Enhancing Impact Measurement of Philanthropic Organisations: A Human-AI Collaboration Framework

Information System Science

Master's Thesis

Author:

Jonah Cabaye

Supervisor:

Jocelyn Husser

04.05.2025

Basel, Switzerland

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 Thesis

Subject: Philanthropy & AI
Authors: Jonah Cabayé
Title: Enhancing Impact Measurement of Philanthropic Organisations: A Human-AI Collaboration Framework
Supervisors: Jocelyn Husser
Number of pages: 68 pages + appendices 20 pages
Date: 05.06.2025

Philanthropic organisations increasingly face pressure to demonstrate the impact of their work, yet existing impact measurement practices remain fragmented, resource-intensive, and often ill-suited to capturing both qualitative and quantitative outcomes. This thesis addresses these challenges by proposing a human–AI collaboration framework designed to enhance the efficiency, traceability, and usefulness of impact data in the nonprofit sector. Building on principles of Design Science Research (DSR), the study integrates semantic technologies (ontology and knowledge graphs), natural language processing (NLP), and automation tools within a prototype system aimed at structuring and querying unstructured impact data.

The research is informed by a two-phase empirical process: initial exploratory interviews to identify key challenges and requirements, followed by evaluative interviews assessing the system's perceived usefulness, usability, and ethical acceptability. The results confirm the relevance of established models such as the Technology Acceptance Model (TAM) and Human-Centered AI (HCAI) in this context, highlighting the importance of transparency, trust, and organisational fit. The proposed framework was found to effectively support common impact measurement needs, such as aggregating indicators, linking data to strategic goals like the SDGs, and making qualitative insights more analysable.

This work contributes both a functional prototype and a set of design recommendations for responsible AI implementation in the social sector. It also responds to documented gaps in the literature regarding integrated, context-sensitive AI tools for nonprofits. The findings underscore the potential of AI to support evidence-based decision-making in philanthropy, provided that technical innovations are embedded within participatory, ethical, and user-centred processes.

Key words: Impact Measurement, Philanthropy, Nonprofit Organisations, Human–AI Collaboration, Knowledge Graph, Ontology Engineering, Natural Language Processing (NLP), Technology Acceptance Model (TAM), Human-Centered AI (HCAI), Design Science Research (DSR), Responsible AI, Semantic Technologies, Sustainable Development Goals (SDGs)

Acknowledgements

I would like to express my deepest gratitude to those who supported me throughout the development of this thesis.

First and foremost, I sincerely thank **Rachel M.**, for her trust, availability, and guidance. Her confidence in my abilities allowed me to immerse myself in a field that was new to me, and her feedback was essential in shaping the direction and depth of this research.

I am also grateful to **Jocelyn Husser** for his valuable guidance and thoughtful advice throughout the process. His insights helped me strengthen the clarity and structure of this work.

A special thanks goes to **Léna G.** for her unwavering support, constructive reviews, and encouragement at each critical stage.

I would like to warmly thank all the **interview participants** and those who introduced me to the world of nonprofit organisations and philanthropy. Their openness and experience brought this research to life and grounded it in real-world relevance.

Finally, I wish to thank my **IMMIT classmates**, as well as my **family** and **friends**, for their constant solidarity and constant support throughout this journey.

To all of you—thank you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

1.1 Background and Context

Philanthropic organisations play a pivotal role in addressing complex social and environmental challenges. However, in recent years, they have faced increasing pressure not only to act, but to demonstrate that their actions lead to measurable, positive change (Eikenberry and Kluver 2009). As expectations from funders, regulators, and the public rise, impact measurement has shifted from a peripheral reporting activity to a strategic imperative (Ebrahim and Rangan 2010). Internally, data is also sought as a basis for learning, improving programme effectiveness, and informing decision-making processes.

Robust impact measurement is therefore essential for both external accountability and internal learning. It contributes to organisational legitimacy and credibility, directly influencing an organisation's ability to attract support and build trust. Initiatives such as those led by the European Venture Philanthropy Association (EVPA) have reinforced the notion that evaluation should be integrated into all stages of programme design and implementation, rather than treated as an afterthought.(European Venture Philanthropy Association 2015)

Yet, despite its recognised importance, impact measurement remains a significant challenge—particularly for organisations with limited resources. Many philanthropic actors operate across multiple programme areas engage with diverse stakeholders, and manage large volumes of unstructured information such as grant reports, evaluations, and beneficiary feedback. Traditional methods—largely manual, fragmented, and narrative-based—often fail to provide timely, structured, and scalable insights (Cordery et al. 2023).

In this context, recent advances in artificial intelligence (AI) and data technologies offer new opportunities. Tools such as Optical Character Recognition (OCR), Natural Language Processing (NLP), Large Language Models (LLMs), and Knowledge Graphs (KGs) present promising avenues to automate data extraction, structure disparate data, and generate analytical insights. These technologies could significantly reduce the burden of reporting while enhancing the quality of information available for impact assessment (Courth 2024).

However, the adoption of AI-driven tools in the philanthropic sector has so far been limited. Barriers include a lack of technical capacity, ethical concerns, and the absence of tailored frameworks that align with philanthropic values and operational realities (Zabłocka-Kluczka, Marciszewska, and Brajer-Marczak 2024). Moreover, it is increasingly acknowledged that technology alone cannot create an impact-driven organisation—it must be accompanied by a culture that values evidence, reflection, and learning (Della Giovampaola et al. 2023a).

This thesis explores how AI technologies, when thoughtfully implemented, can support and strengthen impact measurement practices in philanthropic organisations. It aims to identify both the potential benefits, and the critical conditions required for their effective and responsible use, particularly in organisations facing constraints in resources and technical expertise.

1.2 Problem Statement and Research Gap

Persistent Challenges in Impact Measurement: Philanthropic organisations continue to face persistent challenges in impact measurement, including difficulty attributing outcomes, fragmented data silos, lack of standard metrics, and underutilised qualitative feedback (Della Giovampaola et al. 2023a). These issues often lead to delayed or superficial impact reports, limiting their value for learning and accountability.

Limitations of Current Frameworks: Various impact-measurement frameworks (e.g., logic models, Theory of Change) provide useful structure, but they are labour-intensive to maintain and often fail to scale in practice (Kay Gugerty and Karlan 2018). As a result, a gap persists between what organisations plan to measure and what they can track and learn from.

Literature Gap: Despite growing interest from both the philanthropy and technology communities, there is still no comprehensive framework (Anfossi and Moralis 2024) that integrates AI into the impact measurement process for philanthropic organisations (Courth 2024). Practitioners lack guidance on which AI technologies to use, how to apply them responsibly, and how to embed them into existing monitoring and evaluation workflows (Della Giovampaola et al. 2023b). Additionally, the human and ethical factors (e.g., user acceptance, trust, transparency), explored in much research still not make it clear how to design an AI-driven impact measurement system that is

both effective and readily embraced by philanthropic organisations (Cipriano and Za 2024). Addressing this gap is important for both improving nonprofit practice and advancing information-systems knowledge in a socially meaningful context.

This thesis addresses the gap by proposing to design and evaluate a prototype AI-supported impact measurement framework tailored to philanthropic organisations. The envisioned solution incorporates a suite of AI and data technologies – e.g., OCR, NLP, large language models (LLMs), and knowledge graphs – to streamline data processing and enhance the interpretability of impact information. Crucially, it is guided by principles of Design Science Research Methodology(Hevner et al. 2004), human-centered design (Shneiderman 2020) and AI ethics to ensure that it aligns with philanthropic values (transparency, fairness, privacy) (Geneva Centre for Philanthropy - UNIGE 2025) and is feasible for organisations with limited resources. The goal is not merely a theoretical construct, but a proof-of-concept artifact whose utility can be demonstrated and tested with real stakeholders. By iteratively building and assessing this artifact (Hevner et al. 2004), the research seeks to contribute a novel framework to both practice (a roadmap for philanthropy to leverage AI in impact measurement) and theory (lessons on designing acceptable, ethical AI systems in nonprofit settings).

1.3 Research Objectives and Questions

The primary objective of this study is to design and evaluate an AI-driven framework to improve how philanthropic organisations collect, analyse, and report impact data Essentially, the study asks how AI can help nonprofits not only gather data, but also draw meaningful insights and communicate impact more credibly. To meet this goal, the project will (1) identify the current challenges and user needs, (2) determine which technologies can address these issues, (3) develop a prototype solution, and (4) evaluate its usefulness and acceptability with stakeholders. The study is therefore structured around the creation of an IT artifact (the framework and its prototype implementation) and an assessment of that artifact's usefulness and usability with end-users.

To guide the inquiry, one central research question (RQ) is posed, supported by four sub-questions (RQ1–RQ4):

- **RQ: How can artificial intelligence be effectively applied to optimise impact measurement and jreporting processes in philanthropic organisations?**

The sub-questions break this problem into components, each corresponding to a facet of the design challenge and aligned with specific chapters of the thesis:

- RQ1: *What are the current challenges in impact measurement in philanthropic settings?*

Rationale: Before designing a solution, we must clearly understand the problem. This question focuses on diagnosing issues with how philanthropy currently measures impact (addressed through literature review and stakeholder interviews). Answering RQ1 establishes the requirements the AI framework must meet, by identifying pain points like data fragmentation, lack of standard metrics, manual workload, etc.

- RQ2: *How can AI technologies (e.g., OCR, NLP, knowledge graphs) be integrated to automate and improve these processes?*

Rationale: This question explores the solution space. It asks which specific AI and digital tools could tackle the challenges from RQ1 and how they might fit together. Literature on AI applications and case examples inform this (Chapter 2), and the conceptual framework design directly addresses RQ2 by mapping technologies to impact-measurement needs.

- RQ3: *What design principles should guide the development of an AI-supported framework that is both useful and acceptable to end-users?*

Rationale: Introducing AI in a sensitive, resource-constrained context requires thoughtful design. RQ3 is about the human and ethical factors – ensuring the artifact is user-friendly (addresses real user requirements, easy to use) and responsible (aligns with ethical standards). Theoretical models like the Technology Acceptance Model (TAM)(Davis 1989) and AI ethics guidelines are used to derive these design principles (Chapter 2). RQ3 is answered through both the design requirements in Chapter 4 and the discussion of how theory informed our design choices.

- RQ4: *How do stakeholders perceive the usability, usefulness, and ethical implications of the proposed AI framework?*

Rationale: Ultimately, the success of the artifact depends on end-user evaluation. RQ4 focuses on the evaluation results: when actual philanthropic professionals and experts interact with or observe the prototype, do they find it valuable? Does it actually address

the challenges (usefulness)? Is it easy enough to use (usability)? Do they trust it and feel it aligns with their norms (ethical acceptability)? This question is tackled in Chapter 5 through qualitative feedback from stakeholder interviews, thereby closing the loop to see if the design (informed by RQ1–RQ3) holds up in practice.

Together, RQ1–RQ4 span the problem diagnosis, solution design, and evaluation, ensuring a comprehensive approach to the overarching research question.

1.4 Research Approach

This study adopts a Design Science Research methodology (Peffers et al., 2007) to build and evaluate an artifact (see Chapter 3 for detailed methodology).

Methods and Data: The study employed a mix of qualitative and design methods. It began with a literature review (and expert consultations) to ground the understanding of current impact measurement practices and relevant AI capabilities. Additionally, documents from partner organisations were accessed and analysed to enrich this understanding. Next, two rounds of semi-structured interviews were conducted with philanthropic stakeholders. The first round (exploratory) gathered insights on current processes, needs, and challenges – directly informing the problem definition and design requirements (RQ1 and RQ3). The second round (evaluation) occurred after the prototype was developed, collecting stakeholder feedback on the system's perceived usefulness, ease of use, and ethical acceptability (addressing RQ4). Both phases followed standard ethical research practices (informed consent, anonymity, and data privacy). Overall, the research process was iterative and stakeholder-centered, starting from real-world problems and looping back to practice through evaluation. (Detailed methodology is provided in Chapter 3.)

1.5 Thesis Structure

Following this introduction, Chapter 2 reviews the literature on philanthropic impact measurement and relevant AI technologies, identifying key challenges and gaps. Chapter 3 describes the design science research methodology used. Chapter 4 covers the design and implementation of the AI-driven impact measurement framework. Chapter 5 shows the framework in use and presents stakeholder feedback. Chapter 6 discusses the findings' implications and concludes with the contributions, limitations, and future work, and Chapter 7 concludes the thesis.

2 Literature Review

2.1 Philanthropy and Impact Measurement

Over the years, the social sector has developed various frameworks and models to plan, monitor, and evaluate outcomes. These provide structured approaches that any new solution should complement (not conflict with). For example, logic models and Theories of Change(Funnell, Sue and Rogers, and Patricia 2011; Goldsworthy 2021) chart how inputs and activities are expected to lead to outputs, outcomes, and long-term impacts. They help clarify what to measure and what "success" looks like, and they encourage defining key performance indicators (KPIs) at different levels. A limitation, however, is that they require significant effort to update, so many nonprofits struggle to keep them 'living' as programs evolve. Another common guide is the OECD-DAC evaluation criteria (Relevance, Effectiveness, Efficiency, Impact, Sustainability, Coherence)(OECD 2024), which encourage assessing projects from multiple angles. These criteria are widely adopted by international NGOs and donors to ensure comprehensive evaluations. A modern impact data system should accommodate such dimensions – for instance, by storing information not only on outputs and outcomes but also on contextual factors like those required for assessing sustainability or coherence. Yet, manually tracking all these aspects is difficult with current tools.

In recent years, there has been a push toward standardised metrics to enable comparability across projects. One prominent example is IRIS+, a library of common social and environmental indicators developed by the Global Impact Investing Network (GIIN)(Global Impact Investing Network 2025). IRIS+ provides standardised definitions for metrics in areas like education, health, or poverty alleviation, so that organisations can "speak the same language" when reporting outcomes. For instance, rather than each nonprofit defining its own indicator for jobs created, IRIS+ offers a standard definition.(Global Impact Investing Network 2025) Adopting such standards improves comparability and aggregation of data; Similarly, global frameworks such as the Sustainable Development Goals (SDGs) provide universally recognised impact categories and targets, enabling organisations worldwide to clearly align their outcomes(United Nations 2025). For instance, SDG 4 (Quality Education) explicitly targets universal literacy and numeracy, which helps organisations measure and communicate their contributions against widely recognised benchmarks. Many

organisations now align their core indicators with IRIS+ or SDGs. If our AI framework can incorporate or map to these common taxonomies, it will greatly enhance utility by enabling benchmarking and shared learning. For example, two projects reporting on "Number of children with improved literacy skills" could be recognised as using the same metric if mapped to an IRIS+ or SDG code, allowing the system to aggregate or compare results effortlessly. Alongside formal frameworks, participatory and qualitative approaches are also valued. They engage stakeholders directly in defining what "impact" means to them. The challenge is that qualitative data—stories, interviews, open-ended feedback—is hard to systematically analyse and compare. Organisations often lack tools to make sense of large volumes of narrative information, so such data may be underutilised(Mackinnon 2018). AI techniques offer a promising way to extract themes from qualitative inputs. Each existing impact framework has notable strengths but also limitations. Therefore, an effective modern system should combine their advantages – for example, link data to a Theory of Change structure, adopt standardised indicators for consistency, and incorporate qualitative insights (e.g. via NLP) – to address those gaps. These theoretical foundations inform our design requirements – for example, the importance of linking evidence to outcomes (for traceability) comes from evaluation criteria, and the need to handle both quantitative and qualitative data comes from recognition of mixed-methods approaches.

Common Challenges in Impact Measurement: Despite growing commitment to impact measurement, organisations still face several well-documented challenges in executing impact measurement effectively. The key challenges include:

- **Attribution**: It is difficult to determine whether observed outcomes can be attributed to a specific intervention (Mayne 2012). Social programs operate in complex environments with many contributing factors, and nonprofits often collaborate with governments or other NGOs. Credibly isolating one program's impact (short of expensive methods like randomised trials) is often unfeasible, leading to reliance on qualitative judgments of contribution. Scholars and practitioners alike note that without credible attribution, impact data risks being dismissed as anecdotal or inconclusive (Groß 2024). This remains a fundamental evaluation challenge that no software can fully solve, though a system can help assemble evidence to support plausible contribution claims.

- **Data Fragmentation and Silos**: Impact data is often scattered across multiple documents and systems (Sopact 2024). A foundation might have grant reports in PDF, survey results in Excel, and field updates in emails – with each program using its own formats. Because there is no unified database, staff spend inordinate effort gathering and cleaning data from "ten different places and formats," as one interviewee emphasised. Important insights are easily missed; for example, two projects might be working toward similar outcomes without realising they are tracking the same indicator, due to data being siloed. Any solution must prioritise integrating these disparate data sources so that information can be accessed in one place.

- **Underutilisation of Qualitative Data**: Many organisations collect rich qualitative feedback (beneficiary stories, open-ended survey responses, focus group notes), but this information is not easily analysed alongside numeric metrics (Bamberger, Vaessen, and Raimondo 2016). Systematically reading and coding hundreds of pages of text is labour-intensive and often beyond the capacity of busy Monitoring & Evaluation (M&E) staff. As a result, narrative insights are often left unused. One program manager lamented, "We get a lot of great stories from the field, but we don't have a good way to analyse them, so they end up as a couple of anecdotes in the annual report". This is a lost opportunity, since qualitative data can explain why certain outcomes occurred or provide context for the numbers. The challenge is finding efficient ways to summarise and classify this unstructured data so it can inform decisions alongside quantitative indicators (Hehenberger, Harling, and Scholten 2015).

- **Lack of Standardisation:** Different projects and grantees often use different metrics and definitions, making it impossible to aggregate or compare results across a portfolio (Hehenberger, Harling, and Scholten 2015). For example, one education program might measure student attendance rate while another measures test score improvement – both indicate education outcomes, but they are not directly comparable (Sopact 2024). Without some alignment or translation between metrics, any higher-level analysis becomes an "apples and oranges" problem. As one stakeholder noted, "Each project has its own KPIs, so when we try to roll up the data, it's apples to oranges". Efforts like IRIS+ and the SDGs provide common definitions to mitigate this issue, but adoption of

standards is voluntary and inconsistent. A new framework should facilitate mapping of equivalent concepts (e.g., recognising when two indicators are essentially the same) to support portfolio-wide analytics.

- **Resource Constraints:** Rigorous impact measurement can be expensive and time-consuming, yet nonprofits often lack dedicated monitoring and evaluation (M&E) staff or budget (Kail, Vliet, and Baumgartner 2013). Many program teams must juggle service delivery with data collection and reporting as a secondary task (Cordery et al. 2023). Smaller organisations may have no data analyst on staff, resulting in rudimentary evaluation practices (Kail, Vliet, and Baumgartner 2013). Even when data is collected, there may be little time for in-depth analysis. One interview participant candidly described their reality: "We don't have a full-time data person – evaluation is one more thing on my plate, so we do what we can, but a lot probably falls through the cracks.". This constraint means any solution must be efficient and user-friendly, automating what it can so that limited human capacity is freed up for critical thinking rather than tedious data work. It also means training and support are crucial, since staff tasked with using the tool may not be data specialists.

