POLITECNICO
MILANO 1863

SCUOLA DI INGEGNERIA INDUSTRIALE
E DELL'INFORMAZIONE

# The Added Value of Data Integration on the Decision-Making of Museums: An Integrated Dashboard for Italian Museums

A thesis submitted for the Master of Science degree in Management Engineering

Author: **Giovanni Razzini**

Student ID: 993232
Advisor: Prof. Michela Arnaboldi
Co-advisor: Eng. Paola Riva
Academic Year: 2022-2023

# Acknowledgements

Desidero esprimere la mia gratitudine alla Professoressa Michela Arnaboldi per la sua competenza e gentilezza, nonché per l'assistenza preziosa fornita durante la redazione di questo lavoro. Un particolare ringraziamento va a Paola Riva, la quale mi ha accompagnato con impegno e pazienza, offrendo costantemente parole di incoraggiamento e aiutandomi nel completare al meglio questo progetto.

Desidero altresì ringraziare tutte le persone che mi hanno sostenuto nel mio percorso universitario, specialmente nei momenti difficili.

Innanzitutto, vorrei esprimere la mia gratitudine agli amici conosciuti durante gli anni universitari, i quali sono diventati parte integrante della mia vita. Un ringraziamento speciale va ad Ale, Piè, Dome, Pia e Sara, con i quali è stato un vero piacere condividere questi cinque anni.

Ringrazio anche gli amici di lunga data, coloro con cui sono cresciuto, per il loro immenso sostegno in tutti i momenti della mia vita. Grazie a Ghelluz, Lusa, Pac, Sax, Cico, Brazz, Giuse, Zaz e Piti.

Il mio più profondo ringraziamento va alla mia famiglia, che è stata al mio fianco in ogni fase di questo percorso. Ringrazio di cuore mia mamma Chiara, mio papà Riccardo, mio fratello Emanuele, i miei nonni Dante e Mariella e tutti gli altri parenti, tra cui farei menzione a mia zia Elisa.

Infine, un ringraziamento speciale va a Sara, una persona straordinaria che ha arricchito la mia vita, mi ha sostenuto costantemente e mi ha guidato nel diventare una persona migliore.

# Abstract

Museums face challenges like decreased funding, bureaucratic delays, lack of digital innovation, and a lack of data culture, analysis, and overall data strategy. After the shift of focus from collections to visitors, museums are additionally pushed by stakeholders to seek transparency not only on resource use but also on societal impact. Nonetheless, extant literature on museum management and data-driven decision-making addressing these aspects lacks a holistic view of how museums manage data, measurement, and reporting in decision-making processes. Considering this literature gap, my thesis aims to explicit the interconnections of these subjects, showing how data integration of internal data and open data can be valuable for improving decision-making in museums. Following an extensive literature review on decision-making, museum management and data integration, my thesis proposes the Integrated Decision-Making Framework for Museums to connect these aspects, emphasizing the significance of integrating open data for museum decision-making. To showcase real-world implications of adopting the data integration perspective to support decision making of museums, I integrated via record matching and clustering techniques open data regarding Italian museums with a proprietary survey dataset of the Observatory of Digital Innovation in Culture of the Politecnico di Milano. On top of the resulting integrated data related to 598 Italian museums, I developed a dashboard with nine visuals to support the decision-making of managers of Italian museums and of the Ministry of Culture. It is through this dashboard that I empirically validate the proposed framework showcasing the added value for museums that data integration of external open data has in improving their decision-making. Indeed, my thesis adds to the extant literature by demonstrating the crucial role of data integration of open data to improve decision-making in museums. Additionally, it expands the limited literature of museum management digital indicators by proposing three Key Performance Indicators (KPIs) based on integrated data, namely Online, On-site, and Organizational Readiness. Moreover, the integrated dashboard developed aids museum managers in understanding their digital performance relative to similar museums. The dashboard enables museum managers to benchmark not only on proprietary data dimensions like revenues, visitors, and personnel but also on museum types and time evolution, with the advantage of benchmarking against about 10% of all the Italian museums. The dashboard is also useful to the Ministry of Culture as it provides a comprehensive perspective of the assessment of the Italian museums' digital status and the evolution over time, as the sample of museums represents a significant portion of the Italian museums.

**Keywords:** decision-making; performance management; dashboard; reporting; data integration; open data; digital data; museum management.

# Abstract in Italiano

I musei devono affrontare sfide come la diminuzione dei finanziamenti, i ritardi burocratici, la mancanza di innovazione digitale e la mancanza di cultura, analisi e strategia generale dei dati. Dopo lo spostamento dell'attenzione dalle collezioni ai visitatori, i musei sono ulteriormente spinti dalle parti interessate a cercare trasparenza non solo sull'uso delle risorse, ma anche sull'impatto sociale. I musei affrontano problematiche significative quali la diminuzione nei finanziamenti, ritardi burocratici, mancanza di innovazione digitale e di una strategia generale sui dati. A seguito dello spostamento del punto di attenzione principale dei musei dalla collezione ai visitatori, i musei sono ulteriormente spinti dagli stakeholder a riportare le informazioni in modo trasparente non solo in termini di uso delle risorse ma anche di impatto sociale generato dalle istituzioni. Considerando questa attuale assenza nella letteratura di una visione olistica su come i musei gestiscano i dati, le misurazioni e il reporting nei processi decisionali, la mia tesi si propone di esplicitare le interconnessioni di questi argomenti, mostrando come l'integrazione dei dati interni e dei dati aperti (in inglese *open data*) possa essere preziosa per migliorare il processo decisionale nei musei. A seguito di un'ampia revisione della letteratura riguardante i processi decisionali, la gestione dei musei e l'integrazione dei dati, la mia tesi propone la struttura concettuale dell'*Integrated Decision-Making Framework for Museums*, con lo scopo di collegare questi aspetti, evidenziando l'importanza dell'integrazione dei dati open per i processi decisionali nei musei. Allo scopo di mostrare la reale applicazione di una logica fondata sulla integrazione dei dati, ho eseguito l'integrazione dei dati, mediante tecniche di *record matching* e *clustering*. In particolare, ho integrato un database di open data riguardante i musei italiani con un database proprietario dell'Osservatorio Innovazione Digitale per la Cultura del Politecnico di Milano. Sulla base di questi dati integrati, relativi a 598 musei italiani, ho poi sviluppato e costruito un cruscotto di indicatori composto da nove viste, atte a supportare il processo decisionale dei manager dei musei e del Ministero della Cultura. Il cruscotto valida empiricamente il *framework* proposto, mostrando il valore aggiunto che l'integrazione degli open data ha nei processi decisionali dei musei. Infatti, la mia tesi contribuisce alla letteratura esistente dimostrando il ruolo cruciale dell'integrazione dei dati aperti per migliorare le decisioni nei musei. Inoltre, la mia tesi estende la limitata letteratura sugli indicatori di prestazione chiave (in inglese *Key Performance Indicators*) in ambito digitale per la gestione di musei proponendo tre *Key Performance Indicators* (KPIs) fondati su dati integrati, ovvero *Online, On-site* e *Organizational Readiness*. Il cruscotto integrato consente così ai gestori dei musei di comprendere le loro prestazioni digitali rispetto a musei simili e di confrontarsi tramite *benchmarking* non solo sulle dimensioni ottenute dai dati proprietari, quali i ricavi, i visitatori ed il personale, ma anche sulle tipologie di museo e sull'evoluzione temporale degli indicatori, con il vantaggio di confrontarsi con circa il 10% di tutti i musei italiani. Il campione dei musei presentato nel cruscotto

ne rende accattivante l'uso anche per il Ministero della Cultura, che può facilmente avere un quadro completo per valutare lo stato digitale dei musei italiani e la sua evoluzione nel tempo.

**Parole chiave:** processo decisionale; misurazione di prestazione; dashboard; reporting integrazione di dati; open data; dati digitali; gestione museale; musei.

# List of Contents

# Chapter 1: Introduction

Museums are non-profit organizations (NPOs) that are integral parts of the society, fulfilling essential roles in preserving culture, promoting education, sharing knowledge, and addressing social challenges (ICOM, 2022). Their role in society has been shifting in the past years, from just collecting and preserving artefacts, to being visitor-centered institutes, focused on the experience and enjoyment of their visitors (Anderson, 2004; Welsh, 2005; Giaccardi, 2012; Bonet & Négrier, 2018; Agostino & Arnaboldi, 2021). This shift is one of the many unique challenges that museums encounter.

This new positioning towards their stakeholders places museums in a position where they increasingly face the typical demands of NPOs' stakeholders for transparency and reporting on the achievement of their social mission and the use of resources (Arena et al., 2015; Mehrotra & Verma, 2015; Rainey et al., 2017). Stakeholders, including donors, sponsors, and the community at large, expect these organizations to demonstrate their commitment to social responsibility and their ability to generate both social and economic value (Millar & Hall, 2013; Munik et al., 2021). While economic value is easily measured, social value is more complex and appreciated over the long-term (Scott, 2007; IIRC, 2013, 2021; VRF, 2022; Reimsbach & Braam, 2023). As NPOs, museums should prioritize maximizing positive societal impact rather than profit, leading to a lack of a clear and measurable common goal (Gstraunthaler & Piber, 2012). However, stakeholders' demands for financial sustainability may drive museums to concentrate on the financial side, potentially losing focus on curatorial and qualitative aspects (Gstraunthaler & Piber, 2012; Whelan, 2015; Loach et al., 2017; Tsai & Lin, 2018).

This need for transparency and reporting towards their stakeholders pushes museums to further their Performance Measurement (PM). Though, many authors have made efforts to develop PM tools for NPOs (Moullin, 2002; Bagnoli & Megali, 2011; Ebrahim & Rangan, 2014), their implementation is still facing challenges related to the complexity and the ambiguous meaning of performance in NPOs (De Waal, 2007; Maheshwari & Janssen, 2014; Arena et al., 2015). Indeed, value-based and accounting indicators, common in business, do not suit the objectives pursued by museums and NPOs. The lack of PM in museums also has an impact on decision-making, as PM systems help in making decisions that are informed and driven by performance (Wholey, 1999; Moynihan, 2005). In museums, PM is still based on the experience and expertise of the evaluators (Gstraunthaler & Piber, 2012).

Indeed, a highly neglected activity in museums is data analysis (Agostino et al., 2020), as it is very difficult to find a museum that has a data strategy in place. Data collection often lacks a coherent strategy, relying instead on sporadic surveys addressed to visitors when specific information is needed. In fact, according to the 2021 survey on

the digital innovation of cultural institution from the Digital Innovation Observatories of the School of Management of Politecnico di Milano, 49% of the surveyed Italian cultural institutions collect data in a static way, meaning that data are collected in one-off occasions and are not constantly updated. Moreover, 38% do not collect any data on visitors. Without a comprehensive understanding of their target audience and the communities they serve, these institutions might struggle to create initiatives that really communicate with their visitors (Liu, 2008; Sheng & Chen, 2012). The lack of a consistent approach results in incomplete insights, making it challenging to make well-informed decisions consistently (Yang et al., 2022). This is also connected to the limited financial resources of museums, which results in many museums facing difficulties in carrying out core activities, such as creating exhibits and organizing events (Conn, 2010; Moldavanova, 2016; Camarero et al., 2019; Agostino et al., 2020; Elbashir et al., 2022).

To tackle these challenges, museums and NPOs must embrace a *context-aware* approach to decision-making. As advocated by Berlanga and Nebot (2016), context-awareness means transferring external data inside the organization to implement context outside the organization in the decision-making processes. Understanding the external context and environmental factors surrounding their initiatives is vital for relevance and effectiveness. It allows these organizations to align their actions with the current needs and trends of their target audience, ensuring that their efforts address the most pressing concerns of their communities.

In this context, the utilization of open data stands out as a valuable tool to implement context-awareness even with limited financial resources. Indeed, open data refers to publicly accessible information that "can be freely used, modified, and shared by anyone for any purpose"[1]. Open data is characterized by portability, meaning that the value of data can be realized in a context that is different from the business or industry in which it was originally generated (Pesce et al., 2019). Since open data is publicly available, museums and NPOs can overcome their resource constraints and improve their data collection efforts by exploiting the vast quantity of open databases, available from diverse sources.

Open data integration is cost-effective, a crucial aspect given the often-limited budgets of museums and NPOs (Conn, 2010; Moldavanova, 2016). Unlike purchasing proprietary datasets or conducting extensive surveys, open data is freely available for use, allowing these organizations to allocate their resources more efficiently.

Though in academic literature there is an ongoing debate on these aspects regarding museums, data, measurement, reporting, PM, and decision making, those aspects are mostly considered separately, without a holistic perspective.

[1] https://opendefinition.org/od/2.1/en/

Therefore, the objective of the thesis is to show that the integration of open data with internal data can be a valuable tool for improving decision-making in museums since it allows them to become context-aware and to improve the quality of the insights extracted from data in a cost-effective way.

To do so, in the thesis, a dashboard is constructed, featuring indicators derived from the data integration of open data. After carefully selecting open datasets available on renown portals - e.g., dati.gov.it and Istat -, open data is acquired in the form of 4 datasets, spanning 4 years, each including 4500 Italian museums from ISTAT. The data is then integrated through record matching techniques and clustering methodologies the open data from ISTAT with a proprietary dataset storing the answers to four survey questionnaires addressed to a total of 1402 Italian museums held by the Observatory of Digital Innovation in Culture. The data integration results in an increased quantity of variables available, leading to a significant improvement in the depth and quantity of information and therefore of insights to be gained by the two targeted stakeholders, Italian museum managers and the Ministry of Culture. Specifically, drawing upon existing literature, three novel KPIs are formulated, namely Online, On-site, and Organizational Readiness, proposed to evaluate the digital status of museums and based on the integrated data.

The thesis adds to the extant literature by seeking to prove the crucial role of data integration of open data to improve decision-making in museums, which is a topic yet underexplored in the literature concerned with PM and decision-making in museums. Moreover, the thesis contributes to the extant literature concerning KPIs for museums with the development of three indicators: Online, On-site and Organizational Readiness. The indicators are tailored to assess museums' digital performance across dimensions of online offerings and on-site services and their readiness for the introduction of new technologies.

The thesis encourages museums to develop a data-driven and context-aware approach to decision-making, exploiting the value-added stemming from data integration of open data. Through the dashboard, museum managers can further understand how they are behaving in terms of their digital performance and how it compares to others. Indeed, the dashboard enables external benchmarking, not only on dimensions that originate from proprietary data, such as revenues, number of visitors, and personnel, but also in terms of dimensions obtained thanks to the integration of data, such as the typology of museum, province, or commune. Additionally, the integration of data fosters the visualization of performance over time, enabling internal benchmarking and serving as a tool to assess the progress achieved by museums and the performance of similar museums.

Furthermore, the dashboard also assists the Ministry of Culture in evaluating the status of museums across various dimensions to identify areas requiring improvement. To achieve these improvements, the Ministry can reallocate funds to

areas in greater need or with significant potential for growth. Additionally, the Ministry can also promote initiatives targeted at museums falling within certain specific value ranges of dimensions, such as a specific geographical area, size, or typology, with the aim of fostering the improvement of performance in the museums that fall under the defined parameters.

It promotes the introduction of an integrated thinking approach to support integrated reporting. Additionally, it recommends the implementation to museums of personalized dashboards (equipped with internal and open data) to enhance the comprehension of performance metrics and improve decision-making. The thesis also advocates for increased funding and support directed to enhancing museums' data and digital capabilities.

The thesis is structured as follows. Chapter 2 explains the interaction between decision-making and data integration at a general level, discussing the role of PM, KPIs and dashboarding for decision making but also the relevance of big data (BD), among which open data and data integration, for decision making. Chapter 3 moves the conversation to the specific problems faced by museums when dealing with data for decision making. Starting from the definition of museum, it navigates the reader through the challenges of PM in museums and to museums' hurdles in adopting innovative data sources as digital and social media for visitor engagement and analytical purposes. Building on this literature, Chapter 4 outlines the research objective of the thesis and illustrates the proposed conceptual framework, called Integrated Decision-Making Framework for Museums, which conceptualizes the added value of data integration in connection to data, measurement, reporting, and the human factor. Chapter 5 details the methodological steps pursued in developing the thesis, beginning from the exploration of the datasets, and subsequently delving into data integration practices. Chapter 6 focuses on the development of KPIs and explains how the dashboard was built, focusing on the content of each of the nine views and showing how to interact with them. Chapter 7 summarizes the findings of the thesis, showing the added value that can be extracted from the dashboard thanks to the data integration of Istat open data and the proprietary survey data of the Observatory. Chapter 8 draws conclusions, highlights the limitations of the thesis, and establishes a pathway for future research built upon the findings of the thesis.

# Chapter 2: Decision-making and data integration

This chapter provides an extensive overview of the interconnection between decision-making and data integration. In section 2.1, the concept of decision-making is defined, stemming from definitions that originate from management literature. Section 2.2 introduces PM systems and performance metrics. It emphasizes the critical role of KPIs in reflecting organizational targets and discusses the challenge of selecting suitable indicators. Section 2.3 delves into the concepts of KPI visualization and data visualization tools, focusing on the dashboard. Section 2.4 further defines the interaction between data and decision-making, highlighting the significance of the relation between data quality, humans, data integration and decision-making. Moving the reader's attention to data, section 2.5 introduces the concepts of BD, data management, and data integration. As a special case of these data, section 2.6 introduces the world of open data. At the beginning of the section, the history of open data is explained, and later it explores types, introduces principles, and cites open data portals.

## 2.1. Decision-making

Decision-making is a very complex concept to define. It can be defined in many ways, with interpretations coming from the different fields of psychology and management literature. Since the aim of the thesis is to assess the impact of decision-making in organizations such as museums and NPOs, decision-making is analyzed from the management point of view. In this literature, decision-making has been defined in various ways. For instance, Simon (1947) refers to decision-making as a process in which the actors who take part have a clear objective, they gather information to develop alternatives and then choose the optimal alternative. The decision-maker enters the process with "bounded rationality", meaning that he does not know all the possible information that could influence the decision. The choice of alternative is based on the information gathered and the capacity of the decision-maker (Simon, 1947). Many authors have added to this definition, focusing on different aspects that make up the decision-making process. Cronbach and Gleser (1957) focus on the different natures of the final decisions of the process, which could be terminal or investigatory. A terminal decision ends the decision-making process, while an investigatory decision calls for more information to make a terminal decision. A terminal decision could also be the starting point of a cycle, as the results of such a decision may produce new information that serves to influence the outcome of the terminal decision (Cronbach & Gleser, 1957). Gelatt (1962) supports the cyclicality of

decision-making and proposes a framework based on the scientific method. The author also draws attention to the issues related to the subjectivity of the decision-makers in evaluating choices.

Mintzberg et al. (1976) adds that decision-making is a process that has the goal of making strategic decisions, highlighting the significance in terms of the scale of resources used and actions taken. In fact, these decisions have an influence on the future success (or failure) of an organization (Dean & Sharfman, 1996; Elbanna & Child, 2007a; Walters & Bhuian, 2004). For example, in 2001, Microsoft made a strategic decision to enter the gaming industry with the launch of the Xbox console. This move involved substantial investments in research, development, and marketing. The success of the Xbox had a profound impact on Microsoft's current position in the technology and entertainment market.

Papadakis et al. (1998) focus their work on the factors that influence decision-making. They find that decision-making is impacted by factors originating from both the environment and from the inside of the organization. For example, in 2015, Volkswagen's response to the emissions scandal was influenced by external factors. Volkswagen faced a major backlash when it was revealed that the company had manipulated emissions tests. This led to a strategic decision-making process involving legal actions, recalls, and changes in leadership. The external pressure from regulatory bodies and public scrutiny influenced the company's decisions in response to the scandal. An example of internal factors influencing decision-making is the shift to services of IBM, in the early 2000s. IBM underwent a significant internal shift by transitioning its focus from hardware to services and consulting. This decision involved restructuring the organization and reallocating resources. The move was driven by the recognition of changing market trends and a strategic decision to adapt to a more service-oriented business model.

Considering the information gathered from the literature, decision-making will be referred to as the iterative cyclic process, that has the final goal of making a choice, given the context and the necessary information to make an informed choice between previously formulated alternatives.

## 2.2. Performance Measurement and Key Performance Indicators

To ensure that decision-making is informed and driven by performance, managers develop and deploy PM systems (Wholey, 1999; Moynihan, 2005). PM is the process of measuring the performance of an organization that has the objective of converting the general strategy to measurable operational goals (Neely et al., 2005; Grosswiele et al., 2013). PM is a cyclic process that aims at understanding how to improve one or more macro business targets (e.g., revenues) and also at assessing the results of past actions (Neely et al., 1995). It shows how far the organization has advanced in bridging

the gaps between current performance and target performance (Weber & Thomas, 2005). The performance of an organization tends to improve when PM systems are used correctly (Bisbe & Malagueño, 2012; Abdel-Maksoud et al., 2015; Pollanen et al., 2017). Several PM frameworks have been proposed in the past:

- Performance Measurement Matrix (Keegan et al., 1989): Performance is divided into four dimensions: cost, non-cost, internal, and external.
- The Strategic Measurement Analysis and Reporting Technique Pyramid (Cross & Lynch, 1988): A four-level pyramid of indicators ensuring an effective link between the wider corporate strategy and daily activities and operations (Khan & Shah, 2011). The pyramid is hierarchical: the firm's vision is translated top-down into a series of strategic/business unit/departmental objectives (Euske & Zander, 2005).
- Balanced Scorecard (BSC) (Kaplan & Norton, 1992): The model divides a business into four different dimensions by providing managers with answers to four fundamental questions (Kaplan & Norton, 1992):
    - "How do we look to shareholders?" (Financial Perspective).
    - "What must we excel at?" (Internal Business Perspective).
    - "How do customers see us?" (Customer Perspective).
    - "Can we continue to improve and create value?" (Innovation and Learning Perspective).

The literature is also abundant with definitions of PM systems. One of the most common definitions is the one provided by Neely et al. (1995, p.1229): "The set of metrics used to quantify the efficiency and/or effectiveness of an action". The metrics in the PM system are the most vital part as they support managerial decision-making (Kucukaltan et al., 2016; Elbanna et al., 2020).

In the literature, several types of performance metrics have been described. An important contribution on this topic is the one by Parmenter (2015) who divides performance metrics into two main types: Result Indicators and Performance Indicators. Based on these two types, Parmenter describes four main types of metrics:

- Key Result Indicators: they show the results of many actions that are carried out by different units. They give management a summary of the performance of the organization (e.g., net profit, customer satisfaction, etc.).
- Result Indicators: they show how different teams are working together to deliver results (e.g., number of planned initiatives to be implemented, in-house courses scheduled, etc.). All financial measures are RIs.
- Performance Indicators: they show what teams are delivering (e.g., numbers of innovations implemented by each team, late deliveries by team, etc.).
- Key Performance Indicators (KPIs): they show, with a daily or weekly frequency, how the organization is performing. By taking action on KPIs,

management is able to increase performance dramatically (Weber & Thomas, 2005; Parmenter, 2020).

The distinction between PIs and RIs is not clear in the literature as KPI has been used as a term that includes all four types, considering it as a metric aimed at showing the performance of the organization in its different aspects. For instance, Domínguez et al. (2019) define KPIs as metrics that represent the most important factors for the success of the organization, while Van Looy and Shafagatova (2016) utilize performance measurements, PIs, and KPIs interchangeably. Weber and Thomas (2005) make a distinction between result metrics and performance metrics, but they consider both as KPIs. Many authors argue against this use of KPIs with an overarching definition. Instead, they advocate for a clear differentiation between results and performance, emphasizing that these terms should not be conflated or used interchangeably when discussing metrics and assessments (e.g., Peral et al., 2017; Parmenter, 2020). In particular, Parmenter (2020) states that while PIs are made to improve performance, RIs aim to just show the results, and thus organizations that only use RIs will not improve their performance driven by these indicators. In the thesis, KPIs will be referred to by considering the broader definition that includes both Results and Performance indicators.

The PM process begins with setting the business aim (or target), and then, KPIs are selected and developed. The KPIs are measures that have an impact on the set target and explain the success or failure in reaching the target. They allow the modeling of a generic target through numeric values (del Mar Roldan-García et al., 2021). KPIs aim to show both the short-term results and the long-term strategy related to the target (Parmenter, 2007). They help with visualizing the performance of an organization (Franco-Santos et al., 2012; Domínguez et al. 2019) and improve the ability to make well-informed decisions (Grosswiele et al., 2013). The challenge with KPIs is to find the correct indicators that actually suit the objective (Tenneson & Brocklehurst, 2018). In the literature, many examples of KPIs can be found, however, they are all specific to certain contexts (e.g., Ying et al., 2018; González et al., 2021; Neri et al., 2021) and cannot be used universally.

The choice of KPIs is crucial for an organization, as a well-chosen KPI could improve performance, whereas a poorly selected one has the potential to negatively impact it (Parmenter, 2020). There are many decisions that need to be made about the selection of a KPI: the scope of the metric, measurement frequency, which type of data to use, who should be responsible for the KPI (Gutierrez et al., 2015), and the interrelations between indicators (Kueng, 2000; Gutierrez et al., 2015). The interrelations are trade-offs that may occur between indicators, meaning that a great result in one KPI could cause a bad result in another. An example could be improving *on-time delivery*, which may lead to increased *inventory* (Kueng, 2000).

In the literature, the process of finding useful KPIs has been described in many ways. Authors in the literature often emphasize specific stages of the process for identifying useful KPIs. Kueng (2000) summarizes the process into 4 very general stages and similar versions of this process can be found in the literature (e.g., del-Rey-Chamorro et al., 2003; Parmenter, 2007).

1. Define high-level process goals: The identification of performance indicators begins with the definition of business process goals. The goal is very general and is broken down into smaller goals in later steps. The identification of the business process goals is tied to the concept of Critical Success Factors (CSFs). CSFs are the areas of a business that are vital to its success (Daniel, 1961). Business process goals should be found based on the CSFs, meaning that each CSF should be described by at least one indicator.
2. Derive PIs: In this stage, the objective is to find a measure that reflects the achievement of a goal. "What is measurable and reflects the extent to which a certain goal has been fulfilled?" (Kueng, 2000, p. 76)
3. Derive subgoals: Since goals and their corresponding performance measures can often be quite broad, especially in initial iterations, it becomes essential to break them down. "Which means or actions can be taken by the organization to fulfill a certain goal? The answer normally received has the form of a subgoal." (Kueng, 2000, p. 77)
4. Refine and modify goals: It is imperative that the measurement of the goal does not become more important than the goal itself. To solve this issue, it is important to check if tracking those indicators could cause unexpected problems. It could happen that indicators are generating trade-offs.

## 2.3. KPI visualization and the dashboard

The KPIs can then be organized using data visualization tools. Data visualization (also called information visualization and knowledge visualization) has been seen by many authors as the solution to the problem of information overload (Lurie & Mason, 2007; McCandless, 2009; Gavrilova et al., 2019). The use of data visualization tools has been increasing in recent years (Berinato, 2016; Troise, 2021). It is a recurrent theme in the literature that the visualization of performance can help with better decision-making (Tan & Platts, 2003; Schiuma et al., 2012). It can also foster the creation of knowledge from data (Burkhard, 2004). Visualization improves the communication of knowledge from the organizations to the stakeholders, improving trust and collaboration (Troise, 2021). Data visualization aids in identifying what areas need to be improved and to assess what is currently performing as expected or even better. Without some kind of data visualization, it becomes impossible to make sense of the large amount of data that organizations deal with (Lurie & Mason, 2007; Gavrilova et al., 2019). Moreover, the visualization of data is not a secondary activity to measurement, as it is important to enhance the understanding of performance measures (Piber et al., 2019).

The way that individuals interact with data, information, and knowledge is fundamentally shaped by visualization. According to Schiuma et al. (2022), well-designed visual representations can greatly improve accessibility, meaning, and inspiration in the distribution of knowledge. It is important to recognize the other side, though, as poorly designed visualizations may distort perceptions, give attention to irrelevant information, induce biases, and result in incorrect judgments (Schiuma et al., 2022). Additionally, Troise (2021) points out potential risks highlighted by managers associated with knowledge visualization, such as the ineffectiveness of displaying complex data in simple visualizations.

An example of a data visualization tool is the dashboard. The dashboard is a data visualization tool that organizes KPIs in a unified view to simplify the reading of the results and consequently improve decision-making (Tan & Platts, 2003; Schiuma et al., 2012; Haber & Schryver, 2019). The definition is very broad because its applications are manifold. Every attempt at the visualization of indicators can be considered a dashboard, as the features that make up a dashboard are just a combination of different visualization objects (elements) that are coordinated in a sound way, so as to allow people to see and understand the data behind them (Wexler et al., 2017). The dashboard is crucial part of processes that help guide strategic decision-making, acting as a common source of information and learning tool (Kitchin et al., 2015).

The dashboard is composed of visual objects that visualize indicators in order to make them easier to understand for the user of the tool (Wexler et al., 2017). It uses graphs and colors to improve and facilitate the assessment of KPIs (Kitchin et al., 2015). An in-depth explanation on the process of dashboard development can be found in section 6.2.

## 2.4. Data and Decision-making

In the previous section, the discussion focused on leveraging data visualization techniques to better understand the performance of the organization and to support decision-making. Indeed, data visualization and data as a general concept are greatly connected to decision-making (Madnick et al., 2009). Provost and Fawcett (2013) suggest that by leveraging analytics to obtain insights from relevant data, decision-making would be improved, resulting in better outcomes deriving from correct decisions. Kabir and Carayannis (2013) add that since data are important for everyone in the organization, they need to be taken into consideration within all the decision-making processes. Höchtl et al. (2016) contribute adding that increased availability of data may lead to finding high-quality information and therefore improving the quality of decisions. These publications follow the trend of an increasing emphasis placed on data by companies in the past years (Constantiou & Kallinikos, 2015; Arnaboldi et al., 2017). However, all these favorable claims presume that data used to get insights is of good quality, meaning data that is accurate, complete, updated, and unbiased (Batini & Scannapieco, 2016). The reality is far from that.

## Data quality and decision-making

In the literature, there are several publications that examine the influences that data can have on decision-making. As Bross (1953) suggested, information is the *fuel* for decision-making. The information that enables decision-making comes from data, that needs to be translated from just raw data to meaningful insights (Madnick et al., 2009). Janis & Mann (1977) and D'Zurilla & Goldfried (1971) underline the importance of informational accuracy, as inaccurate information leads to inaccurate insights and decisions. The concept of informational accuracy is very similar to the concept of data quality, which is a common topic in data literature. Many authors acknowledge the importance of data quality in decision-making processes (e.g., Batini & Scannapieco, 2016) and in data analytics (e.g., Sattari et al., 2017). This theme is the main topic of the Quality Declaration of the European Statistical System (ESS): *"…to provide independent high quality statistical information at European, national and regional levels and to make this information available to everyone for decision-making, research and debate."* (ESS, 2016). With this declaration, the ESS presumes a correlation between high-quality data and high-quality decision-making.

## Data, humans, and decision-making

Van der Voort et al. (2018) show that the issues that arise from data-based decision-making come not only from data itself (data quality issues) but also from human interactions with it. The human side of data-based decision-making is crucial since data is just a tool in the hands of the decision-maker, who is the one making the choices and being accountable for them (Gitelman, 2013). If data is dealt with correctly by humans, it becomes a tool to enhance decision-making and help the decision-makers decipher and model reality through figures and numbers (Van der Voort et al., 2018). Other authors support the claim that managers cannot rely on their own expertise and gut feeling to make sound decisions (Eppler & Bresciani, 2013; Aas & Alaassar, 2018), even though it is still standard practice for some (Pfeffer & Sutton, 2006).

Van der Voort et al. (2018) analyze the interconnected dynamics between data and decision-making in the public sector. They find four different theses about the impact of data on public sector decision-making, with two highlighting a positive relationship and the other two a negative one. The *information optimization* and the *decision optimization* theses highlight that data positively impacts decision-making by easing the information exchange between the decision-maker and the data analyst (who is in charge of selecting data, aggregating data, computing indicators, and providing data visualization). Instead, the *politics of algorithms* and the *information market* theses suggest that data analysts and decision-makers may utilize data strategically to pursue their own interests.

Quattrone (2016) warns organizations of the risks of falling for the illusion of having omniscient knowledge thanks to data, the so-called "dream of perfect information" (p. 1). Insights and indicators coming from data should be analyzed with caution, but this

does not always happen, and in the end, we are often left with many numbers without an explanation (Graham, 2008). According to Quattrone (2016), the risks of misuse of data could mislead decision-making processes, making the processes faster but less accurate. The decision-maker becomes a passive observer because every decision is made by algorithms, which could be outdated, not suited for the problems they are trying to solve, or biased (Quattrone, 2016).

While the algorithms that are fit on data are not subject to external pressures as they do not run for office (Zweig et al., 2018), they are still subject to biases since they are made by humans. For example, there have been instances of algorithms making decisions based on race and social class in the justice and security sectors (European Union Agency for Fundamental Rights, 2022). The use of algorithms for decision-making, if the decisions to be made have strong political implications, can also cause a loss of trust in the organization that is using them (de Bruijn et al., 2022). The biases can be present both in the data selection phase and in the interpretation of the results. (Van der Voort et al, 2018). Data selection is the initial phase of both algorithmic decision-making and, in general, every project that has to do with data. Thus, this issue is not just related to algorithm decision-making, but to the whole data-based decision-making (as shown by Arnaboldi, 2018, for example).

In the data selection phase, it is humans that decide subjectively which data should be kept and later used for the generation of results, and which data should be left out (Arnaboldi et al., 2017; Arnaboldi, 2018). If the actors that perform these activities are guided by their own interests, then data and the insights coming from it will be biased (Van der Voort et al., 2018). This degradation in insights quality can have dangerous consequences for the organization and for society in general, especially when this happens in the public sector (Lorenz et al., 2022). Indeed, in the public sector, decision-makers are pressured by different actors, who may try to achieve their own good instead of the organization's (Jensen & Meckling, 1976; Axelrod & Hamilton, 1981; Van der Voort et al., 2018). It may happen that data is used to justify and legitimize an already taken decision, rather than assisting with making that decision (Van der Voort et al., 2018).

It is then more than necessary for data engineers and analysts to have a profound understanding of the business (Kabir & Carayannis, 2013). Indeed, data engineers (data analysts) that are involved in the data acquisition and transformation processes need to be aligned with the decision-makers (management). The data analysts need to have a business understanding of the data required by management and management needs to understand what type of data can provide the needed information (Kabir & Carayannis, 2013). The decision-makers need to be involved in the process as they are the ones making data-driven decisions (Arnaboldi, 2018); most of the time data analysts are not aligned with the final users of data, decision-makers (Quattrone, 2016). The analysts make arbitrary decisions (Arnaboldi et al., 2017) in the data selection

phase and also in modeling data (Bhimani & Willcocks, 2014). If the two actors are not aligned and decision-makers are not involved, then the benefits that data give to decision-making cannot be reaped (Kabir & Carayannis, 2013; Arnaboldi, 2018).

In the literature, most times data is seen as a tool to improve decision-making, following the *information optimization* and *decision optimization* logic proposed by Van der Voort et al. (2018). What is missing is how to practically achieve those improvements (Arnaboldi et al., 2017; Van der Voort et al, 2018).

## 2.5. Big data and data integration

In this section, a broader definition of data is covered by introducing BD. Then the sources from where data is collected, and the practical implication of data integration are discussed.

### Big data and Data Management

BD has been defined in many ways in literature. BD is a high volume (Laney, 2001) of complex data, both structured and unstructured, coming from internal and external sources. BD can support PM and the creation of value for the organization (De Santis & Presti, 2018). BD is a mixture of traditional data, that comes from the organization, and data that comes from external sources, such as social media or open (public) data (Agostino et al., 2020).

The key features of BD have been defined in several ways. One of the first definitions of BD features is the one by Laney (2001). The author acknowledges that BD are characterized by three *Vs*: (high-) volume, (high-) velocity, and (high-) variety.

- Volume: as in the name, *big* data, it refers to the magnitude of data that is generated and acquired continuously.
- Velocity: it refers to the fast way in which the data is collected.
- Variety: it refers to the different nature of data. Data can be structured, semi-structured, and unstructured. BD is composed of all three natures.

In the literature, many more *Vs* were proposed, such as:

- Value: it refers to the potential (economic and social) that BD have if they are dealt with in the correct way (Chang & Grady, 2019)
- Veracity: it refers to the accuracy and the possible presence of biased and untruthful data in the vast plane of BD, that could cause garbage-in, garbage-out effects (Chang & Grady, 2019).
- Variability: it refers to the variation in the data flow rates. Often, BD velocity is not consistent and has periodic peaks and troughs (Gandomi & Haider, 2015).
- Visualization: it refers to the way in which humans can understand data and its analytics. It is composed of exploratory visualization, evaluative visualization, and explanatory visualization. (Chang & Grady, 2019). Visualization is the way

through which managers can deal with the big amount of information coming inside an organization (Gavrilova et al., 2017).

- Volatility: it refers to the continuously changing nature of data (Chang & Grady, 2019). It was also defined as the timeliness of data (Earley & Henderson, 2017).

In the definition proposed by Laney (2001), BD is defined as either structured or unstructured. Structured data is organized, decipherable by machines (IBM, 2021), that uses a predefined format[2]. It enables an easy and fast analysis of data. Unstructured data is qualitative data, difficult to directly process (IBM, 2021), and it comes in many different formats, such as images, audio, video, undefined text.[8]

BD has been given many different features as it is a very ample concept. Most of the features are double-faceted, having both positive and negative implications. To address the negatives and foster the positives, data need to be treated cautiously by implementing data management processes, as without those, with the sole acquisition of data, the organization will never reach the desired targets (Harrison et al., 2019).

Essential for organizations to exploit the potentialities of BD, data management, also referred to as data governance, is defined by Data Management International (DAMA) as the process of planning, implementing, and overseeing strategies and measures to manage, protect, control, and maximize the value of data and information assets throughout their entire lifecycles (Earley & Henderson, 2017). DAMA proposes a framework that divides data management into 11 functional areas, shown in Figure 1.

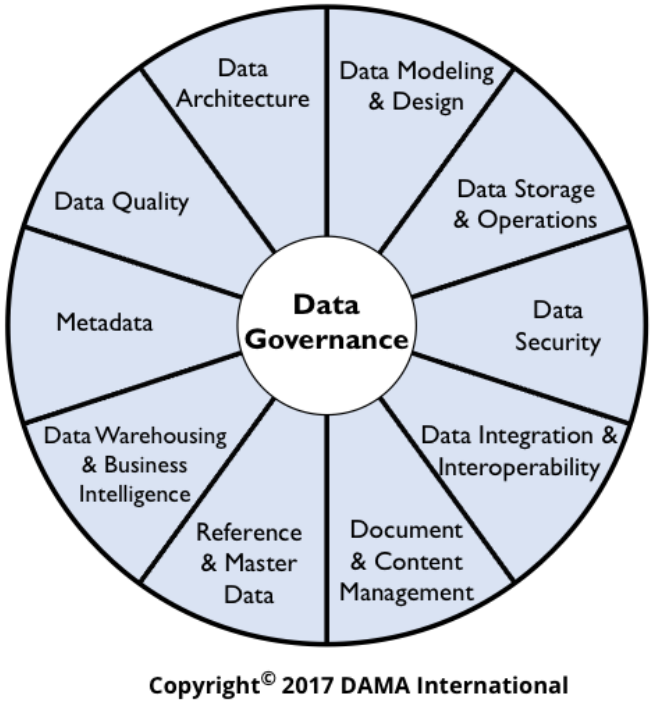


Figure 1 - DMBOK Wheel Framework (Earley & Henderson, 2017)

[2] https://www.oracle.com/se/big-data/structured-vs-unstructured-data/

At this moment in time, most of the organizational efforts are toward the improvement of data analytics, while data management is mostly overlooked (Harrison et al, 2019). The thesis will concentrate on one of the 11 functional areas, specifically Data Integration & Interoperability.

## Data integration

Data integration is the problem of combining data residing at different sources and providing the user with a unified view of these data (Lenzerini, 2002). It is an important topic as it improves the usability of data in the organization because a complete unified view is more consistent and accessible than a fragmented view (Madnick et al., 2009; Kumar et al., 2021). Indeed, to efficiently carry out decisions, data needs to be presented in a unified way (Kumar et al., 2021).This concept is also advocated by Kundra (2010), who adds that "true value lies at the intersection of multiple datasets"[3]. Moreover, as highlighted by Berlanga and Nebot (2016), integrating data is a means to implement context outside the organization in the decision-making processes, in order to make the organization and the decision-makers context-aware. By only feeding the decision-making process with internal data, the organization is making the mistake of not considering the environment in which it works, and a crucial piece of the puzzle. Data integration and data quality are very related concepts, as the sources at which data reside (data sources) are characterized by three main kinds of heterogeneities (Batini & Scannapieco, 2016):

- Technological heterogeneities: related to the use of different database products. This kind of heterogeneity is excluded from the subject of the thesis.
- Schema heterogeneities: caused by the difference in data models between sources and by different data representations.
- Instance-level heterogeneities: conflicting values coming from different data sources, that should represent the same object.

By combining the two definitions, a broader one can be stated, that considers both the data and the user perspective, according to which data integration is the process of combining data residing at different sources with the objective of fixing heterogeneities in data and providing the user with a unified view. The goal of unification is to improve the overview of data and obtain rich and valuable information from it. Within the thesis, this is how data integration is referred to.

### The record linkage problem

The record linkage problem is described as the problem of merging multiple datasets in the absence of a unique identifier (key) that connects (matches) records unambiguously (Kaufman & Klevs, 2021). It was first mentioned by Dunn (1946) in a theoretical publication. In practice, methods to tackle the record linkage problem have been a common topic in the literature (the two main examples being Fellegi & Sunter,

[3] https://obamawhitehouse.archives.gov/blog/2010/05/21/datagov-pretty-advanced-one-year-old

1969 and Jaro, 1989). In the literature, it is also known as object identification, when matches are done between unstructured data sources (Batini & Scannapieco, 2016). The record linkage problem refers to instance-level heterogeneities.

A key is a combination of attributes that can uniquely identify each row of the dataset (Wang & Madnick, 1989). This is also called an identifier (Fellegi & Sunter, 1969; Batini & Scannapieco, 2016; Kaufman & Klevs, 2021). For example, the key that defines univocally a person, in Italy, is *codice fiscale* (Tax ID code in English), which is itself a combination of name, surname, date of birth, gender, and birthplace. The key can be either one attribute or a combination of attributes. In the best-case scenario, when integrating two databases, both share unique identifying columns that unambiguously connect observations across sources (Kaufman & Klevs, 2021). However, the best-case scenario rarely actually happens.

To effectively address the record linkage problem, Batini & Scannapieco (2016) propose a 4-step process named Object Identification process:

1. Preprocessing:
   1.1. Standardization: "reorganization of composed fields, data type checks, and replacement of alternative spellings with a single one." (Batini & Scannapieco, 2016, p. 183)
   1.2. Conversion of upper/lower cases: the strings to be compared are transformed to be uniform in terms of upper and lower cases.
   1.3. Schema reconciliation: "activity that must address all conflict that can occur when data under consideration come from disparate data sources." (Batini & Scannapieco, 2016, p. 184)
2. Search Space Reduction: this step has the goal of removing from the search space all the records that cannot be matched together, entirely avoiding the comparison because of their incompatibility.
3. Comparison & Decision:
   3.1. Comparison: Strings are compared using the distance-based comparison function of choice. There are many distance-based functions and techniques to evaluate the distance between two strings. A brief list is provided by the authors:
      - String-based distance functions: the distance is evaluated on the distance between characters in the string. Examples are: Edit Distance (Levenshtein Distance), n-Grams, Jaro Algorithm, Hamming Distance, Smith-Waterman distance.
      - Item-based distance functions: "the distance is evaluated on strings seen as lists of words" (Batini & Scannapieco, 2016, p.185). Examples are: Jaccard distance and Term Frequency - Inverse Document Frequency (TF-IDF).

3.2. Decision: The actual matching between records is decided. Based on parameters determined by the distance-based function used, records are matched with each other.

4. Quality Assessment: Assess if the result is satisfactory. "The decision on actual matching (M) or unmatching (U) of two records can give rise to two types of errors, false positives (FPs) for records declared as M while actually being U and false negatives (FNs, false unmatches) for records declared as U while actually being M." (Batini & Scannapieco, 2016, p. 209). If the process does not yield a satisfactory result, it can be iterated, either changing some parameters of the distance-based function or substituting the function with a new one.

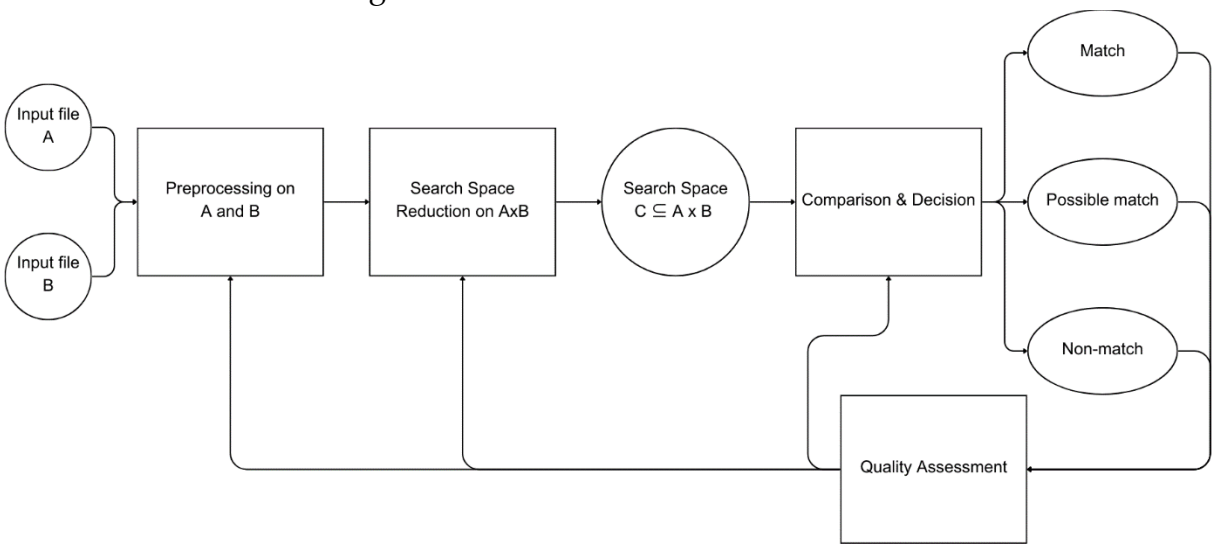


Figure 2 - The Object Identification process (Batini & Scannapieco, 2016)

Kaufman & Klevs (2022) highlight the importance of the collaboration between the computational power of a computer and human intelligence, which is a common theme in many papers (e.g., Norman, 1986; Lazar et al., 2017). This collaboration is also known as Human in the Loop (HITL) when human action is one of the steps of an algorithm, meaning humans are part of the cyclic algorithm.

## 2.6. Open data

Value creation can occur by integrating data from external sources with the traditional data that originates from inside the organization. As Berlanga & Nebot (2016) point out, integrating data is a way to implement context outside the organization in the decision-making processes, with the goal of making the organization and the decision-makers context-aware.

By only feeding the decision-making process with internal data, the organization is not considering the environment in which it works, and so is missing a crucial piece of the puzzle (Berlanga & Nebot, 2016). Indeed, the integration of external data in the decision-making process leads to an increase in volume of information that may enable

better decision-making (Dayal et al., 2009). In fact, the need to integrate external sources has been increasing in importance (Hendler, 2014).

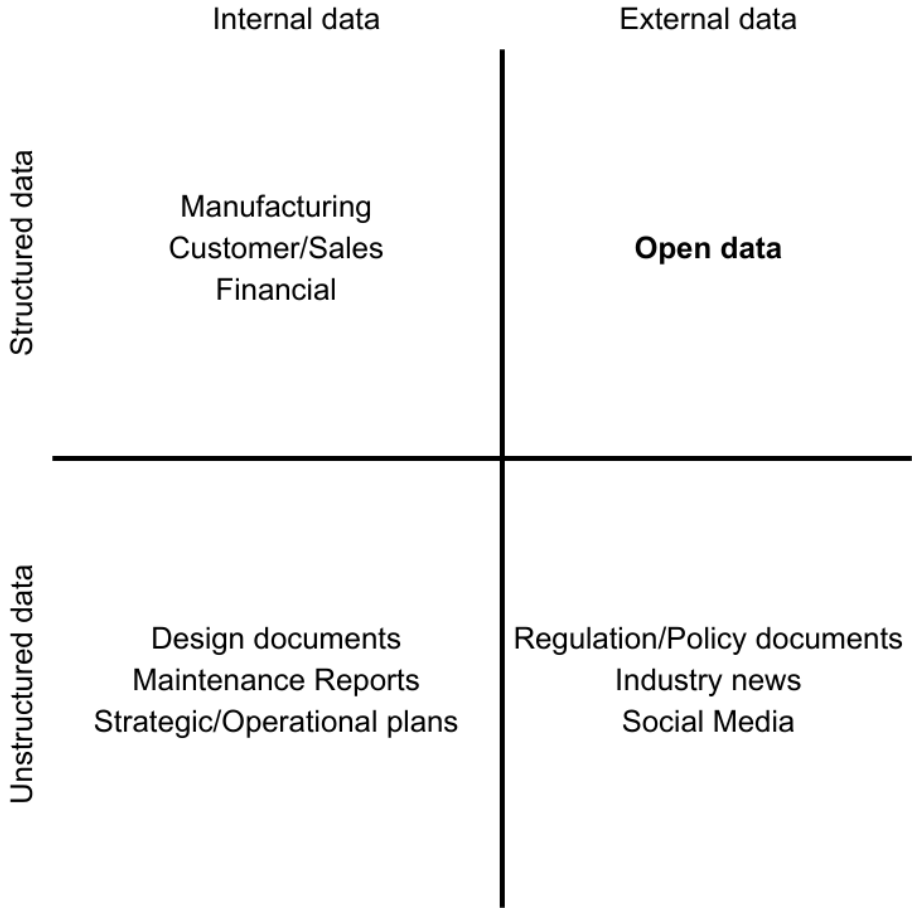


Figure 3 - Examples of data from different sources, readaptation (Hendler, 2014)

Nevertheless, as shown in section 2.5, data integration can be a challenging process. Incorporating these additional data sources is time-consuming and it may lead to delays and inefficient utilization of valuable information (Arputhamary & Arockiam, 2015). Data integration is a valuable activity if the data that is introduced into the organization from the outside is qualitative (Dai et al, 2008; Dayal et al., 2009; Berlanga & Nebot, 2016). The portability of the data is one of the drivers of value creation in the data integration process. Portability means that "the value of data can be realized in a different business/industry context from the one where the data originated." (Pesce et al., 2019, *p.3*). In the next part of this section, the concept of open data is introduced, building upon the preceding discussion of qualitative and portable data originating from external sources.

The definition that the United Nations Educational, Scientific and Cultural Organization (UNESCO) provides for open data is "data that include, among others, digital and analogue data, both raw and processed, and the accompanying metadata, as well as numerical scores, textual records, images and sounds, protocols, analysis code and workflows that can be openly used, reused, retained and redistributed by

anyone, subject to acknowledgement." (UNESCO, 2021, p.140). The Open Knowledge Foundation (OKF) provides a more concise definition, which states: "Open data and content can be freely used, modified, and shared by anyone for any purpose."[4]

## Open data history

The concept of open data dates back from as early as 1942, when Robert King Merton, one of the founders of sociology of science, theorized that science should follow a *communism* (later translated as *communalism*) perspective, meaning that the results of research should be shared and freely accessible to everyone (Merton, 1942). This concept was applied to modern society in 2007, when a set of 8 guidelines for open data was written by thirty Internet activists in a meeting that took place in Sebastopol, California. In 2009, the US Government created their open government data website, initially containing 47 datasets of government information that was previously unavailable to the public. In 2010, the website was populated by over 250000 databases, and it was a success; the value that those newly opened datasets created for the general public was deemed very high thanks to many related initiatives (Kundra, 2010). In 2012, the inventor of the World Wide Web Tim Berners-Lee highlighted the opportunities that opening data could create and urged governments to open data as a means to improve resource efficiency and service delivery to citizens. (Berners-Lee, 2012). In December 2012, the European Union (EU) launched the EU Open Data Portal, a website that was home to public data published by the EU institutions, agencies, and other bodies. In 2013, Manyika et al., through the McKinsey Global Institute published a report on open data called Open data: Unlocking innovation and performance with liquid information. In this report, the focus was put onto the possible economic value that open data could produce, estimated at $3 - $5 trillion, coming from improved efficiency and effectiveness of existing processes and the creation of new products, services, and markets. At this moment, many authors and institutions (e.g., McKiernan et al., 2016; Ziesche, 2023) are still encouraging the opening of data.

Numerous authors, spanning academic literature and reports from various firms and institutions, have underscored the significant potential offered by open data. They emphasize its importance and the opportunities it presents for various sectors and industries. In particular, open data is defined as a tool that can foster participation and social inclusion of their users (citizens). Citizens want to oversee government actions, holding officials accountable, and the opening of government data effectively succeed in increasing transparency between the government and citizens (Bertot et al., 2010; Kundra, 2010; Manyika et al., 2013; Ziesche, 2023). In the science field, the initiative to share scientific data, aligned with Merton's 1942 suggestion, demonstrated remarkable success in addressing the challenges posed by the COVID-19 pandemic. Open

[4] https://opendefinition.org/od/2.1/en/

scientific data reporting played a crucial role in enabling citizens to comprehend the unfolding events at each phase of the pandemic (Ziesche, 2023).

## Open data types

Open data can be of different types, depending on the source:

- Government open data: produced by governments. It is useful for governments to open data to gain the trust of their citizens.
- Scientific open data: produced by scientific institutions. It is important to open scientific data to foster progress (opening increased during the COVID-19 pandemic).
- Private open data: produced by private corporations. This is the typical data that is collected by corporations for profit-maximization purposes. A significant part of these data could be useful "for sustainable development" (Ziesche, 2023, p.22).

Table 1, readapted from Manyika et al. (2013), effectively highlights the differences between data that is completely open and completely closed.

| **Characteristic** | **Completely open data** | **Completely closed data** |
|---|---|---|
| Degree of access | Everyone has access | Access is to a subset of individuals or organizations |
| Machine-readability | Data is available in formats that are easily readable and processable by computers | Data is available in formats that are not easily readable and processable by computers |
| Costs | None | Significant fee |
| Rights | Unlimited rights to reuse and redistribute data | It is forbidden to reuse and redistribute data |

Table 1 - Differences between open and closed data, readaptation of Manyika et al. (2013)

While open data can be published in many different forms, even images and sounds, the most common format is the numerical one (Kalampokis et al., 2016). Open data that comes in that format is also known as Open Statistical Data (OSD) and it is the easiest to process with machines and visualize.

## Open data principles

In the literature, there are many guidelines and rules proposed on open data. The six principles proposed by the Open Data Charter (ODC, 2015) for open data are referenced as they are agreed by many governments (99) and organizations (77) around the world, and they nicely summarize most of the proposals in the literature. The principles were developed by governments, civil society, and experts of the field. They are divided into two parts, with principles 1-4 defining the rules for the publishing of open databases, and principles 5 and 6 stating the goals of open data (ODC, 2015):

1. Open by default: data should be always public. Governments should justify why some datasets are kept closed (e.g., security or privacy reasons).
2. Timely and comprehensive: data should be made public quickly and in a comprehensive manner.
3. Accessible and usable: data should be machine-readable and easy to find. It should also be free to access.
4. Comparable and interoperable: data should be easily comparable between sectors, geographic locations and over time. Data should be structured and standardized to support the integration of multiple datasets. It should also be enriched with accompanying metadata (which is "information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource", NISO, 2004, p. 1).
5. For improved governance and citizen engagement: open data improves the transparency and accountability of governments.
6. For inclusive development and innovation: open data encourages inclusive economic growth and fosters innovation in both governments and the private sector.

Another important set of standards for open data is the FAIR principles (Wilkinson et al., 2016), proposed in 2016 in an article published in Scientific Data by the joint effort Mark D. Wilkinson and 52 other authors, representing academia, business, funding entities, and academic publishers. Later in that same year, those principles were endorsed by the G20 leaders. The FAIR principles are similar in part to the six ODC principles presented below, though the former are related to open data governance (Ziesche, 2023).

