Benefits from implementing a data labelling tool

International Master in Management of IT

Information System Science

Master's thesis

Author:

Väinö Saarinen

Supervisor:

Dr. Timo Leino

6.6.2024

Turku

The originality of this thesis has been checked in accordance with the University of Turku quality assurance system using the Turnitin Originality Check service.

Master's thesis

Subject: Information System Science
Author: Väinö Saarinen
Title: Benefits from implementing a data labelling tool
Supervisor: Dr. Timo Leino
Number of pages: 71 pages + 3 appendices pages
Date: 6.6.2024

The amount of data in the world is constantly increasing, making data management more complex and demanding. Effective data utilization is crucial for in-depth analysis, logical reasoning, and decision-making processes. Data labelling is an essential part of this process, but it has traditionally been labour-intensive and resource-consuming. To manage always scarce resources more efficiently, companies are turning to data labelling tools to automate the process, enhance data management, and extract more value from their data.

This thesis aims to reason the benefits and risks associated with implementing a data labelling tool, specifically Microsoft Purview. The study employs a benefit measurement model and includes a pilot project conducted in a case company. Additionally, interviews with company professionals were conducted to provide further validation and professional insights into the benefits of data labelling.

The findings reveal several notable benefits of data labelling and data labelling tools. Firstly, labelling tools improve the quality and understanding of the data in hand, enhancing its utility. Secondly, automated labelling tools significantly accelerate the labelling process, reducing resource consumption compared to manual methods. Thirdly, data labelling offers broad advantages in data management, data governance, data loss prevention, data security and compliance management and data lifecycle management. Risks related to data labelling tool implementation includes accuracy of labelling, user adoption and engagement and beneficial resource allocation.

Key words: Data labelling, Benefit management, Data labelling tool

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

1.1 Background

We live in a world surrounded by data. Every day, we encounter various types of data from many sources, such as measurements and observations. This data can illustrate the features of biological entities, outline the characteristics of environmental events, encapsulate scientific research findings, or track the operation of industrial equipment. Significantly, this data lays the groundwork for in-depth analysis, logical reasoning, decision-making processes, and, ultimately, the comprehension of various entities and phenomena. Among the numerous analytical tasks, one crucial activity is the categorization, clustering or labelling of data into distinct groups or segments. Objects grouped together in this manner are expected to exhibit comparable characteristics according to specific standards. (Seetha et al., 2018, 2)

According to Jürgen et al. (2021), data labelling involves transforming an unlabelled data instance into a labelled one by assigning it a specific category or descriptor. IBM (ibm.com) explains that this process requires initially identifying raw data, such as images, text files, or videos. Subsequently, a relevant label is added to the data's metadata, which clarifies the data's context and enhances its usability in machine learning models. This step is crucial for training accurate predictive models, as it provides the necessary context for algorithmic interpretation.

The rapid expansion of available data necessitates efficient clustering or labelling to ensure its effective use. Cao and Liang (2011) highlight that despite its necessity, the process of labelling data has remained time consuming. Furthermore, Woodward et al. (2020) discuss how machine learning, as a type of artificial intelligence, has evolved quickly, resulting in applications that can accurately recognize speech and images. However, they point out that there are still many types of data that these techniques haven't fully investigated. They emphasize that labelling is a crucial step in preparing data, particularly when dealing with real-time data collected from single or multiple sensors. They note that labelling this kind of data as it is being collected is currently a cumbersome task, hindered by the few tools available and their lacklustre efficiency. This can negatively affect the performance of machine learning models. Fredriksson et al. (2020) continue by arguing, that according to current research, data preparation and

labelling account for over 80% of engineering activities in machine learning projects, and by 2024, the third-party data labelling industry is predicted to nearly triple.

In today's business landscape, IT and digital data have evolved from mere operational resources to critical components of corporate strategy (Dahlberg & Nokkala, 2015). The substantial growth in datasets used for machine learning tasks has emphasized the importance of addressing issues such as noisy labelling to ensure a robust learning process (Lee et al., 2022). The prevalence of such challenges is often due to industry datasets frequently being incomplete, where some or all instances lack labels, or where available labels are of low quality. These low-quality labels lead to errors or only partially accurate data entries. As a result, ensuring sufficient quality in labels is crucial for supervised machine learning, as the quality of training data directly affects the model's operational performance. (Fredriksson et al., 2020) This strategic emphasis on high-quality data management becomes even more complex and vital when considering the vast and varied environments in which modern data is stored and processed.

Organizations are dealing with increasingly larger amount of data in their operations. As stated by Ahmad et al. (2023), the modern enterprise is often ensnared in a complex web of data sprawl, where information is dispersed across varied environments including on-premises infrastructures, edge computing resources, and multiple cloud platforms. This spreading is not solely a matter of location but also extends to the types of data repositories employed, from traditional relational databases to modern no-SQL systems and unstructured file storage solutions. These repositories serve diverse functions, from day-to-day operational needs to complex analytical queries, further complicating the data governance landscape.

Other core theme in this thesis among data labelling is benefit management, and more specifically benefit management in IS/IT landscape. In the today's business landscape, the effective management of IT implementation and the skilful leverage of its capabilities are becoming paramount in enhancing business performance. This underscores the necessity for business managers to take an active role not only in the selection and prioritisation of IS/IT projects but also in ensuring the realization of their benefits. According to Love and Matthews (2019), businesses frequently spend money on digital technology to provide managers with rapid and high-quality information to improve decision-making, recognise performance trends, and reduce expenses. Nevertheless, this

approach overlooks the fact that managers can have biases and may not effectively use the data, regardless of its accuracy or reliability. Supervisors may even decide to completely ignore the provided information. Thus, it is imperative that organisations put users—those who will understand and interpret the data—at the centre of their digitization operations. Organizations need to critically assess how employees will use data in their decision-making processes and, at the same time, encourage a shift towards relying on formal analysis rather than intuitive judgments.

Benefit management is crucial, as noted by Ward and Daniel (2012, 61), who point out that the consequences of undervaluing IT investments go well beyond any one project. When a company doesn't know how to use IT to its full potential, it becomes less aware of the value it adds. This causes the company to make inconsistent and incorrect decisions about priorities and investments. As such, this deficiency makes it more difficult for the company to identify how IT could be optimally employed to boost performance or facilitate strategic innovations. Furthermore, it frequently leads to the exclusion of IT's potential contributions from conversations about business strategies, ignoring the risks associated with competitors' IT advancements as well as the opportunities presented by emerging IT solutions. This line of thought emphasizes why benefit management is a crucial topic, as it directly influences an organization's strategic alignment with IT and its competitive stance in the industry. (Ward & Daniel, 2012, 61)

Creating a justification for any investment requires evaluating both the benefits and costs. While calculating or estimating costs is typically straightforward, a significant shortcoming in numerous IS/IT investment proposals is the insufficient articulation or examination of the anticipated organizational benefits. (Ward & Daniel, 2012, 129) This gap highlights a critical area where organizations must improve to harness the full potential of IT investments.

Since the inception of IT/IS development projects, there have been continuous reports of IT/IS development project failures (Cule et al., 2000; Lyytinen et al., 1998). Krigsman (2007) discusses in his article how up to two-thirds of IT projects are somehow challenged, with half of those ultimately failing. A project is considered challenged if it does not fulfil all its targeted outcomes, which often include metrics such as target time, budget, and functionality. Based on this premise, it would seem logical to rely on financial metrics to ensure that a new IS/IT project will meet its targets or at least be completed.

However, factors other than financial outcomes should also be considered. It's crucial to appreciate the significant relationship between information technology and economic performance that has emerged from over a decade of research. Brynjolfsson and Hitt (2000) highlight firm-level studies which demonstrate that computers have played a pivotal role in driving economic growth, contributing significantly more than their relative share of capital stock or investment would suggest. The impact of information technology is not only profound but is also expected to increase in the coming years. This evolving understanding highlights the critical need for investment in IT/IS and underscores the importance of recognizing and learning about the benefits of IT/IS projects, guiding more effective strategies for their implementation and management to ensure technology remains a key driver of economic progress.

This thesis includes a pilot study conducted in collaboration with Sitowise, an infrastructure, building, and digital solution consulting company. The pilot utilized Microsoft Purview, a data labelling tool, and involved the author's participation in a pilot creation group alongside Sitowise's IT development team. The initial need for data labelling and the adoption of automated data labelling software at Sitowise arose from the company’s ambition to enhance its applications, increase efficiency, and increase its competitive advantages from software. The pilot study focused on data sensitivity labelling, reflecting Sitowise's emphasis on security considerations in the development of new technologies, such as AI-based chatbots utilizing the company's data. The pilot study is a crucial component of this thesis as it provides insight into data labelling software and offers an opportunity to explore the benefits of its implementation.

Artificial intelligence (AI) tools were employed strictly to enhance grammar and text flow in collaboration with the thesis supervisor and the associated company supervisor. All content, including sources and conclusions, was created, researched, and compiled by the author. The AI tools utilized were Grammarly and ChatGPT-4, strictly for correcting grammatical errors and improving text flow. Specific examples of prompts used with ChatGPT include, "Correct grammar errors," and "How would you make the text more formal/better flowing?" No text was directly used from the AI tools. The author made all final decisions regarding the most suitable phrasing for each occasion.

1.2 Research questions

This thesis aims to firstly discuss data labelling as a concept, including labelling methods and its position in companies' data management functionality. Secondly, it explores benefit management, examining the overall concept and presenting a benefit measurement model. Thirdly, this thesis includes a design science-based pilot study of using the Microsoft Purview tool, with the aim of establishing the various benefits of implementing a data labelling tool. The findings in this thesis are derived from observations and discussions related to the pilot. Additionally, validating interviews were conducted to further consolidate the results. The main research question is: **What are the potential benefits from implementing a data labelling tool?**

The main question is supported by two sub-questions:

What are the benefits of labelling?

What are the risks associated with the labelling tool?

The first sub-question aims to identify the benefits that labelling brings from using the labelling tool, but also in more general terms. The benefits of using the tool derive from the inherent advantages of labelling itself. By exploring these benefits, the aim is to underline the importance of labelling in organizations. In addition to benefits, the second sub-question examines the risks related to utilizing the labelling tool. Understanding these challenges and downsides helps in evaluating the tool's suitability for different organizations and objectives.

Two main themes of this thesis are explored in chapters 2 and 3. Chapter 2 delves into the concept of data labelling, examining its role within organizational data management functions and exploring various data labelling processes and methods. Following this, Chapter 3 shifts focus to benefit management, introducing a benefit measurement model. Chapter 4 outlines the methodology employed throughout the thesis. Chapter 5 explores the pilot study, its aims, the tool used, and how the pilot was practically executed. Chapter 6 consists of interviews and discussions related to the pilot, data labelling and benefit management. Conclusions are in chapter 7, where research questions are answered and information from academic literature, the pilot and interviews are combined. Chapter 8 consists of a summary of what was done in this thesis.

2 Data labelling

2.1 Data management

2.1.1 Data, information and knowledge

In everyday life and even in businesses terms related to data management are used vaguely. Data Management Association (DAMA) defines three important concepts of data management in their data management book of knowledge (DMBOK). Dahlberg and Nokkala (2015) state this book to be as the most "acknowledged cumulative endeavour in data management". Based on DAMA (2009, 2) Data are defined as a representation of facts as text, numbers, graphs, images, sound or videos. Information on the other hand, is defined as data with context. Data becomes significant only when placed in context. Added things are definition, format, timeframe, and relevance. By understanding the surrounding circumstances of data, it can be turned into meaningful information. (DAMA, 2009, 2) March and Smith (1995) supplement this view by stating information being data that has been processed into meaningful form and which provides real or perceived value for current or prospective decisions or actions. Knowledge is augmented by information. Understanding, consciousness, cognizance, situation recognition, and familiarity with its complexity are all components of knowledge. Knowledge is information in perspective, combined with other information and experience. It might also contain beliefs and presumptions regarding the causes. In brief, information turns into knowledge when people add patterns, trends, relationships and assumptions into information. (DAMA, 2009, 3). Figure 1 show evolution from data to knowledge.

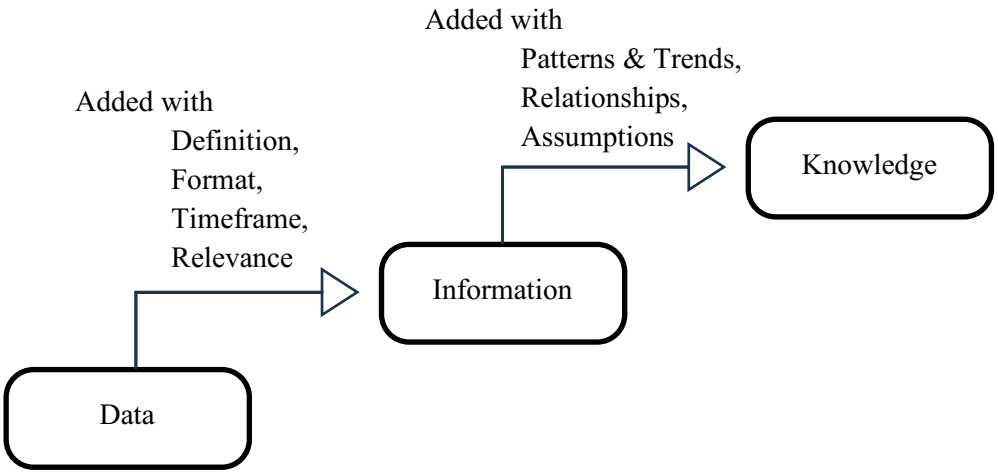

Figure 1. Adaption of Data, Information and knowledge -figure from DAMA (2009, 2).

Labelling data is one way of transforms data into information. As stated above, transforming data to information requires adding definition, format, timeframe or relevance to data. Labelling does just that. It adds metadata to data and data's usability increases.

2.1.2 Metadata

Furner (2020) discusses the definition of metadata in his article, noting that many sources begin with the observation that "metadata" is commonly understood as "data about data." According to him, some even consider "data about data" as a literal definition. Other definitions he mentions include "data that defines and describes other data." Pomerantz (2015, 26) argues that the definition "data about data" is not sufficient, as data is only potential information, raw and unprocessed, before anyone actually becomes informed by it. He suggests that the word "data" should be understood as "potential information" and describes metadata as "a potentially informative object that describes another potentially informative object."