- **Data Quality and Verification**: Impact data is only useful if it's reliable. However, data collected in the field or from partners can be prone to errors and inconsistencies. Surveys might have respondent bias or misunderstanding; different staff might apply indicator definitions inconsistently; or implementers might (consciously or not) overstate results to "look good " (Bamberger, Vaessen, and Raimondo 2016). When consolidating data from multiple sources, ensuring consistent definitions and avoiding double counting is also challenging. A well-designed system could help by enforcing data entry rules and flags for out-of-range values, but users must still be vigilant. Our framework will need to include features for data provenance and quality checks to build confidence in the information (for example, logging the source of each data point and allowing easy cross-checks).

- **Timeliness of Data**: Traditional evaluations are often conducted annually or at project end, so by the time findings are compiled, they are months out of date(Ebrahim and Rangan, V. Kasturi 2014). In fast-moving programs,

managers need more real-time or frequent updates to make timely decisions. If collecting and processing impact data takes too long, insights arrive too late to be actionable. Stakeholders in our study pointed out that by the time an annual impact report is published, the data might be 6–12 months old – "the data is a year old by the time we share it with stakeholders", as one interviewee noted. This lag can hamper organisational agility. There is hope that digitisation and AI could shorten the cycle – for instance, by continuously aggregating new data or enabling on-demand querying of the latest information. A goal for our design is to move toward more up-to-date impact tracking (even if real-time may not be fully attainable, more frequent refreshes are valuable).

In summary, the persistent challenges in impact measurement include difficulties with attribution, fragmented and siloed data, underuse of qualitative information, lack of standard metrics, resource and expertise gaps, data quality issues, and delays in reporting. These pain points form the basis of the problem our research addresses. They directly inform the requirements for our artifact: essentially, the framework should attempt to mitigate these challenges – for example, by integrating data silos, supporting both quantitative and qualitative analysis, promoting standard definitions (through an ontology), reducing manual effort via automation, and enabling traceability of data to its sources. Of course, some challenges (like attribution) cannot be solved by software alone, but the system can at least help assemble and link the evidence needed for credible impact claims.

Finally, approaches to impact measurement differ by organisation type and context. Philanthropic organisations are not monolithic (Mackinnon 2018) – they range from large foundations to small grassroots NGOs, and they operate in diverse sectors (Kail, Vliet, and Baumgartner 2013). These differences provide nuance as we design a framework intended to be broadly useful. For example, large foundations with substantial resources may fund external evaluations and enforce common reporting standards across grantees, whereas small nonprofits might have minimal capacity and rely on informal storytelling for impact evidence. A flexible system must cater to both: it should offer sophistication for those who can use it, but remain simple enough for a small NGO to input basic data or narratives without extensive training. Likewise, organisations that run their own programs (operational NGOs) collect data directly and may benefit from real-time dashboards for internal management, whereas grant-making

foundations must aggregate reports from many independent partners – they need tools for high-level comparison and roll-up of results. Our framework should accommodate both use cases (for instance, drill-down project details for implementers and roll-up portfolio summaries for funders). Sectoral focus matters too: a health-focused charity might track rigorous medical outcomes and integrate with government health data systems, while an arts funder might rely on qualitative indicators of community engagement that are harder to quantify. The framework's ontology needs to be adaptable to different domains and indicator sets, potentially by incorporating external vocabularies (such as SDG goals or sector-specific metrics). Moreover, organisational culture and readiness vary(Kail, Vliet, and Baumgartner 2013). Some organisations have a learning-oriented, tech-forward culture and will eagerly embrace an AI tool, whereas others are more traditional or sceptical of technology. We observed in our stakeholder interviews that user buy-in depends heavily on clearly communicating how the system will help (and not threaten) their work. In short, one-size-fits-all is unlikely in this space. A successful framework needs to be flexible and user-configurable, accounting for differences in scale, data sophistication, sector metrics, and user expectations. This understanding is reflected in our design — for example, we included optional modules and configurable indicators and involved diverse stakeholders in the design process to capture their nuanced needs. Having established the importance of impact measurement, the limitations of current practices, and these contextual nuances, we now turn to how technology and AI might address some of these issues.

2.2 AI and Digital Solutions for Impact Measurement

Developments in digital technology – particularly in artificial intelligence – offer new opportunities to transform how impact is measured and managed (Anfossi and Moralis 2024). The nonprofit sector has been gradually embracing digital tools for monitoring and evaluation, but often in piecemeal ways. Many organisations have moved from paper to online surveys and from ad-hoc spreadsheets to basic databases or dashboards. However, adoption of advanced data analytics and AI remains slow. A recent survey noted that while most impact investors and foundations recognise the value of better data, only a minority have invested in sophisticated analytics capabilities(Kail, Vliet, and Baumgartner 2013). The barriers include cost, lack of in-house expertise, and concerns about data privacy or ethics. Nonetheless, there are success stories: for example, some NGOs reduced the time to compile donor reports from weeks to days by

adopting centralised data platforms and simple automation, significantly improving efficiency (DataKind 2020; Cipriano and Za 2024; UNICEF Office of Research 2024). These cases hint at the gains possible and reinforce the motivation for our research. The question is how emerging AI techniques might be harnessed in a cohesive way to tackle the challenges outlined in Section 2.1.

A variety of AI and digital tools could be leveraged as part of an integrated solution. In our framework design, we consider the following technologies (even if not all are fully implemented in the current prototype) as building blocks for addressing impact measurement problems:

- **Optical Character Recognition (OCR):** OCR converts text from scanned reports and PDFs into machine-readable data, unlocking information trapped in unstructured documents. By digitising archival reports, organisations can make years of content searchable and integrate these legacy records into a unified database (DataKind 2020). This helps address data fragmentation by bringing formerly siloed narrative information into the central system, where it becomes accessible for analysis. (Any extracted content can be reviewed for accuracy to ensure data quality.)

- **Natural Language Processing (NLP):** NLP encompasses techniques for analysing human language, which is particularly useful for nonprofits' qualitative data (e.g. open-ended survey responses and narrative reports). It can automatically identify themes or extract information from large collections of text, dramatically reducing the manual effort of coding and summarising responses. Studies have demonstrated that NLP can sift through extensive textual datasets to find evidence of outcomes while maintaining transparency in how results are obtained (Schmidt et al. 2024). By quickly highlighting common issues or noteworthy results across hundreds of pages, NLP helps ensure that valuable narrative feedback is systematically used in impact measurement rather than overlooked. Crucially, these tools augment human analysis rather than replace it: the AI might perform an initial scan for data to populate databases, but staff must verify and interpret the findings to preserve context and nuance.

- **Large Language Models (LLMs):** LLMs (such as GPT-4) are powerful AI systems that understand and generate text, offering two main possibilities for our

framework: they could serve as a natural-language query interface for users and could automatically summarise complex documents. For example, staff might simply ask the system a question in plain English and receive an answer synthesised from the underlying data, or have a lengthy impact report condensed into a concise summary. However, LLMs also pose challenges: they sometimes hallucinate information (producing plausible-sounding but incorrect answers) if not carefully constrained by real data (Gao et al. 2024). Any use of an LLM in this context must therefore be tightly controlled so that its responses draw only on verified information (Lewis et al. 2021) (ideally with transparency about sources) and that sensitive data is protected. Given these risks and the early stage of this technology, LLM integration is considered an aspirational future enhancement rather than part of our initial prototype – although our system's design could accommodate such features when it becomes feasible (Kang et al. 2023; Courth 2024).

- **Knowledge Graphs and Ontologies:** A knowledge graph is a data structure that represents information as a network of interconnected nodes (entities such as Projects, Outcomes, Indicators, etc.) linked by relationships (IBM 2021a). Using a defined ontology (a formal schema of concepts and relationships), the knowledge graph provides a unifying semantic model for previously siloed data sources. In practice, this means that different datasets can be mapped to common terms – for example, what one report calls "Number of Participants" can be recognised as the same indicator as "ParticipantsCount" in another. By aligning all data to shared ontology definitions, our system can merge and query information across sources that were not originally compatible. This semantic approach offers several benefits. It enables flexible cross-cutting queries that span multiple dimensions of the data without complex manual merging – for instance, one could easily query which projects in a region address a specific outcome, since projects, locations, and outcomes are all linked in the graph. The knowledge graph is also naturally extensible: as metrics or categories evolve in the social sector, new entity types or relationships can be added to the ontology without overhauling the entire data model (Horrocks et al., n.d.). Another key advantage is traceability and transparency: each data point or result in the graph can be linked back to its source document, so users can see exactly where each

piece of information originated. This makes analysis results far more explainable and trustworthy compared to black-box algorithms. Indeed, prior studies have shown that applying knowledge graphs to social-sector data allows analysts to integrate diverse datasets and uncover patterns that remained hidden when data was fragmented in separate silos (Li et al. 2020). *(Detailed design and implementation of the ontology-driven knowledge graph for this project are provided in Chapter 4.)*

- **Robotic Process Automation (RPA**): RPA uses software "bots" to automate repetitive, rules-based tasks by mimicking human actions in digital systems. In our context, RPA acts as integration glue between disparate tools, reducing manual workload(IBM 2021b). For example, a bot could periodically extract data from one system or file (such as downloading metrics from a spreadsheet or email) and then import it into our central database, sparing staff from tedious copy-paste routines. RPA can also automatically generate routine outputs like consolidated reports or updated dashboards by populating templates with the latest data. By offloading these menial tasks, RPA ensures the new AI-driven process does not create double work for staff – a common concern among stakeholders when introducing any new system. In fact, users can continue working in familiar applications (e.g. Excel), and the RPA bots will handle transferring that information into the unified platform in the background. We view RPA as a convenient interim solution for integration until more direct system interfaces (APIs) can be established, allowing our framework to fit into an organisation's existing workflow without demanding immediate changes to all processes (Eikebrokk and Olsen 2020).

- **Machine Learning (ML):** Machine learning (ML) can add powerful analytical capabilities once a consolidated impact dataset is in place(Porter, Xia, and Préel-Dumas 2022). For instance, an ML model could be trained to predict a project's likely outcomes based on patterns learned from past projects, or to flag anomalies (such as unusually high or low reported results) that warrant further review. ML could thus help identify hidden trends or non-obvious factors associated with successful outcomes that human analysts might miss. However, realising these benefits requires a sufficiently large and reliable dataset, and ML must be used carefully to avoid reinforcing any biases present in historical data.

In practice, this means any predictive algorithms should be introduced with transparency and human oversight so that they inform rather than outright drive decisions (Huber et al. 2020).

Given these prerequisites and risks, we consider advanced ML features to be a future extension of the framework rather than an immediate component. Our initial prototype focuses on integrating data and providing basic descriptive analytics, ensuring the information is clean, consistent, and well-structured. This approach makes the system "ML-ready" — by building a solid data foundation now, we enable more sophisticated analyses (predictive models, clustering insights, recommendation systems, etc.) to be added responsibly at a later stage when enough quality data has been accumulated. The key point is that while ML is not a priority for the first implementation, the unified model we create lays the groundwork for leveraging machine learning down the line.

In combining these technologies into one cohesive framework, we are essentially creating a socio-technical system tailored for philanthropic impact measurement. Each tool addresses certain challenges – for example, OCR and data integration tackle silos, NLP/LLMs leverage qualitative data, and RPA/automation reduce manual workload. But technological solutions alone are not enough; their success depends on human factors like user adoption and trust. Both the literature and our findings emphasise that the best outcomes occur when new tools are introduced alongside capacity-building efforts and with users involved in the design (European Commission 2019). This reinforces our participatory design approach. Ultimately, to effectively address the challenges identified earlier, we must ensure not only that we include the right technologies, but also implement them in a way that practitioners will use and trust. Therefore, the next part of our theoretical foundation addresses user adoption and ethical AI design – to ensure the solution is both usable and responsible.

2.3 AI Ethics and Responsible Design Considerations

Even a powerful system yields no benefit if users do not accept and use it. The Technology Acceptance Model (TAM) (Davis 1989) indicates that two factors – how useful a technology appears and how easy it is to use – largely drive user adoption. We therefore design our AI framework to maximise both clear utility and user-friendliness. For example, it automates tedious data aggregation to save staff time (enhancing

perceived usefulness) and offers an intuitive interface that fits into existing workflows (ensuring ease of use). We also plan organisational support and user involvement throughout development so that practitioners feel confident and see the tool as helping rather than disrupting their work. By focusing on these TAM principles from the outset, we aim to increase the likelihood that the system will be embraced in practice.

Equally important, the framework is grounded in core AI ethics principles to address trust and responsibility (European Commission 2019). We draw on widely endorsed guidelines that emphasise transparency, fairness, privacy, accountability, and human agency. In practice, this means making the system's operations explainable and open to scrutiny (transparency), ensuring it does not amplify bias or injustice (fairness), safeguarding sensitive data (privacy), and keeping humans in control of critical decisions (accountability and agency). By upholding these values from the start, we build user trust and align the tool with philanthropic norms, avoiding the pitfalls that caused some past AI interventions to be rejected. In summary, both the user-centric focus (Shneiderman 2020) of TAM and these ethical design imperatives guide our framework so that it is not only technically effective but also acceptable and responsible in its real-world context.

2.4 Summary of Literature Findings and Gap Analysis

The literature review highlights several important findings that shape this research. First, philanthropic organisations face persistent difficulties in impact measurement – such as difficulty attributing outcomes, fragmented data silos, inconsistent metrics, underutilisation of qualitative feedback, and resource constraints – that current frameworks (e.g., logic models, Theory of Change, Social Return on Investment, and standardised metrics like IRIS+) only partially address. No existing approach fully resolves these issues in practice, especially not by leveraging AI, indicating a clear need for innovation.

Second, recent advances in digital and AI tools present new opportunities to improve impact measurement. Techniques like optical character recognition (for digitising documents), natural language processing and large language models (for analysing text at scale), knowledge graphs (for integrating diverse datasets), and robotic process automation (for automating repetitive tasks) could each help tackle aspects of the problem. However, the literature does not yet describe an integrated solution that

combines these technologies into a cohesive system tailored to the nonprofit context and call for research AI for good. This gap represents an opportunity for research to design and test a unified AI-driven framework.

Third, prior studies indicate that even a sophisticated tool will fail if it is not accepted and trusted by its intended users. Many technologies interventions falter when they overlook ease of use, user engagement, or ethical concerns. Thus, any proposed solution must incorporate human-centred design and responsible AI principles from the start, ensuring the system is intuitive, transparent, and fair so that practitioners will actually adopt it.

In response to these gaps, our study proposes and evaluates a novel solution that integrates AI into impact measurement while remaining mindful of both technical and human factors. Specifically, it aims to deliver three key contributions: (1) a practical AI-enhanced impact measurement framework tailored to nonprofit needs; (2) a documented case study demonstrating this framework in a real philanthropic context; and (3) design insights that translate user adoption models and ethical AI guidelines into concrete implementation strategies. By doing so, the research provides a tangible tool to improve practice and also contributes to the academic discourse with a real-world example of responsible AI integration in the social sector. These contributions set the stage for the subsequent chapters, which detail the methodology, design, and evaluation of the proposed framework.

3 Research Methodology – Design Science Research Approach

The research design of this study follows an adaptation of the Design Science Research (DSR) methodology (Hevner et al. 2004)[1]. DSR is appropriate for research that builds and evaluates an artifact to solve an identified problem, ensuring both rigor and practical relevance. In our case, we addressed the real-world problem of fragmented, labour-intensive impact reporting by designing an AI-based framework and then evaluating its effectiveness with stakeholders. This chapter details how we carried out each stage of the DSR process – from initial problem exploration and data collection through artifact development, demonstration, and evaluation – all while upholding ethical research practices and methodological rigor.

3.1 Research Design Overview

The DSR process model by Peffers (Peffers et al. 2007) structures the research. This model consists of six key stages, which our project executed as follows:

1. **Problem Identification & Motivation:** We clearly defined the core problem — philanthropic organisations' struggle with scattered data and time-consuming impact measurement — and articulated why it is important to solve. Initial stakeholder interviews in our partner organisations provided first-hand evidence of these challenges (e.g. data stored in silos, inconsistent metrics and formats, delayed reporting), reinforcing the need for a better solution. The literature review further confirmed a gap in leveraging AI for impact measurement, which motivated us to pursue an AI-driven framework.

2. **Define Solution Objectives:** Based on the problem analysis, we determined the objectives and requirements that a successful solution should meet. These objectives were derived from practitioners' pain points (for example, the need to automate tedious data aggregation and to provide more timely analytics) and informed by opportunities noted in prior research (such as using OCR to digitise

[1] Hevner et al. (2004) established Design Science Research in IS, which seeks to extend human and organisational capabilities by creating and rigorously evaluating innovative artifacts to address important problems. This methodology, later formalised by Peffers et al. (2007), guided our study's artifact-building and evaluation process.

paper reports or NLP to analyse textual feedback). Defining these goals ensured the artifact would address real user needs while remaining technically feasible.

3. **Design & Development:** We iteratively designed and built the artifact — a prototype impact measurement framework. A conceptual architecture was drafted (see Chapter 4) and the core components were developed through multiple cycles of prototyping. In each iteration, a working version of the system (or a subsystem) was implemented and then checked against the defined objectives to make sure we remained on track. Feedback and new insights from each cycle were incorporated into subsequent iterations. This stage resulted in a functional proof-of-concept of the framework by the end of the development phase.

4. **Demonstration:** We demonstrated the prototype in a realistic scenario to show how it solves the problem in practice. For instance, we applied the framework to representative impact data of partner organisations (such as sample project reports) to illustrate its capabilities — automatically extracting key impact indicators, integrating data across projects, and enabling complex queries. This demonstration acted as a proof-of-concept, allowing stakeholders to see the artifact in action within a context relevant to their work.

5. **Evaluation:** We evaluated the artifact's performance and usefulness relative to the objectives. Qualitative feedback was gathered from stakeholders after the demonstrations (through evaluation interviews) to assess the framework's utility, ease of use, and alignment with ethical expectations. We also objectively compared the prototype's outcomes to the initial requirements (e.g. did the system successfully unify data sources? did it produce accurate and meaningful analytics?). This multi-faceted evaluation identified how well the artifact met its goals and what could be improved. (The detailed evaluation results are presented in Chapter 5.)

6. **Communication:** Finally, we communicated the problem, the artifact, and the findings to both academic and practitioner audiences. This thesis document itself is the main vehicle of communication, capturing the design rationale, implementation details, and evaluation outcomes. By disseminating our results (including planned future presentations or publications), we aim to share the knowledge gained with others in the field and contribute to the broader discourse on information systems innovation in the nonprofit sector.

