Data Literacy Framework for upper primary education

WP D2.1

Co-funded by the European Union

Disclaimer
Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or granting authority. Neither the European Union nor the granting authority can be held responsible for them.

Work Package 2:

Conceptual design and teaching/training resources

Deliverable D2.1

Data literacy framework for upper primary education

Co-funded by the European Union

Project Title:	Data Literacy for Upper Primary Schools
Project Acronym:	DALI4US
Project Number	101132912
Grant Agreement Number	101132912 — DALI4US — ERASMUS-EDU-2023-PI-FORWARD
Instrument:	Erasmus+
Deliverable number	D2.1
Work Package	2 Conceptual design and teaching/training resources
Work Package Leader	Service de Coordination de la Recherche et de l'Innovation pédagogiques et technologiques
Work Package Partners	Participating partners: 1. Service de Coordination de la Recherche et de l'Innovation pédagogiques et technologiques, SCRIPT, LU 2. Ministry of Education Slovenia, MVI, SI 3. Dublin West Education Centre, PDST, IR 4. H2 Learning, H2L, IR 5. University of Ljubljana, SI 6. Associated partner: 7. Institut de Formation de l'Éducation nationale, IFEN, Lux
Dissemination level	Confidential Internal Document
Delivery date	2024-06
Status	Final Version
Version	3.0
Author(s)	WP2 partner group: Claude Reuter, Vanessa Hoff, Valerie Peffer
List of contributors	All partners
Deliverable Manager	Claude Reuter, Vanessa Hoff, Valerie Peffer

Co-funded by the European Union

ABSTRACT	The Dissemination and Sustainability Strategy frames the communication and dissemination activities to raise the visibility of the DALI4US project, to engage project stakeholders and to share the project results by defining target groups, communication approaches, promotional materials, communication and dissemination channels. It includes sustainability strategy to ensure sustainability, mainstreaming and multiplication of the project results.
KEYWORDS	Communication, Dissemination, Sustainability, Promotion

Dissemination level		
PU	**Public**	
PP	**Restricted to project partner (including the Commission)**	**X**
RE	**Restricted to a group defined by the consortium (including the Commission)**	
CO	**Confidential, only for members of the consortium (including the Commission)**	

Co-funded by the European Union

Table of Content

Co-funded by the European Union

1. Introduction

In the contemporary era, data is a fundamental component of our daily lives. This connection is rooted in the digital domain, where data is collected, processed, and analyzed, forming the backbone of many technological advancements and everyday applications. It has become an indispensable part of our personal, social, and economic lives. As Robert Burgee (2018) observes, it is becoming increasingly challenging to avoid the digital domain. At the present time, data represents a significant factor in the development of entire industries and is therefore of considerable value[1]. The designation "new oil" (Wikström, 2022) appears to be a self-evident assertion at first glance. However, in contrast to oil, data is not a finite resource; it is reusable and, most importantly, its value is not consistent across all contexts. The value of data is contingent upon the observer's perspective and the specific requirements and context in question. In order to ascertain the intrinsic value of data, it is necessary to refine it in a manner analogous to that employed in the refining of crude oil. The mere presence of unstructured data does not, in itself, facilitate the generation of insights. It is estimated that 80% of data is produced in an unstructured form on a daily basis (Deep Talk, 2021). A series of processes is required in order to achieve the desired outcome, including processing, structuring, filtering, analysis and evaluation. In the absence of such data processes, profitability is unfeasible (Wikström, 2022). Considering the vast quantity of data, the data processes are predominantly conducted on comprehensive digital platforms, which are automated by AI systems and self-learning algorithms. Such systems utilise the raw data to generate representations and models, which in turn serve as a basis for decision-making processes.

The utilisation of AI systems in data-related processes is, at first, challenging for many individuals to comprehend, and the sheer volume of data appears overwhelming (big data). Nevertheless, the concept of big data is not merely a transient phenomenon; according to experts, it is an early indicator of a profound social change that will affect virtually every individual. It is therefore evident that a basic understanding of data management is a prerequisite for maintaining democracy. This understanding will empower citizens to critically evaluate the flow of information that shapes public opinion and policy. Effective data management enables transparency, accountability, and informed decision-making, essential for upholding democratic values. Without it, misinformation and misuse of data could undermine trust in institutions and erode the foundations of civic engagement. Thus, fostering data literacy ensures that individuals can actively participate in democratic processes while safeguarding their rights in a data-driven society. Nevertheless, an examination of digitisation processes, and in particular data analysis processes, should not be approached from a technological standpoint alone. The concepts of "digital self-defense," "echo chambers," "digital sovereignty," and "digital inequality" illustrate the potential risks associated with digitization processes for society. John Morow (2021) posits that a conservative stance, such as absolute abstinence from the disclosure of personal data or complete non-participation in data collection services, is untenable in a data-driven society. These conditions are insufficient for the exercise of sovereign social participation in the context of the digital age. In a democratised society it is imperative

[1] https://www2.deloitte.com/us/en/pages/real-estate/articles/future-real-estate-data-new-gold.html (last access: 13.01.2025)

that all citizens are afforded equal access to data, and that they are able to utilise it in a safe, critical and creative manner. It is therefore vital that they are equipped with the requisite skills to optimise their data utilisation. These required competencies extend beyond the scope of media education and data literacy, encompassing the capacity to 'critically read and understand statistical information in the form of tables, graphs and statistics on important social phenomena in order to make informed decisions in private and public life' (Engel et al., 2019, S. 215). Considering the growing prevalence of misinformation, propaganda and polarising rhetoric, it is imperative to promote the capacity to make well-informed and independent decisions. This implies the ability to critically evaluate data-based assertions and statistical arguments that are often presented as irrefutable truths. Such an approach allows the recognition of potentially misleading information and the formation of an informed opinion (Schiller, 2020).

This document illustrates the interdisciplinary nature of data literacy and argues for a comprehensive basic education that is accessible for all school subjects in a contextual framework and should start at an early age.

2. Needs analysis

DALI4US is a project grounded in the belief that data literacy should be as fundamental as reading and writing, due to the pervasiveness of data in modern life. This section will set out the project requirements for the DALI4US project. The objective is to equip primary school teachers with the skills and knowledge needed to teach data literacy. A comparative analysis of the curricula in Luxembourg, Ireland, and Slovenia reveals that the focus is on the teaching fundamental data literacy skills.

The following section presents potential approaches that could serve as a foundation for the development, implementation, and evaluation of the DALI4US project. This will satisfy the identified needs.

2.1 Curriculum alignment

The curricula of upper primary school classes in the partner countries already place significant emphasis on the development of basic data literacy skills, including the formulation of statistical questions, the collection and analysis of data, and the interpretation of results.

2.2 Teacher training and support

It is evident that there is a pressing need for professional learning opportunities for teachers. The teaching of data literacy presents a number of challenges for teachers, who are often inadequately prepared for this task. In order to ensure that teachers are adequately equipped to teach this subject, it is essential that they receive specific training and access to the requisite resources, which will enable them to gain the necessary knowledge and technical and pedagogical skills. Many educators lack preparation in this area, facing challenges in integrating data concepts into their instruction. Training must align with the Technological Pedagogical Content Knowledge (TPACK) framework from Mishra Punia & Koehler

Co-funded by the European Union

Matthew (2007), emphasizing the interplay of content knowledge (data concepts and ethics), pedagogical knowledge (effective strategies for teaching data literacy), and technological knowledge (proficiency in tools like spreadsheets and visualization software). Teachers need support to design lessons that make data literacy accessible and engaging, ensuring students can critically engage with data. Professional development should provide resources, ongoing guidance, and opportunities for collaboration to build this expertise. By empowering teachers, students will gain the skills to thrive in a data-driven world.

The creation of a common framework enables the design of such training programmes, the creation of resources as well as the development/adaptation of tools (ie Orange).

2.3 Technology integration

Research has shown that there is a necessity for the implementation of appropriate technological tools that can be integrated into the educational process. The needs analysis demonstrates the importance of providing user-friendly digital tools, such as OrangeEDU, for educators and learners, in order to facilitate hands-on learning and the process of exploring data through visual analysis.

2.4 Leadership engagement

It is imperative that those in leadership positions are engaged in multiple facets to ensure the successful implementation of the proposed strategy. This includes providing clear vision and direction, aligning organizational goals with the strategy, and communicating its importance to all stakeholders. They must prioritize resource allocation, ensuring that adequate funding, tools, and training opportunities are available to support the initiative. Additionally, leaders should engage in monitoring progress through regular evaluations and feedback mechanisms to identify and address potential challenges early. By fostering a culture of collaboration and commitment, leaders can inspire collective action and accountability, creating an environment conducive to achieving the strategy's objectives. Their active involvement ensures sustained focus and momentum throughout the implementation process. The needs analysis indicates that school leaders have a pivotal role in influencing the school culture with regard to data literacy.

2.5 Experimentation with frameworks

The conducted needs analysis conducted highlights the relevance of a systematic, iterative approach to designing the data literacy framework. This approach involves cycles of experimentation, feedback and refinement, ensuring that the framework is adaptable and effective in different educational contexts.

2.6 Development of learning activities

The analysis suggests that the development of curriculum-based, project-based learning activities is needed to make data literacy relevant and accessible to students. These activities should emphasise real-world applications and help students engage with data in meaningful ways.

