research data specialists: 79%"* (RDA plenary).

data and open science clouds, making it possible to explore new frontiers in ecological science and support society in addressing the challenges ahead. For example, LifeWatch RI represents the biodiversity and ecosystem community in EUDAT.

Another project is the *Phytoplankton Traits Thesaurus* (from LifeWatch Italy). This is a case of an amount of data that can be found using Metadata included (Semantic Web & Ontology).

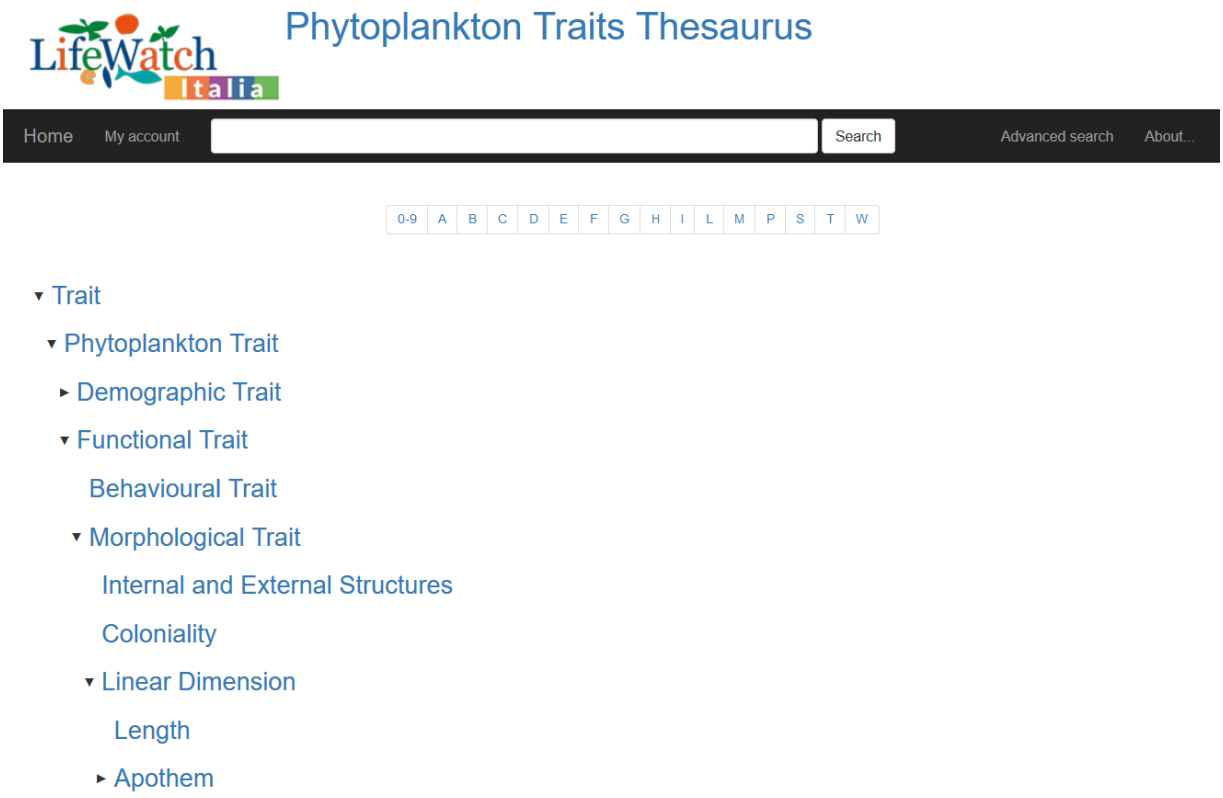


Figure 29. Capture of the Phytoplankton Traits Thesaurus web (LifeWatch Italy)

Figure 30. Some of the collaborating infrastructures using FAIR data principles

References

[1] Data Preprocessing Techniques for Data Mining. From Winter School on *Data Mining Techniques and Tools for Knowledge Discovery in Agricultural Datasets*.

[2] https://towardsdatascience.com/a-brief-overview-of-outlier-detection-techniques-1e0b2c19e561

[3] https://towardsdatascience.com/how-to-handle-missing-data-8646b18db0d4

[4] https://medium.com/@swethalakshmanan14/how-when-and-why-should-you-normalize-standardize-rescale-your-data-3f083def38ff

[5] https://subscription.packtpub.com/book/big_data_and_business_intelligence/9781783982103/1/ch01lvl1sec21/data-transformation-and-discretization

[6] Dra. Eva Méndez, “Cool” metadata for FAIR data. FAIR Data Management, 14-15 November 2016 (Florence, Italy).

[7] https://orcid.org/blog/2018/08/08/building-robust-research-infrastructure-one-pid-time

[8] *Data Fairness International Summer School*. 1-5 July 2019, Lecce (Italy).

[9] https://www.nature.com/articles/sdata201618

[10] https://www.openaire.eu/how-to-make-your-data-fair

[11] https://www.go-fair.org/fair-principles/

[12] https://publications.europa.eu/en/publication-detail/-/publication/7769a148-f1f6-11e8-9982-01aa75ed71a1/language-en/format-PDF/source-80611283

[13] https://www.w3.org/standards/semanticweb/

[14] https://taverna.incubator.apache.org/introduction/

[15] https://galaxyproject.github.io/

[16] https://kepler-project.org/

[17] https://ckan.org/

[18] https://docs.depositar.io/en/6.3.5/user-guide.html#

[19] http://blog.dasheroo.com/kpis-vs-metrics-know-difference/

[20] https://www.linkedin.com/pulse/real-difference-between-metrics-kpis-richard-hatheway/

[21] https://www.e-nor.com/blog/general/kpis-vs-metrics

[22] https://www.doi.gov/sites/doi.gov/files/migrated/news/upload/Hurricane-Sandy-project-metrics-report.pdf

[23] https://www.hanson-sustainability.co.uk/sites/default/files/assets/document/49/3a/summary-of-kpi-performance-against-2020-targets_2.pdf

[24] https://www.arcgis.com/index.html

[25] https://opensource.com/alternatives/arcgis
[26] https://www.qgis.org/en/site/
[27] https://rd-alliance.org/metadata-principles-and-their-use.html
[28] https://velos.com/wp-content/uploads/Strategies-for-Defining-Key-Performance-Indicators-in-Research.pdf
[29] https://bscdesigner.com/innovation-kpis.htm
[30] https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=7921551
[31] https://www.excelcampus.com/vba/macros-explained-part-1/
[32] https://www.openoffice.org/product/calc.html
[33] https://venturebeat.com/2016/09/29/google-updates-calendar-drive-docs-sheets-and-slides-with-machine-intelligence-features/
[34] https://www.ibm.com/uk-en/analytics/spss-statistics-software
[35] Velusamy, Sudha & M Praveena, Ms & Varma, S. Anand & N Ilavarasan, Mr. (2018). PREDICTION OF FUTURE SURFACE WATER QUALITY IN TIRUCHIRAPPALLI DISTRICT USING SPSS SOFTWARE. International Journal of Research. 7. 1229-1235.
[36] https://uk.mathworks.com/company/newsletters/articles/analyzing-and-visualizing-flows-in-rivers-and-lakes-with-matlab.html
[37] http://thesauri.lifewatchitaly.eu/PhytoTraits/index.php
[38] https://en.wikipedia.org

DANUBIUS-PP

Preparatory Phase for the pan-European Research Infrastructure DANUBIUS–RI "The International Centre for advanced studies on river-sea systems"

European Commission

This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreement No 739562