FAIR principles for open data:

- Findable: unique and persistent identifiers should be assigned to data. Metadata should be developed. Data should be easy to search for.
- Accessible: data should be retrievable by their identifier "through a standardized communications protocol", which is "open, free and universally implementable" and potentially includes an "authentication and authorization procedure." (Wilkinson et al., 2016, p. 4)
- Interoperable: it is important that the data are represented in a "formal, accessible, shared and broadly applicable language" (Wilkinson et al., 2016, p. 4) and can be integrated with other data through qualified references.
- Reusable: it is relevant that the data are well described by metadata, meet community standards, and that a clear and accessible data usage license, as well as information about the provenance of the data, their collection method, and their maintenance, are provided.

To summarize, the readaptation of a table from Berlanga & Nebot (2016) (Table 2) explains the differences and the similarities between data that comes from inside the

organization (internal source), data that comes from outside the organization (external source), and data that comes from public open databases that follow the guidelines of open data (defined as open source in this table).

| **Internal sources** | **External sources** | **Open sources** |
|---|---|---|
| Structured data | Un-/Semi- structured | Mostly structured |
| High-quality data | Low-quality data | High-quality data |
| Complete information | Incomplete information | Complete information |
| Historical data | Fresh data (real time) | Historical and fresh data |

Table 2 - Readaptation of a table from Berlanga & Nebot (2016)

## Open data portals

Numerous online platforms and portals have the goal of collecting open data to make it accessible to the public. All the countries in the EU have an official government open data portal (e.g., dati.gov.it in Italy, data.europa.eu in the EU) and an official government statistics office (e.g., ISTAT in Italy). The degree of quality of open data is varied in the EU and the Open Data Maturity Report (Assen et al., 2022) shows that there are still significant differences between the maturity of EU Member countries.

There are also many known institutions, not linked to any government, that make open data available to the public. Most of these institutions are non-for-profit and have high-end goals of improving the world (like UNESCO), focusing on vital issues such as world poverty, underdevelopment, world hunger, etc. Open data is used to make people aware of the problems and to advocate for the actions taken by the organization. Examples of these organizations are:

- The World Bank[5]: a 189-countries partnership that has the purpose of fighting poverty and assisting the development of middle- and low-income countries.
- Organisation for Economic Co-operation and Development (OECD)[6]: an intergovernmental organization with 38 member countries that supports initiatives aimed at improving people's economic and social well-being around the world.
- Food and Agriculture Organization of the United Nations (FAO)[7]: a specialized institution of the United Nations (UN) that helps ensure food security in the world.
- Openpolis[8]: Italian non-profit institution that promotes transparent access to public information and protects democracy.

[5] https://www.worldbank.org/
[6] https://www.oecd.org/
[7] https://www.fao.org/
[8] https://www.openpolis.it/

# Chapter 3: Decision-making and data in museums

This chapter aims to provide background knowledge on the challenges faced by museums in the aspects of decision-making and data. In section 3.1, the chapter begins by outlining the historical evolution of the definition of museums. Then, the role of museums in society is discussed, emphasizing their impact and the challenge in demonstrating to stakeholders the social value they provide. Later, in section 3.2, it explores the challenges of PM and reporting in museums, with the related struggles of measuring social value and balancing financial and non-financial indicators. In section 3.3, the unique role of data in museums is illustrated. It discusses BD, data sources and introduces the concept of data integration in museums.

## 3.1. Museum: purpose and stakeholders

To clarify what a museum is and what is the purpose of museums, it is interesting and relevant to discuss the evolution of the definitions of museum.

In 1974, the International Council of Museums (ICOM) defined a museum as:

*"a non-profit making, permanent institution in the service of the society and its development, and open to the public, which acquires, conserves, researches, communicates, and exhibits, for purposes of study, education and enjoyment, material evidence of man and his environment."* (ICOM, 1974).

In this definition, the non-profit nature of museums is underlined. Moreover, the institution is described as "in the service of society", meaning that its activities are done with the objective of providing value - social value - to society in its entirety. In fact, social value refers to the impact that an action, an event, or an organization can have on society. It is a more comprehensive understanding of value that does not only consider monetary benefits but also social benefits (Social Value UK, 2023) (e.g., the acculturation provided by viewing an exhibition). The definition clearly highlights that the tasks that are at the core of museums are the collection and preservation of cultural heritage. Moreover, the primary functions of museums are listed by the United Nations Educational, Scientific and Cultural Organization (UNESCO) as preservation, research, communication, and education (UNESCO, 2015), which is also in line with the 1974 ICOM definition.

The contents of the definition were never updated until, in 2007, the ICOM General Conference adopted a new definition that introduced the concept of intangible heritage. This change was highly requested by the museums that cared not only for tangible examples of cultural heritage, but also for intangibles (Lehmannová, 2020); they were not represented by the current definition as the nature of the cultural

heritage was never specified. The rest of the definition was not changed, apart from minor adjustments of the wording. The updated definition stated that a museum is:

*"a non-profit, permanent institution in the service of society and its development, open to the public, which acquires, conserves, researches, communicates and exhibits the tangible and intangible heritage of humanity and its environment, for the purposes of education, study and enjoyment."* (ICOM, 2007).

In this new definition, heritage is defined as both tangible and intangible, with intangible heritage meaning "the practices, representations, expressions, knowledge, skills – as well as the instruments, objects, artefacts and cultural spaces associated therewith – that communities, groups and, in some cases, individuals recognize as part of their cultural heritage" (UNESCO, 2003, p.4).

As the museum world kept evolving, the definitions followed along. Over time, more and more members of museums agreed that the current definition did not depict the real form and functioning of museums (Lehmannová, 2020). The main change that happened and was acknowledged during those years, is the shift in the role of museums, moving from collection-driven to visitor-centered institutions (Anderson, 2004). Museums used to be just conservation facilities for artworks (Agostino & Arnaboldi, 2021) but this is changing, as museums are concentrating increasingly on the visitors' experience (Welsh, 2005; Giaccardi, 2012; Bonet & Négrier, 2018). This shift is affirmed by the 2022 definition, that cites the need for participation of the community, enhancing the role of museums in society. This has been a concern as there needs to be a balance between visitors' education and entertainment, with the aim of avoiding the *turistification* of museums (Su & Teng, 2018). The new definitions stated that a museum is:

*"a not-for-profit, permanent institution in the service of society that researches, collects, conserves, interprets and exhibits tangible and intangible heritage. Open to the public, accessible and inclusive, museums foster diversity and sustainability. They operate and communicate ethically, professionally and with the participation of communities, offering varied experiences for education, enjoyment, reflection and knowledge sharing."* (ICOM, 2022).

The evolving role of museums, as highlighted in the 2022 definition, goes beyond the traditional notion of conservation and collection-driven logics seen in the 1974 and 2007 definitions. This transformation not only reflects a shift towards a visitor-centered view (Welsh, 2005; Giaccardi, 2012; Bonet & Négrier, 2018; Agostino & Arnaboldi, 2021) but also introduces the crucial role of museums in society as accessible and inclusive fosterers of diversity and sustainability. The definition also highlights how museums are having an impact on society, which becomes part of the very purpose and mission of the museum.

In fact, over the past years, museums, as NPOs, have increased the impact they have on society (Munik et al., 2021). However, this increased impact was countered by a decrease in funding (Elbashir et al., 2022), unrelated to museums' performance. Instead, it was triggered by the strain on public finances, resulting in cuts to government funding (Naylor, 2016). In fact, museums rely on both earned revenues and external funding to survive (Camarero & Garrido, 2009). The duality of income sources causes the museum to depend on different stakeholders (Lindqvist, 2012). To attract funds, museums need to pursue their social mission and create social value and to report to the investing stakeholders, demonstrating how the organization is positively impacting society (Munik et al., 2021). Even though funding decreased continuously in the last years (Elbashir et al., 2022), museums' stakeholders are requesting more and more transparency and reporting regarding the achievement of their social mission and how they are using resources (Ebrahim & Rangan, 2014; Arena et al., 2015; Mehrotra & Verma, 2015; Rainey et al., 2017). So, they need to show they are meeting specific needs of social responsibility and generating social and economic value (Millar & Hall, 2013; Munik et al., 2021). The issue is that, while economic value is easily measured with simple proxies such as profit and revenues, social value is much more difficult to measure, as it is appreciated in the long-term and it is dimensionless (Scott, 2007). In the literature, there have been some studies on the measurement of social value in the long-term (e.g., Williams, 1997; Matarasso, 1997; Lawlor et al., 2008; Reimsbach & Braam, 2023). As it is said in all the ICOM definitions, museums are NPOs, so they do not have profit maximization as their goal, but rather the maximization of the positive impact they have on society. The absence of a profit maximization view has deprived museums of a clear comprehensive common objective (Gstraunthaler & Piber, 2012).

This lack of a clear direction leads to one of the biggest problems with NPOs: they become caught between the requests of different stakeholders (Gstraunthaler & Piber, 2012). NPOs have many stakeholders that differ in how they evaluate the effectiveness of the organization (Bagnoli & Megali, 2011). Moreover, the stakeholders' demands for better reporting may leave them with figures and numbers that do not have a clear interpretation (Gstraunthaler & Piber, 2012).

The stakeholders' demand for financial indicators could cause museums to concentrate too much on the financial side (Tsai & Lin, 2018) and to lose contact with the curatorial and quality side (Gstraunthaler & Piber, 2012; Whelan, 2015; Loach et al., 2017), essentially going against the very nature and mission of the institution. Museums also rely on their own sources of income (earned revenues) and if external funding decreases, then the museums need to start being more profit-oriented to survive (Gstraunthaler & Piber, 2012).

| Stakeholder | Income type | Source of funding |
|---|---|---|
| Government, public | Allocations | Tax transfer |
| Customers | Donations | Private |
| Donor | Earned income | Private |
| Endowment board | Endowment revenue | Private |
| Public and private grant givers | Grants | Tax transfer, donations |
| Distribution board | Lottery revenue | Private + transfer |
| Sponsors | Sponsorship | Private |
| Friend associations | Support resources | Private |

Table 3 - Readaptation of a table from Lindqvist (2012)

Government, private donors, customers, endowment board, grant givers, distribution board, sponsors, and friend associations are all the external stakeholders listed by Lindqvist (2012). In the context of the thesis, when mentioning stakeholders, the focus will specifically highlight the government and customers since they are the main stakeholders of museums, and they represent the two possible sources of funding: tax transfers and private funding (earned income).

## 3.2. Decision-making and Performance Measurement in museums

In the past years, some efforts have been made to develop tools aimed at PM for NPOs. In fact, to capture the difference in objectives between for-profit organizations and NPOs, the tools used to evaluate the performance of these organizations need to be differentiated (Kaplan, 2001; Bagnoli & Megali, 2011). With the intention of assisting not-for-profit organizations in bridging the gap between a vaguely defined mission, strategic components, and day-to-day operations, Kaplan (2001) presented a modified version of the BSC. Even though the BSC is commonly used in the private sector, its benefits can be found also in NPOs (Moullin, 2017). The work of Kaplan was followed by other authors (e.g., Moullin, 2002; Bagnoli & Megali, 2011; Ebrahim & Rangan, 2014; Moullin, 2017). Moullin (2002), following Kaplan (2001), developed the Public Sector Scorecard, identifying factors that are drivers in managing and improving performance in NPOs. The most important change from the original BSC is the change of the focus, from profit to social good (Euske, 2003; Yeung & Connell, 2006). Another contribution is by Bagnoli & Megali (2011), which developed a multidimensional PM framework for NPOs that still considered aspects related to the social efficacy of organization while also considering economic performance and institutional legitimacy.

De Waal (2007) further observes that the implementation of PM systems in NPOs is not a trivial task because of its inherent complexity and the unclear (or difficult to measure) meaning of performance for these organizations. Even frameworks that are tailored to NPOs, often fail because of their lack of specificity (Arena et al., 2015). In

museums, PM is still based on the experience and expertise of the evaluators (Gstraunthaler & Piber, 2012). As the field is still very unexplored, several publications call for more exploration of the application of PM systems to public organizations and NPOs (Ebrahim & Rangan, 2014; Moustaghfir et al., 2016; Arnaboldi et al., 2017; Garengo & Sardi, 2021).

The main issue with museums and NPOs is that they struggle to effectively evaluate performance beyond classical financial performance (Elbashir et al., 2022), whilst the delivery of services should be the main focus (Hoque & Adams, 2011; Arnaboldi et al., 2015). Moreover, they also struggle to measure the social value generated. The measures that aim at being a proxy for the creation of social value can only be computed subjectively by singular cultural institutions through complex methodologies. A primary example of these kinds of indicators is the Social Return On Investment (SROI).

SROI is based on accounting and cost-benefit analysis methods that attach monetary values to benefits for society and the environment in order to demonstrate value creation (Rotheroe & Richards, 2007). It is intended to comprehend, manage, and report on an organization's social, environmental, and economic value (Nicholls et al., 2009). It is focused on the third sector and makes an explicit effort to incorporate stakeholders at every stage (Arvidson et al., 2010) by analyzing how much they value the service they receive (NPC, 2010). To compute a SROI analysis, there needs to be a continuous dialogue between the data analyst and the cultural institution, as the indicator is computed by merging data coming from the balance sheet, surveys addressed to visitors, and surveys to/ interviews with stakeholders. The indicator has been computed in just few cases, with the most notable being:

- Lega del Filo d'Oro, which is a non-profit foundation that has the purpose of rehabilitation of people with deafblindness and psycho-sensory impairment, reported that after the "SROI analysis we can say that Lega del Filo d'Oro generated an average yearly social return of € 37,9 million"[9], which translates in a SROI result of 1.2 (Vurro & Romito, 2019).
- MUS.E Firenze[10], which is a NPO that has as purpose the enhancement of the heritage of the Florentine Civic Museums and of the city of Florence reported a SROI of 3. (Lombardo, 2018). MUS.E was the first case of measuring the social return on investment of a cultural museum association in Italy (Viganò & Lombardo, 2018).

Unfortunately, SROI cannot be computed in a standard way by museums, and thus it is impossible to use it to benchmark the museums. Moreover, the computation of the measure is very complex and resource-intensive while also being very subjective to the

[9] https://www.legadelfilodoro.it/it/chi-siamo/sroi (Translated in English by the author)
[10] https://musefirenze.it/

interpretation of who computes the indicator. However, the presence of that indicator could very well be a determining factor for the improvement of funding, as stakeholders, who seek a comprehensive overview of the performance of a cultural institution (Gstraunthaler & Piber, 2012), want to know how well their money is spent, and SROI does exactly that, by looking at the grants and funding like they are investments, computing the profitability in a ROI-like fashion.

Often, museums lose contact with the curatorial side, focusing only on financial performance measures (Gstraunthaler & Piber, 2012). Moreover, performance measures do not necessarily need to be quantitative and in many cases, qualitative measures are preferable (Moullin, 2017). The big issue with qualitative measures is that they are difficult to evaluate and are very context-dependent, as a one-size-fits-all approach does not work (Chiaravalloti & Piber, 2011). Labaronne & Piber (2020) also argue that quality, in the arts and cultural sector, cannot be measured methodologically, as those evaluations can only come from specialists and insiders. Though they do not make a case against PM in the cultural sector, they emphasize the importance of understanding the limitations coming from such measures.

The implementation of PM in NPOs is very important. PM is the driver of well-informed decision-making (Wholey, 1999; Moynihan, 2005) and subsequently the driver of improved performance (Bisbe & Malagueño, 2012; Abdel-Maksoud et al., 2015; Pollanen et al., 2017). To address the growing demand for performance transparency, reporting tools, such as the dashboard, play an important role. These tools facilitate the communication of performance metrics to stakeholders.

While in the for-profit world, dashboards are used as means to communicate results to higher-ups and the indicators that make up those dashboards are always related to the business' performance, incorporating values and viewpoints in a PM system is much more difficult for NPOs (Maheshwari & Janssen, 2014). This is due to the different nature of the objectives, with maximization of social value for NPOs and maximization of profit for traditional organizations being two very distinct goals. This difficulty is one of the factors that determine the low usage rate for the dashboard in NPOs, even though they could greatly benefit from its introduction (Maheshwari & Janssen, 2014).

Some museums have caught on to the need for visualization with integrated reporting. Integrated reporting is the act of consolidating financial and visitors information into one report that can holistically capture the impact of museum activities (IIRC, 2013; Accurat, 2021). A great example of this practice can be seen in the 2020 Integrated Report of Museo Egizio di Torino (Fondazione Museo delle Antichità Egizie di Torino, 2021). The report followed the guidelines of the International Integrated Reporting Framework, published by the International Integrated Reporting Council (IIRC) in 2013. This framework is based on the concept of integrated thinking, meaning "the active consideration by an organization of the relationships between its various

operating and functional units and the capitals that the organization uses or affects. Integrated thinking leads to integrated decision-making and actions that consider the creation, preservation or erosion of value over the short, medium and long term" (IIRC, 2013, p.53). This definition highlights the important connection between reporting (which is a facet of data visualization) and decision-making.

The effectiveness of integrated reporting directly depends on how embedded integrated thinking is into organizations (IIRC, 2013; La Torre et al., 2019). Integrated thinking, complemented by integrated reporting, plays a crucial role in enabling organizations to plan, manage, and report in a comprehensive manner (VRF, 2022). In the current literature, integrated thinking is considered challenging to translate in practical terms due to its ambiguity and a limited understanding of how it works (Dumay & Dai, 2017; Feng et al. 2017). In fact, to try to shed light on the ambiguity, the Value Reporting Foundation (VRF) published a guide to help organizations transition toward integrated thinking (VRF, 2022). The Foundation also presented guidelines that should be followed to reap the benefits coming from the correct implementation of integrated thinking into the organization. Following the claim that integrated thinking benefits the organization in the long-term (IIRC, 2013, 2021; VRF, 2022), Reimsbach & Braam (2023) confirm that there is a positive correlation between the implementation of integrated thinking and the creation of long-term value (social, environmental, and financial). However, they add that the implementation of integrated thinking could potentially lead to a short-term decline in financial performance.

## 3.3. Data in museums

The connection between the digital world and museums is very interesting and unique. The impact of digital in museums involves both the internal processes of the organizations and the offer to the public.

The usage of data is often a neglected aspect in museums (Agostino et al., 2020). Research on data usage in museums is limited to very few publications (e.g., Romanelli, 2018; Pesce et al., 2019; Agostino et al., 2020). Romanelli (2018) suggests that museums should increasingly invest in data-driven innovation by introducing and managing BD to develop Intellectual Capital (IC) to create value. Even when museum directors are committed to BD, the issue is that there is a lack of skilled personnel who have data analysis competencies (Agostino et al., 2020), and thus the implementation of BD becomes impossible. In fact, Romanelli (2018) promotes a comprehensive data-driven innovation that should happen at organizational, strategic, and human levels. In fact, humans are as one of the main factors (together with capital investments and creative ideas), that are needed to allow the deployment of a data-driven innovation (Chen & Zhang, 2014). Other authors agree that museums should develop a data-driven strategy as a pattern of innovation (Parmar et al., 2014; Castelnovo, 2017). Günther et al. (2017) add that the advantages of BD can only be encountered when organizations are realigned entirely around BD.

## Data sources: digital technologies in museums

The use of digital technologies in museums has been encouraged by ICOM and the World Federation of Friends of Museums (WFFM) in their jointly written Declaration of Funchal (ICOM & WFFM, 2018). Many authors have argued that museum professionals increasingly rely on Information and Communication Technologies (ICT) to develop new and innovative management practices (Fopp, 1997; Marty, 2006, 2007; Peacock, 2008; Taormina & Baraldi, 2021). Moreover, there has also been research on the positive impact of the use of ICT to support communication and mediation between the museum and the visitors (Kéfi & Pallud, 2011), to enhance visitor engagement (Bertacchini & Morando, 2013; Cerquetti, 2016, Kassahun Bekele et al., 2018; Romanelli, 2018) and to improve visitor experience (Othman et al., 2011; De Bernardi et al., 2019). Currently, ICT, the Internet, and social media are causing a transformation of museums' business models. This transformation is made possible by the growing adoption of Internet of Things (IoT) smart objects and technologies (Camarero & Garrido, 2012; Vicente et al., 2012; Solima, 2016). Immersive technologies—a collective term that includes augmented-, virtual-, and mixed-reality technologies (AR, VR, and MR), that aim at delivering sensory experiences through diverse combinations of real and digital content—have been used in museums since the mid-2000s (Bekele et al., 2018). In addition to those, other digital technologies, including GPS, tagging methods such as Quick Response (QR) codes, Radio Frequency Identification (RFID), Beacons, and a wide range of specialized apps, have proven effective in enhancing the innovative experiences offered by museums (De Bernardi et al., 2019).

Digital technologies generate a vast quantity of data (Quach et al., 2022) that needs to be managed effectively to harness its potential benefits (Pieterson & ICF, 2017). Thanks to visitors' interactions with physical objects (through IoT, AR, VR, GPS, QR, RFID, beacons, and specialized apps), museums have access to a diverse range of information that can be leveraged to obtain valuable insights (Wecker et al., 2015). While there is an ongoing discussion on the positive effects of digital technologies on visitors (Kéfi & Pallud, 2011; Othman et al., 2011; Bertacchini & Morando, 2013; Cerquetti, 2016; De De Bernardi et al., 2019 are just some examples), the same cannot be said for the literature on the exploitation of data generated by those digital technologies (Chianese & Piccialli, 2018). These data not only can provide a better understanding of visitor behaviors and preferences but could also allow museums to create personalized and immersive cultural experiences (Veron & Levasseur, 1983; Wecker et al., 2015), following the new visitor-centered paradigm.

However, most cultural institutions remain conservative about adopting digital innovation due to financial and administrative problems, lack of vision and unstructured strategy, resistance to change, time, and costs (Gombault et al., 2016). Moreover, this conservative behavior is also affected by the lack of knowledge and skilled personnel (Bekele et al., 2018). This behavior is commonly encountered in

Italian institutions (De Bernardi et al., 2019; Agostino & Arnaboldi, 2021; Agostino & Costantini, 2022), where there is a clear lack of workers with digital competencies, a lack of attention for digitalization, and a lack of a long-term vision (Agostino & Costantini, 2022). The adoption of digital technologies is limited to some sporadic initiatives, and it is not integrated into a formalized digital strategy. In most cases, some digital technologies are already implemented, but what is missing is a holistic approach to digital innovation (De Bernardi et al., 2019).

## Data sources: social media in museums

Social media has significantly transformed the museum experience by providing active engagement and entertainment for visitors worldwide (Marty, 2007; Vassiliadis & Belenioti, 2017). Social media platforms foster real-time dialogue, enhance visitor engagement, and facilitate cultural interpretation, thereby fostering participative learning (Russo et al., 2007). In fact, social media followers can engage in dialogues with a museum, which would have been a challenging task prior to the surge of social media (Gonzalez, 2017). They also make museums more accessible by expanding their authenticity, breaking traditional boundaries, and appealing to younger audiences (Hume & Mills, 2011; Jafari et al., 2013). However, despite the potential for dialogue and community building on social media platforms, Jamal & Waters (2011) observe that organizations tend to lean towards a one-way communication model. Although the participative approach has proven to be more effective in attracting new visitors, its implementation can be more expensive; regardless of the chosen approach, the significance of maintaining a social media presence for museums is increasingly evident (Gonzalez, 2017). In fact, according to the Observatory, 2021, 79% of Italian museums had a Facebook account and 68% had an Instagram account. Moreover, past results showcase a great upward trend over the years (Figure 4).

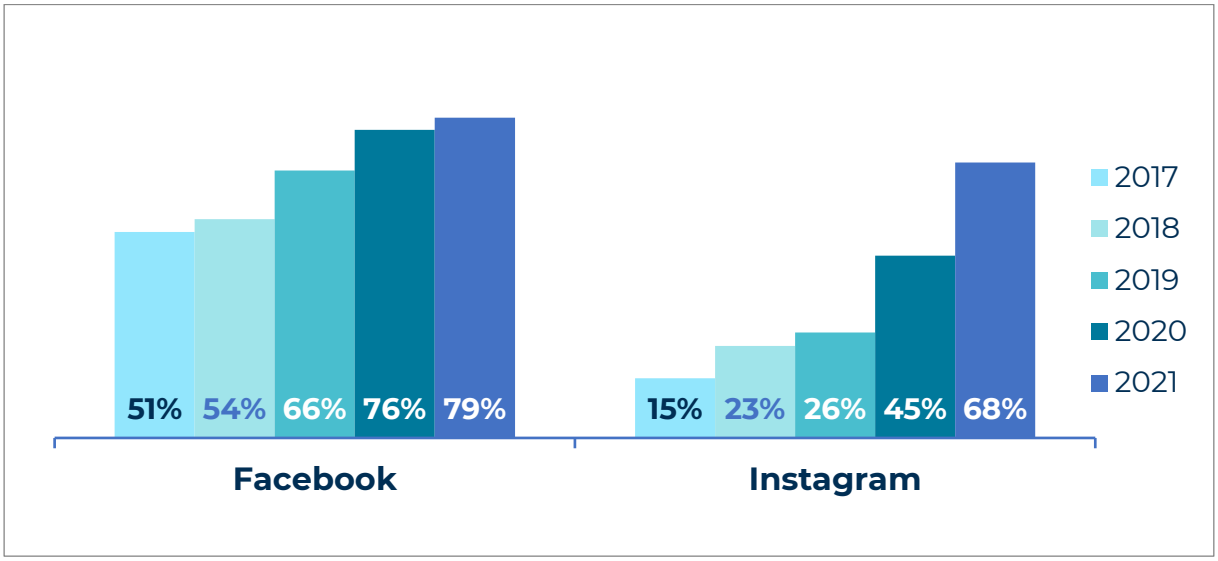


Figure 4 - Italian museums with Facebook and Instagram accounts (2017- 2021), graph from the 2021-22 Survey on the digitalization of Italian cultural institutions by the Observatory

Many authors emphasize the potential of the retrieval of data from social media platforms to analyze visitor interactions, enriching museums' understanding of their audience (Farnadi et al., 2016; Vassilakis et al., 2017; Chianese & Piccialli, 2018). Farnadi et al. (2016) shows the value of the analysis of interacting profiles, while Vassilakis et al. (2017) focus their analysis on social media content (like Facebook and Twitter posts) and other social network-derived data (e.g., Twitter trending topics and Facebook account information). Chianese & Piccialli (2018) combine tweet analysis with geo-referencing, extracting value from data that refers to the location where the tweet was published.

# Chapter 4: Research objective and Conceptual framework

In light of the findings on extant knowledge (Chapters 2 and 3), the first section clearly states the objective of the thesis, which is to show the value residing in the integration of open data for the decision-making process in museums. In the following section, a conceptual framework to address the objective is proposed. First, the four dimensions that revolve around the issue of data-driven decision-making in museums are described: Data, Measuring, Reporting, and Human. Then, a framework that aims at showing the value residing in the integration of open data for the decision-making process in museums is proposed.

## 4.1. Research gaps and objective

**Data management**: the analysis of the literature has shown the importance of data and BD in the cultural heritage field. The usage of data is often a neglected aspect in the field and museum curators are not often committed to data projects (Agostino et al., 2020). Even when museum directors are committed to BD, the issue is that there is a lack of skilled personnel who have data analysis competencies (Agostino et al., 2020). Data collection often lacks a coherent strategy, relying instead on sporadic surveys addressed to visitors when specific information is needed. Romanelli (2018) promotes a comprehensive data-driven innovation that should happen at organizational, strategic, and human levels. The advantages of BD can only be encountered when organizations are realigned entirely around BD (Günther et al., 2017). Data is becoming more and more a vital part of everyday life. Many businesses have realigned around data, while in museums data is still mostly seen as a burden. Because of this conservative behavior (Gombault et al., 2016), museums miss out on insights from data that can lead to improved decision-making (Dayal et al., 2009; Berlanga & Nebot, 2016).

**Measuring**: Measuring performance in museums is much different and more difficult to do than in the for-profit sectors. As a result, the task is overlooked by most museums. PM is still based on the experience and expertise of the evaluators (Gstraunthaler & Piber, 2012) who make decisions that are not based on performance data and insights, but on their subjective perception. Even when PM is applied, it is difficult to effectively evaluate performance beyond financial measures (Elbashir et al., 2022). Performance measures do not necessarily need to be quantitative and in many cases, qualitative measures are preferable (Moullin, 2017). However, qualitative measures are difficult to evaluate (Chiaravalloti & Piber, 2011).

**Reporting**: The act of reporting performance is not common in museums and is still done by only a handful of them. Some of these museums have caught on to the need for visualization with integrated reporting. Integrated reporting is the act of consolidating financial and visitor information into one report that can holistically

capture the impact of museum activities. The act of integrated thinking is defined as "the active consideration by an organization of the relationships between its various operating and functional units and the capitals that the organization uses or affects" (IIRC, 2013, p.53) and it is what constitutes integrated reporting. In turn, integrated reporting influences decision-making, changing the process to integrated decision-making (IIRC, 2013). When integrated thinking thoroughly influences an organization's operations, it improves the flow of information into management reporting, analysis, and decision-making processes, leading to creation of value in the long-term (Reimsbach & Braam, 2023). The coordination of information systems used for communication and reporting, both internal and external, is also improved by this integration (IIRC, 2013). Other reporting tools, like the dashboard, have a low usage rate in museums, even though they could greatly benefit from their introduction (Maheshwari & Janssen, 2014).

**Human**: The human dimension is always preponderant, and it pervades the entirety of the three dimensions explained before. The human side of decision-making is crucial since data is just a tool in the hands of the decision-maker, who is the one making the choices and being accountable for them (Gitelman, 2013). The act of decision-making is inherently human but should be supported by relevant information that reinforces the decisions (Bross, 1953; Provost & Fawcett, 2013; ESS, 2016). It is essential for decision-makers to avoid falling for the dream of perfect information, as striving for it might lead to the disappearance of human judgment from the process (Quattrone, 2016). Nevertheless, there remains a tendency to make decisions without adequately considering performance data (Pfeffer & Sutton, 2006). In the Data Management dimension, the role of humans is crucial as data cannot stand alone, it needs human intervention to become valuable (Gitelman, 2013; Van der Voort et al., 2018). In fact, the main issues concerning data in museums are also related to humans: the lack of commitment by curators and the lack of personnel with digital competencies (Agostino et al., 2020). The data-driven innovation starts from humans (Romanelli, 2018). In the Measuring dimension, humans select and implement PM systems to ensure that decision-making is informed and driven by performance (Wholey, 1999; Moynihan, 2005). While the tool concerns the whole organization, the indicators that make up the PM systems are developed by humans, leveraging their knowledge of the organization and the value drivers that define performance. In the Reporting dimension, humans are the main actors since the tools that are used in reporting are created by humans with the intention of visually explaining the performance of the organization to other humans, either internally or externally. To implement an integrated thinking approach, the people working in the organization need to be aligned on it to foster the creation of long-term value (IIRC, 2013; Reimsbach & Braam, 2023).

In light of the previous considerations, the objective of the research is to study how these aspects connect among each other and how they connect to data integration

through the conceptual framework, which is called Integrated Decision-Making Framework for Museums (IDM Framework for Museums) and that is shown in Figure 5. In the image, the size represents the importance of each dimension, according to the considerations made above.

## 4.2. The Integrated Decision-Making Framework for Museums

The Human dimension is the biggest because it is currently the prominent way to handle the decision-making process in museums. At this moment, the decisions are still made without considering performance data but only personal experience, individual observation, and gut feelings. The Data dimension, together with the Measuring and Reporting dimensions, are much smaller than the Human dimension because they have very little impact on the decision-making process in museums. The absence of a structured data strategy hinders the realization of value that stays hidden and unexpressed in raw data. Moreover, the lack of integration of open data with internal data strips the museums of the much-needed context-awareness to make well-informed decisions.

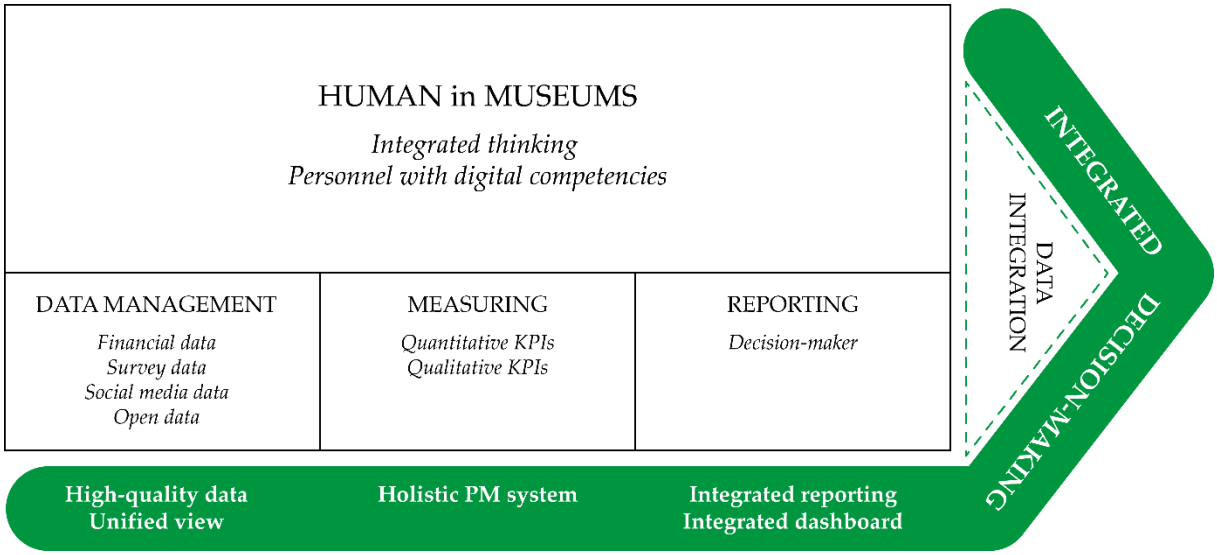


Figure 5 - Integrated Decision-Making Framework for Museums

The Human dimension acts as a complementary support to the various other dimensions within this framework. It operates collaboratively with these dimensions, working in tandem to enhance their effectiveness. The green border signifies the additional value that data integration contributes to the decision-making process. Positioned along this border, beneath each dimension, are the supplementary elements that have been enabled through data integration.

The integration of data has the purpose of improving the decision-making process in museums. Open data are the main resource that make data integration from external sources possible. This is supported by their adherence to the Open Knowledge and

FAIR principles, as discussed in section 2.6. Open databases are a source of truthful data that can be accessed easily and free of charge.

In the next part of this Chapter, the interconnections between data integration and the value added it provides to decision-making are explained.

**Data quality and Unified View**

Data integration is strictly interconnected with the quality aspect of data. Data quality is one of the most important features that humans working should oversee. High-quality data are the foundations to making sensible analyses and developing truthful insights (D'Zurilla & Goldfried, 1971; Janis & Mann, 1977; Batini & Scannapieco, 2016). The data quality issue is encountered in both internal data and external (open) data. While data governance principles should be followed to ensure that internal data are of high quality, the quality of data coming from external sources should be guaranteed by the source (Dai et al, 2008; Dayal et al., 2009; Berlanga & Nebot, 2016) and by the original owners of the open datasets. In fact, transparency and truthfulness are characteristics of open data.

The process of data integration of open data adds value to decision-making as decisions are made taking into consideration the context in which the organization operates (Berlanga & Nebot, 2016). Context-awareness is what enables organizations to respond proactively to changing external dynamics (Berlanga & Nebot, 2016). Without the integration of data from external sources, the decisions are made only considering the internal context, which is often not enough to make well-informed decisions (Papadakis et al., 1998; Berlanga & Nebot, 2016). Moreover, since museums are often dealing with underfunding (Conn, 2010; Moldavanova, 2016), exploiting the free nature of open data can be a sustainable way to achieve well-informed, context-aware, integrated decision-making. Data integration also refers to the integration of data coming from internal sources, which is also an important process. Even though internal data are kept within the organization, if they are not presented in a unified view, it becomes difficult to thoroughly understand and analyze them (Kumar et al., 2021). The lack of data integration can hinder the organization's ability to gain valuable insights from its internal data, potentially resulting in inefficiencies and missed opportunities (Dayal et al., 2009).

**Holistic PM system**

Incorporating values and viewpoints in a PMS is much more difficult (Maheshwari & Janssen, 2014) for museums than for traditional businesses. This is due to the different nature of the objectives, with maximization of social value for museums and maximization of profit for traditional organizations being two very distinct goals. Data integration enables the development of a holistic PM system, allowing the incorporation of measures that originate from diverse data sources (financial

statements, surveys, social media, open data) and of distinct types (quantitative and qualitative).

**Integrated reporting and Integrated dashboard**

Integrated reporting has the objective of presenting data from different sources (performance, financial results, open data) in a consolidated way (IIRC, 2013). This tool is very helpful to communicate results to stakeholders and to the general public (IIRC, 2013; Accurat, 2021). The fact that the report includes data from different sources lets the reader understand thoroughly the environment in which the museum operates and how it operates. In order to develop an integrated report, data should be implemented from many sources in a unified view. Good implementation of integrated reporting is bound to the implementation of integrated thinking (La Torre et al., 2019; VRF, 2022). The museums' employees need to be aligned with integrated thinking principles to enable the correct implementation of integrated reporting (La Torre et al., 2019). Integrated reporting enables the user to holistically understand the performance of the organization, shifting from the traditional approach of strict financial reporting.

The integrated dashboard follows the logic of integrated reporting, as it aims at showcasing consolidated information that comes from different sources (internal data, open data, social media data). This tool should include the developed KPIs and should be used for reporting data internally and externally. The publication of the integrated report does not exclude the creation of an integrated dashboard, and it actually facilitates building the dashboard. The two tools can be used jointly to show information at different granularities and to highlight different topics. Moreover, an integrated report uses visualization techniques together with text explanations, and is typically static, while a dashboard is dynamic and can be interactive.

These concepts interact in an environment in which integrated decision-making happens to generate additional value for the decision maker. The decisions are made considering performance data, which is also integrated with data from external sources (open data). The decisions are supported and measured by KPIs, which are carefully integrated into a PM system. The PM system should be tailored to the specific organization and to the specific decision-maker. The indicators, together with data coming from various sources (both internal and external) are visualized together using tools such as the integrated dashboard and integrated reporting.

The empirical analysis of the thesis (Chapters 5 and 6) shows the effects of integrated decision-making on museums as conceptualized through the Integrated Decision-Making Framework for Museums.

# Chapter 5: Methodology

In this chapter, the methodological procedure followed to build an integrated dashboard for museums is outlined. This involves a data integration process that consists of the integration of 8 datasets, comprising 4 proprietary datasets and 4 open datasets, following the Object Identification process. KPIs are developed to be displayed in the interactive and dynamic integrated dashboard. The intended users of the dashboard are museum managers and supervising museum bodies, such as the Ministry of Culture. The process of data integration, KPI creation, and reporting is undertaken with an integrated thinking logic. The objective of this section is to show the methodological process of data integration of open data, highlighting its potential for the development of useful KPIs and for the generation of better insights in reporting, thereby improving decision-making.

## 5.1. Literature Review approach

The thesis utilizes a comprehensive approach in conducting a literature review focusing on three interconnected topics: PM in museums, decision-making in museums, and the integration of open data to enhance decision-making processes. Recognizing the extensive and diverse nature of the review, the goal was to explore various aspects of these topics.

A specific type of scoping study is implemented, aimed at identifying gaps in existing literature, aimed at drawing conclusions about the overall state of research in those topics. According to Arksey and O'Malley (2005), a scoping study is known for its exploration of broader topics, where various study designs may be fitting. This approach was suitable for this research due to the diverse range of areas investigated, incorporating both academic and non-academic sources such as reports from independent institutions, consulting companies, and proceedings from EU conferences.

The specific type of scoping study utilized involved analyzing existing literature to identify research gaps, subsequently addressing these gaps in the study. Adhering to the framework developed by Arksey and O'Malley (2005), the methodology is structured into five stages: identification of the Research Question, identification of relevant studies, studies selection, data charting, and results report.

### 1. Identification of the Research Question

| Topic | Research Question |
|---|---|
| Performance Measurement in museums | Which are the Performance Measurement systems utilized in museums? |
| Decision-making in museums | How do Performance Measurement systems affect decision-making in museums? |
| Data integration of open data to improve decision-making | How does the integration of open data affect decision-making? |

Table 4 - Identification of the research question

### 2. Identification of relevant studies

The exploration of relevant studies primarily relied on Scopus, chosen for its status as a peer-reviewed database ensuring the inclusion of high-quality papers. Additionally, the search comprised of organizational documents and conference papers due to the limited number of academic studies available and due to their accuracy in answering the research questions. Moreover, the research is enriched by considering materials from consulting firms and European institutions.

### 3. Studies selection

In order to assist in sourcing information pertinent to this research, a set of keywords is selected. The chosen keywords are combined and utilized as a search query within the Scopus database. Table 5 categorizes the specific keywords associated with each respective topic.

| Topic | Keywords associated |
|---|---|
| Performance Measurement in museums | Performance Measurement, Performance Management, Museum |
| Decision-making in museums | Decision-making, Decision-making process, Decision-making cycle, Museum |
| Data integration of open data to improve decision-making | Data integration, Data sources, Open data, Public data |

Table 5 - Topic and associated keywords

In order to further refine the queries and avoid futile papers in the results, in some cases they are limited to the *Arts & Humanities* (SUBJAREA, "ARTS") and the *Business, Management, and Accounting* (SUBJAREA, "BUSI") subjects. For the same reason, the keyword “Account*” is added to some queries.

| Keywords searched | Documents found | Selected |
|---|---|---|
| "Decision-making process" AND "Performance measurement" AND Account* AND "Open data" | 1 | 1 |
| "Decision-making process" AND "Performance management" AND Account* | 27 | 2 |
| "Decision-making process" AND "Performance measurement" AND Account* | 38 | 3 |
| "Performance measurement" AND Museum AND Account* | 9 | 3 |
| "Performance measurement" AND Museum | 25 | 6 |
| "Decision-making" AND "Data integration" AND (LIMIT-TO (SUBJAREA, "BUSI") OR LIMIT-TO (SUBJAREA , "ARTS")) | 59 | 8 |
| "Open data" OR "Public data" AND "Integration" AND (LIMIT-TO (SUBJAREA, "ARTS") OR LIMIT-TO (SUBJAREA ,"BUSI")) | 138 | 7 |

Table 6 - Papers selection

This table shows the queries that resulted in the selections of papers. Considering the extensive range of outputs obtained from the research, a method was employed to select the most representative articles. This involved analyzing solely the titles and abstracts, keeping possibly interesting papers and removing duplicate papers that were already selected in previous queries. After this step, a scrutiny of the Introduction of the candidate papers is conducted, in order to apply a further constraint to the research. Following this process, the final count of relevant papers amounted to 30, forming the foundation for the literature review.

### 4. Data charting

Following the completion of the selection process, all the information about the selected papers is stored in a unified MS Excel document. This facilitated the organization of papers, assigning IDs to papers in order to easily identify them and retrieve information about the query of origin of the paper.

### 5. Results report

Following this process, a comprehensive and structured report detailing the literature was ready. The two graphs below highlight the papers that have a connection between with different topics (Figure 6) and the distribution of papers related to museums (Figure 7).

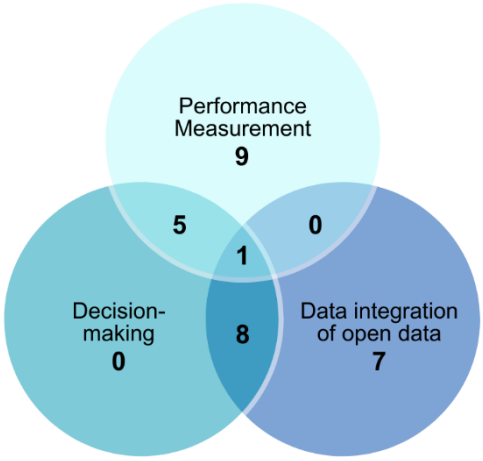


Figure 6 - Venn diagram of number of papers connecting different topics

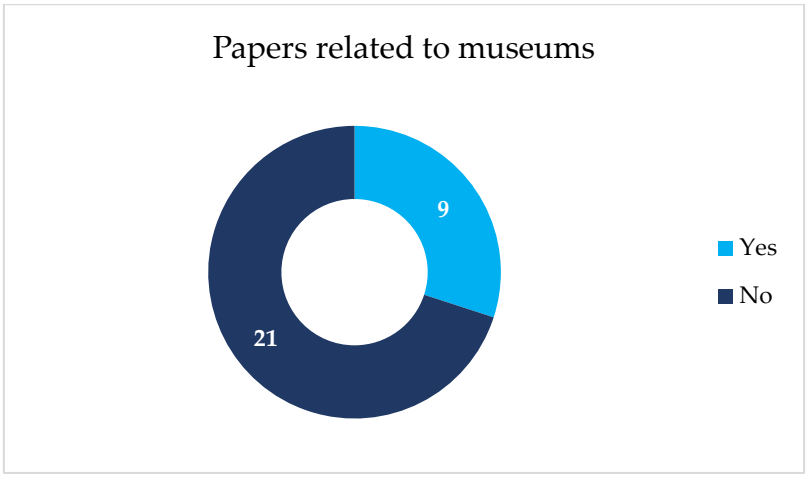


Figure 7 - Ring chart of the distribution of papers related to museums

## 5.2. The Proprietary dataset

In this section, standardization procedures are applied to the four datasets retrieved from the surveys on digital readiness in museums, conducted by the Observatory in the years 2018/2019, 2019/2020, 2020/2021, and 2021/2022. The four datasets are harmonized together using record-matching and clustering techniques.

### Questions mapping

The data used as the foundation of the Methodology originates from four datasets which are composed of all the answers from the Surveys on digital readiness of cultural institutions in Italy from the years 2018/2019, 2019/2020, 2020/2021, and 2021/2022. The surveys are conducted by the Observatory with the purpose of mapping the current state of cultural institutions in Italy, with questions regarding different areas that impact those associations, mostly related to digital innovation in museums. Many of the questions are categorizable under 9 categories, listed below:

- Geographical location: this category of questions inquiries about the region, province, and municipality in which the institution resides. The information is used for sampling the answers in order to have a correct representation of the institutions in the answers, meaning that regions in which there are more cultural associations have more representation in the final dataset.

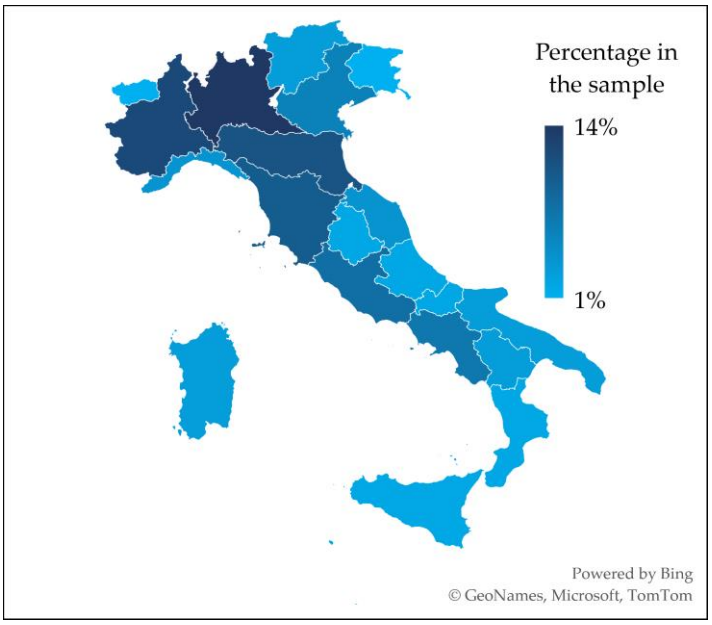


Figure 8 - Sample of museums per region

- Type of institution: the types of institutions surveyed are three: museum, gallery and/or collection, archaeological area or park, and monument or monument complex. There is also an option for *other* types of institutions. The answers to this question are used to filter for *museum, gallery and/or collection* since those type of institutions are the focus of the thesis.
- Numbers of visitors and workers: this category of questions concerns the raw numbers of people that visit the institutions and people that work for the institution.
- Digitalization and digital innovation: this category of questions concerns the state of the digitalization process in the institution.
- Website and social network: this category of questions concerns the online presence of the institution.
- Data collection and analysis: this category of questions concerns the use of data collection and analysis tools by the institution to improve its activities. It also regards the control of visitors' access to the institution.
- Ticket office and reservation of tickets: this category of questions concerns the type of ticket office installed. The questions are about the revenues obtained via the ticket office, the typology of the ticket office (online, physical), the typology

of tickets put out (digital, physical), and the distribution of sales between the channels. Other questions concern the possibility of making a reservation.

- Workers with competencies in the digital sector: this category of questions concerns the current employment of personnel in the institution with competencies in the digital sector and/or are engaged in the digital innovation process.
- Physical collection and digitalized collection: this category of questions concerns the details of the physical collection and the progress of the digitalization process of the collection. Other questions concern the existence of a catalog and the modality of cataloging.
- Activities conducted by the institution: this category of questions concerns the activities that are done by the institution. These are general questions that ask about the importance of some activities and the primary activities considered for future investment.
- Activities offered to visitors: this category of questions concerns the activities that are offered to visitors by the institution. The activities may be offered online or on site (e.g., online tours of the museums, in-presence laboratories, …).
- Marketing to visitors: this category of questions concerns the use of marketing tools to attract more visitors or reach out to previous visitors.
- Network card: this category of questions concerns the benefits offered through the ownership of a network card (i.e., a personal card that grants benefits to the owner; the possible benefits are several: free entry to a museum, reduced price of merchandising, reduced price of transport to get to the museum, …).
- Security: this category of questions concerns the presence of security systems employed to protect the collection and the institution building.

In the surveys, there are additional questions, often specific to a particular survey, that do not align with any predefined categories. The four surveys are very similar in their structure, but they are not identical both in sequence and types of questions asked. Over the years, some questions have been removed and some other questions have been added. Moreover, the number of questions has increased, and the surveys are more specific and more difficult for museum managers to answer. A mapping of the questions year by year can be found in Appendix A.1.

## Preprocessing

Before the integration of open data, the four datasets originating from the surveys need to be unified. The unified dataset, containing all the answers from the four surveys, will be referred to as the Proprietary dataset. Each survey represents its own dataset, which will be referred to as *Year* Proprietary survey and *Year* Proprietary dataset (e.g., the 2018 Proprietary dataset). The process of merging the four Proprietary surveys together is called data harmonization, which is very similar to data integration. Data

harmonization is the process of combining diverse data elements, formats, dimensions, and columns into a unified dataset. The definition and the objective of data harmonization are similar to the one of data integration, as they both are done to enhance the interoperability of data (Porter et al., 2014). While there are several models proposed to solve the data harmonization problems (Avillach et al., 2013; Porter et al., 2014; Firnkorn et al., 2015; are some examples), in the context of the thesis, the datasets are harmonized by hand (using Microsoft Excel and Python), since they are acceptably small, thus a manual approach is preferable. The thesis differentiates between data harmonization and data integration based on a subtle distinction: harmonization involves the process of unifying datasets originating from the same source, whereas data integration refers to the process of unifying datasets originating from diverse sources.

### Standardization

In this section, the implementation of step 1.1 of the Batini & Scannapieco (2016) Object Identification process (Figure 2), which is standardization, is applied to the Proprietary datasets. The variables derived from the mapped questions are standardized in order to ensure a working harmonization of the four Proprietary datasets into a unified Proprietary dataset.

In the four Proprietary surveys, several questions and answers are worded differently between the years, even slightly. To obtain a unified dataset, the questions and answers that regard the same concept need to be unified. For example, the question *Does your institution have workers with digital competencies?* is found in all four surveys, and it has two possible answers, *Yes* or *No*. In the first two datasets (the 2018 and 2019 Proprietary) the answers are collected as either *Yes* or *No*, while in the next two surveys (2020 and 2021), the answers are in binary form (ones and zeroes), where 1 represented *Yes* and 0 represented *No*. To correctly merge the four datasets and have a coherent set of answers, the ones and zeroes need to be converted to *Yes* and *No*.

| | **Workers with digital competencies (binary variable)** | **Workers with digital competencies** |
|---|---|---|
| Museum1 | 1 | Yes |
| Museum2 | 0 | No |
| Museum3 | 1 | Yes |

Table 7 - Harmonization example

The harmonization of the four answers datasets is conducted in Microsoft (MS) Excel and Python. MS Excel is selected due to its adaptability in addressing a wide range of diverse problems and challenges thanks to its easy visualization. Its flexibility makes it a preferred choice for handling various types of issues efficiently. Python is chosen due to its user-friendly interface, vast array of libraries, and the readability of its code compared to other programming languages. Specifically, the software utilized is Jupyter Notebook, a part of the Anaconda library, which operates on the Python

programming language. The list of the functions and features used to assist the harmonization process can be found in Appendix B.1.

After applying a filter on the *Type of institution* question to the complete dataset encompassing four years of responses, only the records categorized under *Museum, collection, and/or art gallery* are retained.

By using a pivot table, it is determined that there are 17 questions repeated across all four surveys. Among them, 5 are related to the identification of the institution, 8 are categorical (multiple choice), and 4 are semi-numerical (multiple choice between different ranges). These are the questions that will be standardized.

The subsequent step involves identifying the question ID assigned to each question corresponding to the survey. This association facilitates the easy retrieval of questions from the respective surveys within the response dataset at a later stage. Table 8 shows the questions that are repeated over the four years, identifying them by their respective question ID.

| Question name | Numerical | Multiple - Choice | Identification | 2018-19 | 2019-20 | 2020-21 | 2021-22 |
|---|---|---|---|---|---|---|---|
| % ticket revenue divided between channels | ✓ | | | Q33 | Q26 | Q34 | Q33 |
| Ticket office | | ✓ | | Q27 | Q22 | Q29 | Q28 |
| Denomination | | | ✓ | Q4 | Q2 | Q2 | Q3 |
| Digitalization of the collection | ✓ | | | Q40 | Q34 | Q19 | Q19 |
| Ticket office revenues | ✓ | | | Q28 | Q23 | Q30 | Q29 |
| Identification | | | ✓ | Q3 | Q1 | Q1 | Q1 |
| Methods of visitor access control | | ✓ | | Q37 | Q29 | Q36 | Q35 |
| Digital innovation plan | | ✓ | | Q13 | Q7 | Q39 | Q39 |
| Workers with digital competencies | | ✓ | | Q44 | Q36 | Q41 | Q40 |
| Data collection on visitors | | ✓ | | Q24 | Q18 | Q21 | Q23 |
| Region | | | ✓ | Q5 | Q38 | Q43 | Q48 |
| Incumbent (Soggetto titolare) | | | ✓ | Q8 | Q41 | Q45 | Q50 |
| Technologies available | | ✓ | | Q26 | Q32 | Q38 | Q38 |
| Number of visitors | ✓ | | | Q9 | Q4 | Q28 | Q27 |
| Type of ticket office | | ✓ | | Q29 | Q24 | Q32 | Q31 |
| Type of institution | | | ✓ | Q7 | Q40 | Q44 | Q49 |
| Workers with digital competencies | | ✓ | | Q44 | Q36 | Q41 | Q40 |

Table 8 - Questions repeated over the years

In the next part of the section, a detailed explanation is provided regarding the standardization process applied to each of the *Numerical* and *Multiple-Choice* variables. The *Identification* questions do not need standardization as they are already

standardized. The transformations involve the handling of Not Assignables (NAs), format standardization, and answers standardization. These transformations are detailed below.

**What is the total revenue from tickets (from the online and/or physical ticket office)?**

In the first two datasets (2018/19 and 2019/20), there is a single column dedicated to the question, where each answer is listed within this column (e.g., Less than 5000€, between 5000€ and 10000€). However, in the latter two datasets (2020 and 2021), there is a presence of multiple columns, each representing a potential answer. This discrepancy in formatting indicates format heterogeneity across the datasets. In fact, the first two datasets follow a *long format* where responses are listed in a single column, whereas the last two datasets follow a *wide format* where each possible response has its own dedicated column[11]. The records indicate a value of 1 if the answer was selected and 0 if it was not selected. This issue is constant for all the questions.