Co-funded by the European Union

3. About Data Literacy

The transition to Big Data is a continuation of humanity's ongoing search for methods to measure, record and analyse the world (Mayer-Schönberger & Cukier, 2013). The combination of big data, artificial intelligence and self-learning algorithms leads to the generation of new knowledge (or new power) that would otherwise be inaccessible to humans. Data that was previously unknown is being used in ways that were not previously intended or expected. Datafication takes place at two different levels: firstly, through direct interaction with digital devices, and secondly, through the unintentional and unconscious generation of data through the use of digital applications such as the internet and social networks (Seemann, 2014). Furthermore, the phenomenon is not limited to adults, but also includes adolescents and children, who are also involved in the production of data, whether consciously or unconsciously. This has led to extensive research into the concept of data literacy. However, there is no agreement on a standardised definition, as these skills are multifaceted, embedded in different scientific fields (e.g. statistics, computer science, coding initiatives and artificial intelligence) and emerge from different empirical contexts. Koltay (2015) and Luci Pangrazio and Neil Selwyn (2018) emphasise that a standardised definition of data literacy does not yet exist, and that the concept is subject to change.

Within the sciences and computing, academic debates define data literacy as a highly technical skill (Sapp Nelson et al., 2011). Pekka Mertala (2020) postulates that data literacy is a transversal skill that all citizens of a data society should possess, not only experts. F. Javier Prado and Miguel Marzal (2013) define data literacy as the ability to find and manage data. The Data-Pop Alliance defines data literacy as “the desire and ability to use data to constructively engage with society” (Bhargava, 2015). Ellen Mandinach and Edith Gummer (2013) define data literacy as the ability to understand and use data in a way that enables informed decision-making. Other attempts to define data literacy refer to it as a willingness to engage with the data-driven systems that characterise contemporary society. These developments have led to the emergence of several sub-competences in the area of data literacy. These include, but are not limited to, data management literacy, critical data literacy, research data literacy, creative data literacy, scientific data literacy, data information literacy and educational data literacy. In his research, Gulsen Guler (2019) addressed the question of which data should be considered within the scope of data literacy. Some definitions focus on the “online aspect of data” (also in the context of big data), while others define data literacy according to the type of data, namely “quantitative and qualitative”.

The definition of data literacy depends consequently on a number of factors, including the understanding of the data, the context and the audience. However, there are also definitions that take a broader approach and can therefore provide a more robust basis because they are more general in nature. Bhargava and D'Ignazio (2015) postulate that data literacy comprises four sub-competencies in the context of data: These sub-competencies can be summarised as follows:

1. read data: involves understanding what data is, and what aspects of the world it represent
2. work with data: involves creating, acquiring, cleaning, and managing it

3. analysing data: involves filtering, sorting, aggregating, comparing, and performing other analytic operations on it
4. arguing with data: involves using data to support larger narratives intended to communicate a message to a particular audience

Annika Wolff et al. (2016) extend the definition to include the sub-competences of communicating stories derived from data and using data as an integral part of the design process. The definition includes the following sub-competencies: the ability to select, clean, analyse, visualise, critique and interpret data; the ability to communicate stories derived from data; and the use of data as an integral part of the design process. (ibid.)

According to Leo Van Audenhove et al. (2020), the existing literature on data literacy focuses primarily on the skills and competencies required to handle, analyse and use data. This corresponds to an instrumental and utilitarian perspective on data literacy. However, they postulate that the increasing relevance of data in society points to the need for a data literacy concept. They define data literacy as a concept, distinct from other forms of literacy such as media, multi- or digital literacy. It encompasses more than the ability to access and use data, analyse it, produce statistics or understand data visualisations. In terms of critical literacy, Pangrazio and Selwyn define data literacy as the need to better support individuals to critically engage with their personal data so that they develop a sense of understanding, control and agency within the dataset.

In 2015, Chantel Ridsdale et al. (2015) conducted a comprehensive analysis of 32 definitions of data literacy. This work provides an almost comprehensive overview of the field. The definition of data literacy as the ability to collect, manage, evaluate and apply data in a critical way is adopted by the authors mentioned. According to the authors, the term “data literacy” can be defined as the ability to collect, manage, evaluate and apply data in a critical way. This definition goes beyond the technical and value-free connotations of data and emphasises the critical aspect. A critical approach to data education aims to raise awareness of the social injustices associated with datafication and to foster students' ability to challenge and question dominant ideologies, beliefs and practices (Pangrazio & Selwyn, 2021).

Jonathan Gray et al (2018) integrate both the instrumental and technical components, as well as the critical and ethical aspects of data literacy in their definition. The authors define data literacy as a combination of information literacy, statistical literacy and technical skills.

Co-funded by
the European Union

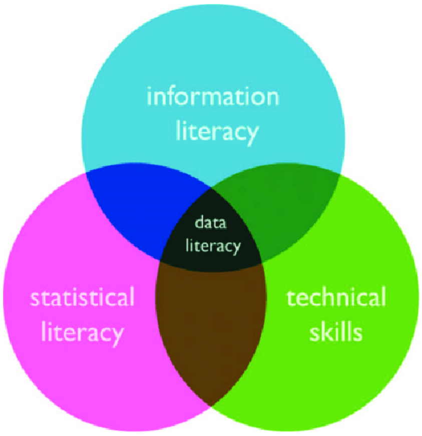

Figure 1: "What is data literacy?" (Source: http://www.undatarevolution.org/datause-availability/)

"Information literacy" is the ability to find, evaluate and use information effectively and critically. "Statistical literacy" involves the accurate understanding and interpretation of statistical data. "Digital literacy" is the mastery of tools and technologies for managing, analysing and visualising data.

4. Data Literacy framework for upper primary

To address the need for a comprehensive approach to data literacy, the DALI4US project focuses on the following areas: curriculum adaptation, teacher training, technology integration, school leadership engagement, and framework testing. These efforts can be based on the concept of data literacy as defined by the United Nations Data Revolution (Gray et al., 2018), which includes three key areas: information literacy (the ability to find and use data), statistical literacy (the ability to interpret data correctly), and technical literacy (the ability to manage data with appropriate tools). Integrating these skills into curricula, providing professional development and using digital platforms will facilitate learning and practical application.

Data literacy is identified in HFD working paper 37/18 as a "central competence for digitisation and the global knowledge society in all sectors and disciplines" (Heidrich et al., 2018). However, while frameworks such as DALI and Data Literacy Education exist for teaching data literacy and machine learning at the post-secondary level, there is a notable gap in the availability of such resources at the primary level. The aim of this project is to address this gap by developing a framework specifically designed for upper primary students. Integrating data literacy at this early stage will not only raise awareness, but also provide students and teachers with the necessary technological and pedagogical tools to develop critical thinking and confidence with data.

Integrating data literacy into the primary school curriculum allows for cross-curricular learning. This approach makes it easier for students to understand practical applications of data and to relate theoretical concepts to real-life scenarios. Initial exposure to informal data-related activities helps students to appreciate the practical value of mathematical concepts before they are formally introduced. As Katie

Co-funded by the European Union

Makar (2018) suggests, learning in problem-based, informal contexts allows students to grasp the value of data-driven insights organically and intuitively, which can increase their motivation and engagement with the subject.

However, the current approach to data literacy often focuses on the acquisition of technical skills, such as the ability to handle or analyse data, without fully considering the broader implications. Van Audenhove et al (2020) argue that data literacy should not only be understood in utilitarian terms but should also empower students to think critically about data and technologies that shape society. This broader understanding of data literacy is in line with the perspective of Heidrich et al. (2018), who describe data literacy as a “core skill for digitalisation and the global knowledge economy across all sectors and disciplines”.

Teaching data literacy at primary level should therefore not be limited to developing technical skills. It is equally important to foster critical thinking and the ability to question societal structures and dynamics. Developing both practical and critical skills enables students to become informed and engaged citizens, able to understand and use data to shape the world around them. This dual focus is particularly important in preparing students for a future in which artificial intelligence (AI) and data-driven technologies will play an increasingly central role. As Long and Magerko (2020) note, data literacy is inextricably linked to AI literacy because of the intrinsic relationship between computational thinking and machine intelligence. Understanding data is essential for grasping how AI systems function, making the development of both data and AI literacy critical for equipping students with the skills needed to navigate and shape a future dominated by AI-driven technologies. Data literacy is most effectively developed if the process is started early, in primary school, and the skills are then gradually improved throughout primary and secondary school. The absence of basic data literacy education represents a critical gap in preparing students to be informed, capable citizens.

4.1 PPDAC Cycle

The PPDAC Data Problem Solving Cycle[2] is a widely recognised framework for statistical literacy and is highly relevant for guiding learners in the ethical use of data to address real-world challenges. The PPDAC framework was originally developed to describe the stages of problem solving using numerical evidence, where the data is typically obtained directly from the learners themselves. However, its scope has been extended to cover scenarios where learners work with existing public datasets, data collected by sensors, and a range of analytical techniques, including machine learning algorithms and traditional statistical methods.