Furthermore, Riley (2017, 1) provides a more detailed definition, stating that metadata is "the information we create, store, and share to describe things." Although all these descriptions aim to convey the same concept, when considering labelling as a method of adding meaning and increasing the usability of data, Riley's definition seems to be the most suitable. This is because labelling involves creating labels and grouping data into those labels, thereby enhancing data's usability by adding metadata.

2.1.3 Data labelling as a concept

Data labelling as a term is unestablished and is used in different environments and context. Term is used similarly as data classification and clustering. Especially "Data labelling" arise in many articles focusing on usage of large data sets in medical research where there are a lot of data in different and complex forms, and it is difficult to analyse it as it is. Diaz-Pinto et al., (2022) researched Interactive segmentation of 3D medical images, Hautz et al. (2021) developed system to limit diagnostic errors by data labelling in routine healthcare data and Chen et al. (2022) explored the automation of topic annotation in COVID-19 literature to improve the efficiency and accuracy of information retrieval in biomedical research.

When talking about less complex data such as numbers and words literature seems to prefer terms classification or clustering terms. Everitt et al. (1974) researched a cluster analysis and provided a comprehensive, non-mathematical introduction to cluster analysis, covering new and developing areas like classification likelihood and neural networks. Jain et al. (1999) on the other hand, provided an overview of pattern clustering methods from a statistical pattern recognition perspective, presenting a taxonomy and identifying cross-cutting themes and recent advances. Besides usage of different terms, the objective is the same: define categories and separate data to different groups and thereby facilitating a more streamlined and efficient analysis and application of the data.

Cambridge dictionary (dictionary.cambridge.org) defines classification as “the act or process of dividing things into groups according to their type”. According to Seetha et al., (2018, 2) a cluster refers to a group of data items that resemble each other but are distinct from items in different clusters. Enormous quantities of data are created and distributed across numerous sources. These definitions support the view of labelling being a more or less synonymous term to classification and clustering.

2.2 Data labelling process

2.2.1 Role of labelling in data management

The Data Management Association (DAMA, 2009, 2) states that managing information assets includes the management of data and metadata. Labelling is a key part of incorporating metadata into data, making it more usable.

Figure 2. “Data management functions” from DAMA (2009, 7), Meta-data Management bolded.

DAMA delineates various data management functions as illustrated in figure 2. Among these, data labelling predominantly falls under the meta-data management function, as labelling principally involves adding metadata to data. However, some might contend that labelling also intersects with data quality management because the quality of data significantly influences labelling practices. Additionally, labelling can be linked to broader data management functions, as evidenced in this thesis, which includes data sensitivity labels, pointing towards data security management function.

DAMA (2009, 7) defines the metadata management function as "integrating, controlling, and providing metadata," which is what data labelling is. It is important to recognize that, in practical scenarios, the functions of data governance often overlap, reflecting the interconnected nature of these activities in real-world applications.

2.2.2 Data collection

One primary requirement for data labelling is having enough data (Roh et al., 2021). This sub-chapter discusses the possibility of labelling data straight when collecting data or afterwards if collecting is done beforehand. A common area in research is the emphasis

that the more additional metadata is included when data is collected, the better the labelling can be done.

Woodward et al. (2020) emphasize the importance of including metadata when collecting raw data, underscoring that metadata serves as a foundational element for understanding and utilizing data effectively. According to DAMA (2009), metadata includes definitions of business-related data and plays a crucial role in clarifying the context of data. Effective management of metadata, therefore, directly contributes to the enhancement of information quality. Building on this basic metadata importance description, Woodward et al. (2020) highlight the necessity of real-time context in accurately labelling data, noting that without captured context, correct labelling might even be impossible. They further elaborate on practical approaches for data collection by presenting three specific methods to ensure thorough and contextually accurate data gathering:

Passive data sensing

Active data sensing

Hybrid data sensing

Passive data sensing uses devices like smartphones to automatically gather data without needing the person to do anything. This method is often used for keeping track of the weather or health. Active data sensing needs people to enter information themselves. For example, used for tracking their health or exercise. Hybrid data sensing merges the strengths of both passive and active methods, allowing for a richer dataset. This approach records data automatically while also enabling users to provide contextual details or corrections, enhancing the accuracy and relevance of the information gathered. It's particularly effective in environments where the background data alone isn't enough to draw comprehensive conclusions, requiring user insights to complete the picture. (Woodward et al., 2020)

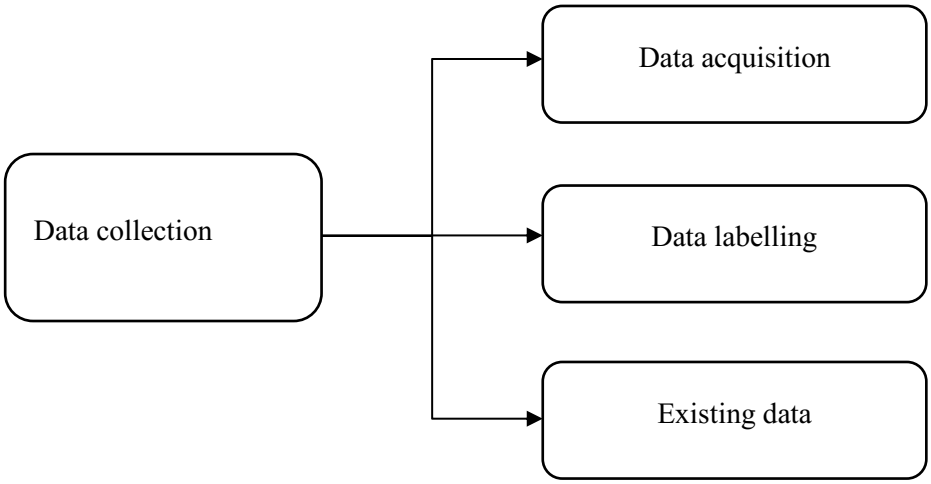

Figure 3. Methods for data collection, adapted from Roh et al. (2021).

In their research Roh et al. (2021) present "a high level research landscape of data collection for machine learning" -figure. Figure 1 shows two highest levels of the picture. It presents methods for data collection. Three main branches are *Data acquisition*, *Data labelling* and *Existing data*. Firstly, if the objective is to disseminate and discover new datasets, then employing data acquisition methods can help in identifying, expanding, or creating such datasets. Secondly, after these datasets are in hand, a variety of techniques for data labelling can be applied to classify the individual pieces of data. This path is further explained in figure 4 in chapter 2.2.4. Thirdly, rather than annotating brand-new datasets, it might be more beneficial to enhance the quality of existing data or to apply additional training to already trained models. While these three approaches are not inherently exclusive, they can be integrated. For instance, one could simultaneously explore and label more datasets while also enhancing the quality of current ones.

2.2.3 Automated versus manual labelling

Data labelling is a costly, labour-intensive, and time-consuming task, leading to large amounts of data remaining untapped during analysis via data mining (Godwin & Matthews, 2014). This view is backed up by Sun et al. (2017). Difficulties are increased with the amount of data that has to be handled in modern time. Desmond et al. (2021) promote AI or algorithmic labelling systems to assist labelling work. They state that especially human-AI collaboration can improve accuracy of labelling task.

Furthermore, the literature extensively discusses the trend towards increasing automation in data labelling processes. A widely adopted method, as discussed by Jürgen et al. (2021), involves carefully selecting a small subset of data instances for manual labelling and then using semi-supervised learning techniques to extend these labels to additional unlabelled instances. This approach splits the labelling task into two distinct parts: first, selecting a candidate instance from a pool of unlabelled data, and then, assigning the appropriate label information to that selected instance. This strategy not only streamlines the labelling process but also significantly enhances the efficiency of preparing data for machine learning applications, underlining the critical importance of data labelling in the field.

Desmond et al. (2021) studied assisting data labelling process via AI tool. AI was used as an assistant for human to label large data sets. Their research revealed that the accuracy of human labellers wasn't significantly impacted by changing the AI's predicted performance. Study participants didn’t over rely on predictions of labelling assistant, which indicates that accuracy of human labelling can be improved by comparatively weak AI or algorithmic based assistant. Noteworthy is the gained labelling speed especially in cases where data or labels were simple. Contrary more complex data needed more human attention to get it correctly labelled.

Challenges may arise when labelled data is noisy or undetectable. In response, Lee et al. (2022) highlights the introduction of various strategies aimed at enhancing the efficiency of learning from data with noisy labels. These strategies are broadly classified into two main approaches: sample-based and model-based. Within the sample-based framework, the focus is on identifying and filtering out inaccurately labelled data within the training set, followed by the development of a classifier trained on the remaining, accurately labelled data.

Desmond et al. (2021) state that having human in data labelling process ensures the most reliable outcome, even if they admit that algorithms and labelling paradigms have involved. They continue by noting that even if the most reliable and trustworthy outcomes come with human actor, humans do mistakes and time of human worker is costly. At least more expensive than self-working computer. Zhang et al. (2022) concur, noting that even when labelling tasks are supported by tools or software, human annotators are essential.

Their role is crucial for ensuring accurate labels and maintaining the quality of the classification algorithms.

Labelling is a crucial aspect of supervised machine learning, essential for effectively utilizing data in various applications. However, in industrial settings, data often arrives unlabelled, complicating its use in machine learning projects. Fredriksson et al. (2020) emphasize the challenges this presents, particularly when data lacks predefined categories that could guide its analysis. Complementing this perspective, Godwin and Mathews (2014) discuss how data collected via different mining methods frequently lacks specific labels. Typically, only basic notes or minimal information accompany the data, detailing the condition of an item when it was removed, repaired, replaced, or serviced. They argue that this reliance on unlabelled data not only slows down operational processes but also leads to costly errors, especially in precision-dependent fields such as maintenance engineering. According to Jürgen et al. (2021) data labelling is an essential prerequisite for conducting supervised machine learning, fundamentally transforming a data instance from unlabelled to labelled. This process assigns specific labels to data, enabling its use as training data for model construction or testing data for evaluating existing models. However, manually labelling a large dataset to support supervised learning is both costly and time-consuming. Recognizing these challenges, significant efforts have been made to accelerate the data labelling process while reducing human involvement.

2.2.4 Data labelling methods

Roh et al. (2021) investigated data collection for machine learning applications and identified data collection and preparation as a significant bottleneck. They reinforce the perspective shared by Desmond et al. (2021), emphasizing that data preparation is both time- and resource-intensive.

In their study, Seetha et al. (2018) explains that organizing data into groups can be done in two ways: either supervised or unsupervised. This distinction is based on whether new pieces of data are assigned to predefined groups (supervised) or formed into new groups without pre-set categories (unsupervised). Woodward et al. (2020) align with this by discussing the process of labelling data. They state that labelling can be done either automatically or by hand. They highlight that labelling process becomes more complex when dealing with individuals performing physical activities. It's important to carefully consider the specifics of each activity when designing user experiences and interfaces, as

well as when choosing and using data sources and applications. Based on these notes, when organizing and labelling data, it's crucial to select the right method and to consider the context in which the data is being used.

Upon successfully collecting sufficient data, the subsequent stage is the annotation of individual data points. For illustrative purposes, consider a dataset comprised of images depicting industrial components used in a smart factory context. In such scenarios, personnel may commence the task of marking the presence of any defects within these components. Often, data collection and annotation occur simultaneously. For example, when information is extracted from the internet to form a knowledge base, each piece of information is typically presumed to be accurate, thereby receiving an implicit 'true' label. In the scholarly discussion surrounding data annotation, it is beneficial to distinguish it from data collection, as the methodologies employed in each case can vary significantly. (Roh et al., 2021)

Roh et al. (2021) provide three point category which they believe to provide sufficient understanding of data labelling landscape. Categories are *Utilization of pre-existing labels, Crowd-based techniques* and using *Weak labels.*

> Utilization of Pre-existing Labels: One foundational approach to data labelling capitalizes on the use of already available labels. This notion is central to the extensive literature on semi-supervised learning, which focuses on utilizing existing labels to predict the labels of unlabelled data.
>
> Crowd-based: This category includes methods that leverage the collective input of many individuals. A basic form of this approach involves numerous participants' straightforward annotation of data points. A more nuanced method incorporates active learning techniques to refine the selection of questions posed to annotators. Recent advancements in crowdsourcing methodologies have introduced novel techniques to increase participants' labelling efficiency and accuracy.
>
> Weak Labels: When the constant generation of precise labels is economically untenable, an alternate strategy is to produce a large volume of less accurate or weak labels to counterbalance their lower precision. This approach has gained

increased acceptance, particularly in fields with a need for labelled data for new applications.

Friedriksson et al. (2020) assert that, despite the presence of well-established labelling techniques such as crowdsourcing, active learning, and semi-supervised learning, these methods fail to deliver accurate and reliable labels for all industrial machine learning applications. Consequently, the industry continues to depend significantly on manual data annotation and labelling. Roh et al. (2021) attempts to explain possible ways to label data either by hand or more automatically as illustrated in figure W.W.

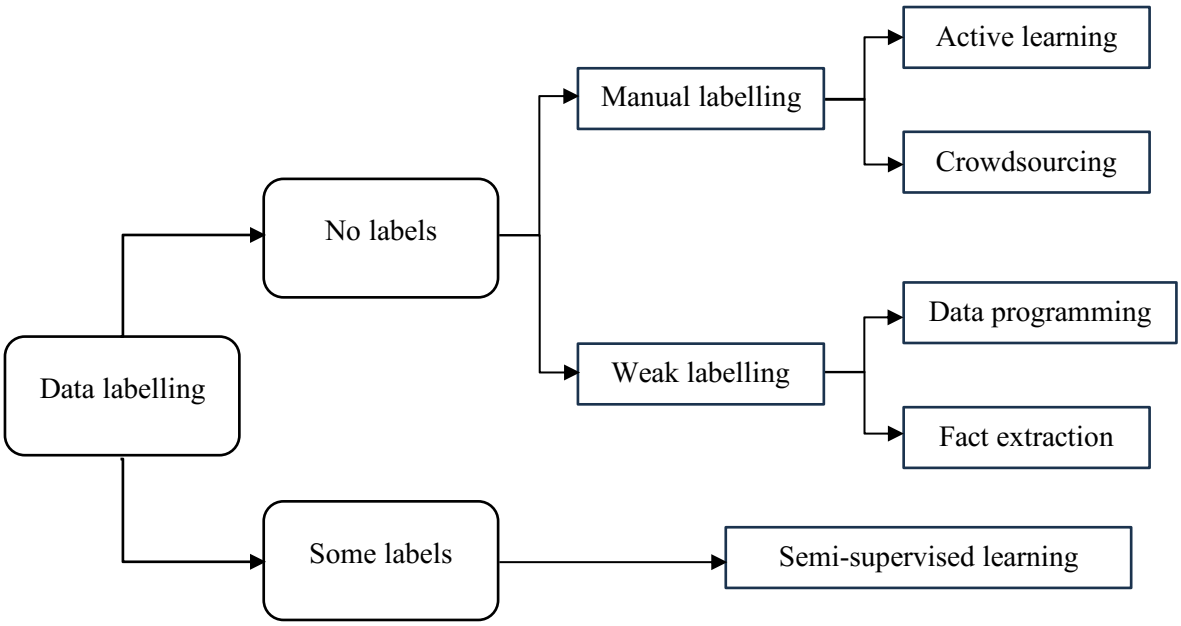

Figure 4. Data labelling methods in different situations, adapted from Roh et al. (2021).