The process of converting the dataset from a wide format to a long format is referred to as *format standardization*. The standardization process involves utilizing the MS Excel XLOOKUP function. This function executes the search for 1 in each row, representing a museum response, and retrieves the corresponding column name where this value appeared. Certain records did not contain any answer to this question. In that case, the museum does not charge anything to visit, hence the answer is transformed into *No fees*.

| Less than 5.000 € | 5.001 - 10.000 € | 10.001 - 20.000 € | 20.001 - 50.000 € | 50.001 - 100.000 € | 100.001 - 500.000 € | 500.001 - 1 million € | 1 - 3 million € | 3 - 5 million € | More than 5 million € | **Long format** |
|---|---|---|---|---|---|---|---|---|---|---|
| 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | **20.001 - 50.000 €** |
| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **Less than 5.000 €** |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10.001 - 20.000 €** |
| 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | **10.001 - 20.000 €** |

Table 9 - Revenue from tickets format standardization example

[11] https://towardsdatascience.com/long-and-wide-formats-in-data-explained-e48d7c9a06cb

After solving the format standardization issue, another heterogeneity needs to be solved, which is the standardization of the answers to the question, referred to as *answer standardization*. This is a typical issue that will be commonly encountered during the process of data harmonization. The issue is that, while the questions in the Proprietary surveys are the same, the available options for the answers are different. The unified dataset will contain only the intersection of these answers, meaning that there is going to be a trade-off between the level of detail and the level of harmonization of data.

<table>
<tr><th>2018-2019</th><th>2019-2020</th><th>2020-2021</th><th>2021-2022</th></tr>
<tr><td>No fees</td><td>No fees</td><td>No fees</td><td>No fees</td></tr>
<tr><td>Less than 1.000 €</td><td rowspan="3">Less than 5.000 €</td><td rowspan="3">Less than 5.000 €</td><td rowspan="3">Less than 5.000 €</td></tr>
<tr><td>1.000 - 2.500 €</td></tr>
<tr><td>2.501 - 5.000 €</td></tr>
<tr><td>5.001 - 10.000 €</td><td>5.001 - 10.000 €</td><td>5.001 - 10.000 €</td><td>5.001 - 10.000 €</td></tr>
<tr><td>10.001 - 20.000 €</td><td>10.001 - 20.000 €</td><td>10.001 - 20.000 €</td><td>10.001 - 20.000 €</td></tr>
<tr><td>20.001 - 50.000 €</td><td>20.001 - 50.000 €</td><td>20.001 - 50.000 €</td><td>20.001 - 50.000 €</td></tr>
<tr><td>50.001 - 100.000 €</td><td>50.001 - 100.000 €</td><td>50.001 - 100.000 €</td><td>50.001 - 100.000 €</td></tr>
<tr><td>100.001 - 500.000 €</td><td>100.001 - 500.000 €</td><td>100.001 - 500.000 €</td><td>100.001 - 500.000 €</td></tr>
<tr><td>500.001 - 1.000.000 €</td><td>500.001 - 1 million €</td><td>500.001 - 1 million €</td><td>500.001 - 1 million €</td></tr>
<tr><td rowspan="4">More than 1 million €</td><td>1 - 2 million €</td><td rowspan="2">1 - 3 million €</td><td rowspan="2">1 - 3 million €</td></tr>
<tr><td>2 - 3 million €</td></tr>
<tr><td>3 - 5 million €</td><td>3 - 5 million €</td><td>3 - 5 million €</td></tr>
<tr><td>More than 5 million €</td><td>More than 5 million €</td><td>More than 5 million €</td></tr>
</table>

Table 10 - Revenue from tickets answers before standardization

| **Standardized answers** |
|---|
| No fees |
| Less than 5.000 € |
| 5.001 - 10.000 € |
| 10.001 - 20.000 € |
| 20.001 - 50.000 € |
| 50.001 - 100.000 € |
| 100.001 - 500.000 € |
| 500.001 - 1.000.000 € |
| More than 1 million € |

Table 11 - Revenue from tickets standardized answers

For this question, it is possible to get a very good level of detail by considering the last 3 years; if 2018 is also considered, the information about the museums that earn the most is lost.

**Do you have any workers specialized in digital innovation?**

For this question, there are issues in some surveys because of the lack of answers, as some rows are NAs, showing neither *Yes* nor *No.* Nevertheless, these are very few and are handled according to logic by setting them as *No*. In some cases, museums answer positively to the question but then do not specify which professionals are working. The answers to this question are standardized to be just *Yes* and *No*. The variable is already correctly standardized and needs no transformations, apart from the already seen transformation from wide format to long format (format standardization).

**Is there a ticket office, either online or physical, in the institution?**

In the 21/22 survey, if the respondent answered *No* to this question, it indicated their inability to answer the subsequent question regarding revenue from ticket sales. This is based on the understanding that without a ticket office, ticket sales cannot happen. The variable is already correctly standardized and needs no transformations, apart from the already seen transformation from wide format to long format (format standardization).

**How many people have visited the museum?**

<table>
<tr><th>2018-2019</th><th>2019-2020</th><th>2020-2021</th><th>2021-2022</th></tr>
<tr><td>Less than 5.000 people</td><td>Less than 5.000 people</td><td>Less than 5.000 people</td><td>Less than 5.000 people</td></tr>
<tr><td>5.001 - 10.000 people</td><td>5.001 - 10.000 people</td><td>5.001 - 10.000 people</td><td>5.001 - 10.000 people</td></tr>
<tr><td>10.001 - 50.000 people</td><td>10.001 - 50.000 people</td><td>10.001 - 50.000 people</td><td>10.001 - 50.000 people</td></tr>
<tr><td>50.001 - 100.000 people</td><td>50.001 - 100.000 people</td><td>50.001 - 100.000 people</td><td>50.001 - 100.000 people</td></tr>
<tr><td>100.001 - 500.000 people</td><td>100.001 - 500.000 people</td><td>100.001 - 500.000 people</td><td>100.001 - 500.000 people</td></tr>
<tr><td>500.001 - 1.000.000 people</td><td>500.001 - 1.000.000 people</td><td>500.001 - 1.000.000 people</td><td>500.001 - 1.000.000 people</td></tr>
<tr><td rowspan="2">More than 1.000.000 people</td><td rowspan="2">More than 1.000.000 people</td><td>1.000.001 - 3.000.000 people</td><td>1.000.001 - 3.000.000 people</td></tr>
<tr><td>More than 3.000.000 people</td><td>More than 3.000.000 people</td></tr>
</table>

Table 12 - Museum visitors answers before standardization

The answers need to be standardized. However, since there are no museums that answered *More than 3 million people* in the last two surveys, the answers are standardized over all years by considering *Over 1 million people* as the only option, without losing any information.

| **Standardized answers** |
|---|
| Less than 5.000 people |
| 5.001 - 10.000 people |
| 10.001 - 50.000 people |
| 50.001 - 100.000 people |
| 100.001 - 500.000 people |
| 500.001 - 1.000.000 people |
| More than 1.000.000 people |

Table 13 - Museum visitors standardized answers

Format standardization is also applied to the variable.

**Do you collect data on visitors?**

In the 2018 Proprietary survey, the respondent could only answer one of the three options: *Yes, in digital*, *Yes, in print*, and *No*. In the following surveys, there is a multiple-response option, considering the case where a museum collects visitor data in both paper and digital versions. To solve this answer standardization problem, if an answer includes both digital and paper, it is replaced with *Both*, using a nested XLOOKUP on MS Excel.

| **2018-2019** | **2019-2020** | **2020-2021** | **2021-2022** |
|---|---|---|---|
| No | No | No | No |
| Yes, in digital | Yes, in digital | Yes, in digital | Yes, in digital |
| Yes, by paper | Yes, by paper | Yes, by paper | Yes, by paper |
| | Both | Both | Both |

Table 14 - Answers before standardization

**How do you check the access of visitors?**

| 2018-2019 | 2019-2020 | 2020-2021 | 2021-2022 |
|---|---|---|---|
| Entrance ticket detachment | Entrance ticket detachment | Entrance ticket detachment | Entrance ticket detachment |
| Barcode (gun) | Barcode (gun) on physical ticket | Barcode (gun) on physical ticket | Barcode (gun) on physical ticket |
| | Barcode (gun) on display | Barcode (gun) on display | Barcode (gun) on display |
| QR code | QR code | QR code | QR code |
| Turnstiles or people counting gates | Turnstiles or people counting gates | Turnstiles or people counting gates | Turnstiles or people counting gates |
| Paper sheet | Paper sheet | Paper sheet | Paper sheet |
| Digital sheet | Digital sheet | Digital sheet | Digital sheet |
| No method | No control on visitors' access | No control on visitors' access | No control on visitors' access |
| Other access | Other access | Other access | Other access |
| Other access (specify) | Other access (specify) | | |

Table 15 - Visitor access answers before standardization

For this question, the answers need to be standardized. To do so, the use of barcodes is considered in general terms, without specifying if it was used on a physical ticket or on a display. Moreover, the *Other access (specify)* answer is discarded because it was discontinued in the 2020/21 and 2021/22 surveys. In addition, the *No method* answer, found the 2018 Proprietary survey is transformed to *No control on visitors' access*, which has the same meaning.

| Standardized answers |
|---|
| Entrance ticket detachment |
| Barcode |
| QR code |
| Turnstiles/ people counting gates |
| Paper sheet |
| Digital sheet |
| No control on visitors' access |
| Other access |

Table 16 - Visitor access standardized answers

**Does a digital innovation plan exist?**

This question only requires a transformation from wide format to long format (format standardization).

**Which of these technologies is available in the institution?**

The answers need to be standardized. *Wi-Fi* is one of the available options for the 2018/19 Proprietary survey, while it is not an option in the other surveys. The presence

of Wi-Fi technology within the institution was specifically asked for in the 2019/20 survey, while it was not asked in any other surveys.

The first two surveys (2018/19 and 2019/20) ask whether the technology is currently present, whether the museum is planning to implement it, whether it is not present and the institution will not implement it, and whether it was present in the past and then removed. The last two surveys simply ask about the current existence of the technology in the institution. To standardize the answers, from the 2018 and 2019 surveys only the information about current technologies is kept.

| 2018-2019 | 2019-2020 | 2020-2021 | 2021-2022 |
|---|---|---|---|
| Wi-Fi | Wi-Fi | | |
| Audioguide | Audioguide | Audioguide | Audioguide |
| AR | AR | AR | AR |
| VR | VR | VR | VR |
| QR code | QR code/ Beacon | QR code/ Beacon | QR code/ Beacon |
| Beacon | | | |
| ChatBot | ChatBot | ChatBot | ChatBot/ Virtual Assistant |
| Videogames | Videogames | | |
| NFC | 3D display | 3D display | 3D display |
| LIS Video | Interactive installations | Interactive installations | Interactive installations |
| Blockchain | | App | App |
| | | Touch screen | Touch screen |
| | | Mixed Reality | Mixed Reality |
| | | | Holographics |
| | | | 4D elements |
| | | | Podcast |

Table 17 - Available technologies answers before standardization

| Standardized answers |
|---|
| Audioguide |
| AR |
| VR |
| QR code |
| ChatBot |

Table 18 - Available technologies standardized answers

**What is the percentage distribution of ticket sales from the following channels?**

| 2018-2019 | 2019-2020 | 2020-2021 | 2021-2022 |
|---|---|---|---|
| In situ | In situ | In situ | In situ |
| Website | Website | Website | Website |
| Travel agency and tour operator (online) | Other online websites or apps | Other online websites or apps | Other online websites or apps |
| Tourist guides with online integration | | | |
| Accomodations with online integration | | | |
| *Enti locali* or tourist offices | Other physical channels | Other physical channels | Other physical channels |
| Physical accomodations facilities | | | |
| Travel agency and tour operator (physical) | | | |
| Tourist guides (physical) | | | |
| Other | Other | Other | Other |
| | Other (specify) | Other (specify) | Other (specify) |

Table 19 - Sales channels answers before standardization

The answers of the 2018/2019 need to be standardized. To update the answers from that survey, the percentage values for both online and physical channels are aggregated.

| Standardized answers |
|---|
| In situ |
| Website |
| Other online websites or apps |
| Other physical channels |
| Other |

Table 20 - Sales channels standardized answers

The answers have another issue as there are some NAs. A column is added to check whether the answers were not given. The *No answer* column checks with an IF function if the sum of all channels is 100. If it is not 100, it means that the row is empty. The sum of the answers is always either 100 or 0. So, if the sum is not 100, in that row the *No answer* column will have the value 1. *No answer* refers to museums that do not earn

from tickets because the visits are free, so those institutions are not required to answer the question.

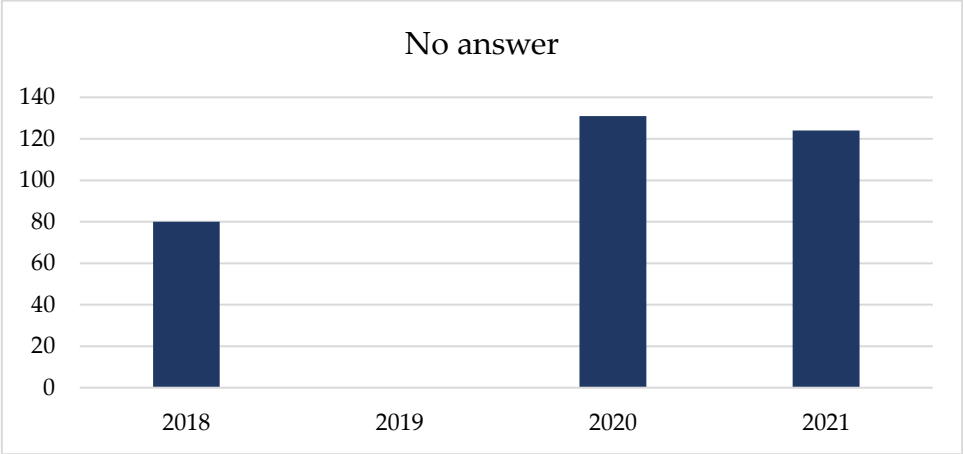


Figure 9 - Distribution of *No answer* over the surveys

Although no prior results for other survey questions have been disclosed, the results for this specific question are presented to illustrate a discrepancy in the answering protocols. This specific case underscores an issue attributed to a change in the way respondents could answer this question in the 2019/20 survey. In fact, the results are unusual for the 2019/20 survey and that is because the respondents could not avoid answering this question in that year (as shown in Figure 9). This makes a harmonized analysis considering that survey difficult, and assumptions need to be made. A logical assumption might be that if a museum answered by putting 100 in *Other* and 0 to all other channels, then most likely that museum has no admission fees. There could also be a museum that simply sells tickets exclusively through platforms that are not mentioned in the other responses, but this would be extremely rare and negligible, as in the other three surveys (2018/19, 2020/21, and 2021/22) this happens only 4 times. In addition, most respondents that put 100% in *Other* specified that museum admission is free, and no one wrote that they sell tickets through a different platform.

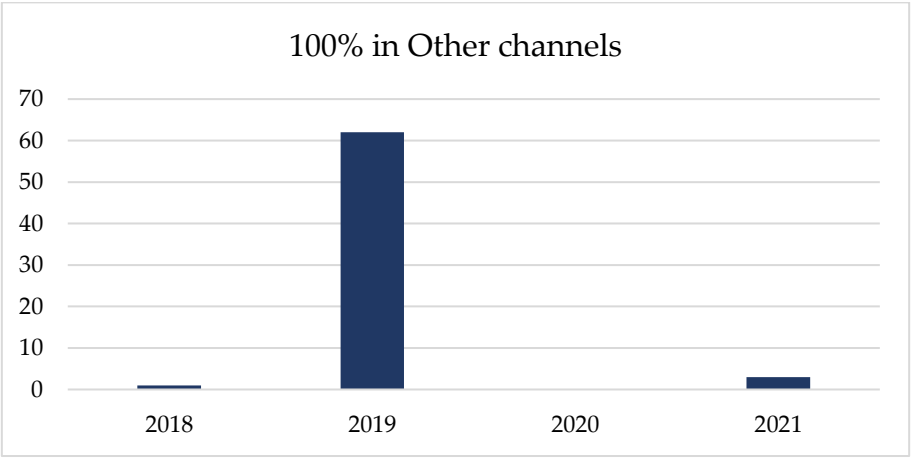


Figure 10 - Distribution of answers with 100 in *Other*

Considering these numbers, the 62 responses that indicated an answer of 100% in *Other* in the 2019/20 survey are changed to *No entrance fee*. In this way, the results are consistent with other years:

| Survey | In situ | Other physical channels | Website | Other online websites or apps | Other |
|---|---|---|---|---|---|
| 2018-2019 | 90% | 3% | 4% | 1% | 1% |
| 2019-2020 | 88% | 7% | 2% | 2% | 1% |
| 2020-2021 | 82% | 8% | 6% | 4% | 1% |
| 2021-2022 | 82% | 8% | 6% | 3% | 2% |

Table 21 - Distribution of Sales channels per survey

**What is the percentage of the digitalization of the collection?**

| 2018-2019 | 2019-2020 | 2020-2021 | 2021-2022 |
|---|---|---|---|
| > 75% | > 75% | > 75% | 100% |
| | | | > 75% |
| 51% - 75% | 51% - 75% | 51% - 75% | 51% - 75% |
| 25% - 50% | 25% - 50% | 25% - 50% | 25% - 50% |
| < 25% | < 25% | < 25% | < 25% |
| No | No | No | No |

Table 22 - Digitalized collection answers before standardization

The answers need to be standardized. To do so, the *100%* answer from the 2021/22 survey are merged with the *>75%* answer.

| Standardized answers |
|---|
| 100% |
| > 75% |
| 51% - 75% |
| 25% - 50% |
| < 25% |
| No |

Table 23 - Digitalized collection standardized answers

**What is the type of ticket office in the institution?**

| 2018-2019 | 2019-2020 | 2020-2021 | 2021-2022 |
|---|---|---|---|
| Detach-ticket from paper with paper accounting | Paper ticket (or receipt) with paper accounting | Paper ticket (or receipt) with paper accounting | Paper ticket (or receipt) with paper accounting |
| Detach-ticket with accounting on electronic system | Paper ticket with accounting on electronic system (or database) | Paper ticket with accounting on electronic system (or database) | Paper ticket with accounting on electronic system (or database) |
| Ticket printed on-site with electronic database | | | |
| Ticket purchasable online and printed at home | Ticket purchasable online and printed at home | Ticket purchasable online and printed at home | Ticket purchasable online and printed at home |
| Ticket purchasable online and not printed | Ticket purchasable online and not printed | Ticket purchasable online and not printed | Ticket purchasable online and not printed |
| Ticket purchasable online with pre-sale option | | | |
| Ticket purchasable online without pre-sale option | | | |

Table 24 - Type of ticket office answers before standardization

The answers need to be standardized as the selection of possible answers is broader in the 2018/19 survey. Two specific answers from the 2018 survey *(Detach-ticket with accounting on electronic system* and *Ticket printed on-site with electronic database*) are standardized into one single answer that is also present in the following surveys: *Paper ticket with accounting on electronic system or database*. Moreover, two answers (*Ticket purchasable online with pre-sale option* and *Ticket purchasable online and not printed*) from the 2018 survey cannot be standardized with the following surveys' answers to this question, and so are removed.

| Standardized answers |
|---|
| Paper ticket (or receipt) with paper accounting |
| Paper ticket with accounting on electronic system (or database) |
| Ticket purchasable online and printed at home |
| Ticket purchasable online and not printed |

Table 25 - Type of ticket office standardized answers

Table 26 shows a summary of the transformation applied to the variables retrieved from the questions.

| **Question** | **Transformations applied** |
|---|---|
| What is the total revenue from tickets (from the online and/or physical ticket office)? | Format standardization, answers standardization |
| Do you have any workers specialized in digital innovation? | NAs handling, format standardization |
| Is there a ticket office, either online or physical, in the institution? | Format standardization |
| How many people have visited the museum? | Format standardization, answers standardization |
| Do you collect data on visitors? | Format standardization, answers standardization |
| How do you check the access of visitors? | Format standardization, answers standardization |
| Does a digital innovation plan exist? | Format standardization |
| Which of these technologies is available in the institution? | Format standardization, answers standardization (twice) |
| What is the percentage distribution of ticket sales from the following channels? | NAs handling, format standardization, answers standardization |
| What is the percentage of the digitalization of the collection? | Format standardization, answers standardization |
| What is the type of ticket office in the institution? | Format standardization, answers standardization |

Table 26 - Questions and transformations applied

### Conversion of upper/lower cases

All characters are capitalized to avoid errors caused by different capitalizations used over the years. This follows step 1.2 of the Batini & Scannapieco (2016) Object Identification process (Figure 2), which is the conversion to upper/lower cases.

### Schema reconciliation

There is an issue related to the presence of excess spacings at the end of the string. This issue concerns step 1.3, which is schema reconciliation, as this issue is solely related to the different formats in which the datasets were saved and encoded. The museums' names in the 2018 Proprietary dataset all share this problem. To deal with it, the excess spacings are removed using Python, specifically the rstrip() function, which removes any set trailing characters (characters at the end of a string).

```
data['Museum'].rstrip()
```

### Search Space Reduction

The search space reduction step (step 2) is not implemented at this moment (it is implemented in Recursive Integration process, p.78). Moreover, search space reduction is done to ensure that the application of the distance-based function is not the bottleneck of the Object Identification process (Figure 2, Batini & Scannapieco, 2016). In this scenario, the dataset's size allows for rapid comparisons, even without significant computational resources, eliminating concerns regarding bottlenecks.

### Enhanced Comparison & Decision and Quality Assessment

The next task to tackle in the harmonization process is to find out how many museums have answered the surveys multiple times over the four years.

For the preliminary analysis, a search is conducted using a pivot table to identify individual museums that appeared multiple times across different years. The initial analysis results are unsatisfactory as certain museums do not appear consistently over the years, despite the naming suggesting they are the same institution. This discrepancy arises from variations in the denomination, even if the differences are slight, observed across the surveys. These differences cannot be solved using a Pivot table. The reasons behind these discrepancies are many. For example, it is common to refer to a museum with multiple names (e.g., *Triennale Milano* is also known as *Triennale di Milano* or simply *Triennale*), or there could be some difference in the way the name was spelled (e.g., *Cubo - Museo D'impresa Del Gruppo Unipol* and *Cubo, Museo D'impresa Gruppo Unipol* are the same museum but because of spelling they are mislabeled).

To address this issue in the analysis, clusters of similar museum names are identified using text analytics. The specific issue to solve in this case is called *fuzzy string matching problem*, which is categorized as a record linkage problem, and it is specific to the case in which the identifier is a string. The implementation of a distance-based function concerns step 3 of the Batini & Scannapieco (2016) Object Identification process (Figure 2), which is the Comparison and Decision step.

To proceed with data harmonization, a human-computer interaction logic is followed, as when dealing with finding similar names, computers are very good at finding many plausible matches that could be correct, but it is only humans that can identify what matches make sense and what matches are correct by a computational point of view, but they do not reflect reality (Batini & Scannapieco, 2016).

The Object Identification process, explained in section 2.5, defines the matching between two datasets on a common key. However, in the context of this harmonization process, the objective is slightly different, as matches must be found within all 4 Proprietary datasets, meaning that every museum in every dataset should be compared with every museum in every other dataset, finding a score for each comparison.

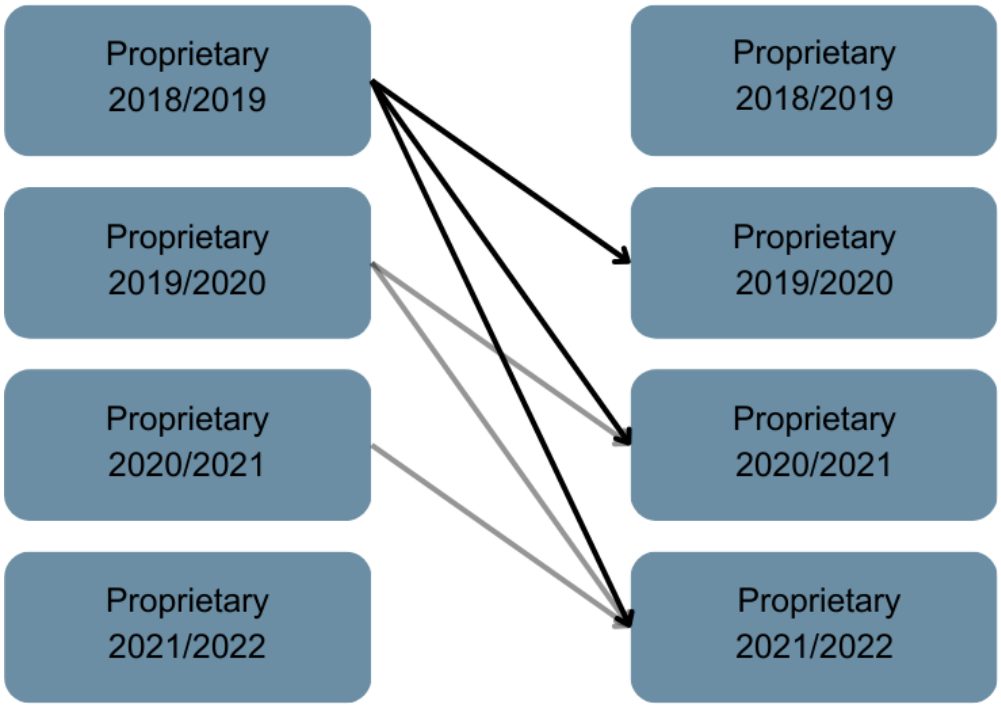


Figure 11 - Comparisons within datasets

While the Object Identification process is extensively followed, the Comparison step is implemented differently (hence the name *Enhanced Comparison*). In fact, after finding comparisons scores using a distance-based function, the scores are clustered to find the denominations that are repeated over the four Proprietary datasets.

During this stage of the process, the decision of what type of distance-based function to implement is in favor of an item-based distance function. These types of functions are great at matching records that do not have many typos, whereas string-based distance functions perform better in the opposite situation (Batini & Scannapieco, 2016). A combination of TF-IDF, cosine similarity and clustering methods is employed to find matches between the four Proprietary surveys.

The initial step involves converting the museum names into a (TF-IDF) format. TF refers to the amount of times a word is present in a *document* (the full name of the museum). IDF refers to the frequency of the word in the entire *corpus* (the complete dataset). The multiplication of TF and IDF is the score that is stored for each word. Then, that information is stored into a *Bag-of-Words* and this *bag* is used later for clustering.

$TF(A, d) = f(A, d)$, where A is the word, d is the document and f is the frequency

$IDF(A, c) = ln(\frac{c}{number\ of\ d\ containing\ A})$, where c is the corpus (total number of documents)

| MUSEO DI MINERALOGIA (MUST - Museo Universitario Sapienza di Scienze della Terra) | | | | | |
|---|---|---|---|---|---|
| **Word** | **TF = f(A,d)** | | **IDF = N/number of d containing A** | | **TFIDF** |
| **della** | 1/11 | 0.09 | LN(1401/193) | 1.98 | **0.18** |
| **di** | 2/11 | 0.18 | LN(1401/566) | 0.91 | **0.16** |
| **mineralogia** | 1/11 | 0.09 | LN(1401/2) | 6.55 | **0.60** |
| **museo** | 2/11 | 0.18 | LN(1401/961) | 0.38 | **0.07** |
| **must** | 1/11 | 0.09 | LN(1401/3) | 6.15 | **0.56** |
| **sapienza** | 1/11 | 0.09 | LN(1401/6) | 5.45 | **0.50** |
| **scienze** | 1/11 | 0.09 | LN(1401/22) | 4.15 | **0.38** |
| **terra** | 1/11 | 0.09 | LN(1401/7) | 5.30 | **0.48** |
| **universitario** | 1/11 | 0.09 | LN(1401/8) | 5.17 | **0.47** |

Table 27 - Example of a TF - IDF transformation

The final dataset is a matrix, in which each row represents a museum, and each column represents a word. In the intersection between word and museum, there's the TFIDF value.

| **Museum** | **archeologico** | **bra** | **chianti** | **del** | **giocattolo** | **museo** | **senese** |
|---|---|---|---|---|---|---|---|
| **MUSEO DEL GIOCATTOLO (BRA)** | 0 | **0.68** | 0 | **0.29** | **0.66** | **0.14** | 0 |
| **MUSEO ARCHEOLOGICO DEL CHIANTI SENESE** | **0.27** | 0 | **0.68** | **0.27** | 0 | **0.12** | **0.62** |

Table 28 - Example of a TF-IDF representation

To transform data into TD-IDF format, the Python function TfidfVectorizer() from the library sklearn.feature_extraction.text was used.

```
vectorizer = TfidfVectorizer()
museum_vectors = vectorizer.fit_transform(museum_names)
```

After obtaining the TDIDF matrix, the second part of the process begins, which is the calculation of the cosine similarity. This is the measure of distance that is used to identify matches. For this calculation, every row is considered as a vector, for example, the first row of Table 28 would be $A = (0, 0.68, 0, 0.29, 0.66, 0.14, 0, 0, 0)$. Then, for each row, the cosine similarity is calculated between that row and every other row. The formula is:

$$cosine\_similarity(A, B) = \frac{A \cdot B}{||A|| \cdot ||B||}$$

This value ranges between -1 and 1, with 1 indicating that the vectors are in the same direction, 0 indicating that they are orthogonal (not related), and -1 indicating that they are in opposite directions. Considering that the starting values are all positive because TF-IDF is the product of two positive numbers, the values for the cosine similarity in this case range from 0 (not related) to 1 (very related). For example, the cosine

similarity between row 1 MUSEO DEL GIOCATTOLO (BRA) and PALAZZO TOZZONI is 0 since they do not have any words in common. The similarity between rows MUSEO DEL GIOCATTOLO (BRA) and MUSEO ARCHEOLOGICO DEL CHIANTI SENESE is bigger than 0 (it is 0.095) since the words *museo* and *del* are in common. To compute the cosine similarity in Python, the cosine_similarity function from the sklearn.metrics.pairwise library was used.

```
similarity_matrix = cosine_similarity(museum_vectors)
```

After the procedure, a similarity matrix is computed. The similarity matrix is a square matrix where each cell (i, j) holds the cosine similarity value between the i-th and j-th vectors (museums). The diagonal of the matrix is only 1's since a vector is always perfectly similar to itself. Because of how it is computed, the matrix is symmetrical, meaning that (i, j) = (j, i). The similarity matrix is important because higher values indicate greater similarity between pairs of museum names, meaning that if two museum names have high similarity, then it is possible that they refer to the same museum, even though the names are not identical.

| | MUSEO CIVICO DI SCIENZE NATURALI MARIO REALINI | MUSEO CIVICO DI SCIENZE NATURALI 'MARIO REALINI' MALNATE | MUSEO CIVICO DI SCIENZE NATURALI 'MARIO REALINI' MALNATE VA | PALAZZO TOZZONI |
|---|---|---|---|---|
| MUSEO CIVICO DI SCIENZE NATURALI MARIO REALINI | **1** | 0.87 | 0.77 | 0 |
| MUSEO CIVICO DI SCIENZE NATURALI 'MARIO REALINI' MALNATE | 0.87 | **1** | 0.89 | 0 |
| MUSEO CIVICO DI SCIENZE NATURALI 'MARIO REALINI' MALNATE VA | 0.77 | 0.89 | **1** | 0 |
| PALAZZO TOZZONI | 0 | 0 | 0 | **1** |

Figure 12 - Example of the Similarity matrix

The color helps visualize the similarity coefficients; the deeper the shade of blue in the cell, the greater the similarity observed between the two museums. In this example, the first two rows (MUSEO CIVICO DI SCIENZE NATURALI MARIO REALINI and MUSEO CIVICO DI SCIENZE NATURALI 'MARIO REALINI' MALNATE) have a similarity coefficient of 0.87 because TF-IDF does not consider punctuation marks and symbols and they differ only by one word. The two museums will most likely be clustered together. The resulting similarity matrix is a 1401x1401 symmetrical and square matrix. The process of hierarchical clustering is applied to the matrix. An

explanation of the concepts of clustering and hierarchical clustering can be found in Appendix B.2.

In the context of the thesis, a hierarchical-based algorithm is employed, specifically with an agglomerative approach.

The goal is to group similar museum names together using hierarchical clustering based on their cosine similarity scores, looking for names of museums in the four surveys. In the first iteration of the process, no constraints are set on the maximum number of members that could be in a cluster, which is 4. This ended up not being a problem since the clusters found are all inhabited by a maximum of four members, apart from one.

The clustering is done in Python, using the AgglomerativeClustering function from the library sklearn.cluster. The AgglomerativeClustering function has three main parameters that need to be set:

- Number of clusters: how many clusters should be found in the data. This can be set to *None* if that number is unknown a priori, as in this case.
- Distance threshold: the linkage distance threshold at or above which clusters will not be merged. 0.5 is a very common number to use for this parameter.
- Type of linkage: refers to the way the distance between clusters is computed. *average* uses the average of the distances of each observation of the two sets.

```
clusterer = AgglomerativeClustering(n_clusters=None,
     distance_threshold=0.5, linkage='average')

clusters = clusterer.fit_predict(similarity_matrix)
```

### First iteration of Enhanced Comparison & Decision and Quality Assessment

The results of the clustering show that in the complete dataset, there are 1087 unique museums, with 852 having only one entry and 235 with at least two.

| Population | Number of clusters |
|---|---|
| 1 | 852 |
| 2 | 170 |
| 3 | 54 |
| 4 | 10 |
| 7 | 1 |

Table 29 - Population and number of clusters after the first iteration

Since they are a feasible number, the accuracy of the 235 multiple-members clusters is checked manually. There are no false positives, meaning museums that are clustered together, even though they refer to different museums. One of the encountered issues is that some clusters are composed of entries that belong to the same survey, like the POLO MUSEALE DELL'ABRUZZO example cited before. This kind of instance-level

heterogeneity may be due to wrongful multiple answers by respondents or to the answers being relative to different parts of the institution which are seen as an independent body but share the same name. This issue is found in just 6 clusters. To address these concerns, the removal of duplicated records is necessary to acquire a single record for each combination of museum and year. The record that is clearly wrong (e.g., the answer is blank) is removed and, if it is not clear which is the wrong one, the record deriving from the answer that was given more recently is kept.

| **Denomination** | **2018-2019** | **2019-2020** | **2020-2021** | **2021-2022** |
|---|---|---|---|---|
| POLO MUSEALE DELL'ABRUZZO | | **7** | | |
| DIREZIONE REGIONALE MUSEI PUGLIA | | | **4** | |
| MUSEO ARCHEOLOGICO STATALE DI ASCOLI PICENO | | 1 | **2** | 1 |
| DIREZIONE REGIONALE MUSEI CAMPANIA | | **2** | 1 | |
| PROVINCIA DI POTENZA | | | | **2** |
| DIREZIONE REGIONALE MUSEI DELLA BASILICATA | | | **2** | |

Table 30 - Instance-level heterogeneities in the Proprietary dataset

A unique name is assigned to each cluster, resulting in a total of 1087 distinct museums. This is accomplished using the XLOOKUP function in MS Excel.

To improve the results obtained, a refinement phase is necessary. This step is cyclical, and it stops only when an acceptable result is achieved. It refers to step 4 of the Batini & Scannapieco (2016) Object Identification process (Figure 2), which is quality assessment.

There are some museums that are not clustered together even though they represent the same museum, which are false negatives. One example of this problem is MUSEO ARCHIVIO DELLA MEMORIA, which is called in three different ways in the surveys: MUSEO ARCHIVIO DELLA MEMORIA, MUSEO ARCHIVIO DELLA MEMORIA - BAGNONE and MAM - MUSEO ARCHIVIO DELLA MEMORIA.

### Second iteration of Enhanced Comparison & Decision and Quality Assessment

To solve this issue, two elements that are not considered in the first iteration are now introduced: the region and the e-mail of the respondent. While the e-mail has the issue that it may change over the years because the respondent is different, the region in which the museum resides does not change with time. The results obtained by adding the region to the clustering can be improved by checking those e-mails that are present in multiple entries during the years but have not been clustered together.

The dataset that is used for this step is the resulting dataset from the first iteration of clustering since no false positives are found. There could be instances of names that are correctly clustered together but missed one or two names that should have been in the same cluster.

To correctly add the region to the name, the name of the regions should be compacted by removing the spacings (Emilia Romagna becomes EmiliaRomagna). This is implemented to avoid the region having too much importance in the clustering and skewing the results, as museums with short names would become very similar to other museums with short names in the same region (e.g., MUSEO LADIN Trentino Alto Adige with CINE' MUSEO Trentino Alto Adige). The same record-matching methods are applied.

The results remained largely unchanged, as 1085 unique museums are identified.

### Third iteration of Enhanced Comparison & Decision and Quality Assessment

In an attempt to further identify additional museums that might belong to the same clusters, the dataset is enriched by the inclusion of the e-mail addresses of the respondents.

The results are better as 1000 unique museums are identified. The clusters showed the presence of false positives (7), with most of these being museums belonging to the same body (polo museale) with similar names (e.g., MUSEO DI FISICA and MUSEO DI IDRAULICA both belonging to Polo Museale La Sapienza). The encountered false positives are moved to their correct allocation by hand. Subsequently, e-mails in single-member-clusters are analyzed, with the goal of discovering more clusters that are not merged together automatically by the clusterer. 19 museums, that should've been clustered together but are not, have been found using this method.

### Harmonization of the clusters

Next, the results of this clustering iteration are matched with those found in the first iteration (by clustering using only the denominations of the museums). This is done to ensure the best overall result, comparing clusters found during the two iterations and keeping the correct ones. The dataset is composed as follows: ID of the museum, number of the cluster with the *only name* clustering, number of members in that cluster, number of the cluster with the *name, region and e-mail* clustering, number of members in that cluster.

| *Name* Cluster | Members | *Region and E-mail* Cluster | Members |
|---|---|---|---|
| 992 | 1 | 978 | 1 |
| 36 | 3 | 133 | 3 |
| 128 | 2 | 26 | 3 |
| 1059 | 1 | 95 | 2 |

Table 31 - Sample of the inital dataset

A check on the discrepancies between the members in clusters is employed. The goal is to find the clusters that have the same members, meaning those that do not need a change, and give a unique cluster ID to those that have different number of members. The clusters that have the most members have priority, so if in the *name* clustering a museum is a member of a 4-members cluster, that cluster is chosen over the one found in the *name, region and e-mail* clustering. The implementation of this task is carried out using MS Excel. The IF function, in conjunction with the CONCAT function, is employed for this purpose. The CONCAT function is used because the clusters' IDs are not unique, so there are duplicate numbers in both the columns related to the IDs. The function is used to add a letter at the end of the ID, signaling the origin of the chosen ID. An *X* is used to signal that the cluster is the same, an *N* if the chosen cluster originated from the *name* clustering and an *E* for the *name, region and e-mail* clustering.

A check that must be applied is related to the region. In each cluster, there should be one and only one region since each cluster represents one museum. This is not the case for 1 record, and so it is fixed.

| *Name* Cluster | Members | *Region and E-mail* Cluster | Members | Difference | **Cluster** |
|---|---|---|---|---|---|
| 992 | 1 | 978 | 1 | 0 | **992_X** |
| 36 | 3 | 133 | 3 | 0 | **36_X** |
| 128 | 2 | 26 | 3 | 1 | **26_E** |
| 1059 | 1 | 95 | 2 | 1 | **95_E** |

Table 32 - Sample of the final dataset

### Resulting dataset

In this way, a total of 951 unique museums are found, with 645 having only one entry and 306 with at least two. These numbers are an improvement from the 1087 unique found before.

| **Population** | **Number of clusters** |
|---|---|
| 1 | 645 |
| 2 | 200 |
| 3 | 87 |
| 4 | 18 |

Table 33 - Population and number of clusters after the second iteration

## 5.3. The Open dataset

In this section, the unified Istat Microdata dataset is crafted using data harmonization techniques. The bulk of the section revolves around integrating and clustering the four Istat Microdata datasets to create a unified dataset of museums, ready for the integration with the Proprietary dataset.

The insights coming from the dashboard are useful if and only if the data that is fed into it is from trustworthy sources (Dai et al, 2008) and coherent with the objective of the dashboard. In the context of the thesis, data come from diverse sources and some transformations need to be applied in order to obtain a usable building block for dashboard construction. This is a scheme of the process of transforming data to obtain a final dataset that is usable for dashboard construction.

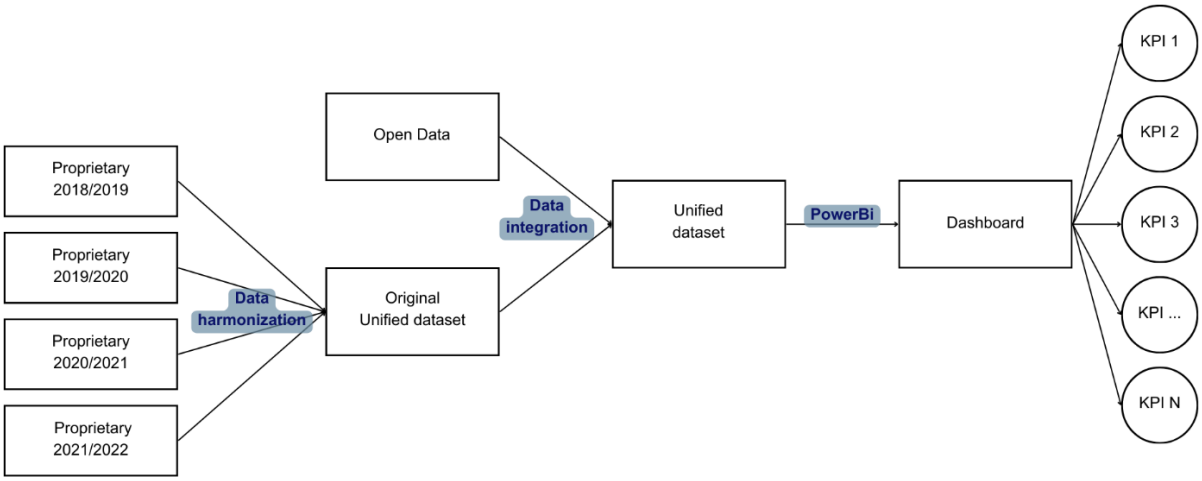


Figure 13 - Open data integration and dashboard building process

Open data are very important for improving the context-awareness of an organization (Berlanga & Nebot, 2016) and it is crucial that open data come from trustworthy sources (Dai et al, 2008). If the source is not credible, then all the analysis and implications will not be credible as well. There are many trustworthy portals (cited before) where open data is stored. Portals that host datasets that relate to museums are researched to identify datasets that could potentially be integrated and add value to the Proprietary dataset.

## Open data portals exploration

The first website explored is the Italian Government's official open data website: *dati.gov.it*. The website is in Italian, and it hosts 61635 datasets. On the website, there is a *search by keywords* option and a *navigate by category* option. There are 13 categories available, and they are not exclusive, meaning that a dataset could belong to more than one category. Every dataset belongs to at least one category. The list of categories can be seen in Figure 14.


Figure 14 - Categories from dati.gov.it

The website gives the user the possibility to search by keywords and confine the results to a category, and it is also possible to search a keyword only under a category. It is also possible to refine the search even more by filtering by source (*Cataloghi* on the website). The datasets can be easily downloaded without the need for identification which is in compliance with the easy accessibility that all open data should be characterized by (per the Open Knowledge and FAIR principles). Moreover, every user can assign a score to the datasets, ranging from 1 to 5 stars. This peer evaluation can help both the user in the selection of the datasets and the creator of the dataset in possibly improving it or changing something if it is wrong. This is the translated list of the categories: Agriculture, fisheries, forestry and food products; Economy and finance; Education, culture and sports; Energy; Environment; Government and public sector; Health; International issues; Justice, legal system and public safety; Regions and cities; Population and society; Science and technology; Transportation.

Since the categories are very broad, the queries could be refined by category. The exploration began by searching for the keyword *Musei* (Italian word for "museums"). The results using these keywords are 250 datasets. The results are censuses of museums from many different sources, from regions to municipalities geographical areas. The datasets are very broad in the variables they are composed of. Some of these datasets contain basic census information about the museums (denomination, address, website), and these data alone are already interesting since in the Proprietary dataset there are no specific addresses present. With this added information, more precise geostatistics could be computed. Other datasets are more specific and contain information about the number of visitors, the revenues, and the artifacts showcased. These datasets could be used to create interesting benchmarking dashboards to compare museums over many different aspects. Some other datasets even show information at different granularities, like visitors and revenues by month, or information regarding long periods of time. These could be useful in assessing the evolution over time. The problem is that this interesting information is scattered between datasets that are specific to different geographical areas, and there is not one summary dataset that includes all the museums in Italy. This issue is what Kalampokis et al. (2016) described as open data fragmentation. They defined it as a situation where

"collections of relevant open data are broken down into many pieces that are not close together" (Kalampokis et al., 2016, p. 34). They also added that, in their own research, they found that fragmentation was actually very prominent in the open data databases rather than just an exception. Additionally, there is a lack of 8 out of 20 regions in the results, which prevents the integration of all datasets to form a comprehensive unified dataset. To summarize, the results are great for putting out an analysis of specific geographical locations, while they are not useful for analyzing the situation in the whole country. For this reason, no datasets from this portal are chosen.

The second portal explored is *istat.it*. The portal is owned by National Institute of Statistics ("Istituto Nazionale di Statistica" in Italian, shortened as "Istat"), which is the most important public body of research in statistics in Italy. This portal is a collection of several datasets related to Italian demographics, economics, and industry. The datasets are compiled by taking the answers from surveys that target individuals, families, and organizations. The answers are then aggregated and filtered by some factors like age and region. Istat also conducts surveys in the museums sector. In fact, on the website, it is possible to access to a census on the Italian museum. The concern regarding the dataset is its lack of updates, with the most recent one dating back to 2015, approximately 8 years ago. This situation contradicts one of the fundamental principles of open data, which emphasizes regularly updated datasets. The data in that database may be inaccurate because of the changes that may have happened during the last 8 years. Moreover, the datasets are aggregated by region, meaning data that is specific to single museums cannot be retrieved.

Istat also carries out surveys that are aimed at collecting microdata, useful for research purposes and for the purpose of the thesis. Istat Microdata is data that "contains information on individuals, households or enterprises"[12]. The Survey On Museums And Other Cultural Institutions: Public Use Micro.Stat Files is a collection of information about 4500 Italian institutes (museums, galleries, archaeological sites and parks, monuments, and historic buildings)[13]. The survey takes place every year and data are uploaded on the Istat website with a two-year lag from the census, meaning that the most recent database downloadable is the 2021 database. The databases that are useful for the purpose of the thesis are those that refer to the years that are represented in the Proprietary dataset, thus the period from 2018-2019 to 2021-2022. The surveys are composed of questions that vary over the years. In the 2018 survey, the questions were 62, in 2019, 27, in 2020, 23, and in 2021, 28. Some questions repeat over the four years, while others are unique or only repeated in some years. The questions are divided into sections based on the nature of them. The number of sections also varies over the years. The answers are organized into datasets. The metadata that accompanies the variables and the questions of the surveys are

[12] https://ec.europa.eu/eurostat/web/microdata

[13] https://www.istat.it/it/archivio/167568

downloaded in the same folder as the dataset. Every possible answer to the questions is codified with an ID and all the metadata about each question is in another document.

The dataset selected is the Survey On Museums And Other Cultural Institutions: Public Use Micro.Stat Files because they allow for the integration of individual museums. The integration will be of high value because each museum will be enriched singularly, adding more information to the Proprietary dataset as many questions that are in the Istat Microdata datasets are not found in the Proprietary dataset.

Before beginning with the integration of this dataset, the same process applied to the Proprietary dataset has to be applied to the Istat microdata dataset. In this Chapter, the harmonized dataset will be referred to as *Istat Microdata dataset* and the dataset representing the answers will be referred to as *Year Istat Microdata dataset*.

## Questions mapping

The questions vary in number and typology over the years, but there are many that are repeated over the years. All the questions are codified with a unique ID, and some of these IDs are repeated over the years, but some are different even though they refer to the same question in different years. To standardize the questions, a more manual approach needs to be implemented. The questions are analyzed manually and those that are repeated over the four years are selected. Istat provides an MS Excel file with all the required metadata only for the 2018 survey. The metadata dataset is composed of 9 columns and 5 are important for the selection task: section, number of the question, name of the variable, description of the variable-question, and format (numeric, text). After 2018, Istat provides metadata in *pdf* form with a three-column table that shows info about the name of the variable, the number of the question, and the description of the question. A variable *section* is added that summarizes the questions that are related to the same area. The surveys are already divided into sections by Istat, that information is just included in the dataset to help visualize similar questions.

The names of the sections vary over the years, however, some of them regard the same topic: official denomination of the institution, type of institution, localization, juridical nature, management, admission to the institution, visitors, and activities. Instead, some types of questions are unique to one survey, like the topic of eco-museums in the 2018 survey or the topic of the COVID-19 pandemic (emergency) in the 2020 survey. Table 34 lists the name of the sections, translated in English from the original Italian.

| 2018 | 2019 | 2020 | 2021 |
|---|---|---|---|
| External variable | External variable | External variable | External variable |
| Name And Location | Name, Location, And Contact Information | Name, Location, And Contact Information | Location, Contact Information, And Opening |
| Typology | Typology | Typology, Legal Nature, And Eligibility Requirements | Typology, Ownership, And Management |
| Ecomuseums | Personnel | Opening And Accessibility | Mode Of Admission And Visitors |
| Legal Nature And Forms Of Management | Legal Nature | Personnel | Personnel |
| Methods Of Admission And Visits | Mode Of Admission And Visitation | Covid-19 Emergency Management | Activities And Services |
| Characteristics And Assets | Web Services | | Accessibility |
| Personnel | Digital Activities | | |
| Financial Resources | Supports, Fruition Services And Activities | | |
| Facilities, Fruition Supports, Activities And Services | | | |
| Relationship With The Territory | | | |

Table 34 - Istat Microdata sections names translated in English from original Italian

In this section, the objective is to create a unified Istat Microdata dataset. In section 5.4, it will be integrated with the unified Proprietary dataset to create a unified dataset that is composed of records and variables originating from both datasets. The chosen key for the harmonization process of the Istat Microdata dataset is a combination of the official denomination of the museum and the address of the museum, keeping an eye on the issue of shared addresses within museums belonging to the same body, representing museum centers (“polo museale” in Italian).

## Preprocessing

While in the Proprietary dataset the standardization step was more difficult, in the Istat Microdata dataset the variables are for the most part already correctly standardized.

### Standardization

The records in the Istat Microdata dataset present the same spacing problem also seen in the Proprietary datasets, so the rstrip() Python function is applied to remove the unnecessary spacing. Then, each of the four Istat Microdata datasets are filtered by the type of the institution, keeping only the museums. This leads to a total number of 14485 records between the four datasets.

### Conversion of upper/lower cases

The transformation applied to the dataset is the conversion the text to upper case, same as the conversion applied to the Proprietary dataset, following step 1.2 of the Object Identification process (Figure 2, Batini & Scannapieco, 2016).

### Schema reconciliation

An issue related to schema reconciliation is that the address of the institution in the 2021 survey presents the province and municipality in numbers instead of letters. The enumeration follows the Elenco dei codici e delle denominazioni delle unità territoriali (List of codes and denominations of geographical units in English), a document by Istat that assigns an ID to each province and municipality. In order to harmonize the information in the 4 surveys, the Elenco dei codici e delle denominazioni delle unità territoriali dataset should be integrated with the Istat Microdata dataset. The integration is completed using the XLOOKUP function on MS Excel and searching for the ID number of the province and for the progressive ID of the municipality, since the unique ID that identifies a single municipality is composed as ID province + progressive ID of the municipality.

Then, the key is assembled by creating a new column that contains in each row the name and the full address of an institution. This is implemented in the workspace by using the CONCAT function which allows to concatenate multiple cells into a single cell. The result is a cell that is composed as follows: Denomination – Municipality – Province – DUG[14] – Toponym. An example of the cell:

MUSEO DEL TESSILE E DELLA TRADIZIONE INDUSTRIALE – BUSTO ARSIZIO – VARESE – VIA ALESSANDRO VOLTA

The subsequent task involves determining the number of unique museums within the dataset. Following this, the museums that share the same name will be clustered

[14] DUG (Denominazione Urbanistica Generica) refers to the generic urban designation that precedes the specific toponym (such as Via, Piazza, etc.) in an address. The toponym is the actual name of the street.

together and then, the dataset will be integrated with the previously harmonized Proprietary dataset.

### Search Space Reduction

By using the *Full Address* column created before, a preliminary analysis can be applied to find the museums that were for sure surveyed 4 years in a row and that never changed their name (and no typo is made about denomination and address).

| Population | Number of clusters |
|---|---|
| 1 | 1423 |
| 2 | 1175 |
| 3 | 1258 |
| 4 | 2752 |

Table 35 - Population and number of clusters before record matching (Istat Microdata)

After this preliminary analysis, the number of unique museums between the four datasets is at least 6608. The objective is to lower this number by finding museums that should be clustered together and are not right now (false negatives). To complete this task, a combination of TF-IDF and cosine similarity methods are applied. The methods will only be applied to those museums that are not in 4-members clusters. So, they are filtered out before proceeding with the clustering, effectively reducing the matching space. After filtering, the dataset is composed of 8793 rows and 5185 unique museums.

### Enhanced Comparison & Decision and Quality Assessment

In the upcoming part of the section, multiple iterations will be conducted to the dataset in order to refine clustering as much as possible. The TF-IDF and cosine similarity method will be applied to the column *Full address*.

#### First iteration of Enhanced Comparison & Decision and Quality Assessment

These are the results obtained after this first iteration of the process.

| Population | Number of clusters |
|---|---|
| 1 | 2286 |
| 2 | 1201 |
| 3 | 1180 |
| 4 | 140 |
| 5 | 1 |

Table 36 - Population and number of clusters after the first iteration (Istat Microdata)

There is one cluster that is bigger than it should be, however, after looking at the museums involved, it is clear that the problem is in the Istat Microdata dataset, rather than in the clustering process. In fact, the five records are all related to one museum (MUSEO DIOCESANO SABINO - PIAZZA MARIO DOTTORI - POGGIO MIRTETO (RI)) and there is a duplicate in the 2019 survey. This is a kind of instance-level heterogeneity that was also found in the Proprietary dataset (e.g., POLO MUSEALE

DELL'ABRUZZO has 7 entries in the 2019 Proprietary survey). By looking at the data regarding the two rows, it is clear that the first of the two (ID: 8002) should be removed as it presents some NAs in the answers, while the second row (ID: 8003) is complete.

Considering that the dataset description from Istat does not explicitly mention consistency in surveyed museums across the years, the obtained results can be deemed acceptable. Nevertheless, to improve the quality and interoperability of the dataset, some checks are implemented.

### Second iteration of Enhanced Comparison & Decision and Quality Assessment

The first check concerns the similarity in the addresses. The museums that are in different clusters but share the same address are the focus of this check (only the geographical address, without considering the denomination of the museum). This check should be done only on museums that are not in 4-member clusters yet. This check is not applied to the Proprietary dataset since there is no information about the addresses of the institutions in all the Proprietary datasets.

| Number of museums with a shared address | Occurrences |
|---|---|
| 14 | 1 |
| 12 | 1 |
| 11 | 4 |
| 10 | 1 |
| 9 | 2 |
| 8 | 8 |
| 7 | 11 |
| 6 | 40 |
| 5 | 18 |
| 4 | 293 |
| 3 | 1032 |
| 2 | 851 |
| 1 | 1689 |

Table 37 - Number of museums with a shared address

The records that need to be checked are those with more than 3 occurrences since the 4-member clusters were already filtered out. These are 379 unique addresses, which translates to 1749 rows, that might need to be unified. A subset that only contains those museums is created. After arranging the dataset by address, only the instances in which museums share the same address and exhibit similar names but are not grouped together through automated clustering are reviewed. There are many false negatives, so museums that should've been clustered together but are not, an example can be seen in Table 38.

| Denomination | Cluster | Survey |
|---|---|---|
| MUSEO BAILO | 1135 | 2018 |
| MUSEO BAILO | 1135 | 2019 |
| MUSEO BAILO | 1135 | 2020 |
| MUSEI CIVICI TREVISO - MUSEO LUIGI BAILO | 3332 | 2021 |

Table 38 - Example of False Negatives

Clustering is implemented on this newly found dataset since 1749 are too many to check manually. For this clustering, the threshold value is adjusted to be more lenient (higher than the typical 0.5), in order to generate more clusters.

By adjusting the threshold, the results improve. However, despite these adjustments, certain museums that the algorithm should group together still remain in different clusters.