[2] https://dataschools.education/about-data-literacy/ppdac-plan-stage/ (last access: 14.10.2024)

Co-funded by the European Union

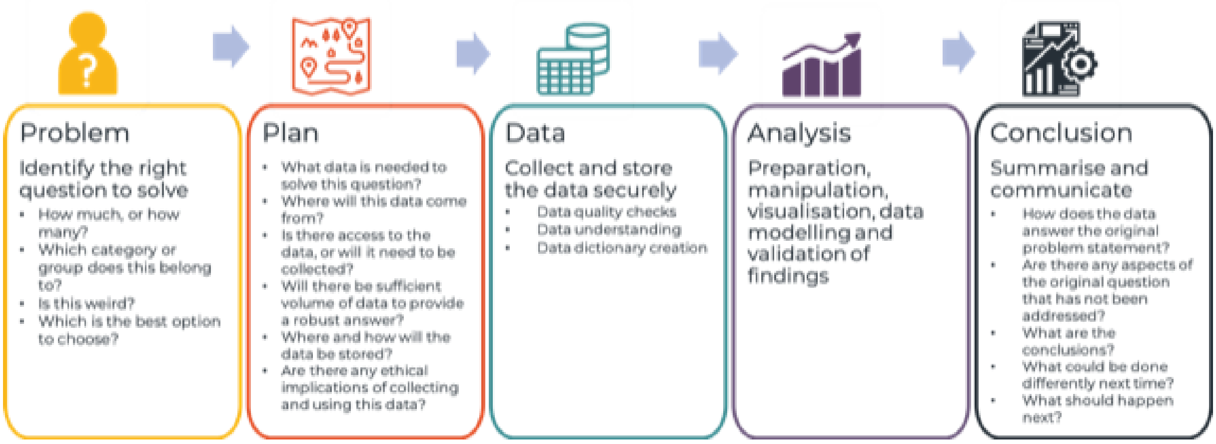

Figure 2:PPDAC cycle (Source:https://dataschools.education/about-data-literacy/ppdac-the-data-problem-solving-cycle/)

4.1.1 Problem

The first step is to identify and define the problem to be addressed. This includes formulating a statistical research question that will guide the entire investigation.

4.1.2 Plan

The next stage is the planning stage. In this phase a plan is drawn up for collecting the data needed to answer the research question. This includes selecting appropriate data sources, determining appropriate data collection methods, and selecting the tools and technologies to be used.

4.1.3 Data

In this phase, data is collected according to the plan established earlier. This may involve collecting new data through surveys or experiments, or using pre-existing data sets originally collected for other purposes.

4.1.4 Analysis

Once the data have been collected, they are analysed to extract valuable insights. This phase involves using statistical methods and tools to interpret the data, identify patterns and draw meaningful conclusions.

4.1.5 Conclusion

The final stage is to evaluate the results of the analysis and draw conclusions. This involves reflecting on the original research question, evaluating the results in relation to the research question, and considering the wider implications of the results.

Co-funded by the European Union

By following the PPDAC cycle, learners can approach data-driven challenges in a systematic and ethical way. This method not only improves their statistical skills, but also enables them to navigate the complexities of working with data in the modern world.

4.2 Gaise II Framework

The GAISE II framework (Guidelines for Assessment and Instruction in Statistics Education)[3] provides a systematic methodology for developing statistical literacy and problem-solving skills in students from pre-school to grade 12.

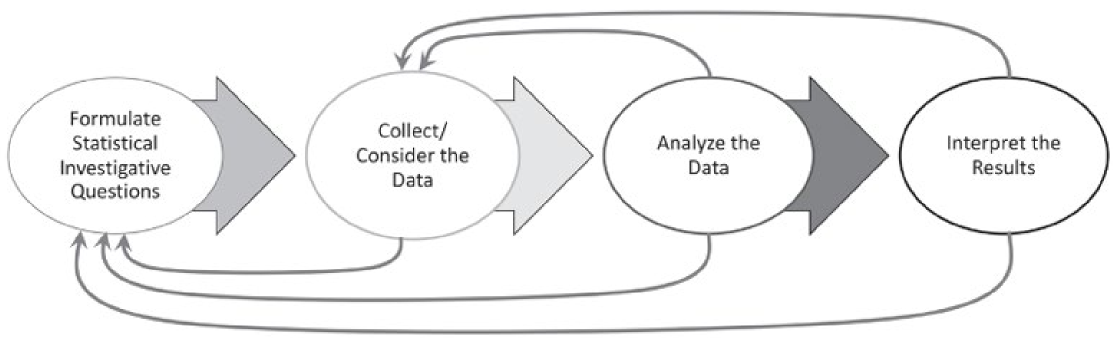

Figure 3: GAISE II Framework (Source: https://www.nctm.org/uploadedFiles/GAISEIIPreK-12_Full.pdf)

The GAISE II framework advocates a systematic approach to problem solving that includes formulating research questions, collecting or considering data, analysing data and interpreting results. This process is crucial to developing students' understanding and ability to work with data.

Thie first level A is for younger students (e.g. in lower secondary school) and focuses on the development of basic statistical skills. The process includes the following steps

4.2.1 Formulate statistical investigative questions

The formulation of statistical questions is a fundamental aspect of statistical literacy. It is important to understand when and how to formulate statistical research questions that are relevant and manageable for students, often using data they have collected or have access to.

4.2.2 Collect/Consider data

The next step is to collect and analyse the data. The ability to collect data through surveys, experiments or examination of existing data sets is developed, as is an understanding of the different types of variables (categorical or quantitative).

4.2.3 Analyze the data

[3] https://www.nctm.org/Standards-and-Positions/GAISE-II/ (last access: 14.10.2024)

Co-funded by the European Union

This stage involves the analysis of the data. Appropriate graphical representations such as tables, scatterplots and bar charts are used to visualise the data. Students are introduced to basic statistical concepts such as mean, median, range and variability.

4.2.4 Interpret the results

The final stage of the process is to interpret the results. Students reflect on their results, evaluate them in the context of their initial question and learn about the limitations of generalising conclusions beyond the data set. The interpretation of results is an essential aspect of data analysis. At this stage, students begin to extrapolate their results beyond the immediate sample, taking into account uncertainty and variability in their conclusions.

Level B is for older students (intermediate to advanced) and builds on Level A. It involves formulating more complex questions, advanced concepts and more sophisticated analytical tools in data analysis and visualisation.

The focus of the learning process should be on inquiry. GAISE II advocates an inquiry-based methodology for teaching statistics, in which students take on the role of active participants in the learning process. They explore data, ask questions and draw conclusions, guided by teachers in a structured way.

The framework emphasises the value of using authentic data to increase the relevance of learning. Pupils are encouraged to work with data from a variety of sources, including public datasets, and to apply statistical concepts to address practical challenges.

GAISE II recognises the importance of technology in the teaching and learning of statistics. It advocates the use of tools and software to enhance data analysis and visualisation in order to provide students with more effective methods for understanding and interpreting data.

4.3 DALI4US Framework

The DALI4US Data Literacy Framework focuses on developing critical thinking skills and problem-solving strategies among upper primary school students, demonstrating how data literacy intersects with the core curriculum and various subjects. The framework serves as the foundation for all subsequent developments and integrations within the project. It is based on the principles of the GAISE II framework and the PPDAC cycle (Annex A illustrates, for each of the four stages, a description of the corresponding competences to be developed). It provides a structured approach to enable students to effectively collect, analyse and transform data into applicable knowledge. The framework is closely aligned with international standards such as the DIGCOMP framework (Vuorikari et al., 2022), which emphasises key skills such as filtering, searching, analysing, storing, managing and manipulating data.

Co-funded by the European Union

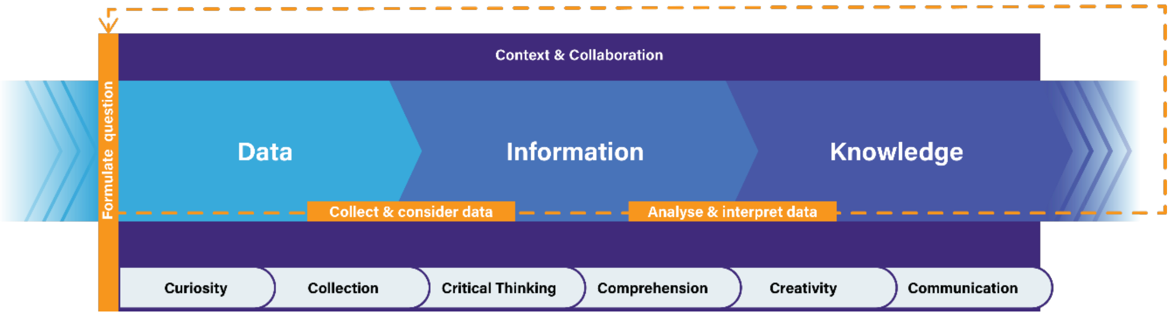

Figure 4: DALI4US Data Literacy Framework (own representation)

The DALI4US framework also places a strong emphasis on the ability to communicate, reason, solve problems and put data into context. Critical thinking is also emphasised throughout.

4.3.1 Data → Information → Knowledge: The continuum of understanding

The DALI4US framework is based on Nathan Shedroff's concept of the continuum of understanding, also visualised in the DIKW pyramid[4] (Data, Information, Knowledge, Wisdom).