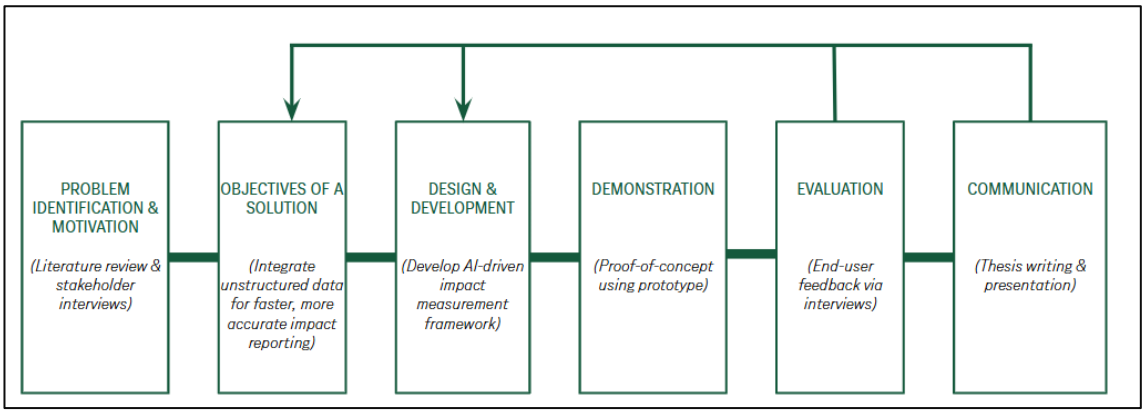


Figure 1 Iterative cycle of the DSR methodology

These stages are depicted as an iterative cycle rather than a strictly linear sequence. In practice, we embraced the iterative nature of DSR: for example, early feedback from stakeholders during the evaluation stage looped back to inform refinements in the design. Given the time constraints of a thesis project, we completed essentially one full cycle of the DSR process; however, we incorporated iteration by conducting two rounds of data collection (an exploratory phase and a subsequent evaluation phase) and by revisiting design decisions based on what we learned. This structured yet flexible approach ensured a strong link between the problem identified and the solution developed.

With this DSR framework guiding our approach, the methodology can be described in two main parts: (1) data collection and analysis to ground the design in real-world insights (and later evaluate the artifact), and (2) artifact development through iterative prototyping. The following sections describe these components in detail.

3.2 Data Collection Methods

Given the exploratory and evaluative needs of our study, we employed a primarily qualitative approach to data collection. We conducted semi-structured stakeholder interviews in two key phases, aligned with the early and late stages of the DSR cycle, and supplemented these with document analysis and literature review for triangulation. In all data collection activities, participants gave informed consent, and anonymity was maintained to ensure ethical standards.

Phase 1: Exploratory Interviews (Problem Exploration). In the first phase, corresponding to problem identification and requirements definition, we carried out semi-structured interviews to gather insights into current practices and challenges. Five staff members from partner philanthropic organisations were interviewed, selected to represent a range of roles from program manager to senior management. Each interview (approximately 45–60 minutes) followed an open-ended guide covering several themes: how impact data is currently collected and reported (to pinpoint pain points like manual effort or data silo issues), attitudes toward using technology and AI in their workflow (to gauge openness or concerns about an AI-driven tool), any ethical or trust considerations (such as worries about data privacy or algorithmic bias), and the interviewees' "wish list" for an ideal solution (features or improvements they would value). These conversations provided rich, first-hand information on the day-to-day difficulties in impact measurement and what users would want from a new system. Common problems raised included fragmented data storage across multiple files, inconsistency in metrics and formats, and the significant time required to prepare impact reports.

All Phase 1 interviews were recorded (with permission) and transcribed for analysis. We then performed a thematic analysis of the transcripts, following the approach of Braun and Clarke[2] (Braun and Clarke 2006). This involved coding the data for key issues and ideas, grouping similar codes into broader themes, and iteratively refining those themes. For example, if multiple interviewees mentioned difficulties in aggregating data from different sources, those remarks were coded under a theme of

[2] Braun & Clarke (2006) define thematic analysis as "a method for identifying, analysing, and reporting patterns (themes) within data," offering a flexible 6-phase process for qualitative coding. We applied their recommended steps (familiarisation, coding, theme development, review, definition, and reporting) to extract salient themes from the interviews.

data fragmentation; comments about having to manually compile reports fell under a labour-intensity theme. We compared these emergent themes with the challenges identified in the literature (Chapter 2) to check for alignment or any gaps. Indeed, many practitioner-raised issues (such as siloed data and delayed reporting) echoed issues documented in the nonprofit IT literature, which strengthened our confidence that the problem was well understood. The outcome of this analysis was a clear set of user requirements and design objectives for the artifact. For instance, the Phase 1 findings indicated the need for a centralised repository for all impact information, tools for automatically extracting data from reports, and an interface that non-technical staff could use to query data easily. These requirements directly informed the design goals and features we pursued in developing the prototype (see Chapter 4).

Phase 2: Evaluation Interviews (Artifact Assessment). After building the prototype, we proceeded to a second phase of interviews aimed at evaluating the artifact in practice. This evaluative phase involved a smaller number of interviewees, chosen for their ability to provide informed feedback on the working system. We conducted interviews with three key stakeholders: (1) an AI/Knowledge Management expert familiar with data systems, to offer a technical perspective on the framework's design and feasibility; and (2) two Senior Managers (decision-makers) from partner organisations, to provide a user-centric and strategic perspective on the artifact's usefulness and acceptability. Each evaluation session lasted around one hour. We began by giving a live demonstration of the prototype's functionality — walking the participant through how the system could ingests data, organises information in the knowledge graph, and allows querying for insights.

Following the demonstration, we used a semi-structured interview format to gather feedback, structured around key evaluation criteria, covering:

- **Perceived Usefulness:** Whether the system provides real value in solving their problems (e.g., saving time or improving reporting quality).
- **Ease of Use:** Whether the system is intuitive to use, and what learning curve or usability issues they anticipate.

- **Implementation Feasibility:** What it would take to implement the system in their organisation (e.g., technical integration needs, staff training, potential resistance).

As with Phase 1, we recorded and transcribed all feedback with consent.

We analysed the Phase 2 interview data using the same thematic coding approach, categorising feedback according to the criteria above. This allowed us to see how the artifact performed on each dimension. We looked for points of consensus and divergence: for example, if all interviewees agreed that the tool can significantly reduce the time needed to compile reports, that was strong evidence of usefulness; if the expert praised the system's data model but one of the managers founds the interface confusing, that indicated a usability issue to address. We also noted all specific suggestions for improvement (such as adding a particular visualisation feature or simplifying a workflow step), as these are valuable for refining the design. The insights from these stakeholder evaluations are summarised in Chapter 5. In short, the Phase 2 feedback highlighted the prototype's strengths (e.g. its ability to integrate data from disparate sources was appreciated) as well as areas for enhancement (e.g. the need for a more user-friendly front-end), guiding recommendations for the next iteration of the artifact.

Additional Data Sources: Alongside interviews, we utilised other data sources to inform the design and evaluation. We reviewed relevant organisational documents provided by a partner (such as past impact reports, data collection templates, and strategy papers) to ground our understanding of the existing context. Analysing these documents revealed what kinds of metrics and information the organisation typically collects and how they report outcomes, as well as the formats in which data is stored. This ensured our framework would handle the actual metrics and mix of qualitative/quantitative data the organisation uses. Document analysis also highlighted the variety of formats (PDF reports, spreadsheets, etc.) our data ingestion processes needs to accommodate (for example, reinforcing the need for OCR capabilities for scanned reports).

We also continuously consulted the academic literature (Chapter 2) throughout the project to reinforce our design decisions. For example, prior research on NLP techniques informed our text-processing features, TAM literature reinforced the importance of a simple user interface, Human-centered Artificial Intelligence and AI

ethics guidelines influenced us to incorporate transparency, privacy safeguards and boundaries in the design. Triangulating insights from interviews, documents, and literature ensured our artifact design was evidence-based and holistically informed, and it reflects DSR's emphasis on grounding the artifact in a well-understood problem context. This multi-source approach adds rigor: by reducing reliance on any single source of data, it increases the validity of our findings and helps ensure the artifact is well-aligned with both user needs and established theory.

Note on Transcript Confidentiality: Due to the limited number of interviewees involved in the study, some of whom are colleagues or otherwise professionally connected, it is not feasible to include direct transcript excerpts without potentially compromising participant anonymity and confidentiality. Maintaining strict anonymity and confidentiality was a critical ethical commitment made to participants, especially important given the close professional relationships among them *(More details are provided in Appendices B & E).*

3.3 Artifact Development Method

The development of the AI-driven impact measurement artifact followed an iterative prototyping strategy consistent with DSR principles. Rather than attempting to build the final system in one go, we progressed through multiple cycles of design, implementation, and refinement. In each iteration, we developed a functional prototype or a component of the framework, then tested it against the requirements and gathered feedback (either through self-testing or informal stakeholder input) and used those insights to improve the next version. This cyclical process allowed us to learn and adjust continuously, ensuring that the artifact evolved in step with our deepening understanding of the problem and user expectations. It also meant that problems could be identified and addressed early — before we committed to a large-scale design that might miss the mark. This approach embodies the DSR ethos of *learning through building* and helped maintain a tight feedback loop between design and evaluation.

During development, we leveraged **existing technologies and tools** whenever appropriate to accelerate progress, ensure reliability, and focus our efforts on the innovative aspects of the framework. For example, we used a private OCR software to handle text extraction from documents (avoiding the need to write our own OCR from scratch), and we utilised the Protégé ontology editor for building the domain ontology

(taking advantage of its robust features for ontology management). For the data storage and querying layer, we selected GraphDB (a stable semantic graph database) as our knowledge graph repository. These tools were chosen because they are well-established in their respective areas, which lent credibility and robustness to our implementation. By assembling the artifact from these building blocks, we ensured that each component was grounded in solid technology, allowing us to concentrate on integrating them in a novel way to meet our research goals.

A cornerstone of our artifact's design was the creation of a domain ontology to formally model the key concepts and relationships in philanthropic impact data. Guided by insights from the literature, organisational documents and the Phase 1 interviews, we identified essential entities that the system needs to represent — such as *Project*, *Outcome*, *Indicator*, *Beneficiary*, *Location*, and *Funder*. We defined how these entities relate to each other in the context of impact measurement. For example, a **Project** *achieves* certain **Outcomes**, each **Outcome** is *measured by* one or more **Indicators**; projects have associated **Locations**, **Beneficiaries**, and **Funders**. We also allowed linking **Outcomes** to broader goals (e.g., relevant Sustainable Development Goals) to connect project-level results to higher-level objectives. The ontology was encoded using standard semantic web languages (OWL/RDF), ensuring interoperability with semantic tools and adherence to best practices. We kept the model as lean as possible in the first iteration, including the core classes and properties needed to cover our case without over-complicating the schema. This cautious scope helped avoid unnecessary complexity at the start. We refined the ontology in an iterative manner: early versions were tested with sample data and reviewed, and we adjusted definitions or added elements if we found that certain real-world data points did not fit neatly into the structure. The final ontology for the prototype comprises 7 main classes and their key relationships and attributes. This semantic schema became the backbone of the artifact: it dictates how all impact data is represented in the system, enabling different sources of information to be integrated under a common vocabulary.

With the ontology in place, we built the knowledge graph that serves as the unified data store for the framework. We instantiated an RDF triplestore using GraphDB and loaded our ontology to enforce the data schema. We then populated the knowledge graph with a small test dataset to validate that the system can represent and interlink impact information correctly. This dataset consisted of a few fictional project entries with their

associated outcomes, indicators, locations, and funders, designed to mirror realistic scenarios encountered by philanthropic organisations. We deliberately included some overlapping elements across the sample projects (for instance, two projects sharing a common indicator or goal) to ensure the graph could capture connections between projects, as well as distinct elements to verify it could differentiate contexts. We then tested the knowledge graph's capabilities by running example queries (e.g., finding all projects addressing a certain SDG, or identifying which projects shared a particular indicator) and inspecting the results. These tests confirmed that the integrated data could answer cross-project questions and reveal connections that would be hard to see if the data were not unified. This validation demonstrated that our core architecture (the ontology plus graph database) was functioning as intended and provided a sound foundation for the next steps.

In parallel to building the core system, we explored an automated data ingestion approach to see how unstructured data (like free-text reports) might be fed into the knowledge graph in the future. In a small experiment, we took a sample project report document and applied OCR to convert it into machine-readable text. Then, we used a large language model (GPT-4 via an API) to parse the text and extract key information corresponding to our ontology (for example, identifying mentions of specific outcomes achieved or quantitative results reported). This exercise showed that an AI could potentially assist in populating the knowledge graph. We did not integrate this capability into the current prototype (due to time constraints), but the experiment informed our vision for a future automated data pipeline and highlighted challenges to address (like verifying AI-extracted information). For the present prototype, data was entered manually or via simple scripts, but this trial helped us anticipate how a more advanced ETL (Extract-Transform-Load) process with AI support could work in subsequent iterations.

Throughout the development process, we remained mindful of end-user interaction with the system, even though creating a full user interface was beyond the scope of this thesis. We conducted informal usability thought exercises by using GraphDB's interface ourselves to simulate typical user queries. For instance, writing SPARQL queries underscored that non-technical users would need a far more intuitive query interface (e.g., a simple dashboard or guided search). Our experience also highlighted the value of visualising relationships (tracing from a broad outcome down to specific project

data), suggesting that future interfaces should include features for easy navigation of the graph's information. While we did not implement a custom front-end, we sketched out ideas for a user-friendly dashboard and even considered how an AI-driven assistant (e.g., a chatbot) might allow users to ask questions in natural language and retrieve answers from the system. Importantly, we designed the artifact's architecture to be modular and extensible, making it possible to plug in such components later. For example, the knowledge graph and ontology layers operate as a back-end service that can support multiple front-end applications, and by using standard query languages (SPARQL) we ensure that an AI assistant or other interface could interact with the data without changing the core system. By anticipating these needs, we ensured the artifact's core architecture could accommodate such additions without major redesign.

In summary, our artifact development method delivered a working prototype that serves as a proof-of-concept for AI-assisted impact measurement. By iterating through design-and-build cycles and leveraging reliable tools, we were able to create a solid foundation — a unified knowledge base with semantic querying capabilities — within the project's timeframe. Each decision in the development was guided by both theoretical frameworks (for instance, applying TAM principles to consider user-friendliness, and adhering to ethical AI guidelines for transparency) and stakeholder input (reflecting users' requirements and concerns). This integrative, iterative approach ensured that the resulting system is not just technically sound but also aligned with the needs and values of its intended users.

Demonstration and Evaluation Strategy: Finally, as part of the DSR methodology, we planned how to demonstrate the artifact and evaluate its success. For demonstration, we prepared a realistic use-case scenario in collaboration with partner organisations to showcase the prototype in action. This involved using actual or representative impact data to simulate how the framework would operate in a real organisational context. The demonstration (detailed in Chapters 4 and 5) allowed us to verify that the artifact could indeed address the identified problem when applied to real-world data.

For evaluation, we established clear criteria and procedures to assess the artifact against its objectives and to capture stakeholder reactions. As noted above, the Phase 2 stakeholder interviews (with an AI/Knowledge Management expert and two senior managers) were central to our evaluation, focused on the core questions of usefulness,

ease of use, trust, and feasibility. In addition, we systematically checked the system's capabilities against each major design requirement from Chapter 4. For example, in this first iteration one requirement was to "integrate data from multiple sources", so we examined whether the knowledge graph successfully combined information from different documents during the demonstration; another goal was to "enable cross-project queries," we verified that such queries were possible and yielded meaningful results. Any requirement that the prototype only partially met or did not meet was flagged for future improvement.

Throughout the evaluation, we adhered to ethical best practices and took steps to ensure the validity of our findings. All evaluation participants provided informed consent, and their feedback was anonymised to protect privacy. In the end, we could judge not only whether the system works technically but also whether it would likely be accepted and effective in the organisation's context. These evaluation outcomes, which close a loop of the DSR cycle by linking back to the original problem, presented in Chapter 5.

4 Design and Development of the AI-Driven Impact Measurement Framework

In this chapter, we detail the design and development of the AI-driven impact measurement framework – the artifact resulting from this research. Guided by earlier findings and stakeholder input, we set out key design objectives, propose a four-layer conceptual architecture, and implement a functional prototype. An illustrative use-case scenario demonstrates how the framework operates in practice. (Detailed technical specifications, such as the full ontology, example RDF triples, and SPARQL queries, are provided in the appendices for reference.)

4.1 Design Requirements and Objectives

Based on the challenges identified in current impact measurement practices (and user needs gathered in previous chapters), we defined three primary design requirements to guide the framework's development. These requirements address data integration and analysis issues while embedding usability and ethical principles to ensure the solution is practical and acceptable to end-users:

- **Unified Data Integration:** Consolidate fragmented impact data from multiple sources into a single repository. The framework should ingest diverse formats (text reports, spreadsheets, databases, etc.) through an ETL process, creating one "source of truth" knowledge base. This integration reduces siloed data and manual effort in compiling reports, ensuring all projects' data can be accessed in one place.

- **Semantic Cross-Project Querying:** Enable rich queries across projects and datasets. Users should be able to ask complex questions that span multiple projects, time periods, or themes and get aggregated answers. Achieving this requires storing data with explicit relationships (e.g. linking projects to their outcomes, indicators, locations) so that a single query can traverse all relevant data. A knowledge graph with a well-defined ontology was chosen to fulfil this need, as it allows semantic links between data points and can answer questions that would be tedious or impossible with disconnected documents or spreadsheets.

- **Human-Centered and Ethical AI:** Design the system as a decision-support tool that augments (rather than replaces) human work, adhering to responsible AI guidelines. The framework must be intuitive and fit user workflows (maximising **usability** for adoption), handle data ethically (ensuring **fairness** and **privacy**, e.g. protecting any sensitive information and avoiding bias in AI analyses) and provide transparency for **accountability** (so users can trace each result back to its source). These principles ensure users trust the system and feel in control – a critical factor for adoption in a resource-constrained, trust-sensitive context.

These requirements guided the framework's design. They directly address the core problems (integrating data silos and enabling cross-project analysis) while ensuring the solution remains user-friendly and trustworthy. Next, we describe the high-level system architecture developed to meet these goals.

4.2 Conceptual Framework Architecture

To satisfy the above requirements, we designed the framework as a layered architecture with modular components. Organising the system into distinct layers allows each to handle specific functions (and to be improved independently) while still working together as a whole.

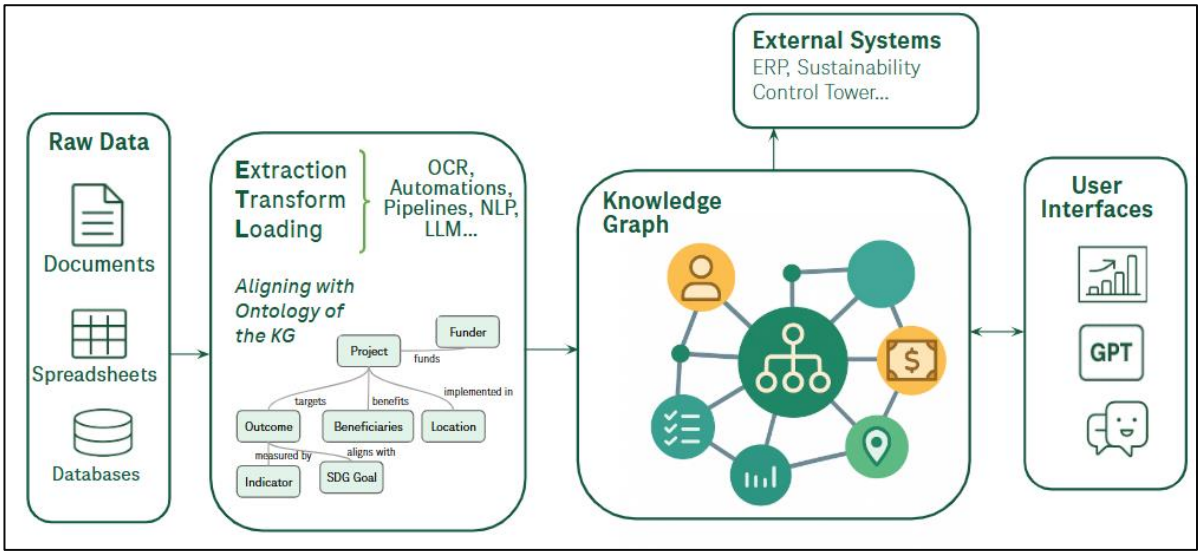


Figure 2 Conceptual Framework

Figure 2 illustrates the four main layers of the architecture, described below:

1. **Data Integration & ETL Layer:** This bottom layer handles importing and preprocessing data from diverse sources into the system. It comprises Extract-Transform-Load processes to gather raw data (e.g. text from reports, spreadsheet tables, database records) and convert it into a unified structured format. For example, an OCR+NLP pipeline could extract key facts from PDF reports, and scripts could normalise spreadsheet data (standardising date formats, units, and terminology) to align with the framework's schema. The ETL layer funnels all these heterogeneous inputs into the common data model. In our prototype, a fully automated ETL is not yet implemented (data was prepared manually for the demo), but the architecture is designed to accommodate automated data pipelines in the future. This layer is crucial for Unified Data Integration, ensuring that no matter how varied the incoming data, it ends up in a single integrated knowledge repository.