### Third iteration of Enhanced Comparison & Decision and Quality Assessment

One potential issue could be the presence of repeated words throughout the whole dataset, such as the words for museum/museums (*Museo* and *Musei*) and civic (*Civico*/*Civici*). The words are removed from the museums' denomination and clustering is applied. To proceed with the removal, the *Find and replace* function is used, and the repeating words are replaced with a blank space (""). The new names are put in another column, named *Full address - no repeating*.

| Population | Number of clusters |
|---|---|
| 1 | 162 |
| 2 | 165 |
| 3 | 196 |
| 4 | 165 |
| 6 | 1 |

Table 39 - Population and number of clusters after the third iteration

These results are found by clustering on the column *Full address - no repeating* and using a threshold value of 0.7. After a careful analysis of the clusters with 4 members it is clear that this iteration of the clustering caught more false negatives, that are now true positives. There are currently 196 unique museums in 3-member clusters, 165 in 2-member clusters, and 162 museums that are alone in their clusters. They are sorted by hand to find if there are some museums that should be clustered together. Most of the museums that are manually edited had a name change in one of the four years which led to the misclustering.

| Population | Number of clusters |
|---|---|
| 1 | 58 |
| 2 | 74 |
| 3 | 116 |
| 4 | 296 |
| 5 | 1 |
| 6 | 1 |

Table 40 - Population and number of clusters after manual filtering

The process applied in the iterations are summarized in these five steps:

- Filtered out the museums that already formed 4-member clusters.
- Clustered the remaining museums (first iteration).
- Clustered the museums that were not in the same cluster but had the same address (second iteration).
- Removed repeating words (third iteration).
- Changed by hand the leftover museums.

This process follows the suggestions from the literature as the human interaction is present thoroughly and it is used to help the algorithm find better results (HITL).

### Harmonization of the clusters

The correct cluster number must be assigned to all museums identified. This is done as already seen for the Proprietary dataset within the first three iterations.

### Results

| Population | Number of clusters |
|---|---|
| 1 | 2062 |
| 2 | 984 |
| 3 | 1029 |
| 4 | 1842 |

Table 41 - Final population and number of clusters (Istat Microdata)

The results show that 5917 unique museums are found, with 1842 found in all 4 years. This is a big improvement from the 6608 unique museums found with the first superficial analysis. Now, there are still some transformations that need to be applied to the Istat Microdata dataset to prepare it for integration with the Proprietary dataset.

## 5.4. The Data Integration

In this section, the integration between the Istat Microdata and the Proprietary dataset takes place. The integration process followed a recursive matching approach that involved iterative comparisons of museum names year by year. Through a comprehensive set of validations and checks, a unified view is obtained, containing information from 599 unique museums across the eight datasets integrated.

The process to be implemented is different from the ones that were implemented in the previous sections. The difference that stands out the most is that the integration is implemented between two datasets that come from different sources, Osservatorio Innovazione Digitale and Istat. This implies dealing with many issues related to the fact that the two datasets were not created to be integrated. As outlined in the methodological section of the thesis, museums frequently undergo name changes, and those changes, even if minor, can pose challenges when clustering museums based on their names. Another difference is in the type of integration that is implemented. During the data harmonization processes the objective is to find similar names in a column of a dataset and cluster them together to find repeating museums. The objective with this integration is to match the denominations of the museums in the Proprietary dataset with the denominations in the Istat Microdata dataset and then merge the two datasets using the identification of the matches as a key.

## Preprocessing

At this moment, neither dataset is ready for data integration.

- The Proprietary dataset is in the wrong format. The records of all the years are one after the other. To prepare it to suit data integration, the records that belong in the same cluster have to be unified and the variables should be divided by year. The dataset is transformed by pivoting the *Year* columns values into separate columns and appending the corresponding variable values as new columns. The transformation changes the dataset from a long format to a wide format. This is done to enable matching of the names between the two datasets. Table 42 explains the current situation.

| Name | Cluster | Year | Var1 | Var2 |
|---|---|---|---|---|
| Museum1 | 1 | 2018 | … | … |
| Museum1 | 1 | 2019 | … | … |
| Museum2 | 2 | 2018 | … | … |
| Museum2 | 2 | 2019 | … | … |

Table 42 - Format of the Proprietary dataset

- The Istat Microdata dataset is already unified, but only for the questions relative to the denomination. So, the first task to apply should be to add the clusters as a new column in each of the four Istat Microdata databases. Then, they are joined together in order to get a result like in Table 43.

| Name | Cluster | Var1_2018 | Var2_2018 | Var1_2019 | Var2_2019 |
|---|---|---|---|---|---|
| Museum1 | 1 | … | … | … | … |
| Museum2 | 2 | … | … | … | … |

Table 43 - Format of the reshaped dataset

For both transformations, Python is utilized due to the large number of variables involved in the process. In this case, the computational power of Python is more important than the visualization power of MS Excel.

The transformation of the Proprietary dataset is done in two steps:

1. Four subsets are created, with every subset containing the answers from one of the four years.

```
Survey_2018 = data[data['Year']=2018]
```

2. The four subsets are merged using the identification of the cluster as the key. The function used is the merge function, which is in the Pandas library. The function's parameters are: right (the databases that should be merged together), on (the key), how (the type of merge to be performed). There are other parameters, but they are not important in this situation. The *how* parameter has five possible choices, and it defaults to *inner*, which means that only the common keys are kept (so the museums that do not appear in all four surveys are excluded). That is why the parameter is changed to *outer*, which keeps all the keys and leaves blank spacings where there is no data. The merges will be done two by two, as shown below in Figure 15.

```
Unified_df0 = pd.merge(Survey2018, Survey2019,
          on='cluster', how='outer')
Unified_df1 = pd.merge(Survey2020, Survey2021,
          on='cluster', how='outer')
Unified_df = pd.merge(Unified_df0, Unified_df1,
          on='cluster', how='outer')
```

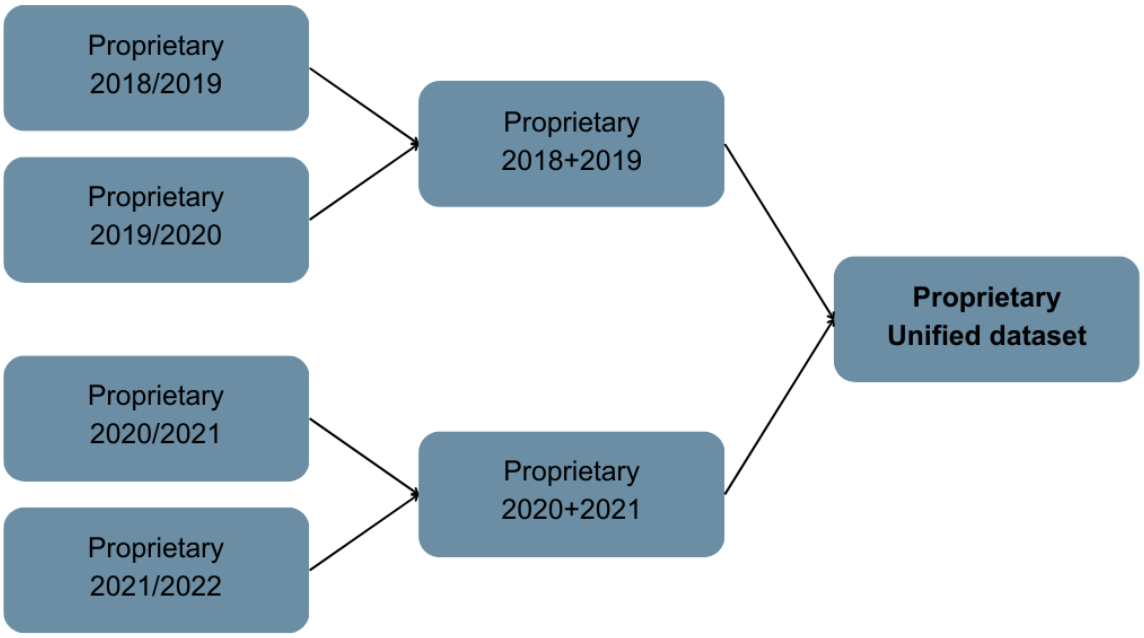


Figure 15 - Transformation of the Proprietary dataset

The transformation of the Istat Microdata dataset is done in 2 steps:

1. A new column is created in each of the Istat Microdata datasets, called *Cluster*. The cluster identifier is assigned to the museum it represents.
2. The datasets are unified as shown in Figure 15.

Now that the datasets are in the correct form, data integration can be implemented.

### Search Space Reduction

The matching will be implemented only between museums that are in the same region. This idea follows step 2 of the Batini & Scannapieco (2016) Object Identification process (Figure 2), which is search space reduction. To implement this in Python, a for-loop is utilized, iterating through the names of the regions. This loop ensures that matching occurs exclusively between museums belonging to the same region.

### Recursive Integration process

The task of matching the records from one dataset to another is not trivial as it needs to be considered that the integration will be done using as a key the name of the museum, which may have typos or variations in the values due to name changes or mistakes in the compilation of the surveys (instance-level heterogeneities). Museums are matched across different years by following a sequential approach, ensuring accurate integration step by step:

1. Create a dataset that contains only the names of the museums from the 2018 Proprietary survey dataset and the 2018 Istat Microdata dataset.
2. Match the names between the two, prioritizing the Proprietary dataset (merging the museums by only keeping the records coming from the Proprietary dataset).
3. Check the results and solve any matching issues.
4. Create a dataset that contains only the names of the museums from the 2019 Proprietary survey dataset and the 2019 Istat Microdata dataset.
5. Remove the records belonging to clusters that were already successfully matched in the past matching.
6. Match the names and repeat from step 3, adding one year in step 4.

This recursive plan of action is very similar to the Adaptive Fuzzy String Matching algorithm proposed by Kaufman & Klevs (2021). The algorithm proposed is a cyclic system that starts with finding matches using a matching algorithm. The results are checked by a human, after which the datasets are updated, and the cycle repeats.

To implement this process practically, both MS Excel and Python are utilized. The first step is done on MS Excel using simple filtering. Then, the dataset is uploaded to Python to proceed with step 2, the matching. To match the records, the logic followed will be similar to the one used for clustering the museums in the Proprietary and Istat Microdata datasets. Since it is very likely that a museum that is in the Proprietary dataset is also in the Istat Microdata dataset (because of the fact that the Proprietary

dataset has fewer museums than the Istat Microdata dataset), a one-to-one matching is implemented.

The chosen methods are still TF-IDF and cosine similarity, but the final clustering is omitted and just the best match will be kept, meaning the match with the highest similarity. Every museum from the Proprietary dataset will be matched with a museum from the Istat Microdata dataset, and the matches that have low similarity will be manually checked to decide if they need to be excluded from the dataset because the match is not correct, meaning that the two names refer to different museums.

### Implementation of Recursive integration and Quality Assessment

The results show that the average similarity is 0.89, which is very high and in fact, 63% of the matches show a similarity of 1 (perfect match) and 82% show a similarity of at least 0.7 which is a standard threshold for matching records. The remaining 18% (62 museums) will be checked by hand, while the 37% that are not a perfect match will also be scrutinized. In the dataset, 48 (14%) museums that should not be matched together are found. This means that either the name is not found correctly, or the museum is not present in the Istat Microdata dataset (which contains 4500 museums in Italy, so it is possible that some museums are not surveyed and databased). These 48 museums are lost and cannot be matched correctly. Subsequently, an attempt will be made to identify the accurate matches for the museums that were excluded during this and the previous four iterations. This situation is an example of the challenges encountered when performing data integration using keys that exhibit differences between the two datasets that should be integrated (instance-level heterogeneities).

Once steps 2 and 3 have been completed, the focus shifts to steps 4 and 5. Having created the dataset in step 4, the subsequent task involves eliminating museums from the records associated with clusters that have already been successfully matched in the 2018 iteration. This process entails identifying the clusters to which the matched museums belong and filtering out these clusters from the 2019 Proprietary dataset. Following this operation, it is observed that 232 out of 310 museums are left. The matching process will be carried out solely on the remaining subset, while the other 78 museums will be excluded from the dataset requiring matching, as they have already been correctly matched.

The results of the matching of the 2019 denominations show an average similarity of 0.79, with 66% of matched museums having a similarity of at least 0.7. These results are worse than the last match, but this is as expected since 78 museums that had a good match in the last matching iteration were removed from this iteration, thus removing museums that would have likely found a match.

Upon reviewing the museums with similarity scores below 0.7, it is discovered that 69 museums, constituting 30% of the total, are not correctly matched. The unmatched

museums are greater than the first iteration. As previously mentioned, subsequent attempts will be made to find matches for the remaining 117 museums (69 from this iteration and 48 from the initial iteration) that were not successfully matched in previous iterations.

The records belonging to clusters that were already successfully matched in the 2018 and 2019 matching are removed. Out of the initial 384 museums, there are still 262 museums that need to be matched.

The results of this matching iteration show an average similarity of 0.73, with 55% having a similarity of at least 0.7. After checking the museums that have a similarity below and slightly above 0.7, 108 (41%) are not correctly matched.

After deleting the entries belonging to clusters that were already successfully matched in the 2018, 2019, and 2020 matching, 214/363 museums that need to be matched are left.

The average similarity for the 2021 iteration is 0.75, with 62% having a similarity of at least 0.7. After a manual inspection, 87 museums (41%) were found to be incorrectly matched.

| Iteration | Number of museums | Average Similarity | Matched | Not matched |
|---|---|---|---|---|
| Match 2018 | 344 | 0.89 | 296 | 48 |
| Match 2019 | 232 | 0.79 | 163 | 69 |
| Match 2020 | 262 | 0.73 | 154 | 108 |
| Match 2021 | 214 | 0.75 | 127 | 87 |

Table 44 - Summary of the iterated integration process

## Leftovers matching

After the iterations, the attention is directed toward the 312 remaining records that were not matched successfully. To reduce this number, the initial step involves searching for these denominations across the entire unified Istat Microdata dataset, instead of matching year by year. The cardinality of the matching process is changed from one-to-one to one-to-many. The difference is shown in Figure 16.

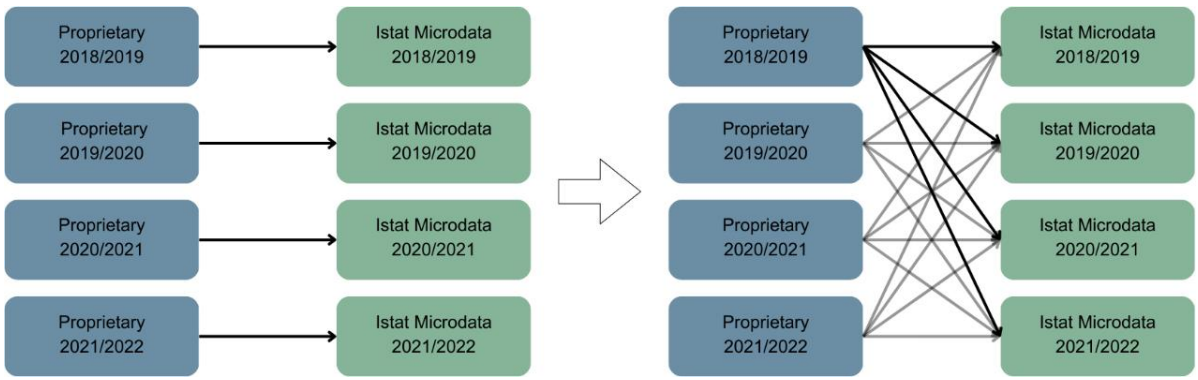


Figure 16 - Leftovers matching scheme

### First iteration of Comparison & Decision and Quality Assessment on the Leftovers dataset

The initial step involves creating a dataset exclusively comprising the remaining museums, which will be referred to as *Leftovers dataset*. This process entails filtering the *Cluster* column to include only entries labeled as *No cluster* (no match) within the dataset containing the matching outcomes. Subsequently, the identical TF-IDF and cosine similarity methodologies, as seen previously, will be applied. The process to be applied is now again comparable to the Object Identification process (Figure 2).

The average similarity found is 0.66, which is very high considering that the matching is on the leftovers records. In fact, there are 131 (42%) museums that have a similarity of at least 0.7 and even 29 (9%) perfect matches. After manually verifying the matching, 229/312 did not match correctly. The outcome is worse than the ones achieved in the integration but is expected, as this process involves museums that likely do not exist in either dataset, making it challenging to achieve a complete matching for all museums. The new leftovers dataset is composed of 229 museums without a match.

### Second iteration of Comparison & Decision and Quality Assessment on the Leftovers dataset: fuzzy matching

To try to reduce this number, a new distance-based function will be implemented: fuzzy matching (approximate string matching). Fuzzy matching is a method that aims at identifying similar strings between two or more datasets. The main difference between this method and the combination of TF-IDF and cosine similarity is that fuzzy matching is used when trying to match two sets of data that may have typos because they are human-made. In fact, fuzzy matching refers to the category of string-based distance functions. For example, the words MUSEO and MUESO are completely different for the TF-IDF cosine similarity method, while they are considered similar and get a score higher than 0 when applying a fuzzy matching logic. Fuzzy matching works by calculating the distance between two strings. One of the most common distances used is the Levenshtein distance (LD), also known as Edit distance, which counts the letters that are *substituted* or *inserted* between one word and another. Substitutions and insertions are called *edits*. The distance is computed as the sum of the minimum edits it takes to change one word into the other. The objective of fuzzy matching is to find matches that minimize the distance. These are three examples that explain LD:

1. The distance between the words MUSEO and MUSEI is 1 since there needs to be one substitution and no insertions (the final *O* of MUSEO is substituted for an *I*).
2. The distance between the words MUSE and MUSEO is still 1 since there needs to be an insertion and no substitutions (an *O* needs to be added at the end).

3. The distance between the words MUESO and MUSEO is 2 since the *E* from word 1 needs to be changed to an *S* and the *S* needs to be changed to an *E*, thus making two substitutions.

These examples show the power of the LD in finding typos, but they do not show the limitation in matching records that are subsets of one another. For example, the two records MUSEI CAPITOLINI and MUSEI CAPITOLINI ROMANI obtain an LD of 7 (the distance includes the spacing), which is very high. For these pairs of records, commonly found in the datasets being used, the combination of TF-IDF and cosine similarity proves to be highly efficient in identifying matches. This approach has been implemented to maximize the identification of matches across the datasets.

Fuzzy matching is implemented in Python using the *fuzzywuzzy* library. The library relies on fuzzy matching and it is equipped with many different functions that fit diverse problems relative to the matching of strings. The main function is the *extractOne* function from the sub-library *process*. *process.extractOne* returns the best match and its similarity score for each entry in the first dataset and the second one. The key parameter of the function is scorer, which determines the way the similarity score is computed for a pair of records. For this analysis, the *token_set_ratio* parameter is selected, as it is the best choice for matching strings with words in different orders.

```
process.extractOne(Proprietary, Istat Microdata,
          scorer=fuzz.token_set_ratio)
```

Unfortunately, the first phase of fuzzy matching is not helpful as the results were very poor, even for those museums that have a very high score. In fact, only 6 more museums were matched, while 12 possible matches were also found, but they needed further examination. To verify the accuracy of these matches, the address found in the Istat Microdata dataset can be cross-referenced with the name in the Proprietary dataset to find coherent matches. For example, if *MUSEI CIVICI Lombardia* in the Istat Microdata dataset has the address that locates it in the Municipality of Busto Arsizio, then the matching between *MUSEI CIVICI DI BUSTO ARSIZIO Lombardia* and *MUSEI CIVICI Lombardia* would be correct and added to the main dataset that contains the matches. Table 45 shows the results of the cross-reference.

| Name in Proprietary dataset | Name in Istat Microdata | Address |
|---|---|---|
| MUSEI CIVICI COMUNE DI SONDRIO. MVSA, CAST E MUMIV Lombardia | MUSEI CIVICI Lombardia | VIA REGINA TEODOLINDA - MONZA (MB) |
| **MUSEI CIVICI DI MONZA /COMUNE DI MONZA Lombardia** | **MUSEI CIVICI Lombardia** | **VIA REGINA TEODOLINDA - MONZA (MB)** |
| MUSEI CIVICI DI BUSTO ARSIZIO Lombardia | MUSEI CIVICI Lombardia | VIA REGINA TEODOLINDA - MONZA (MB) |
| MUSEI CIVICI DI BUSTO ARSIZIO Lombardia | MUSEI CIVICI Lombardia | VIA REGINA TEODOLINDA - MONZA (MB) |
| MUSEI CIVICI DI JESI Marche | MUSEI CIVICI Marche | VIA GIACOMO LEOPARDI - SARNANO (MC) |
| MUSEI CIVICI DI JESI Marche | MUSEI CIVICI Marche | VIA GIACOMO LEOPARDI - SARNANO (MC) |
| **MUSEO ARCHEOLOGICO SAN LORENZO Lombardia** | **MUSEO ARCHEOLOGICO Lombardia** | **VIA SAN LORENZO - CREMONA (CR)** |
| MUSEO ARCHEOLOGICO FERRUCCIO BARRECA SANT'ANTIOCO Sardegna | MUSEO ARCHEOLOGICO Sardegna | VIA FRAU - VILLASIMIUS (SU) |
| **MUSEO CIVICO DI CASTEL BOLOGNESE Emilia Romagna** | **MUSEO CIVICO Emilia Romagna** | **VIALE UMBERTO PRIMO - CASTEL BOLOGNESE (RA)** |
| MUSEO CIVICO CASTELBUONO - RETE MUSEA - ECOMUSEO DELLE MADONIE Sicilia | MUSEO CIVICO Sicilia | PIAZZA CARLO MARIA CARAFA-GRAMMICHELE (CT) |
| COMUNE DI LIVORNO- MUSEO CIVICO FATTORI Toscana | MUSEO CIVICO Toscana | VIA CRESCI - MONTAIONE (FI) |
| **MUSEO DI ARTE CONTEMPORANEA (G. ET. DAL VERME) Lombardia** | **MUSEO DI ARTE CONTEMPORANEA Lombardia** | **VIA CARLO DAL VERME - ZAVATTARELLO (PV)** |
| **MUSEO D'ARTE CONTEMPORANEA 'GIUSEPPE E TITINA DAL VERME' Lombardia** | **MUSEO DI ARTE CONTEMPORANEA Lombardia** | **VIA CARLO DAL VERME - ZAVATTARELLO (PV)** |
| **MUSEO ARTE CONTEMPORANEA GIUSEPPE E TITINA DAL VERME Lombardia** | **MUSEO DI ARTE CONTEMPORANEA Lombardia** | **VIA CARLO DAL VERME - ZAVATTARELLO (PV)** |
| MUSEO BOTANICO, SAPIENZA UNIVERSITÀ DI ROMA Lazio | MUSEO DI ROMA Lazio | PIAZZA DI SAN PANTALEO - ROMA (RM) |
| ARCHIVIO MUSEO STORICO DI FIUME IN ROMA Lazio | MUSEO DI ROMA Lazio | PIAZZA DI SAN PANTALEO - ROMA (RM) |
| MUSEO DIOCESANO DI ARTE SACRA DI LAMEZIA TERME Calabria | MUSEO DIOCESANO Calabria | PIAZZA SANT'EUSEBIO DA CASSANO - CASSANO ALL'IONIO (CS) |
| MUSEO ETNOGRAFICO SANT'ANTIOCO Sardegna | MUSEO ETNOGRAFICO Sardegna | VIA NAZIONALE - PALAU (55) |
| **PINACOTECA NAZIONALE DI SIENA Toscana** | **PINACOTECA NAZIONALE Toscana** | **VIA DI SAN PIETRO - SIENA (SI)** |

Table 45 - Actual matches (in bold) found using the address

Thanks to the cross-referencing, 5 records (in bold) are added to the dataset containing the matches.

### Third iteration of Comparison & Decision and Quality Assessment on the Leftovers dataset

The same function is applied to the leftovers dataset with a different scorer parameter. The *fuzz.ratio* parameter, which minimizes LD as a criterion for optimal matches, led to the discovery of a final match to include in the dataset containing the matches.

## Quality Assessment and improvement of results

The unified dataset is composed of the name of the museum in the Proprietary dataset, the name of the museum in the Istat Microdata dataset, the identifier for the cluster of the Proprietary dataset, the identifier for the cluster of the Istat Microdata dataset, the identifier for the matching of the clusters (which will be later used for the integration).

Within this dataset, several checks are essential to ensure alignment with the intended structure. Initially, it is crucial to verify the presence of duplicate names in the Istat Microdata column. The existence of duplicates would imply that names in Istat Microdata were matched with more than one cluster of museums in Proprietary. This discrepancy could stem from an oversight in the matching process or a clustering issue in the Proprietary dataset, indicating a failure to group museums that should belong to the same cluster. Thus, this is a further check to find false negatives.

Upon inspection, 107 duplicate clusters are identified. Among these, 7 clusters were inaccurately matched, while 100 clusters were actually correctly matched, meaning that they were false negatives of the Object Identification process applied to the Proprietary dataset. To resolve this discrepancy, the museums that should have been clustered together are assigned the same cluster identifier, determining the cluster assignment based on which cluster has a higher number of members.

Due to the methodology utilized in prior iterations of the integration process, false negatives from the initial clustering are identified and corrected in this phase. Given the nature of this record linkage problem, that is, when the common column for the integration of data does not include perfect matches, the act of consistently conducting checks to detect errors before they impact subsequent analyses on the integrated data is crucial. Issues in the datasets would introduce errors and possible bias to future evaluations stemming from the integrated data (Kaufman & Klevs, 2021).

## The unified dataset

The leftovers matching process led to the output of a dataset that contains the information about the matching between the Proprietary and the Istat Microdata dataset. In the next part of the section, this dataset, referred to as *Matches dataset*, is used as the basis for the actual integration of the Proprietary and Istat Microdata datasets.

The final step of the integration is done in Python. The first task to be completed is the creation of an unambiguous identifier (key) that will be used to match the integration identifier with the correct record in each of the five datasets (the unified Proprietary

dataset and the 4 Istat Microdata datasets). The objective in creating the variable is to ensure uniqueness for each row by combining the museum's denomination, the region, and the year of the survey that the record pertains to. The variable is called *ID_Denomination_Oss* for the museums in the Proprietary dataset and *ID_Denomination_Micro* for the Istat Microdata ones. A column, containing the integration identifier, is added to each of the datasets. This is done by joining the integration identifier column from the Matches dataset to each of the 5 datasets, using as the common column the correct denomination, which is *ID_Denomination*. In Python, the *merge* function from the Pandas library is used, keeping the type of matching as *inner* as some records were left out of the integration (not matched with any museum of the Istat Microdata dataset).

With each dataset equipped with an identifier that links it with the others, the datasets can be integrated one by one to obtain the integrated dataset. The merging is done by prioritizing the Proprietary dataset, so the first four groups of columns come from Proprietary and the last four come from the Istat Microdata dataset. The merge is implemented with an *outer* logic, since all the matched museums should be kept, even those that only have one year match. The integrated dataset is composed of 598 records, with every row representing one unique museum, and 1532 columns, representing all the answers given in the 8 surveys. Table 46 shows how many museums are matched and for how many years.

| Population | Number of clusters |
|---|---|
| 2 | 353 |
| 4 | 154 |
| 6 | 68 |
| 8 | 23 |

Table 46 - Population and number of clusters in the unified dataset

Figure 17 summarizes the integration process implemented on Python.

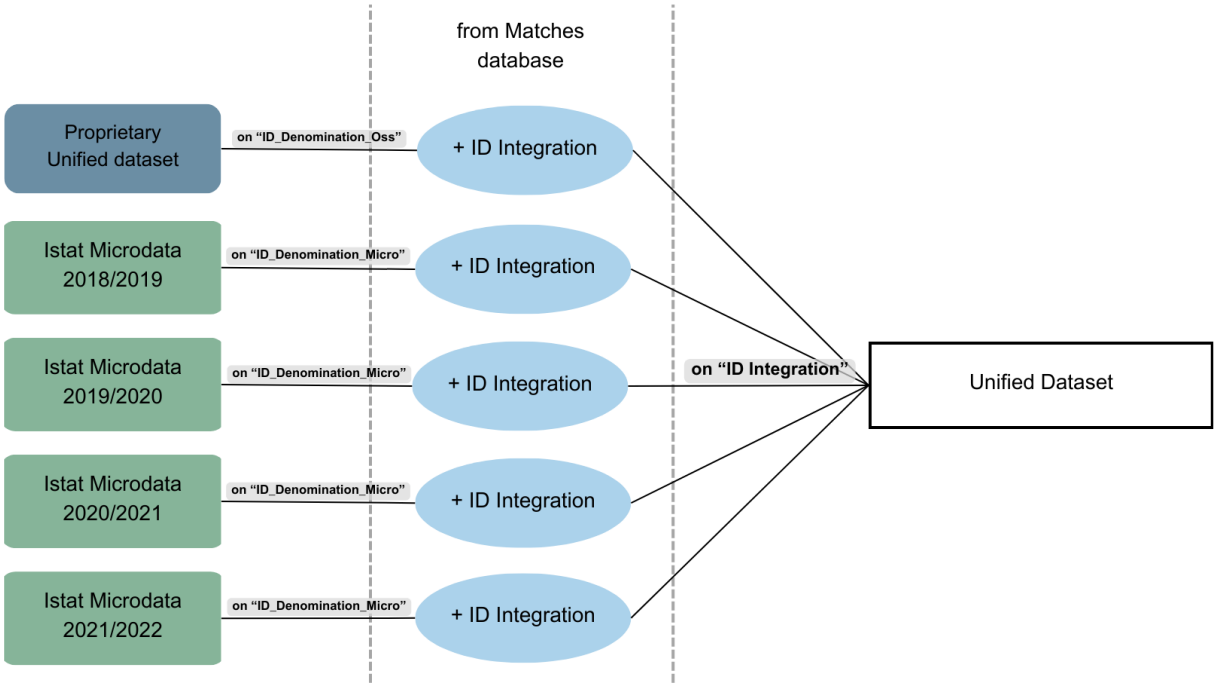


Figure 17 - Integration process schema on Python

Figures 18, 19, and 20 summarize the integration process, showing every step implemented to integrate the Proprietary and Istat Microdata datasets into a unified dataset.

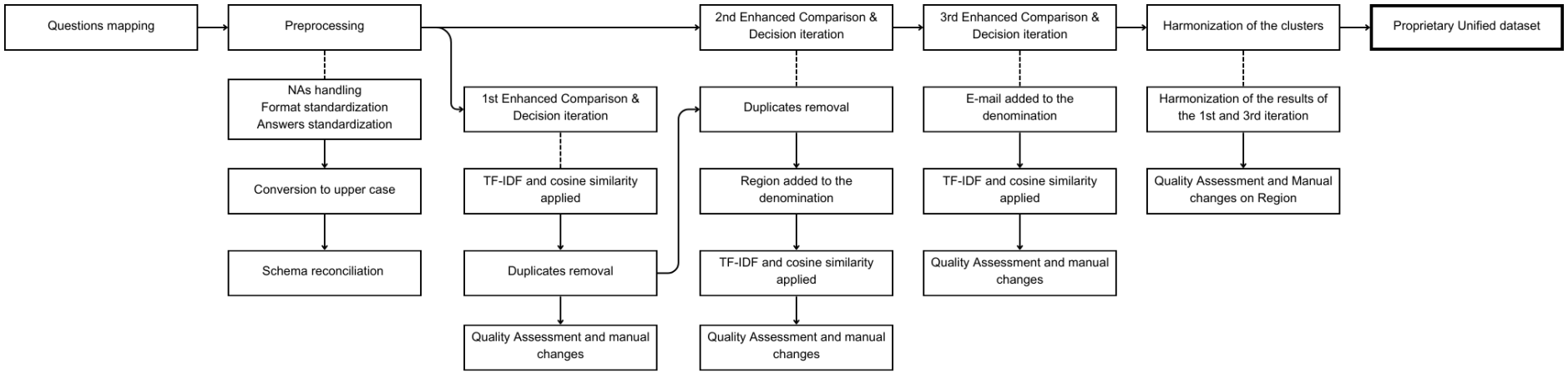


Figure 18 - Harmonization of the Proprietary dataset

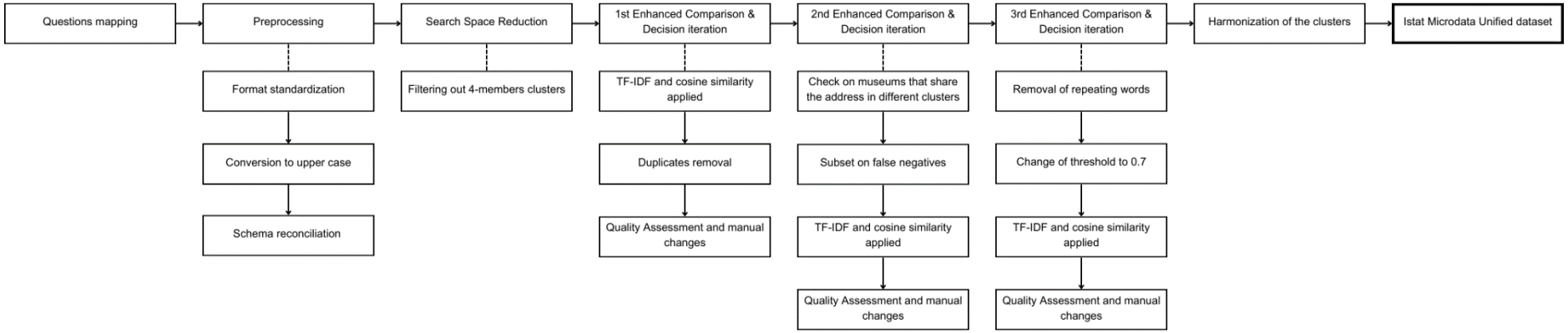


Figure 19 - Harmonization of the Istat Microdata dataset

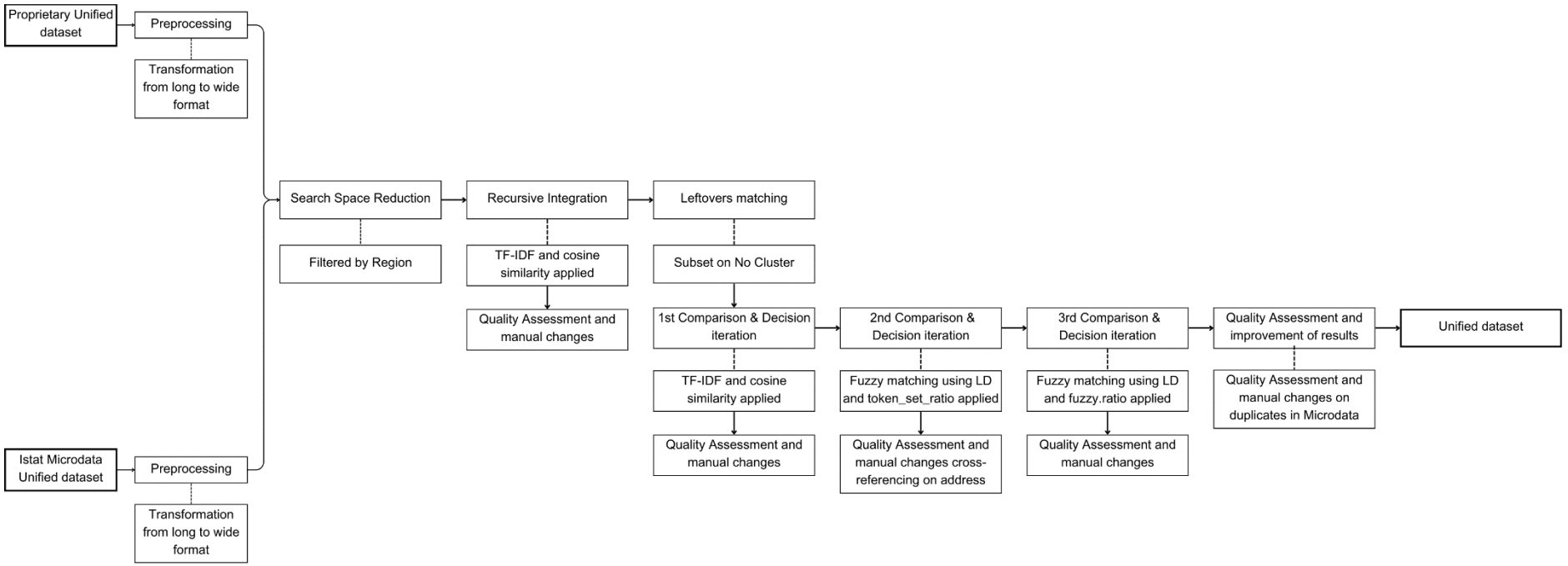


Figure 20 - Data integration of Proprietary and Istat Microdata datasets

# Chapter 6: The dashboard

In this Chapter the process of building the dynamic dashboard is shown. The dashboard is a tool that can assist the practices of museum managers and the Italian Ministry of Culture. The tool is designed to be used by individuals with competencies in both the economic and cultural sectors. Powered by MS Power BI, this dashboard serves as a valuable tool for visualizing KPIs and gaining insights thanks to impactful visual objects. The dashboard is structured around two main views: the Descriptive View and the Digital View, each offering a unique set of KPIs, which will be elaborated upon in this chapter.

In order to evaluate the achieved performance of an indicator, a target is set (Ghalayini & Noble, 1996). The target is based on the level of the competitors or is based on the historical value of the indicator in the organization. This aspect is also interesting for museums because in this way they can compare themselves to other museums and analyze the indicators by contextualizing them with similar museums. This aspect is called benchmarking (Saul, 2004), and it is very important for the dashboard. Museums can be considered competitors to other similar museums, as similarity is found in various aspects such as geographic area, revenues, visitors... Benchmarking is a fundamental part of the dashboard because for many measures there is no defined best practice, therefore, it is best to see the level of the individual museum compared to similar ones. Benchmarking is implemented in the dashboard by showing the difference between the average of a KPI, calculated on similar museums, and the value of the KPI for that museum. Another way that the dashboard enables benchmarking is through quantiles, which allow the museum to see what level the museum achieves compared to other museums.

Therefore, the objectives of the dashboard are twofold:

- Exploratory objective, to serve:
    - Italian museums, in exploring their digital status based on the Organizational Readiness, Online and On-site KPIs.
    - The Italian Ministry of Culture, showing an integrated overview of the situation of museums in Italy.
- Explanatory objective, to serve:
    - Italian museums, specifically museum managers, showing them the level of service offered, and comparing it to that of other similar Italian museums.
    - The Italian Ministry of Culture, showing an integrated overview that can be diced and sliced on many dimensions to analyze specific situations of groups of museums.

## 6.1. KPI selection

### Descriptive View

The first view is created to show the museums that are included in the dataset and the basic characteristics. Because of this, I called the view *Descriptive View*. The Descriptive View aims to show a generalized overview of the situation of museums in Italy. The indicators used allow the user to see the museums' main characteristics (e.g., demographics, the average number of visitors, revenues...). Subsequently, these indicators will serve as a filtering mechanism in various dashboard Views, allowing museum managers to identify museums similar to their own. This functionality enables effective benchmarking among museums that share similarities across one or more dimensions. The indicators shown will come from the integrated data, i.e., from the integration of both the Proprietary dataset and the Istat Microdata dataset, where the second dataset enriches the Descriptive View with more information that cannot be found in the Proprietary dataset alone.

#### Integrated data to build the Descriptive View

The data integration enriches the Descriptive View with 4 new variables that provide more information. Firstly, the geographical location is more precise thanks to the addition of Province and Municipality which can be used for geostatistics and can also be keys for further integration (e.g., integrating the number of residents in the city). The *Typology* of museum is another important addition as it shows the context of the museum and it creates the possibility of benchmarking between museums of the same *Typology*. There are 12 possible typologies of museums, each highlighting a different focus of the museum's offering: arts, modern arts, religion, archeology, history, natural science, science and technology, ethnography, thematic, industrial, museum-house, and other. The *Typology* options do not change over the 4 years of the Istat Microdata survey, meaning they can be used without needing harmonization. Moreover, the number of workers (*Personnel*) in the institution is an important variable as it is an indicator of the size of the institution, which can work together with the revenues and the number of visitors, showing different dimensions that define the size in different ways.

Table 47 shows the list of the variables and the questions they refer to in the surveys. The variables in *italics* are obtained thanks to data integration.

| Variable | Proprietary dataset | | | | Istat Microdata | | | |
|---|---|---|---|---|---|---|---|---|
| | 2018/19 | 2019/20 | 2020/21 | 2021/22 | 2018/19 | 2019/20 | 2020/21 | 2021/22 |
| Region | Q5 | Q38 | Q43 | Q48 | Ext | Ext | Ext | Ext |
| Revenue from tickets | Q28 | Q23 | Q30 | Q29 | 46 | / | / | / |
| Number of visitors | Q9 | Q4 | Q28 | Q27 | 30.1 | 15.1 | 11 | 14.1 |
| *Province* | / | Q39 | / | / | 1.2 | 2.5 | 2.5 | 2.5 |
| *Municipality* | Q6 | / | / | / | 1.3 | 2.6 | 2.6 | 2.6 |
| *Typology* | / | / | / | / | 5 | 5 | 5 | 5 |
| *Personnel* | Q10 | / | / | / | 43.1 | 6 | 12.1.A - 13.2 | 16.1 - 16.5 |

Table 47 - Descriptive View variables in the sourcing surveys

The Region variable in the Istat Microdata surveys is not included as a question but it is included in the dataset. This is because the variable is added to the dataset in a prior integration, done by Istat.

The variables Number of workers (*Personnel*) in the 2020 and 2021 Istat Microdata surveys need to be harmonized as those two surveys do not ask for the total number of workers directly, but they ask for the total number of internal staff, external staff, consultants volunteers, and civil service operators. The sum of these variables makes up the total number of workers.

The *Typology* variable needs to be harmonized in all four years as, in the dataset, the answers are codified with numbers from 1 to 12, with each number representing a different type of museum. The match between number and type can be found in the metadata provided by Istat.

### KPIs in the Descriptive View

The Descriptive View is composed of KPIs that are typical of PM literature in museums. The number of visitors has been proposed several times (Bishop & Brand, 2003; Guccio et al., 2018) and has always been collected, mostly to comply with procedures (Agostino & Arnaboldi, 2021) rather than for use as a performance indicator. The revenues from tickets indicator is another common financial KPI that has been traditionally collected by museums (Agostino et al., 2020), often found aggregated into financial metrics originating from the balance sheet. The number of workers (*Personnel*) is an important indicator in defining the size of the museum. Indeed, the two indicators presented before present some issues when benchmarking museums: the number of visitors and the revenue from tickets depend on the access to the museum (e.g., closing periods due to renovations, closing periods due to COVID-19, etc.). Moreover, the revenues from tickets depend on the price of the tickets, which

in many museums are free during promotional periods (like Domenica al Museo[15] in Italy). *Personnel* can be a useful variable to avoid the issues related to visitors and revenues.

To define the developed KPIs, tables that are a standard practice in PM and reporting, recognized and accepted in academic and business contexts, are utilized. The structure chosen is a readaptation of the ISO 22400 standard, which defines an industry-neutral framework for the definition of KPIs (ISO, 2014).

| | |
|---|---|
| **Name** | Number of visitors |
| **Description** | The indicator quantifies the number of visitors who visit the museum |
| **Metric** | No formula |
| **Unit of measure** | No Unit |
| **Performance Manager** | Marketing Department, PR Specialists, Curator |
| **Frequency** | Yearly |
| **Distribution list** | Marketing Employees |
| **Dimension** | Visitor |
| **Values measured** | Number of visitors |
| **Target** | The target would change by the size and type of a museum. There is no one-size-fits-all target. |

Table 48 - Number of visitors KPI summary table

| | |
|---|---|
| **Name** | Revenue from tickets |
| **Description** | The indicator quantifies the revenue generated from ticket sales |
| **Metric** | $\sum_i Price_i \cdot Number\ of\ ticket\ sold_i$ |
| **Unit of measure** | Euros (€) |
| **Performance Manager** | Financial officers, Curator |
| **Frequency** | Yearly |
| **Distribution list** | Ticketing officer |
| **Dimension** | Organization – Visitor |
| **Values measured** | Ticket price, Number of tickets sold |
| **Target** | The target would change by the size and type of a museum. There is no one-size-fits-all target. |

Table 49 - Revenue from tickets KPI summary table

[15] https://cultura.gov.it/domenicalmuseo

| | |
|---|---|
| **Name** | Number of workers – Personnel |
| **Description** | The indicator measures the number of workers employed at the museum |
| **Metric** | No formula |
| **Unit of measure** | No unit |
| **Performance Manager** | HR Department, Curator |
| **Frequency** | Yearly |
| **Distribution list** | HR Recruiters |
| **Dimension** | Organization |
| **Values measured** | Number of people that work for the museum |
| **Target** | The target would change by the size and type of a museum. There is no one-size-fits-all target. |

Table 50 - Personnel KPI summary table

## Digital View

The second view has the goal of showing how pervasive are digital technologies and ICT in both the internal processes of the organization and the museum's offerings to the general public. Because of this, I called the view *Digital View*. The View looks at museums from the aspect of organizational readiness and the services offered to the visitor.

### Integrated data to build the Digital View

The data integration enriches the Digital View with 3 new variables that provide more information.

Firstly, a variable related to the presence of a dedicated website is found in each of the four Istat Microdata datasets, while that is not the case for the Proprietary dataset (the variable is present only in the 2018-2019 and the 2019-2020 survey). This is a value added to the Proprietary dataset and this information can now be used for the creation of a KPI that spans four years.

This is also true for the presence of the museum on social media; in the Proprietary dataset there are no questions about the social media presence, and there is only one question that asks if the social networks are monitored (in surveys 2018-2019 and 2019-2020). In the Istat Microdata dataset, the variable regards if the museum owns a social media page (on any social media), which is very different and gives information that was not available before. Moreover, this question is found in all four Istat Microdata datasets.

Another variable that is possible to assess thanks to the integration is the presence of online tours. The question asks if the museum offers the service. The question is present in all four years of the Istat Microdata dataset, while it is only in two (2020-2021 and 2021-2022) of the Proprietary datasets.

The list of variables used for the Digital View is composed of 8 variables, shown in Table 51. The last three are in italics because they are obtained thanks to data integration.

| Variable | Proprietary dataset | | | | Istat Microdata | | | |
|---|---|---|---|---|---|---|---|---|
| | 2018/19 | 2019/20 | 2020/21 | 2021/22 | 2018/19 | 2019/20 | 2020/21 | 2021/22 |
| Digitalization of the collection | Q40 | Q34 | Q19 | Q19 | 42.2, 42.3B | 21, 22 | 16.1 | 25.1, 25.2 |
| Presence of a digital innovation plan | Q13 | Q7 | Q39 | Q39 | / | / | / | / |
| Workers with digital competencies | Q44 | Q36 | Q41 | Q40 | 45.6 | / | / | 17.6 |
| Available technologies | Q26 | Q32 | Q38 | Q38 | 52.8 | 26.8 | / | 21 |
| % ticket revenue divided between channels | Q33 | Q26 | Q34 | Q33 | / | / | / | / |
| Presence of a dedicated website | Q18 | Q11 | / | / | 1.11, 55.1 | 17 | 2.7 | 2.7 |
| Social media presence | Q21 | Q14 | / | / | 55.5 | 20 | 16.4 | 23.2 |
| Online tours | / | / | Q7 | Q9 | 55.4 | 19.1 | 16.6 | 23.3, 23.4 |

Table 51 - Digital View variables in the sourcing surveys

### KPIs in the Digital View

The Digital View comprises newly suggested KPIs outlined in the thesis, intended to assess the digital state of museums.

The first proposed KPI is Organizational Readiness. Organizational readiness is defined as a multidimensional construct for assessing the extent to which resources and implementation conditions favor an initiative's success (Shahrasbi & Paré, 2015). This concept is used to describe the readiness of an organization to adopt new technologies. The readiness entails both structural aspects, associated with access to resources, and psychological aspects, linked to the attitudes of the members (Raguseo et al., 2018; Shahrasbi & Paré, 2015). Many papers have been published on the factors that influence organizational readiness, with different authors emphasizing different factors (for example Mehrtens et al., 2001; Iacovou et al., 1995). The focus of organizational readiness has been the development of IT knowledge and structures in the organization (e.g., Mehrtens et al., 2001; Molla & Licker, 2005, Grandon & Pearson,

2004) and only recently it has shifted towards the importance of data and digital transformation in the organization (Raguseo et al., 2018). In this section, organizational readiness is considered as the foundation on which the integration of digital technologies into the museum is based. A high degree of organizational readiness facilitates the implementation of new digital technologies and assists in the creation of a digital strategy.

The formalization of the digital strategy through the draft of a specific document that supports and plans digital innovation is still not widespread among cultural institutions. The lack of vision and the lack of a structured strategy are factors that limit digital innovation in these institutions (Gombault et al., 2016). Another constraint that hinders digital innovation is the lack of personnel with digital competencies (Bekele et al., 2018; Agostino & Arnaboldi, 2021). These two issues explain the lack of organizational readiness of many museums. These backstage factors are not directly observed by the visitors, but they indirectly affect them.

Following the work of De Bernardi et al. (2019) and Guccio et al. (2020), two other KPIs are proposed: Online and On-Site. The Online category refers to the digital technologies that are used both before (antecedent) and after (subsequent) the on-site visit. The On-Site category refers to digital technologies that are used during the on-site visit (concurrent) (Nigro et al., 2016; De Bernardi et al., 2018). Guccio et al. (2020) also makes the distinction between in-situ and online services, adding that in-site services improve the visitor's experience during the visit, while online services prepare for the visit, or even can substitute the visit (e.g., online tours). The two macro-categories are converted into indicators designed to display the museums' performance in terms of Online and On-site offerings in the dashboard.

| KPI | Variables in the dataset | Reference |
|---|---|---|
| Organizational readiness | Presence of a digital innovation plan | Gombault et al. (2016) |
| Organizational readiness | Workers with digital competencies | Bekele et al. (2018); Agostino & Arnaboldi (2021) |
| Online | Digitalization of the collection | De Bernardi et al. (2019); Pesce et al. (2019) |
| Online | Online ticketing | De Bernardi et al. (2019) |
| Online | Online tours | De Bernardi et al. (2019) |
| Online | Presence of a dedicated website | De Bernardi et al. (2019); Camarero & Garrido (2012) |
| Online | Social media presence | De Bernardi et al. (2019); Marty (2007) |
| On-Site | Available technologies: Audioguide | De Bernardi et al. (2019); Guccio et al. (2020) |
| On-Site | Available technologies: Augmented Reality | De Bernardi et al. (2019); Guccio et al. (2020) |
| On-Site | Available technologies: Virtual Reality | De Bernardi et al. (2019); Guccio et al. (2020) |
| On-Site | Available technologies: QR code | De Bernardi et al. (2019); Guccio et al. (2020) |
| On-Site | Available technologies: ChatBot | De Bernardi et al. (2019); Guccio et al. (2020) |

Table 52 - Digital View KPIs, variables and references

This is the proposal for the metrics of the indicators:

### *Organizational Readiness*

The indicator shows how much the institution is advanced on the subject of organizational readiness. The indicator is a combination of the presence of a digital innovation plan and the presence of workers with digital competencies. The last three variables are binary, with 1 representing *Yes* and 0 being *No*. Equal weight is assigned to both variables in the computation since the thesis does not go in-depth into the impact and significance of each variable.

| Variable | Values | Description |
|---|---|---|
| Presence of a digital innovation plan | 0 | There's no digital innovation plan |
| | 1 | There is a formalized digital innovation plan |
| Workers with digital competencies | 0 | There are no workers with digital competencies |
| | 1 | There is at least one worker with digital competencies |

Table 53 - Organizational Readiness metrics

The proposed formula for computing the KPI is an average of the two variables.

$$Organizational\ Readiness = \frac{Digital\ innovation\ plan + Workers\ with\ digital\ competencies}{2}$$

*Online*

The indicator reflects how advanced is a museum on the subject of online offerings. The five variables that make up the online presence are the digitalization of the collection, the presence of online ticketing, the possibility of doing virtual tours of the museum, the social media presence, and the presence of a dedicated website. The digitalization of the collection variable has five possible values: 0%, <25%, 25-50%, 51-75%, and >75%. These values represent the percentage of the collection that is been digitalized (on the total number of goods in the collection). To make this variable computable, an integer is given to every number, which incrementally represents how much of the collection is digitalized, starting from 0. The online ticketing variable is represented as the percentage of ticket sales that happen through online channels (owned website + other online websites and apps). The last three variables are binary, with 1 representing *Yes* and 0 being *No*. Equal weight is assigned to all variables in the computation since the thesis does not go in-depth into the impact and significance of each variable.

| **Variable** | **Values** | **Description** |
|---|---|---|
| Digitalization of the collection | 0 | The collection has not been digitalized |
| | 1 | The collection has been digitalized for less than 25% |
| | 2 | The collection has been digitalized for 25%-50% |
| | 3 | The collection has been digitalized for 51%-75% |
| | 4 | The collection has been digitalized for more than 75% |
| Online ticketing | % | The percentage of ticket sales that happen through online channels |
| Online tours | 0 | There are no online tours in the museum offering |
| | 1 | There is the possibility of doing an online tour of the museum |
| Presence of a dedicated website | 0 | The museum does not have a dedicated website |
| | 1 | The museum has a dedicated website |
| Social media presence | 0 | The museum does not own any social media page |
| | 1 | The museum owns at least one social media page |

Table 54 - Online metrics

The proposed formula for computing the KPI is an average of the five variables.

$$Online = \frac{Digitalization/_{4} + Online\ ticketing/_{100} + Social\ media + Website + Online\ tours}{5}$$

*On-site*

The indicator shows how much the on-site visit can be enhanced thanks to digital technologies. The indicator is composed of the available in-situ technologies: audio guide, augmented reality, virtual reality, QR code/ beacon, and ChatBot. The variables are five and are all binary. Equal weight is assigned to all variables in the computation since the thesis does not go in-depth into the impact and significance of each variable.

| Variable | Values | Description |
|---|---|---|
| Available technologies: Audioguide | 0 | There are no audioguides in the on-site offering of the museum |
| | 1 | There are audioguides in the on-site offering of the museum |
| Available technologies: Augmented Reality | 0 | There are no AR technologies in the on-site offering of the museum |
| | 1 | There are AR technologies in the on-site offering of the museum |
| Available technologies: Virtual Reality | 0 | There are no VR technologies in the on-site offering of the museum |
| | 1 | There are VR technologies in the on-site offering of the museum |
| Available technologies: QR code | 0 | There are no QR codes in the on-site offering of the museum |
| | 1 | There are QR codes in the on-site offering of the museum |
| Available technologies: ChatBot | 0 | There are no ChatBots in the on-site offering of the museum |
| | 1 | There are ChatBots in the on-site offering of the museum |

Table 55 - On-site metrics

The proposed formula for computing the KPI is an average of the five variables.

$$On - site = \frac{\sum Available\ technology}{5}$$

*Average Digital KPI*

This indicator is a summary of the three KPIs proposed. It is computed as the average of Online, On-site, and Organizational Readiness. It is useful to benchmark on a singular dimension and to give a comprehensive overview of the Digital status of a

museum. The proposed formula for computing the KPI is an average of the three Digital KPIs. Equal weight is assigned to the three KPIs used in the computation since the thesis does not go in-depth into the impact and significance of each KPI.

$$Average\ Digital\ KPI = \frac{Organizational\ Readiness + Online + On - site}{3}$$

| | |
|---|---|
| **Name** | Online |
| **Description** | The indicator measures the number of services offered online |
| **Metric** | $\frac{Digitalization/_4 + Online\ ticketing/_{100} + Social\ media + Website + Online\ tours}{5}$ |
| **Unit of measure** | No unit |
| **Performance Manager** | Digital content curator, Employees who deal with digital content |
| **Frequency** | Yearly |
| **Distribution list** | Online ticketing manager, Website manager, Social media manager |
| **Dimension** | Organization – Visitor |
| **Values measured** | Digitalization of the collection (% of total collection), % of sales through the online channels, presence on social media, presence of a dedicated website, and possibility of doing online tours of the museum. |
| **Target** | The target would change by the size and type of a museum. There is no one-size-fits-all target. |

Table 56 - Online KPI summary table

| | |
|---|---|
| **Name** | On-site |
| **Description** | The indicator measures the quantity of available technologies on-site (during the visit) |
| **Metric** | $\frac{\sum Available\ technology}{5}$ |
| **Unit of measure** | No Unit |
| **Performance Manager** | ICT Manager |
| **Frequency** | Yearly |
| **Distribution list** | Digital technologies Manager |
| **Dimension** | Organization – Visitor |
| **Values measured** | Presence of audioguides, augmented reality, virtual reality, QR code/ beacon, and ChatBot |
| **Target** | The target would change by the size and type of a museum. There is no one-size-fits-all target. |

Table 57 - On-site KPI summary table

| | |
|---|---|
| **Name** | Organizational Readiness |
| **Description** | The indicator measures the Organizational Readiness of the museum in the adoption of digital technologies |
| **Metric** | $\frac{Digital\ innovation\ plan + Workers\ digital\ competencies}{2}$ |
| **Unit of measure** | No Unit |
| **Performance Manager** | ICT Manager, HR Department |
| **Frequency** | Yearly |
| **Distribution list** | Digital innovation manager, HR Recruiters |
| **Dimension** | Organization |
| **Values measured** | Presence of a digital innovation plan, Workers with digital competencies |
| **Target** | The target would change by the size and type of a museum. There is no one-size-fits-all target. |

Table 58 - Organizational Readiness KPI summary table

| | |
|---|---|
| **Name** | Average Digital |
| **Description** | The indicator measures the Digital status of the museum |
| **Metric** | $\frac{Organizational\ Readiness + Online + On - site}{3}$ |
| **Unit of measure** | No Unit |
| **Performance Manager** | ICT Manager, Curator |
| **Frequency** | Yearly |
| **Distribution list** | Digital innovation manager, HR Recruiters, Digital technologies Manager, |
| **Dimension** | Organization |
| **Values measured** | Presence of a digital innovation plan, Workers with digital competencies, Online ticketing manager, Website manager, Social media manager |
| **Target** | The target would change by the size and type of a museum. There is no one-size-fits-all target. |

Table 59 - Average Digital KPI summary table

## 6.2. Dashboard building

In building the dashboard, rules and guidelines are followed to ensure an unbiased and pleasing visualization. The rules followed originate from a combination of sources:

- Suggestions from an expert in data visualization
- The Big Book of Dashboards - Visualizing Your Data Using Real-World Business Scenarios (Wexler et al., 2017)

- Chart Suggestions—A Thought-Starter[16]

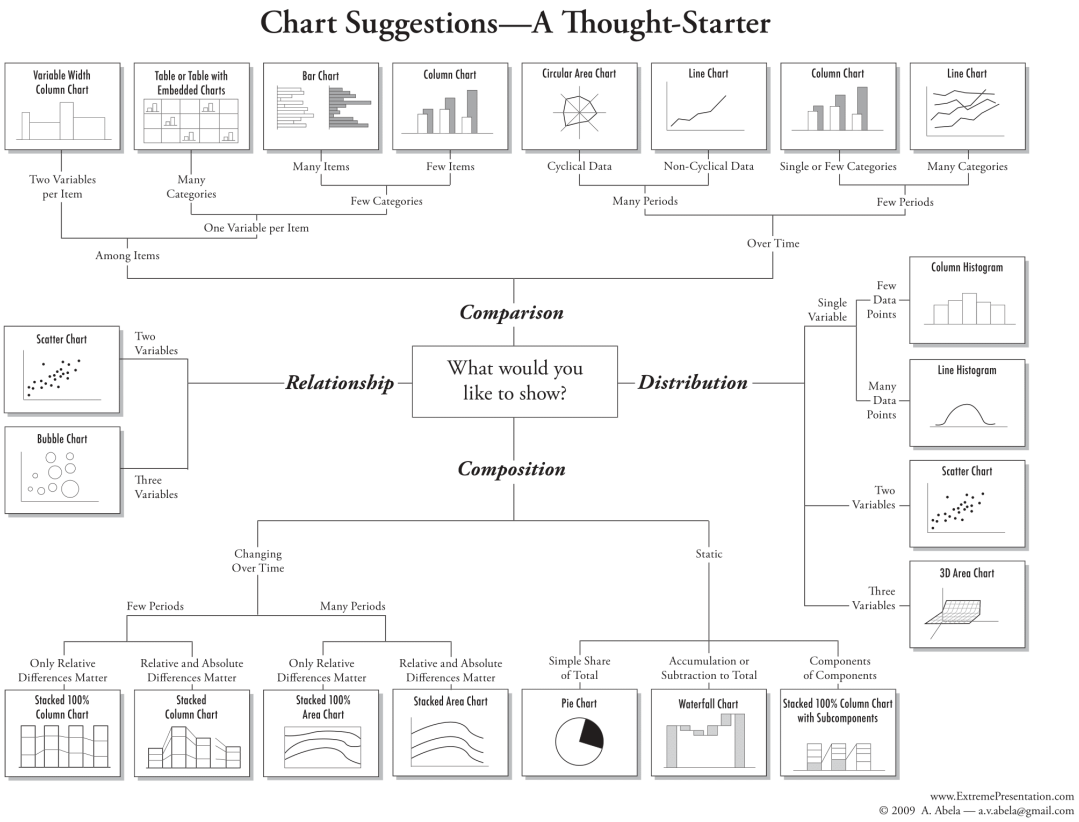


Figure 21 - Abela's (2009) Chart Suggestions

There are 5 main steps[17] that need to be followed to build a dashboard.