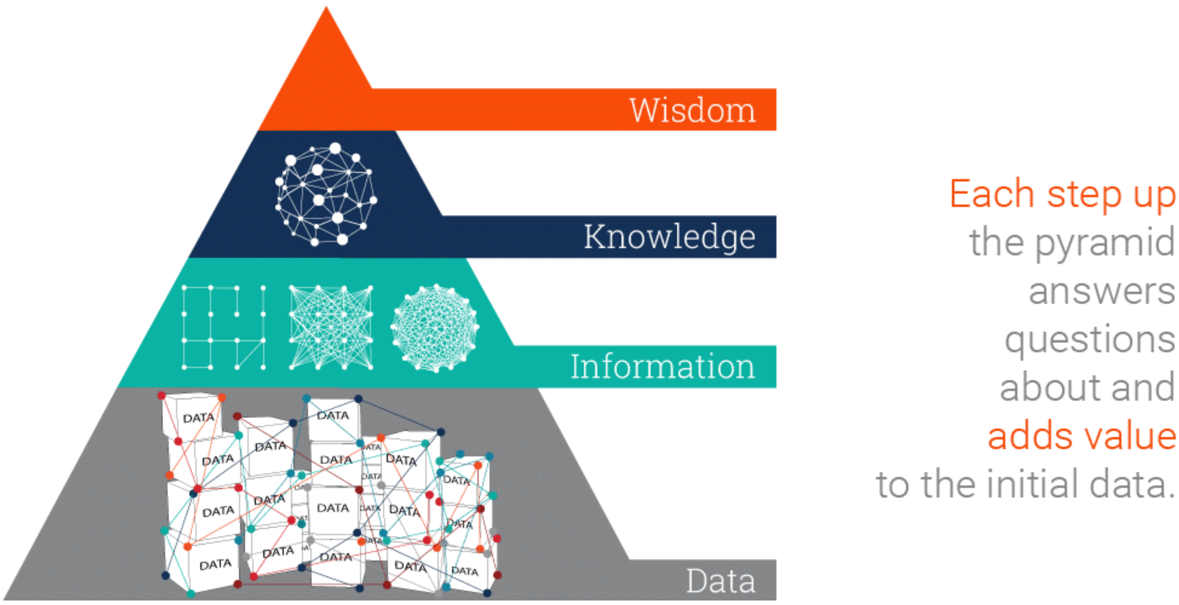

Figure 5: DIKW Pyramid (Source: https://www.ontotext.com/knowledgehub/fundamentals/dikw-pyramid/)

The process begins with the collection of data, which are considered hard facts or figures. This data is then analysed and placed in a meaningful context to generate information. Information is interpreted data that has been given meaning through analysis. When this information is processed and applied to specific situations, knowledge is created. Knowledge enables informed decision making and action. The pyramid

[4] https://www.ontotext.com/knowledgehub/fundamentals/dikw-pyramid/ (last access: 14.10.2024)

Co-funded by the European Union

ends with wisdom, which is the result of extensive knowledge and experience and provides deep insight into the nature of problems and their solutions.

In essence, the model describes the ascent from raw data to a high level of understanding and judgement. Data provides the foundation, while knowledge is created through the application of information. This developmental process illustrates how understanding occurs in stages, with each stage enabling a deeper and more complex understanding of the world.

4.3.2 Processes and steps

The process of data analysis is central in the modern world because it allows valuable information and knowledge to be extracted from a wide range of data sources. The PPDAC cycle (Problem, Plan, Data, Analysis, Conclusion) provides a well-established structure for this process, while the GAISE II framework (Guidelines for Assessment and Instruction in Statistics Education) provides clear guidelines for conducting statistical investigations accurately and effectively. Both models are closely linked and provide a structured approach that can be divided into four essential steps: formulating the question, collecting and examining data, analysing the data and interpreting it. Each of these steps is critical to successful data analysis and ensures that data is transformed into valuable insights and knowledge.

1. Formulate question

The first step in any data-driven analysis process is to formulate questions. This step is embedded in both the PPDAC cycle and the GAISE II framework. In the PPDAC cycle, this step is addressed in the "Problem" and "Plan" phases, while in the GAISE II framework it is referred to as "Formulate Statistical Investigative Questions". Formulating the right questions is the starting point of the whole process as it sets the direction of the investigation and clarifies what information is needed to solve a specific problem or answer a question.

When formulating questions, it is important to be curious and to develop a deep understanding of the problem. Questions such as "what do I want to find out" and "what data do I need to answer these questions" drive the process. This curiosity helps in the search for information that can be analysed later. The GAISE II framework emphasises that the ability to formulate relevant and precise questions is crucial to sharpening the focus of the analysis. Without clear questions it is easy to get lost in irrelevant or inaccurate data.

2. Collect and consider data

Once the questions have been clearly formulated, the data collection and analysis phase follows. This step corresponds to the "Data" phase in the PPDAC cycle and the "Collect and consider data" phase in the GAISE II framework. The aim of this phase is to collect relevant data that will help to answer the statistical research questions.

The type of data collected will depend on the research question and may include quantitative or qualitative data. It is important to consider different types of data and variables. For example, data may be

Co-funded by the European Union

collected through surveys, interviews, experiments or observations. The GAISE II Framework emphasises the need for careful planning at this stage to ensure that the data collected are reliable and representative. Only accurate and well-designed data collection can provide the basis for meaningful analysis.

Another important aspect of this stage is multivariate thinking. This means that when planning and collecting data, several variables should be considered at the same time in order to understand complex relationships and interactions between variables. This type of thinking allows for a completer and more accurate picture of the issue under investigation.

3. Analyse and interpret data

After data collection, the data analysis and interpretation phase follow. This is the 'analysis' phase in the PPDAC cycle and the "analyse the data" phase in the GAISE II framework. In this phase, the collected data is systematically analysed to identify patterns, test hypotheses and find answers to the questions posed.

An essential part of data analysis is probabilistic thinking. This means that uncertainty and randomness are always taken into account when analysing data, as most real-world data sets are characterised by randomness and variability. Quantifying randomness is a key concept in understanding the extent to which the patterns observed in the data are random or systematic.

Modern technology also plays a crucial role at this stage. The use of tools such as OrangeEDU enables efficient data analysis by processing and visualising large amounts of data quickly and accurately. These tools help to identify patterns and trends in the data that might otherwise be missed. They also enable multivariate analysis, which is particularly useful when looking at multiple variables at the same time.

Once the data has been analysed, it needs to be interpreted. This involves putting the results of the analysis into context and understanding their relevance to the original research question. In the PPDAC cycle this corresponds to the "Conclusion" phase, while in the GAISE II framework this phase is called 'Interpret the results'. In this phase, the statistical results are communicated in a clear and understandable way. It is not only about what the data show, but also how they are to be interpreted and what conclusions can be drawn from them to gain knowledge.

Interpreting data often requires clear communication of statistical information. Data are used to tell a story, and that story needs to be told in a concise and understandable way. In many cases, data is a model of reality that is used to explain or predict complex phenomena. The ability to draw the right conclusions from the data and to communicate them in an understandable way is crucial for the application of the knowledge gained in practice.

4.3.3 Context and collaboration

Another key element of the DALI4US framework is context. Data does not exist in a vacuum; it only has meaning if it is collected and interpreted in a specific context. The PPDAC cycle and the GAISE II framework emphasise the need to understand the variability and distribution of data in the context of the question being studied. Variability and context are critical at all stages of the data process. Collaboration and knowledge sharing are also essential components of the framework in order to draw informed conclusions and create shared understanding.

4.3.4 Key competences: The C's of data literacy

The DALI4US framework emphasises that certain skills are needed to manage the data analysis process effectively. These include the “three C's of data literacy”[5] : comprehension, communication and critical thinking. These three pillars are critical not only for understanding data, but also for communicating results effectively and thinking critically about the implications of the results. An alternative view of the three Cs, as proposed by Jordan Morrow (2021), adds curiosity, creativity and critical thinking, emphasising the importance of innovative and open-minded thinking.

Curiosity drives the formulation of questions, while data collection requires the ability to effectively gather relevant information. Critical thinking is needed to understand the meaning of the data and draw informed conclusions. Creativity is needed to develop new solutions or approaches, and communication is essential to share and disseminate lessons learned.

4.3.5 Iterative and cyclical nature of the process

The whole PPDAC process is iterative and circular, meaning that each stage of the process may lead to a return to and revision of previous stages. Insights from data analysis may lead to new questions and investigations, while conclusions may reveal new aspects of planning. This circular nature of the PPDAC process makes it flexible and dynamic, allowing it to respond to new insights and challenges and to continually refine the process.

In summary, the DALI4US framework provides a comprehensive and structured basis for promoting data literacy, based on both proven statistical methods and innovative technological approaches. It combines essential key competences with a flexible, iterative approach to ensure that learners and professionals are able not only to understand data, but also to use and communicate it effectively. The framework allows for a transversal approach to data literacy, enabling its integration into all school subjects through a constructivist pedagogical approach. This approach encourages learners to actively construct their own understanding of data in different contexts, fostering deeper engagement and making data literacy a core component across the curriculum.

5. Extension of the DALI4US framework

The present chapter explores the potential of extending the DALI4US framework to incorporate elements of exploratory data analysis (EDA). The DALI4US framework, presented in chapter 4.3 and designed to

[5] https://atlan.com/what-is-data-literacy-and-why-is-it-important/#:~:text=Overall%2C%20the%203C's%20of%20Data,various%20professional%20and%20personal%20contexts. (last access: 14.10.2024)

Co-funded by the European Union

teach data literacy in upper primary education, is based on the GAISE II framework and the PPDAC cycle. This framework provides a basis for guiding students through structured processes of problem definition, planning, data collection, analysis and communication of conclusions. The incorporation of elements of iteration, curiosity, and critical thinking is a fundamental aspect of this framework, but it relies primarily on a confirmatory approach to data analysis. The expansion of the framework to encompass exploratory data analysis (EDA) serves to enhance its ability to promote open-ended inquiry, creativity, and flexibility, and consequently provides a more comprehensive and practical approach to the teaching of data literacy.