In their study, Roh et al. (2021) discussed how collected data can be labelled with different labelling approaches. Figure 4 illustrates their idea from a data labelling viewpoint. The first note is to know whether there are already created labels or not. If there are already at least some labels identified or created, according to them, labelling is done in a semi-supervised manner. This means that existing labels are used to predict other needed labels. In semi-supervised learning, already labelled and yet unlabelled data make predictions to predict needed labels and in which new data belongs. Semi-supervised learning falls into utilization of pre-existing labels -category discussed above.

Not having any existing labels makes labelling more complicated. Roh et al (2021) provide two ways to do labelling in this situation: manual and weak labelling. Manual labelling means that the human operator creates labels. In active learning, labelling experts choose the most “interesting” unlabelled data examples and create labels from there. The assumption is that human experts are very accurate in creating labels as those will be the baseline for machine learning algorithms to label data. In the crowdsourcing method, there is a larger pool of human operators making labels. The idea is to have more opinions to create less bias and more accurate labels. In this method, the assumption is that humans make mistakes, and the crowd will fix individual mistakes. It is noteworthy that manual labelling always requires humans to be in action in both active and crowdsourcing labelling duties. Weak labelling is based on situations where there are large amounts of data, manual labelling is infeasible and reasonable high accuracy of labels is sufficient. The main aim is to make weak labelling as automated as possible. Data programming deals with weak labelling by creating multiple weak labels based on the data set and then combining them into a sufficient number of labels. Fact extraction works by finding critical information from data and labelling it based on that critical information.

3 Benefit management

3.1 Benefit management as a concept

Ward and Daniel (8, 2012) define benefit management in scope of information technology (IT) or information system (IS) as the process of organizing and managing such that the potential benefits arising from the use of IS/IT are actually realized. Even if benefits in IS/IT projects are widely researched Aubry et al. (2021) state performing and realizing these benefits in projects are in many cases still difficult and a key challenge for project management.

The literature on benefit management dates back to the 1980s, primarily focused on IT/IS development from its inception. This focus has persisted because presenting the benefits of developments in IT/IS beforehand has proven challenging. Although many studies link benefit management to measurable outcomes, the definition of "benefit" often remains vague. Various perspectives are used to evaluate benefits, including return on investment, value management, and performance metrics. However, no single perspective has emerged as dominant (Aubry et al., 2021). Love and Matthews (2019) state that value from technology assets can be measured using either quantitative or qualitative methodologies. They emphasize that organizations should understand value when making investment decisions, defining value in rough terms as "positive (benefits) and negative (dis-benefits) impacts, which can be categorized as being financial or non-financial." Aubry et al. (2021) continues by asserting that despite changes in methodologies and perspectives over the years, the primary aim of benefit management has remained consistent: to improve investment decisions in IT/IS projects.

The primary goal in preparing a justification for an IS/IT project (often called a “business case”) often involves securing funding. However, in many cases, justification encompasses more than just financial reasoning. Ward and Daniel (2012, 127) suggest that reasons other than financial ones should guide organizations in both deciding on the investment and managing the project effectively to its completion. Financial validation typically focuses on immediate, tangible performance gains and overlooks non-financial advantages. They acknowledge that a narrow scope focusing mostly on financial aspects was sufficient when benefits in IS/IT development projects were mainly about automating manual tasks. Love and Matthews (2019) support this view, reminding that IS/IT has

brought automation to many business fields, increasing competitiveness, productivity, and safety. However, Ward and Daniel (2012, 127) argue that, considering the crucial role of IS/IT in modern organizations, focusing solely on automation is no longer sufficient. This approach overlooks the broader advantages, particularly how technology supports an organization's growth and strategic objectives over the long term, creating new possibilities. This perspective is echoed by Manwani (2008, 12), who states that information technology has become a key factor in business success for many organizations. At the same time, Love and Matthews (2019) note that all IT investments should be demand-driven. They claim that organizations often face pressure, for example from clients, to implement new systems. Systems that do not have benefits clearly defined beforehand and whose demand is unclear can become financial ‘black holes’ for organizations, as such systems may not provide any benefits at all but still uses financial resources continuously.

Only securing funding for a project is not sufficient for the success of IS/IT projects. Sustained support throughout their entire lifespan is equally vital. Keil et al. (1998) identified the lack of top management's commitment as the most significant risk factor for the failure of IS/IT projects. Additionally, the difficulty in securing user commitment ranks as the second highest risk, indicating that belief in the project's benefits must extend beyond top management. Supporting this perspective, Krigsman (2007) underscores that the primary responsibility for the success of IT projects lies with project managers and top management. He points out that overly ambitious projects, shifting targets, and managerial challenges are frequent sources of failure in IT/IS projects.

The Standish Group International (2013, 5-7) published the CHAOS Manifesto report, which found that in small projects, the two most important factors for successful project execution are executive management support and user involvement. They evaluated project success based on three categories: time, cost, and features. A project is considered successful if it meets the goals for all three categories, if one or more category goals are not met, the project is classified as challenged and if the project is cancelled or never used, it is deemed a failure. The CHAOS Manifesto also states that having smaller projects with reduced and more manageable scopes significantly increases the success rate compared to very large and complicated projects. Furthermore, they emphasize that companies should focus on developing the most important features in software

development, as only 20% of all features are used frequently. This has the potential to maximize investment and improve overall end-user satisfaction.

IT/IS investments that are meant to change to the way business is performed and to deliver major IT/IS benefits requires funding outside of normal IT maintenance. Brynjolfsson and Hitt (2000) assert that fully leveraging IS/IT assets often requires an investment that is five times greater than the initial purchase cost of the technology, once hidden costs such as employee training, configuration, and setup are counted in.

Based on Ward and Daniel (2012, 132-133), it should be clear for people working in business setting that change is essential in the business environment. Love and Matthews (2019) asserts managers to generally know why changes in digital assets need to be made, the methods and how things should be done to realize needed benefits, is more often less known area. Ward and Daniel (2012, 132-133) further state that the anticipated benefits derived from accomplishing each objective need to be categorized according to the primary type of change required for their realization. While it might appear oversimplified to attribute each benefit to one of merely three origins, it's important to note that the majority of benefits emerge from these specific causes. These causes are foundation for gaining benefits in business setting. Without change there are neither evolvement nor benefits. Three main types are presented as follows:

1. The organization can do new thing or things in new ways

2. The organization can do better things it already does.

3. The organization can stop doing things

Love and Matthews (2019) agree with the idea that change is essential when seeking evolution or benefits. However, they emphasize the importance of continuity as change is a continuous process. While it may be sufficient to make temporary changes to test new things, achieving long-term benefits requires ongoing efforts. For example, it is usually not sufficient to train and educate employees only once to implement a change. Learning new ways of doing things on a daily basis requires continuous acknowledgement, support, and resources. Continuous involvement of managers and a willingness to change can also

reveal inefficient practices or bad habits and may bring additional benefits compared to the original goals.

In their article, Peppard et al. (2007) provide reasons and list the advantages of utilizing a benefit management approach when making decisions related to IT investments. In their research, they identified several benefits in different organizations. The benefit management approach resulted in clearer planning as the benefits and the methods to achieve them were better understood. This led to more accurate planning during the implementation phase of the investment and improved future evaluations. The benefit management approach also improved the relationship between IT staff and other employees, as people needed to work more closely together, thereby increasing mutual respect. This approach also led companies to make wiser investments. It doesn't only increase the value of investments but also helps avoid projects that would not deliver the expected benefits. Furthermore, and perhaps most importantly, the benefit management approach increases the realized benefits of investments. Using this approach greatly increases the likelihood that the expected benefits from an IT project are realized. This is very important because it is mostly the reason why the investment was made in the first place.

In the next chapter causes of change are combined with benefit levels forming a benefit measurement model.

3.2 Benefit measurement model

Ward and Daniel (2012, 133) present a compelling framework for evaluating investments through their benefit measurement model. Their model categorizes benefits according to four degrees of 'explicitness'. Levels of benefits are *financial*, *quantifiable*, *measurable* and *observable*. These levels are illustrated in table 1. This categorization is based on the ability to quantify a benefit and the extent of current understanding about potential enhancements. In this model any wanted perspective of benefit can be used. As stated by Aubry et al. (2021) different actors may look into different perspectives of benefits including return on investment, value management, and performance metrics. In their study, they found 56 examples of benefits, which they grouped based to their nature. Those benefit groups are operational, financial, social in external environment and users, and social in internal environment. However, this is not comprehensive listing and benefits not fitting to these groups may exist. Manwani (2008, 113) argues starting the

categorisation by considering is benefit's effects financial or non-financial. When considering their model, Ward and Daniel (2012, 133) suggest chosen benefit to be initially classified as either observable or measurable, provided it satisfies the minimum criteria. Barrier between these two levels is critical as shown later, but in nutshell benefit transforms from measurable, being only an assumption or based to good guesses to quantifiable, a concrete, meaningful benefit in business perspective. Following the initial classification, the placement of a benefit within one of the four specified degrees is refined based on the depth of knowledge concerning the benefit and the ascertainable factors. This subsequent assessment evaluates the extent of existing or potentially attainable insights, which could lead to a more precise categorization within the model.

Degree of Explicitness	**Do New Things**	**Do Things Better**	**Stop Doing Things**
Financial	By applying a cost/price or other valid financial formula to a quantifiable benefit a financial value can be calculated		
Quantifiable	Sufficient evidence exists to forecast how much improvements/benefit should result from the changes		
Measurable	This aspect of performance is currently being measured or an appropriate measure could exist. But it is not possible to estimate by how much performance will improve when the changes are completed.		
Observable	By use of agreed criteria, specific individuals/groups will decide, based upon their experience or judgement, to what extent the benefit has been realized		

Table 1. "Classifying the benefits by the explicitness of the contribution" from Ward and Daniels (2012, 134)

3.2.1 Observable benefits

Manwani (2008, 114) states that the first step in evaluating any benefit is ensuring that the benefit can be observed in some manner. Ward and Daniel (2012, 134-135) continue by stating that observable benefits necessitate a well-defined set of criteria for evaluation and the identification of individuals who are best suited to conduct these assessments objectively. This approach is crucial for accurately determining the realization of 'softer' benefits, like enhanced staff morale or increased customer satisfaction. Observable benefits are often classified as non-quantifiable, however some observable benefits can still be measured over time through surveys and feedback mechanisms. Therefore, it is

possible to quantify these benefits rather than just observe them. Although these benefits alone may not justify an investment, for example a new system, they play a crucial role in affecting numerous stakeholders whose behavioural changes are the key to achieving broader organizational goals. Consequently, these benefits should be consistently included in the business case and benefits plan, maintaining their consideration even in scenarios where financial and quantifiable benefits are predominant. (Ward & Daniel, 2012, 134-135) Observable benefits provide essential insights into the impacts of organizational initiatives on stakeholder attitudes and behaviours, which are key to achieving long-term success. Properly recognizing and documenting these benefits, along with implementing structured assessment processes, can greatly enhance strategic decision-making and project evaluation within organizations.

3.2.2 Measurable benefits

A measurable benefit is defined as an aspect of performance that is already under measurement or could feasibly have measurement criteria applied to assess enhancements after implementation. However, accurately predicting the exact scale of performance improvement before making changes is challenging. The ability to understand and measure change requires that previous performance (or another measurable level) be known and quantifiable. It is important to note that the metrics used should not only be relevant to the benefit itself but should also relate directly to the necessary changes for realizing these benefits, ensuring that any improvements can be clearly attributed to specific actions. (Ward & Daniel, 2012, 135) It is also possible that one measure is not enough to understand a big picture. For example, consider the scenario in operational efficiency improvements: if an upgrade in machinery leads to a faster production line, it's essential to assess not just the increase in output but also factors like the quality of the products and the maintenance downtime. It is necessary to think questions like: Has the increased production speed resulted in a higher rate of quality issues? Or has the new equipment effectively reduced downtime? Further, if the new process allows products to be assembled faster and with fewer defects, this could mean a decrease in the number of necessary quality checks, thereby saving time and resources. These considerations highlight the need for measures that reflect both the direct benefits and the operational changes needed to achieve these benefits, making it possible to directly link reported improvements to the implemented strategies.

When measuring benefits, Manwani (2008, 119) suggests using key performance indicators (KPIs) to track the progress of changes and integrate them into the company's metrics. Challenges may arise if the company lacks appropriate metrics to measure the benefits. KPIs can also be used at a quantifiable level if they are found suitable. Ward and Daniel (2012, 136) highlight one important notion, there is no need to forcibly elevate a benefit from observable level to a measurable level if it does not naturally lend itself to such classification. This is particularly relevant in scenarios where introducing or adapting measurement systems might be overly complex or expensive. Additionally, when focusing on metrics like KPIs, striving to achieve specific numerical targets can sometimes shift attention away from the actual desired benefits.

Similarly as it was stated in observable level, also in measurable level getting meaningful measures takes time between making the change and observing their effects. Ward and Daniel (2012, 136) note that the necessary time to see benefits varies. Some may appear almost immediately, while others might take weeks or months before any benefits or noticeable degree of change. They also emphasize the importance of estimating a time frame in which these benefits are expected to become evident and measurable before any further changes are made.