1. Define goal and audience: the dashboard needs to have a clear goal and needs to be understandable for the audience that is going to use it. This means that the visualization needs to be adapted to the user and its understanding.
2. Selecting visual format: each element that a dashboard is composed of needs to be chosen by keeping in mind its readability. If an object is readable, then the information that it represents can be understood by the user. Abela (2009)[16] published these guidelines to help in selecting the best graph (element) in any given context (Figure 21). Another tool that can be used for the same purpose is the Visual Vocabulary by Financial Times[18]. In this tool, nine categories of graphs are

[16] Abela (2009). Retrieved from: https://extremepresentation.typepad.com/files/choosing-a-good-chart-09.pdf

[17] Bresciani, Sabrina. "Visualizing data". Class lecture, Advanced Performance Measurement, Politecnico di Milano, Milano, 19 October 2023.

[18] https://www.ft.com/content/c7bb24c9-964d-479f-ba24-03a2b2df6e85

found, and 73 specific graphs are explained, highlighting their context of use and how they work.

3. Accuracy of encoding: the dashboard should be coherent in its colors and shapes found in the graphs. In general, to ensure an easier visualization, the coloring of objects to signal their greatness should be done logically. Color should be used purposefully, not just to make a visualization look more vivid (Wexler et al., 2017).

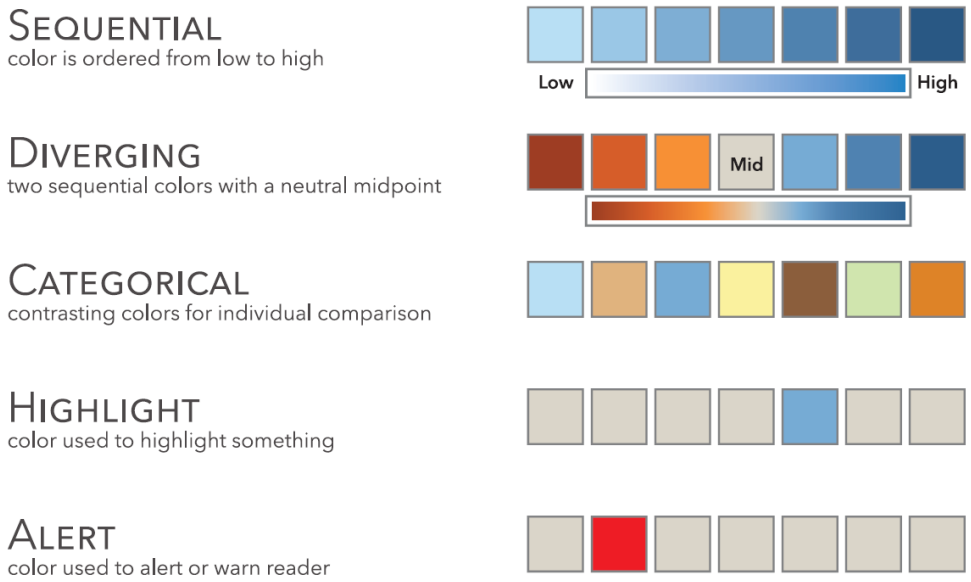


Figure 22 - Use of Color in Data Visualization (from Wexler et al., 2017)

4. Inclusive design:
    - When choosing colors, the creator of the dashboard needs to be mindful of its readability to color-deficient people. In fact, more than 300 million people in the world are born with a color deficiency[19], and 99% of them have trouble distinguishing red from green[20]. That's why the combination of red and green should not be used, while the combination of blue and orange is the preferred choice.
    - From 2025 onwards, every product in the EU must comply with the European Accessibility Act (EAA), which is an EU law that requires products and services to be accessible for persons with disabilities (EU, 2019). The dashboard needs to be built in compliance with the standards written in the EAA.
5. Testing: the best way to understand if the dashboard is working as intended is to test it with a sample of future users. There are many kinds of testing possible, from trivial to sophisticated methods:
    - Comprehension test: it checks the misalignment between perceived comprehension of the dashboard vs the actual comprehension, assessed by

[19] https://www.colorblindguide.com
[20] https://www.color-blindness.com

asking both if the user understood and testing their understanding with specific questions on the dashboard.

- Usability evaluation: it asks the user questions regarding the functionality and usability of the dashboard (e.g., Do you evaluate the dashboard as practical and functional?; Is the meaning of this graph clear?).
- Think-aloud test: the user tests the dashboard by expressing their opinion aloud during their experience (stream of consciousness).
- Eye-tracking test: a tool that tracks the movement of the eye to find what is looked at first, more, last, and least in a dashboard.
- A/B testing: different prototypes (versions) of the dashboards are given to different groups of people and the two are evaluated separately. Then, results from the evaluations show which one is the best prototype.

### Development of the dashboard

The constructed dashboard comprises the KPIs highlighted in section 6.1. It consists of two primary Views: Descriptive View, featuring four visuals, and Digital View, featuring five visuals. This division is done to avoid the overcrowding of a singular View, which would result in an unpleasing and chaotic visual.

The dashboard is interactive, meaning the user is engaged and can autonomously choose the area of interest to analyze. The Power BI Document featuring the dashboard can be found as an attachment to the thesis.

The choice of colors follows the proposition of Wexler et al. (2017). The main color used in the dashboard is blue, as it is a standard color for data visualization. The selection between using a sequential scale or a diverging scale depends both on the type of data that needs to be visualized and on a personal choice. The two scales can be used interchangeably, however, a diverging scale emphasizes the difference from a midpoint, while a sequential scale just visually shows the size of a value. A diverging scale is often used when the midpoint is meaningful. The chosen colors for the diverging scales are orange, blue, and grey. The selection of these three colors is intended to enhance the inclusivity and comprehensibility of the dashboard for individuals with color deficiency. In every graph encoded with a diverging scale, orange represents the lowest value, grey the middle value, and blue the highest value. The color chosen for the sequential scale is blue.

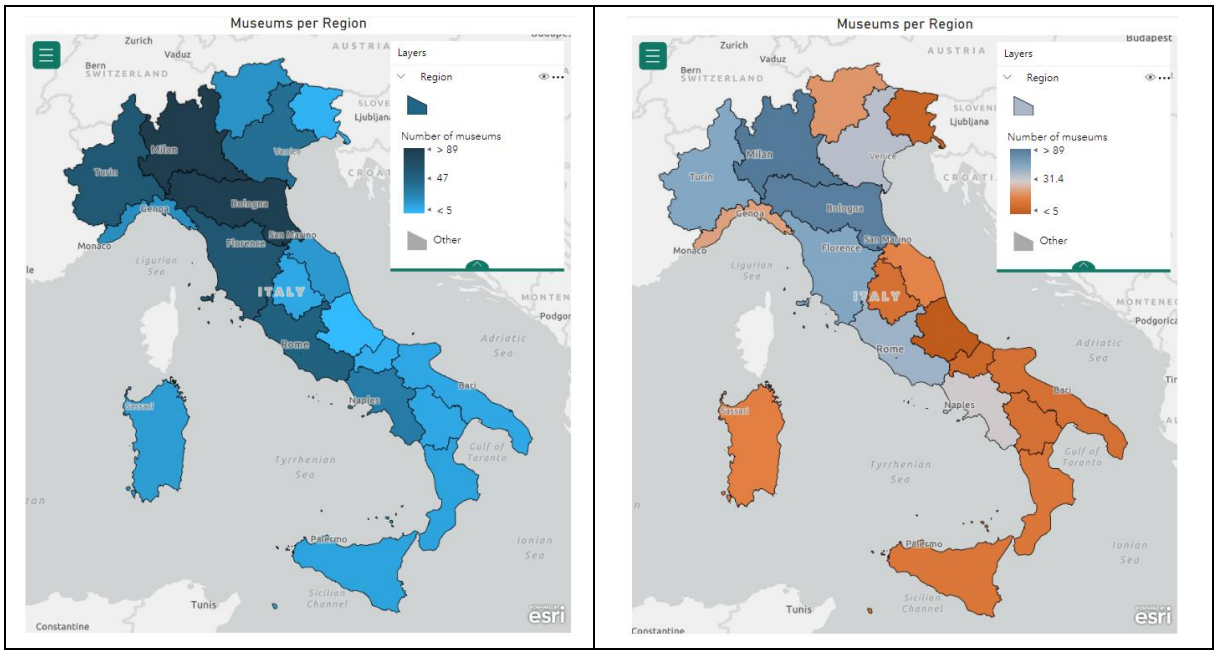


Figure 23 - Difference between sequential and diverging scale

This example shows the two scales applied to a geographical visual object (a choropleth map from ArcGIS Maps for Power Bi). In this case, since the midpoint is not a meaningful value, the sequential visualization would be preferrable.

The choice of visual objects to represent the KPIs is done following Abela's (2009) Chart Suggestions and the Financial Times Visual Vocabulary. A summary of the choices of visual objects made in every View can be found in Appendix A.2.

## The Views

In this sub-section, the structure of the Views is illustrated, detailing their interactive functionalities and the questions they aim to address.

### Benchmarking page

This page is incorporated into the dashboard to enable benchmarking across the Views. Although it is not classified as a View due to its non-visual objective within the dashboard, this page plays a critical role by enabling filtering across all Views, essentially ensuring the dashboard's functionality. Indeed, through this page, the user can filter the other Views of the dashboard by a vast quantity of variables: Number of visitors, Revenues from tickets, *Personnel*, Online KPI, On-site KPI, Organizational Readiness KPI, *Typology*, Region, *Province*, *Municipality*.

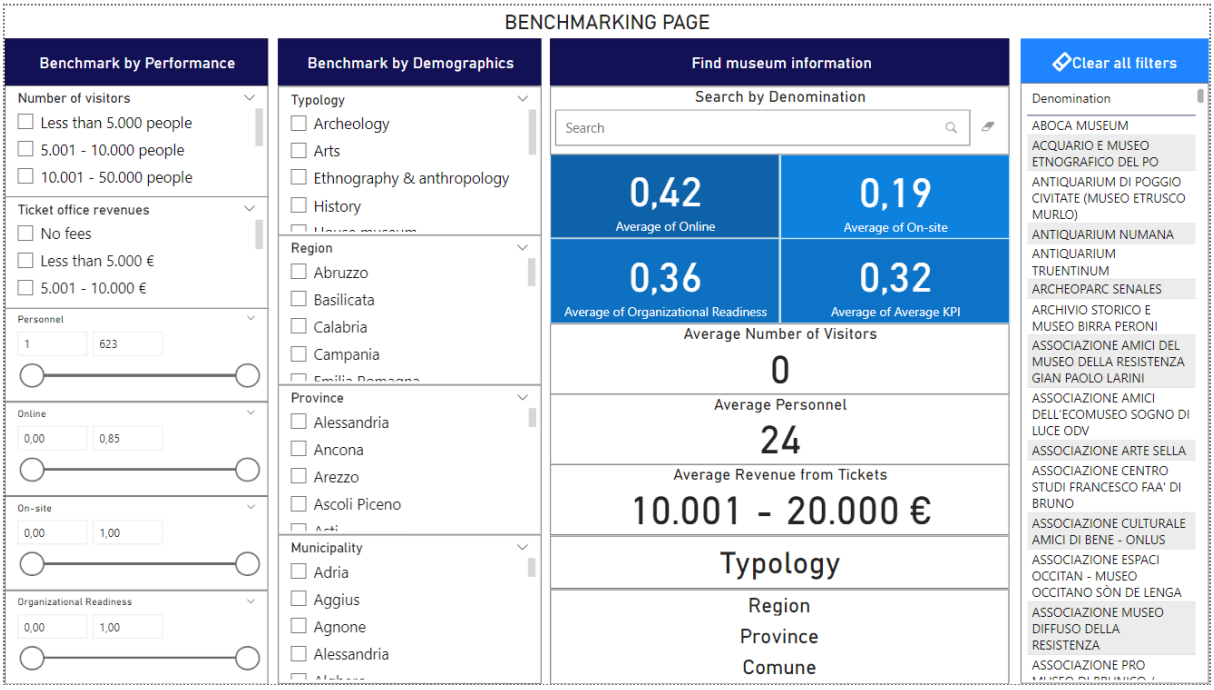


Figure 24 - Benchmarking page

The page is divided into two parts. The first part is composed of 10 filters and can be used to benchmark either by performance or by the demographics selected. The second part lets the user search for a museum and displays all the information stored in the dataset about that museum. This is done to ensure that a museum manager can search for information about its own museum that can be used later for filtering and benchmarking against museums that share similar characteristics. The possibility to search for a specific museum is not exclusive to the Benchmarking page, as it is found in every View.

## Descriptive View

This View has the objective of showing an initial general overview of the dataset. In this View, the user can find information about the demographics and the *Typology* of the museums.

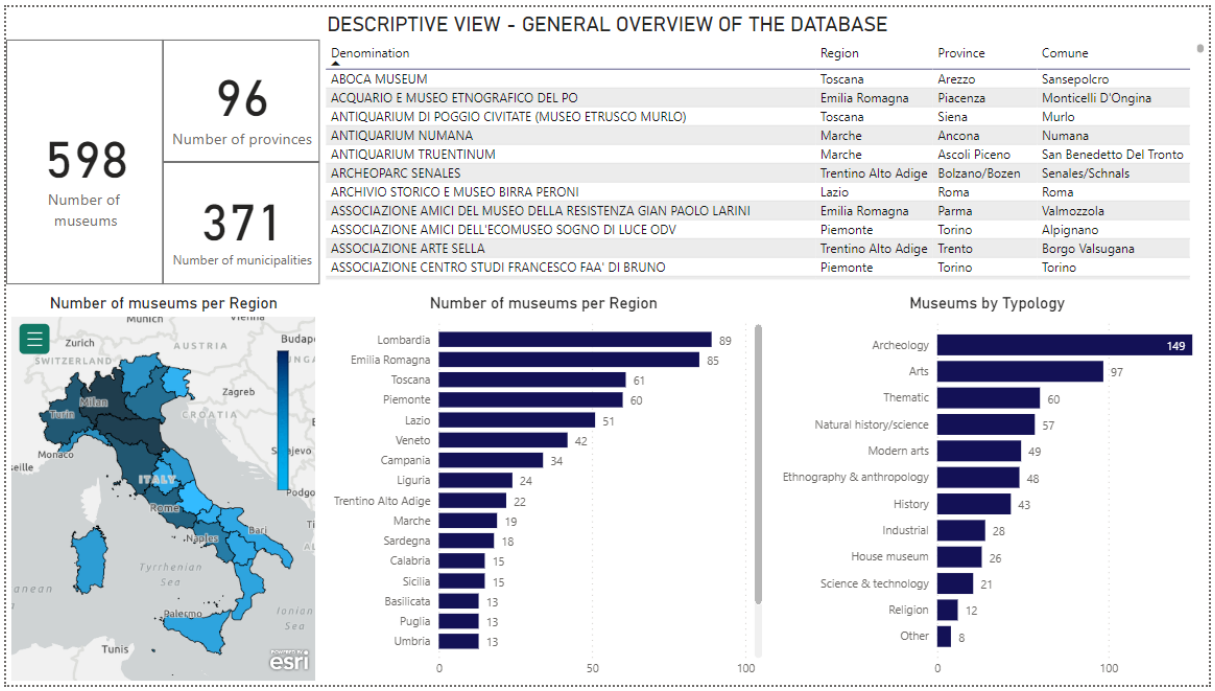


Figure 25 - Descriptive View

The View is composed of three *cards* (number of museums, number of provinces, number of municipalities), a table that shows the name and geographical location (region, *Province*, and *Municipality*) of all the museums, a choropleth map (encoded with a sequential scale to show the number of museums per region), and two bar charts, showing the number of museums per region and the number of museums by *Typology*.

The View can be filtered by region by interacting with the choropleth map and with the *Number of museums per Region* bar chart. It can also be filtered by the *Typology* of museums by interacting with the *Museums by Typology* bar chart. Every object of the View is dynamic and can be used to filter, apart from the first three *cards* which change dynamically but cannot be used to filter.

Questions answered by the View:

- How many museums are in the unified dataset?
- In how many provinces and municipalities are the museums located?
- Where is a specific museum located (region, province, municipality)?
- How many museums are located in a set region?
- What are the typologies of museums?
- How many museums are of a particular typology? Where are they located?

## Descriptive View - Personnel

This View has the objective of giving information about the situation of the workers in museums. This View is meant to be used to benchmark the average *Personnel* between regions and Italy as a whole, and to benchmark within the region.

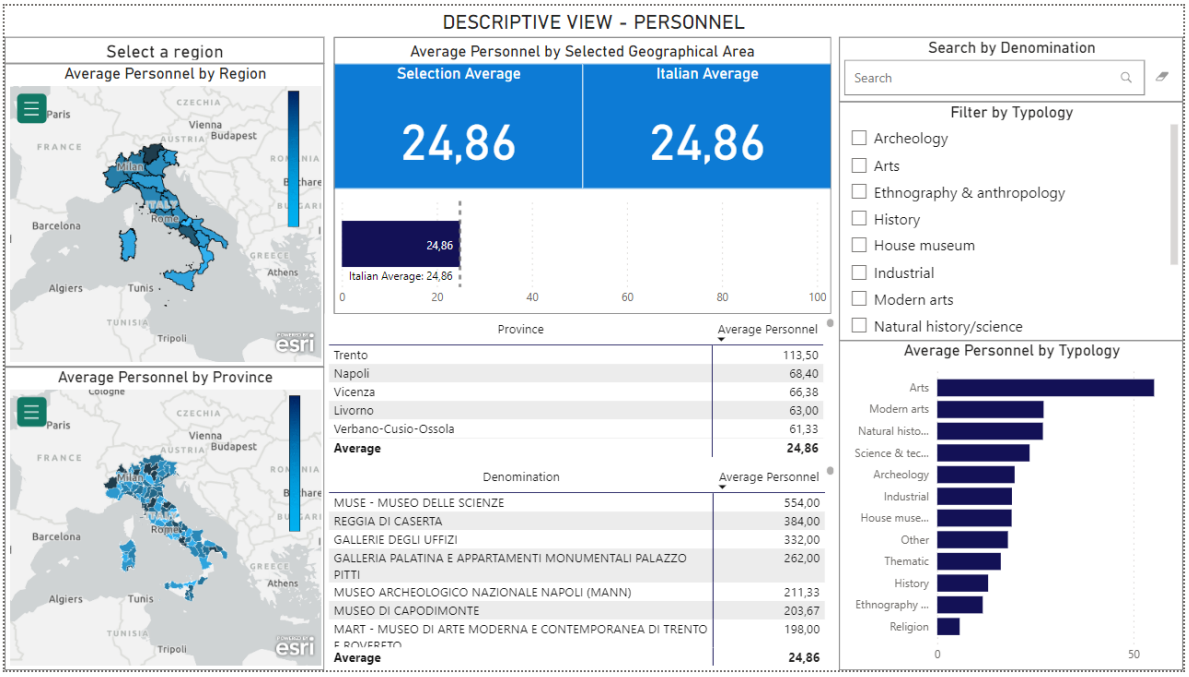


Figure 26 - Descriptive View - Personnel

The View is composed of two choropleth maps (that show the average museum *Personnel* by region and *Province*), two numerical cards (showing the selected regional average and the Italian average), two tables (showing average *Personnel* by *Province* and by museum), a text card (highlighting the selected region), a filter object (to filter by *Typology* of museum), and two bar charts (the first showing the selected geographical location average *Personnel* against the Italian average *Personnel*, the second highlighting the average *Personnel* by *Typology* of museum).

The View can be filtered by region and *Province* by interacting with the choropleth maps. To make filtering by provinces easier, it is best to begin by filtering by region. This will cause the maps to zoom in on the selected region and thus make it easier to select the *Province* of choice. The provinces can also be selected using the *Province - Average Personnel* table. If a region has not yet been selected, the text card will show a message telling the user to *Select a region*. If a region has been selected, then the text card will show the name of the selected region. The View can also be filtered by the *Typology* of museums by interacting with the *Filter by Typology* filter object or with the *Average Personnel by Typology* bar chart. Lastly, the *Denomination - Average Personnel* is an informative table that shows the best museums by *Personnel*. It can also be used to filter, but it is meant to be a static chart. The objects that cannot be used to filter are the numerical cards, of which one changes dynamically (the regional average of the

selected region) and the other stays constant (the Italian average), and the text card, which only shows what is the region that is selected. The two numerical cards and the bar chart below them are used to benchmark against the Italian average. The cards are colored with a sequential scale to visually facilitate benchmarking. The bar chart shows the Italian average as a dashed line, which stays constant, regardless of applied filters.

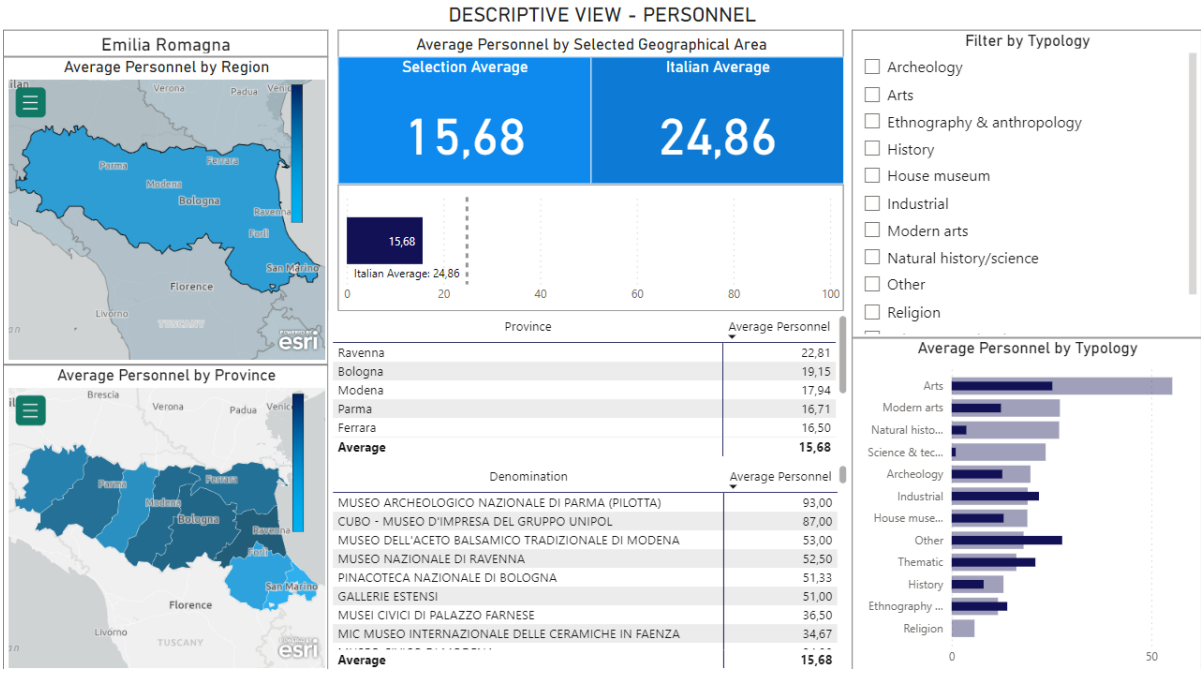


Figure 27 - Example of the View filtered by selecting Emilia Romagna as Region

Questions answered by the View:

- What is the average personnel of museums in Italy?
- What is the average personnel of museums in a specific region or province?
- How does the average personnel of a specific region or province compare to the Italian average personnel?
- How does the average personnel of a specific province compare to the average personnel of the region in which the province is located?
- What is the average personnel of a specific typology of museums in Italy (or in a specific province or region)?
- Which are the museums that have the most personnel in Italy (or in a specific province or region)?

## Descriptive View - Revenues from Tickets

This View has the objective of giving information about the number of visitors of the museums. This View is meant to be used to benchmark the average revenues between regions and Italy as a whole, and also to benchmark within the region.

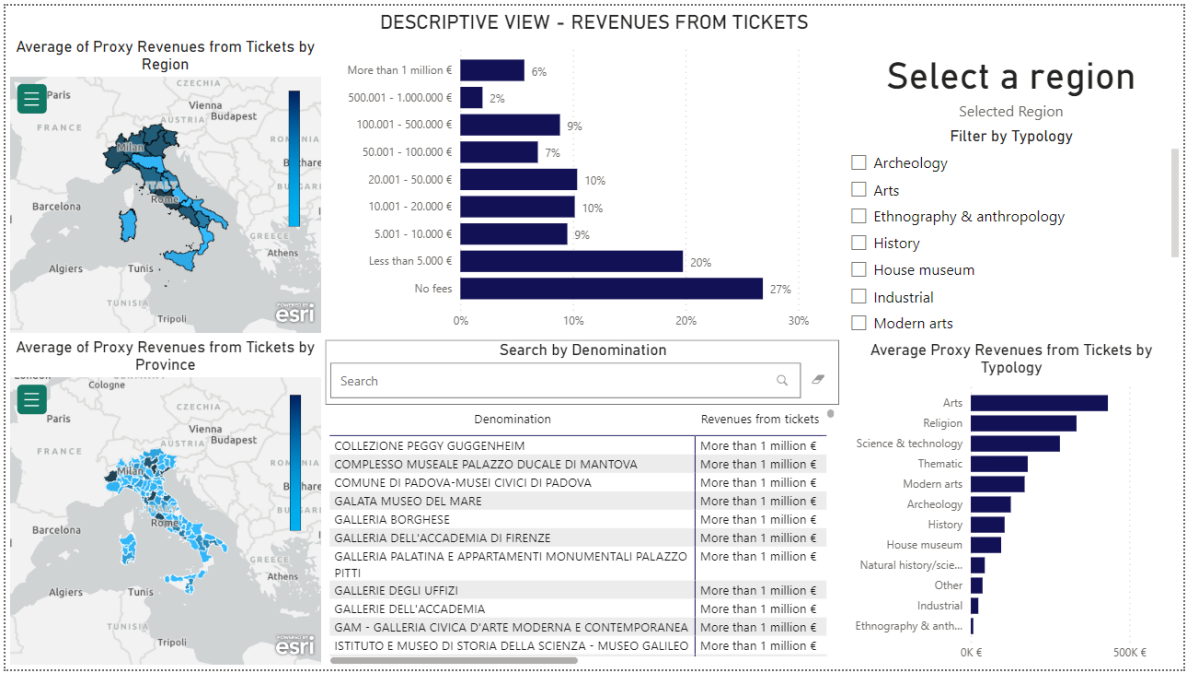


Figure 28 - Descriptive View - Revenues from Tickets

To compute the average revenues from tickets, a proxy variable is computed since in the dataset there are no actual revenue numbers, but only ranges. The proxy is computed by using the average value of the revenues in each range. The proxy indicator will be referred to as *Proxy Revenues from tickets* and to the original ranged indicator as *Ranged Revenues from tickets*. The View is composed of two choropleth maps (that show the average proxy revenue from tickets by region and *Province*), one table (showing the ranged revenues from tickets by museum), a text card (highlighting the selected region), a *filter* object (to filter by *Typology* of museum), and two bar charts (one showing the percentages of museums belonging to each range of ranged revenues from tickets, the other showing the average proxy revenue from tickets by *Typology* of museum).

The View can be filtered by region and *Province* by interacting with the choropleth maps. To make filtering by provinces easier, it is best to begin by filtering by region. This will cause the maps to zoom in on the selected region and thus make it easier to select the *Province* of choice. If a region has not yet been selected, the text card will show a message telling the user to *Select a region*. If a region has been selected, then the text card will show the name of the selected region. The View can be filtered by the *Typology* of museums by interacting with the *Filter by Typology* filter object or with the *Average Proxy Revenues from Tickets by Typology* bar chart. Moreover, the View can also

be filtered by the range of Ranged Revenues from Tickets by interacting with the related bar chart. When a region is selected, the bar charts will show the updated values, filtered on the selected region, while the Italian averages are kept constant behind the new bars, thus enabling benchmarking.

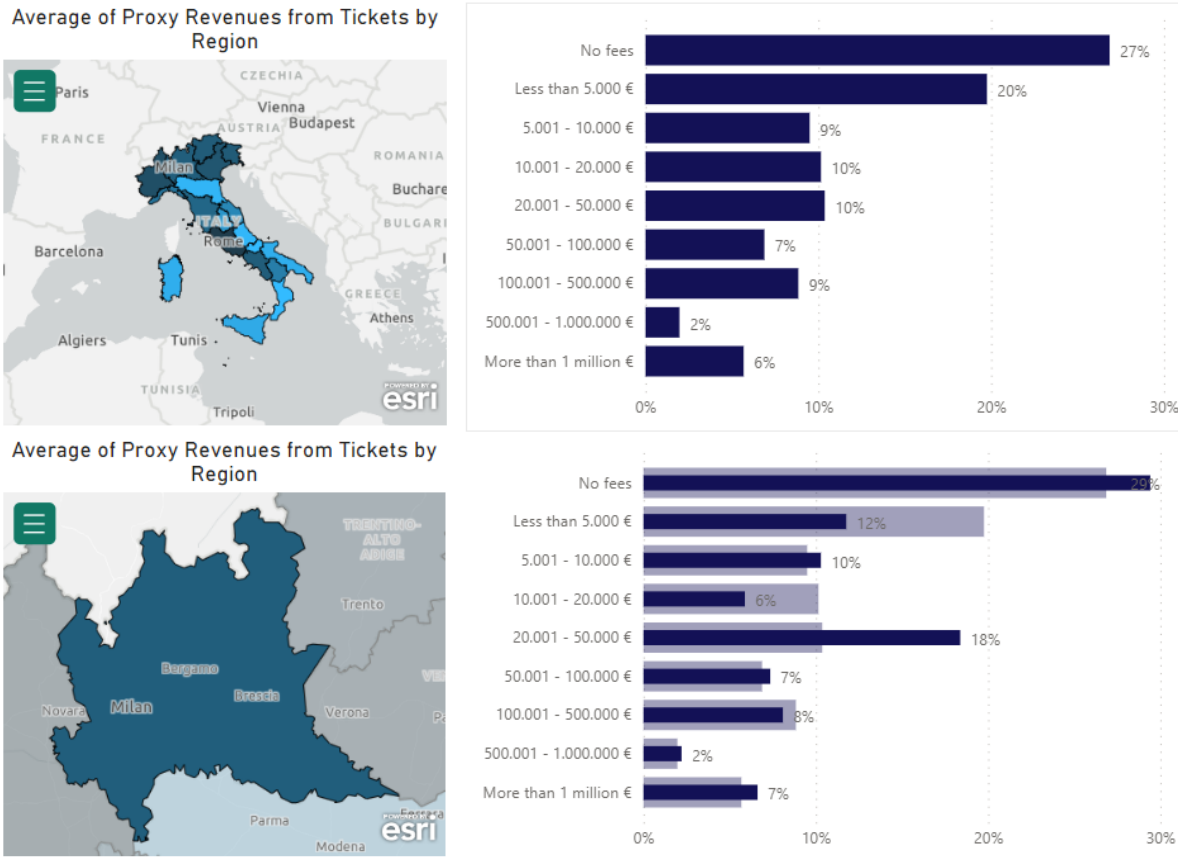


Figure 29 - Example of benchmarking, filtered by selection Lombardia as Region

Questions answered by the View:

- What is the average revenue earned from tickets by museums in Italy?
- What is the average revenue earned from tickets by museums in a specific region or province?
- How do the revenues earned from tickets by museums in a specific region or province compare to the Italian average?
- How does the average personnel of a specific province compare to the average personnel of the region in which the province is located?
- What is the average revenue from tickets of a specific typology of museums in Italy (or in a specific province or region)?
- Which are the museums that earn the most revenue from tickets in Italy (or in a specific province or region)?

## Descriptive View - Visitors

This View has the objective of giving information about the revenues earned by the museums from tickets. This View is meant to be used to benchmark the average visitors between regions and Italy as a whole, and also to benchmark within the region. This View is organized in the same way as the *Descriptive View - Revenues from Tickets*, as they share the same main features.

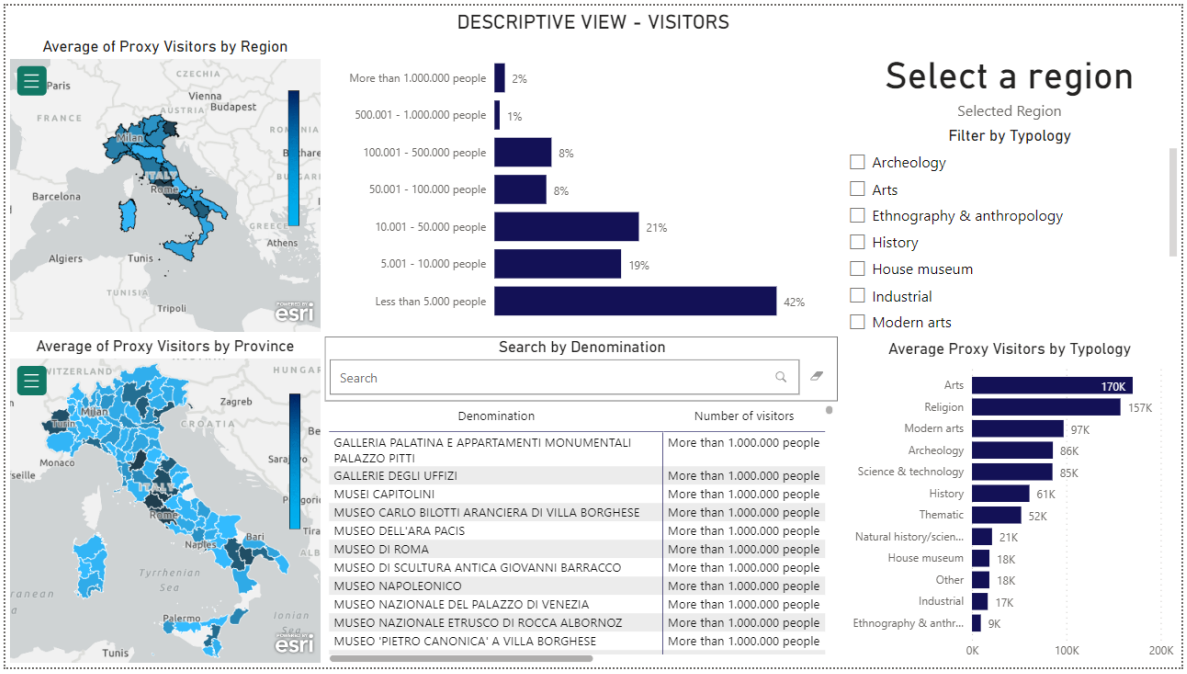


Figure 30 - Descriptive View - Visitors

To compute the average visitors, a proxy variable must be developed since in the dataset there are no actual visitor numbers, but only ranges. The proxy is computed by taking the average of the range as the value. The proxy indicator is referred to as *Proxy Visitors* and the original ranged indicator is referred to as *Ranged Visitors*. The View is composed of two choropleth maps (that show the average proxy visitors by region and *Province*), one table (showing the ranged visitors by museum), a text card (highlighting the selected region), a filter object (to filter by *Typology* of museum), and two bar charts (one showing the percentages of museums belonging to each range of ranged visitors, the other showing the average proxy visitors by *Typology* of museum).

The View can be filtered by region and *Province* by interacting with the choropleth maps. To make filtering by provinces easier, it is best to begin by filtering by region. This will cause the maps to zoom in on the selected region and thus make it easier to select the *Province* of choice. If a region has not yet been selected, the text card shows a message telling the user to *Select a region*. If a region has been selected, then the text card will show the name of the selected region. The View can be filtered by the *Typology* of museums by interacting with the *Filter by Typology* filter object or with the *Average Proxy Visitors by Typology* bar chart. Moreover, the View can also be filtered by the

range of Ranged Visitors by interacting with the related bar chart. When a region is selected, the bar charts show the updated values, filtered on the selected region, while the Italian averages are kept constant behind the new bars, thus enabling benchmarking, as shown before in Figure 29.

Questions answered by the View:

- What is the average number of visitors in Italy?
- What is the average number of visitors in a specific region or province?
- How does the average number of visitors in a specific region or province compare to the Italian average?
- How does the average number of visitors of a specific province compare to the average personnel of the region in which the province is located?
- What is the average number of visitors of a specific typology of museums in Italy (or in a specific province or region)?
- Which are the museums that have the highest number of visitors in Italy (or in a specific province or region)?

### Digital View - Online

This View has the objective of giving an overview of the Online KPI in Italy. The indicator is computed as a combination of digitalization of the collection, online ticketing, online tours, presence of a dedicated website, and social media presence. This View shows the evolution of the Online KPI over the four years in analysis. The View is meant to show the change over time in specific museums and to enable benchmarking between the museum and the average Online KPI in its region and of its *Typology*.

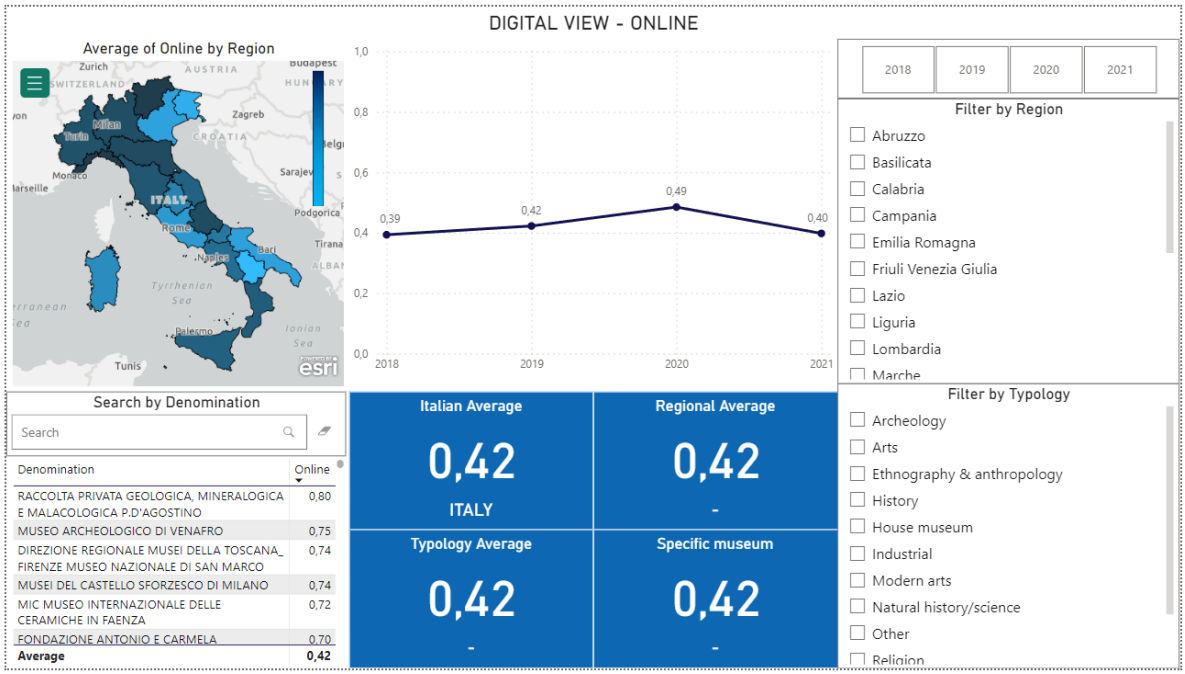


Figure 31 - Digital View - Online

The View is composed of a choropleth map (that shows the average Online KPI by region), a table (showing the Online KPI by museum), a line chart (that shows the evolution over time of the Online KPI), four numerical cards (highlighting the average Online KPI in Italy, the selected region, and the selected *Typology*, and the actual Online KPI for the selected museum), two filter objects (to filter by *Typology* of museum and region), and a slicer (to filter by a specific year).

The View can be filtered by region by interacting with the choropleth map and with the *Filter by Region* object. The View can be filtered by the *Typology* of museums by interacting with the *Filter by Typology* filter object. The View can also be filtered by year, transforming it from an overview of the evolution of the Online KPI to a static visualization of a single year. The four numerical cards enable benchmarking as the user can see the differences in the average Online KPI between Italian, regional, and specific typologies.

Questions answered by the View:

- What is the average Online KPI in Italy?
- What is the average Online KPI in a specific region?
- How does the Online KPI in a specific region compare to the Italian average?
- What is the average Online KPI of museums of a specific typology?
- How does the Online KPI of museums of a specific typology compare to the Italian average?
- How does the Online KPI of a specific museum compare to the average KPI of its region?
- How does the Online KPI of a specific museum compare to the average KPI of museums of the same typology?
- Which are the museums that have the highest Online KPI in Italy?
- How did the average Online KPI change over time?
- How did the Online KPI for a specific museum change over time?
- How was the Online KPI situation in a specific year?

### Digital View - On-site

This View has the objective of giving an overview of the On-site KPI in Italy. The indicator is computed as a combination of five on-site digital technologies (audioguide, augmented reality, virtual reality, QR code, chatbot). This View shows the evolution of the On-site KPI over the four years in analysis. The View is meant to show the change over time in specific museums and to enable benchmarking between the museum and the average On-site KPI in its region and of its *Typology*.

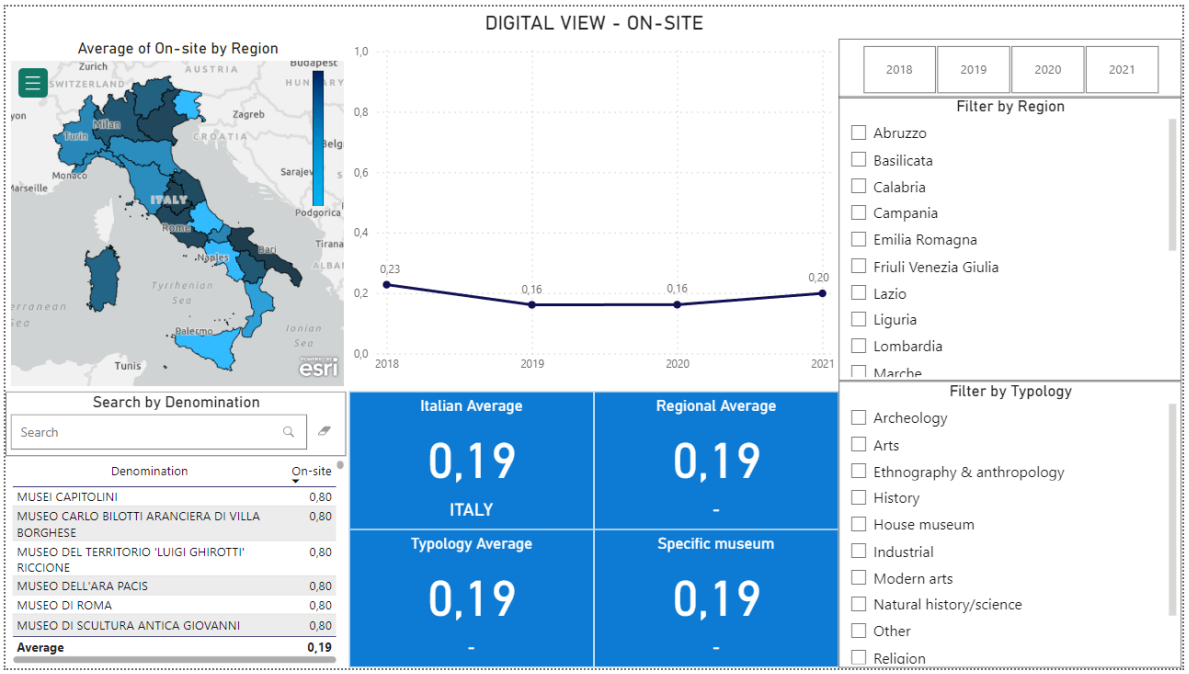


Figure 32 - Digital View - On-site

The View is composed of a choropleth map (that shows the average On-site KPI by region), a table (showing the On-site KPI by museum), a line chart (that shows the evolution over time of the On-site KPI), four numerical cards (highlighting the average On-site KPI in Italy, the selected region, and the selected *Typology*, and *the actual On-site KPI* for the selected museum), two filter objects (to filter by *Typology* of museum and region), and a slicer (to filter by a specific year).

The View can be filtered by region by interacting with the choropleth map and with the *Filter by Region* object. The view can be filtered by the *Typology* of museums by interacting with the *Filter by Typology* filter object. The view can also be filtered by year, transforming it from an overview of the evolution of the On-site KPI to a static view of a single year. The four numerical cards enable benchmarking as the user can see the differences in the average On-site KPI between Italian, regional, and specific typologies.

Questions answered by the View:

- What is the average On-site KPI in Italy?
- What is the average On-site KPI in a specific region?
- How does the On-site KPI in a specific region compare to the Italian average?
- What is the average On-site KPI of museums of a specific typology?
- How does the On-site KPI of museums of a specific typology compare to the Italian average?
- How does the On-site KPI of a specific museum compare to the average KPI of its region?
- How does the On-site KPI of a specific museum compare to the average KPI of museums of the same typology?
- Which are the museums that have the highest On-site KPI in Italy?
- How did the average On-site KPI change over time?
- How did the On-site KPI for a specific museum change over time?
- How was the On-site KPI situation in a specific year?

### Digital View - Organizational Readiness

This View has the objective of giving an overview of the Organizational Readiness KPI in Italy. The indicator is computed as a combination of two variables: Presence of a Digital Innovation Plan and Workers with Digital Competencies. This View shows the evolution of the Organizational Readiness KPI over the four years in analysis. The View is meant to show the change over time in specific museums and to enable benchmarking between the museum and the average Organizational Readiness KPI in its region and of its *Typology*.

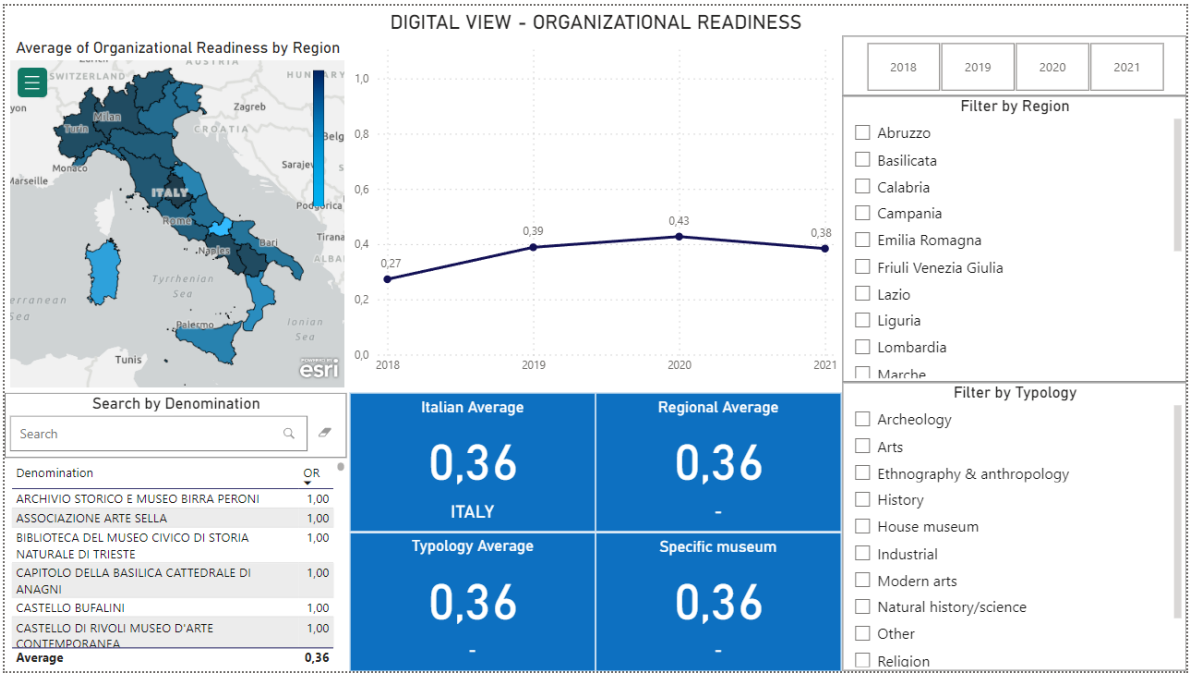


Figure 33 - Digital View - Organizational Readiness

The View is composed of a choropleth map (that shows the average Organizational Readiness KPI by region), a table (showing the Organizational Readiness KPI by museum), a line chart (that shows the evolution over time of the Organizational Readiness KPI), four numerical cards (highlighting the average Organizational Readiness KPI in Italy, the selected region, and the selected *Typology*, and the actual Organizational Readiness KPI for the selected museum), two filter objects (to filter by *Typology* of museum and region), and a slicer (to filter by a specific year).

The View can be filtered by region by interacting with the choropleth map and with the *Filter by Region* object. The View can be filtered by the *Typology* of museums by interacting with the *Filter by Typology* filter object. The View can also be filtered by year by interacting with either the slicer or the line chart, transforming it from a view of the evolution of the Organizational Readiness KPI to a static visualization of a single year. The four numerical cards enable benchmarking as the user can see the differences in the average Organizational Readiness KPI between Italian, regional, and specific typologies.

Questions answered by the View:

- What is the average Organizational Readiness KPI in Italy?
- What is the average Organizational Readiness KPI in a specific region?
- How does the Organizational Readiness KPI in a specific region compare to the Italian average?
- What is the average Organizational Readiness KPI of museums of a specific typology?
- How does the Organizational Readiness KPI of museums of a specific typology compare to the Italian average?
- How does the Organizational Readiness KPI of a specific museum compare to the average KPI of its region?
- How does the Organizational Readiness KPI of a specific museum compare to the average KPI of museums of the same typology?
- Which are the museums that have the highest Organizational Readiness KPI in Italy?
- How did the average Organizational Readiness KPI change over time?
- How did the Organizational Readiness KPI for a specific museum change over time?
- How was the Organizational Readiness KPI situation in a specific year?

### Digital View - Evolution Average

This View has the objective of showing the evolution of the Average Digital (AD) KPI in Italy. The indicator is computed as the average of the three Digital KPIs: Online, On-site, and Organizational Readiness. This View shows the evolution of the AD KPI over the four years in analysis. The View is meant to show the change over time in specific museums and to enable benchmarking between the museum and the AD KPI in its region and of its *Typology*.

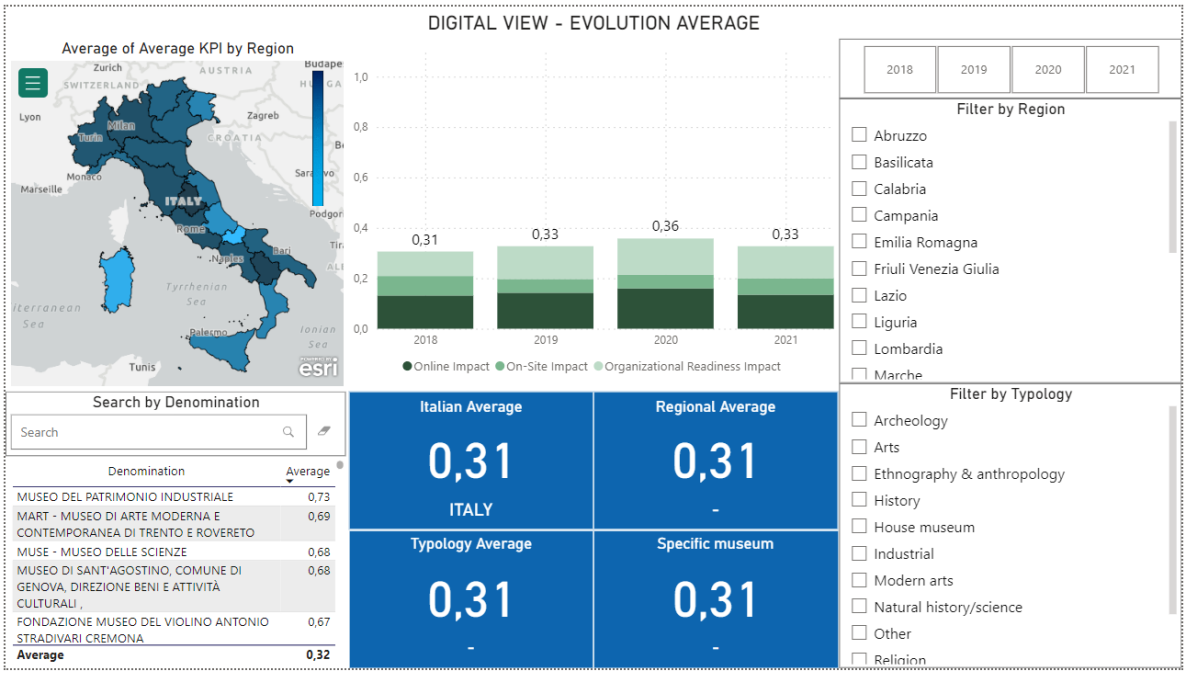


Figure 34 - Digital View - Evolution Average

The View is composed of a choropleth map (that shows the average AD KPI by region), a table (showing the AD KPI by museum), a stacked column chart (that shows the evolution over time of the AD KPI, with a visual representation of the three Digital KPIs that it is composed of), four numerical cards (highlighting the average AD KPI in Italy, the selected region, and the selected *Typology*, and the actual AD KPI for the selected museum), two filter objects (to filter by *Typology* of museum and region), and a slicer (to filter by a specific year).