5.1 The necessity of updating the DALI4US-framework

The structured methodologies of PPDAC and GAISE II are invaluable for teaching fundamental data analysis processes. GAISE II provides multiple opportunities to experience data science concepts (Weiland & Engledowl, 2022). However, due to the predominantly linear nature of PPDAC and GAISE II, they risk oversimplifying the iterative and dynamic aspects of real-world exploration (Smith, 2003). In practice, questions often arise during the analysis, and new insights may lead to a revision of the initial plan or question (Jones, 1987). This dynamic process, which is central to effective data exploration, is not fully captured by a purely confirmatory approach (Brown, 1992). EDA addresses this limitation by complementing the structured elements of PPDAC and GAISE II with a discovery process, prioritising unstructured exploration of data, contrasting with the confirmatory data analysis (CDA), which employs statistical methodologies to test predefined hypotheses.

The contemporary data landscape is distinguished by the integration of a diverse array of data sources, the management of incomplete or distorted data, and the identification of patterns within high-dimensional or complex data sets. Emerging technologies such as machine learning and artificial intelligence present transformative opportunities for data exploration (Jones, 2023). These technologies facilitate automated pattern recognition, anomaly detection, and predictive modelling that extend and enhance conventional EDA methods (Smith, 2003).Visualisation techniques have also undergone substantial advancement, with tools now supporting interactive and real-time data exploration (Behrens, 1997). These technologies have the potential to enhance the efficiency and effectiveness of data analysis for both educators and learners. Furthermore, consideration of ethical aspects, including privacy, impartiality, and transparency, is crucial in modern data practices. The integration of these principles into the framework enables educators and learners to recognise and eliminate bias, which lays the foundation for responsible and reproducible data collection processes.

Alongside the confirmatory processes delineated in the CDA, EDA offers substantial advantages. While CDA is hypothesis-driven and serves to answer specific questions, EDA is an iterative process that aims to gain insights and uncover patterns. EDA uses tools such as histograms, scatter plots and box plots, as well as summary statistics such as means, medians and variances, to examine relationships, trends and anomalies. Unexpected clusters or nonlinear trends can be uncovered, leading to further investigation or refinement of the hypothesis.

Co-funded by the European Union

The flexibility of EDA is particularly effective in dealing with problems that arise during data collection and analysis, as it allows for iterative examination of data, identification of inconsistencies, and correction of missing values or outliers that could affect subsequent analyses. This proactive approach ensures that issues with data quality are addressed early on, thereby increasing the reliability of subsequent validation analyses. EDA fosters critical thinking by encouraging students to question assumptions and engage deeply with the data. In contradistinction to CDA, which adheres to predefined models and strict assumptions, EDA invites experimentation and customisation, helping students to develop an intuitive understanding of the data. This openness renders EDA a powerful tool for building data literacy in educational contexts.

The emphasis on context and collaboration in the former DALI4US-framework highlights the need for an updated approach to EDA. The analysis of contemporary data challenges underscores the imperative for multidisciplinary and cross-functional teams, wherein insights from diverse perspectives are paramount. Therefore the promotion of collaboration and contextual understanding can facilitate the development of an updated framework and enable adaptive problem solving.

5.2 Complementary roles of EDA and CDA

EDA and CDA are not competing methods, but rather complementary approaches that collectively form a comprehensive analytical toolkit. EDA offers a foundational framework for CDA by formulating hypotheses and illuminating the data structure, while CDA employs statistical methods to test these hypotheses or make predictions. John W. Tukey (1977) discussed data analysis as a continuum from EDA to CDA. The integration of exploratory and confirmatory approaches to data analysis underscores the necessity for EDA's iterative insights to inform the structured processes of CDA. “(a) both exploration and confirmation are important, (b) exploration comes first, (c) any given study can, and usually should, combine both”. (J. Tukey, 1980) The absence of the creative flexibility of EDA in confirmatory analyses can result in incomplete or misdirected results, while the absence of exploratory findings from the rigour of CDA can compromise validation. The integration of these approaches is therefore pivotal in ensuring a balanced and comprehensive understanding of the data, enabling analysts to transition seamlessly from the discovery phase to the validation stage. The incorporation of EDA within the DALI4US framework serves to move it beyond a linear process, thereby better reflecting the dynamic nature of real-world investigations. This balanced approach equips learners with the skills necessary to navigate complex data sets and fosters curiosity, critical thinking, and adaptability, ensuring that data is explored creatively and iteratively, preparing learners for the challenges of a data-driven world.

5.3 Exploratory data analysis (EDA): definition and role

The concept of exploratory data analysis (EDA), as introduced and popularised by Tukey in the 1970s, occupies a central position in the fields of data science and statistics. In contradistinction to confirmatory data analysis (CDA), which is hypothesis-based and aims to test pre-established hypotheses, the focus of EDA is on open-ended discovery. Its objective is to reveal latent insights, patterns, and relationships

Co-funded by the European Union

within a data set, unencumbered by the rigidity of preconceived hypotheses or expectations. This preliminary phase of data analysis enables the data to be familiarised, anomalies to be identified, and meaningful questions to be formulated, thereby guiding the subsequent analysis strategies. EDA is not simply a set of techniques but an attitude towards data. “EDA is an attitude, a state of flexibility, a willingness to look for those things that we believe are not there, as well as those we believe to be there” (J. W. Tukey, 1984) The fundamental principle of EDA is its iterative and flexible character, which stands in contrast to the more rigid and linear approach of CDA. Rather than following a predetermined analysis path, EDA encourages repeated examination and re-analysis of data. An exploratory analysis looks at data from as many angles as possible (Morgenthaler, 2009). Through the use of visualisations and statistical summaries, the understanding of a data set can be iteratively refined. EDA makes extensive use of visual tools such as histograms, scatter plots, box plots, and line charts to examine data distributions, identify outliers, and detect trends. These graphical methods are complemented by summary statistics, including measures of central tendency (e.g., mean and median) and spread (e.g., variance and interquartile range), which provide a quantitative overview of the data structure. EDA is distinguished by its emphasis on creativity and adaptability, as it frees from the constraints of predefined models, encouraging an experimental approach to understanding the data. This approach facilitates the identification of inconsistencies or anomalies that might otherwise remain unnoticed. Furthermore, the EDA plays a pivotal role in the assessment of data quality, facilitating the identification of missing values, inconsistencies, or measurement errors at an early stage of the analysis process, thereby ensuring that subsequent analyses are based on a reliable foundation.

Beyond its technical applications, the EDA fosters a profound comprehension of data, thereby integrating it into educational curricula and decision-making processes. Within the educational milieu, the EDA serves as a pedagogical instrument, imparting critical data literacy skills such as trend interpretation and variability discernment. To illustrate, in the context of analysing meteorological data, students can employ line plots to discern seasonal patterns or scatter plots to investigate the correlation between temperature and precipitation. These exercises cultivate active engagement with data and nurture the development of analytical intuition.

In practice, EDA serves as a conduit between raw data and actionable insights, laying the foundation for confirmatory analysis by formulating hypotheses and influencing model selection. The significance of EDA is accentuated by its congruence with the progression of computational tools and software, with the pervasive adoption of interactive visualisation tools rendering EDA more accessible and versatile. These tools facilitate the efficient exploration of large data sets, experimentation with diverse visualisations, and the dynamic refinement of analyses.

EDA can thus be regarded as both a philosophy and a practical framework for data analysis. According to Tukey, the philosophy is basic common sense: one must learn about data before trying to learn from the data. “There is no right or wrong way to conduct an EDA process. The key is to keep an open mind and to test different modelling techniques until new information about the data is uncovered.” (Courtney, 2021)

Co-funded by the European Union

5.4 Extended DALI4US-framework

According to Philipp Morrison (1983) EDA can be seen as a response to the uncritical teaching and application of confirmatory statistics. Introducing students to confirmatory statistics at the start of their training is often inappropriate because it assumes that they already possess a deep understanding of the problem, a well-defined theory, and a properly derived model to test. Integrating exploratory data analysis (EDA) into school education is a crucial step in equipping students with the data literacy and analytical skills needed to succeed in a data-driven world. With data at the heart of decision-making in fields ranging from science to social policy, teaching EDA to students at an early age helps them develop critical thinking and problem-solving skills, as well as the ability to interpret and communicate information effectively. Consequently the integration of EDA in the curricula for grades 1 to 12 is in line with the findings from the needs analysis (cf chapter 2).

In today's world, data is ubiquitous, whether in social media, environmental studies or personal health, and students need to be prepared to make informed decisions based on data. EDA teaches students to examine datasets, spot patterns and ask questions, encouraging a deep engagement with information. Furthermore EDA taps into the natural curiosity of young learners. Through hands-on activities, students learn to use data as a tool to answer questions about the world around them. For example, they can explore topics such as climate change, population growth or sports performance, making learning relevant and exciting. This connection to real-world problems helps them develop an appreciation of data as a means of understanding and influencing their environment. The use of real-world datasets makes learning engaging and relevant, motivating students and improving retention. EDA supports cross-curricular learning by integrating with subjects such as mathematics, where students can explore statistical measures like mean and median, science through experimental data analysis, and social studies by examining historical or demographic trends.