3.2.3 Quantifiable benefits

As noted in table 1, a benefit is considered quantifiable when there is sufficient evidence to predict the level of improvement that should result from the implemented changes. The key distinction between measurable and quantifiable benefits lies in the ability to estimate improvement before implementation. For quantifiable benefits, there must be enough evidence to calculate the expected amount of improvement. Once the most suitable measures for a benefit have been identified, ones again it is important to be able to establish a current baseline. This baseline is vital as it provides the reference point from which any future performance improvements can be measured and predicted. As in other levels this foundational step ensures that the anticipated enhancements can be accurately quantified and tracked over time. (Ward & Daniel, 2012, 136-137) There are various ways to define baseline. In factory setting it could be straight forward to measure finished products in old methods and later, after changes, compare if any improvement has happened.

Ward and Daniel (2012, 137-140) discuss the often challenging transition from measurable to quantifiable benefits within project frameworks. To bridge this gap, they introduce several methods that can legitimately quantify the impact of change projects. One such method, piloting, is utilized in this research. Additionally, they outline other techniques such as using reference sites, engaging in external benchmarking, applying modelling or simulation, and compiling detailed evidence, although these are not elaborated upon in this context.

3.2.4 Financial benefits

Ward and Daniel (2012, 144) state that on the financial level, benefits should be calculable, and valid financial goals should be set. From this model's perspective, a clear link between the benefit and financial improvement is expected. Manwani (2008, 115) further defines financial benefits as those that directly or indirectly either increase an organization's income or decrease its costs. He notes that only a few business change investments can be solely and directly linked to an increase of income. Regarding financial benefits and their measurement, Manwani (2008, 117) notes that financial metrics are often backward-looking, as they typically compare current financial figures to past data. Understanding how a particular change yields financial benefits requires careful consideration of whether the change directly impacts financial numbers. Additionally, if a change is justified by anticipated financial benefits, it should be clear that these benefits result from the change itself and not from other factors.

Ward and Daniel (2012, 144-145) state that organizations, especially companies, should aim to maximize financial benefits from changes. However, they also highlight several issues that may arise if organizations rely solely on financial metrics in their IS/IT investment decisions. They point out the following concerns:

> Not promoting innovation in IS/IT projects because the financial benefit is uncertain.
>
> Focusing only on individual IS/IT projects, without considering the broader (financial) impacts across the organization.
>
> Employing financial metrics in decision-making when it is not suitable.

Allocating only as much resources to a change project as the expected financial gain.

Minimizing costs in the system by sacrificing some functionalities or training.

Making assumptions that justify sufficient financial benefits to provide the necessary return on investment relative to the costs.

To summarize the discussed levels in the benefit measurement model, it is clear that each level has its place and is necessary. While more predictable outcomes and benefits of a change project are highly valued in business settings because they help justify the benefits before changes are made, there is also a significant role for observable and measurable levels. This is because not everything can be quantified in financial terms, as demonstrated by the potential issues arising from relying solely on financial metrics.

4 Research methodology

4.1 Qualitative and quantitative methods

Research methodologies typically fall into two main categories: qualitative and quantitative. While distinct, these approaches are complementary rather than mutually exclusive. Qualitative research focuses on understanding behaviours, events, and causations within specific groups, offering insights into the nuanced dynamics of particular situations. On the other hand, quantitative research aims to verify hypotheses and establish empirical, measurable facts. The influence of researchers also varies between the two; in qualitative research, outcomes may be shaped by the researcher's experiences and subjective interpretations, whereas quantitative studies demand the researcher's objectivity, minimizing their impact on the findings. Furthermore, quantitative data tends to be more replicable compared to qualitative data. (Ghauri et al., 2020, 95-98) The chosen methodology for this study is qualitative.

Data collection in qualitative research can utilize various sources, including interviews, questionnaires, observations, and existing documents. According to Tuomi and Sarajärvi (2002, 73), these are among the most prevalent methods. This particular study will primarily employ interviews as the main tool for gathering information. Notably, interviews allow for dynamic question adjustment and a deeper interpretation of responses, unlike the more rigid format of questionnaires (Tuomi & Sarajärvi, 2002, 75).

Ghauri et al. (2020, 115) describes interviews as ranging from structured to unstructured, with a third, hybrid form known as semi-structured interviews. Structured interviews are methodical and may include predetermined responses, aligning them closer to quantitative techniques. Unstructured interviews, conversely, offer extensive freedom for interviewees to express their views and reactions. Semi-structured interviews blend elements from both, providing preselected, open-ended questions that facilitate a flowing dialogue. This allows the interviewer to delve deeper into certain topics as the conversation progresses. For the purposes of this study, semi-structured interviews are deemed most appropriate due to their balance of structured guidance and conversational flexibility, potentially unveiling unforeseen insights related to the research theme.

Using interviews is particularly effective for this research as it seeks a profound comprehension of a subject that has not been extensively explored. This approach is

optimal for capturing individual perspectives and emotions, areas where quantitative or more rigid qualitative methods, such as surveys with preset response options, might fall short.

4.2 Design science

In core of design science is to understand the main problem domain and create problem solving artefact (Hevner et al., 2004). Natural science may be seen to provide explanations and understanding reality, design science attempt to create things that serve human purposes. Noteworthy is also design science's nature as technology oriented research, as stated – the attempt is to create something that solves problems. (March & Smith, 1995) Hevner et al. (2004) further position design science by suggesting that information system science encompasses two principal research paradigms: behavioural science and design science. Behavioural science aims to develop and validate theories that explain and predict human or organizational behaviour. In contrast, design science focuses on expanding human and organizational capabilities through the creation of innovative artefacts. They emphasize that both paradigms are crucial for a comprehensive examination of information systems from human, organizational, and technological perspectives.

March and Smith (1995) identified two design processes in information system research: building and evaluating. They describe these two processes as follows: "Building is the process of constructing an artefact for a specific purpose, and evaluation is the process of determining how well the artefact performs." One critical difficulty in design science, as raised by March and Smith (1995), is that the created artefact, its performance, and the possible benefits gained are almost always related to the environment. This interdependence has significant side effects and complicates unbiased evaluation. It may be challenging to assess outcomes, as the environment and users almost always influence the results.

Table 2 showcases design science research guidelines. In this thesis, the author designed a viable artefact, a data labelling instrument, as illustrated later in Figure 5. This instrument was implemented using Microsoft Purview tool. Tool selection and its capabilities are further discussed in Chapter 5.2.

Guideline	Description
Design as an Artefact	Design-science research must produce a viable artefact in the form of a construct, a model, a method, or an instantiation.
Problem relevance	The aim of design-science research is to develop technology-driven solutions for important and relevant business issues.
Design evaluation	The effectiveness, quality, and efficacy of a design artefact must be thoroughly validated through robust evaluation techniques.
Research contribution	Effective design-science research should make clear and verifiable advancements in the design artefact itself, design foundations, and/or methodologies.
Research Rigor	Design-science research relies upon the application of rigorous methods in both developing and evaluating the design artefact.
Design as a Search Process	The search for a suitable artefact involves leveraging available resources to achieve desired outcomes while satisfying to the constraints of the problem environment.
Communication of Research	It is essential for design-science research findings to be communicated effectively to audiences both in the technological and managerial domains.

Table 2. Design science research guidelines, adopted from Hevner et al. (2004)

The main objective of the pilot is to recognize the benefits of the tool used for labelling. The amount of data in organisations has increased massively, and new tools' methods for managing data are needed. The Purview pilot attempts to assist with labelling in the case company, Sitowise. The instrument is created based on the needs of the company. While seeking benefits, there is a goal to set them into a benefit management model in order to understand in what level benefits of data labelling and usage of data labelling software can be evaluated. The instrument's effectiveness and quality were evaluated through rigorous pilot testing and interviews which includes interviews of various professionals

inside Sitowise and also broader thinking what use cases are possible with labelling software in Sitowise. These evaluations, detailed in chapter 6, highlight the practical application of Microsoft Purview tool and provide insights into its benefits and potential use cases. The iterative design process leveraged available resources to create an instrument that meets the specific needs and constraints of Sitowise's data management environment. The created pilot lays the foundation for the future development of utilizing labelling software in Sitowise.

4.3 Piloting

Ward and Daniel (2012, 139) justify using piloting as a meaningful research method by stating piloting being an opportunity to test new technology and sometimes even more importantly possibility to evaluate the benefits that may be achieved from the new system or ways of working. Piloting is viable method also when new system is found working but there is no other feasible way of determining the degree of improvement that could result from the changes. They continue by stating that pilot is in many cases the needed proof of benefits new system or way of working provides or is expected to provide.

Piloting happens almost always with using smaller sample size as total benefits are more easily identified. Providing the best evidence requires comparing results from comparable group which works as before or comparing new system or way of doing to historical way of doing of the same thing. (Ward & Daniel 2012, 139)

As presented in Table 1, there are four degrees of benefits: Financial, Quantifiable, Measurable and Observable. Ward and Daniel (2012, 136-138) recognise the difficulty for organisations to get benefits – and broader, impacts – to quantifiable level. The table tells the difference between measurable and quantifiable benefit. The noticeable difference is that for measurable benefit pre-implementation benefit is not possible to make estimation of improvement’s size or degree. In many cases, without legitimate quantification it is very difficult or even impossible to count a realistic financial value of new implementation. They present piloting along with external benchmarking, modelling, simulating and others being a way to convert measurable benefits to quantifiable ones. By piloting a system benefits become much more realistic and observable, which helps making business related decisions and eventually to decide if the new system or way of working is implemented to day-to-day environment.

Part of this thesis is piloting data labelling tool Microsoft Purview, which is illustrated in chapter 5.

5 Case description

5.1 Sitowise

Sitowise Oy, a company listed on the Finnish stock exchange, specializes in building, infrastructure, digital solutions and environmental consulting and planning. It operates in Finland, Sweden, Estonia, and Latvia. Main customers are municipalities, government owned building, infrastructure and geographic information projects and building and infrastructure companies.

According to Kauppalehti (kauppalehti.fi), in 2022, Sitowise reported revenue of 152 million euros, with an operating result of 9.4 million euros. The company's workforce expanded from 1,224 employees at the end of 2019 to 1,673 by the end of 2022. Sitowise's headquarters is located in Espoo, Finland, and its business spans several sectors, including Buildings, Infrastructure, and Digital Solutions. This growth trajectory highlights the company's rapid expansion and underscores the necessity for innovative business practices and advanced tools to support its development.

The initial need for data labelling and the adoption of automated data labelling software at Sitowise rise from the company's ambition to enhance its applications, increase efficiency, and increase its competitive advantages from software. As the company has expanded, the volume of data it handles has also grown significantly. Moreover, strict data privacy regulations and certifications, such as the ISO 27001 in information security management system, require Sitowise to maintain rigorous data oversight. The company is keen to broaden its suite of applications, particularly by integrating AI-based tools like Microsoft Copilot. Additionally, Sitowise is exploring the development of proprietary applications utilizing their own data sets and large language models (LLMs).

At the initial stages of developing AI-based applications, Sitowise places a high priority on the security classifications of their data. This is a particularly critical concern when designing tools like chatbots that utilize company information. It is vital to ensure that these tools do not unintentionally grant access to sensitive data. This concern is backed up by Sağlam et al. (2021), who report that individuals, regardless of age, are apprehensive about the use of their sensitive data in AI-based tools such as chatbots. Given the potential organizational risks if users access data not intended for them, the issue of data sensitivity must be taken seriously.

Consequently, the objective is to implement labelling for all company data. Manually, by hand labelling each file would be impractical and resource-intensive, requiring continuous allocation of significant human resources also in the future. This underscores the benefits of adopting an automated solution. Discussions of implementing data labelling system and labelling all data files brought up several advantages. Labelling data tool helps restricted access to specific files, enhancing the company's control over its information assets. Furthermore, labelling data may open possibilities to implement other positively contributing systems, for example providing data loss and theft prevention, data lifecycle management (DLM) by clarifying security classifications, and ultimately an enhancement in data quality as whole.

5.2 Microsoft Purview

Microsoft Purview is a comprehensive data governance solution that provides tools for managing and governing data across a company's entire data estate. Microsoft Purview, as it is currently known, integrates the capabilities of the former Azure Purview and Microsoft 365 compliance solutions into a unified platform. It is plays pivotal role in assisting companies with data security, governance, and compliance tasks. (Microsoft Purview.com, 2024) Ahmed et al (2023) adds that Purview is distinguished by its ability to automate data management features such as data discovery, classification, and policy enforcement across a varied data landscape, which includes on-premises, edge, and cloud environments. This is noteworthy as in many situations' companies have data in different places and forms, therefore it is crucial to have solution handling all data in the same tool.

In the market, there are other similar data labelling software as Microsoft Purview such as IBM Watson Knowledge Catalog, Informatica's AI-Powered Cloud Data Management solution, and SAP Master Data Governance. Microsoft Purview's deep integration with the Microsoft ecosystem is valuable for firms already utilizing Microsoft products. Financially, it is often more viable as it seamlessly integrates without additional costs if Microsoft's tools and systems are already in place. Ahmed et al (2023) highlights Purview's integration with Office 365 products. It enhances data governance capabilities, allowing for consistent application of governance policies across various data formats and systems. Furthermore, Purview's capability to manage both on-premises and multi-cloud environments through Azure provides a flexible solution that adapts to various

organizational needs, setting it apart from solutions with more limited multi-cloud capabilities.

This thesis exploits Microsoft Purview's data labelling capabilities. Users can add their own sensitivity types and labels, leveraging these to govern the data processed within the platform. Purview's extensive functionality spans data governance, security, and compliance tasks. While its applications in data governance are broad, this thesis primarily focuses on compliance and security aspects, particularly on labelling company's documents with chosen sensitivity classes.

5.3 Project description

5.3.1 Project plan

Core of the pilot is to test labelling software's suitability to Sitowise's need for project labelling based on ISO 27001 data security standard with Microsoft Purview software. It is done by defining different labels, choosing and revising labelling rules and perfecting labelling based on software's results. The pilot is focusing on Microsoft's data formats such as .docx, .xlsx, .pptx and .pdix files, but also less platform dependent PDF files. The pilot is done in co-operation with Sitowise's IT development team – in which the author is included – and Nordic data consultancy firm Epical.