2. **Ontology Layer:** At the heart of the system is the ontology – a formal semantic schema defining the key entities and relationships for impact data. This layer provides a shared vocabulary for projects, outcomes, indicators, locations, funders, etc., and how they interrelate (e.g. a **Project** *achieves* certain **Outcomes**; an **Outcome** *isMeasuredBy* an **Indicator**; a **Project** *alignsWith* an **SDG Goal**, *implementedIn* a **Location**, *fundedBy* a **Funder**, and so on). By enforcing a consistent data model, the ontology ensures that all data adheres to the same structure and meanings. This consistency enables data from different projects to be compared and aggregated, directly addressing data inconsistency issues. Moreover, the ontology encodes domain knowledge that allows for inference – for instance, if the ontology knows **Kenya** is in **Africa**, a query for projects in Africa can automatically include projects located in Kenya. We developed the ontology using standard OWL/RDF technologies. *(see Appendix D for the full ontology specification)*

3. **Knowledge Graph Layer:** This layer is the central data repository – the knowledge graph itself, stored in an RDF triplestore. After data is processed by the ETL layer and structured according to the ontology, it is loaded as a collection of RDF triples (subject–predicate–object statements) into the graph database (for our prototype we used GraphDB). The knowledge graph holds all integrated information and can be queried using SPARQL to retrieve insights. Because all project data and their connections live in one graph, users (or applications) can ask questions that join across what were once separate silos. For example, a single query can find "all projects in region X addressing outcome Y" by traversing the relationships in the graph, rather than manually merging spreadsheets. The graph database also supports "reasoning": given a certain ontology, it also could infer new facts (e.g. deducing a project is in Africa because its location is Kenya, as per the ontology's hierarchy). In the prototype, we kept the ontology and dataset relatively simple, so heavy reasoning wasn't required, but the infrastructure allows adding more complex logic as needed. An important advantage of the knowledge graph is interoperability – external tools or systems can access the data via standardised endpoints. For instance, the graph's SPARQL endpoint or an API could feed a business intelligence dashboard or integrate with other datasets. This ensures the framework's data is not a dead-end silo but a hub that can exchange information. By centralising all data and logic here, we also guarantee everyone uses the same consistent data; any analytical tool or interface draws from the one unified source (preventing the discrepancies that occur when different teams use separate data copies).

4. **Application Layer:** The top layer includes all user-facing applications and interfaces through which end-users interact with the framework. Its purpose is to deliver the knowledge graph's insights in an accessible way and to allow users to query or input data without needing technical expertise. In our conceptual design, this could take multiple forms – for example: (a) a web-based dashboard that visualises key metrics (showing how many projects are active, their distribution across SDGs or regions, aggregated outcome indicators, etc., with filters for the user to slice the data by year, country, theme, etc.), and (b) a natural-language query assistant (chatbot-style interface) that lets users ask

questions in plain language *(e.g. "Which projects in 2020 had the highest number of beneficiaries?")* and receive answers with relevant data and source citations. These front-end tools would translate user actions or questions into queries to the knowledge graph behind the scenes. Although building a full application layer was beyond the scope of the prototype, we planned for it in the architecture. We ensured, for example, that the ontology includes human-readable labels and that provenance information is stored, so a future interface can display friendly names and allow users to drill down to original sources ("show me where this number came from"). The layered design means these interface components can be developed later and plugged in without changing the underlying data layer. In essence, the Application layer is what will fulfil the usability and human-in-the-loop requirements – by providing intuitive access to the data (through visuals or conversational queries) and enabling users to trust and act on the insights (via transparency features like explanations and source links).

Overall, this four-layer architecture (ETL, Ontology, Knowledge Graph, Application) provides a modular and scalable structure that meets our design objectives. Data flows up from raw sources to the integrated graph, and users interact through applications that draw on the graph's unified intelligence. The layers also allow flexibility: one can improve or swap out components (e.g. use a more advanced NLP extractor in the ETL layer, or upgrade the database technology) without overhauling the whole system, as long as interfaces between layers adhere to standards. This architecture thus lays a solid foundation for our solution, addressing the need to **unify data**, enable **semantic queries** and **traceability**, and incorporate **user-centric design** principles.

4.3 Prototype Implementation

To validate the proposed design, we built a proof-of-concept prototype focusing on the core layers (ontology and knowledge graph) and basic query capabilities. This prototype demonstrates key functionalities on a small scale and provides a basis for gathering feedback for future enhancements. Here we summarise the implementation steps *(detailed technical artifacts are provided in Appendix C & D)*:

- **Ontology construction:** We developed the impact ontology using Protégé, defining the main classes and properties as outlined in the architecture. Key

entity classes included **Project**, **Outcome**, **Indicator**, **Beneficiary**, **Location**, **Funder**, and **SDG_Goal**. Key relationships were implemented as object properties (e.g. *achieves*, *isMeasuredBy*, *implementedIn*, *fundedBy*, etc., corresponding to the conceptual model). We also added a few data properties (such as *hasStartDate*/*hasEndDate* for project timelines or *has number of beneficiaries*). The final ontology contains seven core classes and a similar number of primary relationships, sufficient to model the essential aspects of impact data. We ensured the ontology was logically consistent (Protégé's reasoner was used to check for errors) and exported it in a standard OWL/Turtle format. *(A complete ontology definition is provided in Appendix D.)*

- **Knowledge graph setup and data loading:** We deployed an instance of GraphDB (an RDF triplestore) to serve as the knowledge graph layer. The ontology was loaded into GraphDB to act as the schema for the data. We then prepared a small **demo dataset** of impact information to populate the graph. This dataset included five fictional projects (call them *Project 1*, *Project 2*, *Project 3…*) along with related outcomes, indicators, and other entities. The data was crafted to mirror a realistic use case (for example, two projects in Africa aligned with SDG 4, sharing a common indicator, etc.). We encoded the dataset in RDF (using Turtle syntax for brevity) and imported it into the GraphDB repository. This resulted in a set of interconnected triples representing the projects and their relationships. For instance, Project 1 is linked to *Outcome "Improved Health Access"*, which is linked to *Indicator "Mortality Rate Reduced* and *SDG 3 (Good Health and Well-being)*, and *Project 1* is also linked to its location (Kenya), funder, and beneficiary group. *Following projects* were similarly added, reusing certain entities where appropriate (e.g. Projects 2 and 5 both are related to SDG 4 (Quality Education) *(A snippet of the RDF triples for the demo dataset is provided in Appendix C as an example.)*

- **Query testing and functionality verification:** After loading the data, we used GraphDB's SPARQL query interface to test that the system could answer the kinds of questions we envisioned. For example, we executed queries to list all projects and their associated outcomes and indicators, and a specific query to retrieve "all beneficiaries per project and per type". We also tested traceability queries: given indicators (e.g. *Literacy Rate*) or SDGs, fetch all projects that use

it and the description of the outcome it measures – it successfully returned the relevant projects and outcome descriptions from our data, demonstrating that one can trace a metric across projects. These tests validated that the framework could handle cross-project queries and link related information as intended. *(For reference, SPARQL queries and its outputs are included in Appendix C.)*

- **Prototype scope and limitations:** It should be noted that the prototype's user interface was minimal – we primarily interacted with the knowledge graph through the GraphDB workbench and simple scripts to simulate what an application layer would do. Features like a full dashboard UI or a conversational assistant were not fully implemented due to time and scope constraints. However, the prototype was designed with these future extensions in mind. For instance, we stored human-readable labels and source references in the data so that a front-end could display names and allow source tracing. Similarly, while our data ingestion for the demo was manual, the process mimicked what an automated ETL module would do, illustrating feasibility for later development. In summary, the prototype confirmed the feasibility and value of the knowledge graph approach on a small scale: it successfully unified data from multiple sources and answered complex queries with traceable results. These core functions lay the groundwork for building out the remaining features (automated data pipelines, interactive dashboards, AI assistants) in future iterations.

5 Demonstration and Evaluation

This chapter presents a demonstration of the prototype in a realistic organisational setting and an evaluation of its performance with input from stakeholders. Section 5.1 describes how the AI-driven framework was showcased using a plausible scenario and dataset, illustrating its key functionalities in practice. Section 5.2 then synthesises feedback from stakeholders who observed and interacted with the prototype, focusing on the tool's perceived usefulness, ease of use, trustworthiness, and implementation feasibility. Finally, Section 5.3 reflects on how well the artifact met its intended objectives and discusses its limitations, providing an academically rigorous assessment of the prototype's utility and areas for improvement.

5.1 Demonstration in Organisational Context

To evaluate the practical applicability of the proposed human–AI collaboration framework, we demonstrated a prototype within partner philanthropic organisations. The demonstration focused on key functionalities of the framework – such as semantic querying across projects, integrating data from multiple sources, and ensuring traceability of information – using a carefully curated subset of data rather than a live production system. We assembled a small sample dataset that included a few structured records (e.g. selected project entries with associated indicators and outcomes) as well as representative unstructured content from partner organisations (e.g. an example project report in PDF form). This setup kept the scenario realistic while acknowledging that some processes were simulated due to the limited scope of the prototype.

We were transparent about which components of the system were fully implemented and which were mocked or conceptual. For instance, the framework's data ingestion pipeline (Extract-Transform-Load) was not automated at this stage. Key details from source documents were extracted by the researchers beforehand and pre-loaded into the knowledge graph, rather than being pulled in automatically by the system. By clarifying it, we avoided implying that a production-ready integration of all data sources was already achieved. Instead, the demonstration leveraged pre-loaded data to show how the system would operate once an automated integration is in place. To compensate for the incomplete ETL implementation, we introduced an external AI tool to simulate the kind of text-processing module that the framework would eventually include.

Specifically, we employed *Google NotebookLM* – an experimental AI-powered research assistant – as a stand-in to illustrate automated extraction of information from unstructured documents. During the demonstration, we uploaded an example PDF project report to NotebookLM and queried it for relevant details (for example, asking for a summary of the report's key outcomes and beneficiary numbers). The AI assistant parsed the document and returned a concise summary with references to the original text, effectively mimicking how our framework could use OCR/NLP techniques to pull facts from raw reports. This exercise demonstrated that even though our prototype lacked a built-in OCR/NLP component, existing AI technology can fulfil that role: NotebookLM accurately identified outcome indicators and impact figures from the report, and it provided source citations for transparency. This outcome reinforced the viability of integrating such an AI assistant into the framework's pipeline in the future, aligning with our design principle of traceability (users can see exactly where each piece of data came from).

In parallel with the unstructured data example, we showed how the system handles structured data queries through the knowledge graph. Using the prototype's query interface (built on a graph database with SPARQL capabilities), we executed queries addressing real information needs identified by the organisation. For instance, we demonstrated a query to retrieve "all education projects in African countries that contribute to SDG 4 (Quality Education) and the indicators they use." The system returned results within seconds, listing the relevant projects along with their associated outcome indicators. These results were visualised using a node-link graph view, helping the audience intuitively see relationships – for example, highlighting that multiple projects shared a common metric (a literacy rate indicator) and even a common donor. This scenario illustrated how the integrated framework can connect disparate data points (projects, goals, locations, metrics) on demand, something that would be tedious and time-consuming using the organisation's traditional manual methods. Stakeholders observing the demo noted that what normally required sifting through numerous documents was accomplished almost instantly via the unified query, underscoring the potential efficiency gains.

Throughout the demonstration, we emphasised which parts of the workflow were fully functional, and which were prototypes or mock-ups. This candid approach maintained academic rigor and managed expectations, ensuring the stakeholders understood the

prototype's current capabilities versus its aspirational features. Overall, the controlled demonstration successfully showcased the framework's core value proposition in a plausible context: it illustrated how combining a semantic knowledge base with AI-driven tools could streamline impact data management. Even with certain features simulated (such as the automated data ingestion), participants could clearly envision the framework's promised end-to-end functionality – from ingesting a raw report to querying consolidated insights (*For technical details illustrating the demonstration, see Appendix C.).* This sets the stage for the next section, where we report on stakeholders' impressions and evaluation of the prototype following the demonstration.

5.2 Stakeholder Feedback and Evaluation Findings

Following the demonstration, we conducted in-depth interviews with three key stakeholders: two senior managers from partner organisations (a potential end-user of the system) and an AI/knowledge management expert familiar with such technologies. These semi-structured interviews aimed to evaluate the prototype's usability, usefulness, ethical acceptability, and fit with the organisation's needs. Instead of listing each individual comment, we summarise the feedback thematically, grouping insights into major categories. This thematic analysis provides a clear picture of how stakeholders perceived the artifact's strengths and weaknesses.

Utility and Relevance: All stakeholders indicated that the AI-driven framework addresses pressing needs in the organisation's impact reporting process. They agreed that the system would alleviate a major pain point by consolidating information that is currently scattered across multiple reports and files. This consolidation was seen as likely to save significant time when compiling data for donors or strategic reviews. The framework was also praised for its potential to generate new insights: by enabling cross-project and cross-program analysis (which was previously impractical), the tool could help identify patterns such as common success factors or underperforming areas across an organisation's portfolio. In terms of alignment with organisational goals, the stakeholders felt the framework directly supports the mission to use data for learning and accountability. For example, having up-to-date outcome metrics "at their fingertips" would inform evidence-based decision making at the management level. That said, two interviewees cautioned that the system's adoption would hinge on clear practical value rather than novelty. In their view, the prototype must demonstrably improve existing

processes – for instance, by *truly* reducing manual work or improving the quality of impact reports – to justify the effort of implementation. This emphasis on tangible usefulness echoes established technology acceptance theory, which posits that users are more likely to adopt a new system if it clearly enhances their job performance. In summary, the feedback on utility was very positive: the stakeholders saw the artifact as highly relevant and potentially transformative for their impact measurement tasks, provided it delivers real efficiency and insight benefits in practice.

Usability and User Experience: Overall, the prototype's user interface and interaction design were received as promising, especially given the complexity of the technology behind it. The non-technical managers were able to follow the demonstration without confusion, suggesting that our design choices (such as using plain-language filters, intuitive dashboards, and explanatory tooltips) are on the right track for an audience with limited IT expertise. All stakeholders did, however, stress that adequate training and user support would be crucial for successful adoption. They recommended offering workshops or manuals to ensure all staff – including those less accustomed to data software – can become comfortable with this tool. The AI expert added that even an excellent system benefits from a structured onboarding process, proposing that early pilot users could act as in-house champions to help train their colleagues. In terms of interface functionality, one area for improvement concerned advanced querying. At present, the prototype requires manually constructing certain complex queries (e.g. writing a SPARQL query to answer a very specific question), which typical end-users are unlikely to do. The stakeholders suggested integrating more user-friendly query mechanisms, such as natural language query capability or a set of pre-built query templates for common questions. This recommendation aligns with our roadmap to incorporate an AI assistant for querying, allowing users to ask questions in plain English instead of dealing with technical query languages. In short, the stakeholders found the prototype easy to use for basic tasks and were impressed that no coding was needed for the core features. However, they highlighted that achieving a polished user experience will require simplification of advanced tasks and robust user education. Ensuring high perceived ease-of-use will be as important as usefulness in driving adoption across the organisation.

Trust and Ethical Acceptability: A significant portion of the feedback focused on whether users would trust the AI-generated outputs and how the system aligns with

ethical norms. Two of the three stakeholders were cautiously optimistic about relying on the tool's analyses, but they underscored that transparency is absolutely critical to earn and maintain trust. In practice, this means users should be able to understand why the system produced a given result. The stakeholders reacted very positively to the prototype's traceability features that we demonstrated – for example, the ability to click on a statistic or insight and immediately see the original source data (the specific project report or document from which that piece of information was derived). A senior manager noted that being able to verify each figure back to its source is essential for credibility; they would be unlikely to trust an AI-based system for impact reporting if it did not provide such justification for its outputs. This feedback validates our emphasis on explainability and aligns with ethical AI principles discussed earlier in the thesis (e.g. the need for explicability and accountability in AI systems).

Data privacy and security were also raised as important concerns. Philanthropic organisations often handle sensitive information (such as personal data on beneficiaries or budget details), so the stakeholders wanted to know how the framework would protect such data. We clarified that, in our prototype, the knowledge graph is deployed in a secure environment and that any personally identifiable information can be anonymised or excluded as needed. In an actual implementation, strict access controls and compliance with data protection regulations would be put in place. A senior manager suggested establishing clear governance policies if the tool were adopted – for instance, guidelines on acceptable data use and an oversight process to periodically review the AI's recommendations – to ensure the technology is used responsibly and does not stray from the organisation's values.

Another ethical aspect discussed was the potential for algorithmic bias. The AI expert noted that if we incorporate machine learning or intelligent assistants more deeply into the framework, we must guard against biases in how information is processed or presented. All interviewees agreed that keeping a human in the loop is vital: the system should support human judgment, not replace it. A manager explicitly described the envisioned AI tool as *an assistant rather than a decision-maker*, reinforcing our human-centric approach to design.

In summary, the stakeholders found the framework's current approach to transparency and user control to be on the right track for building trust. They felt the tool would be ethically acceptable and welcomed the efficiencies it offers, so long as it continues to prioritise explainability, privacy, and human oversight. Their expectations closely mirror widely accepted guidelines for trustworthy AI (calling for transparency, fairness, data privacy, and human agency), indicating that our design principles align well with what end-users will require to feel comfortable using an AI-driven system in a sensitive domain like philanthropy.

Feasibility and Implementation Challenges: When reflecting on how the prototype could be rolled out in practice, the stakeholders provided a realistic assessment of implementation challenges and enablers. On the technical side, the expert stakeholder believed the concept is sound and built on mature technologies (such as OCR, natural language processing, and graph databases) that are already available. They did not identify any fundamental technical barrier that would prevent the framework from working at scale. However, they emphasised that data preparation could become a bottleneck. For the system to be populated, an organisation might need to invest significant effort into converting historical reports and files into the required digital format (combining automated OCR for paper/PDF documents with manual data cleaning for inconsistencies). This preparatory step could be time-consuming, especially for organisations with years of archival data, and would need to be planned for in any implementation project.