Strategies for teaching EDA include introducing basic data concepts early in primary school through simple data collection and visualization activities, such as bar graphs or pictograms. Age-appropriate tools, like for example OrangeEdu, enhance accessibility and skill development. Introducing concepts of data quality, such as handling missing values or outliers, highlights the importance of clean data and sharpens problem-solving skills.

As mentioned in the previous chapter, John Tukey (1977, 1980) redefined the process of data analysis by shifting from a traditional linear framework to a more iterative and dynamic approach. The former DALI4US-framework presented in chapter 4.3 outlines a straightforward methodology:

1. Formulate a question: Clearly define the problem to investigate.
2. Collect & consider data: Plan a structured approach to answer the question, collect relevant data based on the experimental design.
3. Analyse and interprete data: Analyze the data using statistical methods and conclude and communicate findings.

Co-funded by the European Union

This linear progression forms the core of a confirmatory process, with a clear trajectory from a question (or a hypothesis) to conclusion. However, the former DALI4US-framework already highlighted that the progression is not always straightforward and included an element of iteration. When a question cannot be answered with the available data, or when further clarification is needed, the process naturally adopts an iterative approach. This means that the inquiry revisits earlier stages, collecting additional data or refining hypotheses. New questions may arise during this process, prompting further investigation and modification of the initial trajectory.

In contrast, Tukey's exploratory data analysis (EDA) introduces a more flexible, investigative methodology. Describing EDA as "detective work – numerical detective work – or counting detective work – or graphical detective work" (J. Tukey, 1977), the author emphasizes its project-based and engaging nature. EDA integrates exploration and discovery into the analytical process:

1. Start with an idea: Begin with a broad concept or area of interest rather than a rigid question.
2. Iterate between question and design: Refine both the research questions and designs in tandem, allowing for adjustments as insights emerge.

This iterative process moves beyond the constraints of predefined questions or hypotheses, encouraging a dynamic interplay between ideas, design, and data. Tukey's reconceptualization aligns data analysis with real-world exploration, where flexibility and adaptability are essential for uncovering patterns, generating hypotheses, and advancing knowledge.

However this cannot occur in isolation. It requires a carefully designed and systematically researched learning environment—imbued with a genuine spirit of EDA. The challenge lies in developing a coherent sequence of learning situations, complemented by appropriate materials and computerized tools, to transform the teaching and learning of statistics. Such an environment should be designed to have a significant impact on daily classroom practices. The ultimate goal is to move classrooms away from being academic work factories, where students simply complete assigned tasks under teacher management, toward becoming communities of learning and interpretation. In these communities, students are empowered to take ownership of their learning, engage deeply with data, and actively participate in the process of inquiry. Reflective practice, involving students, teachers, and researchers, should be an integral component of this transformation, fostering a culture of mutual learning and collaboration. This approach not only aligns with the principles of EDA but also reimagines classrooms as dynamic spaces where statistical literacy and critical thinking can thrive.

The literature underscores the challenges educators encounter when transitioning to inductive and exploratory teaching methodologies. Effective teacher preparation programs are essential to address these challenges and to support teachers in creating dynamic, inquiry-driven learning environments. Key objectives (Ben-Zvi, 2001) for such programs include:

- Shifting the teaching paradigm: Transitioning teachers from the role of a "sage on the stage," where they primarily deliver knowledge, to a "guide on the side," where they facilitate student-driven inquiry and exploration.

Co-funded by the European Union

- Promoting interactive, investigative methods: Encouraging educators to adopt teaching practices that prioritize interaction, collaboration, and hands-on exploration of data.

Professional development initiatives play a critical role in achieving these goals. For instance, framing EDA as "detective work" can empower teachers to cultivate investigative mindsets in their classrooms. Such approaches help teachers design and facilitate environments where students engage deeply with data, ask meaningful questions, and uncover patterns through guided exploration. This preparation, which is part of the DALI4US-project, is pivotal for enabling teachers to effectively support students in mastering the principles and practices of EDA, fostering both statistical literacy and critical thinking skills.

There is a need to stress the importance of using real data and focusing less on mathematical and probabilistic theory, while prioritizing interpretation and communication (Cobb & Moore, 1997). Moore (1990) elaborates that statistics combines computational activity in meaningful settings with the exercise of judgment in choosing methods and interpreting results, highlighting its practical and applied nature.

In his synergy on statistical education, Moore (1997) integrates content, pedagogy, and technology to create a cohesive learning environment. Content and pedagogy are intertwined, with a focus on hands-on data analysis, communication, and conceptual understanding, while reducing the emphasis on rigorous mathematical proofs. Pedagogy is enhanced by technology, which facilitates visualization through multiple representations, automates calculations, and supports active learning through multimedia resources. Technology, in turn, transforms content by enabling advanced data analysis, diagnostics, and simulations, providing alternatives to traditional proofs and allowing for the exploration of broader and more complex concepts. This framework underscores the transformative potential of technology in reshaping statistical education, enabling a more interactive and practical approach to learning. By combining content, pedagogy, and technology effectively, educators can create dynamic environments that prioritize real-world application, critical thinking, and meaningful engagement with data. Moores synergy on statistical education has a lot of overlaps with the TPACK (Technological Pedagogical Content Knowledge) framework. This framework describes the integration of technology, pedagogy, and content knowledge to enhance teaching and learning. It emphasizes the need for teachers to understand how to use technology effectively in conjunction with pedagogy and subject matter to create engaging and meaningful educational experiences. Aligned with these goals, DALI4US aims to function as a digital ecosystem that integrates framework development, software, and teacher education.

The revised DALI4US framework incorporates a series of updated core principles and considerations aimed at enhancing its focus on exploratory data analysis and fostering comprehensive data literacy. The core principles emphasize

- **democratization**, ensuring equitable access to data, tools, and knowledge for all learners, regardless of background,
- the principle of **real-world context** highlighing the importance of grounding data literacy in practical, everyday applications, making data analysis meaningful and relevant,
- a **cross-curricular approach** being integral to the framework, promoting the integration of data literacy skills across diverse subjects to support interdisciplinary learning,

Co-funded by the European Union

- the **ubiquity of data**, emphasizing the pervasive presence of data in various forms, such as text, images, and other digital formats, to foster awareness and engagement in modern life.

The framework also integrates several key considerations essential to its successful implementation:

- A **foundational understanding of data** is prioritized, ensuring learners can identify and comprehend the different forms of data and recognize its omnipresence in the digital age.
- Attention is given to **skills and tools**, equipping learners with practical techniques and familiarity with **visual data analysis** tools, such as CODAP and OrangeEDU, to enable effective data exploration and analysis.
- The framework highlights the importance of innovative **pedagogical approaches**, encouraging inquiry-based learning and interactive exploration to engage students actively.
- Furthermore, the development of **critical thinking** skills is emphasized as a core competency, enabling learners to analyze, interpret, and evaluate data thoughtfully and responsibly.
- Finally, the revised framework places significant emphasis on **the ethical use of data**, ensuring that learners understand the broader implications of data usage and are equipped to navigate ethical considerations in their analysis and application of data.

By integrating these principles and considerations, the new DALI4US-framework provides a robust foundation for fostering data literacy in a manner that is accessible, relevant, and grounded in both technical and ethical dimensions. This approach seeks to empower learners to critically engage with data and apply their skills in diverse, real-world contexts.

The following picture illustrates the process:

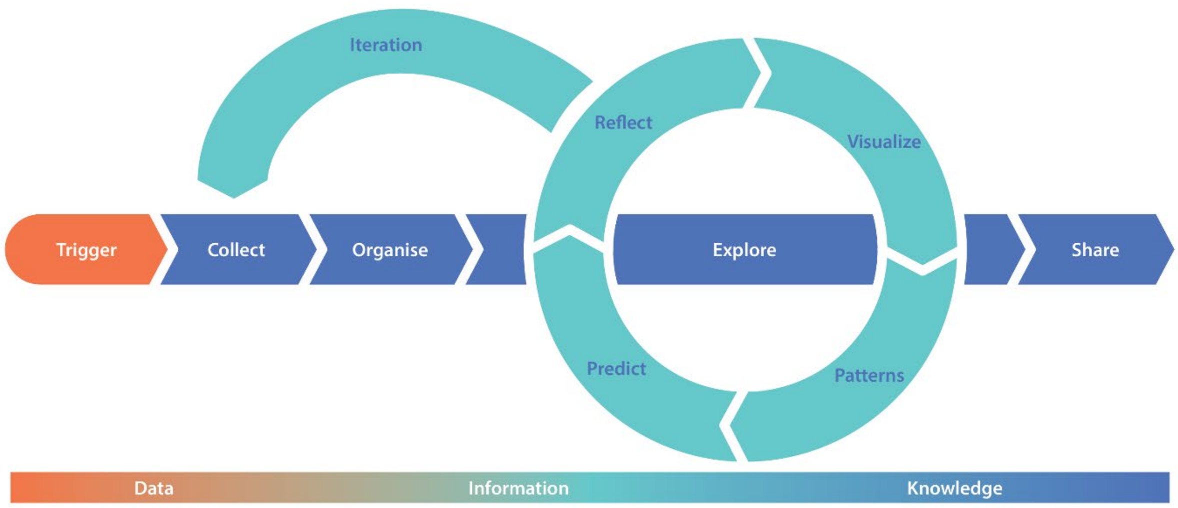

Figure 6: The extended DALI4US framework (own representation)

Co-funded by the European Union

This extended DALI4US-framework reflects the core ideas from Moore's synergie in statistical education, integrating content, pedagogy, and technology to foster an interactive and practical approach to learning. This cyclical process highlights the iterative nature of data exploration and interpretation, essential for modern data literacy frameworks:

- **Trigger:** This stage initiates curiosity, encouraging students to ask meaningful questions and engage with real-world data—aligning with Moore and Cobb's emphasis on using real data and fostering interpretation.
- **Collect:** Gathering data from the world mirrors Moore's recommendation to ground learning in meaningful, authentic contexts, integrating computational tools as part of the process.
- **Organize:** Structuring data to facilitate easier analysis is consistent with the pedagogical goal of fostering skills in managing and preparing data for exploration and interpretation.
- **Explore:** Students uncover initial insights, consistent with Tukey's emphasis on exploratory data analysis (EDA) as detective work, where patterns and trends are sought iteratively to build understanding.
- **Patterns:** Identifying trends and relationships within the data reflects Moore's call for a focus on higher-order thinking, problem-solving, and making connections within the data.
- **Predict:** Developing predictive models builds on this understanding, encouraging application and judgment, as highlighted in Moore's focus on combining computational activity with interpretation.
- **Reflect:** Reviewing insights, predictions, and findings emphasizes critical thinking and self-assessment, encouraging a deeper engagement with data.
- **Share:** Communicating and presenting findings aligns with Moore's stress on communication as a vital component of statistical practice.