5.3.2 Use of ISO 27001

Based on publisher's website "ISO 27001 is the world's best-known standard for information security management systems (ISMS)." The ISO 27001 standard is essential for organizations of all sizes and industries, providing a solid framework for setting up, implementing, maintaining ISMS. By adopting this standard, organizations show their commitment to managing data security risks effectively, following established best practices and principles. As cyber threats continue to increase, ISO 27001 helps organizations proactively detect and tackle vulnerabilities, creating a culture of risk awareness. It advocates a comprehensive approach to information security that includes careful review of policies, technologies, and personnel. This broad view not only improves cyber resilience but also promotes operational excellence, making ISO 27001 a key element in protecting corporate information assets. Framework was firstly published

in 2005. Moment of writing the most resent iteration is third and it is from the year 2022. (ISO/IEC 27001, 2022)

Podrecca and Sartor (2023) researched implementing ISO 27001 standard and they found implementation process requiring significant amount of resources. Particularly companies need to invest a lot of time to their staff activities and meeting related to setting up information systems according to this standard. They also state standard to provide only limited advice for accomplishing needed things to get the certificate. Rezaei et al. (2014) add that many companies need to hire external consultant to accomplish any certification implementation to their information system, which increases costs. Podrecca and Sartor (2023) continue by inserting that difficulties arise many cases when working with external environment like other companies if those don't have same standard implemented. Lack of common standard may even prevent doing cooperation.

Despite its costs and challenges, ISO 27001 offers significant advantages to companies. Al-Karaki et al. (2022) assert that the certification reduces risk levels within information systems. Rezaei et al. (2014) emphasize its benefits in enhancing business continuity. Perhaps the most compelling reason for companies to adopt this standard is its impact on competitiveness. Podrecca and Sartor (2023) describe the ISO 27001 standard as an essential prerequisite for market entry, suggesting that certified companies are considered viable partners for collaboration in some fields and markets. Generally, having an ISO 27001 certificate is often seen as an advantage when selecting business partners because it serves as proof of adhering to certain levels of information security management. The widespread international recognition of the certificate further enhances its attractiveness in the marketplace.

According to their website (Sitowise, 2023), Sitowise was granted an ISO 27001 certificate for their Information Security Management System (ISMS) in 2022. They underscore the certification's recognition as a global standard for information security. They also highlight their position as the largest operator in the environmental planning and consulting field in Finland with this certificate. Consequently, their data labelling processes are required to match to the standards set by this certification.

In ISO 27001 standard there are many sections related to information security, cybersecurity and privacy protection. This thesis focuses on the classification of information, particularly as detailed in Annex A 5.12, which stipulates: "Information

should be classified according to the information security needs of the organisation based on confidentiality, integrity, availability, and relevant interested party requirements" (ISO/IEC 27001, 2022).

5.3.3 Data levels according to ISO 27001

ISO 27001 suggests the use of a four level data classification standard to support these needs, which includes *Public data*, *Internal data*, *Confidential data*, and *Restricted data*. Sitowise adopts these levels and are specified subsequently. These classifications allow Sitowise to identify data at various levels and ensure that each type is handled according to its sensitivity and security requirements. This approach is crucial for defining clear ownership of data across the company, thus facilitating better accountability and control over information assets. Furthermore, these classification standards are the core in responding to customer requirements and regulatory expectations, as they enable Sitowise to establish stringent controls tailored to the confidentiality and integrity demands of in all usages of data. The classifications also dictate processing requirements that span the entire lifecycle of the data, from creation to disposal, ensuring that data handling processes are consistent and secure.

The implementation of these classification levels enables Sitowise to establish standardized processes and essential criteria for managing data, which are crucial for ensuring strong data security and aiding the company's compliance with ISO 27001 standards. This systematic approach to data management not only safeguards the integrity and security of the data but also increases the trust and confidence of clients and stakeholders in Sitowise's practices.

In the following, we offer more detailed explanations to four levels of data classifications. The names Sitowise used for each level are in brackets and those are used later on when discussing about different levels. Descriptions for levels can be found from multiple sources and core of descriptions is also largely matching. For the purpose of this thesis and pilot, descriptions are from Sitowise's practical guides. In order for the pilot to meet the company's objectives, it is important to understand exactly how different levels are utilized at Sitowise.

Public (Public)

Data or information that is meant for external communication. Public information may include commercial and marketing material, material about clients that is already publicly available and general business information in company level. Core idea is that published information or data may not harm parties outside or inside the company and may even bring advantages to the company or clients when published.

Internal (Private)

Data or information that is meant for internal usage. Information is available for all internal stakeholders, excluding specific stakeholders whose/which access is restricted in their contract. Internal information may include general project information, information of existing customer relationship, employee qualification information, offers and contracts and results of opinion and feedback surveys. Core idea is that internal data or information may marginally harm the company or clients if made publicly available.

Confidential (Restricted)

Data or information defined as confidential by either the client or the company and which is not classified by authorities. Data or information is not available for all inside the company rather only to specific user group. Confidential information may include personal information, project information, financial information such as profitability calculations and billings, specific contract information and any information specified as confidential and not meant to be published for whole company. Core idea is that sharing confidential data or information outside of limited user group may significantly harm the company or cause breach of contract.

Restricted (Secret)

Data or information encrypted by the company, a partner of the company, or a client, or material classified or required to be classified as restricted by authorities. Restricted data or information may include sensitive information related to clients, information regarding confidential negotiations, specific personal information such as personal identification number or home address and information specified as restricted. With restricted data operating practices are always defined separately according to the specific characteristics of the data.

5.3.4 Project execution

Initial discussion of the need for data labelling software took place in February and March 2024. During these discussions, it was discovered that Microsoft offered a relatively new tool called Purview, which had seen significant enhancements recently combining old Azure Purview and Microsoft Compliance manager under one solution. By beginning of March, it was discovered that the security department at Sitowise would greatly benefit from applying sensitivity labels to project data. This need was further emphasized by the requirements of sustaining the ISO 27001 standard. The discussions at this stage involved multiple stakeholders inside the company, including the chief of IT development, the security lead, cybersecurity lead, a cloud specialist, and a system architect.

On March 14, 2024, Sitowise held a meeting with the Nordic consultancy firm Epical. This collaboration was sought to leverage Epical's expertise in implementing data labelling with Microsoft Purview, which would help avoid potential pitfalls in setting up the new system. It was decided that the IT development team at Sitowise would take responsibility for setting up the new system, with ongoing support from Epical.

The Purview pilot project kicked off on May 6, 2024. The pilot included three sessions, each lasting three hours. During these sessions, the basic capabilities of Purview were established, including the creation of sensitive info types and sensitivity labels. The labelling process was tested with a created policy and a test group using sample data.

It is noteworthy that although this pilot project was brief and focused on labelling, the use and exploitation of Microsoft Purview is an ongoing process that will continue beyond the pilot. Microsoft Purview offers numerous capabilities for data management after labelling, including Data Loss Prevention (DLP) policies, access control, information lifecycle management, and advanced analytics and metrics, as illustrated in figure 5.

5.4 Research model

5.4.1 Labelling process

In the pilot, a labelling instrument was created. The focus was on labelling test documents with four data security classifications: public, private, restricted, and secret, with the aim of making labelling mostly automated.

There were a few steps before Purview could label data. First, we created sensitive info types, as explained in chapter 5.4.2. These info types were used to create four sensitivity labels, detailed in chapter 5.4.3. To label data, a policy needs to be created and published. The policy dictates which labels are effective, what type of data, and from which source labelling is done. Creation of labelling policy is illustrated in chapter 5.4.4.

In Purview, labels are created by humans. This type of labelling is categorized as supervised labelling, as discussed in chapter 2. Whether labelling is performed automatically or manually, the groups within which data are categorized are predefined, and the software does not independently determine these groups. The created policy determines the default label for each file type. According to Woodward et al. (2020), the labelling process using Microsoft Purview involves both automatic and manual elements. Labelling is automatic when integrating already created data into the database, with the system assigning labels automatically according to predefined policy. The process is manual when creating new data entries, with the label assigned by the document's creator.

Figure 5 illustrates the construction of the Purview labelling process. The piloting phase included several key steps: defining information types, creating labels, establishing and publishing policies, and observing how effectively Purview labels the test data.

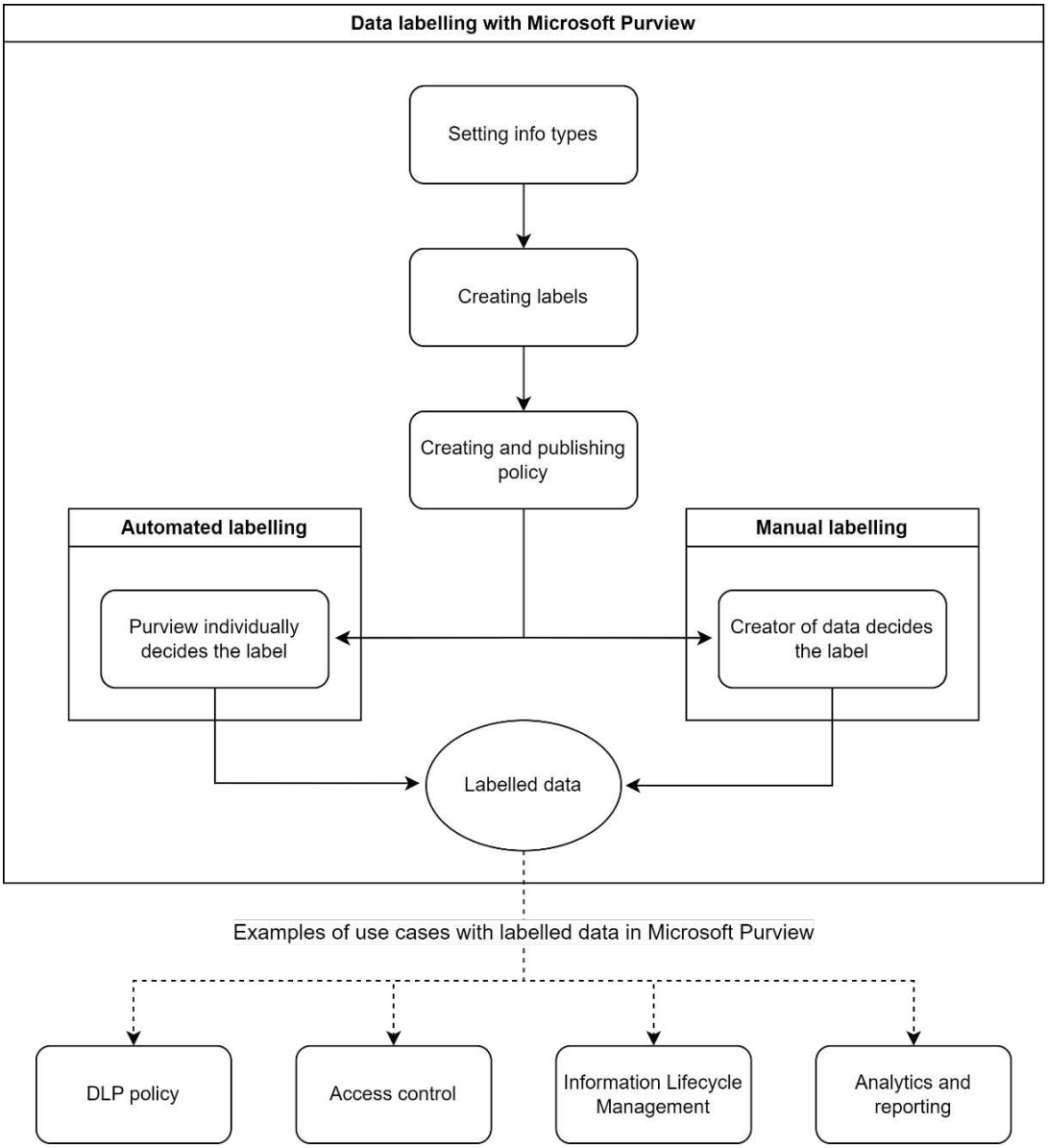

Figure 5. Labelling process using Microsoft Purview.

Manual labelling in Purview is done by the data creator. Purview asks the document creator what sensitivity label she/he wants to implement when she/he decides to save the document for the first time. After the initial labelling, changing label requires giving reason for the change.

Automated labelling occurs whenever an unlabelled document is opened, saved, or moved to a place where a labelling policy is set to work. Automated labelling works according to the published policy and included labels. Both automated and manual labelling operate simultaneously in Purview. Automated labelling detects certain

elements, such as Finnish personal IDs, but in some cases, manually setting or adjusting the label is necessary to ensure a practical and high-quality labelling outcome.

Data Loss Prevention (DLP) policy, access control, information lifecycle management, and advanced analytics and metrics of data are examples of how labelled data could be further utilized in Microsoft Purview. This thesis does not concentrate on that part further.

In the following three sub-chapters, the creation of sensitive info types, sensitivity labels, and policy creation is explained step-by-step with practical examples.

5.4.2 Creating new sensitive info type

Using Microsoft Purview, administrators can identify specific data elements within datasets by leveraging sensitive info types. For instance, in Finland, Purview offers predefined sensitive info types such as those for driver's license numbers, personal IDs, and Passport Numbers. These can be readily used or adjusted to better meet specific needs.

System administrators have the privilege to create new sensitive information types from scratch within Purview. Figure 6 depicts how new sensitivity info type is created.

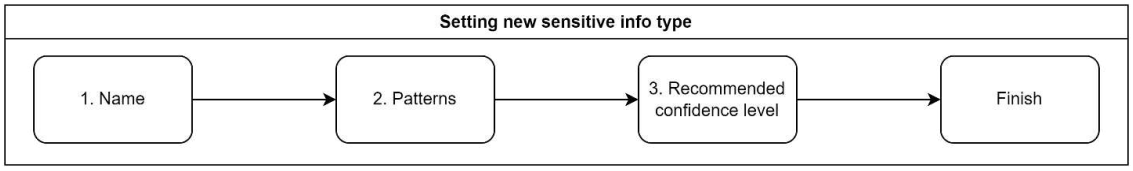

Figure 6. Illustration of creating a new sensitive info type.

The first step is the naming phase (1.), where the administrator assigns a name and description to the information type. The name and description should clearly convey the nature of the data the info type is designed to detect.

During the patterns phase (2.), the administrator defines patterns that the info type will use to identify data. These patterns may include regular expressions, keywords, keyword lists, or functions that match the data. Each info type must have at least one pattern, but multiple patterns can be used if needed to enhance detection accuracy. The creation of a new pattern starts with setting a confidence level, which determines the precision with which the pattern must match in the data. Options for confidence levels are high, medium, or low. Following this, the primary element of the pattern is set—it is the principal

component that the information type aims to detect and can be a regular expression, keyword, keyword list, or function. The administrator then configures the character proximity, specifying how close the primary and any supporting elements must be to trigger detection.