The View can be filtered by region by interacting with the choropleth map and with the *Filter by Region* object. The View can be filtered by the *Typology* of museums by interacting with the *Filter by Typology* filter object. The View can also be filtered by year by interacting with either the slicer or the stacked column chart, transforming it from an overview of the evolution of the AD KPI to a static View of a single year. The four numerical cards enable benchmarking as the user can see the differences in the average AD KPI between Italian, regional, and specific typologies. The stacked column chart visually represents the impact of the three Digital KPIs. The actual values of the Digital KPIs can be seen in the next View, i.e., *Digital View - Positioning*.

Questions answered by the View:

- What is the average AD KPI in Italy?
- What is the average AD KPI in a specific region?
- What is the impact of the three Digital KPIs for a specific museum?
- How does the AD KPI in a specific region compare to the Italian average?
- What is the average AD KPI of museums of a specific typology?
- How does the AD KPI of museums of a specific typology compare to the Italian average?
- How does the AD KPI of a specific museum compare to the average KPI of its region?
- How does the AD KPI of a specific museum compare to the average KPI of museums of the same typology?
- Which are the museums that have the highest AD KPI in Italy?
- How did the average AD KPI change over time?
- How did the AD KPI for a specific museum change over time?
- How was the AD situation in a specific year?

## Digital View - Positioning

This View has the objective of showing the positioning of the museums based on the three Digital KPIs in Italy. The View is meant to show visually where specific museums are positioned with respect to others. Moreover, benchmarking between the museums is done using quantiles and a ranking.

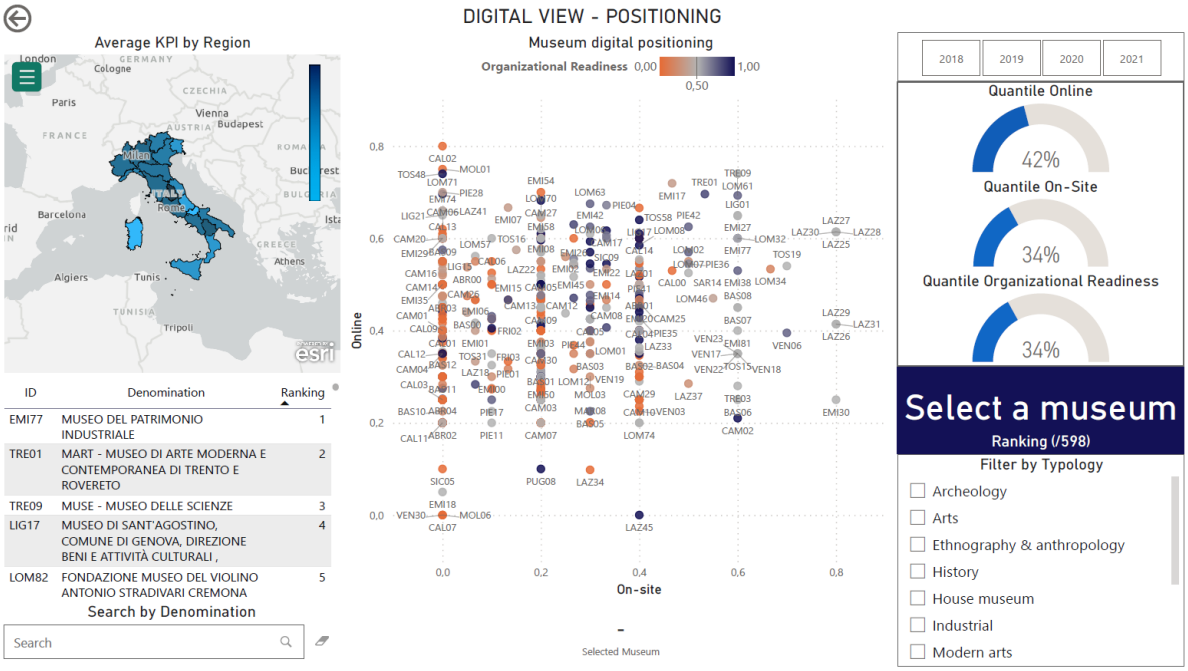


Figure 35 - Digital View - Positioning

The View is composed of a choropleth map (that shows the average AD KPI by region), a table (showing the ID, denomination, and ranking of museums), a scatter chart with color mapping (that incorporates the dimension of the three Digital KPIs on a 2D chart by using color to encode the third dimension), three gauges (highlighting the quantile of the specific museum, based on the three Digital KPIs), two filter objects (to filter by *Typology* of museum and region), a numerical card (showing the ranking of the selected museum), a text card (showing the denomination of the selected museum), and a text filter (allowing the search of museums).

The View can be filtered by region by interacting with the choropleth map, by the *Typology* of museums by interacting with the *Filter by Typology* filter object, and by year by interacting with the slicer, transforming it from a view of the average of the three Digital KPIs to a View of a single year. The three gauges enable benchmarking as the user can see how the museums are positioned with respect to the others, using as dimensions the three Digital KPIs. Moreover, the numerical card shows the ranking of the selected museum. The *bubble* chart visually represents the museums on the dimensions of the three Digital KPIs. The actual values of the Digital KPIs can be seen by hovering over the museum of interest. In the chart, the museums are represented by IDs (first three letters of the region + two numbers), this is done to avoid long

denominations that would make the chart chaotic. To access the information about ranking and quantiles, the museum can be selected with the table, which also assigns IDs to denominations. The denomination of the museum can also be searched with the Text filter object, below the table.

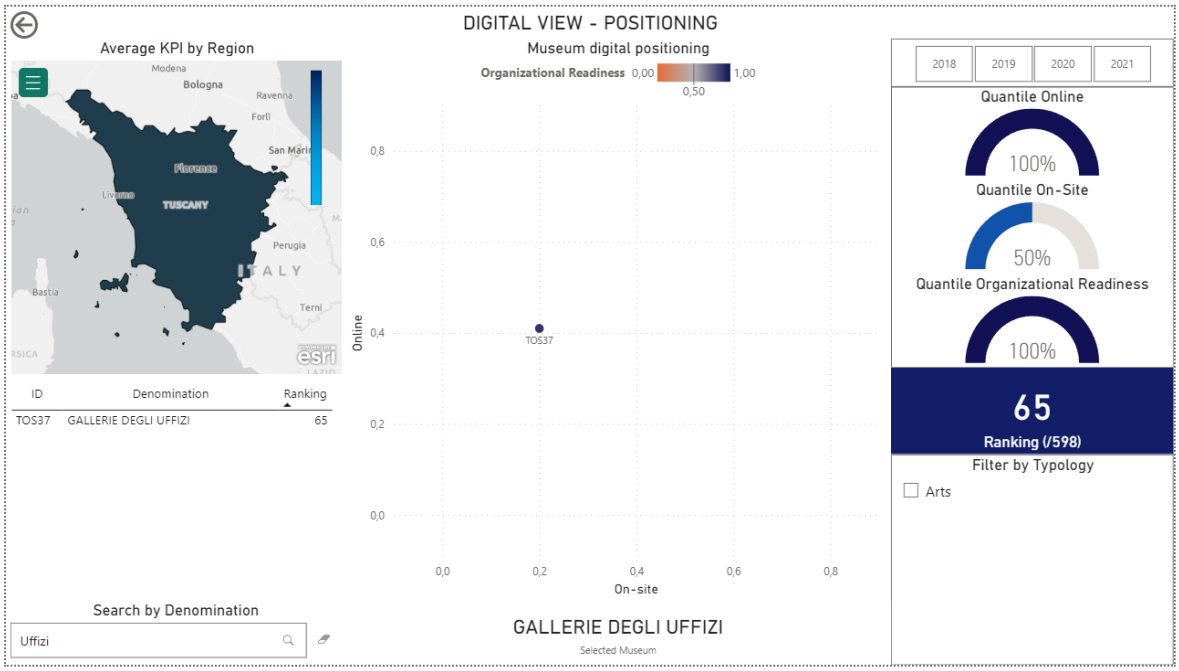


Figure 36 - Example of the View, filtered by searching *Uffizi*

Questions answered by the View:

- What is the average AD KPI in a specific region?
- What are the values of the three Digital KPIs for a specific museum?
- How does a specific museum rank with respect to the others?
- Where is a museum positioned based on the three Digital KPIs?
- How does a specific museum rank with respect to other museums in the same region? Or of the same typology?
- How good is a museum in the Online KPI, with respect to the others?
- How good is a museum in the On-site KPI, with respect to the others?
- How good is a museum in the Organizational Readiness KPI, with respect to the others?
- How was the situation in a specific year?

# Chapter 7: Findings

This Chapter aims to act on, with practical implications, the objective of the thesis, which is to show how the dimensions of the IDM Framework for Museums are connected among each other and how they connect to data integration.

Specifically, this chapter shows the practical implications and the effects of the data integration of open data with proprietary data for two decision makers. These two sources are represented in the empirical setting of the thesis by the the Istat Microdata dataset and the Proprietary dataset of the Observatory. This is done providing a comprehensive explanation regarding the significance of the dashboard based on integrated data and of its application for both the intended users of the dashboard: museum managers and the Ministry of Culture. The objective of this chapter is therefore to answer the following questions: How does the integration of open data affect decision-making in museums? How does the integrated dashboard improve decision-making for museums and for the Ministry of Culture?

The Chapter elaborates on the value that the specific integrated dashboard developed brings to museum managers and the Ministry of Culture, focusing on the concepts of internal and external benchmarking (Saul, 2004) and showing the alignment of the findings with the IDM Framework for Museums. This is corroborated by the visual representation of the added value of the integration of open data on the Views is shown, highlighting its practical implications.

## 7.1. The value added by each View

In this section, a visual representation of the value added obtained thanks to the integration of open data is provided, showing how each View of the dashboard would have looked like without the integration of open data. Additionally, a table displaying what are the variables added thanks to the integration of open data is included. Subsequently, the added value that the Views provide to museum managers and the Ministry of Culture is clarified, based on the question answered by the Views, presented above, in Chapter 6.

### Descriptive View

This View has the objective of showing an initial general overview of the dataset. In this View, the user can find information about the demographics of museums, meaning that the geographical location (Region - *Province* - *Municipality*) of all the museums are showcased. The user can obtain information about the number of museums per Region and by *Typology*. Since this is the first View, the user can learn about the sample of museums per Region and the Typologies that are considered in this dashboard. The map shows a purely visual representation of the number of museums per Region, while the bar chart is more precise, showing the actual values.

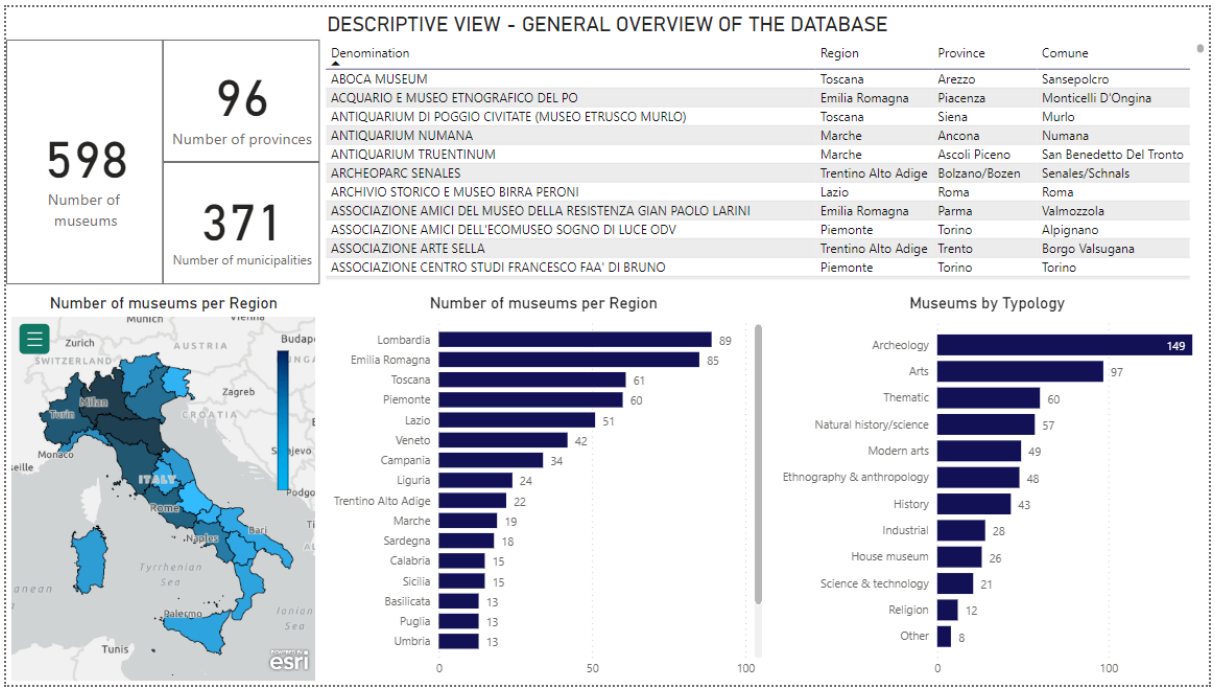


Figure 37 - Descriptive View

| **Question answered by the View** | **Value added for museum managers** | **Value added for the Ministry of Culture** |
|---|---|---|
| How many museums are in the unified *dataset*? | Information about the sample | Information about the sample |
| In how many provinces and municipalities are the museums located? | Information about the sample | Information about the sample |
| Where is a specific museum located (region, province, municipality)? | / | Opportunity to conduct personalized or tailored analysis for individual museums |
| How many museums are located in a set region? | Understanding the number of museums located in their vicinity | Information about geographical distribution of museums |
| What are the typologies of museums? | Information about the sample | Information about the sample |
| How many museums are of a particular typology? Where are they located? | Understanding the number of museums that share similar features to the museum they manage | Information about the overall typology distribution of museums |

Table 60 - Questions answered and value added of the Descriptive View

The integration of open data has enabled museum managers and the Ministry of Culture to assess and benchmark institutions across dimensions that were inaccessible previous to the integration. This value added can be appreciated in every View. Specifically to this View, the integration has improved the precision of benchmarking on the geographical dimension by allowing a breakdown into Region - *Province* - *Municipality*, consequently refining the analysis granularity, and allowing users an increased flexibility in their analysis. Without this integration, relying solely on the broader Region dimension severely limits precision in analysis. Additionally, the *Typology* dimension holds significance in benchmarking since museums belonging to distinct typologies may have different features that lead to different performance. For a museum manager, the possibility of benchmarking its museum against museums that belong to the same *Typology* is important as it greatly refines the similarity of museums, thus allowing for better external benchmarking. Figure 38 shows visually the information that would be lost without data integration of open data and Table 61 summarizes the variables added to the View solely thanks to the integration of open data.

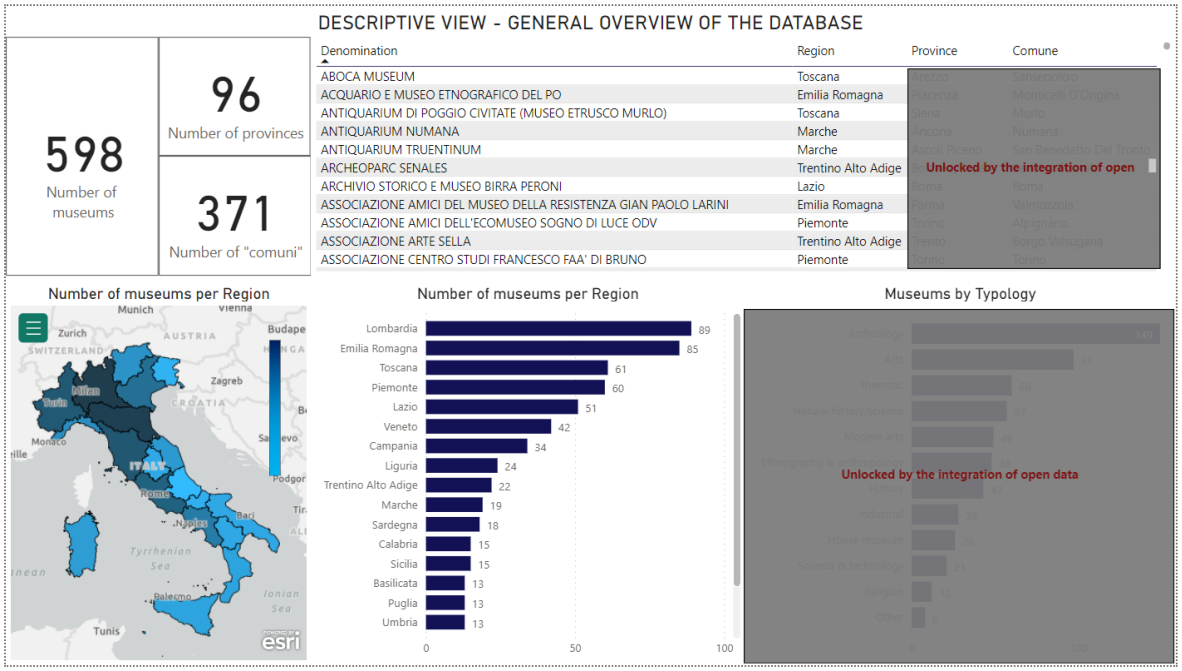


Figure 38 - Descriptive View without open data

| View | Variable added through data integration |
|---|---|
| Descriptive View | Province |
| | Municipality |
| | Typology |

Table 61 - Variables added through data integration in the Descriptive View

### Descriptive View - Personnel

This View has the objective of giving information about the situation of the Number of Workers (*Personnel)* in museums. In this View, the user can obtain information about the precise *Personnel* employed in museums. The View shows, through various visual objects, the distribution of the *Personnel* in museums over the Italian territory. It is structured to facilitate benchmarking, allowing users the flexibility to focus either on a regional level or delve deeper into a provincial level. As per all the other Views, the user can also filter by *Typology*.

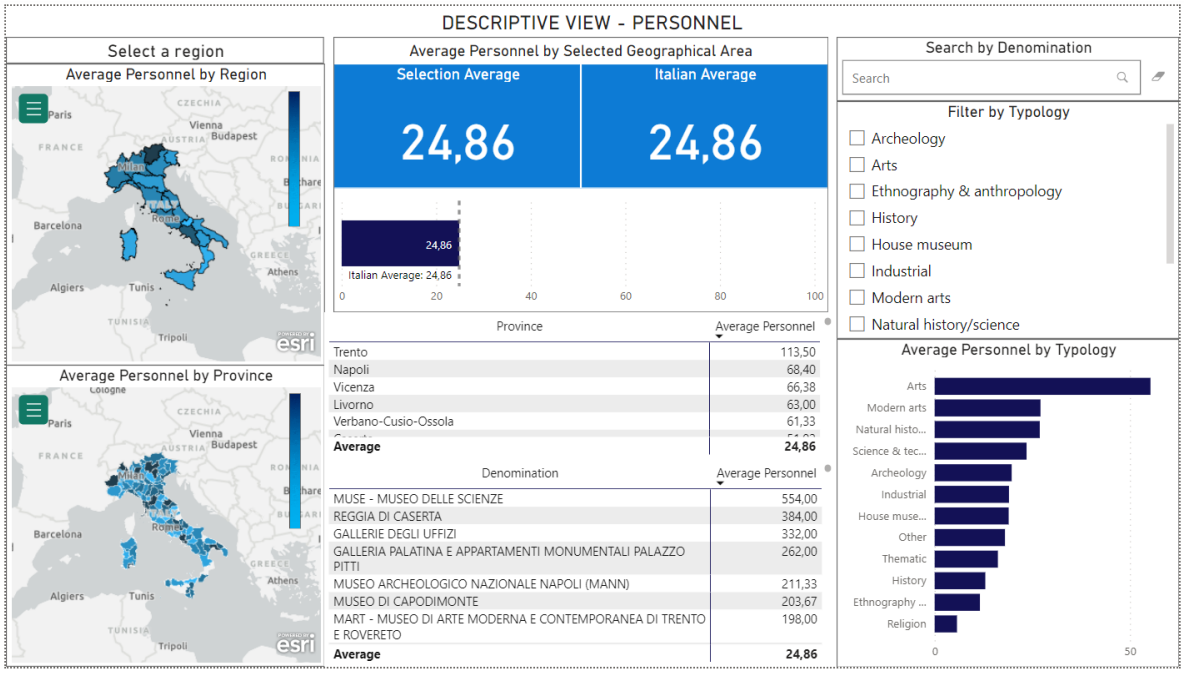


Figure 39 - Descriptive View - Personnel

| **Question answered by the View** | **Value added for museum managers** | **Value added for the Ministry of Culture** |
|---|---|---|
| What is the average *Personnel* of museums in Italy? | Possibility of comparing their museum's *Personnel* against the Italian average | Opportunity to conduct personalized or tailored analysis for specific geographical areas |
| What is the average *Personnel* of museums in a specific region or province? | Possibility of comparing the museum's *Personnel* against the regional or provincial average | Opportunity to conduct personalized or tailored analysis for specific geographical areas |
| How does the average *Personnel* of a specific region or province compare to the Italian average *Personnel*? | Possibility of comparing the museum's *Personnel* against the regional or provincial average | Opportunity to conduct personalized or tailored analysis for specific geographical areas |
| How does the average *Personnel* of a specific province compare to the average *Personnel* of the region in which the province is located? | / | Opportunity to conduct personalized or tailored analysis for specific geographical areas |
| What is the average *Personnel* of a specific *Typology* of museums in Italy (or in a specific province or region)? | Possibility of comparing the museum's *Personnel* against the *Typology* average | Opportunity to conduct personalized or tailored analysis for specific typologies of museums |
| Which are the museums that have the most *Personnel* in Italy (or in a specific province or region)? | Identifying the top performers in terms of *Personnel* and how their museum compares against them; Possibility of implementing new practices by emulation | Identifying the top performers in terms of *Personnel*; Possibility of promoting the adoption of best practices to non-top performing museums |

Table 62 - Questions answered and value added of the Descriptive View - Personnel

The integration has improved the precision of benchmarking on the geographical dimension by allowing a breakdown into Region - *Province*, consequently refining the analysis granularity, and allowing users an increased flexibility in their analysis. Without this integration, relying solely on the broader Region dimension severely limits precision in analysis. Additionally, the *Typology* dimension holds significance in benchmarking since museums belonging to distinct typologies may have different features that lead to different performance. For a museum manager, the possibility of benchmarking its museum against museums that belong to the same *Typology* is important as it greatly refines the similarity of museums, thus allowing for better external benchmarking. Furthermore, although the *Personnel* dimension is already available in the original dataset, the integration brings an improvement by accessing *Personnel* information in numerical form rather than the range form, found in the original dataset. This enhancement ensures that benchmarking can be done on personalized ranges (instead of predefined ranges) that can better reflect the different needs of the users. Figure 40 shows visually the information that would be lost without

data integration of open data and Table 63 summarizes the variables added to the View solely thanks to the integration of open data.

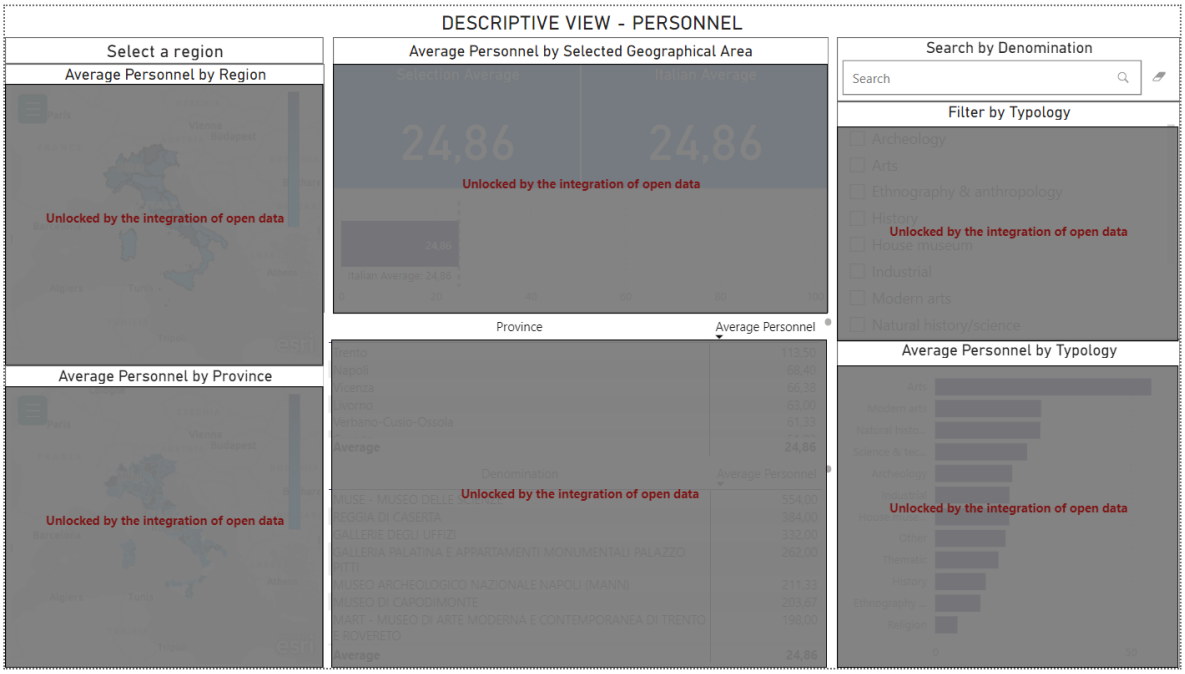


Figure 40 - Descriptive View - Personnel without open data

| View | Variable added through data integration |
|---|---|
| Descriptive View – Personnel | Province |
| | Typology |
| | Personnel |

Table 63 - Variables added through data integration in the Descriptive View - Personnel

### Descriptive View - Revenue from tickets

This View has the objective of giving information about the situation of the Revenue from Tickets in museums. In this View, the user can obtain information about range of revenues earned from tickets by museums. The View shows, through various visual objects, the distribution of the Revenue from Tickets earned in museums over the Italian territory. It is structured to facilitate benchmarking, allowing users the flexibility to focus either on a regional level or delve deeper into a provincial level. As per all the other Views, the user can also filter by *Typology*.

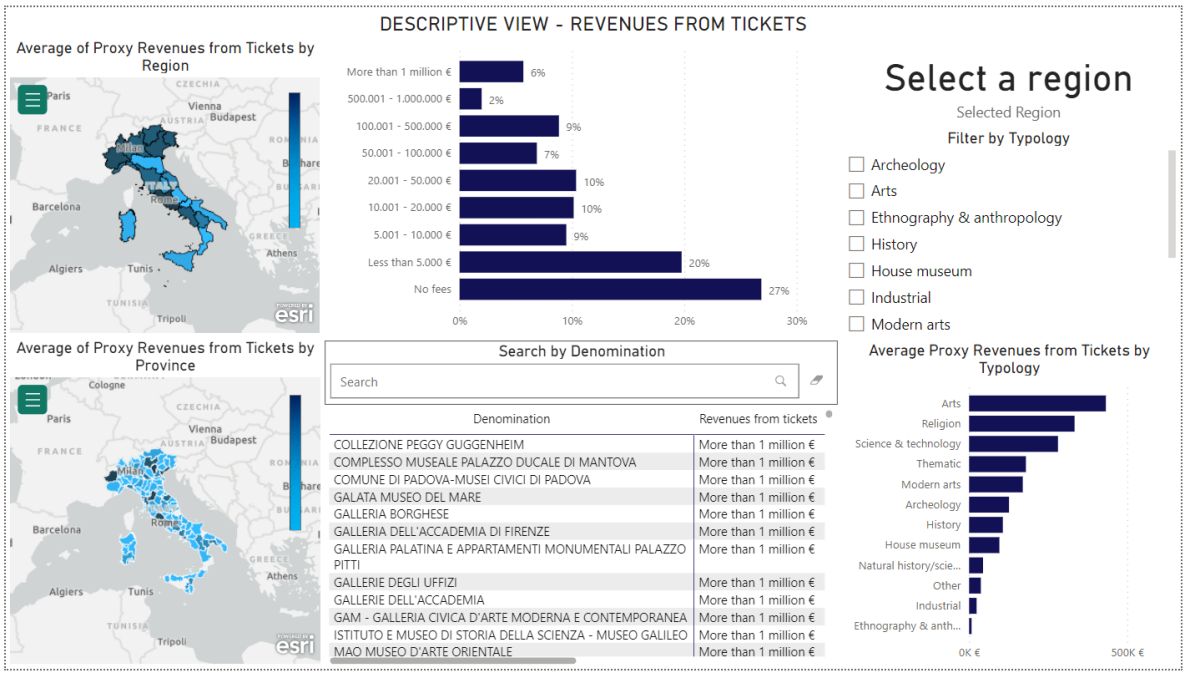


Figure 41 - Descriptive View - Revenue from tickets

| **Questions answered by the View** | **Value added for museum managers** | **Value added for the Ministry of Culture** |
|---|---|---|
| What is the average revenue earned from tickets by museums in Italy? | Possibility of comparing their museum's Revenue from Tickets against the Italian average | Gain insights into sector-wide performance |
| What is the average revenue earned from tickets by museums in a specific region or province? | Possibility of comparing the museum's Revenue from Tickets against the regional or provincial average | Opportunity to conduct personalized or tailored analysis for specific geographical areas |
| How do the revenues earned from tickets by museums in a specific region or province compare to the Italian average? | Possibility of comparing the museum's Revenue from Tickets against the regional or provincial average | Opportunity to conduct personalized or tailored analysis for specific geographical areas |
| How does the average Revenue from Tickets of a specific province compare to the average Revenue from Tickets of the region in which the province is located? | / | Opportunity to conduct personalized or tailored analysis for specific geographical areas |
| What is the average revenue from tickets for a specific typology of museums in Italy (or in a specific province or region)? | Possibility of comparing the museum's Revenue from Tickets against the *Typology* average | Opportunity to conduct personalized or tailored analysis for specific typologies of museums |
| Which are the museums that earn the most revenue from tickets in Italy (or in a specific province or region)? | Identifying the top performers in terms of Revenue from Tickets and how their museum compares against them; Possibility of implementing new practices by emulation | Identifying the top performers in terms of Revenue from Tickets; Possibility of promoting the adoption of best practices to non-top performing museums |

Table 64 - Questions answered and value added of the Descriptive View - Revenue from tickets

The integration has improved the precision of benchmarking on the geographical dimension by allowing a breakdown into Region - *Province*, consequently refining the analysis granularity, and allowing users an increased flexibility in their analysis. Without this integration, relying solely on the broader Region dimension severely limits precision in analysis. Additionally, the *Typology* dimension holds significance in benchmarking since museums belonging to distinct typologies may have different features that lead to different performance. For a museum manager, the possibility of benchmarking its museum against museums that belong to the same *Typology* is important as it greatly refines the similarity of museums, thus allowing for better external benchmarking. Figure 42 shows visually the information that would be lost

without data integration of open data and Table 65 summarizes the variables added to the View solely thanks to the integration of open data.

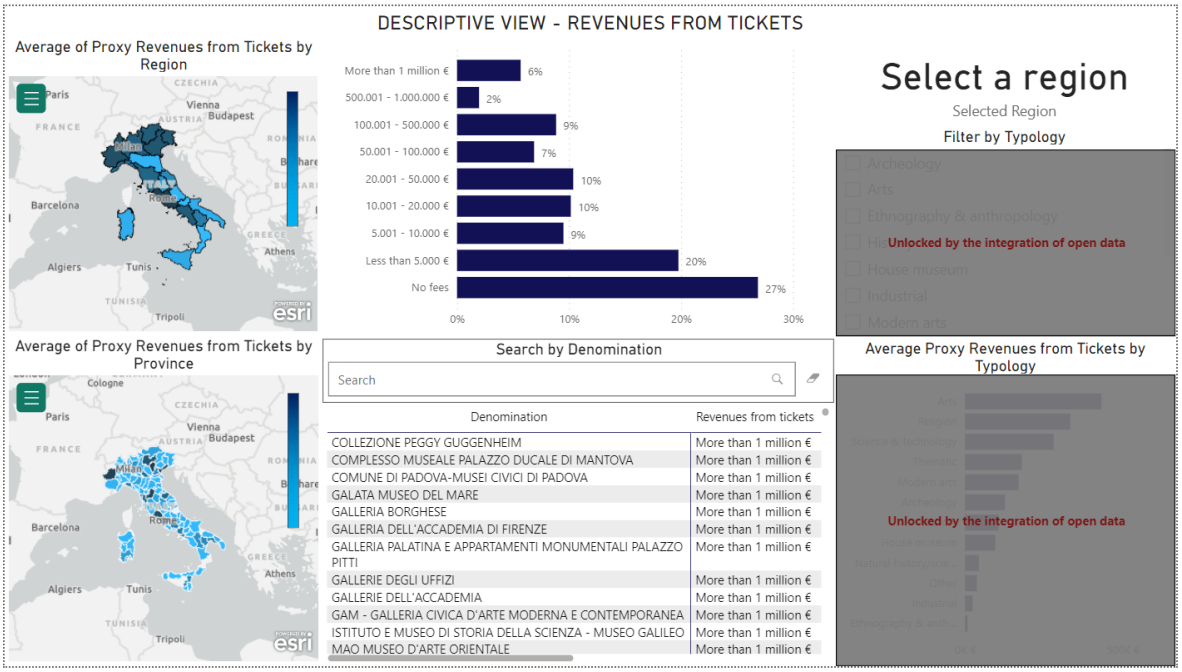


Figure 42 - Descriptive View - Revenue from tickets without open data

| View | Variable added through data integration |
|---|---|
| Descriptive View – Revenues | Province |
| | Typology |

Table 65 - Variables added through data integration in the Descriptive View - Revenue from tickets

## Descriptive View - Visitors

This View has the objective of giving information about the situation of the Visitors of museums. In this View, the user can obtain information about range of Visitors in museums. The View shows, through various visual objects, the distribution of the Visitors that visit museums over the Italian territory. It is structured to facilitate benchmarking, allowing users the flexibility to focus either on a regional level or delve deeper into a provincial level. As per all the other Views, the user can also filter by *Typology*.

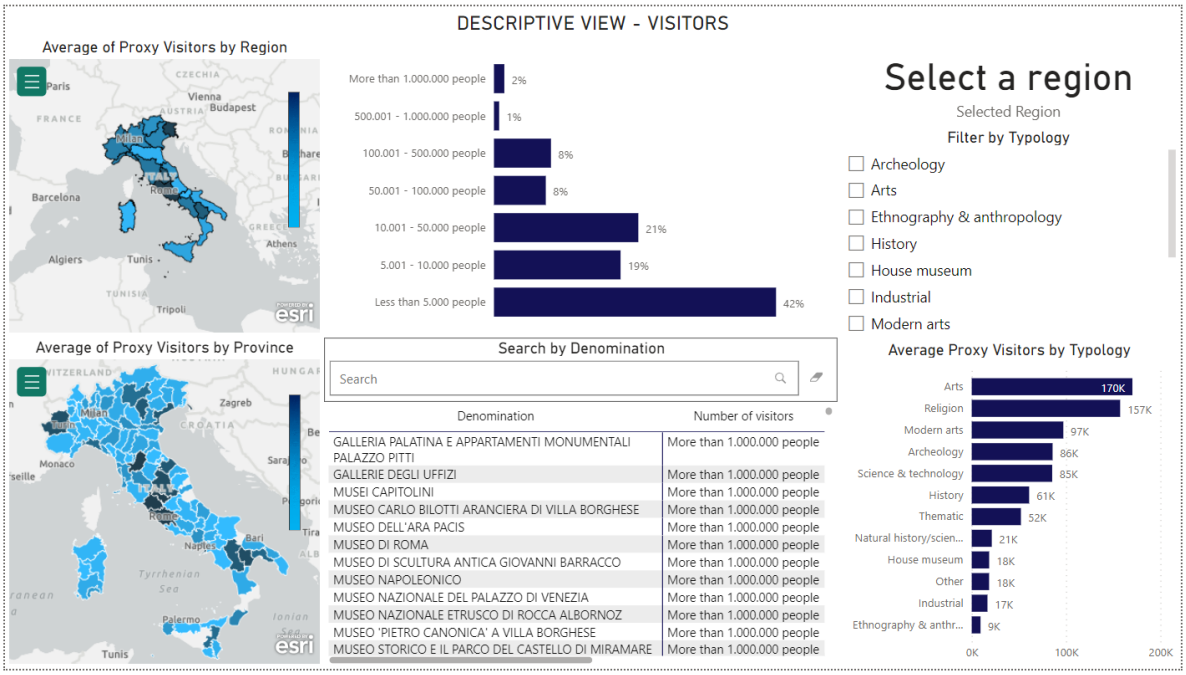


Figure 43 - Descriptive View - Visitors

| **Questions answered by the View** | **Value added for museum managers** | **Value added for the Ministry of Culture** |
|---|---|---|
| What is the average Number of Visitors in Italy? | Possibility of comparing their museum's Number of Visitors against the Italian average | Gain insights into sector-wide performance |
| What is the average Number of Visitors in a specific region or province? | Possibility of comparing the museum's Number of Visitors against the regional or provincial average | Opportunity to conduct personalized or tailored analysis for specific geographical areas |
| How does the average Number of Visitors in a specific region or province compare to the Italian average? | Possibility of comparing the museum's Number of Visitors against the regional or provincial average | Opportunity to conduct personalized or tailored analysis for specific geographical areas |
| How does the average Number of Visitors of a specific province compare to the average personnel of the region in which the province is located? | / | Opportunity to conduct personalized or tailored analysis for specific geographical areas |
| What is the average Number of Visitors of a specific *Typology* of museums in Italy (or in a specific province or region)? | Possibility of comparing the museum's Number of Visitors against the *Typology* average | Opportunity to conduct personalized or tailored analysis for specific typologies of museums |
| Which are the museums that have the highest Number of Visitors in Italy (or in a specific province or region)? | Identifying the top performers in terms of Number of Visitors and how their museum compares against them; Possibility of implementing new practices by emulation | Identifying the top performers in terms of Number of Visitors; Possibility of promoting the adoption of best practices to non-top performing museums |

Table 66 - Questions answered and value added of the Descriptive View - Visitors

The integration has improved the precision of benchmarking on the geographical dimension by allowing a breakdown into Region - *Province*, consequently refining the analysis granularity, and allowing users an increased flexibility in their analysis. Without this integration, relying solely on the broader Region dimension severely limits precision in analysis. Additionally, the *Typology* dimension holds significance in benchmarking since museums belonging to distinct typologies may have different features that lead to different performance. For a museum manager, the possibility of benchmarking its museum against museums that belong to the same *Typology* is important as it greatly refines the similarity of museums, thus allowing for better external benchmarking. Figure 44 shows visually the information that would be lost

without data integration of open data and Table 67 summarizes the variables added to the View solely thanks to the integration of open data.

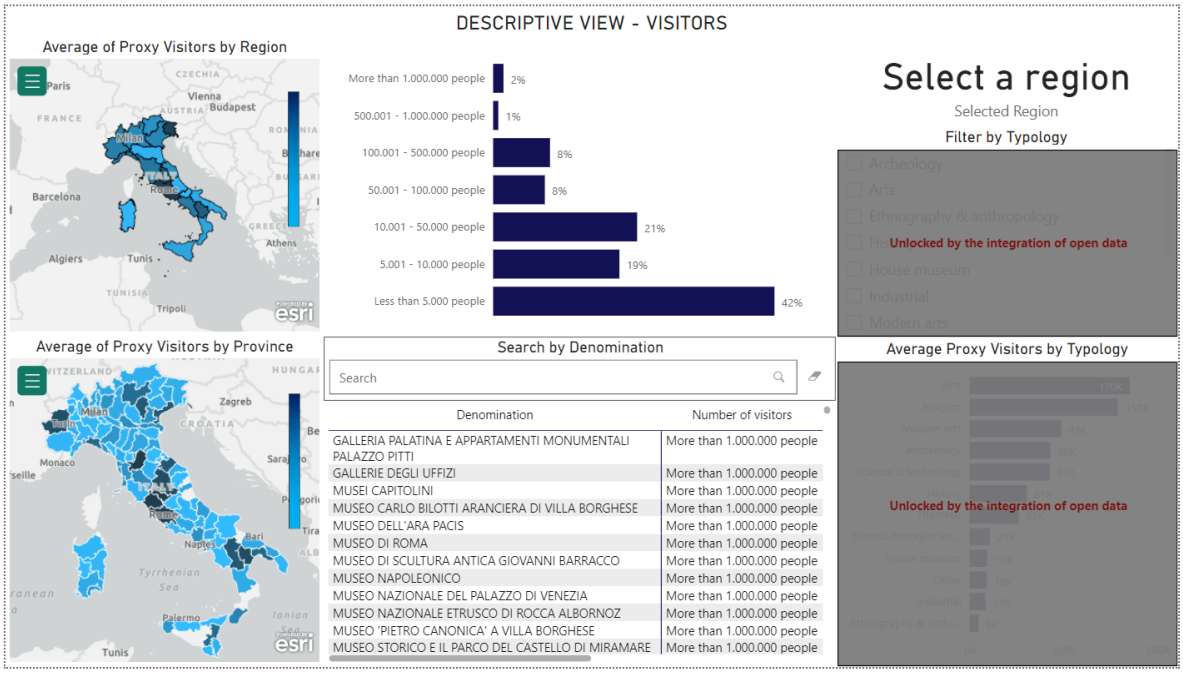


Figure 44 - Descriptive View - Visitors without open data

| View | Variable added through data integration |
|---|---|
| Descriptive View – Visitors | Province |
| | Typology |

Table 67 - Variables added through data integration in the Descriptive View - Visitors

### Digital View - Online

This View has the objective of giving an overview of the Online KPI in Italy. The indicator is computed as a combination of digitalization of the collection, online ticketing, online tours, presence of a dedicated website, and social media presence. The View shows the evolution of the Online KPI over time. The View shows, through various visual objects, the distribution of the Online KPI in the Italian territory and over time. It is structured to facilitate benchmarking, allowing users the flexibility to focus on a specific Region, *Typology* and/or year. In this View, users are empowered with the flexibility to choose between focusing on a specific year or exploring the evolution of the indicator over time. This feature allows users to tailor their analysis based on their preferences and requirements. Whether considering a particular year's data for in-depth insights or observing the trend and changes in the indicator across multiple years, this functionality provides users with a comprehensive understanding of the temporal aspect of the indicator. As per all the other Views, the user can also filter by *Typology*.

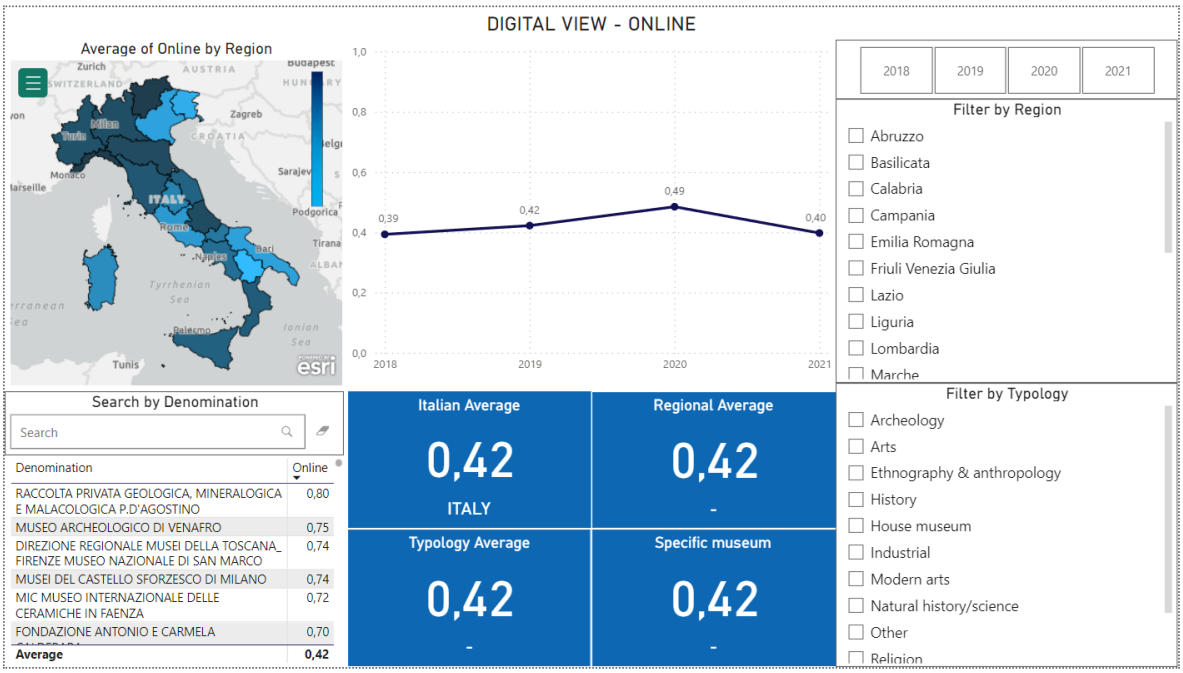


Figure 45 - Digital View - Online

| **Questions answered by the View** | **Value added for museum managers** | **Value added for the Ministry of Culture** |
|---|---|---|
| What is the average Online KPI in Italy? | Possibility of comparing their museum's Online KPI against the Italian average | Gain insights into sector-wide performance |
| What is the average Online KPI in a specific region? | Possibility of comparing the museum's Online KPI against the regional or provincial average | Opportunity to conduct personalized or tailored analysis for specific geographical areas |
| How does the Online KPI in a specific region compare to the Italian average? | Possibility of comparing the museum's Online KPI against the regional or provincial average | Opportunity to conduct personalized or tailored analysis for specific geographical areas |
| What is the average Online KPI of museums of a specific typology? | Possibility of comparing the museum's Online KPI against the *Typology* average | Opportunity to conduct personalized or tailored analysis for specific typologies of museums |
| How does the Online KPI of museums of a specific typology compare to the Italian average? | / | Opportunity to conduct personalized or tailored analysis for specific typologies of museums |
| How does the Online KPI of a specific museum compare to the average KPI of its region? | Possibility of comparing the museum's Online KPI against the regional average | / |
| How does the Online KPI of a specific museum compare to the average KPI of museums of the same typology? | Possibility of comparing the museum's Online KPI against the *Typology* average | / |
| Which are the museums that have the highest Online KPI in Italy? | Identifying the top performers in terms of Online KPI and how their museum compares against them; Possibility of implementing new practices by emulation | Identifying the top performers in terms of Number of Visitors; Possibility of promoting the adoption of best practices to non-top performing museums |
| How did the average Online KPI change over time? | Possibility of comparing the trend of their museum against the Italian trend | Understanding the history of the performance in the Online KPI |
| How did the Online KPI for a specific museum change over time? | Understanding the history of the performance in the Online KPI | / |
| How was the Online KPI situation in a specific year? | Understanding the performance of the Online KPI in a specific year | Understanding the history of the performance in the Online KPI, focusing on a specific year |

Table 68 - Questions answered and value added of the Digital View - Online

The integration introduced the *Typology* dimension, which holds significance in benchmarking since museums belonging to distinct typologies may have different features that lead to different performance. For a museum manager, the possibility of benchmarking its museum against museums that belong to the same *Typology* is important as it greatly refines the similarity of museums, thus allowing for better external benchmarking. Moreover, the integration of data made it possible to visualize the Online KPI. In fact, the KPI is composed of five variables, three of which are obtained thanks to the integration of open data (*Presence of a dedicated website, Social media presence,* and *Online tours*). Lastly, the *Time* dimension adds value to decision-makers by introducing the dimension that enables internal benchmarking. The *Time* dimension is very important for decision-making as it makes the museums' managers aware of the internal performance over the years. It is also helpful for the Ministry of Culture to evaluate overall performance over time and to assess the effectiveness of implemented initiatives. Figure 46 shows visually the information that would be lost without data integration of open data and Table 69 summarizes the variables added to the View solely thanks to the integration of open data.

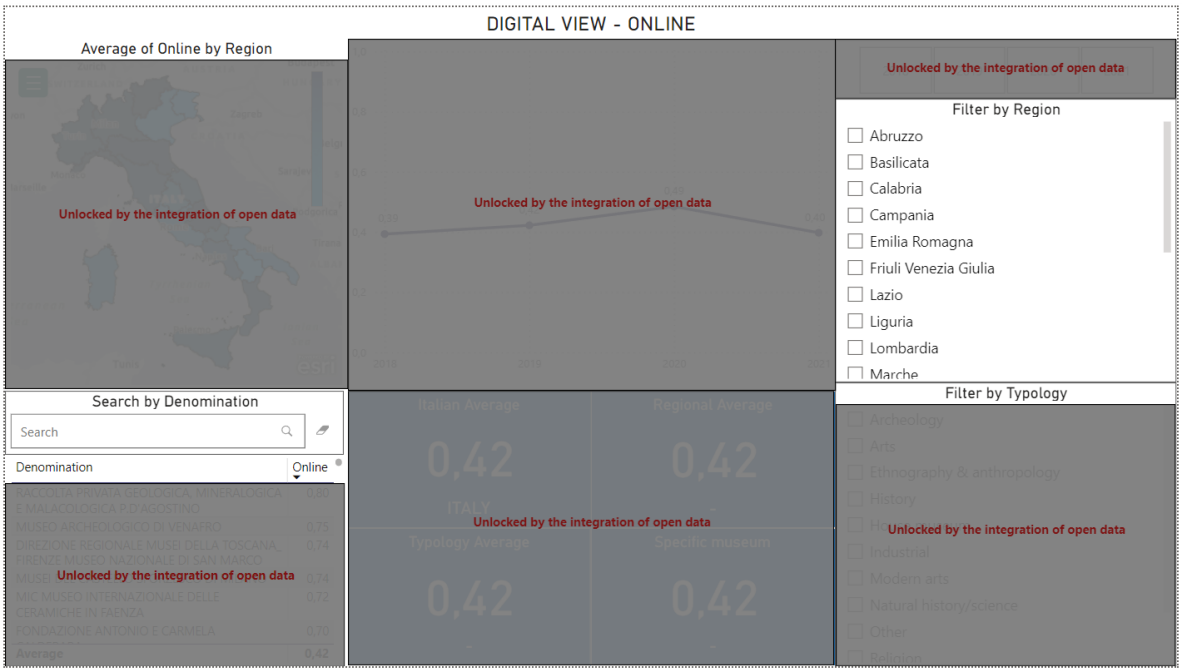


Figure 46 - Digital View - Online without open data

| View | Variable added through data integration |
|---|---|
| Digital View – Online | Typology |
| | Presence of a dedicated website |
| | Social media presence |
| | Online tours |
| | Time |

Table 69 - Variables added through data integration in the Digital View - Online

### Digital View - On-site

This View has the objective of giving an overview of the On-site KPI in Italy. The indicator is computed as a combination of five on-site digital technologies (audio guide, augmented reality, virtual reality, QR code, chatbot). The View shows the evolution of the On-site KPI over time. The View shows, through various visual objects, the distribution of the On-site KPI in the Italian territory and over time. It is structured to facilitate benchmarking, allowing users the flexibility to focus on a specific Region, *Typology* and/or year. In this View, users are empowered with the flexibility to choose between focusing on a specific year or exploring the evolution of the indicator over time. This feature allows users to tailor their analysis based on their preferences and requirements. Whether considering a particular year's data for in-depth insights or observing the trend and changes in the indicator across multiple years, this functionality provides users with a comprehensive understanding of the temporal aspect of the indicator. As per all the other Views, the user can also filter by *Typology*.

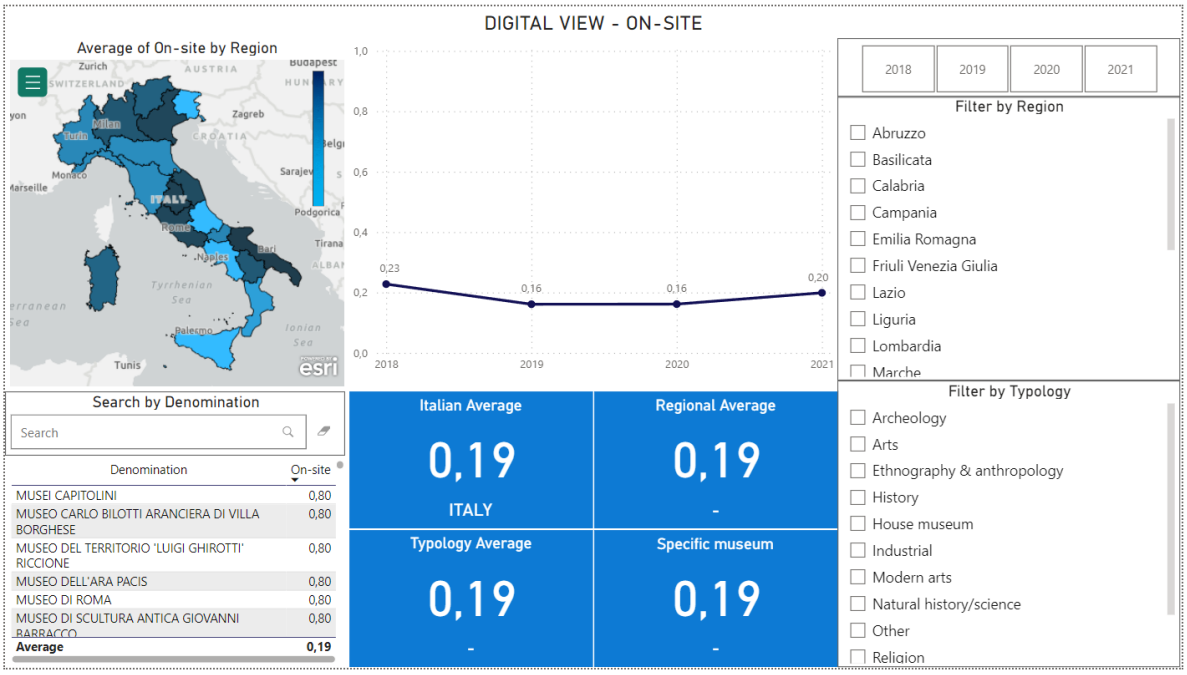


Figure 47 - Digital View - On-site

| **Questions answered by the View** | **Value added for museum managers** | **Value added for the Ministry of Culture** |
|---|---|---|
| What is the average On-site KPI in Italy? | Possibility of comparing their museum's On-site KPI against the Italian average | Gain insights into sector-wide performance |
| What is the average On-site KPI in a specific region? | Possibility of comparing the museum's On-site KPI against the regional or provincial average | Opportunity to conduct personalized or tailored analysis for specific geographical areas |
| How does the On-site KPI in a specific region compare to the Italian average? | Possibility of comparing the museum's On-site KPI against the regional or provincial average | Opportunity to conduct personalized or tailored analysis for specific geographical areas |
| What is the average On-site KPI of museums of a specific typology? | Possibility of comparing the museum's On-site KPI against the *Typology* average | Opportunity to conduct personalized or tailored analysis for specific typologies of museums |
| How does the On-site KPI of museums of a specific typology compare to the Italian average? | / | Opportunity to conduct personalized or tailored analysis for specific typologies of museums |
| How does the On-site KPI of a specific museum compare to the average KPI of its region? | Possibility of comparing the museum's On-site KPI against the regional average | / |
| How does the On-site KPI of a specific museum compare to the average KPI of museums of the same typology? | Possibility of comparing the museum's On-site KPI against the *Typology* average | / |
| Which are the museums that have the highest On-site KPI in Italy? | Identifying the top performers in terms of On-site KPI and how their museum compares against them; Possibility of implementing new practices by emulation | Identifying the top performers in terms of Number of Visitors; Possibility of promoting the adoption of best practices to non-top performing museums |
| How did the average On-site KPI change over time? | Possibility of comparing the trend of their museum against the Italian trend | Understanding the history of the performance in the On-site KPI |
| How did the On-site KPI for a specific museum change over time? | Understanding the history of the performance in the On-site KPI | / |
| How was the On-site KPI situation in a specific year? | Understanding the performance of the On-site KPI in a specific year | Understanding the history of the performance in the On-site KPI, focusing on a specific year |

Table 70 - Questions answered and value added of the Digital View - On-site

The integration introduced the *Typology* dimension, which holds significance in benchmarking since museums belonging to distinct typologies may have different features that lead to different performance. For a museum manager, the possibility of benchmarking its museum against museums that belong to the same *Typology* is important as it greatly refines the similarity of museums, thus allowing for better external benchmarking. Moreover, the *Time* dimension adds value to decision-makers by introducing the dimension that enables internal benchmarking. The *Time* dimension is very important for decision-making as it makes the museums' managers aware of the internal performance over the years. It is also helpful for the Ministry of Culture to evaluate overall performance over time and to assess the effectiveness of implemented initiatives. Figure 48 shows visually the information that would be lost without data integration of open data and Table 71 summarizes the variables added to the View solely thanks to the integration of open data.