This framework integrates technology throughout, supporting visualization, automation, and interactivity, as suggested in Moore's approach. It promotes a dynamic, hands-on learning environment where students can engage deeply with data while fostering critical thinking and practical application, making it an effective tool for enhancing statistical education. DALI4US further positions itself as a digital ecosystem that integrates these elements to transform classrooms into collaborative communities of learning and exploration.

Furthermore, it builds upon foundational ideas from the preceding DALI4US-framework (cf chapter 4.3) by integrating modern concepts like Nathan Shedroff's continuum of understanding, often visualized in the DIKW pyramid (Data, Information, Knowledge, Wisdom). This continuum forms the conceptual backbone of DALI4US, illustrating the progression from raw data to meaningful insights and actionable knowledge. Another core principle carried forward from the earlier framework is the importance of **context**. In the DALI4US framework, data is not treated as an isolated entity; instead, it gains meaning only when interpreted within its specific context. The DALI4US framework also incorporates the "three C's of data literacy" - comprehension, communication, and critical thinking - expanded with curiosity and creativity as proposed by Morrow (2021).

Co-funded by the European Union

6. List of figures

7. Bibliography

Behrens, J. T. (1997). Principles and procedures of exploratory data analysis. *Psychological Methods*, *2*(2), 131–160. https://doi.org/10.1037/1082-989X.2.2.131

Ben-Zvi, D. (2001). *Children as Explorers: Exploratory Data Analysis by Junior High School Students in a Computer Assisted Environment*. https://www.researchgate.net/publication/258342340_Children_as_Explorers_Exploratory_Data_Analysis_by_Junior_High_School_Students_in_a_Computer_Assisted_Environment

Bhargava, R. (2015). *Designing Tools and Activities for Data Literacy Learners*. https://www.semanticscholar.org/paper/Designing-Tools-and-Activities-for-Data-Literacy-Bhargava/12157fe54d862929e301f7fdb2c6233c1ae4502a

Bhargava, R., & D'Ignazio, C. (2015). *Designing Tools and Activities for Data Literacy Learners*. MIT Media Lab. https://www.media.mit.edu/publications/designing-tools-and-activities-for-data-literacy-learners/

Brown, A. L. (1992). Design Experiments: Theoretical and Methodological Challenges in Creating Complex Interventions in Classroom Settings. *Journal of the Learning Sciences*, *2*(2), 141–178. https://doi.org/10.1207/s15327809jls0202_2

Burgee, R. (2018, Februar 26). *The History of Digital Content [INFOGRAPHIC]*. MerlinOne. https://merlinone.com/history-of-digital-content-infographic/

Cobb, G. W., & Moore, D. S. (1997). Mathematics, Statistics, and Teaching. *The American Mathematical Monthly*, *104*(9), 801–823. https://doi.org/10.1080/00029890.1997.11990723

Courtney, M. B. (2021). Exploratory Data Analysis in Schools: A Logic Model to Guide Implementation. *International Journal of Education Policy and Leadership*, *17*(4), Article 4. https://doi.org/10.22230/ijepl.2021v17n4a1041

Co-funded by the European Union

Deep Talk. (2021, Oktober 21). 80% of the world's data is unstructured. *Medium*. https://deep-talk.medium.com/80-of-the-worlds-data-is-unstructured-7278e2ba6b73

Engel, J., Biehler, R., Frischemeier, D., Podworny, S., Schiller, A., & Martignon, L. (2019). Zivilstatistik: Konzept einer neuen Perspektive auf Data Literacy und Statistical LiteracyCivic Statistics: Concept of new perpectives on data literacy and statistical literacy. *AStA Wirtschafts- und Sozialstatistisches Archiv*, *13*. https://doi.org/10.1007/s11943-019-00260-w

Gray, J., Gerlitz, C., & Bounegru, L. (2018). Data infrastructure literacy. *Big Data & Society*, *5*, 205395171878631. https://doi.org/10.1177/2053951718786316

Guler, G. (2019). *Data literacy from theory to reality: How does it look?* https://doi.org/10.13140/RG.2.2.27537.35680

Heidrich, J., Bauer, P., & Krupka, D. (2018). *Strukturen und Kollaborationsformen zur Vermittlung von Data-LiteracyKompetenzen – Stand der Forschung. Arbeitspapier Nr. 32*. 10.5281/zenodo.1408600

Jones, L. V. (1987). *The Collected Works of John W. Tukey: Philosophy and Principles of Data Analysis 1965-1986, Volume IV* (1st edition). Chapman and Hall/CRC.

Koltay, T. (2015). Data literacy: In search of a name and identity. *Journal of Documentation*, *72*, 401–415. https://doi.org/10.1108/JD-02-2014-0026

Long, D., & Magerko, B. (2020). What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (S. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376727

Makar, K. (2018). Theorising Links Between Context and Structure to Introduce Powerful Statistical Ideas in the Early Years. In A. Leavy, M. Meletiou-Mavrotheris, & E. Paparistodemou (Hrsg.), *Statistics in Early Childhood and Primary Education: Supporting Early Statistical and Probabilistic Thinking* (S. 3–20). Springer. https://doi.org/10.1007/978-981-13-1044-7_1

Mandinach, E., & Gummer, E. (2013). A Systemic View of Implementing Data Literacy in Educator Preparation. *Educational Researcher*, *42*, 30–37. https://doi.org/10.3102/0013189X12459803

Mayer-Schönberger, V., & Cukier, K. (2013). *Big Data: A Revolution That Will Transform How We Live, Work, and Think*. Eamon Dolan / Houghton Mifflin Harcourt.

Mertala, P. (2020). Data (il)literacy education as a hidden curriculum of the datafication of education. *Journal of Media Literacy Education*, *12*(3), 30–42. https://doi.org/10.23860/JMLE-2020-12-3-4

Mishra, P., & Koehler, M. J. (2007). *Technological Pedagogical Content Knowledge (TPCK): Confronting the Wicked Problems of Teaching with Technology*. 2214–2226. https://www.learntechlib.org/primary/p/24919/

Moore, D. S. (1997). New Pedagogy and New Content: The Case of Statistics. *International Statistical Review / Revue Internationale de Statistique*, *65*(2), 123–137. https://doi.org/10.2307/1403333

Co-funded by the European Union

Morgenthaler, S. (2009). Exploratory data analysis. *WIREs Computational Statistics*, *1*(1), 33–44. https://doi.org/10.1002/wics.2

Morrison, P. S. (1983). Teaching Exploratory Data Analysis. *New Zealand Journal of Geography*, *75*(1), 2–6. https://doi.org/10.1111/j.0028-8292.1983.tb00140.x

Morrow, J. (2021). *Be Data Literate: The Data Literacy Skills Everyone Needs To Succeed* (1. Aufl.). Kogan Page.

Pangrazio, L., & Selwyn, N. (2018). “It's Not Like It's Life or Death or Whatever”: Young People's Understandings of Social Media Data. *Social Media + Society*, *4*(3), 2056305118787808. https://doi.org/10.1177/2056305118787808

Pangrazio, L., & Selwyn, N. (2021). Towards a school-based ‘critical data education’. *Pedagogy, Culture & Society*, *29*(3), 431–448. https://doi.org/10.1080/14681366.2020.1747527

Prado, J. C., & Marzal, M. Á. (2013). Incorporating Data Literacy into Information Literacy Programs: Core Competencies and Contents. *Libri*, *63*(2), 123–134. https://doi.org/10.1515/libri-2013-0010

Ridsdale, C., Rothwell, J., Smit, M., Ali-Hassan, H., Bliemel, M., Irvine, D., Kelley, D., Matwin, S., & Wuetherick, B. (2015). *Strategies and Best Practices for Data Literacy Education: Knowledge Synthesis Report* [Report]. https://doi.org/10.13140/RG.2.1.1922.5044

Sapp Nelson, M., Carlson, J., Fosmire, M., & Miller, C. (2011). Determining Data Information Literacy Needs: A Study of Students and Research Faculty. *Portal-libraries and The Academy - PORTAL-LIBR ACAD*, *11*, 629–657. https://doi.org/10.1353/pla.2011.0022

Schiller, A. (2020). *Mathematische Bildung und Demokratie: Statistical Literacy als eine Basiskompetenz zivilgesellschaftlicher Partizipation* [Pädagogische Hochschule Ludwigsburg]. https://phbl-opus.phlb.de/frontdoor/index/index/docId/679

Seemann, M. (2014). *Das neue Spiel: Strategien für die Welt nach dem digitalen Kontrollverlust* (New Edition). orange-press.