The third phase involves selecting a recommended confidence level (3.) for the info type. This setting can be also later adjusted by the administrator as needed and serves as guidance for using the information type effectively within Purview, such as when creating labels. The confidence levels available are high, medium, and low, where the highest level indicates the fewest false positives and the most false negatives. The lowest confidence level does the opposite.

The final phase is finishing step (4.), where the administrator reviews all settings to ensures everything is configured correctly and according to the intended specifications.

By following these steps, administrators can effectively tailor sensitive info types to meet the specific security and compliance needs of their organization.

5.4.3 Creating sensitivity label

Figure 7 depicts how a new sensitivity label is created in Microsoft Purview. Labels are created by administrator.

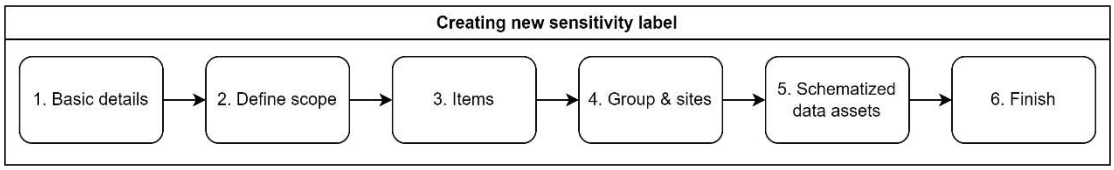

Figure 7. Illustration of creating a new sensitivity label.

Creating sensitivity label in Microsoft Purview begins by setting basic details (1.), including details defining the label's identity and purpose. Administrator gives a unique name visible to administrators and a display name that users will see in their applications. The label priority helps determine its position compared to other labels. Four created labels are Public, Internal, Restricted and Secret. They were set 1 to 4 in priority level. The higher priority level the higher level of security. Descriptions are provided separately for both users and administrators to understand the label’s function and scope. Additionally, a label colour can be chosen to make the label visually distinct. As an

example, we created restricted label according to Sitowise's specification. Its naming is straightforward, "Restricted" for both users and administrators. Priority level is three as restricted label is third of our four labels. Description defines label of being third of four sensitivity labels and it should be used according to Sitowise's security instructions. Lesser note is that we gave purple colour for this label.

The next step involves defining the label's scope (2.), which is crucial as it determines where the label can be applied. Labels can be attached directly to items such as files, emails, and meetings, as well as to larger containers like SharePoint sites, Microsoft Teams environments, and complex data structures like fabric and Power BI items. Decisions are made regarding specific applications, such as Microsoft 365 files, Outlook emails, and meetings scheduled via Outlook and Teams, where the label will be active. For our four labels, we chose to include all relevant applications.

The following step is configuring protection settings for the selected items (3.). Options include controlling access to ensure only authorized users can view the labelled items and applying content markings like custom headers, footers, and watermarks. Specific markers, such as tailored footer text, can be automatically appended to emails or documents. If no marking choices are made, the label will act invisibly, tagging the content without altering its appearance, with the marking visible only in the item's metadata. We decided not to add any watermarks or footer text to labelled documents, but users can still see which label is applied to each document.

Under item selection, there are auto-labelling settings. Auto-labelling settings are an important aspect of the label configuration process. This feature allows for the automatic application of labels based on detected sensitive info types within the content. Info type creation was explained in chapter 5.4.3, and default Microsoft info types can also be used. Administrators can set the label to be applied automatically or only recommend its application, providing customer advisory text to guide users when sensitive info types are detected. For our "Restricted" label, we used an existing Finnish national ID check with high confidence. This means that if Purview detects one or more Finnish IDs in a scanned document, it assigns the "Restricted" label to it. This phase is especially important for automatic labelling capabilities.

The groups and sites (4.) phase dictate internal and external user access to labelled teams and Microsoft 365 groups. We set that if a group or site has a default label other than

"Public", Purview applies that label to all documents in the group. We also set that only group members (not everyone in the company) can add users to groups with default labels included. Additionally, we configured that external people can be added to groups as guests.

Auto-labelling schematized data assets (5.) is still in preview status. This phase makes it possible to automate labelling for all company data governed by Microsoft's tools like Microsoft Purview. This capability helps detecting specific information and map where in the data it is located. This is especially useful if company decided to implement labelling to all company data and wants to know where sensitive data is and in what form. Our pilot is concentrated only on our test group and therefore applying created labels to all company data is not intended at this stage.

Finally, in the finishing phase (6.), upon completing the label creation, the administrator can review and verify all settings before saving the label. Even after saving, the label is not active. Created labels are stored in Purview and it makes is easy to configure many labels and only in policy creation phase choose which to use. Using labels requires policy creation, which is explained in the next sub-chapter.

5.4.4 Creating policy

Policy dictates which data types and from which location are labelled according to selected labels. Policy also dictates which data is managed and labelled automatically and which data requires human actor to decide the label. Sensitive info types or labels alone do not take effect without an applied policy, making the creation of these policies a critical step and is part of administrators work when setting up Microsoft Purview.

This chapter outlines the procedure for creating a new policy in Microsoft Purview, as illustrated step by step in Figure 8. Each step of the policy creation process is thoroughly explained. Additionally, this chapter describes the decisions that were made regarding the pilot study.

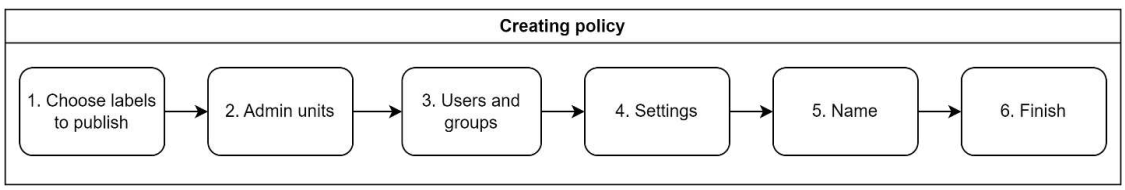

Figure 8. Illustration of creating new policy.

The process of establishing a new policy in Microsoft Purview begins by selecting labels to publish (1.). These labels, once published, become available to specified groups in all applicable Microsoft Office applications including Word, Excel, PowerPoint, SharePoint, Team sites, and Microsoft 365 groups. In the pilot study for this thesis, we incorporated those sensitivity labels, which were determined during the sensitivity label creation phase. Labels are: Public, Private, Restricted, and Secret.

The next phase involves defining admin units (2.), where permissions can be delegated to administrators, such as office managers, regional chiefs, or department managers. This facilitates the distribution of workload among administrators and allows the implementation of different labels tailored to distinct units. This selection is made for big corporation with regionally changing requirement in labelling. There was no need for delegation in our pilot study.

Following this, the users and groups (3.) selection is made, determining which users and groups will have access to the chosen labels. Options include importing Teams groups, individual users via Microsoft credentials, or all users within the organization. For our pilot, we utilized our internal Purview piloting group, with plans to extend label access to various project working groups in the future.

In the settings (4.) phase, configurations for the included labels are established. Choices here can range from requiring users to justify any downgrading or removal of a label's classification, to mandating that labels be applied to all relevant documents and emails upon their next save if they are not already labelled. Moreover, settings can enforce that labels are added to content in Fabric and PowerBI, and provide a link to a custom help page whenever labelling is required. Our pilot adopted the first option, necessitating justification for any reduction or removal of existing label levels, with specific reasons required for changes. Options for a justification of change are "Previous label no longer applies," "Previous label was incorrect," or an open explanation subject to admin review. Default labels were also set; documents and emails were labelled as Public by default, with emails containing attachments labelled according to their content. Public status was assigned by default to meetings and calendar events, while Fabric and PowerBI content were set to Private.

The process culminates in the name (5.) phase, where the policy is given a descriptive title and a description, followed by a finish (6.) phase, which includes a reviewing page that displays all the decisions made throughout the policy creation process.

6 Data collection and discussion

6.1 Interviews

For the purpose of this thesis, six employees were interviewed regarding data labelling. All interviewees work at Sitowise and are somehow involved with data labelling in their roles. The first two informants were directly part of the pilot made with Microsoft Purview. The third informant played a key role in the decision-making process and discussions leading up to the pilot. The remaining three informants were not connected to the Purview pilot.

The interviews were conducted via Microsoft Teams, and Microsoft Copilot was used to transcribe the discussions, allowing the interviewer to focus on the conversations. The interviews were held in semi-structured manner. The question framework is provided in Appendix 1. Table 3 lists all six interviewees, their titles, and brief job descriptions.

	Title	Brief job description
Informant 1 (i1)	System architect	Data analytics and product ownership
Informant 2 (i2)	Cloud infrastructure lead	Cloud infrastructure, access management, cybersecurity development and telecommunications.
Informant 3 (i3)	IT development manager.	Responsible for IT development in Sitowise.
Informant 4 (i4)	Security lead	Responsible for Sitowise's security in whole, including daily operative security, compliance and ensuring business continuity.
Informant 5 (i5)	Innovation manager	Responsible for the development of new services for the business.
Informant 6 (i6)	Technology manager in infrastructure business area	Leads technological development in the infrastructure business area and develops new and existing technologies.

Table 3. Interviewees' titles and brief job descriptions.

6.2 Labelling as a concept

Informants 1 and 2 mentioned that labelling is not a new concept, but i1 believes that its usage across all businesses will likely increase. I3 noted that while the term “labelling” might seem new or fresh, the same functions have been performed for a long time. This view aligns with the findings in chapter 2, where terms like “labelling,” “clustering,” and “classification” appear to be used interchangeably, although “labelling” has gained prominence in recent academic literature.

6.3 Direct and indirect benefits of labelling

Throughout all the interviews, there was a common theme of improved data quality through labelling, which broadens the usability of data. Labelling also plays a significant role in machine learning (ML). The informants consistently mentioned that as ML, artificial intelligence (AI), and large language models (LLMs) become more prevalent, the need to make data suitable for these technologies’ increases, with labelling being a crucial part of this process.

Informant 1 emphasized that labelling data does not have intrinsic value by itself. The value is derived from how the labelled data is used, particularly through policies and monitoring. With ML and AI becoming ubiquitous, the necessity for labelled data grows. Informant 5 highlighted that “labelling is a necessary function that has to be done to utilize any data assets. Including metadata to data is the starting point. It lays the foundation for what we can do with data, increases understanding of what we already have, and what we are still missing.”

Informant 4 describes labelling as a very positive manner. He states that, as a security lead, labelling is generally a simple solution to complex problems in data security. He explains that when companies transitioned from physical papers to digital formats, data labelling for security was not a significant issue; access rights managed who got access to what data. Nowadays, relying solely on access control is not feasible because the amount of data is enormous, it exists in multiple formats, and it is located in various locations. Therefore, the data itself must contain information about its type, which can then inform security decisions. Labelling serves precisely this purpose. I4 states, “Metadata is the only way to make this happen.”

Informant 5 also asserted that labelling provides direct benefits by making it easier for systems to locate specific data. I6 elaborated: “If data is labelled so that the system knows which picture has, for example, animals and which document has a certain element, it helps the system to find needed information more accurately.” This benefit extends to future applications, such as AI-based chatbots, where labelled data ensures responses are appropriate to the security level of the data and the user’s permissions.

Another significant benefit mentioned by I5 is related to data governance and management. He stated: “A direct benefit from labelling a company’s data is that it provides the opportunity to govern and manage the data the company has, including how much data there is to base business decisions on. Labelling dictates how data-driven a company can be.” This sentiment underscores the role of labelling in enhancing data-driven decision-making processes.

I2 brought up the advantages labelling brings to data lifecycle management (DLM). As discussed in the first chapter, the amount of data is constantly increasing, and individuals often neglect to delete or manage their data. Labelling can mitigate this issue by embedding metadata that dictates when data should be deleted or moved to a specific storage. Knowing the type of data contained within documents helps organizations manage their data more effectively.

Overall, the interviews reveal that people in Sitowise believe labelling to offer numerous direct and indirect benefits. It improves data quality, facilitates advanced technologies like ML and AI, enhances security, and supports better data management and governance. These benefits collectively contribute to the efficiency and effectiveness of organizational data handling.

6.4 Labelling related to benefit measurement model

Almost all direct benefits mentioned in the interviews were related to increasing data quality, thereby enhancing data's usability and functionality. All the commented benefits were at least at an observable level. However, there was also consensus that most of the benefits that fit into this model are indirect. For instance, applications like AI and ML gain advantages from data labelling, but directly attributing these benefits to labelling is not straightforward. Labelling helps achieve benefits, but usually indirectly.

I2 states that some labelling benefits can be measurable. For instance, if a certain dataset is labelled, it can be monitored for how storage is used with that dataset. This can also have direct financial effects, as identifying old or unused data that can be deleted will save money on data storage costs.

I2 brings up one challenge related to measuring the benefits of labelling. He states that, since there is no direct data from before labelling to compare against, measuring the change is based on assumptions. Even if we can observe that benefits exist, measuring them is a more complicated task. In some cases, the additional change in benefits could be measurable.

Increasing understanding of used data through labelling minimizes potential misclassifications and the resulting sanctions. For instance, when working with customer data, having their sensitive information in a vulnerable position may end up breaching a contract. Labelling data can provide direct financial benefits by preventing possible sanctions.

The results of the interviews showed that the main benefits of labelling are improved data quality, usability, and functionality. Most of the observed direct benefits are at an observable level; however, indirect benefits from labelling can be positioned at other levels as well. Applications like AI and ML benefit from data labelling, although directly attributing these gains to labelling is complex. Some measurable benefits include monitoring dataset storage, which can lead to financial savings by identifying and deleting old or unused data. However, measuring these benefits is challenging without baseline data for comparison, making the measurement at least partly based on assumptions. This makes moving from a measurable level to a quantifiable benefit level difficult. Labelling also minimizes misclassifications and potential sanctions, particularly when handling sensitive customer data, thereby providing direct financial benefits by preventing contractual breaches and associated penalties.

6.5 Benefits and risks of Microsoft Purview in labelling

Before the pilot, we discussed the potential risks associated with the Purview system. We identified several reasons why pursuing the pilot was worthwhile. Firstly, piloting the system was relatively straightforward and didn’t necessarily require deep knowledge of the system beforehand. Additionally, the option to get help from our partner firm, which

had experience with similar pilots, made us feel more secure about moving forward with Purview. The pilot didn't require extensive resources, and since the tool was part of Microsoft's licensing package that Sitowise already had, a permanent company-wide implementation seemed feasible. Furthermore, Purview was tested to bring benefits to the company rather than to fix ongoing problems, which would have increased demands on the pilot's success.