The stakeholders also discussed resource requirements. Successfully deploying this AI framework would likely require dedicated resources – for example, budget for software development or licenses (if using commercial AI services or cloud infrastructure), and possibly hiring or upskilling staff to maintain the system and manage the data. Smaller organisations or those with very limited IT budgets might find this challenging, so part of evaluating feasibility is assessing whether the expected benefits justify the investment in technology and training.

Integration with existing workflows and systems was noted as another critical factor. The framework should ideally connect to tools and databases staff already use to avoid creating extra work. For instance, if the organisation currently stores project data in a certain database or in SharePoint/Excel files, the AI system should be able to import

from those sources (or interface with them) so that staff do not have to duplicate data entry. Stakeholders felt that the closer the tool aligns with current processes (at least initially), the smoother its adoption would be. If the AI framework is too isolated or requires completely new data management processes, it could face pushback.

The human and organisational aspect of change management was a recurring theme as well. All interviewees anticipated some resistance to change among staff if a new system is introduced, which is a common dynamic in nonprofit environments where many employees are accustomed to established ways of working. There may be a learning curve and even scepticism about an AI-driven approach. To mitigate this, the stakeholders suggested securing *early wins* and buy-in: for example, start by piloting the system on one or two projects and use those successes (such as significantly faster reporting or interesting new insights discovered) as case studies to demonstrate the tool's value to the rest of the team. Showing concrete improvements can help convince others of the framework's usefulness and build momentum for wider adoption. Additionally, identifying champions or power-users within the organisation who are excited about the tool can help in peer-to-peer advocacy and training.

In summary, the stakeholders concluded that implementing the framework is feasible in a philanthropic organisation, but it requires careful planning beyond the technical deployment. An organisation considering this tool would need a clear implementation strategy that addresses data onboarding, user training, integration with current systems, and ongoing maintenance and support. Moreover, leadership commitment is needed to allocate resources and to foster an open mindset toward data-driven innovation. The feedback here highlights that success of the artifact will depend not just on its technical merit, but also on the organisational readiness and change management efforts that accompany its introduction.

Overall**,** the evaluation feedback from stakeholders was encouraging, pointing to considerable potential value in the AI-supported impact measurement framework. In their view, the prototype directly tackles real challenges in impact data management and could improve current practices substantially, assuming the refinements and support discussed above are put in place. Crucially, all interviewees indicated they would be open to adopting a system like this in their work, provided it demonstrably improves their workflow and maintains the transparency and ethical standards they expect. Their

insights validated many of our design decisions (e.g. focusing on usability and traceability) while also identifying areas for improvement (such as simplifying advanced queries and planning for integration, scalability and training). This stakeholder input will inform final adjustments to the prototype and guides our recommendations for scaling the framework in practice. With the qualitative evaluation complete, we next examine how well the artifact, in its current state, meets the research objectives and discuss the limitations that must be acknowledged.

5.3 Reflection on Artifact Utility and Limitations

Considering the demonstration and stakeholder feedback presented above, we now assess the artifact's overall utility and its limitations relative to the goals of this research. The central question for evaluation was whether the developed AI-driven framework helps improve impact data management and reporting in the philanthropic context. Based on the evidence gathered, our design artifact shows itself to be a **viable solution** to the identified problem, albeit at an early prototype stage with clear limitations that temper its scope.

On the positive side, the framework fulfilled its core purpose as envisioned. It successfully integrates information from multiple, previously siloed sources into a unified knowledge base, enabling more efficient retrieval and analysis of impact data. This directly addresses the problem of fragmented data and laborious manual aggregation identified at the outset of the thesis (recalling RQ1 from Chapter 2). The demonstration illustrated a concrete use-case of a manager obtaining cross-project insights in seconds – something that would have previously required manually compiling data from numerous documents over days. Stakeholders corroborated this utility by confirming that the tool targets real needs in their reporting workflow. They believed that, if implemented, the system would likely improve how impact is measured (by surfacing insights and patterns across programs) and how it is communicated (by making it faster and easier to compile credible reports with source-backed data). In terms of the specific requirements and objectives derived from our problem analysis and design (Chapters 2 and 4), the prototype met many of the key criteria: it enabled cross-project comparisons, provided traceability from high-level metrics back to original sources, supported flexible querying of the data, and incorporated ethical design principles like transparency and access control. This indicates a strong alignment

between the artifact's capabilities and the intended design objectives. Moreover, by involving actual end-users in the evaluation phase, we ensured that the artifact's relevance and usefulness were validated by more than just theoretical reasoning – this user feedback serves as an important checkpoint confirming that the solution can indeed deliver value in practice, not just in concept.

However, given that this is a prototype developed under academic time and resource constraints, there are several notable limitations to acknowledge. First, there are technical scope limitations: certain envisioned functionalities were only partially implemented or remain as future work. For example, as discussed, the ETL (Extract-Transform-Load) process for automatically ingesting data was not fully built out – much of the data integration in the demo was performed manually behind the scenes. Likewise, the idea of a conversational AI assistant for natural language querying exists in our design, but the current prototype only has a basic querying interface and relied on an external tool (NotebookLM) for a subset of those capabilities. In practice, this means the current artifact is a proof-of-concept rather than a deployable product. Users interacting with it today would not yet experience the seamless, automated data updating or the AI-driven Q&A functionality that a full implementation would aim to provide. These gaps between the prototype and the ideal feature set highlight where further development is needed. It is important to note that while these missing features limit the artifact's immediate functionality, they were consciously left out due to scope and are identified as priorities for future enhancement rather than fundamental flaws in the design.

Second, the evaluation scope was limited in scale. We gathered in-depth feedback from three stakeholders in a controlled demonstration setting. While their insights are highly valuable, the sample size is very small, and the evaluation did not include hands-on use of the system over an extended period or with a broader user base. Thus, our findings on user acceptance and performance are indicative but cannot be generalised to all potential users or scenarios. We did not, for instance, quantitatively measure improvements such as time saved or error reduction through a comparative study (e.g. comparing teams using the tool vs. not using it), which would be ideal to conclusively demonstrate efficiency gains. Those kinds of summative evaluations were beyond the scope of this thesis but remain necessary to rigorously quantify the benefits (for example, we hypothesised that the tool could reduce impact reporting time by a certain percentage,

but this was not empirically verified). Additionally, because the evaluation was short-term, we could not observe long-term adoption factors – such as whether initial enthusiasm for the tool would sustain over months of regular use, or whether any novelty effect would wear off, revealing new challenges. The participants' positive reception could be influenced by knowing they were part of a research pilot (which sometimes makes evaluators more forgiving or enthusiastic). We attempted to mitigate bias by encouraging candid feedback and critical perspectives, but some degree of positive response bias may still be present. In short, the limited and qualitative nature of our evaluation means the results should be interpreted as preliminary evidence of success, calling for larger-scale testing in future work.

Third, there are contextual and organisational limitations affecting how broadly our results can be applied. This research was essentially a single-case study co-developed with specific partner organisations that were moderately prepared and interested in data innovation. The prototype was somewhat tailored to their context (including the particular data they provided and their reporting processes). Therefore, the degree to which the framework would fit a different philanthropic organisation – especially one of a very different size, region, or technological maturity – is uncertain. For example, what works for a mid-sized European foundation in our case may not translate seamlessly to a small grassroots NGO or a large international charity with different infrastructure. The successful demonstration assumes certain enabling conditions such as the availability of digital data, willingness to invest in data processes, and openness to AI tools. If an organisation lacked these conditions (say, if they had very low digitisation or a culture averse to analytics), implementing this framework could face substantial hurdles or might need significant adaptation. In our case, the partner organisations' interests and basic IT infrastructures created a relatively favourable environment that likely biased the outcomes positively. In a less prepared environment, the challenges could be greater, and the impact of the tool might be more limited. These considerations remind us that while the artifact appears to work well in our study setting, broader generalisation requires caution and possibly further design modifications to suit other contexts.

Despite these limitations, the overall evaluation outcome is encouraging. The design science research cycle in this project has come full circle: we identified a significant problem in practice, proposed an innovative IT artifact to address it, and then evaluated

that artifact against real-world criteria of success (utility, usability, etc.). The evidence suggests that our artifact can indeed contribute to improving impact measurement practices in philanthropy. The stakeholders' feedback serves as an existence proof that such a tool can be acceptable and valuable to practitioners, even as it also highlights what would be needed to fully realise that value. In essence, the study demonstrates a feasible path forward for philanthropic organisations to leverage AI in their impact measurement efforts, confirming many of the anticipated benefits while also shedding light on the practical constraints and requirements. These findings will be taken into account in the next chapter, where we discuss the broader implications of the work. Before concluding the thesis, we use the insights gained to refine our understanding of AI's role in this domain, consider how user acceptance and ethical design played out, and outline recommendations for both practice and future research.

6 Discussion and Conclusion

This concluding chapter synthesises the findings of this study and discusses their implications for AI design, nonprofit impact measurement, and responsible innovation. It also outlines the key contributions of the research, acknowledges its limitations, and suggests directions for future work. Together, these elements provide a final perspective on what has been achieved and how it can inform both theory and practice moving forward.

6.1 Alignment with Existing Literature and Frameworks

Our findings confirm and extend several key theories from the literature. In line with Shneiderman's Human-Centered AI (HCAI) framework, our participants responded positively to a system that combined high human control with high automation(Shneiderman 2020). As Shneiderman argues, systems designed with *both* substantial human oversight and powerful computer assistance tend to be reliable, safe, and ultimately more adoptable(Shneiderman 2020). Similarly, our human-in-the-loop design was seen as useful and trustworthy by stakeholders, consistent with the HCAI claim that AI tools should function as "powerful, tool-like" aids rather than autonomous agents.

At the same time, classical technology-adoption theory is also borne out by our data. In line with Davis's (1989) Technology Acceptance Model (TAM), participants emphasized perceived usefulness and ease of use as critical to adoption. Our system's efficiency gains and intuitive interface echo Davis's finding that *ease-of-use influences adoption indirectly through usefulness* (Davis 1989). In sum, the core design outcomes of our project – a transparent, human-centered AI framework underpinned by knowledge integration – directly support these existing theoretical principles(Shneiderman 2020; Kang et al. 2023). Rather than contradicting past work, our empirical results confirm HCAI and knowledge-augmentation theories and extend them to the nonprofit sector.

Lastly, we also look to support for theories of knowledge-augmented AI. By integrating a domain-specific knowledge graph, our system seeks to avoid factual errors that purely language-based models often make. This aligns with recent work (Kang et al., 2023) showing that retrieving relevant knowledge from a graph leads to *high-quality,*

knowledge-faithful outputs(Kang et al. 2023). In practice, enabling the AI to query structured domain knowledge will improve accuracy and user trust – exactly as Kang et al. describe. In this way, our future results should extend prior findings on knowledge-grounded models: we seek to validate that augmenting LLMs with a formal ontology improves performance in the philanthropy context, which had not previously been studied.

6.2 Implications for Nonprofit Impact Measurement Practice

The study's findings suggest that well-designed AI tools can indeed act as "force multipliers" for nonprofit impact measurement, but only under the right conditions. Consistent with TAM, all users in our evaluation noted that clear usefulness was a prerequisite for any new technology (Davis 1989). When stakeholders saw how the prototype could aggregate data across projects or automate tedious reporting tasks, they recognized real value – confirming that improved job performance (efficiency, insight) drives acceptance as TAM predicts (Davis 1989). In practical terms, this means that philanthropic organisations can leverage AI to handle fragmented data (e.g. disparate narrative reports) and perform cross-program analytics much faster than by hand. For example, our system made it easy to run complex queries (e.g. roll-ups by theme or country) that were previously infeasible, indicating that AI can both streamline reporting and deepen strategic insight.

However, we also found clear caveats, echoing prior literature on technology adoption. For nonprofits with limited budgets and technical capacity, adoption requires a convincing cost-benefit narrative. Even if an AI tool is powerful, it will only gain traction if it demonstrably saves time or improves outcomes. Stakeholders in our study insisted on a clear *return on investment* – for instance, quantifying how much reporting time the tool saves or how many additional insights it yields – before committing scarce resources. This practical emphasis on measurable benefits is consistent with adoption research and with management advice: new systems must deliver tangible "quick wins" and fit smoothly into existing workflows. Accordingly, our experience underscores that successful implementation depends not only on the technology but on effective change management and training, as seen in literature on nonprofit innovation.

In short, our study confirms that AI has great potential to enhance impact measurement, but realising those potential demands attention to organisational readiness, usability, and clear demonstration of value.

6.3 Ethical and Governance Implications for Responsible AI

Deploying AI for social impact also brings well-known ethical requirements into sharp focus. Our findings align closely with established responsible-AI guidelines. First, transparency and explainability emerged as non-negotiable. Stakeholders repeatedly asked to see *how* the system arrived at its insights, mirroring the EU's Trustworthy AI requirement that decisions be explainable (European Commission 2019). We therefore implemented explainable outputs (e.g. traceable links from statistics back to source documents), which was highly appreciated by users. This supports the broad view (e.g. Miller, 2019) that AI systems must provide interpretable evidence of their reasoning. Second, concerns about fairness and bias also surfaced. Although our prototype did not make predictive decisions, users rightly warned that future predictive models could inherit biases from data. This reflects the international consensus that AI systems require governance mechanisms (such as bias audits and stakeholder review) to prevent discriminatory outcomes (European Commission 2019). Third, **privacy and data protection** were seen as critical: any misuse of sensitive beneficiary information would betray trust and violate ethical norms. In line with ethical principles, we designed safeguards (access controls, anonymisation) and emphasized the need for GDPR compliance.

Ultimately, our study illustrates that core principles – transparency, fairness, privacy, and human control – must be embedded from the ground up. The EU Trustworthy AI guidelines, for example, specify transparency (explainability), non-discrimination, and accountability as baseline requirements (European Commission 2019; Jobin, Ienca, and Vayena 2019). Users in our project clearly expect these features; for instance, they insisted that AI should *augment* rather than override human judgment. By upholding these values in design, we not only built user trust but also aligned the tool with philanthropic norms of accountability and respect for stakeholders. In summary, our findings confirm that responsible AI in the nonprofit domain is not optional: features like explainability, access control, and human-in-the-loop design are practical necessities to ensure adoption and mitigate ethical risks (European Commission 2019).

6.4 Contributions of the Research

This research makes several novel contributions to scholarship and practice. First, it delivers a concrete *Human-AI Collaboration Framework* for nonprofit impact measurement, filling the gap noted in prior studies (which called for integrated AI solutions tailored to social-sector needs). By combining OCR, NLP, and a semantic knowledge graph into one system, we provide a blueprint that addresses persistent challenges (data fragmentation, inconsistent metrics, underused qualitative data) identified in the literature. This artifact extends theory by showing how abstract principles (e.g. human-centered design and ontology use) can be instantiated in practice.

Second, our empirical evaluation yields new insights on technology adoption in the philanthropic context. We confirm that classical factors from TAM (perceived usefulness, ease of use) remain influential in this domain, but we also highlight that ethical alignment and organisational culture are equally pivotal. In other words, our findings extend adoption theory by illustrating how trust and ethical fit become critical antecedents for nonprofits (e.g. transparent tools were preferred because of accountability to donors and beneficiaries). For practitioners, this means our study offers actionable lessons: NGOs should invest in staff training, secure leadership buy-in, and communicate clear efficiency gains when rolling out AI for impact reporting.

Third, our work contributes methodologically by demonstrating the value of a Design Science approach in a mission-driven setting. The iterative, stakeholder-engaged process not only produced an innovative IT artifact but also generated generalizable design principles (such as “ensure transparency and integration with existing workflows”). Sharing these guidelines adds to the academic conversation about how to build acceptable, ethical AI tools.

In summary, this thesis provides both a *working solution* and a *set of insights* that connect back to scholarly gaps and recommendations. By mapping each of our contributions to the literature (as in Table 1), we show how the framework “completes” what was missing in prior work and how our findings “confirm” or “extend” the theories of human-centered AI, knowledge integration, and technology adoption discussed above.

Table 1 Contributions in relation to prior scholarship and literature gaps

Thesis Contribution/Key Finding	Relevant Scholar(s) / Literature Gap (from thesis)	Nature of Relationship	Brief Explanation of the Connection
Development of an integrated Human-AI Collaboration Framework for impact measurement.	Gap in integrated, nonprofit-tailored AI solutions and calls for addressing nonprofit-specific system needs (Sec. 1.2, 2.4).	Completes / Addresses	The thesis develops and prototypes an explicit Human-AI framework (Ch. 4) that specifically addresses the previously identified gap. This framework provides a coherent, nonprofit-specific AI solution, thus initiating the practical fulfillment of this identified research need.
Framework design directly targets persistent challenges like data fragmentation.	Persistent challenges in impact measurement (e.g., data fragmentation, underutilised data) (Sec 2.1, 2.4).	Addresses	The framework explicitly employs a knowledge graph architecture and NLP tools (Ch. 4), tailored to effectively mitigate persistent operational difficulties documented in philanthropic impact measurement, particularly issues of fragmented and underutilised data.
Stakeholder evaluation findings on PU & PEOU align with TAM.	Davis (1989) - Technology Acceptance Model (Sec 2.3).	Confirms / Extends	Positive stakeholder feedback on potential usefulness and ease of use (with training) (, Sec 5.2) validates TAM's applicability in the philanthropic AI context. It extends TAM by highlighting the amplified role of trust and transparency as critical antecedents in this specific domain.
Emphasis on transparency, traceability, and human oversight in the AI framework's design.	Ethical AI principles (EU Trustworthy AI, OECD guidelines, Sec. 2.3); calls for responsible and human-centered AI design (Sec. 1.2, 2.4).	Completes / Provides Practical Insights	By explicitly incorporating transparency, traceability, and human oversight features (Ch. 4, Sec. 4.1-4.2), the framework practically demonstrates how abstract ethical principles from literature can be operationalised. Positive stakeholder evaluation (Sec. 5.2) confirms the practical value and necessity of these ethical considerations in nonprofit AI design.
Application of DSR to develop a socio-technical solution for a real-world nonprofit problem.	Methodological literature on DSR (e.g., Peffers et al., 2007, cited in , Sec 1.4, 3.1).	Offers Nuance / Provides Contextual Example	The thesis provides a detailed case (, Ch 3, 4, 5) of applying DSR in the philanthropic sector, offering nuanced insights into the iterative process of designing and evaluating an AI artifact in a value-driven, resource-constrained environment , contributing a practical example to the DSR body of knowledge.

6.5 Limitations

Despite its contributions, this study has certain limitations that must be acknowledged:

- **One-case context:** The project was conducted with a single type of organisation (mid-sized philanthropic foundation) as the primary context. As a result, the findings and the developed framework are tailored to that specific setting. What works for this case may not generalise to all nonprofits – different organisations (e.g. small community NGOs or large international charities) have varied resources, cultures, and data practices. Thus, caution should be taken in extending our conclusions universally without further validation in other contexts.

- **Prototype stage of development:** The AI framework we built is a proof-of-concept rather than a fully mature product. Some functionalities (such as more advanced automation in data extraction or a polished user interface) were only partially implemented. Therefore, while our evaluation indicates potential efficiency gains and usability, these results are preliminary. A more extensive deployment might expose performance bottlenecks, integration issues, or user experience problems that we did not encounter in the controlled setting. In other words, the positive outcomes we observed assume an idealised environment and may not all hold at scale or over extended periods.