Smith, T. (2003). Health psychology. In J. A. Schinka & W. F. Felicer (Hrsg.), *Handbook of psychology: Vol. 2. Research methods in psychology* (S. 241–270). Wiley. https://www.wiley.com/en-au/Handbook+of+Psychology%2C+Volume+2%2C+Research+Methods+in+Psychology%2C+2nd+Edition-p-9780470890646

Tukey, J. (1977). *Exploratory Data Analysis* (1. Aufl.). Pearson.

Tukey, J. (1980). We Need Both Exploratory and Confirmatory. *The American Statistician*, *34*(1), 23–25. https://doi.org/10.2307/2682991

Tukey, J. W. (1984). *The Collected Works of John W. Tukey: Philosophy and principles of data analysis, 1949-1964*. Wadsworth Advanced Books & Software.

Co-funded by the European Union

Van Audenhove, L., Broeck, W. V. den, & Mariën, I. (2020). Data literacy and education: Introduction and the challenges for our field. *Journal of Media Literacy Education*, *12*(3), 1–5. https://doi.org/10.23860/JMLE-2020-12-3-1

Vuorikari, R., Kluzer, S., & Punie, Y. (2022). *DigComp 2.2: The Digital Competence Framework for Citizens - With new examples of knowledge, skills and attitudes*. Publications Office of the European Union. https://doi.org/10.2760/115376

Weiland, T., & Engledowl, C. (2022). Transforming Curriculum and Building Capacity in K–12 Data Science Education. *Data Science Education*, *4*. https://doi.org/10.1162/99608f92.7fea779a

Wikström, J. (2022). *If data is the new oil, the Data Platform collects and refines it*. https://www.kaito.fi/blog/if-the-data-is-a-new-oil-the-data-platform-collects-and-refines-it

Wolff, A., Gooch, D., Montaner, J. J. C., Rashid, U., & Kortuem, G. (2016). Creating an Understanding of Data Literacy for a Data-driven Society. *The Journal of Community Informatics*, *12*(3), Article 3. https://doi.org/10.15353/joci.v12i3.3275

Co-funded by the European Union

Annex 1: Gaise II Framework

Process Component	Level A	Level B	Level C
I. Formulate Statistical Investigative Questions	Understand when a statistical investigation is appropriate Pose statistical investigative questions of interest to students where the context is such that students can collect or have access to all required data Pose summary (or descriptive) statistical investigative questions about one variable regarding small, well-defined groups (e.g., subset of a classroom, classroom, school, town) and extend these to include comparison and association statistical investigative questions between variables Experience different types of questions in statistics: those used to frame an investigation, those used to collect data, and those used to guide analysis and interpretation	Recognize that statistical investigative questions can be used to articulate research topics and that multiple statistical investigative questions can be asked about any research topic Understand that statistical investigative questions take into account context as well as variability present in data Pose summary, comparative, and association statistical investigative questions about a broader population using samples taken from the population Pose statistical investigative questions that require looking at a variable over time Understand that there are different types of questions in statistics: those used to frame an investigation, those used to collect data, and those used to guide analysis and interpretation Pose statistical investigative questions for data collected from online sources and websites, smartphones, fitness devices, sensors, and other modern devices	Formulate multivariable statistical investigative questions and determine how data can be collected and analyzed to provide an answer Pose summary, comparative, and association statistical investigative questions for surveys, observational studies, and experiments using primary or secondary data Pose inferential statistical investigative questions regarding causality and prediction

Co-funded by the European Union

Process Component	Level A	Level B	Level C
II. Collect Data/ Consider Data	Understand that data are information; recognize that to answer a statistical investigative question, a person may collect data themselves specifically for that purpose, or a person may use data that have been collected by other people for another purpose Understand how to collect and record information from the group of interest using surveys and measurements collected from observations and simple experiments Understand that a variable measures the same characteristic on several individuals or objects and results in data values that may fluctuate Understand that within a data set there can be different types of variables (e.g., categorical or quantitative) Interrogate the data set to understand the context of the variables as they may relate to statistical investigative questions Understand that data are not always pristine but may contain errors, have missing values, etc., and that decisions have to be made about how to account for these issues	Understand that data are information collected and recorded with a purpose and can be organized and stored in a variety of structures (e.g., spreadsheets) Understand that a sample can be used to answer statistical investigative questions about a population. Recognize the limitations and scope of the data collected by describing the group or population from which the data are collected Understand that data can be used to make comparisons between different groups at one point in time and the same group over time Recognize that data can be collected using surveys and measurements, and develop a critical attitude in analyzing data collection methods Understand that quantitative variables may be either discrete or continuous Understand how to interrogate the data to determine how the data were collected, from whom they were collected, what types of variables are in the data, how the variables were measured (including units used), and the possible outcomes for the variables Understand that data can be collected (primary data) or existing data can be obtained from other sources (secondary data) Understand how random assignment in comparative experiments is used to control for characteristics that might affect responses	Word as: Apply an appropriate data collection plan when collecting primary data or selecting secondary data for the statistical investigative question of interest. Distinguish between surveys, observational studies, and experiments Understand what constitutes good practice in designing a sample survey, an experiment, and an observational study Understand the role of random selection in sample surveys and the effect of sample size on the variability of estimates Understand the role of random assignment in experiments and its implications for cause-and-effect interpretations Understand the issues of bias and confounding variables in observational studies and their implications for interpretation Understand practices for handling data that enhance reproducibility and ensure ethical use, including descriptions of alterations, and an understanding of when data may contain sensitive information Understand how concerns about privacy and human subjects may affect the collection and distribution of data Understand that in some circumstances, the data collected or considered may not generalize to the desired population, or this data may be the entire population

Co-funded by the European Union

Process Component	Level A	Level B	Level C
III. Analyze the Data	Understand that the distribution of a categorical variable or quantitative variable describes the number of times a particular outcome occurs Represent the variability of categorical variables or quantitative variables using appropriate displays (e.g., tables, picture graphs, dotplots, bar graphs) Describe key features of distributions for quantitative variables, such as: ∘ center: mean as the equal share, and median as the middle-ordered value of the data ∘ variability: range as the difference between the greatest and least value, and dispersion as how many units from the equal share value ∘ shape: number of clusters, symmetric or not, and gaps Recognize distributions can be used to compare two groups Observe whether there appears to be an association between two variables	Represent the variability of quantitative variables using appropriate displays (e.g., dotplots, boxplots) Learn to use the key features of distributions for quantitative variables, such as: ∘ center: mean as a balance point, and median as the middle-ordered value ∘ variability: interquartile range and mean absolute deviation (MAD) ∘ shape: symmetric or asymmetric and number of modes Use reasoning about distributions to compare two groups based on quantitative variables Explore patterns of association between two quantitative variables or two categorical variables: ∘ measures of correlation: quadrant count ratio (QCR) ∘ comparison of conditional proportions across categorical variables	Use technology to subset and filter data sets and transform variables, including smoothing for time series data Identify appropriate ways to summarize quantitative or categorical data using tables, graphical displays, and numerical summary statistics, which includes using standard deviation as a measure of variability and a modified boxplot for identifying outliers Summarize and describe relationships among multiple variables Understand how sampling distributions (developed through simulation) are used to describe the sample-to-sample variability of sample statistics Develop simulations to determine approximate sampling distributions and compute *p*-values from those distributions Describe associations between two categorical variables using measures such as difference in proportions and relative risk Describe the relationship between two quantitative variables by interpreting Pearson's correlation coefficient and a least-squares regression line Use simulations to investigate associations between two categorical variables and to compare groups Construct prediction intervals and confidence intervals to determine plausible values of a predicted observation or a population characteristic

Co-funded by the European Union

Process Component	Level A	Level B	Level C
IV. Interpret Results	Use statistical evidence from analyses to answer the statistical investigative questions and communicate results through structured answers with teacher guidance Make statements about the group or population from which the data were collected, recognizing that conclusions are limited to these groups and cannot be generalized to other groups Describe the difference between two groups with different conditions	Use statistical evidence from analyses to answer the statistical investigative questions and communicate results with comprehensive answers and some teacher guidance Acknowledge that looking beyond the data is feasible Generalize beyond the sample providing statistical evidence for the generalization and including a statement of uncertainty and plausibility when needed Recognize the uncertainty caused by sample to sample variability State the limitations of sample information (e.g., a sample may or may not be representative of the larger population, measurement variability) Compare results for different conditions in an experiment	Use statistical evidence from analyses to answer the statistical investigative questions and communicate results through more formal reports and presentations Evaluate and interpret the impact of outliers on the results Understand what it means for an outcome or an estimate of a population characteristic to be plausible or not plausible compared to chance variation Interpret the margin of error associated with an estimate of a population characteristic Acknowledge the presence of missing values and understand how missing values may add bias to an analysis Use multivariate thinking to understand how variables impact one another Communicate statistical reasoning and results to others in a variety of formats (verbal, written, visual) Understand how to interpret simulated p-values appropriately

Co-funded by the European Union