Despite these advantages, a couple of risks were recognized in the initial discussions. There is always a chance that the pilot could be a waste of time and resources. Additionally, there was no extensive competitive selection of the system, partly because Microsoft products are so widely used at Sitowise. Piloting a similar competing system didn't make sense, even though other systems might have been better or more suitable for the company's needs. Furthermore, relying heavily on one supplier for most IT software always carries some risk. Another challenge discussed was end-user involvement. Often, IT develops solutions that are theoretically useful, but users must "buy into" the change and the new system for it to be successful. This is especially important in sensitivity labelling. Even if the goal is to use automation as much as possible, end users still need to align with the new system operating behind the scenes. For users to accept it, the new system should not make their work harder or require too much additional effort.

One problematic theme discovered while making the pilot was the situation with the limited number of labels. As discussed in chapter 2, having too many different labels can present challenges. In a company with a lot of data and many files requiring a confidential label, problems arise because numerous people should have permission to files that are all labelled as confidential. One solution could be to create unique labels for all groups that have common confidential information, but this would require a substantial number of labels for essentially the same purpose. Another option is to combine labelling with other restricting methods, such as using Teams groups with restricted access. When considering whether labelling is worth the resource investment, I1 stated that indirect benefits from labelling documents in Purview include increased data security and compliance. He mentioned that the pilot was successful in demonstrating what Purview can do. However, another question is whether it meets the goals and objectives set by Sitowise. One major observation during the pilot was that Sitowise's project-based work makes it difficult to increase data security with only four security labels (public, private, restricted and secret). In most projects, access to highly sensitive data is needed, and if an

employee has access to one piece of highly sensitive data, they have access to all similarly labelled data. Another point raised was the difficulty in governing data assets due to the frequent exchange of data and documents with clients and other stakeholders.

A significant benefit that Microsoft Purview could provide in the near future, according to i1, is data loss prevention (DLP). Direct benefits from labelling documents with Purview using sensitivity labels include increased understanding of what kind of data Sitowise has. For the system to be effective, adjusting sensitivity labels needs to be easy for the user. The suitability of the Microsoft Purview system is further supported by the fact that Sitowise already extensively uses Microsoft services and file types.

As discussed in chapter 2, there are two ways of doing labelling: automatically or manually. I1 supports using automation as much as possible, as changing user habits and assigning labelling tasks to them would be very difficult and problematic. I2 complements this view but states that users should have the option to change labels if needed. He asserts that the way Purview operates is a good starting point, requiring both manual decision-making and automated label setting. I3 notes that for manual labelling to happen, there needs to be some benefit for the user themselves. If labelling relies too much on manual efforts, it may negatively affect the quality of labelling.

Labelling data with sensitivity labels in Purview would particularly help in managing Sitowise's internal data, including employee personal data, sales data, and financial data, among others. I1 and i2 both believe that Purview is a capable system with many (perhaps even too many) functionalities. One issue is that users don't necessarily use Purview, which might make it seem more engineering-oriented than user-friendly. Another point is that not all functionalities are needed by every company, so each organization can choose which features are meaningful for them and deploy those. Chapter 3 discussed that smaller projects are often more successful and that it is beneficial to concentrate on the most used features in software development. With Purview, the story could be similar: if Purview is harnessed as a companywide labelling software, Sitowise should perhaps concentrate on the most important features and if not maximize user friendliness, at least ensure that users' daily work is not interrupted too much.

I3 highlights a significant risk he sees as the IT development manager in the wide implementation of Microsoft Purview: "The broad deployment can be jeopardized if the system is unable to identify files with sufficient accuracy and extent, leading to users

having to do most of the labelling themselves.” This concern underscores the importance of the system's accuracy and ease of use to ensure successful adoption and implementation.

The pilot for Microsoft Purview at Sitowise showcased Purview’s potential benefits for Sitowise, including enhanced data security, compliance, and data loss prevention (DLP), facilitated by its integration within the existing Microsoft ecosystem. Despite its advantages, challenges such as limited label management, potential over-reliance on a single supplier, and the potential complexity of user adoption emerged. Effective use of Purview requires balancing between automation and manual labelling. Also, user involvement is required to maintain high-quality labelling, which is crucial for Sitowise's project-based work and frequent data exchanges. Overall, while Purview offers significant capabilities, its success hinges on overcoming user engagement and system accuracy issues to ensure broad and effective implementation.

6.6 Additional comments

When considering allocating resources to data labelling, it is important to keep the benefits and main goals in mind. Labelling can enhance data quality, but it is only beneficial if it is done thoughtfully and in alignment with those benefits. For instance, labelling data based on its size improves data quality by providing clear categorization. However, if the goal is to improve the accuracy of a predictive model, labelling data by size may not be helpful. Instead, labelling should be relevant to the features that impact the model's performance, such as categorizing data based on user behaviour, product type, or sentiment. In this example, resources are wasted, and labelling does not bring any benefits.

Labelling doesn’t necessarily have to be visible to users, but its effects often force some adaptation in behaviour. Being aware of labelling happening may increase employees' understanding of what kind of input they give to documents and other data. I4 notes that this is only a good thing. Based on interviews, informants unanimously think that acknowledging that sensitivity labelling is done would not change employees' behaviour, but it might make data usage more conscious among staff.

Labelling can also have both positive and negative effects simultaneously. Informant 6 provides an example related to recording and allocation of working hours: "More detailed

recording produces better and more accurate data for reporting and identifying time use, but at the same time, it consumes more of the employees' working hours. Then we as a company must decide which is the greater, the cost or the benefit."

A key consideration is that labelling can become a 'self-feeding circle.' If it is possible to demonstrate to users the benefits and how labelling increases the ability to be more data-driven, it boosts the willingness and acceptance to do labelling with a broader scope or greater resources.

Project and change management are essential when implementing data labelling to a company. I1 emphasizes that IT provides the tools, but implementation must be done with the business side in mind and requires user involvement. I1 asserts that IT most likely has to be always onboard, as data is almost always in some systems. Labelling would also be a change management process when users are given a chance to influence labelling.

Informant 4 notes that the more complex the data used in an organization, the more difficult it is to label it meaningfully. Nonetheless, I5 believes that if users have some tasks to perform themselves, it increases their understanding and the importance of data labelling. I5 also states that while human involvement is crucial, humans should not have too much power in deciding sensitivity labelling, as they are the weakest link in cybersecurity.

Effective data labelling requires thoughtful resource allocation, user involvement, and a balance between automated and manual processes. It should be an integral part of project and change management, considering both the direct and indirect benefits while mitigating potential risks.

In these interviews, it was noteworthy that interviewees working closely with technology emphasized different aspects compared to those focusing on other areas in their daily work life.

7 Conclusions

This master thesis examined data labelling, benefits management, and the benefits of data labelling tools for a company. In order to understand any benefits data labelling can provide, the concept and role in the organisation have to be understood. As discussed in chapter 2, data labelling means setting groups or labels that categorise data. Labelling identifies specific similarities or characteristics of data and places them into labels.

Another core part of this thesis was benefit management. Based on Aubry et al (2021), the primary aim of benefit management is to improve investment decisions in IT/IS projects. There is a debate in the literature about what should guide IT/IS investment decision-making. Financial factors have governed decision-making. This is understandable in the sense that in business, it is expected that investments are beneficial in financial terms. However, Ward and Daniel (2012, 127) suggest that factors other than financial ones should also guide investment decision-making. Peppard et al. (2007) found that analysing benefits more broadly than only in financial terms increases the value of investment, prevents failure and increases the likelihood for expected benefits to be realised.

To assist benefit management evaluation, Ward and Daniel (2012, 133) presented a benefit measurement model. It assists companies in quantifying a benefit and the extent of current understanding about potential enhancements. The model consists of four levels of benefits: observable, measurable, quantifiable, and financial. Each level has its place and necessity. Benefit evaluation is vital as any wanted benefit always begins with some change. Change is essential for any improvement to happen. There are three types of changes: doing new things, doing things better and doing things. Then, with the four-level measurement model, it is possible to better evaluate in which level change can be noticed.

Part of this thesis was pilot creation for the case company Sitowise. The data labelling tool was Microsoft Purview. When combining academic literature and the way Purview works, it is possible to conduct how labelling works in the tool. Considering Roh et al.'s (2021) high-level figure (Figure 4), Purview falls into the manual and active learning labelling category. Labels need to be created by hand, and it is expected that labels are created correctly and accurately to ensure the tool to label data correctly. The requirement of manual label creation means that the Purview tool is not the most automated solution

available for labelling. A noteworthy mention that came up in interviews was that even if label creation requires effort, after successful creation, Purview offers a number of useful applications where labelled data can be utilised.

The main research question in this thesis is, what are the potential benefits of implementing a data labelling tool? This thesis attempts to answer this question based on academic sources, piloting Microsoft Purview software and interviews with Sitowise's employees. Interviews revealed many potential benefits. Labelling tools improve data quality, making data more usable and functional. Labelling also increases the organisation's understanding of available data and helps make data-driven business decisions. As stated in chapter 2, labelling tools allow labelling to be more automated, even if cooperation with manual labelling is still often necessary.

Increased automation, speed, and capacity allow new use cases for labelled data. When talking within a scope of sensitivity, labelling and tools such as Microsoft Purview offer broad benefits in data management, governance and security. As stated in chapter 6, the sensitive data labelling tool allows advanced data lifecycle management capabilities, data loss prevention management, and data security and compliance capabilities. Overall, the tool provides benefits in increased data quality on many levels. Some benefits are direct, but most benefits can be acquired indirectly when labelled data are used to support other data use cases.

When considering the benefits of implementing a data labelling tool in relation to Ward and Daniel's (2012, 133) benefit measurement model, it is necessary to look into the specifics observed in our pilot. In the pilot, data was labelled according to four sensitivity levels: public, private, restricted, and secret. This approach suggests that benefits are related to data security, accessibility, and compliance. A common discussion in interviews was that many benefits related to data labelling tools are indirect. Benefits are acquired when labelled data is utilized for specific purposes. The Purview tool provides numerous applications for using labelled data once it is labelled. Examples include knowing where company data is stored, preventing company data from being moved outside the company, detecting attempts to move sensitive data, and restricting access to specific sensitive data within the company to only those with a high enough sensitivity rating.

Certainly, all these benefits can be observed. Equally increased data quality can be observed as data becomes more usable as its context becomes more explicit. The setting is more complicated when going to the next levels – measurable, quantifiable and financial benefit levels. For setting benefits to a measurable level and beyond, there should be some base level where to compare the change. The situation may be that data is not labelled, and no one really knows where and how sensitive data is located and who has access to it. It is easy to observe that there is a beneficial improvement but measuring it may not be easy. At the same time there are benefits that can be linked to measurable or even straight to financial benefit level. For instance, by adding monitoring to data by labelling it, the company can measure how much storage capacity it uses; this might lead to savings as spare capacity can be used for something else, or additional space can be sold to someone else. Financial benefits can be obtained when setting sensitivity labelling to increase data security. This straightly minimises the company's risk of incurring possible sanctions for contract breaches related to data loss or theft.

The first sub-question is stated as: What are the benefits of labelling? On a broader level, labelling increases data quality for numerous use cases. Some informants in interviews stated that labelling itself does not provide any benefits, and the value is generated from data, which quality is improved by labelling. However, there are clear benefits in labelling itself. Labelled data is a baseline for emerging artificial intelligence and machine learning capabilities, providing the necessary context for data interpretation. Labelling data increases understanding of owned datasets and improves the usability of data by clearly categorising it.

The second sub-question is stated as follows: What are the risks associated with the labelling tool? The main research question is all about the benefits labelling tools can provide to organisations. However, there is always another side to the coin. When an IT related project is involved, there is a risk of project failure and lost resources. As discussed in chapter 3, the CHAOS manifesto (2013, 5-7) states that small projects centred around the most used functionalities have the highest success rate. Therefore, extensive and complex projects should be avoided. Microsoft Purview can be complex and full of capabilities; it requires knowledge of the system and an understanding of the needs of the organisation to implement its use successfully. Another risk is relying too much on automation. Systems are rarely fault-proof. Especially when talking about sensitivity labelling, as was done in this thesis pilot, too much reliance on the system may cause

more harm than good, as incorrectly labelled data is worse than not labelled data. As stated above, for labelling purposes, Purview requires manual working, and therefore, it is unlikely that the system is to blame if labelling is incorrect. However, this brings another risk, as relying too much on expert setting labels may cause harm if a human makes a mistake. Roh et al. (2021) detailed different labelling methods. Based on their observations and suggestions, the crowdsourcing method could tackle human error risk. Additionally, one major risk arises from connecting systems and people. It is essential that end users fully endorse and adopt the new system for its successful implementation. In sensitivity labelling, some labelling has to be done by humans. If people in the organization do not believe in the benefits of labelling, those potential benefits may never materialize.

8 Summary

In this master thesis we created an instrument and tested it in a pilot where Microsoft Purview was used as a data labelling tool. The aim was to identify the benefits associated with data labelling tools and data labelling itself, as well as to understand the risks related to their use. The research is grounded in academic literature and supported by interviews that validate the results.

The literature used is predominantly recent and relevant, with most sources from the 21st century and many from the 2020s. However, there appears to be a scarcity of research in this area. Data labelling is a relatively under-researched topic that has gained more attention in the past decade. Research on the benefits of data labelling tools is particularly limited. Many interviewees suggested that labelling and grouping data have become increasingly timely as the volume of data grows and the quality of data becomes more critical for the effectiveness of artificial intelligence and machine learning models.

The findings highlight the advantages of data labelling and its tools. Firstly, these tools enhance data quality and comprehension, making the data more useful. Secondly, automated labelling tools speed up the process, saving resources compared to manual methods. Thirdly, data labelling brings broad benefits in various aspects of data management, including governance, security, and compliance. Risks include accuracy issues, user engagement, and resource allocation concerns during implementation.

The author developed an instrument, which was tested in a pilot within Sitowise. Despite the limited data used for testing, discussions and interviews indicated that the company viewed this pilot as a promising foundation for further development. The pilot successfully demonstrated Purview's capabilities in automating data labelling, suggesting it as a beneficial tool for enhancing data management and security.