- **Limited evaluation scope:** Our evaluation involved a small number of participants – three individuals in the formal Phase 2 evaluation, supplemented by five interviewees in the exploratory Phase 1. The depth of feedback was rich, but the sample size is very limited. This means the insights should be interpreted as indicative rather than conclusive. We did not measure quantitative performance metrics (for example, exact time saved in report preparation or accuracy improvements), which would be valuable to substantiate the qualitative impressions. Moreover, because the evaluation was short-term and conducted in a simulated scenario (with stakeholders imagining use cases based on a demonstration), we could not observe long-term adoption behaviour or organisational changes.

6.6 Future Work

Building on this research, there are several avenues for future work to further advance AI-enabled impact measurement and address the open questions remaining:

- **Real-world deployment and longitudinal evaluation:** A natural next step is to deploy the AI framework in a live nonprofit environment and observe its use over an extended period. By integrating the tool into an organisation's actual reporting workflow (as opposed to a pilot demonstration), researchers and practitioners could assess long-term metrics such as sustained time savings, improvements in data quality, user adoption rates, and effects on decision-making or learning within the organisation. A longitudinal field study would reveal how the tool performs under regular use, what adoption challenges arise over months or years, and whether initial enthusiasm translates into tangible improvements in organisational effectiveness and mission impact.

- **Broader user testing across contexts:** Future studies should involve a larger and more diverse set of nonprofit organisations to validate and refine the framework. Testing the system with different types of nonprofits (varying in size, focus area, and technical capacity) and with various user roles (from data analysts to program managers to executives) would help identify any context-specific needs or constraints. A broader evaluation could determine which elements of our solution are generalisable and what modifications might be required for other settings. It would also increase confidence in the findings by seeing if similar benefits and challenges occur elsewhere.

- **Technical enhancements to the framework:** There is considerable scope to improve the prototype's technical capabilities. One priority is to further automate the data processing pipeline – for example, employing more advanced OCR and NLP models (potentially trained on nonprofit-specific documents) to extract information with higher accuracy and less human intervention. Another enhancement is to implement a natural language query interface powered by state-of-the-art language models, so that users can interact with the impact data by asking questions in plain language. Additionally, richer visualisation tools (interactive dashboards, maps, etc.) could be integrated to help users explore the analysed data intuitively. Each of these improvements should be developed and

evaluated carefully to ensure they truly add value for end-users without introducing complexity or opacity that might hinder adoption.

- **Comparative and impact studies:** Further research could compare the AI-driven approach presented here with other methods of managing impact data. For example, it would be useful to evaluate how our knowledge graph-based system stacks up against more traditional database systems or even improved manual processes in terms of efficiency, insight generation, and user satisfaction. Such comparative studies would clarify the specific value added by the AI components. Additionally, researchers should investigate the downstream effects of using AI in impact measurement: does having more accessible and analysable data actually lead to different strategic decisions or better outcomes for programs and beneficiaries? Answering these questions may require longitudinal and mixed-method approaches (combining quantitative performance metrics with qualitative interviews over time), but it would provide powerful evidence of whether tools like ours ultimately help nonprofits increase their social impact or organisational learning.

- **Managerial recommendations**: In parallel with technical developments, managerial readiness and internal capacity-building should not be overlooked. Organisations aiming to adopt AI-based impact tools should prioritise staff training, not only on technical functionalities but also on interpreting results and understanding data ethics. Pilot phases should be used to foster a culture of experimentation and iterative learning, ensuring staff buy-in and reducing change resistance. Furthermore, embedding feedback mechanisms into tool usage can help refine the system over time and reinforce trust. Finally, leadership plays a crucial role in championing the adoption and ensuring alignment between the AI tool and the organisation's values, goals, and reporting frameworks.

By pursuing these future directions—including technical refinement, broader user engagement, real-world deployment, and organisational preparedness—researchers and practitioners can build on the foundation laid by this thesis. The project has illustrated the potential of combining AI technologies with human-centred design to strengthen impact measurement practices in the philanthropic sector. However, it represents only

an early contribution. Realising the full value of such systems will require sustained, multidisciplinary efforts involving not only technical innovation but also capacity-building, ethical safeguards, and managerial alignment. Through rigorous evaluation and collaborative development, this work can evolve into a mature and adaptable tool that supports more transparent, strategic, and accountable action in the nonprofit field. Ultimately, the goal remains to equip organisations with actionable insights that amplify their social impact.

7 Conclusion

This thesis sets out to bridge a clear gap between advanced AI technologies and the practical needs of philanthropic organisations for impact measurement. In summary, we developed and evaluated an AI-driven framework that enables nonprofits to transform unstructured impact reports into a structured, queryable repository of insights. The key findings are that (1) combining human-centered design with knowledge-graph augmentation produced a system that stakeholders judged to be useful, accurate, and trustworthy, (2) classical adoption factors (perceived usefulness and ease of use) remain essential, but ethical considerations (transparency, fairness, human oversight) are equally decisive for nonprofits, and (3) well-designed AI tools can greatly enhance efficiency and analytical depth in impact reporting if they are introduced with clear demonstrations of value and adequate support.

The contributions are both practical and theoretical. Practitioners now have a concrete model for how to apply OCR, NLP, LLM and knowledge graphs to impact data, along with documented lessons on user engagement and governance. Academically, the study confirms and extends existing theories (such as HCAI and TAM) to the social-sector domain, and it fills a noted literature gap by providing a unified AI-framework tailored for nonprofit impact analysis. Importantly, the research highlights that responsible design cannot be an afterthought: transparency and human control must be embedded throughout.

In closing, this work demonstrates that a responsible human–AI collaboration approach can indeed strengthen impact measurement in philanthropy. By rigorously aligning technology with ethical imperatives and user needs, we provide a pathway for nonprofits to leverage AI’s power without sacrificing trust or mission. Future work will build on this foundation to create even more robust, real-world solutions, but the present study lays an important groundwork for bringing AI into service of social good.

References

Anfossi, Alberto, and Delphine Moralis. 2024. ‘Data Science, AI and Data Philanthropy in Foundations - on the Path to Maturity’. Philea. 2024. https://philea.eu/opinions/data-science-ai-and-data-philanthropy-in-foundations-on-the-path-to-maturity/.

Bamberger, Michael, Jos Vaessen, and Estelle Raimondo. 2016. *Dealing With Complexity in Development Evaluation: A Practical Approach*. SAGE Publications, Inc. https://doi.org/10.4135/9781483399935.

Cipriano, Michele, and Stefano Za. 2024. ‘Exploring the Evolution of Digital Transformation Research in Non-Profit Organisations: A Bibliometric Analysis’. *Lecture Notes in Information Systems and Organization*. https://ideas.repec.org//h/spr/lnichp/978-3-031-52120-1_13.html.

Cordery, Carolyn J., Galina Goncharenko, Tobias Polzer, Danielle McConville, and Ataur Belal. 2023. ‘NGOs’ Performance, Governance, and Accountability in the Era of Digital Transformation’. *The British Accounting Review*, The British Accounting Review, 55 (2): 101239. https://doi.org/10.1016/j.bar.2023.101239.

Courth, Christoph. 2024. ‘Exploring the Role of Artificial Intelligence in Philanthropy’. The Pictet Group. 2024. https://www.pictet.com/cn/en/insights/role-of-artificial-intelligence-in-philanthropy.

DataKind. 2020. ‘Get Back on the Road: How Riders for Health Intends to Use Computer Vision to Digitize Health Forms’. 2020. https://www.datakind.org/2020/10/27/get-back-on-the-road-how-riders-for-health-intends-to-use-computer-vision-to-digitize-health-forms/.

Davis, Fred D. 1989. ‘Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology’. *MIS Quarterly* 13 (3): 319–40. https://doi.org/10.2307/249008.

Della Giovampaola, Camilla, Maria Cristina Tudor, Lucia Gomez, and Giuseppe Ugazio. 2023a. ‘Current and Potential AI Use in Swiss Philanthropic Organizations - Survey Results’. *SwissFoundations* (blog). 14 June 2023. https://www.swissfoundations.ch/fr/actualites/current-and-potential-ai-use-in-swiss-philanthropic-organizations-survey-results/.

———. 2023b. ‘Current and Potential AI Use in Swiss Philanthropic Organizations - Survey Results’. *SwissFoundations* (blog). 14 June 2023. https://www.swissfoundations.ch/fr/actualites/current-and-potential-ai-use-in-swiss-philanthropic-organizations-survey-results/.

Ebrahim, Alnoor S., and V. Kasturi Rangan. 2010. ‘The Limit of Nonprofit Impact: A Contingency Framework for Measuring Social Performance’. http://www.ssrn.com/abstract=1611810.

Ebrahim, Alnoor S., and Rangan, V. Kasturi. 2014. ‘What Impact? A Framework for Measuring the Scale and Scope of Social Performance’. 2014. https://journals.sagepub.com/doi/10.1525/cmr.2014.56.3.118.

Eikebrokk, Tom Roar, and Dag Håkon Olsen. 2020. ‘Robotic Process Automation and Consequences for Knowledge Workers; a Mixed-Method Study’. *Responsible Design, Implementation and Use of Information and Communication Technology* 12066 (March):114–25. https://doi.org/10.1007/978-3-030-44999-5_10.

Eikenberry, Angela, and Jodie Kluver. 2009. ‘Social Impact Measurement and Non-Profit Organisations: Compliance, Resistance, and Promotion’. *ResearchGate* 15 (3): 85–101. https://doi.org/10.1007/s11266-013-9373-6.

European Commission. 2019. ‘Ethics Guidelines for Trustworthy AI’. 2019. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

European Venture Philanthropy Association. 2015. ‘A Practical Guide to Measuring and Managing Impact’. EVPA. https://www.oltreventure.com/wp-content/uploads/2015/05/EVPA_A_Practical_Guide_to_Measuring_and_-Managing_Impact_final.pdf.

Funnell, Sue and Rogers, and Patricia. 2011. *Purposeful Program Theory: Effective Use of Theories of Change and Logic Models.* Jossey-Bass. San Francisco, CA. https://www.researchgate.net/publication/259999058_Purposeful_Program_Theory_Effective_Use_of_Theories_of_Change_and_Logic_Models.

Gao, Yunfan, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. ‘Retrieval-Augmented Generation for Large Language Models: A Survey’. arXiv. https://doi.org/10.48550/arXiv.2312.10997.

Geneva Centre for Philanthropy - UNIGE. 2025. ‘Philanthropy and Artificial Intelligence’. 2025. https://www.unige.ch/philanthropie/en/research-publication/research/ai-and-philanthropy.

Global Impact Investing Network. 2025. ‘IRIS+ System’. 2025. https://iris.thegiin.org/.

Goldsworthy, Kathryn. 2021. ‘What Is Theory of Change ?’ Australian Institute of Family Studies.

Groß, Marie-Christine. 2024. ‘Doing Good or Doing Well? Drivers and Strategies for Impact Measurement at Foundations’. SSRN Scholarly Paper 4811927. Rochester, NY: TUM School of Management, Technical University of Munich. https://papers.ssrn.com/abstract=4811927.

Hehenberger, Lisa, Anna-Marie Harling, and Peter Scholten. 2015. *A Practical Guide to Measuring and Managing Impact*. Brussels: European Venture Philanthropy Association.

Hevner, Alan R, Salvatore T March, Jinsoo Park, and Sudha Ram. 2004. ‘Design Science in Information Systems Research’ 28 (1): 75–105.

Horrocks, Tamma, Pan, and Jiménez-Ruiz. n.d. ‘Knowledge Graphs’. The Alan Turing Institute. https://www.turing.ac.uk/research/interest-groups/knowledge-graphs.

Huber, B. Rose, Woodrow Wilson School of Public, International Affairs on March 31, 2020, and 10:02 A.m. 2020. ‘Multi-Year Datasets Suggest Projecting Outcomes of People’s Lives with AI Isn’t so Simple’. *Princeton University News*, 2020. https://www.princeton.edu/news/2020/03/31/multi-year-datasets-suggest-projecting-outcomes-peoples-lives-ai-isnt-so-simple.

IBM. 2021a. ‘What Is a Knowledge Graph’. 2021. https://www.ibm.com/think/topics/knowledge-graph.

———. 2021b. ‘What Is Robotic Process Automation (RPA)? | IBM’. 2021. https://www.ibm.com/think/topics/rpa.

Jobin, Anna, Marcello Ienca, and Effy Vayena. 2019. ‘The Global Landscape of AI Ethics Guidelines’. *Nature Machine Intelligence* 1 (9): 389–99. https://doi.org/10.1038/s42256-019-0088-2.

Kail, Angela, Alex Van Vliet, and Lena Baumgartner. 2013. ‘Impact Measurement Practices among Funders in the UK’.

Kang, Minki, Jin Myung Kwak, Jinheon Baek, and Sung Ju Hwang. 2023. ‘Knowledge Graph-Augmented Language Models for Knowledge-Grounded Dialogue Generation’. arXiv. https://doi.org/10.48550/arXiv.2305.18846.

Kay Gugerty, Mary, and Dean Karlan. 2018. ‘Ten Reasons Not to Measure Impact—and What to Do Instead’. 2018.

https://ssir.org/articles/entry/ten_reasons_not_to_measure_impact_and_what_to_do_instead.

Lewis, Patrick, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, et al. 2021. 'Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks'. arXiv. https://doi.org/10.48550/arXiv.2005.11401.

Li, Ying, Vitalii Zakhozhyi, Daniel Zhu, and Luis J. Salazar. 2020. 'Domain Specific Knowledge Graphs as a Service to the Public: Powering Social-Impact Funding in the US'. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2793–2801. Virtual Event CA USA: ACM. https://doi.org/10.1145/3394486.3403330.

Mackinnon, Anne. 2018. 'Measuring What Counts: Meaningful Evaluation for Family Foundations'.

Mayne, John. 2012. 'Contribution Analysis: Coming of Age?' *Evaluation* 18 (3): 270–80. https://doi.org/10.1177/1356389012451663.

OECD. 2024. 'Evaluation Criteria'. 2024. https://www.oecd.org/en/topics/sub-issues/development-co-operation-evaluation-and-effectiveness/evaluation-criteria.html.

'PDF'. n.d. Accessed 4 June 2025. https://unstats.un.org/sdgs/files/report/2024/SG-SDG-Progress-Report-2024-advanced-unedited-version.pdf.

Peffers et al. 2007. 'A Design Science Research Methodology for Information Systems Research.' *Journal of Management Information Systems*, 2007.

Porter, Kristin, Samantha Xia, and Camille Préel-Dumas. 2022. 'The Value of Predictive Analytics and Machine Learning to Predict Social Service Milestones: Lessons Learned from Career Pathways and Child First'. MDRC. https://www.mdrc.org/work/publications/value-predictive-analytics-and-machine-learning-predict-social-service-milestones.

Saunders, Benjamin, Jenny Kitzinger, and Celia Kitzinger. 2015. 'Anonymising Interview Data: Challenges and Compromise in Practice'. *Qualitative Research* 15 (5): 616–32. https://doi.org/10.1177/1468794114550439.

Schmidt, Lena, Pauline Addis, Erica Mattellone, Hannah O'Keefe, Kamilla Nabiyeva, Uyen Kim Huynh, Nabamallika Dehingia, Dawn Craig, and Fiona Campbell. 2024. 'Evaluation Synthesis Analysis Can Be Accelerated through Text Mining, Searching, and Highlighting: A Case-Study on Data Extraction from 631 UNICEF Evaluation Reports'. medRxiv. https://doi.org/10.1101/2024.08.27.24312630.

Shneiderman, Ben. 2020. 'Human-Centered Artificial Intelligence: Reliable, Safe & Trustworthy', February. https://doi.org/10.48550/arXiv.2002.04087.

Sopact. 2024. 'Social Impact Assessment Guide'. 2024. https://www.sopact.com/guides/social-impact-assessment.

UNICEF Office of Research. 2024. 'An Operational Framework for Machine Learning in Evaluation'. 2024. https://www.unicef.org/evaluation/documents/operational-framework-machine-learning-evaluation.

United Nations. 2025. 'United Nations Sustainable Development Goals'. 2025. https://sdgs.un.org/goals#implementation.

Zabłocka-Kluczka, Anna, Anna Marciszewska, and Renata Brajer-Marczak. 2024. 'Non-Profit Organisations in the Context of Digital Transformation: Research Results from Poland'. *Organizacja i Kierowanie* 195 (1): 23–44. https://doi.org/10.33119/OiK.2024.195.1.2.

Appendices

Appendix A: Interview Guide

This appendix presents the structured interview guide employed during stakeholder interviews. The guide was developed in alignment with Technology Acceptance Models (TAM and UTAUT), AI ethics frameworks, and established impact measurement methodologies. Interviews were conducted in two rounds: an exploratory phase and an evaluative phase, designed to first capture initial perceptions about AI adoption and subsequently evaluate the AI-driven impact measurement framework developed in this research.

Round 1: Exploratory Interviews

Purpose: Understand internal perceptions regarding AI adoption, ethical considerations, feasibility, and organisational readiness before introducing the specific AI framework.

Participants: Staff members from diverse roles (Project Managers, Head of Department, Operational Managers).

Duration: Approximately 45-60 minutes.

Interview Questions:

1. **Introduction & Warm-Up**

Q1. “Can you briefly describe your role in the organisation and your main responsibilities?”

Q2. “What experience do you have with new technologies in your work? For example, have you used data analytics tools, automation software, or any kind of AI?”

2. **Perceptions of AI Usefulness & Adoption – Exploring attitudes using TAM/UTAUT concepts.**

Q3. “In your own words, what do you think ‘AI’ could do in the context of our organisation’s work?”

Q4. “What tasks or challenges in your work do you believe AI might help with, if any?”

Q5. “Do you feel using an AI tool would make your job easier or more efficient? Why or why not?”

Q6. “How easy or difficult do you imagine it would be to learn and regularly use an AI-based system in your role?”

Q7. “What might influence your decision to actually start using an AI tool at work?”

3. **Ethical Concerns & Trust in AI – Using AI ethics frameworks to guide discussion.**

Q8. “What concerns, if any, do you have about using AI in our work?”

Q9. “One common concern with AI is ethical issues – for example, fairness (treating all groups equally) or transparency (understanding how AI makes decisions). Do any such issues worry you in our context?” – (Encourage them to consider specific ethics:

- Bias/Fairness: “Do you think an AI could unintentionally favor certain communities or types of projects over others?”
- Transparency: “Would you trust an AI’s recommendation if you don’t know how it arrived at it, or would you need an explanation?”
- Accountability: “If an AI-made suggestion led to a bad outcome, who is responsible?”
- Privacy: “Are you comfortable with AI analysing data about our projects/beneficiaries?”)

Q10. “How important is it for you to have control or oversight over the AI’s decisions or recommendations?”

Q11. “What would help you trust an AI system in your work?”

4. **Feasibility & Organisational Readiness – Discussing practical considerations for AI adoption.**

Q12. “Do you think implementing AI in our organisation is feasible in the near future? Why or why not?”

Q13. “What challenges or barriers do you anticipate in adopting AI here?” – (If not already covered:

• Technical: “We might not have the right data or IT systems.”

• Skills: “Staff might not have the skills to use or interpret AI outputs.”

• Cultural: “There might be resistance to relying on algorithms.”