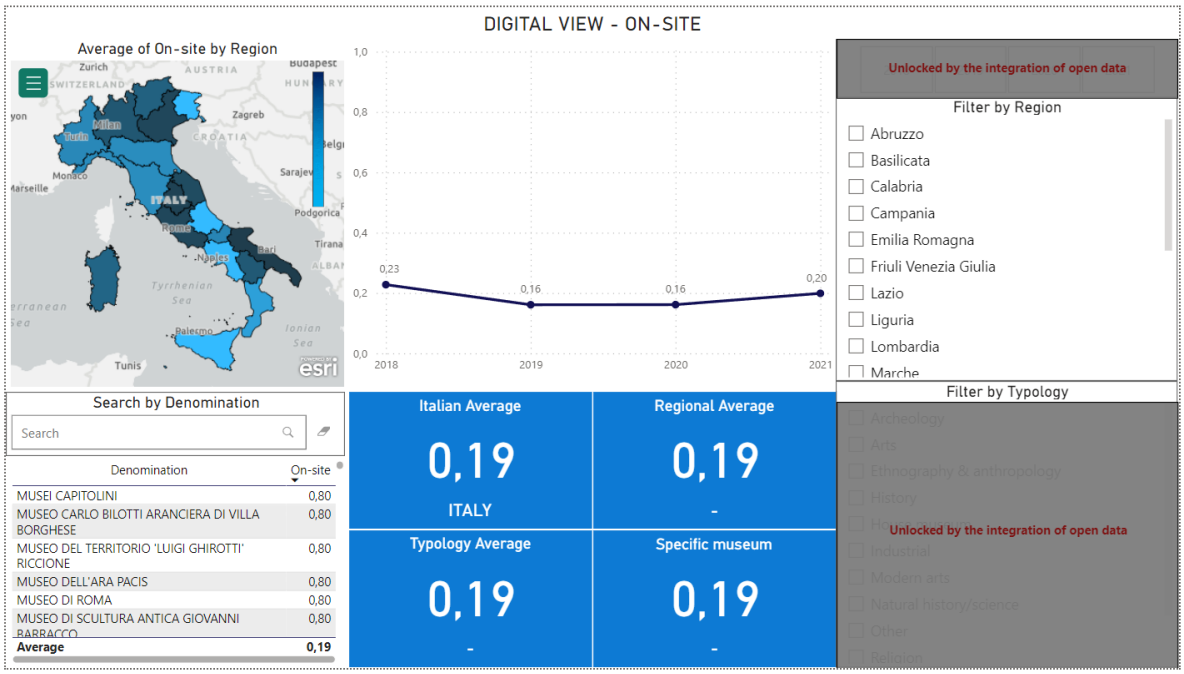


Figure 48 - Digital View - On-site without open data

| View | Variable added through data integration |
|---|---|
| Digital View – On-site | Typology |
| | Time |

Table 71 - Variables added through data integration in the Digital View - On-site

### Digital View - Organizational Readiness

The View has the objective of giving an overview of the Organizational Readiness KPI in Italy. The indicator is computed as a combination of two variables: Presence of a Digital Innovation Plan and Workers with Digital Competencies. The View shows the evolution of the Organizational Readiness KPI over time. The View shows, through various visual objects, the distribution of the Organizational Readiness KPI in the Italian territory and over time. It is structured to facilitate benchmarking, allowing users the flexibility to focus on a specific Region, *Typology* and/or year. In this View, users are empowered with the flexibility to choose between focusing on a specific year or exploring the evolution of the indicator over time. This feature allows users to tailor their analysis based on their preferences and requirements. Whether considering a particular year's data for in-depth insights or observing the trend and changes in the indicator across multiple years, this functionality provides users with a comprehensive understanding of the temporal aspect of the indicator. As per all the other Views, the user can also filter by *Typology*.

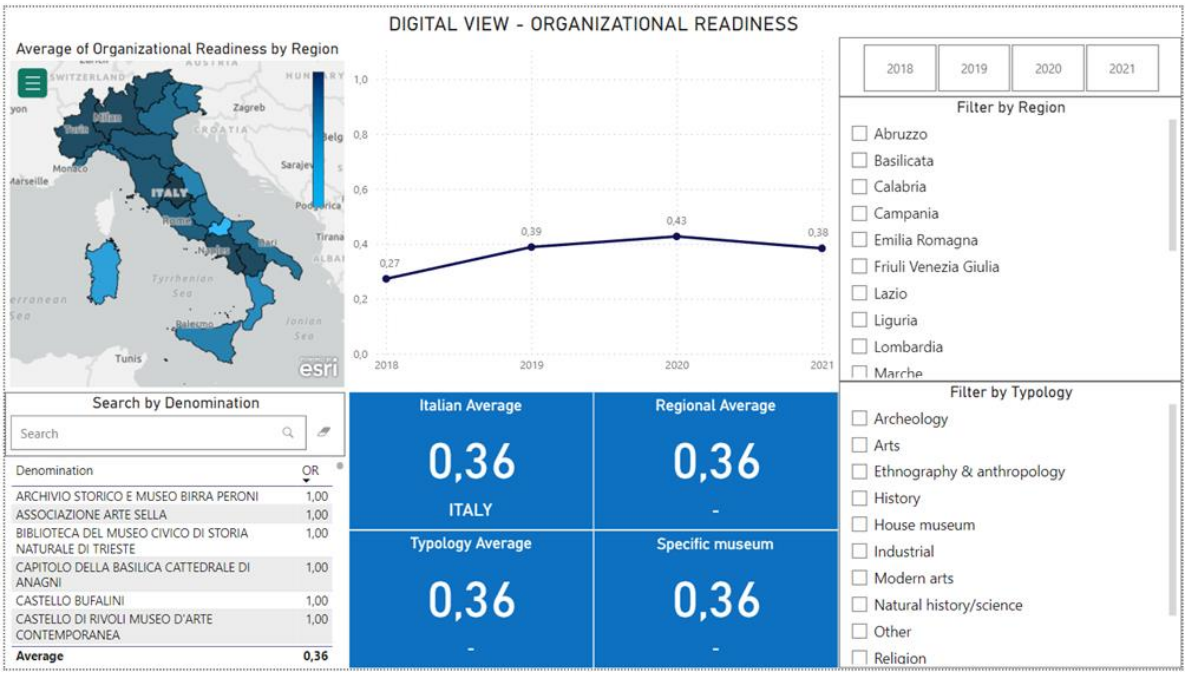


Figure 49 - Digital View - Organizational Readiness

| **Questions answered by the View** | **Value added for museum managers** | **Value added for the Ministry of Culture** |
|---|---|---|
| What is the average Organizational Readiness KPI in Italy? | Possibility of comparing their museum's Organizational Readiness KPI against the Italian average | Gain insights into sector-wide performance |
| What is the average Organizational Readiness KPI in a specific region? | Possibility of comparing the museum's Organizational Readiness KPI against the | Opportunity to conduct personalized or tailored analysis for specific geographical areas |

| Questions answered by the View | Value added for museum managers | Value added for the Ministry of Culture |
|---|---|---|
| | regional or provincial average | |
| How does the Organizational Readiness KPI in a specific region compare to the Italian average? | Possibility of comparing the museum's Organizational Readiness KPI against the regional or provincial average | Opportunity to conduct personalized or tailored analysis for specific geographical areas |
| What is the average Organizational Readiness KPI of museums of a specific typology? | Possibility of comparing the museum's Organizational Readiness KPI against the *Typology* average | Opportunity to conduct personalized or tailored analysis for specific typologies of museums |
| How does the Organizational Readiness KPI of museums of a specific typology compare to the Italian average? | / | Opportunity to conduct personalized or tailored analysis for specific typologies of museums |
| How does the Organizational Readiness KPI of a specific museum compare to the average KPI of its region? | Possibility of comparing the museum's Organizational Readiness KPI against the regional average | / |
| How does the Organizational Readiness KPI of a specific museum compare to the average KPI of museums of the same typology? | Possibility of comparing the museum's Organizational Readiness KPI against the *Typology* average | / |
| Which are the museums that have the highest Organizational Readiness KPI in Italy? | Identifying the top performers in terms of Organizational Readiness KPI and how their museum compares against them; Possibility of implementing new practices by emulation | Identifying the top performers in terms of Number of Visitors; Possibility of promoting the adoption of best practices to non-top performing museums |
| How did the average Organizational Readiness KPI change over time? | Possibility of comparing the trend of their museum against the Italian trend | Understanding the history of the performance in the Organizational Readiness KPI |
| How did the Organizational Readiness KPI for a specific museum change over time? | Understanding the history of the performance in the Organizational Readiness KPI | / |
| How was the Organizational Readiness KPI situation in a specific year? | Understanding the performance of the Organizational Readiness KPI in a specific year | Understanding the history of the performance in the Organizational Readiness KPI, focusing on a specific year |

Table 72 - Questions answered and value added of the Digital View - Organizational Readiness

The integration introduced the *Typology* dimension, which holds significance in benchmarking since museums belonging to distinct typologies may have different

features that lead to different performance. For a museum manager, the possibility of benchmarking its museum against museums that belong to the same *Typology* is important as it greatly refines the similarity of museums, thus allowing for better external benchmarking. Moreover, the *Time* dimension adds value to decision-makers by introducing the dimension that enables internal benchmarking. The *Time* dimension is very important for decision-making as it makes the museums' managers aware of the internal performance over the years. It is also helpful for the Ministry of Culture to evaluate overall performance over time and to assess the effectiveness of implemented initiatives. Figure 50 shows visually the information that would be lost without data integration of open data and Table 73 summarizes the variables added to the View solely thanks to the integration of open data.

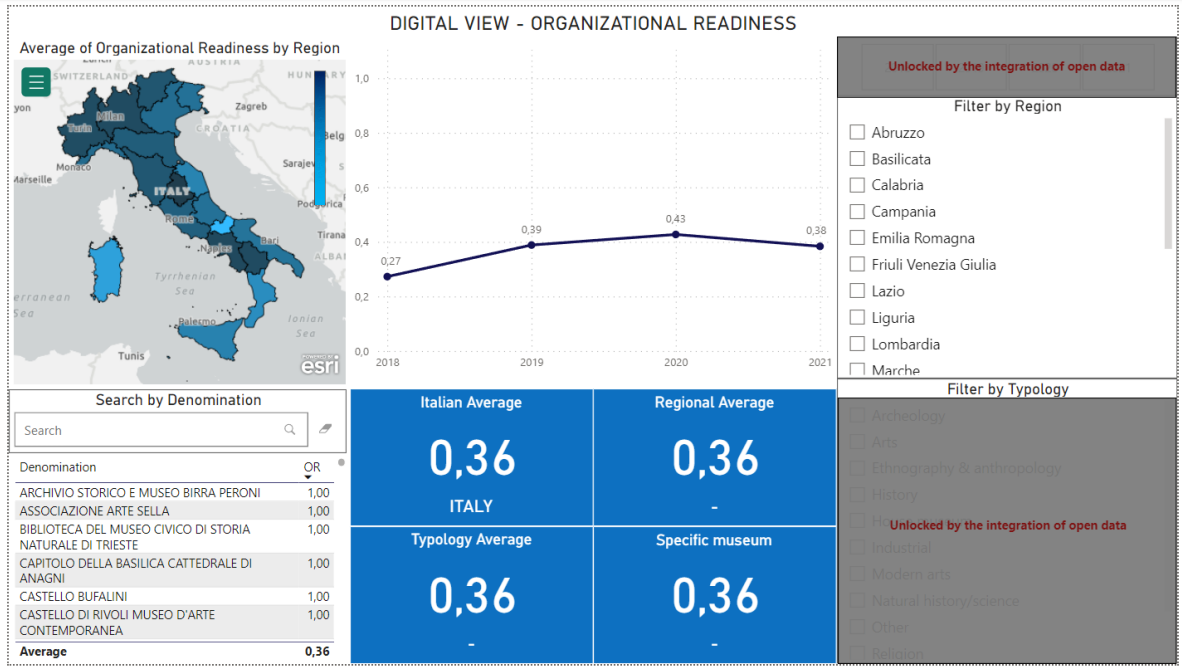


Figure 50 - Digital View - Organizational Readiness without open data

| View | Variable added through data integration |
|---|---|
| Digital View – Organizational Readiness | Typology |
| | Time |

Table 73 - Variables added through data integration in the Digital View - Organizational Readiness

### Digital View - Evolution Average

This View has the objective of showing the evolution of the AD KPI in Italy. The indicator is computed as the average of the three Digital KPIs: Online, On-site, and Organizational Readiness. The View provides insights into the composition of the AD KPI by displaying the individual impact of each of the three Digital KPIs that constitute it. This feature allows users to understand and analyze how the Online, On-site, and Organizational Readiness components contribute to the overall progression of the AD KPI, offering a comprehensive view of their respective influences on the indicator's evolution over time. The View shows, through various visual objects, the distribution of the AD KPI in the Italian territory. It is structured to facilitate benchmarking, allowing users the flexibility to focus on a specific Region, *Typology* and/or year. In this View, users are empowered with the flexibility to choose between focusing on a specific year or exploring the evolution of the indicator over time. This feature allows users to tailor their analysis based on their preferences and requirements. Whether considering a particular year's data for in-depth insights or observing the trend and changes in the indicator across multiple years, this functionality provides users with a comprehensive understanding of the temporal aspect of the indicator. As per all the other Views, the user can also filter by *Typology*.

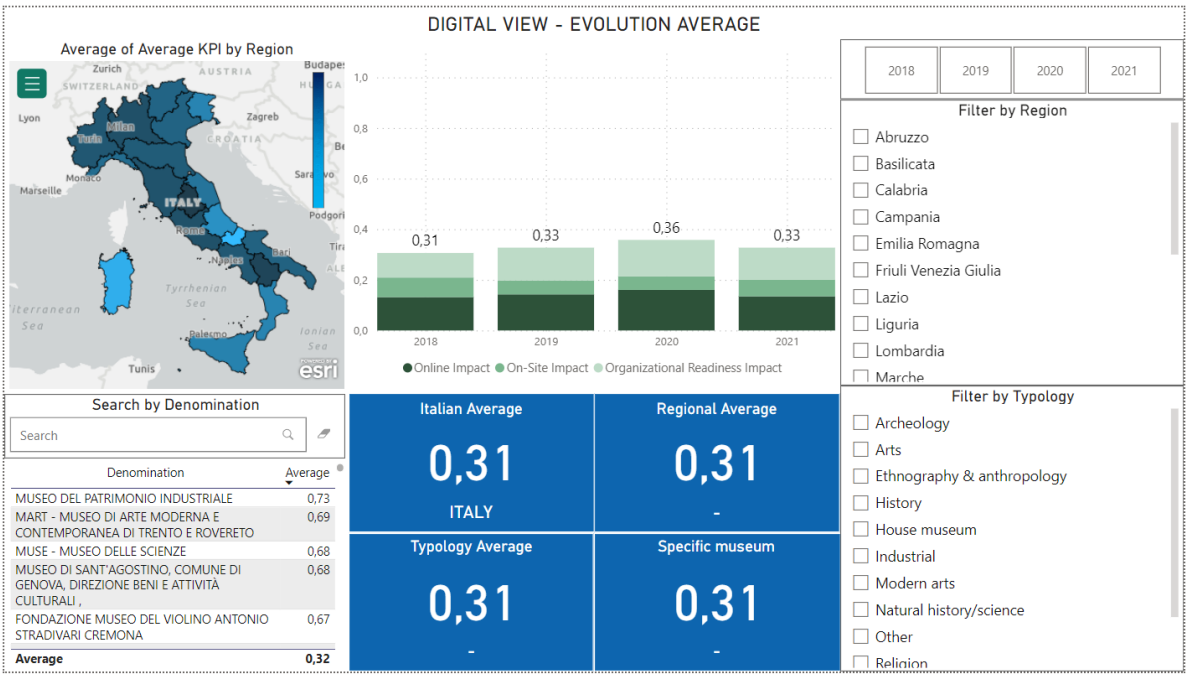


Figure 51 - Digital View - Evolution Average

| **Questions answered by the View** | **Value added for museum managers** | **Value added for the Ministry of Culture** |
|---|---|---|
| What is the average AD KPI in Italy? | Possibility of comparing their museum's AD KPI against the Italian average | Gain insights into sector-wide performance |
| What is the average AD KPI in a specific region? | Possibility of comparing the museum's AD KPI against the regional or provincial average | Opportunity to conduct personalized or tailored analysis for specific geographical areas |
| What is the impact of the three Digital KPIs for a specific museum? | Possibility of comparing the museum's AD KPI against the regional or provincial average | Opportunity to conduct personalized or tailored analysis for specific geographical areas |
| How does the AD KPI in a specific region compare to the Italian average? | Possibility of comparing the museum's AD KPI against the *Typology* average | Opportunity to conduct personalized or tailored analysis for specific typologies of museums |
| What is the average AD KPI of museums of a specific typology? | / | Opportunity to conduct personalized or tailored analysis for specific typologies of museums |
| How does the AD KPI of museums of a specific typology compare to the Italian average? | Possibility of comparing the museum's AD KPI against the regional average | / |
| How does the AD KPI of a specific museum compare to the average KPI of its region? | Possibility of comparing the museum's AD KPI against the *Typology* average | / |
| How does the AD KPI of a specific museum compare to the average KPI of museums of the same typology? | Identifying the top performers in terms of AD KPI and how their museum compares against them; Possibility of implementing new practices by emulation | Identifying the top performers in terms of Number of Visitors; Possibility of promoting the adoption of best practices to non-top performing museums |
| Which are the museums that have the highest AD KPI in Italy? | Possibility of comparing the trend of their museum against the Italian trend | Understanding the history of the performance in the AD KPI |
| How did the average AD KPI change over time? | Understanding the history of the performance in the AD KPI | / |
| How did the AD KPI for a specific museum change over time? | Understanding the performance of the AD KPI in a specific year | Understanding the history of the performance in the AD KPI, focusing on a specific year |

Table 74 - Questions answered and value added of the Digital View - Evolution Average

The integration introduced the *Typology* dimension, which holds significance in benchmarking since museums belonging to distinct typologies may have different features that lead to different performance. For a museum manager, the possibility of

benchmarking its museum against museums that belong to the same *Typology* is important as it greatly refines the similarity of museums, thus allowing for better external benchmarking. Moreover, the integration of data made it possible to compute the Online KPI. In fact, the KPI is composed of five variables, three of which are obtained thanks to the integration of open data (*Presence of a dedicated website*, *Social media presence*, and *Online tours*). The Online KPI is needed for the computation of the AD KPI, as it is one of the three Digital KPIs that formulate it. Lastly, the *Time* dimension adds value to decision-makers by introducing the dimension that enables internal benchmarking. The *Time* dimension is very important for decision-making as it makes the museums' managers aware of the internal performance over the years. It is also helpful for the Ministry of Culture to evaluate overall performance over time and to assess the effectiveness of implemented initiatives. Figure 52 shows visually the information that would be lost without data integration of open data and Table 75 summarizes the variables added to the View solely thanks to the integration of open data.

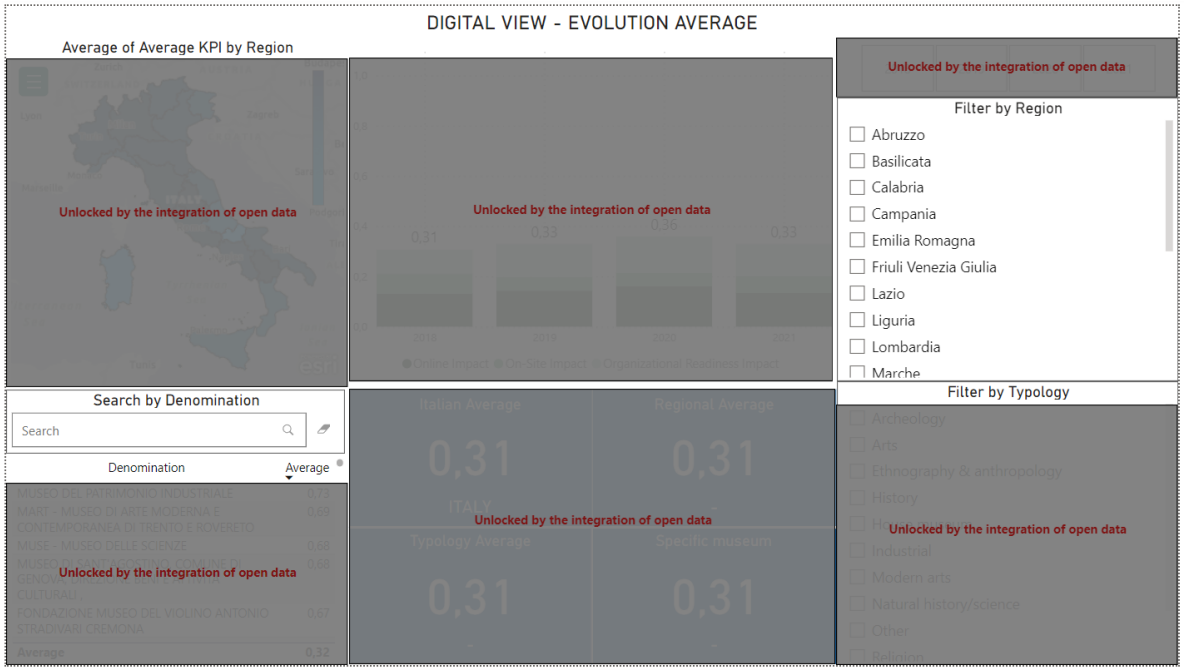


Figure 52 - Digital View - Evolution Average without open data

| View | Variable added through data integration |
|---|---|
| Digital View – Evolution Average | Typology |
| | Presence of a dedicated website |
| | Social media presence |
| | Online tours |
| | Time |

Table 75 - Variables added through data integration in the Digital View - Evolution Average

## Digital View - Positioning

This View as the objective of showing the positioning of the museums based on the three Digital KPIs in Italy. The View provides insights into the positioning of museums, using a 3-dimensional chart that shows the performance of each museum in the three dimensions (the Digital KPIs). Additionally, it allows users to assess the performance of a particular museum by showcasing the percentage of museums that perform worse on the three Digital KPIs. It is structured to facilitate benchmarking, allowing users the flexibility to focus on a specific Region, *Typology* and/or year. In this View, users are empowered with the flexibility to choose between focusing on a specific year or exploring the evolution of the indicator over time. This feature allows users to tailor their analysis based on their preferences and requirements. Whether considering a particular year's data for in-depth insights or observing the trend and changes in the indicator across multiple years, this functionality provides users with a comprehensive understanding of the temporal aspect of the indicator. As per all the other Views, the user can also filter by *Typology*.

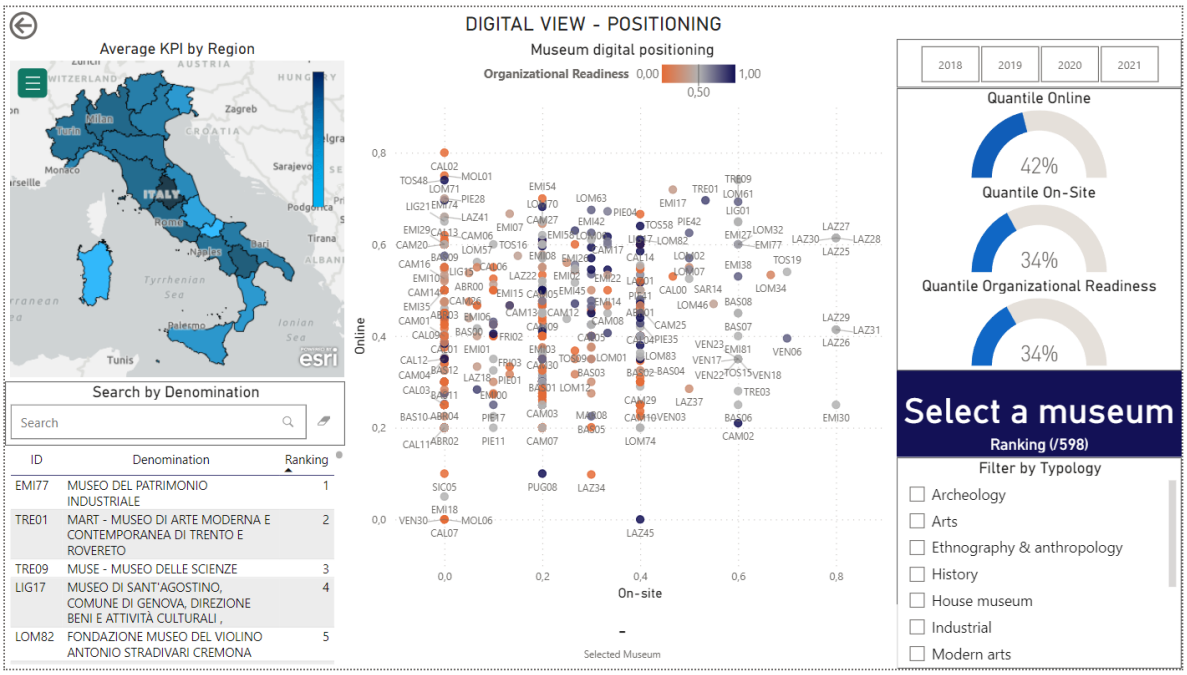


Figure 53 - Digital View - Positioning

| Questions answered by the View | Value added for museum managers | Value added for the Ministry of Culture |
|---|---|---|
| What is the average AD KPI in a specific region? | Possibility of comparing the museum's AD KPI against the regional average | Opportunity to conduct personalized or tailored analysis for specific geographical areas |
| What are the values of the three Digital KPIs for a specific museum? | Understanding the performance of their museum in the Digital KPIs | / |

| **Questions answered by the View** | **Value added for museum managers** | **Value added for the Ministry of Culture** |
|---|---|---|
| How does a specific museum rank with respect to the others? | Understanding how the performance of their museum compares to the Italian museums | Opportunity to conduct personalized or tailored analysis for specific museums that are top performers or laggards |
| How does a specific museum rank with respect to other museums in the same region? | Understanding how the performance of their museum compares to the other museums, similar in terms of geographical area | Opportunity to conduct personalized or tailored analysis for specific museums that are top performers or laggards |
| How does a specific museum rank with respect to other museums of the same *Typology*? | Understanding how the performance of their museum compares to the other museums, equal in terms of *Typology* | Opportunity to conduct personalized or tailored analysis for specific museums that are top performers or laggards |
| Where is a museum positioned based on the three Digital KPIs? | Understanding visually how the performance of their museum compares to the Italian museums | Opportunity to conduct personalized or tailored analysis for specific museums that are top performers or laggards |
| How good is a museum in the Online KPI, with respect to the others? | Understanding the performance of their museum in the Online KPI, benchmarking against other museums | Opportunity to conduct personalized or tailored analysis for specific museums that are top performers or laggards |
| How good is a museum in the On-site KPI, with respect to the others? | Understanding the performance of their museum in the On-site KPI, benchmarking against other museums | Opportunity to conduct personalized or tailored analysis for specific museums that are top performers or laggards |
| How good is a museum in the Organizational Readiness KPI, with respect to the others? | Understanding the performance of their museum in the Organizational Readiness KPI, benchmarking against other museums | Opportunity to conduct personalized or tailored analysis for specific museums that are top performers or laggards |
| How was the situation in a specific year? | Understanding the performance of the Digital KPIs in a specific year | Understanding the history of the performance in the AD KPI, focusing on a specific year |

Table 76 - Questions answered and value added of the Digital View - Positioning

The integration introduced the *Typology* dimension, which holds significance in benchmarking since museums belonging to distinct typologies may have different features that lead to different performance. For a museum manager, the possibility of

benchmarking its museum against museums that belong to the same *Typology* is important as it greatly refines the similarity of museums, thus allowing for better external benchmarking. Moreover, the integration of data made it possible to compute the Online KPI. In fact, the KPI is composed of five variables, three of which are obtained thanks to the integration of open data (*Presence of a dedicated website*, *Social media presence*, and *Online tours*). The Online KPI is needed for the computation of the AD KPI, as it is one of the three Digital KPIs that formulate it. At last, the *Time* dimension adds value to decision-makers by introducing the dimension that enables internal benchmarking. The *Time* dimension is very important for decision-making as it makes the museums' managers aware of the internal performance over the years. It is also helpful for the Ministry of Culture to evaluate overall performance over time and to assess the effectiveness of implemented initiatives. Figure 54 shows visually the information that would be lost without data integration of open data and Table 77 summarizes the variables added to the View solely thanks to the integration of open data.

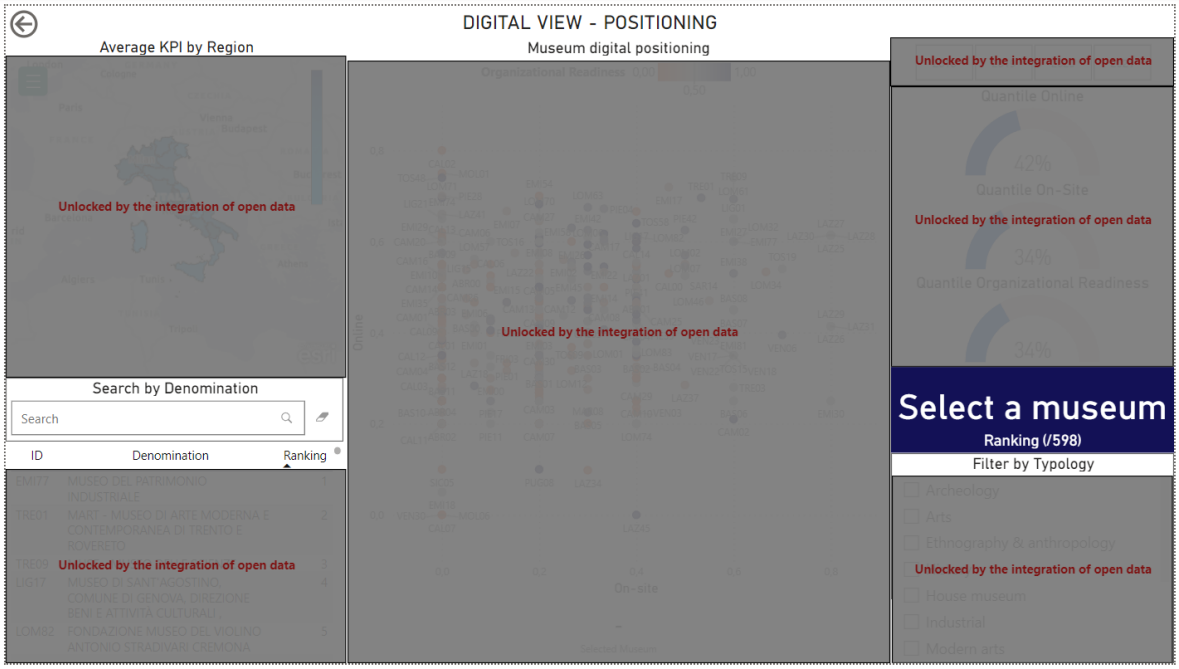


Figure 54 - Digital View - Positioning without open data

| View | Variable added through data integration |
|---|---|
| Digital View – Positioning | Typology |
| | Presence of a dedicated website |
| | Social media presence |
| | Online tours |
| | Time |

Table 77 - Variables added through data integration in the Digital View - Positioning

## 7.2. The value added of the dashboard

This section elaborates on the IDM Framework for Museums (see also Chapter 4) with a specific focus on the additional value offered by the dashboard to museum managers and the Ministry of Culture.

The development of the dashboard was guided by an integrated thinking approach. The outcome is a dashboard that is interconnected with all the aspects of the IDM Framework for Museums. Specifically, the Data dimension is embodied in the integration of open data which is the driver of the value added in the dashboard. The development of the Digital KPIs - i.e., Online, On-site, and Organizational Readiness - used to evaluate the Digital state of museums, is aligned with the Measuring dimension, while the Reporting dimension pertains directly to the creation of the integrated dashboard, meaning a reporting tool that consolidates measures from different sources in a unified view. Ultimately, the entire process is steered by the Human dimension, employing an integrated thinking logic.

The dashboard empowers museum managers in Italy to compare their museums with similar institutions through an approach known as "external benchmarking" (Saul, 2004, p. 6). It allows museum managers to assess the performance of similar museums using KPIs that are equal for all the museums. This assists the museum managers in identifying areas of strength and weaknesses for improvement. This aspect relates to the Measuring dimension of the IDM Framework for Museums, as external benchmarking of museums is enabled thanks to KPIs and data integration. External benchmarking can increase an organization's performance, fostering proactive decisions (Saul, 2004). In fact, the integration of data serves as a method to incorporate external context into decision-making processes, ensuring that the organization and its decision-makers are considering the environment within their decisions, thus being context-aware (Berlanga & Nebot, 2016). This aspect is aligned with the Data Management dimension of the IDM Framework for Museums as better decision-making is achieved thanks to data integration. The external benchmarking aspect of the dashboard supports a better informed and context-aware decision-making by museum managers.

The benchmarking dimensions are showcased in the Benchmarking Page of the dashboard (Figure 24). The benchmarking dimensions relate to geographical characteristics (Region, *Province, Municipality*), size (*Personnel*, Revenues, Visitors), and *Typology*. Among these dimensions, those highlighted in *italics* have been incorporated into the dashboard thanks to the integration of the Istat Microdata dataset. This means that the integration of open data has enabled museum managers and the Ministry of Culture to assess and benchmark institutions across dimensions that were inaccessible before the integration (Data Management dimension). Specifically, the integration has improved the precision of benchmarking on the geographical dimension by allowing a breakdown into *Region - Province - Municipality*, consequently refining the analysis

granularity, and allowing users an increased flexibility in their analysis (improving benchmarking, pertaining to the Measuring dimension). Although the *Personnel* dimension was available in the Proprietary dataset, the integration brings an improvement by accessing *Personnel* information in numerical form rather than in a range form in the Proprietary dataset. This enhancement ensures that benchmarking can be done on personalized ranges that can better reflect the different needs of the users. Finally, the *Typology* dimension holds significance in benchmarking since museums belonging to distinct typologies may exhibit varying performances primarily due to their classification within a specific typology (Camarero et al., 2011). The selection of benchmarking dimensions enables museum managers to benchmark against similar museums on various dimensions or combinations of dimensions (e.g., History museums in Lombardia).

The improvements to decision-making, benefiting both the museum managers and the Ministry of Culture, are directly fostered by the dashboard. In turn, the dashboard is enhanced by the integration of open data, which enables the inclusion of new and more precise benchmarking dimensions and the development of the Digital KPIs.

In the dashboard, the performance of museums is exclusively focused on the digital domain. The three KPIs outlined in section 6.1 - Online, On-site, and Organizational Readiness - are derived from a combination of variables that originate from both the Proprietary and Istat Microdata datasets. In particular, the computation of the Online KPI is possible only thanks to the integration of the Istat Microdata dataset, since three of the five variables that compose it originate from that dataset. Consequently, the integration of data allows the computation of an indicator that would have been impossible to compute by considering solely the Proprietary dataset. The development of KPIs that are computed thanks to the integration of different sources pertains to the Measuring dimension of the IDM Framework for Museums, as it enables the development of a holistic PM system for the control of quantitative and qualitative KPIs.

The value added of data integration does not only reside in incorporating new variables into a unified view; it also implements an additional and crucial dimension: the time dimension. The time dimension is very important as it allows for a visualization of a dynamic view of performance. It enables museums to benchmark their performance against their own historical data, a practice referred to as “internal benchmarking" (Saul, 2004, p.6). This new feature empowers museum managers to visualize the performance history of their institutions, improving decision-making. In this case, the concepts shift from context-awareness to history-awareness. In fact, museum managers need to be aware of the environment in which they operate but also aware of the performance history of their institutions in order to have a clear overview and make sound decisions based on performance information. Indeed, decision-making is mainly influenced by internal factors (Papadakis et al., 1998;

Berlanga & Nebot, 2016). The dashboard empirically shows that context-awareness and a unified view of internal data can be obtained through data integration. Moreover, the time dimension proves valuable for the Ministry of Culture as it serves as a tool to evaluate the impact of initiatives, showcasing changes in performance within specific areas over time. The time dimension may provide valuable insights into the effectiveness and outcomes of various interventions or programs and suggest trends and ideas for future initiatives.

Table 78 summarizes the value added to the user of the dashboard, illustrating both the museum managers and the Ministry of Culture perspectives:

| **Value added to the decision-maker** | **Implementation in the dashboard** | **Museum perspective** | **Ministry of Culture perspective** | **Relevance to the IDM Framework for Museums** |
|---|---|---|---|---|
| Capability to benchmark against similar museums (External benchmarking) | Visual objects that enable filtering of the dashboard | Context-awareness leading to improved decision-making | Identification of areas that need improvement | Human, Data Management, Measuring, Reporting |
| Capability to track performance over time and benchmark against its history of performance (Internal benchmarking) | Views that show performance over time | History-awareness leading to improved decision-making | Assessment of trends and initiatives over time | Human, Data Management, Measuring, Reporting |
| Capability to assess the Digital performance of museums through performance measures | Online, On-site, and Organizational Readiness KPIs | Standardized assessment of performance | Standardized assessment of performance | Human, Data Management, Measuring |
| Improved benchmarking capabilities and precision | Integration of new benchmarking dimensions: Province, Municipality, Personnel, Typology | Improved identification of comparable museums, based on different dimensions | Improved capabilities of analyzing museums falling under specific parameters | Human, Data Management, Measuring, Reporting |

Table 78 - Value added to the user

Table 83 shows each variable incorporated in the dashboard thanks to the integration of the Istat Microdata dataset and its value added to the user of the dashboard.

| **Variable** | **Value added to the decision-maker** |
|---|---|
| Province | It adds value by introducing a new dimension for benchmarking geographically with higher precision, from a regional perspective to a provincial one. It lets the user identify visually, on a map, the location of the museum and enables external benchmarking on the geographical dimension. |
| Municipality | It adds value by introducing a new dimension for benchmarking geographically with higher precision. It lets the user identify visually, on a map, the location of the museum and enables external benchmarking on the geographical dimension. |
| Typology | It adds value by introducing a new dimension that enables external benchmarking against similar museums. *Typology* is one of the most important dimensions to benchmark on because the typology of a museum deeply influences its characteristics and features. |
| Personnel | It adds value by introducing a new dimension that enables improved external benchmarking against similar museums. It enriches the dashboard by refining benchmarking accuracy, shifting from predefined ranges to custom ones. |
| Presence of a dedicated website | It is a component used to calculate the Online KPI. It adds value by introducing the aspect of the online presence of the museum, meaning that the museum owns its own website. The connection with the time dimension makes it possible to see if a museum created its website during the four years analyzed. |
| Social media presence | It is a component used to calculate the Online KPI. It adds value by introducing the aspect of the social media presence of the museum, meaning that the museum owns at least one social media page. The connection with the time dimension makes it possible to see if a museum has landed on social media for the first time during the four years analyzed. |
| Online tours | It is a component used to calculate the Online KPI. It adds value by introducing the aspect of the possibility of making online tours in the museum. The connection with the time dimension makes it possible to see if a museum has started providing the service for the first time during the four years analyzed. |
| Time | It adds value by introducing a new dimension that enables internal benchmarking. The *Time* dimension is very important for decision-making as it makes the museums' managers aware of the internal performance over the years. It is also helpful for the Ministry of Culture to evaluate overall performance over time and to assess the effectiveness of implemented initiatives. |

Table 79 - Value added of the variables

# Chapter 8: Conclusions

This chapter concludes the thesis presenting the main conclusions and limitations of this study and recommending paths for further research based on the presented results.

## 8.1. Main conclusions

In the current scenario, museums are pressured by:

- The demands of stakeholders for more transparency and better reporting, which may lead them to focus only on profit-oriented actions, overlooking their social mission, and may leave stakeholders with indicators they do not fully understand.
- The constant limitations in funding, which are the underlying cause of many management problems. In the thesis, emphasis is placed on some of them, such as the decrease in the quality of the offer, the limited prospect of new paid hirings, the inability to improve supporting activities and to implement new ones.
- The usually limited digital competencies of personnel, which is due to both a lack of funding and the conservative view of most curators who do not believe that digital competencies are needed in museums.
- The lack in digital innovation processes, which is a direct consequence of the lack of skilled personnel with digital competencies.

The issue of human resources is crucial for decision-making in museums. Without the acquisition of knowledge by current employees or the employment of new personnel with digital competencies, museums will always lack digital innovation and consequently data capabilities. The employment of digitally competent people is hindered by both the conservativism of curators and the lack of funding in the cultural heritage sector. The funds are directed to the core activities of the museum, while the supporting ones are overlooked.

Through the literature review I conducted within the thesis, I found that two main topics were either just briefly or completely not featured in the literature and yet needed further exploration (Table 80).

| **Literature gaps addressed in the thesis** | **Inspiration for research** |
|---|---|
| LG1: Open data and its integration for improving decision-making in museums | Berlanga & Nebot, 2016 |
| LG2: Development of Key Performance Indicators specifically tailored to museums | De Bernardi et al., 2019 |

Table 80 - Literature gaps

To address these two main gaps, I developed a framework to conceptually describe the added value generated by data integration on the main dimensions of the decision-making. The framework is called IDM Framework for Museums.

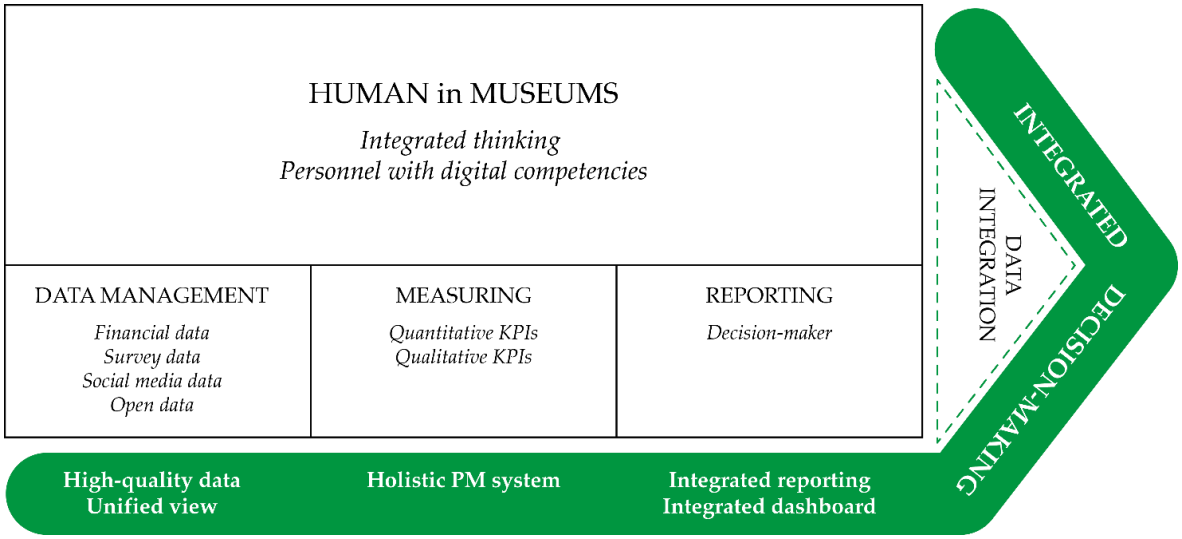


Figure 55 - Integrated Decision-Making Framework for Museums

Then, working on the specific case of digitalization of Italian museums, I empirically validated the IDM Framework for Museums by integrating two datasets, one open and one proprietary, to develop a dashboard to be used by museum managers and the Ministry of Culture.

In terms of Human dimension, the methodological part of the thesis that explains how to integrate data so as to obtain the relevant information to measure KPIs to be displayed in the dashboard, demonstrates how important the Human dimension of the IDM Framework is for supporting the other dimensions and improving decision-making. An example is the required competencies in data analysis that enabled the author of the thesis to harmonize and integrate the Proprietary and Istat Microdata datasets.

In terms of Data Management, the thesis also shows the added value of the data integration in terms of added value to quality of data, as shown in the methodological section by the various added information that are available from the integration of the two dataset, open and proprietary. Indeed, data integration of open data led to the enrichment of the proprietary data with many variables. Out of these, seven variables, namely Typology, Province, Municipality, Personnel, Presence of a dedicated website, Social media presence, and Online tours, were incorporated into the dashboard because of their relevance on the insights offered by the dashboard for decision makers.

Still in terms of data, the thesis highlights a crucial issue in data integration: the instance-level heterogeneities that do not allow the perfect matching of records. In this work, some museums have been lost in the integration because of issues with their

denominations. The issues encountered were the change of names over the years, typos in the survey answer of the denomination, and the use of a very general denomination that is not unique (e.g., there are many *Museo Civico* in Italy, and even in the same region). This integration issue can be solved with either a great amount of manual work, which may even include contacting the museums, or with the introduction of a univocal identifier that should be assigned to each museum. The identifier would be used as a key to enable the perfect matching of databases of museums and also to univocally identify a museum.

In terms of Measurement, the thesis also shows the contribution of the data integration to the dimension, as I developed three KPIs, specifically tailored to assess the digital situation in museums were developed, and an integrated dashboard, composed of data originating from different sources, was built. This is illustrated in Chapter 6Chapter 6: The dashboard by showing which variables are used in the dashboard and how the indicators are obtained. Indeed, the indicators are made as a combination of both variables from the proprietary dataset and variables from the open dataset. This combination is only possible to achieve thanks to the integration of open data with proprietary data.

In terms of Reporting, the thesis also shows the contributions of data integration to the dimension, as in the dashboard various visuals would not be available to the decision makers limiting their decisions and their benchmarking possibilities. Showing how the dashboard would have been without data integration and the information that could have been displayed, in the empirical case of museums' survey data, supports the objective of showing the effects of data integration.

The thesis adds to the extant literature in terms of *LG2* by introducing three general indicators, built following little insights found in the literature. These indicators may serve as the basis for the development of other KPIs, that should be tailored to the specific institution. The indicators divide the performance of museums into three dimensions: Online, On-site, and Organizational Readiness. The three KPIs are used to define the state of digital in a museum. The indicators should be used as a way to benchmark museums against other similar museums, but they can also provide information to museum managers and the Ministry of Culture by themselves.

The measurement of the Online indicator was made possible thanks to the integration of open data (from Istat) into the Proprietary dataset. This process highlights the importance of integrating data coming from external sources with proprietary datasets, to make the organization context-aware and to enable the possibility of benchmarking with similar museums, enriching knowledge on *LG1*. In fact, the thesis also serves the purpose of sparking an interest in the topic of data integration of open data in cultural institutions, which has not been discussed in the literature and should be investigated further.

Moreover, the dashboard serves as a tool for both the museums and the Ministry of Culture. The dashboard serves as an example of the impact of the integration of open data with proprietary data, which can be a valuable process for both museums and the Ministry of Culture. The integration possibilities are many and can be tailored to the needs and wants of the user. In the thesis, a broad approach was taken, providing a dashboard that gives a general overview of the state of digital in museums in Italy.

The Ministry of Culture can evaluate the status of museums across various dimensions to identify areas requiring improvement. To achieve these improvements, the Ministry can reallocate funds to areas in greater need or with significant potential for growth. Additionally, the Ministry can also promote initiatives targeted at museums falling within certain specific value ranges of dimensions, such as a specific geographical area, size, or typology, with the aim of fostering the improvement of performance in the museums that fall under the defined parameters.

It is also a helpful tool for museum managers to compare the performance of their museum with the performance of similar museums, to understand how they are positioned in the landscape, which may be based on the size of the museum, the typology, the revenue range, the geographical area, or a combination of the mentioned dimensions. Through these assessments, museum managers can strategically address areas where their institutions may be lacking compared to similar ones, implementing an external benchmarking logic. These informed decisions are grounded in data, allowing managers to take targeted actions based on performance data and avoid making decisions based on feelings and interpretation.

Another element of value that was introduced by data integration and harmonization is time. The harmonization of the four Proprietary surveys made it possible to analyze the performance of individual museums over time. In a context in which museums often do not collect data about their visitors or about their performance, the visualization of the evolution of performance over time is crucial for gaining insights into the changes happening in the museum, implementing an internal benchmarking logic. The visualization of performance over time is the best way to gauge the signs of progress that museums are achieving over time. The time dimension is also important for the general overview of museums, as it highlights the differences between regions and provinces over time and how those changed.

The importance of open data for this work is significant.

In the thesis, Istat was selected as source for open data, however, open data sources are many and they provide different information. Moreover, by reducing the geographical granularity of the analysis, the open databases that can be integrated with the proprietary databases increase in quantity and quality of variables and insights. This means that a more personalized and information-rich dashboard can be created by reducing the geographical scope of the analysis.

The thesis was developed with the idea of proving that data integration of open data can be an important process for museums. Museums should investigate the data integration of open data to enrich their proprietary databases and ultimately improve their ability to make decisions. While in the thesis a dashboard was created that is meant to be given as an external tool to both the Ministry of Culture and the museums, it's important to highlight the important task of the creation of a personalized dashboard by the museums themselves, following examples of integrated reporting. The dashboard is very broad in its focus, and it gives a very general overview of the museum situation in Italy. The museums could be interested in other aspects that were not considered when developing the dashboard. In this final section of the thesis, the development of a personalized dashboard by museums is encouraged. The tool can help them better visualize and understand the performance of their institution and can also enable benchmarking against similar museums. However, the museums should be supported in this task by increasing funding and by helping them to increase their data and digital capabilities. Unfortunately, this decrease seems to be an unavoidable and atavistic issue that has never been fixed and just keeps getting worse. The issue is not caused by the inefficiencies of museums but directly depends on the substantial cuts made by governments to the Ministry of Culture budget, especially in Italy.

## 8.2. Limitations and further developments

Though the thesis has been rigorously conducted, the research presents at least four limitations.

First, the variables in the dataset have a yearly frequency. This leads to the creation of indicators that are not rapidly actionable but rather show the performance of the whole year. Moreover, the datasets are composed of mostly binary variables that represent a *Yes* or *No* answer, overlooking everything that is in between. This leads to the creation of indicators that are limited in terms of the difference in quality (e.g., the presence of a dedicated website does not show the quality and effort put into the website). Future research should consider collecting data more frequently or integrating data with higher detail on frequency to better show the change in performance over time. Moreover, broadening the range of data categories from binary to more categories or even extend to other types of data formats beyond numerical values, such as text data, would assist in capturing more qualitative aspects and further enrich the decision making of museums.

Second, the indicators used in the dashboard are all performance-based, while it has not been possible to construct quality-based indicators. The absence of quality-based indicators limits the depth and comprehensiveness of the dashboard. Without these indicators, the evaluation focuses solely on performance metrics, overlooking crucial aspects related to the quality of museum services and offerings. This limitation restricts the ability to provide a more nuanced and holistic understanding of museums' overall effectiveness and impact. Moreover, there are no indicators that act as a proxy for the

social value provided by the museums. Though the literature so far claims that those indicators cannot be used for benchmarking because they are computed personally by museums and their computation is very subjective, future research should focus on the development of social value proxy measures that can be computed in a standard way and, thus, used for benchmarking. Indeed, indicators that can be used as proxies for measuring the social value of museums exist, but they currently lack the possibility to use them as comparing tools between museums. This is because museums themselves compute these indicators using varying metrics. To enable true comparisons between museums based on the social value they provide to society, there is a need for either the development of a new, standardized, and objective indicator or the establishment of an independent entity that is responsible for the computation and reporting of social value indicators.

Third, within the thesis, the integration was limited to a specific set of datasets, namely the Istat Microdata datasets. However, the integration can be expanded by introducing more open databases, especially if the geographical scope of the dashboard is reduced. That is due to the fact that open databases are very fragmented and mostly specific to certain geographical areas. Future research could consider refining the geographical scope of the dashboard, as narrowing down the geographical area would allow for more comprehensive integration of open databases and improve the quality of insights.

Fourth, the context of the empirical research of the thesis is limited to Italian museums, but the research could extend to other geographical contexts, not necessarily in the national scope. Future research should consider expanding the scope to the EU to get a broader overview of the museum situation and to also enable benchmarking of Italian museums against museums that belong to the EU. It could also be interesting to see the differences throughout the EU and to grasp the effects caused by different amounts of funding to culture in the different EU nations.

# Sitography

https://cultura.gov.it

https://ec.europa.eu

https://extremepresentation.typepad.com

https://musefirenze.it

https://towardsdatascience.com

https://www.colorblindguide.com

https://www.color-blindness.com

https://www.fao.org

https://www.ft.com

https://www.istat.it

https://www.legadelfilodoro.it

https://obamawhitehouse.archives.gov

https://www.oecd.org

https://obamawhitehouse.archives.gov

https://www.openpolis.it

https://www.oracle.com

https://www.worldbank.org

# Bibliography

Aas, T. H., & Alaassar, A. (2018). The impact of visual performance management on decision-making in the entrepreneurial process. International Journal of Innovation Management, 22(05), 1840002.

Abdel-Maksoud, A., Elbanna, S., Mahama, H. and Pollanen, R. (2015), The use of performance information in strategic decision making in public organizations, International Journal of Public Sector Management, Vol. 28 No. 7, pp. 528-549.

Accurat. (2021). Project Showcase: Integrated Report for Museo Egizio. Retrieved from: https://medium.com/accurat-in-sight/project-showcase-integrated-report-for-museo-egizio-ab11273a34de

Agostino, D., & Arnaboldi, M. (2021). From preservation to entertainment: Accounting for the transformation of participation in Italian state museums. Accounting History, 26(1), 102-122.

Agostino, D., & Costantini, C. (2022). A measurement framework for assessing the digital transformation of cultural institutions: the Italian case. Meditari Accountancy Research, 30(4), 1141-1168.

Agostino, D., Arnaboldi, M., & Carloni, E. (2020). Big data for decision making: are museums ready?. Management, Participation and Entrepreneurship in the Cultural and Creative Sector, 61-78.

Anderson, G. (2004). Reinventing the Museum: Historical and Contemporary Perspectives on the Paradigm Shift. AltaMira Press.

Arena, M., Azzone, G., & Bengo, I. (2015). Performance Measurement for Social Enterprises. VOLUNTAS: International Journal of Voluntary and Nonprofit Organizations, 26(2), 649–672.

Arksey, H., & O'Malley, L. (2005). Scoping studies: towards a methodological framework. International journal of social research methodology, 8(1), 19-32.

Arnaboldi, M. (2018). The Missing Variable in Big Data for Social Sciences: The Decision-Maker. Sustainability, 10(10), 3415.

Arnaboldi, M., Busco, C., & Cuganesan, S. (2017). Accounting, accountability, social media and big data: revolution or hype?. Accounting, auditing & accountability journal, 30(4), 762-776.

Arnaboldi, M., Lapsley, I., & Steccolini, I. (2015). Performance management in the public sector: The ultimate challenge. Financial Accountability & Management, 31(1), 1-22.

Arputhamary, B., & Arockiam, L. (2015). Data integration in Big Data environment. Bonfring International Journal of Data Mining, 5(1), 1-5.