The thesis acknowledges certain limitations. The pilot study was relatively small in scale, conducted within a single company, and focused on labelling only a specific type of data. Pilot was also surrounded in sensitivity labelling. Testing in different companies and environments with various data types and needs would have provided more robust conclusions. Furthermore, evaluating different labelling tools could have offered insights into whether the benefits observed with Microsoft Purview are generalizable to other tools. The interviews conducted involved six employees from the same company, albeit

from different roles and departments. A larger number of interviews and a broader selection of professionals would likely have preferably brought more diverse perspectives, more fresh ideas and more common agreements.

Future research could focus on exploring more benefits of using data labelling tools as academic research related to that area mostly doesn't exist. Investigating use cases of labelled data and the direct benefits could be valuable. Additionally, there are numerous potential applications for labelled data, but further research is needed to evaluate the most effective use cases and the contexts in which they are viable. Evaluating different data labelling tools in various environments could enhance our understanding of the essential features and capabilities that an effective labelling tool should possess.

References

Ahmad, S. R., Arumugam, D. D., Bozovic, S., Degefa, E., Duvvuri, S., Gott, S., Gupta, N., Hammer, J., Kaluskar, N., Kaushik, R., Khanduja, R., Mujumdar, P., Malhotra, G., Naik, P. S., Ogg, N., Parthasarthy, K. K., Ramakrishnan, R., Rodríguez, V., Sharma, R., . . . Wolter, A. (2023). Microsoft Purview: A System for Central Governance of Data. Proceedings of the VLDB Endowment, 16(12), 3624–3635.

Al-Karaki, J. N., Gawanmeh, A., & Elyassami, S. (2022). GoSafe: On the practical characterization of the overall security posture of an organization information system using smart auditing and ranking. Journal of King Saud University. Computer and Information Sciences/Mağalaẗ Ǧam'aẗ Al-malīk Saud : Ùlm Al-ḥasib Wa Al-ma'lumat, 34(6), 3079–3095.

Aubry, M., Boukri, S. E., & Sergi, V. (2021). Opening the black box of benefits management in the context of projects. Project Management Journal, 52(5), 434–452.

Brynjolfsson, E., & Hitt, L. M. (2000). Beyond computation: information technology, organizational transformation and business performance. The Journal of Economic Perspectives, 14(4), 23–48.

Cambridge dictionary. Classification. Cambridge Dictionary. Retrieved 2.4. 2024, from https://dictionary.cambridge.org/dictionary/english/classification

Cao, F., & Liang, J. (2011). A data labeling method for clustering categorical data. Expert Systems With Applications, 38(3), 2381–2385.

Chen, Q., Allot, A., Leaman, R., Doğan, R. I., Du, J., Li, F., Wang, K., Xu, S., Zhang, Y., Bagherzadeh, P., Bergler, S., Bhatnagar, A., Bhavsar, N., Chang, Y. C., Lin, S., Tang, W., Zhang, H., Tavchioski, I., Pollak, S., . . . Lu, Z. (2022). Multi-label classification for biomedical literature: an overview of the BioCreative VII LitCovid Track for COVID-19 literature topic annotations. Database, 2022.

Cule, P. E., Schmidt, R. C., Lyytinen, K., & Keil, M. (2000). Strategies for Heading Off is Project Failure. Information Systems Management, 17(2), 61–69.

Dahlberg, T., & Nokkala, T. (2015). A framework for the corporate governance of data – theoretical background and empirical evidence. Business, Management and Education, 13(1), 25–45.

Desmond, M., Müller, M., Ashktorab, Z., Dugan, C., Duesterwald, E., Brimijoin, K., Finegan-Dollak, C., Brachman, M., Sharma, A., Joshi, N. N., & Qian, P. (2021). Increasing the Speed and Accuracy of Data Labeling Through an AI Assisted Interface. IUI ’21: 26th International Conference on Intelligent User Interfaces.

Diaz-Pinto, A., Mehta, P., Alle, S., Asad, M., Brown, R., Nath, V., Ihsani, A., Antonelli, M., Palkovics, D., Pintér, C., Alkalay, R. N., Pieper, S., Roth, H. R., Xu, D., Dogra, P., Vercauteren, T., Feng, A., Quraini, A., Ourselin, S., & Cardoso, M. J. (2022). DeepEdit: Deep Editable learning for interactive segmentation of 3D medical images. In Lecture Notes in Computer Science (pp. 11–21).

Everitt, B. S., Landau, S., & Leese, M. (1974). Cluster analysis. Retrieved 23.3.2024, from https://openlibrary.org/books/OL21531058M/Cluster_analysis

Fredriksson, T., Mattos, D.I., Bosch, J., Olsson, H.H. (2020). Data Labeling: An Empirical Investigation into Industrial Challenges and Mitigation Strategies. In: Morisio, M., Torchiano, M., Jedlitschka, A. (eds) Product-Focused Software Process Improvement. PROFES 2020. Lecture Notes in Computer Science(), vol 12562. Springer, Cham.

Furner, J. (2020). Definitions of “Metadata”: A Brief survey of international standards. Journal of the Association for Information Science and Technology, 71(6).

Ghauri, P., Grønhaug, K., & Strange, R. (2020). Research methods in business studies.

Godwin, J., & Matthews, P. (2014). Robust statistical methods for rapid data labelling. In Advances in data mining and database management book series (pp. 107–141).

Hautz, W. E., Kündig, M. M., Tschanz, R., Birrenbach, T., Schuster, A., Bürkle, T., Hautz, S. C., Sauter, T., & Krummrey, G. (2021). Automated identification of diagnostic labelling errors in medicine. Diagnosis, 9(2), 241–249.

Hevner, March, Park, H., & Ram. (2004). Design science in Information Systems Research. Management Information Systems Quarterly, 28(1), 75.

International Organization for Standardization (ISO) & International Electrotechnical Commission (IEC). (2022). ISO/IEC 27001:2022. Retrieved April 21, 2024, from https://www.iso.org/obp/ui/en/#iso:std:iso-iec:27001:ed-3:v1:en

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering. ACM Computing Surveys, 31(3), 264–323.

Jürgen, Bernard; Hutter, Marco; Sedlmair,Michael; Zeppelzauer, Matthias & Munzner, Tamara. (2021). A taxonomy of property measures to unify active learning and human-centered approaches to data labeling. ACM Transactions on Interactive Intelligent Systems, 11(3–4), 1–42.

Kauppalehti.fi, Sitowise Oy | Yritys- ja taloustiedot. Tärkeimmät Talousuutiset | Kauppalehti. Retrieved 23.2.2024, from https://www.kauppalehti.fi/yritykset/yritys/sitowise+oy/2335445-0#:~:text=Sitowise%20Oy%20(2335445%2D0),nettotulosprosentti%20oli%204%2C36%25.

Keil, M., Cule, P. E., Lyytinen, K., & Schmidt, R. C. (1998). A framework for identifying software project risks. Communications of the ACM, 41(11), 76–83.

Krigsman, M. (2007). New IT project failure metrics: is Standish wrong? ZDNET. Retrieved 25.4.2024, from https://www.zdnet.com/finance/new-it-project-failure-metrics-is-standish-wrong/

Lai, V. T., & Tan, C. (2019). On Human Predictions with Explanations and Predictions of Machine Learning Models. FAT* ’19: Proceedings of the Conference on Fairness, Accountability, and Transparency.

Learn about Microsoft Purview. (2024). Microsoft Purview. Retrieved 9.5.2024, from https://learn.microsoft.com/en-us/purview/purview

Lee, H., Lee, H., Hong, H., & Kim, J. (2022). Noisy Label Classification Using Label Noise Selection with Test-Time Augmentation Cross-Entropy and NoiseMix Learning. In Lecture Notes in Computer Science (pp. 74–82).

Love, P. E., & Matthews, J. (2019). The ‘how’ of benefits management for digital technology: From engineering to asset management. Automation in Construction, 107, 102930.

Lyytinen, K., Mathiassen, L., & Ropponen, J. (1998). Attention Shaping and Software Risk—A categorical analysis of four classical risk management approaches. Information Systems Research, 9(3), 233–255.

Manwani, S. (2008). IT - enabled business change: successful management. (1st ed., pp. xxvi–xxvi). Swindon: BCS, The Chartered Institute for IT.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. Decision Support Systems, 15(4), 251–266.

Peppard, J., Ward, J. M., & Daniel, E. M. (2007). Managing the realization of business benefits from IT investments. MIS Quarterly Executive, Pp. 1–11., 6(1), 3. Retrieved 23.4.2024, from http://dblp.uni-trier.de/db/journals/misqe/misqe6.html#PeppardWD07.

Podrecca, M., & Sartor, M. (2023). Forecasting the diffusion of ISO/IEC 27001: a Grey model approach. the TQM Journal, 35(9), 123–151.

Pomerantz, J. (2015). Metadata. MIT Press.

Rezaei, G., Ansari, M., Memari, A., Zahraee, S. M., & Shaharoun, A. M. (2014). A huiristic method for information scaling in manufacturing organizations. Jurnal Teknologi/Jurnal Teknologi, 69(3).

Riley, J. (2017). Understanding Metadata: What is Metadata, and What is it For? Retrieved 3.6.2024, from https://digital.library.unt.edu/ark:/67531/metadc990983/m2/1/high_res_d/understanding_metadata.pdf

Roh, Y., Heo, G., & Whang, S. E. (2021). A survey on Data Collection for Machine Learning: A Big Data - AI Integration Perspective. IEEE Transactions on Knowledge and Data Engineering, 33(4), 1328–1347.

Sağlam, R. B., Nurse, J. R. C., & Hodges, D. (2021). Privacy concerns in Chatbot interactions: When to trust and when to worry. In Communications in computer and information science (pp. 391–399).

Seetha, H., Murty, M. N., & Tripathy, B. K. (2018). Modern technologies for big data classification and clustering. In Advances in data mining and database management book series.

Sitowise.fi, The Smart City Company | Sitowise. (2020). Retrieved 23.2.2024, from https://www.sitowise.com/sitowise,

Sitowiselle harvinainen tietoturvasertifikaatti. (2023). sitowise.com. Retrieved 21.4. 2024, from https://www.sitowise.com/fi/uutiset/sitowiselle-harvinainen-tietoturvasertifikaatti

Sun, Y., Lank, E., & Terry, M. (2017). Visualizing the likelihood of machine learning classifier's success during data labeling. IUI '17: Proceedings of the 22nd International Conference on Intelligent User Interfaces.

The Data Management Association (DAMA). (2009). The DAMA guide to the Data Management Body of Knowledge (DAMA-DMBOK guide). 1st ed. Bradley Beach, New Jersey: Technics Publications.

The Standish Group International. (2013). CHAOS MANIFESTO 2013. Retrieved 12.5.2024, from https://www.pm-partners.com.au/wp-content/uploads/2020/08/Chaos-Manifesto-2013.pdf

Tuomi, J., & Sarajärvi, A. (2002). Laadullinen tutkimus ja sisällönanalyysi. Helsinki: Kustannusosakeyhtiö Tammi.

What is data labeling? | IBM. Retrieved 12.3.2024, from https://www.ibm.com/topics/data-labeling

Woodward, K., Kanjo, E., & Οικονόμου, A. (2020). LabelSens: enabling real-time sensor data labelling at the point of collection using an artificial intelligence-based approach. Personal and Ubiquitous Computing, 24(5), 709–722.

Zhang, Y., Wang, Y., Zhang, H., Zhu, B., Chen, S., & Zhang, D. (2022). OneLabeler: a flexible system for building data labeling tools. CHI Conference on Human Factors in Computing Systems.

Appendices

Appendix 1 Template used in interviews. (Translated from Finnish)

Interviews were done in semi-constructed manner and therefore questions and topics presented in here are only a baseline for discussion. Also last topic “Microsoft Purview” were discussed only with informants knowing the system (informants 1, 2 and 3).

Introduction (these were told to interviewees)

Thank you for agreeing to participate and finding the time for this interview.

If it suits for you, I will record this interview by using Copilot transcription to document our conversation, allowing me to focus on our discussion and accurately record observations later.

The objective is to gather professional insights and discuss the benefits of data labelling. The interview aims to let you share your views beyond the prepared questions, so feel free to express any thoughts that come to mind.

My thesis focuses on data labelling and its benefits. At Sitowise, we piloted the Microsoft Purview system to test both automatic and manual document security classification. The categories used were public, internal, restricted, and encrypted. The questions in this interview cover data labelling in general and in the end we can discuss about our Purview system pilot.

Definition of data labelling (this is the same as stated in chapter 1.1)

“According to Jürgen et al. (2021), data labelling involves transforming an unlabelled data instance into a labelled one by assigning it a specific category or descriptor. IBM (ibm.com) further explains that this process requires initially identifying raw data, such as images, text files, or videos. Subsequently, a relevant label is added to the data's metadata, which clarifies the data's context and enhances its usability in machine learning models. This step is crucial for training accurate predictive models, as it provides the necessary context for algorithmic interpretation.”

Basic Information

Please start by stating your job title and a brief description of your role or work responsibilities.

General Questions

What is your general perception of data labelling? Has it appeared in your work, and is it a prominent topic? Do you have a positive, negative, or neutral view of it?

What direct and indirect benefits do you see in data and document labelling?

Can you think of ways in which data labelling could change current practices?

Is labelling always an IT project, or can it be a change management project?

Should labelling always be automatic, or should users have the power to make decisions? Why?

(Following figure shown, from chapter 3.2)

Degree of Explicitness	**Do New Things**	**Do Things Better**	**Stop Doing Things**
Financial	By applying a cost/price or other valid financial formula to a quantifiable benefit a financial value can be calculated		
Quantifiable	Sufficient evidence exists to forecast how much improvements/benefit should result from the changes		
Measurable	This aspect of performance is currently being measured or an appropriate measure could be. But it is not possible to estimate by how much performance will improve when the changes are completed.		
Observable	By use of agreed criteria, specific individuals/groups will decide, based upon their experience or judgement, to what extent the benefit has been realized		

Are the benefits, in your opinion, observable, measurable, quantifiable, or measurable by financial metrics? Why?

Can you give any examples regarding data labelling, which you could place into this model?

Is there a risk that labelling consumes more resources than its benefits?

Are there benefits to labelling itself, or do the benefits only come from using the labelled information for certain purposes?

Regarding Purview

Does it seem like a useful tool?

What are its strengths and weaknesses?