• Resource: “It could be costly or time-consuming to implement.”)

Q14. “How do you feel the leadership and your colleagues would react to an AI initiative?”

Q15. “What do you think would make AI adoption successful in our organisation?”

5. **Current Impact Measurement Practices – Understanding how impact is assessed today (context for AI solution).**

Q16. “In your role, how do you currently measure or evaluate the impact of the projects or programs you’re involved in?”

Q17. “What tools or processes do you use for impact measurement and reporting?”

Q18. “What are the biggest challenges in measuring impact right now?”

6. **Potential Role of AI in Impact Measurement – Bridging to the AI framework concept (though not presented yet).**

Q19. “In what ways do you think AI could assist in measuring or improving our impact?”

Q20. “Do you have any concerns about using AI for impact measurement or decision-making?”

Q21. “What would an AI-driven tool need to show you to trust its analysis of impact?”

7. **Wrap-Up**

Q22. “Is there anything else you’d like to add about the idea of using AI in our work – any hopes, expectations, or worries we haven’t discussed?”

Round 2: Evaluative Interviews

Purpose: Evaluate the AI-driven impact measurement framework in terms of feasibility, usability, scalability, and organisational fit.

Participants: Senior decision-makers and technical experts.

Duration: Approximately 60 minutes.

Interview Questions:

1. **Overall Impression & Relevance**

Q1. “Now that you’ve seen the framework, what are your initial impressions?”

Q2. “Do you think this tool will help improve how we measure or achieve our impact? Can you give an example?”

2. **Ease of Use & User Adoption**

Q3. “How user-friendly does the framework appear? Could staff with limited tech experience use it comfortably? – what could make it more user friendly?”

Q4.“Based on the organisational culture, do you anticipate staff would actually use this tool? What factors might encourage or discourage them

3. **Feasibility & Technical Considerations**

Q5. “Do you foresee any resource or cost challenges in implementing and maintaining this AI framework?”

Q6. “How well do you think this AI tool would integrate into our existing workflows?”

Q7. “Are there any technical risks or limitations you can identify with this framework?”

4. **Scalability & Organisational Fit**

Q8. “Can you see this framework scaling up to more projects or other departments within the organisation?”

5. **Value and Impact Alignment**

Q9. “If this framework is successful, what would it mean for the organisation long-term? Can it transform how we operate or make decisions?”

Q10. “Do you think the insights from this tool will influence decision-making effectively (e.g., funding decisions, strategy adjustments)?”

6. **Recommendations & Final Feedback**

Q11. “What improvements or changes would you suggest for this framework?”

Q12. “Given everything we’ve discussed, would you personally support the rollout of this framework in the organisation? Why or why not?”

Q13. “Any final thoughts or advice you’d like to share regarding this AI framework and its future in our organisation?”

Appendix B: Detailed Thematic Analysis of Interviews Transcripts

This appendix outlines the thematic analysis conducted according to Braun and Clarke's (2006) six-step methodology.

Methodological Process:

7. **Familiarisation with Data:**
   Transcripts were reviewed repeatedly to ensure thorough familiarity.
8. **Generating Initial Codes:**
   Data was systematically coded, highlighting key concepts and statements.
9. **Searching for Themes:**
   Codes were grouped into broader thematic categories.
10. **Reviewing Themes:**
    Themes were refined and cross-checked with coded data for coherence and consistency.
11. **Defining and Naming Themes:**
    Finalised themes were succinctly defined.
12. **Producing the Report:**
    Representative examples were selected to illustrate the findings clearly.

Methodological Integration of NotebookLM: In the early stages of thematic analysis, NotebookLM (an AI-based qualitative analysis tool) was employed to perform a preliminary exploratory analysis. This tool automatically generated visual representations (concept maps) from interview data, offering an intuitive overview of potential thematic connections and clusters. It facilitated rapid identification of recurrent concepts and provided initial insights, which were subsequently verified and refined through manual coding following Braun & Clarke's (2006) methodology.

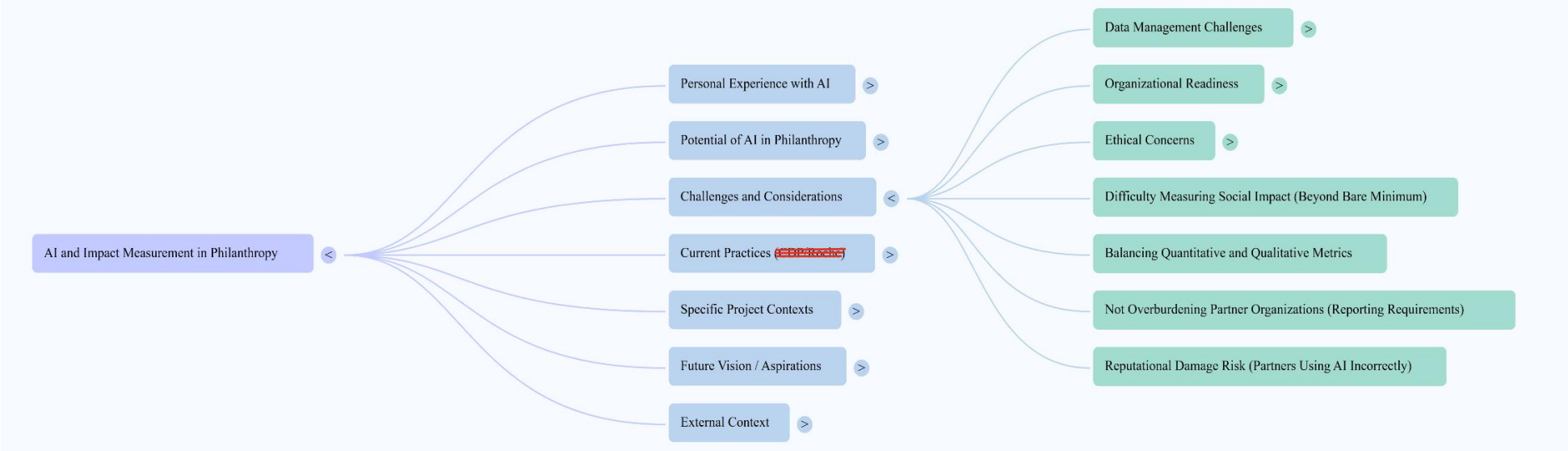


Figure 3 Concept tree visualisation for thematic analysis

Ethical note on Confidentiality and Methodological Choices: The qualitative data collected in this research required careful management due to the interconnected nature of the participant group and the sensitivity of the topics discussed. Confidentiality has thus been rigorously maintained through distinct methodological strategies adapted to the characteristics of each interview round.

For the exploratory phase (Round 1), given the participants' close professional relationships within the organization, direct attribution of quotations posed significant ethical and confidentiality concerns. To mitigate this risk, findings are presented through composite excerpts—constructed quotes synthesizing insights from multiple participants—and aggregate thematic tables. This methodological decision aligns with recommendations from qualitative researchers such as Saunders, Kitzinger, and Kitzinger (2015)(Saunders, Kitzinger, and Kitzinger 2015), who argue that employing composite excerpts and aggregate thematic analyses is essential when managing highly sensitive or closely interconnected participant groups, thereby ensuring participant anonymity without sacrificing the integrity of qualitative findings.

Conversely, the evaluative phase (Round 2) allowed for greater granularity and transparency in coding due to the nature of the content, primarily focusing on evaluating the practical and technical aspects of the developed AI-driven framework. Participants in Round 2 had defined, distinct roles, reducing the risk of re-identification from detailed thematic tables.

By adopting these tailored approaches—composite excerpts and aggregated thematic analyses for Round 1, and more detailed coding tables for Round 2—the research upholds ethical standards, ensuring participant confidentiality is maintained while delivering qualitative analysis.

Round 1: Composite excerpts by theme

Table 2 Refined Thematic Table Round 1

Theme	Description	Representative Insight
Current Impact Measurement Practices	Current methods and metrics used by organisations, and associated practical challenges in impact measurement.	
Perceptions of AI Usefulness & Adoption	Stakeholder attitudes and perceptions toward adopting AI solutions, informed by perceived usefulness, ease of use, and adoption readiness.	*"I can see AI making our impact measurement more efficient by crunching data much faster than we ever could manually."* *"AI sounds promising, but I'm not sure it would solve our basic problems – our data is unstructured to begin with."* *"There might be resistance from staff – not everyone is comfortable with new technology, and using AI would be a big shift for us."*
Ethical Concerns & Trust in AI	Participant concerns regarding ethical implications, data privacy, transparency, and trust-building in AI use for social impact measurement.	*"I wouldn't blindly trust an AI's analysis; I'd feel the need to double-check all its results."* *"One concern is bias – an algorithm might skew the findings or miss important nuances that a human would catch."* *"We also have to consider privacy – feeding sensitive beneficiary data into an AI system could pose risks."*
Feasibility & Organisational Readiness	Practical considerations for implementation including organisational capacity, resource availability, training needs, and data readiness.	*"We don't have any in-house AI expertise, so we'd likely need outside help or new hires to make it work."* *"We'd also need better data and probably more funding to use AI effectively, which is a stretch for us at the moment."*
Potential Role of AI in Impact Measurement	Stakeholders' views on potential benefits and specific roles AI could play in improving accuracy, efficiency, and insights of impact measurement practices.	*"AI could help us spot patterns and trends in our impact data that we might not notice on our own."* *"It might also handle the tedious data crunching, letting us spend more time on understanding the results and taking action."* *"Maybe it could even help predict which projects will have the most impact early on, so we can prioritise our efforts."*

Table 3 Aggregate Coding Table Round 1

Theme	Codes	Prevalence	Explanation
Current Measurement Challenges	Subjective outcomes	High	Impact measurement viewed as subjective/hard to quantify and analyse
	Resource-intensive process	High	Evaluation described as lengthy and costly
	Data collection hurdles	Moderate	Difficulties gathering comprehensive data
	Lack of standard metrics	Lower	Absence of universally accepted impact measures

Round 2: Composite excerpts by theme

Table 4 Refined Thematic Table Round 2

Theme	Description	Representative Insight
Framework Utility & Relevance	Perceived value and relevance of the artifact to organisational reporting needs.	"The framework could give us a clear structure for impact measurement – it can help organise our process." "Some parts of the framework seems useful, but a few elements feel too complicated to fit our organisation's context." "Having an AI analyse and summarise the data would be helpful"
Usability & Interface Concerns	Importance of simplicity and ease of use for non-technical staff.	"It's promising, but we'll need a more intuitive query interface for broad adoption." "The interface is a bit confusing at first – I would like to hunt around to discover certain features." "The design should be more intuitive – especially for peoples who aren't very tech-savvy."
Trust & Ethical Acceptability	Importance of traceability to build trust in AI-generated insights.	"I don't always understand how an AI is making its decisions, which made me hesitant to fully trust recommendations." "If the system explains its reasoning or shows the data behind its suggestions, I'll be more confident about using it."
Feasibility & Implementation Challenges	Practical considerations and resources needed for integration.	"We'd have to train people to use this tool properly – not everyone is comfortable with new software." "Our data isn't very well organised right now, so we'd need to clean it up before." "With our limited budget and time, rolling this out across the organisation would be pretty tough at the moment."
	Necessity of leadership support and organisational culture to drive adoption.	"Leadership buy-in is crucial; without executive backing, adoption won't happen." "Without strong support from the top, an AI tool like this would likely never get off the ground here."

Table 5 Aggregate Coding Table Round 2

Theme	Codes	Frequency (out of 3)	Explanation
Framework Utility	Time-saving confirmed	3	AI seen as significantly speeding up analysis
	Pattern/trend identification	2	AI effectively highlighted key data trends
	Data visualisation helpful	3	Appreciation for clear visual representation of insights
Usability & Interface Concerns	User-friendly interface	2	Moderate feedback on intuitive interface design
	Feature suggestion	2	Recommendations for tutorials and clearer UI labels
	Learning curve	1	Initial difficulty in understanding new metrics
Trust & Ethical Acceptability	Need for explanations	3	Necessity of clearly explained AI outputs for trust
	Human verification	3	Consensus on necessity of human oversight of AI outputs
	Accuracy concerns	2	Questions regarding reliability of AI-generated insights
	Transparency of algorithm	2	Need for understandable AI processes
	Maintaining ethical standards	3	Ensuring AI recommendations adhere to ethical guidelines
Implementation Challenges	Integration with current systems	3	Concerns about compatibility with existing processes
	Data readiness	2	Need for high-quality, standardised data
	Training needs	3	Importance of targeted training for effective AI use
	Gradual rollout	1	Recommended phased implementation strategy

This thematic analysis demonstrates the methodological rigor, ensuring a transparent connection between raw data and interpretive findings while keeping anonymity of the interviewees. This appendix serves as evidence of the thorough qualitative grounding of the thesis.

Appendix C: Proof-of-Concept Material (GraphDB visuals & queries)

This appendix supplements the main thesis by demonstrating the prototype's capability for querying, aggregating, and visualising structured philanthropic data in an RDF knowledge graph. The following figures present screenshots from GraphDB, showing how various queries and visualisations were used in the proof-of-concept implementation.

Knowledge Graph Visualisation

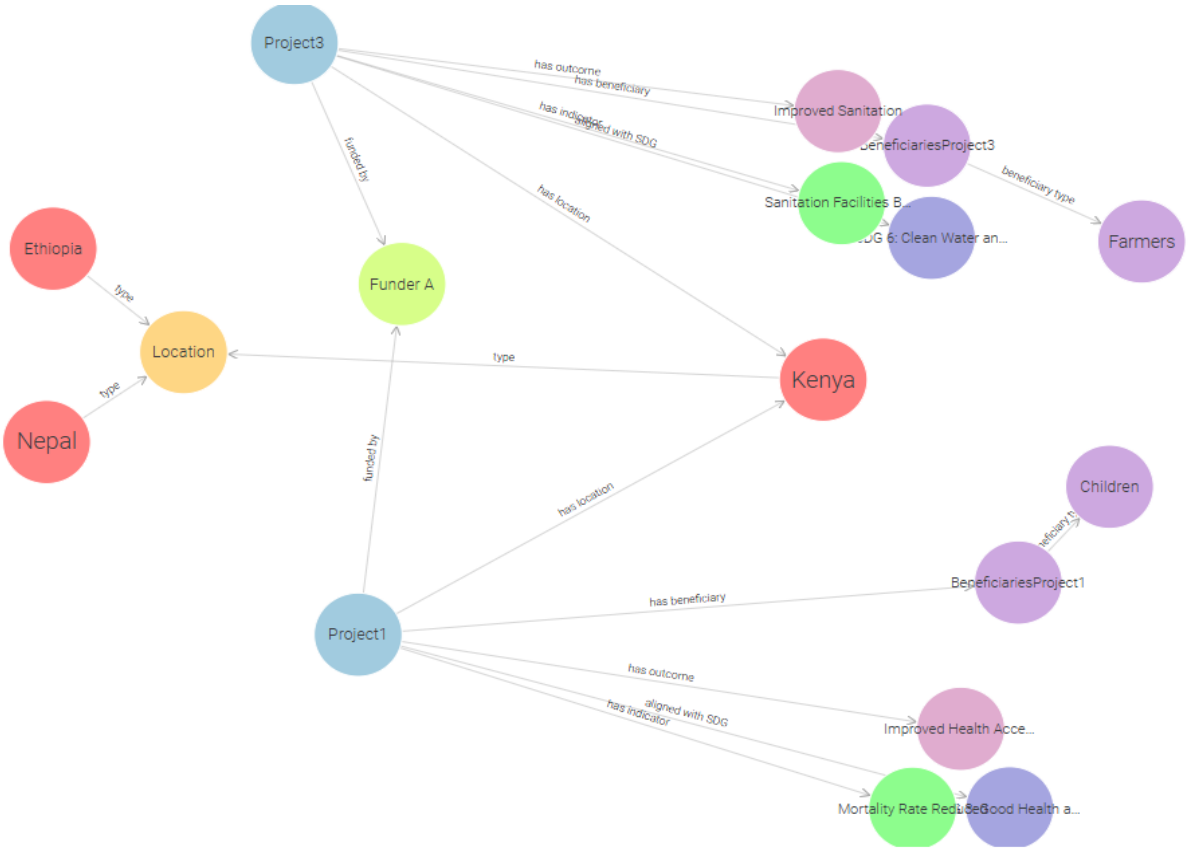


Figure 4: Knowledge graph screenshot

Visualisation showing relationships between projects, geographic locations (countries), funding organisations (funders), Sustainable Development Goal (SDG) targets, and beneficiary groups. Each node represents an entity (e.g., a project, a country, a funder, an SDG goal, or a beneficiary demographic), and edges denote the relationships between them (for example, a project is implemented in a location, funded by a particular funder, addresses a specific SDG, or serves a certain beneficiary group). This visual illustrates how diverse project-related information is interlinked in the graph, enabling holistic queries across these connected dimensions.

SPARQL for Beneficiary Counts

```
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>
PREFIX ex: <http://example.org/philanthropy#>

SELECT ?project ?beneficiaryType ?numberOfBeneficiaries
WHERE {
  ?project a ex:Project ;
           ex:hasBeneficiary ?beneficiary .
  ?beneficiary ex:beneficiaryType ?beneficiaryTypeInstance ;
               ex:numberOfBeneficiaries ?numberOfBeneficiaries .
  ?beneficiaryTypeInstance rdfs:label ?beneficiaryType .
}

```

Figure 5 SPARQL query example 1

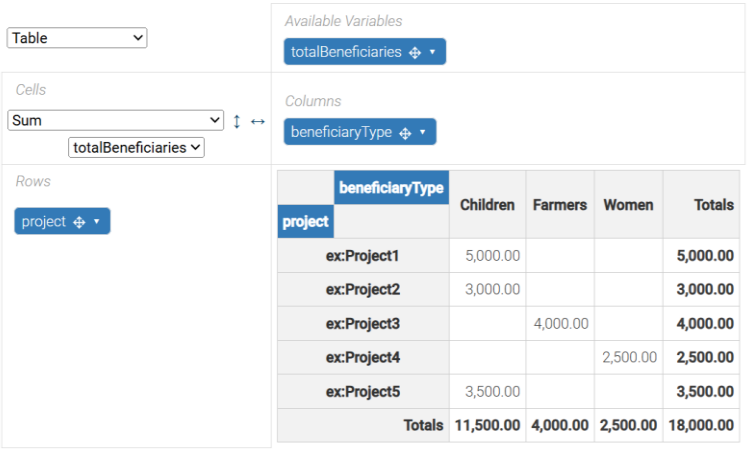
This query retrieves, for each project, the number of beneficiaries grouped by their type. It selects all resources of type ex:Project and follows their ex:hasBeneficiary relationship to identify associated beneficiaries. For each beneficiary, the query extracts the numeric value specified by ex:numberOfBeneficiaries and the label of their type via ex:beneficiaryType and rdfs:label. The result provides a structured view of how many individuals of each beneficiary type are associated with each project. This information supports comparative analysis of reach across projects and beneficiary groups.