Arvidson, M., Lyon, F., McKay, S., & Moro, D. (2010). The ambitions and challenges of SROI. Third Sector Research Center.

Avillach, P., Coloma, P. M., Gini, R., Schuemie, M., Mougin, F., Dufour, J. C., ... & Trifirò, G. (2013). Harmonization process for the identification of medical events in eight European healthcare databases: the experience from the EU-ADR project. Journal of the American Medical Informatics Association, 20(1), 184-192.

Axelrod, R., & Hamilton, W. D. (1981). The evolution of cooperation. Science, 211(4489), 1390–1396.

Bagnoli, L., & Megali, C. (2011). Measuring performance in social enterprises. Nonprofit and Voluntary Sector Quarterly, 40(1), 149-165.

Batini, C., & Scannapieco, M. (2016). Data and information quality: Dimensions, principles and Techniques. Springer.

Bekele, M. K., Pierdicca, R., Frontoni, E., Malinverni, E. S., & Gain, J. (2018). A survey of augmented, virtual, and mixed reality for cultural heritage. Journal on Computing and Cultural Heritage (JOCCH), 11(2), 1-36.

Berinato, S. (2016). Good charts: The HBR guide to making smarter, more persuasive data visualizations. Harvard Business Review Press.

Berlanga, R., & Nebot, V. (2016). Context-Aware Business Intelligence. In Business Intelligence: 5th European Summer School, eBISS 2015, Barcelona, Spain, July 5-10, 2015, Tutorial Lectures 5 (pp. 87-110). Springer International Publishing.

Berners-Lee, T. (2012, November 9). Raw data, now!. Wired. Retrieved from https://www.wired.co.uk/article/raw-data

Bertacchini, E., & Morando, F. (2013). The future of museums in the digital age: New models for access to and use of digital collections. International journal of arts management, 15(2), 60-72.

Bertot, J. C., Jaeger, P. T., & Grimes, J. M. (2010). Using ICTs to create a culture of transparency: E-government and social media as openness and anti-corruption tools for societies. Government Information Quarterly, 27(3), 264–271.

Bhimani, A., & Willcocks, L. (2014). Digitisation, 'Big Data' and the transformation of accounting information. Accounting and Business Research, 44(4), 469–490.

Bisbe, J., & Malagueño, R. (2012). Using strategic performance measurement systems for strategy formulation: Does it work in dynamic environments?. Management Accounting Research, 23(4), 296-311.

Bishop, P., & Brand, S. (2003). The efficiency of museums: a stochastic frontier production function approach. Applied Economics, 35(17), 1853-1858.

Bonet, L., & Négrier, E. (2018). The participative turn in cultural policy: Paradigms, models, contexts. Poetics, 66, 64–73.

Bross, I. D. (1953). Design for decision.

Burkhard, R. A. (2004, July). Learning from architects: the difference between knowledge visualization and information visualization. In Proceedings. Eighth International Conference on Information Visualisation, 2004. IV 2004. (pp. 519-524). IEEE.

Camarero, C., & Garrido, M. J. (2012). Fostering innovation in cultural contexts: Market orientation, service orientation, and innovations in museums. Journal of service research, 15(1), 39-58.

Camarero, C., & Garrido, M.-J. (2009). Improving Museums' Performance Through Custodial, Sales, and Customer Orientations. Nonprofit and Voluntary Sector Quarterly, 38(5), 846-868.

Camarero, C., Garrido, M. J., Vicente, E., & Redondo, M. (2019). Relationship marketing in museums: influence of managers and mode of governance. Public Management Review, 21(10), 1369-1396.

Camarero, C., Garrido, M.J. & Vicente, E. (2011). How cultural organizations' size and funding influence innovation and performance: the case of museums. J Cult Econ 35, 247–266

Castelnovo, W. (2017), I musei alla sfida della innovazione data-driven: come creare valore nell'universo digitale, Prospettive in Organizzazione, Vol. 8.

Cerquetti, M. (2016). More is better! Current issues and challenges for museum audience development: a literature review. Current Issues and Challenges for Museum Audience Development: A Literature Review (December 1, 2016). JOURNAL OF CULTURAL MANAGEMENT & POLICY, 6(1).

Chang, W. L., & Grady, N. (2019). Nist big data interoperability framework: Volume 1, definitions.

Chen, C.P. and Zhang, C.Y. (2014), Data-intensive applications, challenges, techniques and technologies: a survey on big data, Information Sciences, Vol. 275, pp. 314-347.

Chianese, A., & Piccialli, F. (2018). A perspective on applications of in-memory and associative approaches supporting cultural big data analytics. International Journal of Computational Science and Engineering, 16(3), 219-233.

Chiaravalloti, F., & Piber, M. (2011). Ethical implications of methodological settings in arts management research: The case of performance evaluation. The Journal of Arts Management, Law, and Society, 41(4), 240-266.

Conn, S. (2010). Do museums still need objects?. University of Pennsylvania Press.

Constantiou, I. D., & Kallinikos, J. (2015). New games, new rules: big data and the changing context of strategy. Journal of Information Technology, 30, 44-57.

Cronbach, L.J. and Gleser, G.C. (1957). Psychological tests and personnel decisions. Urbana: University of Illinois Press.

Cross, K.F. & Lynch, R.L. (1988). The "SMART" Way To Define and Sustain Success. National Productivity Review, 8 (1), 23-33.

Dai, C., Lin, D., Bertino, E., & Kantarcioglu, M. (2008). An approach to evaluate data trustworthiness based on data provenance. In Secure Data Management: 5th VLDB Workshop, SDM 2008, Auckland, New Zealand, August 24, 2008. Proceedings 5 (pp. 82-98). Springer Berlin Heidelberg.

Daniel, D.R. (1961). Management information crisis. Harvard Business Review, 39 (5). 111-121.

Dayal, U., Castellanos, M., Simitsis, A., & Wilkinson, K. (2009). Data Integration Flows for Business Intelligence. Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, 1–11. Presented at the Saint Petersburg, Russia.

De Bernardi, P., Bertello, A., & Shams, S. M. (2019). Logics hindering digital transformation in cultural heritage strategic management: An exploratory case study. Tourism Analysis, 24(3), 315-327.

De Bernardi, P., Gilli, M., & Colomba, C. (2018). Unlockingmuseum digital innovation. Are 4.0 Torino museums? In V. Cantino, F. Culasso, & G. Racca (Eds.), Smart tourism (pp. 453–471). New York, NY: McGraw Hill Education.

de Bruijn, H., Warnier, M., & Janssen, M. (2022). The perils and pitfalls of explainable AI: Strategies for explaining algorithmic decision-making. Government Information Quarterly, 39(2), [101666].

De Santis, F., & Presti, C. (2018). The relationship between intellectual capital and big data: a review. Meditari Accountancy Research, 26(3), 361-380.

de Waal, A.A. (2007), Strategic Performance Management, a Managerial and Behavioural Approach, Palgrave Macmillan, London.

Dean, J. W., & Sharfman, M. P. (1996). Does Decision Process Matter? A Study of Strategic Decision-Making Effectiveness. The Academy of Management Journal, 39(2), 368–396.

del Mar Roldán-García, M., García-Nieto, J., Maté, A., Trujillo, J., & Aldana-Montes, J. F. (2021). Ontology-driven approach for KPI meta-modelling, selection and reasoning. International Journal of Information Management, 58, 102018.

del-Rey-Chamorro, F.M., Roy, R., van Wegen, B. and Steele, A. (2003), A framework to create key performance indicators for knowledge management solutions, Journal of Knowledge Management, Vol. 7 No. 2, pp. 46-62.

Domínguez, E., Pérez, B., Rubio, Á. L., & Zapata, M. A. (2019). A taxonomy for key performance indicators management. Computer Standards and Interfaces, 64, 24-40.

Dumay, J., & Dai, T. (2017). Integrated thinking as a cultural control?. Meditari Accountancy Research, 25(4), 574-604.

Dunn, H. L. (1946). Record linkage. American Journal of Public Health and the Nations Health, 36(12), 1412-1416.

D'Zurilla, T. J., & Goldfried, M. R. (1971). Problem solving and behavior modification. Journal of Abnormal Psychology, 78(1), 107–126.

Earley, S., & Henderson, D. (2017). Dama-DMBOK: Data Management Body of Knowledge. Basking Ridge, NJ: Technics Publications.

Ebrahim, A., & Rangan, V. K. (2014). What Impact? A Framework for Measuring the Scale and Scope of Social Performance. California Management Review, 56(3), 118-141.

Elbanna, S. and Child, J. (2007), Influences on strategic decision effectiveness: Development and test of an integrative model. Strat. Mgmt. J., 28: 431-453.

Elbanna, S., Thanos, I. & Jansen, R. (2020). A Literature Review of the Strategic Decision-Making Context: A Synthesis of Previous Mixed Findings and an Agenda for the Way Forward. M@n@gement, 23, 42-60.

Elbashir, M.Z., Sutton, S.G., Arnold, V. and Collier, P.A. (2022), "Leveraging business intelligence systems to enhance management control and business process performance in the public sector", Meditari Accountancy Research, Vol. 30 No. 4, pp. 914-940.

Eppler, M. J., & Bresciani, S. (2013). Visualization in management: From communication to collaboration. A response to Zhang. Journal of Visual Languages & Computing, 24(2), 146–149.

European Union Agency for Fundamental Rights, (2022). Bias in algorithms : artificial intelligence and discrimination, Publications Office of the European Union.

Euske, K. J. (2003). Public, private, not-for-profit: everybody is unique?. Measuring Business Excellence, 7(4), 5-11.

Euske, K.J. & Zander, L.A. (2005). History of Business Performance Measurement. Encyclopedia of Social Measurement, 2, 227-232.

Farnadi, G., Sitaraman, G., Sushmita, S., Celli, F., Kosinski, M., Stillwell, D., ... & De Cock, M. (2016). Computational personality recognition in social media. User modeling and user-adapted interaction, 26, 109-142.

Fellegi, I. P., & Sunter, A. B. (1969). A theory for record linkage. Journal of the American Statistical Association, 64(328), 1183-1210.

Feng, T., Cummings, L., & Tweedie, D. (2017). Exploring integrated thinking in integrated reporting–an exploratory study in Australia. Journal of Intellectual Capital, 18(2), 330-353.

Firnkorn, D., Ganzinger, M., Muley, T., Thomas, M., & Knaup, P. (2015). A generic data harmonization process for cross-linked research and network interaction. Methods of information in medicine, 54(05), 455-460.

Fondazione Museo delle Antichità Egizie di Torino. (2021). Report Integrato 2020. Torino. Retrieved from: https://drive.google.com/file/d/1uecM6ha96-YEJPnrOBCSkoxIo9Gej3XT/view. Accessed 17/11/2023.

Fopp, M. A. (1997). The Implications of Emerging Technologies for Museums and Galleries. Museum Management and Curatorship, 16(2), 143–153.

Franco-Santos, M., Lucianetti, L., & Bourne, M. (2012). Contemporary performance measurement systems: A review of their consequences and a framework for research. Management Accounting Research, 23(2), 79-119.

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. International journal of information management, 35(2), 137-144.

Garengo, P., & Sardi, A. (2021). Performance measurement and management in the public sector: state of the art and research opportunities. International Journal of Productivity and Performance Management, 70(7), 1629-1654.

Gavrilova, T. A., Alsufyev, A. I., & Grinberg, E. Y. (2017). Knowledge visualization: critique of the St. Gallen School and an analysis of contemporary trends. Business Informatics, (3), 7-19.

Gavrilova, T., Kudryavtsev, D., & Grinberg, E. (2019). Aesthetic knowledge diagrams: Bridging understanding and communication. In D. Carlucc (Ed.), Knowledge management, arts, and humanities (pp. 97–117). Springer

Gelatt, H. B. (1962). Decision-making: A conceptual frame of reference for counseling. Journal of Counseling Psychology, 9(3), 240–245.

Ghalayini, A. M., & Noble, J. S. (1996). The changing basis of performance measurement. International Journal of Operations & Production Management, 16(8), 63-80.

Giaccardi, E. (Ed.). (2012). Heritage and social media: Understanding heritage in a participatory culture. Routledge.

Gitelman, L. (Ed.) (2013). "Raw Data" Is an Oxymoron. (Infrastructures series). MIT Press. http://mitpress-ebooks.mit.edu/product/raw-data-oxymoron

Gombault, A., Allal-Chérif, O., & Décamps, A. (2016). ICT adoption in heritage organizations: Crossing the chasm. Journal of Business Research, 69(11), 5135-5140.

González, A., Wilby, M., Díaz, J., Pozo, R., & Ávila, C. (2021). Utilization rate of the fleet: a novel performance metric for a novel shared mobility. Transportation, 1-17.

Gonzalez, R. (2017). Keep the Conversation Going: How Museums Use Social Media to Engage the Public. The Museum Scholar. 1(1).

Graham, C. (2008). Fearful asymmetry: The consumption of accounting signs in the Algoma Steel pension bailout. Accounting, Organizations and Society, 33(7), 756–782.

Grandon, E. E., & Pearson, J. M. (2004). Electronic commerce adoption: an empirical study of small and medium US businesses. Information & management, 42(1), 197-216.

Grosswiele, L., Roeglinger, M., & Friedl, B. (2013). A Decision Framework for the Consolidation of Performance Measurement Systems. Decision Support Systems, 54, 1016–1029.

Gstraunthaler, T., & Piber, M. (2012). The performance of museums and other cultural institutions (Vol. 42). Presented at the International Studies of Management and Organization.

Guccio, C., Martorana, M., Mazza, I., Pignataro, G., & Rizzo, I. (2020). An analysis of the managerial performance of Italian museums using a generalised conditional efficiency model. Socio-Economic Planning Sciences, 72, 100891.

Guccio, C., Mignosa, A., & Rizzo, I. (2018). Are public state libraries efficient? An empirical assessment using network Data Envelopment Analysis. Socio-Economic Planning Sciences, 64, 78-91.

Günther, W. A., Mehrizi, M. H. R., Huysman, M., & Feldberg, F. (2017). Debating big data: A literature review on realizing value from big data. The Journal of Strategic Information Systems, 26(3), 191-209.

Gutierrez, D. M., Scavarda, L. F., Fiorencio, L., & Martins, R. A. (2015). Evolution of the performance measurement system in the Logistics Department of a broadcasting company: An action research. International Journal of Production Economics, 160, 1-12.

Haber, J., & Schryver, C. (2019). How to create key performance indicators. The CPA Journal, 89(4), 24-30.

Harrison, T., F. Luna-Reyes, L., Pardo, T., De Paula, N., Najafabadi, M., & Palmer, J. (2019, June). The data firehose and AI in government: Why data management is a key to value and ethics. In Proceedings of the 20th annual international conference on digital government research (pp. 171-176).

Heads of the National Statistical Institutes, & Director-General of Eurostat. (2016). Quality Declaration of the European Statistical System. European Statistical System. Retrieved from: https://ec.europa.eu/eurostat/web/products-catalogues/-/ks-02-17-428

Hendler, J. (2014). Data Integration for Heterogenous Datasets. Big Data, 2(4), 205–215.

Höchtl, J., Parycek, P., & Schöllhammer, R. (2016). Big data in the policy cycle: Policy decision making in the digital era. Journal of Organizational Computing and Electronic Commerce, 26(1-2), 147-169.

Hoque, Z., & Adams, C. (2011). The rise and use of balanced scorecard measures in Australian government departments. Financial Accountability & Management, 27(3), 308-334.

Hume, M., & Mills, M. (2011). Building the sustainable iMuseum: is the virtual museum leaving our museums virtually empty?. International Journal of Nonprofit and Voluntary Sector Marketing, 16(3), 275-289.

Iacovou, C. L., Benbasat, I., & Dexter, A. S. (1995). Electronic data interchange and small organizations: Adoption and impact of technology. MIS quarterly, 465-485.

IBM (2021). Structured vs. unstructured data: What's the difference? IBM Blog. https://www.ibm.com/blog/structured-vs-unstructured-data/

IIRC. (2013). Integrated reporting: the international framework. Retrieved from: https://integratedreporting.org/wp-content/uploads/2013/12/13-12-08-THE-INTERNATIONAL-IR-FRAMEWORK-2-1.pdf. Accessed 17/11/2023.

IIRC. (2013). Integrated reporting: the international framework. Retrieved from: https://www.integratedreporting.org/wp-content/uploads/2022/08/IntegratedReportingFramework_081922.pdf. Accessed 17/11/2023.

Internation Council of Museums, World Federation of Friends of Museums. (2018). Museums, Social Landmarks: Declaration of Funchal at the European Year of Cultural

Heritage. Funchal. Retrieved from: https://icom-europe.mini.icom.museum/wp-content/uploads/sites/24/2019/01/Declaration_of_Funchal__English__02.pdf

International Council of Museums. (1974). ICOM Statutes: Adopted by the Eleventh General Assembly of Icom. Copenhagen. Retrieved from https://books.google.it/books?id=W47SxgEACAAJ

International Council of Museums. (2007). Resolutions Adopted By Icom's 22nd General Assembly. Wien. Retrieved from: https://icom.museum/wp-content/uploads/2018/07/ICOMs-Resolutions_2007_Eng.pdf

International Council of Museums. (2022). Extraordinary General Assembly. Retrieved from: https://icom.museum/en/resources/standards-guidelines/museum-definition/. Accessed 17/11/2023.

International Organization for Standardization. (2014). Automation systems and integration — Key performance indicators (KPIs) for manufacturing operations management — Part 1: Overview, concepts and terminology (ISO Standard No. 22400-1). Retrieved from: https://www.iso.org/standard/56847.html

Jafari, A., Taheri, B., & Vom Lehn, D. (2013). Cultural consumption, interactive sociality, and the museum. Journal of Marketing Management, 29(15-16), 1729-1752.

Janis, I. L., & Mann, L. (1977). Decision making: A psychological analysis of conflict, choice, and commitment. New York: Free Press.

Jaro, M. A. (1989). Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. Journal of the American Statistical Association, 84(406), 414-420.

Jensen, M. C., & Meckling, W. H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. Journal of Financial Economics, 3(4), 305–360.

Kabir, N., & Carayannis, E. (2013, January). Big data, tacit knowledge and organizational competitiveness. In proceedings of the 10th international conference on intellectual capital, Knowledge Management and Organisational Learning: ICICKM (p. 220).

Kalampokis, E., Tambouris, E., Karamanou, A., Tarabanis, K. (2016). Open Statistics: The Rise of a New Era for Open Data?. In: Scholl, H.J., et al. Electronic Government. EGOV 2016. Lecture Notes in Computer Science(), vol 9820. Springer, Cham.

Kaplan, R.S. & Norton, D.P. (1992). The Balanced Scorecard - Measures that Drive Performance. Harvard Business Review, 70 (1), 71-79.

Kaplan, R.S. (2001), "Strategic performance measurement and management in non-profit organisations", Non-profit Management and Leadership, Vol. 11 No. 3, pp. 353-70.

Kaufman, A. R. & Klevs, A. (2021). Adaptive Fuzzy String Matching: How to Merge Datasets with Only One (Messy) Identifying Field. Political Analysis, 30(4), 590-596.

Kaufman, A. R., & Klevs, A. (2022). Adaptive Fuzzy String Matching: How to Merge Datasets with Only One (Messy) Identifying Field. Political Analysis, 30(4), 590-596.

Keegan, D. P., Eiler, R. G., & Jones, C. R. (1989). Are your performance measures obsolete?. Strategic Finance, 70(12), 45.

Kéfi, H., & Pallud, J. (2011). The role of technologies in cultural mediation in museums: an Actor-Network Theory view applied in France. Museum Management and Curatorship, 26(3), 273-289.

Khan, K. & Shah, A. (2011). Understanding performance measurement through the literature. African Journal of Business Management, 5 (35), 13410-13418.

Kitchin, R., Lauriault, T. P., & McArdle, G. (2015). Smart cities and the politics of urban data. Smart urbanism: Utopian vision or false dawn, 16-33.

Kriegel, H. P., Kröger, P., Sander, J., & Zimek, A. (2011). Density-based clustering. Wiley interdisciplinary reviews: data mining and knowledge discovery, 1(3), 231-240.

Kucukaltan, B., Irani, Z., & Aktas, E. (2016). A decision support model for identification and prioritization of key performance indicators in the logistics industry. Computers in Human Behavior, 65, 346-358.

Kueng, P. (2000). Process performance measurement system: A tool to support process-based organizations. Total Quality Management, 11(1), 67-85.

Kumar, G., Basri, S., Imam, A. A., Khowaja, S. A., Capretz, L. F., & Balogun, A. O. (2021). Data Harmonization for Heterogeneous Datasets: A Systematic Literature Review. Applied Sciences, 11(17), 8275.

Kundra, V. (2010, May 21). Data.gov: Pretty Advanced for a One-Year-Old [web log]. Retrieved from https://obamawhitehouse.archives.gov/blog/2010/05/21/datagov-pretty-advanced-one-year-old

Kwokwah Yeung, A., & Connell, J. (2006). The Application of Niven's Balanced Scorecard in a Not-For-Profit Organization in Hong Kong: What Are the Factors for Success?. Journal of Asia Business Studies, 1(1), 26-33.

La Torre, M., Bernardi, C., Guthrie, J., & Dumay, J. (2019). Integrated reporting and integrating thinking: practical challenges. Challenges in managing sustainable business: Reporting, taxation, ethics and governance, 25-54.

Labaronne, L., & Piber, M. (2020). Performance measurement and evaluation in the arts and cultural sector: State-of-the-art in theory and practice and prolegomena for

further developments. Management, participation and entrepreneurship in the cultural and creative sector, 219-240.

Laney, D. (2001). 3D Data Management: Controlling Data Volume, Velocity, and Variety (). META Group.

Lawlor, E., Neitzert, E., & Nicholls, J. (2008). (publication). Measuring value: a guide to Social Return on Investment (SROI) (2nd ed.). The New Economics Foundation.

Lazar, J., Feng, J. H., & Hochheiser, H. (2017). Research methods in human-computer interaction. Morgan Kaufmann.

Lehmannová, M. (2020). (rep.). 224 YEARS OF DEFINING THE MUSEUM. International Council of Museums. Retrieved from https://icom.museum/wp-content/uploads/2020/12/2020_ICOM-Czech-Republic_224-years-of-defining-the-museum.pdf

Lenzerini, M. (2002, June). Data integration: A theoretical perspective. In Proceedings of the twenty-first ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems (pp. 233-246).

Lindqvist, K. (2012). Museum finances: challenges beyond economic crises. Museum Management and Curatorship, 27(1), 1-15.

Liu, W. C. (2008). Visitor study and operational development of museums. Museology Quarterly, 22(3), 21-37.

Loach, K., Rowley, J., & Griffiths, J. (2017). Cultural sustainability as a strategy for the survival of museums and libraries. International journal of cultural policy, 23(2), 186-198.

Lombardo, G. (2018). L'impatto Generato E La Valutazione Del Social-Roi. In Annual Report 2018 (pp. 176-193). MUS.E.

Lorenz, L., van Erp, J., & Meijer, A. (2022). Machine-learning algorithms in regulatory practice: Nine organisational challenges for regulatory agencies. Technology and Regulation, 2022, 1–11.

Lurie, N. H., & Mason, C. H. (2007). Visual representation: Implications for decision making. Journal of marketing, 71(1), 160-177.

Madnick, S. E., Wang, R. Y., Lee, Y. W., & Zhu, H. (2009). Overview and framework for data and information quality research. Journal of data and information quality (JDIQ), 1(1), 1-22.

Maheshwari, D., & Janssen, M. (2014). Reconceptualizing measuring, benchmarking for improving interoperability in smart ecosystems: The effect of ubiquitous data and crowdsourcing. Government Information Quarterly, 31, S84-S92.

Manyika, J., Chui, M., Groves, P., Farrell, D., Van Kuiken, S., & Almasi Doshi, E. (2013). (rep.). Open data: Unlocking innovation and performance with liquid information. McKinsey Global Institute.

Marty, P. F. (2006). Finding the skills for tomorrow: Information literacy and museum information professionals. Museum Management and Curatorship, 21(4), 317–335.

Marty, P.F. (2007), The changing nature of information work in museums. J. Am. Soc. Inf. Sci., 58: 97-107.

Matarasso, F. (1997). Use or ornament. The social impact of participation in the arts, 4(2), 34-41.

McCandless, D. (2012). The visual miscellaneum: A colorful guide to the world's most consequential trivia. New York: Harper Design.

McKiernan, E. C., Bourne, P. E., Brown, C. T., Buck, S., Kenall, A., Lin, J., … Yarkoni, T. (2016). How open science helps researchers succeed. eLife, 5, e16800.

Mehrotra, S. and Verma, S. (2015), "An assessment approach for enhancing the organizational performance of social enterprises in India", Journal of Entrepreneurship in Emerging Economies, Vol. 7 No. 1, pp. 35-54.

Mehrtens, J., Cragg, P. B., & Mills, A. M. (2001). A model of Internet adoption by SMEs. Information & management, 39(3), 165-176.

Merton R.K. (1942), Science and technology in a democratic order, «Journal of legal and political sociology», 1, 115-126

Millar, R., & Hall, K. (2013). Social Return on Investment (SROI) and Performance Measurement. Public Management Review, 15(6), 923–941.

Mintzberg, H., Raisinghani, D., & Théorêt, A. (1976). The Structure of "Unstructured" Decision Processes. Administrative Science Quarterly, 21(2), 246–275.

Moldavanova, A. (2016). Two narratives of intergenerational sustainability: A framework for sustainable thinking. The American Review of Public Administration, 46(5), 526-545.

Molla, A., & Licker, P. S. (2005). eCommerce adoption in developing countries: a model and instrument. Information & management, 42(6), 877-899.

Moullin, M. (2002), Delivering Excellence in Health and Social Care, Open University Press, Buckingham.

Moullin, M. (2007). Performance measurement definitions: Linking performance measurement and organisational excellence. International journal of health care quality assurance, 20(3), 181-183.

Moullin, M. (2017). Improving and evaluating performance with the Public Sector Scorecard. International Journal of Productivity and Performance Management, 66(4), 442-458.

Moustaghfir, K., Schiuma, G., & Carlucci, D. (2016). Rethinking performance management: a behaviour-based perspective. International Journal of Innovation and Learning, 20(2), 169-184.

Moynihan, D.P. (2005), Goal-Based Learning and the Future of Performance Management. Public Administration Review, 65: 203-216.

Munik, J., Pinheiro de Lima, E., Deschamps, F., Gouvea Da Costa, S.E., Van Aken, E.M., Almeida Prado Cestari, J.M., Moura, L.F. and Treinta, F. (2021), "Performance measurement systems in nonprofit organizations: an authorship-based literature review", Measuring Business Excellence, Vol. 25 No. 3, pp. 245-270.

Naylor, R. (2016), "Measuring the value of museums: options, tips and benefits". Network of European Museum Organisations. Money Matters: The Economic Value of Museums. 24. 27-29. Retrieved from: https://www.ne-mo.org/fileadmin/Dateien/public/NEMO_documents/NEMOAC2016_EcoVal.pdf.

Neely, A., Gregory, M. and Platts, K. (1995), "Performance measurement system design: A literature review and research agenda", International Journal of Operations & Production Management, Vol. 15 No. 4, pp. 80-116.

Neely, A., Gregory, M., & Platts, K. (2005). Performance measurement system design: A literature review and research agenda. International Journal of Operations & Production Management, 25, 1228–1263.

Neri, A., Cagno, E., Lepri, M., & Trianni, A. (2021). A triple bottom line balanced set of key performance indicators to measure the sustainability performance of industrial supply chains. Sustainable Production and Consumption, 26, 648-691.

New Philanthropy Capital (NPC) (2010) Social return on investment. London: New Philanthropy Capital.

Nicholls, J., Lawlor, E., Neitzert, E. and Goodspeed, T. (2009) A guide to social return on investment. London: Office of the Third Sector, The Cabinet Office.

Nigro, C., Iannuzzi, E., Petracca, M., & Montagnano, V. (2016, June), L'adozione delle ICT in un campione di musei europei. Paper presented at the XXVIII Sinergie Annual Conference "Management in a Digital World. Decisions, Productions, Communications", Udine.

NISO. (2004). (publication). Understanding Metadata. NISO Press. Retrieved from https://web.archive.org/web/20141107022958/http:/www.niso.org/publications/press/UnderstandingMetadata.pdf

Norman, D. (1986). User centered system design. New perspectives on human-computer interaction.

Open Data Charter (2015). International Open Data Charter. Open Data Charter. Retrieved from https://opendatacharter.net

Open Knowledge Foundation. (2015). Open Definition 2.1. Retrieved from https://opendefinition.org/od/2.1/en/

Othman, M.K., Petrie, H., Power, C.: Engaging visitors in museums with technology: scales for the measurement of visitor and multimedia guide experience. In: Campos, P., Graham, N., Jorge, J., Nunes, N., Palanque, P., Winckler, M. (eds.) INTERACT 2011. LNCS.

Papadakis, V. M., Lioukas, S., & Chambers, D. (1998). Strategic Decision-Making Processes: The Role of Management and Context. Strategic Management Journal, 19(2), 115–147.

Park, N. H., & Lee, W. S. (2004). Statistical grid-based clustering over data streams. Acm Sigmod Record, 33(1), 32-37.

Parmar, R., Mackenzie, I., Cohn, D. and Gann, D. (2014), "The new patterns of innovation", Harvard Business Review, Vol. 92 No. 2, pp. 86-95.

Parmenter, D. (2007) Key Performance Indicators: Developing, Implementing and Using Winning KPIs. John Wiley & Sons, Inc., Hoboken, 236 p.

Parmenter, D. (2015) Key Performance Indicators-Developing, Implementing, and Using Winning KPIs. 3rd Edition, Wiley, Hoboken.

Parmenter, D. (2020). Key performance indicators: Developing, implementing, and using winning Kpis (4th ed.). Hoboken, NJ: John Wiley & Sons, Inc.

Peacock, D. (2008). Making ways for change: Museums, disruptive technologies and organisational change. Museum management and curatorship, 23(4), 333-351.

Peral, J., Maté, A., & Marco, M. (2017). Application of data mining techniques to identify relevant key performance indicators. Computer Standards & Interfaces, 54, 76-85.

Pesce, D., Neirotti, P., & Paolucci, E. (2019). When culture meets digital platforms: value creation and stakeholders' alignment in big data use. Current Issues in Tourism, 22(15), 1883–1903.

Pfeffer, J., & Sutton, R. I. (2006). Evidence-based management. Harvard business review, 84(1), 62.

Piber, M., Demartini, P., & Biondi, L. (2019). The management of participatory cultural initiatives: Learning from the discourse on intellectual capital. Journal of Management and Governance, 23, 435-458.

Pieterson, W., & ICF. (2019). DIGITAL TECHNOLOGIES AND ADVANCED ANALYTICS IN PES. Luxembourg: Publications Office of the European Union.

Pollanen, R., Abdel-Maksoud, A., Elbanna, S., & Mahama, H. (2017). Relationships between strategic performance measures, strategic decision-making, and organizational performance: empirical evidence from Canadian public organizations. Public Management Review, 19(5), 725–746.

Porter, C. H., Villalobos, C., Holzworth, D., Nelson, R., White, J. W., Athanasiadis, I. N., ... & Jones, J. W. (2014). Harmonization and translation of crop modeling data to ensure interoperability. Environmental modelling & software, 62, 495-508.

Provost, F., & Fawcett, T. (2013). Data science and its relationship to big data and data-driven decision making. Big data, 1(1), 51-59.

Publications Office of the European Union, Assen, M., Cecconi, G., Carsaniga, G. (2022). Open data maturity report 2022, Publications Office of the European Union.

Publications Office of the European Union. (2019). Directive (EU) 2019/882 of the European Parliament and of the Council of 17 April 2019 on the accessibility requirements for products and services. Retrieved from: https://eur-lex.europa.eu/eli/dir/2019/882/oj

Quach, S., Thaichon, P., Martin, K. D., Weaven, S., & Palmatier, R. W. (2022). Digital technologies: tensions in privacy and data. Journal of the Academy of Marketing Science, 50(6), 1299–1323.

Quattrone, P. (2016). Management accounting goes digital: Will the move make it wiser?. Management accounting research, 31, 118-122.

Raguseo, E., Pigni, F., & Piccoli, G. (2018). Conceptualization, operationalization, and validation of the digital data stream readiness index. Journal of Global Information Management (JGIM), 26(4), 92-112.

Rainey, S., Wakunuma, K. & Stahl, B. (2017). Civil Society Organisations in Research: A Literature-Based Typology. Voluntas 28, 1988–2010

Reimsbach, D., & Braam, G. (2023). Creating social and environmental value through integrated thinking: International evidence. Business Strategy and the Environment, 32(1), 304–320.

Romanelli, M. (2018), "Museums creating value and developing intellectual capital by technology: From virtual environments to Big Data", Meditari Accountancy Research, Vol. 26 No. 3, pp. 483-498.

Rotheroe, N., & Richards, A. (2007). Social return on investment and social enterprise: transparent accountability for sustainable development. Social Enterprise Journal, 3(1), 31-48.

Russo, A., Watkins, J., Kelly, L., & Chan, S. (2007). Social media and cultural interactive experiences in museums. Nordisk Museologi, (1), 19-19.

Sattari, M. T., Rezazadeh-Joudi, A., & Kusiak, A. (2017). Assessment of different methods for estimation of missing data in precipitation studies. Hydrology Research, 48(4), 1032-1044.

Saul, J. (2007). Benchmarking Basics. In Benchmarking for nonprofits: How to measure, manage, and improve performance (pp. 1–12). Saint Paul, MN: Fieldstone Alliance.

Schiuma, G., Carlucci, D., & Sole, F. (2012). Applying a systems thinking framework to assess knowledge assets dynamics for business performance improvement. Expert Systems with applications, 39(9), 8044-8050.

Schiuma, G., Gavrilova, T., & Carlucci, D. (2022). Guest editorial: Knowledge visualisation for strategic decision-making in the digital age. Management Decision, 60(4), 885-892.

Scott, C. (2007). Measuring social value. In Museum Management and Marketing (pp. 181–194).

Shahrasbi, N., & Paré, G. (2015). Inside the "Black Box": Investigating the Link between Organizational Readiness and IT Implementation Success.

Sheng, C.-W., & Chen, M.-C. (2012). A study of experience expectations of museum visitors. Tourism Management, 33(1), 53–60.

Simon, H. A. (1947). Administrative behavior: A study of the decision-making process in administrative organization (1st ed.). New York: Macmillan.

Social Value UK. (2023, April 5). What is social value? Social Value Definition and meanings. Social Value UK. Retrieved from: https://socialvalueuk.org/what-is-social-value/

Solima, L. (2010). Social Network: verso un nuovo paradigma per la valorizzazione della domanda culturale. Sinergie, 82, 47–74.

Su, Y., & Teng, W. (2018). Contemplating museums' service failure: Extracting the service quality dimensions of museums from negative on-line reviews. Tourism Management, 69, 214–222.

Tan, K. H., & Platts, K. (2003). Linking objectives to actions: A decision support approach based on cause–effect linkages. Decision sciences, 34(3), 569-593.

Taormina, F., & Baraldi, S. B. (2022). Museums and digital technology: a literature review on organizational issues. European Planning Studies, 30(9), 1676-1694.

Tenneson, C., & Brocklehurst, G. (2018). Digital business KPIs: Defining and measuring success for tech CEOs (Online). Accessed 17/11/2023. https://www.gartner.com/doc/3891236/digital-business-kpis-defining-measuring

Troise, C. (2022). Exploring knowledge visualization in the digital age: an analysis of benefits and risks. Management Decision, 60(4), 1116-1131.

Tsai, P.-H., & Lin, C.-T. (2018). How Should National Museums Create Competitive Advantage Following Changes in the Global Economic Environment? Sustainability, 10(10), 3749.

UNESCO (2003) Convention for the safeguarding of the intangible Cultural Heritage. Paris: UNESCO. Retrieved from: https://ich.unesco.org/doc/src/01852-EN.pdf

UNESCO. (2015). Proposal for a non-binding standard-setting instrument on the protection and promotion of various aspects of the role of museums and collections. Paris, France: UNESCO. Retrieved from: https://unesdoc.unesco.org/ark:/48223/pf0000233892

UNESCO. (2021). Resolutions: Recommendation on Open Science (Annex VI, pp. 137-149). Paris, France: UNESCO. Retrieved from: https://unesdoc.unesco.org/ark:/48223/pf0000380399

Value Reporting Foundation. (2022). Transition to integrated thinking: A guide to getting started. Retrieved from: https://www.integratedreporting.org/wp-content/uploads/2022/07/VRF-ITP-GettingStartedGuide.pdf

Van der Voort, H., Klievink, A. J., Arnaboldi, M., & Meijer, A. J. (2018). Rationality and politics of algorithms. Will the promise of big data survive the dynamics of public decision making? Government Information Quarterly, 36.

Van Looy, A., & Shafagatova, A. (2016). Business process performance measurement: a structured literature review of indicators, measures and metrics. SpringerPlus, 5(1), 1-24.

Vassilakis, C., Antoniou, A., Lepouras, G., Poulopoulos, V., Wallace, M., Bampatzia, S., & Bourlakos, I. (2017). Stimulation of reflection and discussion in museum visits through the use of social media. Social Network Analysis and Mining, 7, 1-12.

Vassiliadis, C., & Belenioti, Z. C. (2017). Museums & cultural heritage via social media: an integrated literature review. Tourismos, 12(3), 97-132.

Veron, E. and Levasseur, M. (1983) Ethnographie de l'Exposition. Bibliothque publique d'Information, Centre Georges Pompidou, Paris.

Vicente, E., Camarero, C., & Garrido, M. J. (2012). Insights into Innovation in European Museums: The impact of cultural policy and museum characteristics. Public Management Review, 14(5), 649-679.

Viganò, F., & Lombardo, G. (2018). Misurare l'impatto sociale generato dai musei: L'applicazione della metodologia del Ritorno Sociale sull'investimento (SROI). In Ambienti digitali per l'educazione all'arte e al patrimonio (pp. 332-350). FrancoAngeli.

Vurro & Romito (2019). La valutazione economica degli impatti sociali della Lega del Filo d'Oro: l'estensione dei confini dell'analisi SROI. Milano: VITA.

Walters, B. A., & Bhuian, S. N. (2004). Complexity absorption and performance: A structural analysis of acute-care hospitals. Journal of Management, 30(1), 97-121.

Wang, Y. R., & Madnick, S. E. (1989, February). The Inter-Database Instance Identification Problem in Integrating Autonomous Systems. In ICDE (pp. 46-55).

Waters, R. D., & Jamal, J. Y. (2011). Tweet, tweet, tweet: A content analysis of nonprofit organizations' Twitter updates. Public relations review, 37(3), 321-324.

Weber, A., & Thomas, R. (2005). Key performance indicators. Measuring and Managing the Maintenance Function, Ivara Corporation, Burlington.

Wecker, A., Kuflik, T. and Stock, O. (2015) 'AMuse – an initial plan to associate museum visits to outdoor cultural heritage activities', Proceedings of PATCH 2015, the 8th Workshop on Personal Access to Cultural Heritage.

Welsh, P. (2005). Re-configuring museums. Museum Management and Curatorship, 20, 103–130.

Wexler, S., Shaffer, J., & Cotgreave, A. (2017). The big book of dashboards: Visualizing your data using real-world business scenarios. Hoboken: Wiley.

Whelan, G. (2015). Understanding the social value and well-being benefits created by museums: A case for social return on investment methodology. Arts & Health, 7(3), 216–230.

Wholey, J. S. (1999). Performance-Based Management: Responding to the Challenges. Public Productivity & Management Review, 22(3), 288–307.

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3, 160018

Williams, D., (1997) How The Arts Measure Up: Australian research into the Social Impact of the Art, Stroud, Comedia.

Yang, Z., Xue, F., & Lu, W. (2021). Handling missing data for construction waste management: Machine learning based on aggregated waste generation behaviors. Resources, Conservation and Recycling, 175, 105809.

Ying, F., Tookey, J., & Seadon, J. (2018). Measuring the invisible: A key performance indicator for managing construction logistics performance. Benchmarking, 25(6), 1921-1934.

Ziesche, S. (2023). Open data for AI: what now? Unesco. Retrieved from https://unesdoc.unesco.org/ark:/48223/pf0000385841.

Zweig, K. A., Wenzelburger, G., & Krafft, T. D. (2018). On chances and risks of security related algorithmic decision making systems. European Journal for Security Research, 3, 181-203.

# Appendix A

## A.1. Question mapping

| Question number | 2018-2019 | 2019-2020 | 2020-2021 | 2021-2022 |
|---|---|---|---|---|
| Q1 | Consent 1 | Identification | Identification | Identification |
| Q2 | Consent 2 | Denomination | Denomination | Institution (Ente) |
| Q3 | Identification | Institution (Ente) | Institution (Ente) | Denomination |
| Q4 | Denomination | Number of visitors | Digital technologies invested in | Digital technologies invested in |
| Q5 | Region | Jointly managed activities | Priority activity to invest in | Priority activity to invest in |
| Q6 | Municipality | E-mail | Online services | Other revenue-generating services |
| Q7 | Type of institution | Digital innovation plan | Online tours | Online services |
| Q8 | Incumbent (Soggetto titolare) | Digital technologies invested in | Online workshops | Images for research, reproduction or commercial purposes |
| Q9 | Average number of visitors | % investiment in digital | Advanced training online courses | Online tours |
| Q10 | Number of workers | Priority activity to invest in | Online laboratories | Online workshops |
| Q11 | Managed activities and booking modes | Dedicated website | Videogames | Advanced training online courses |
| Q12 | % displayed collection | App | Podcast | Online laboratories |
| Q13 | Digital innovation plan | App features | Images for research, reproduction or commercial purposes | Videogames |
| Q14 | Importance of activities | Account social | Public response to activities offered for free | Podcast |
| Q15 | % investiment in digital | Social network monitoring | Satisfaction of free activities | Public response to activities offered for free |
| Q16 | Computer-based management system | Review platforms | Public response to paid activities | Public response to paid activities |
| Q17 | Computer-based management system activities | Reviews monitoring | Satisfaction of paid activities | Why are you satisfied |

| Question number | 2018-2019 | 2019-2020 | 2020-2021 | 2021-2022 |
|---|---|---|---|---|
| Q18 | Dedicated website | Data collection on visitors | Other revenue-generating services | Why are you unsatisfied |
| Q19 | App | Which data are collected | Digitalization of the collection | Digitalization of the collection |
| Q20 | App features | Marketing and communication activities | Digitalized collection activities | Reasons for the digitalization of the collection |
| Q21 | Account social | Contracts with tour operators | Data collection on visitors | Digitalized collection activities |
| Q22 | Social network monitoring | Ticket office | Which data are collected | NFT |
| Q23 | Reviews monitoring | Ticket office revenues | Data collection modalities | Data collection on visitors |
| Q24 | Data collection on visitors | Type of ticket office | Use of collected data | Use of collected data |
| Q25 | Newsletter | *Skip-the-line* tickets | % visitors consent for marketing actions | Software CRM |
| Q26 | Technologies available | %ticket revenue divided between channels | Software CRM | Marketing and communication activities |
| Q27 | Ticket office | Other revenue-generating services | Marketing and communication activities | Number of visitors |
| Q28 | Ticket office revenues | Network card | Number of visitors | Ticket office |
| Q29 | Type of ticket office | Methods of visitor access control | Ticket office | Ticket office revenues |
| Q30 | Main channel for online ticket sales | Computer-based management system activities | Ticket office revenues | % revenues split between sources |
| Q31 | *Skip-the-line* tickets | Wi-Fi | % revenues split between sources | Type of ticket office |
| Q32 | Social network to website link | Technologies available | Type of ticket office | *Skip-the-line* tickets |
| Q33 | % ticket revenue divided between channels | % catalogued collection | *Skip-the-line* tickets | % ticket revenue divided between channels |
| Q34 | Ticket reservation | Digitalization of the collection | % ticket revenue divided between channels | Channel manager |
| Q35 | Reservation channels | Publishing of digitalized collection | Channel manager | Methods of visitor access control |
| Q36 | Network card | Workers with digital competencies | Methods of visitor access control | Investments in security |

| Question number | 2018-2019 | 2019-2020 | 2020-2021 | 2021-2022 |
|---|---|---|---|---|
| Q37 | Methods of visitor access control | Which type of workers | Investments in security | Future investments in security |
| Q38 | Catalog type | Region | Technologies available | Technologies available |
| Q39 | Cataloguing frequency | Province (Provincia) | Digital innovation plan | Digital innovation plan |
| Q40 | Digitalization of the collection | Type of institution | *2020 Piano Opportunità* revision | Workers with digital competencies |
| Q41 | Metadata | Incumbent (Soggetto titolare) | Workers with digital competencies | Digital innovation priority |
| Q42 | Publishing of digitalized collection | | Which type of workers | Strengthening of Research |
| Q43 | % published/digitalized collection | | Region | Strengthening of Care and asset management |
| Q44 | Workers with digital competencies | | Type of institution | Strengthening of Education and involvement |
| Q45 | Which type of workers | | Incumbent (Soggetto titolare) | Strengthening of Communication and promotion |
| Q46 | | | Consent | Strengthening of Conservation and security |
| Q47 | | | | Strengthening of Governance |
| Q48 | | | | Region |
| Q49 | | | | Type of institution |
| Q50 | | | | Incumbent (Soggetto titolare) |
| Q51 | | | | Consent |

Table 81 - Mapping of the questions of the proprietary dataset

## A.2. Charts selected by View

| View | KPI | Visual object/ Chart | Scale |
|---|---|---|---|
| Descriptive View | Number of museums per Region | Basic choropleth (FT) | Sequential |
| Descriptive View | Number of museums per Region | Bar chart (Abela, 2009), Ordered bar (FT) | None |
| Descriptive View | Number of museums per Typology | Bar chart (Abela, 2009), Ordered bar (FT) | None |
| Descriptive View | Number of museums | Card (Power Bi) | None |
| Descriptive View | Number of provinces | Card (Power Bi) | None |
| Descriptive View | Number of municipalities | Card (Power Bi) | None |
| Descriptive View - Personnel | Average Personnel by Region | Basic choropleth (FT) | Sequential |
| Descriptive View - Personnel | Average Personnel by Province | Basic choropleth (FT) | Sequential |
| Descriptive View - Personnel | Selection Average | Card (Power Bi) | Sequential |
| Descriptive View - Personnel | Italian Average | Card (Power Bi) | Sequential |
| Descriptive View - Personnel | Selection and Italian Average | Bar chart (Abela, 2009), Ordered bar (FT) | None |
| Descriptive View - Personnel | Average Personnel by Typology | Bar chart (Abela, 2009), Ordered bar (FT) with constant line (Power Bi) | None |
| Descriptive View - Revenues from tickets | Average Revenues from Tickets by Region | Basic choropleth (FT) | Sequential |
| Descriptive View - Revenues from tickets | Average Revenues from Tickets by Province | Basic choropleth (FT) | Sequential |
| Descriptive View - Revenues from tickets | Number of museums by Range of Revenues from Tickets | Bar chart (Abela, 2009), Ordered bar (FT) | None |
| Descriptive View - Revenues from tickets | Average Revenues from Tickets by Typology | Bar chart (Abela, 2009), Ordered bar (FT) | None |

| View | KPI | Visual object/ Chart | Scale |
|---|---|---|---|
| Descriptive View - Visitors | Average Visitors by Region | Basic choropleth (FT) | Sequential |
| Descriptive View - Visitors | Average Visitors by Province | Basic choropleth (FT) | Sequential |
| Descriptive View - Visitors | Number of museums by Range of Visitors | Bar chart (Abela, 2009), Ordered bar (FT) | None |
| Descriptive View - Visitors | Sum of Visitors by Typology | Bar chart (Abela, 2009), Ordered bar (FT) | None |
| Digital View - Online | Online KPI per Region | Basic choropleth (FT) | Sequential |
| Digital View - Online | Online KPI per Year | Line Chart (Abela, 2009), Line (FT) | None |
| Digital View - Online | Italian Average | Card (Power Bi) | Sequential |
| Digital View - Online | Regional Average | Card (Power Bi) | Sequential |
| Digital View - Online | Typology Average | Card (Power Bi) | Sequential |
| Digital View - Online | Specific museum | Card (Power Bi) | Sequential |
| Digital View – On-site | On-site KPI per Region | Basic choropleth (FT) | Sequential |
| Digital View – On-site | On-site KPI per Year | Line Chart (Abela, 2009), Line (FT) | None |
| Digital View – On-site | Italian Average | Card (Power Bi) | Sequential |
| Digital View – On-site | Regional Average | Card (Power Bi) | Sequential |
| Digital View – On-site | Typology Average | Card (Power Bi) | Sequential |
| Digital View – On-site | Specific museum | Card (Power Bi) | Sequential |
| Digital View – Organizational Readiness | Organizational Readiness KPI per Region | Basic choropleth (FT) | Sequential |
| Digital View – Organizational Readiness | Organizational Readiness KPI per Year | Line Chart (Abela, 2009), Line (FT) | None |
| Digital View – Organizational Readiness | Italian Average | Card (Power Bi) | Sequential |

| View | KPI | Visual object/ Chart | Scale |
|---|---|---|---|
| Digital View – Organizational Readiness | Regional Average | Card (Power Bi) | Sequential |
| Digital View – Organizational Readiness | Typology Average | Card (Power Bi) | Sequential |
| Digital View – Organizational Readiness | Specific museum | Card (Power Bi) | Sequential |
| Digital View – Evolution Average | Average KPI per Region | Basic choropleth (FT) | Sequential |
| Digital View – Evolution Average | Average KPI per Year (composed) | Stacked Column Chart (Abela, 2009) | Categorical |
| Digital View – Evolution Average | Italian Average | Card (Power Bi) | Sequential |
| Digital View – Evolution Average | Regional Average | Card (Power Bi) | Sequential |
| Digital View – Evolution Average | Typology Average | Card (Power Bi) | Sequential |
| Digital View – Evolution Average | Specific museum | Card (Power Bi) | Sequential |
| Digital View – Positioning | Average KPI per Region | Basic choropleth (FT) | Sequential |
| Digital View – Positioning | Online KPI, On-site KPI, Organizational Readiness KPI | Bubble Chart[21] (Abela, 2009), Bubble (FT) | Diverging |
| Digital View – Positioning | Quantile of Online KPI | Gauge (Power Bi) | None |
| Digital View – Positioning | Quantile of On-site KPI | Gauge (Power Bi) | None |
| Digital View – Positioning | Quantile of Organizational Readiness KPI | Gauge (Power Bi) | None |
| Digital View – Positioning | Ranking of Selected museum | Card (Power Bi) | Sequential |

Table 82 - Charts selected by View

[21] The bubble chart output is confusing because of the quantity of data points. Instead of using bubble to encode size, a diverging color scale is implemented.

# Appendix B

## B.1. List of functions and features used

This is the list of MS Excel functions and features used to assist the harmonization process:

- XLOOKUP: Using XLOOKUP, it is possible to search within one column for a specific term and retrieve a result from another column located in the same row. XLOOKUP is the evolution of one of the most used functions in MS Excel, VLOOKUP.

| Audioguide | AR | VR | QR/ Beacon | ChatBot | Answer |
|---|---|---|---|---|---|
| 0 | 1 | 0 | 0 | 0 | AR |
| 0 | 0 | 0 | 1 | 0 | QR/ Beacon |

Table 83 - Example of XLOOKUP application

  In this example, the XLOOKUP function searches for a 1 in every row and outputs the headers. In this way, data can be transformed from a binary format (ones and zeroes) to a categorical format (data that can be grouped through a common element).
- IF: The IF function outputs a specified value if a set condition is met. For example, the result of Table 83 was achieved using an IF function (IF(A1=0; "No"; "Yes")).
- CONCAT: The CONCAT function concatenates strings.
- Pivot table*:* this functionality serves both as an analytical and reporting tool. It generates a summary table from a bigger table, in which the elements of the bigger table are grouped by categories. This tool is intuitive and very customizable, and it is used extensively in dataset exploration and other activities related to dataset analysis.

| Etichette di riga | Count of Data collection on visitors |
|---|---|
| No | 218 |
| Yes, by paper | 612 |
| Yes, in digital | 273 |
| Both | 167 |
| **Totale complessivo** | **1270** |

Figure 56 - Example of pivot table output

- Filter*:* this functionality is applied to a table. All the records that meet a certain condition, such as *being equal to* a number or text, or *higher/lower than* a number, are kept, while the other records are excluded.

The Python libraries used are:

- Pandas: one of the most important Python libraries, used for data manipulation. In the library, among many others, the function read_csv is present, which lets the user upload a database into the workspace.
- Scikit-Learn (sklearn): it is a machine learning (ML) based library. It features various classification, regression and clustering algorithms. It is composed of many sub-libraries which are focused on different areas of ML. The sub-libraries are used for many kinds of data analysis and manipulation. In fact, they are used for clustering, changing the format of data from text to TF-IDF, and computing the cosine similarity between two matrices.
- Fuzzywuzzy: it is used to deploy record matching techniques. The library is used in Leftovers matching (p.80).

## B.2. Hierarchical clustering

Clustering is the process of combining data points into different groups, called clusters. The objective of clustering is to populate the clusters in a way that the members inside are similar to other members of the same cluster and are different from members of other clusters. This similarity is computed based on a determined way of computing the difference between data points, which is the distance. Clustering is an unsupervised approach, meaning that data is unlabeled, and the goal is to find a pattern in the data. The clustering algorithm lacks the capability to assess the quality of its results; only human users possess the ability to determine the goodness of the clustering results. Even for humans, the goodness of the results is solely determined by the objective. There are several methods used to apply clustering to a set of data:

- Density-based methods: the space in which the variables reside is divided into high-density spaces (where many points are close to each other) and low-density spaces (where only a few points are). Clusters are created with members belonging to the same high-density region of space. Data points that are located in low-density areas are typically considered noise or outliers (Kriegel et al., 2011) and this means that not every point belongs to a cluster.
- Hierarchical methods: the objective of these methods is to build a hierarchy of clusters, organized visually with a dendrogram (tree diagram). The solutions of this type of clustering are multiple as the definitive number of clusters is chosen by the user and determined by where the tree is cut. There are two categories of hierarchical methods:
    - Agglomerative (bottom-up): each data point belongs to its own cluster. Then, clusters are merged by moving down the hierarchy.

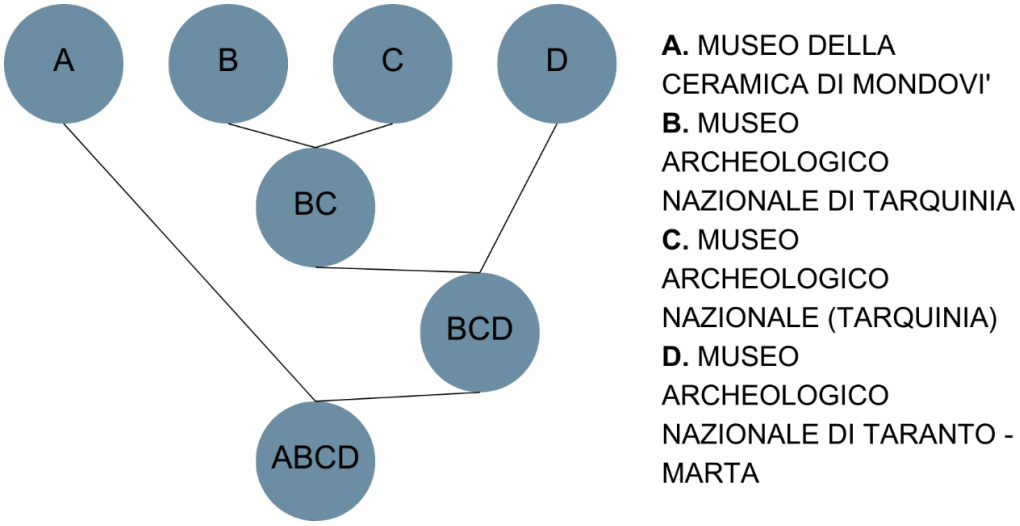


Figure 57 - Example of agglomerative clustering

  - Divisive (top-down): each data point belongs to the same cluster. Then, clusters are divided by moving down the hierarchy.
- Partitioning methods: the data points are divided into a predetermined number of clusters. The difference with the other methods is that the number of clusters is known before running the algorithm.
- Grid-based methods: the data space is divided into a set of initial cells that are the same size and are mutually exclusive. When the support of a cell becomes high enough (high-level), the cell is split into two new cells (Park & Lee, 2004).

# List of Tables

# List of Figures