After running the beneficiary count query, GraphDB returns a results table that lists each project alongside the count of its beneficiaries in each group. These raw results can be further aggregated (instantly into GraphDB) to gain insights across the entire portfolio of projects. Figure 6 below shows the direct query output, Figure 7 presents a pivot table summarising the totals by beneficiary group and Figure 8 a stocked bar chart visualisation of beneficiary counts per group.

	project	beneficiaryType	totalBeneficiaries
1	ex:Project1	"Children"	"5000"^^xsd:integer
2	ex:Project2	"Children"	"3000"^^xsd:integer
3	ex:Project3	"Farmers"	"4000"^^xsd:integer
4	ex:Project4	"Women"	"2500"^^xsd:integer
5	ex:Project5	"Children"	"3500"^^xsd:integer

Figure 6 Query result table example 1

Query result from GraphDB, listing projects and number of beneficiaries for each beneficiary group in those projects. Each row corresponds to a single project–beneficiary group, showing how many beneficiaries of that group are associated with the project. This tabular output provides the detailed breakdown necessary for granular analysis of impact across different groups.



project \ beneficiaryType	Children	Farmers	Women	Totals
ex:Project1	5,000.00			5,000.00
ex:Project2	3,000.00			3,000.00
ex:Project3		4,000.00		4,000.00
ex:Project4			2,500.00	2,500.00
ex:Project5	3,500.00			3,500.00
Totals	11,500.00	4,000.00	2,500.00	18,000.00

Figure 7 Pivot table derived from the query results example 1

Pivot table aggregating the total number of beneficiaries for each beneficiary group across all projects. In this summary view, each beneficiary group is listed with the cumulative count of beneficiaries from all projects. This allows easy comparison of which groups are most served by the project portfolio overall.

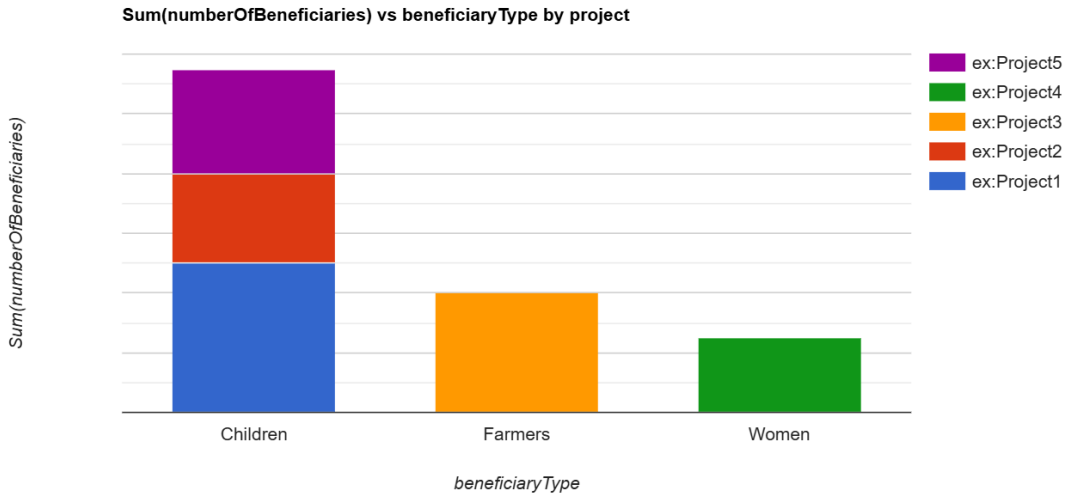


Figure 8 Stocked bar chart visualisation example 1

A stocked bar chart plotting each beneficiary group along the horizontal axis and the total number of beneficiaries on the vertical axis. The height of each bar reflects the aggregate number of beneficiaries in that category, providing a quick visual comparison of impact across different beneficiary demographics. This visual aid helps stakeholders immediately identify which beneficiary groups have the largest reach in the project portfolio.

SPARQL for Projects aligned with Sustainable Development Goals (SDGs)

```
PREFIX ex: <http://example.org/philanthropy#>
PREFIX rdfs: <http://www.w3.org/2000/01/rdf-schema#>

SELECT ?project ?sdgGoal
WHERE {
  ?project a ex:Project ;
           ex:alignedWithSDG ?sdg .
  ?sdg rdfs:label ?sdgGoal .
}
ORDER BY ?sdgGoal ?project
```

Figure 9 SPARQL query example 2

Query to retrieve a structured list of all projects explicitly aligned with specific Sustainable Development Goals (SDGs). This query extracts each project's name alongside the related SDG goal, providing a clear overview of how each project contributes to global sustainability objectives. Executing this query provides valuable data to understand and report on strategic alignment at the organisational or programmatic level.

	project	sdgGoal
1	ex:Project1	"SDG 3: Good Health and Well-being"
2	ex:Project4	"SDG 3: Good Health and Well-being"
3	ex:Project2	"SDG 4: Quality Education"
4	ex:Project5	"SDG 4: Quality Education"
5	ex:Project3	"SDG 6: Clean Water and Sanitation"

Figure 10 Query result table example 2

Query result from GraphDB showing each project and its linked SDG. Each row corresponds to a project–SDG combination, facilitating precise insight into which projects address specific global sustainability goals. This tabular presentation offers transparency and clarity for sustainability reporting and strategic analysis.

Table

Available Variables: project

Cells: Count

Columns: sdgGoal

Rows

sdgGoal	SDG 3: Good Health and Well-being	SDG 4: Quality Education	SDG 6: Clean Water and Sanitation	Totals
Totals	2	2	1	5

Figure 11 Derived pivot table example 2

Pivot table summarising the total number of projects addressing each SDG. This aggregated view clearly displays each SDG and the cumulative number of projects aligned with it, offering an immediate visual summary of organisational impact and strategic focus regarding sustainable development.

Appendix D: Ontology Specification

This appendix describes the structure, classes, relationships, and evolution of the ontology developed for the philanthropic impact measurement framework, explicitly aligned with the semantic standards OWL and RDF, and implemented using Protégé. The ontology is designed to be interoperable and queryable through semantic graph databases like GraphDB, enabling comprehensive and traceable impact analysis. While this ontology represents a foundational structure, it is anticipated to evolve to address additional use cases and emerging requirements.

Class Hierarchy

Table 6 Classes and Descriptions

Class	Description
Project	Philanthropic initiative or intervention
Outcome	Measurable social or environmental effect of a project
Indicator	Specific metric or measure assessing an outcome
Beneficiary	Individuals or groups benefiting from a project
Location	Geographic area of project implementation
Funder	Organisation providing financial resources
SDG_Goal	Sustainable Development Goal addressed by a project

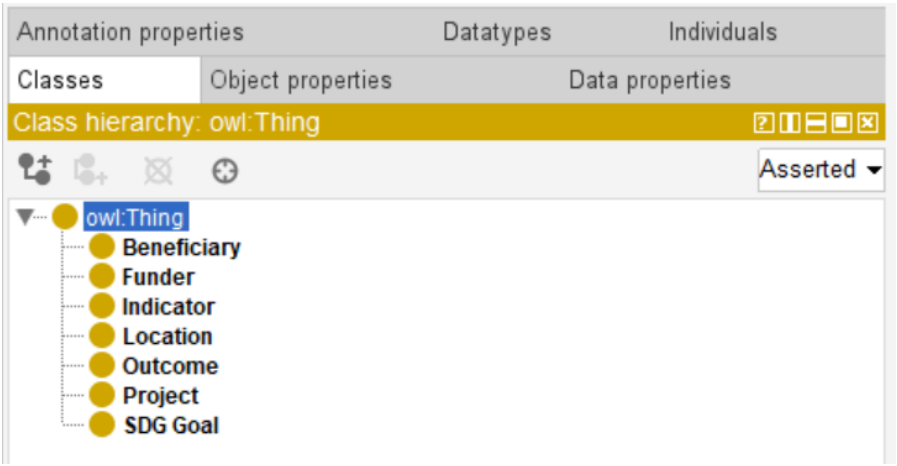


Figure 12 Class hierarchy

Class hierarchy clearly displaying defined core classes (Project, Outcome, Indicator, Beneficiary, Location, Funder, SDG_Goal).

Object Properties (Relationships)

The ontology explicitly defines the following object properties to represent relationships clearly:

Table 7 Object Properties and Descriptions

Property	Domain → Range	Description
hasBeneficiary	Project → Beneficiary	Identifies beneficiaries of a project
fundedBy	Project → Funder	Links projects to their funders
alignedWithSDG	Project → SDG_Goal	Associates projects with relevant SDGs
hasIndicator	Outcome → Indicator	Connects outcomes to specific indicators
hasOutcome	Project → Outcome	Links projects to targeted outcomes
hasLocation	Project → Location	Indicates the geographical area of projects
beneficiaryType	Beneficiary → Beneficiary	Defines categories/types of beneficiaries

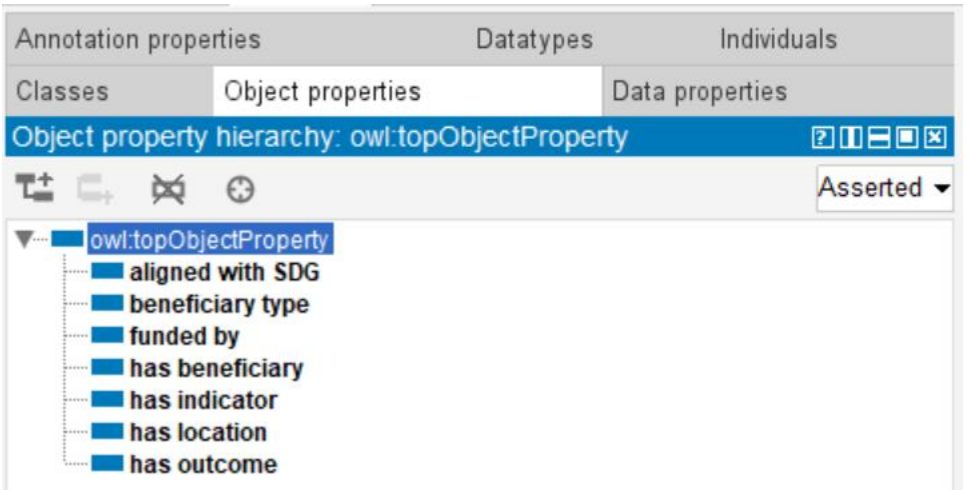


Figure 13 Object properties

Object properties clearly illustrating relationships (hasBeneficiary, fundedBy, alignedWithSDG).

Data Properties (Numeric Values)

The ontology explicitly captures numeric values critical for impact measurement using the following data property:

Table 8 Data Properties and Descriptions

Data Property	Domain	Range	Description
numberOfBeneficiaries	Beneficiary	Integer	Quantifies the beneficiaries impacted

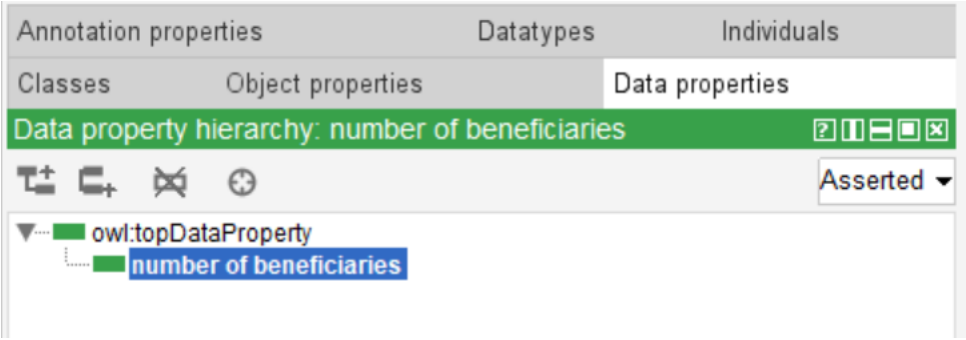


Figure 14 Data properties

Explicit Individual Instances

The ontology explicitly defines multiple individual project instances clearly linked to their associated entities, exemplified by the following:

Example: Project 1

- **Beneficiaries:** 5000 (Children)
- **Funder:** Funder A
- **SDG:** SDG 3 (Good Health and Well-being)
- **Outcome:** Improved Health Access
- **Indicator:** Mortality Rate Reduction
- **Location:** Kenya

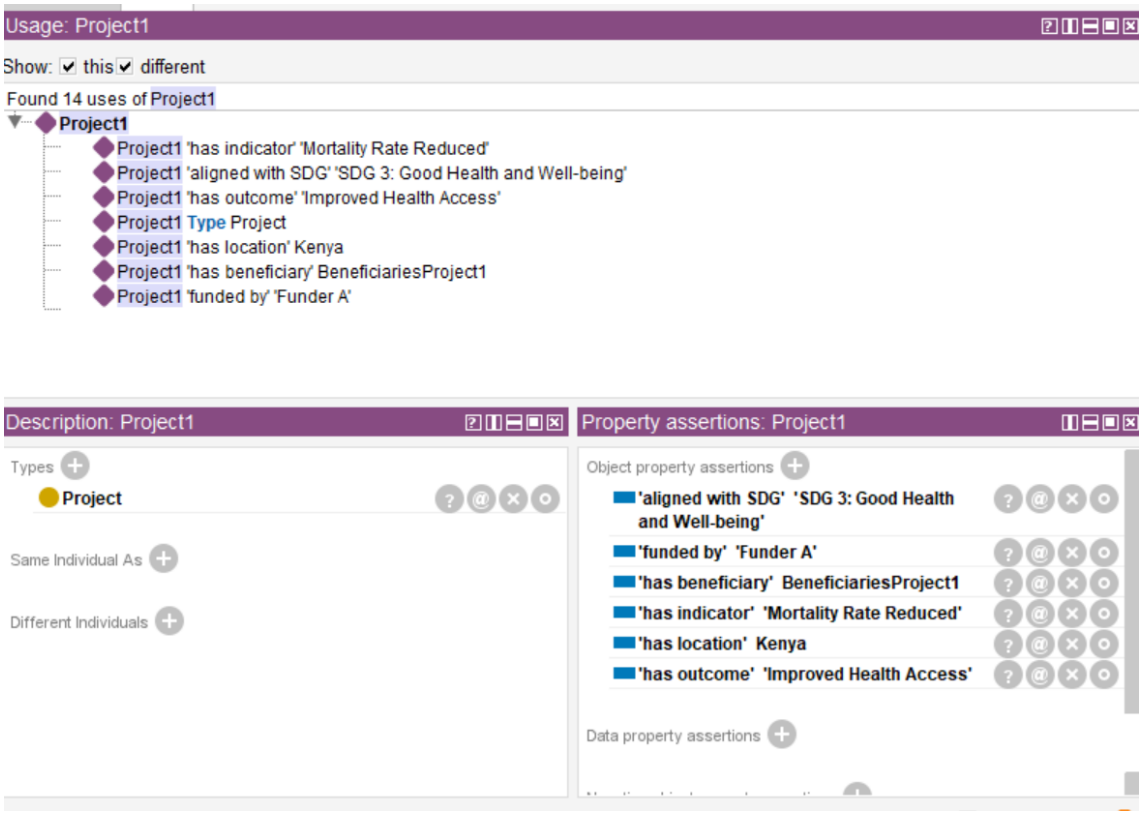


Figure 15 Individual instance view

Individual instance view exemplifying Project 1 with clearly linked beneficiaries, outcomes, indicators, and location.

Additional projects similarly incorporate diverse beneficiaries (Women, Farmers), funders (Funder B, Funder C), and SDGs (SDGs 4, 6), demonstrating extensive coverage and applicability.

Iterative Design Rationale

The ontology's design was iteratively refined through DSR methodology. Initial versions were streamlined for demonstration purposes. Subsequent iterations introduced explicit relationships, expanded beneficiary types, numeric measures, and alignments with global impact standards (SDGs). This iterative approach ensures relevance and adaptability to evolving impact measurement needs.

Final Notes on Ontology Use and Applicability

This ontology explicitly supports precise semantic queries (via SPARQL), robust impact analysis, and transparent reporting. Its clear definitions of beneficiaries, outcomes, indicators, and alignment with SDGs enable scalable, accurate, and practical impact measurement and reporting, directly supporting organisational learning and accountability in philanthropic contexts.

Appendix E: Consent form for participation in scientific research

Consent for participation in scientific research

Designing an AI-Driven Framework to Optimise Impact Measurement in Philanthropic Organisations

Jonah Cabayé
Master's Student – International master's in management of IT
Université Aix Marseille – Turku University – Tilburg University
Email: jonah.cabaye@gmail.com

Purpose of the Research

You are invited to participate in a research study exploring how AI technologies can support impact measurement and reporting in philanthropic organisations.

This second round of interviews is part of the evaluation phase of the study. It aims to gather your feedback on a prototype framework developed based on prior research and interviews. Your insights will help assess its relevance, usability, and potential value to your organisation and the wider sector.

Participation Details

- Participation is voluntary.
- The interview will last approximately 45–60 minutes.
- You can withdraw at any time without giving a reason and without any negative consequences.
- Your responses will be kept confidential and anonymised in all documents and results.
- No identifying information about your role, team, or organisation will be published.
- Audio recordings will be used solely for transcription and analysis, then deleted after transcription is complete.
- Transcripts will be stored securely and only accessed by the researcher.
- You may request a copy of the final thesis upon completion.
- The thesis will be made publicly accessible (e.g., university repository), but no individual or organisation will be identifiable.

Consent Statements

Please read and check the boxes that apply:

☐ I have read and understood the information above.

☐ I understand that my participation is voluntary and that I may withdraw at any time.

☐ I give permission for the interview to be audio recorded for transcription purposes.

☐ I agree that anonymised quotes from my interview may be used in the thesis.

☐ I understand that my identity and the name of the organisation will not be disclosed.

☐ I consent to participate in this research interview.

Name of Participant: _____
Date: _____
Signature:

Appendix F: Use of Generative AI Tools in the Thesis Process

This appendix documents the responsible and transparent use of generative AI tools throughout the development of this thesis. AI tools were used solely to support the writing, editing, and information structuring process. All critical thinking, data interpretation, and analytical decisions were performed exclusively by the researcher. The following tools were employed during different stages of the research:

Table 9 Use of Generative AI in the Thesis Process

Purpose	Tool	Usage Description	Researcher Control
Writing assistance; sentence structuring; clarity and tone refinement	ChatGPT (OpenAI)	Used to improve academic tone, cohesion, and fluency of self-written text in the introduction, methodology, literature review, discussion, and conclusion. Also used to rephrase definitions and clarify relationships between theories.	Used only to enhance structure and grammar. All content, analysis, and interpretation originated from the researcher.
Literature clarification and terminology review	ChatGPT (OpenAI) NotebookLM (Google) Gemini (Google)	Assisted in comparing key concepts (e.g. knowledge graphs vs. taxonomies), and helped refine explanations after the researcher reviewed and understood academic sources.	AI suggestions were paraphrased and adapted. No text was generated without prior input or review.
Interview transcript formatting and simplification	NotebookLM (Google)	Used to clean and restructure long text passages in post-interview documentation. Helped highlight coherent segments in researcher-generated transcripts.	No automatic interpretation or thematic coding was performed.
Theme surfacing and content structuring from uploaded transcripts	NotebookLM (Google)	Supported identification of recurrent themes and alignment across interviews, to assist in the organisation of coding tables and summaries.	Coding and interpretation were exclusively manual. AI only assisted with structural clarity.
Appendix and technical section editing	ChatGPT (OpenAI)	Used to improve descriptions of SPARQL queries, ontology structures, and formatting of interview-related materials in appendices.	Technical content was authored and reviewed by the researcher; AI helped refine grammar and coherence.