Master's Aptitude Thesis

AI-Driven Impact Measurement and Management:

Design and Evaluation of a Framework using the Inluma Case

under the Design Science Research Methodology

Dean Didion

Matriculation Number: IU14148408

Study Program: M.Sc. Applied Artificial Intelligence

Module: Preparation – Master for Professionals

Date: December 23, 2025

Abstract

Measuring social and economic impact has become increasingly important in public sector innovation, as organizations seek to demonstrate accountability, optimize resource use, and align their actions with public value objectives (for Economic Co-operation & Development, 2020b; (GIIN), 2023). Amid this development, artificial intelligence (AI) offers new possibilities for automating data analysis, improving transparency, and supporting evidence-based decision-making in Impact Measurement and Management (IMM).

This thesis applies the **Design Science Research (DSR)** methodology to design, develop, and evaluate an AI-supported IMM framework. The research is situated within the *Public Value Hub* in Leipzig and contributes to the development of the *Inluma* platform for measuring and managing social impact. Drawing on established IMM frameworks from Phineo and UnternehmerTUM, the work derives design requirements that integrate both technical feasibility and public value alignment.

The resulting artefact is implemented in a prototypical form as an initial instantiation and evaluated according to criteria of feasibility, usability, transparency, and comparability. Through this iterative DSR process, the study bridges theory and practice, demonstrating how AI-driven approaches can responsibly enhance impact measurement and strengthen accountability in the public sector.

The expected contribution is threefold: (1) a scientifically grounded artefact design for AI-supported IMM, (2) a methodological illustration of Design Science Research in the public innovation domain, and (3) practical insights for social enterprises and public organizations seeking to operationalize data-driven impact management.

Contents

List of Figures

Chapter 1

Introduction

1.1 Motivation and relevance

Innovation in the public sector is increasingly seen as essential for tackling complex societal challenges. As governments and public institutions explore new ways to deliver services, assess policy outcomes, and engage with citizens, the question of **impact** becomes central. While private sector innovations often measure success through profit and efficiency, public sector innovation needs to be evaluated against broader societal value, which is a much more nuanced and multidimensional goal.

Artificial Intelligence (AI) has emerged as a powerful tool for analyzing vast datasets, identifying complex patterns, and supporting evidence-based decision-making (Marr, 2018; Russell & Norvig, 2016). In the public sector, AI holds significant promise for enhancing transparency, accountability, and responsiveness. A recent study (see Bright et al., 2024) carried out among UK public service professionals showed that about 22% actively use generative AI and 45% are aware of AI tools in their area. However, AI is still *not routinely applied* to assess the impact of innovation initiatives.

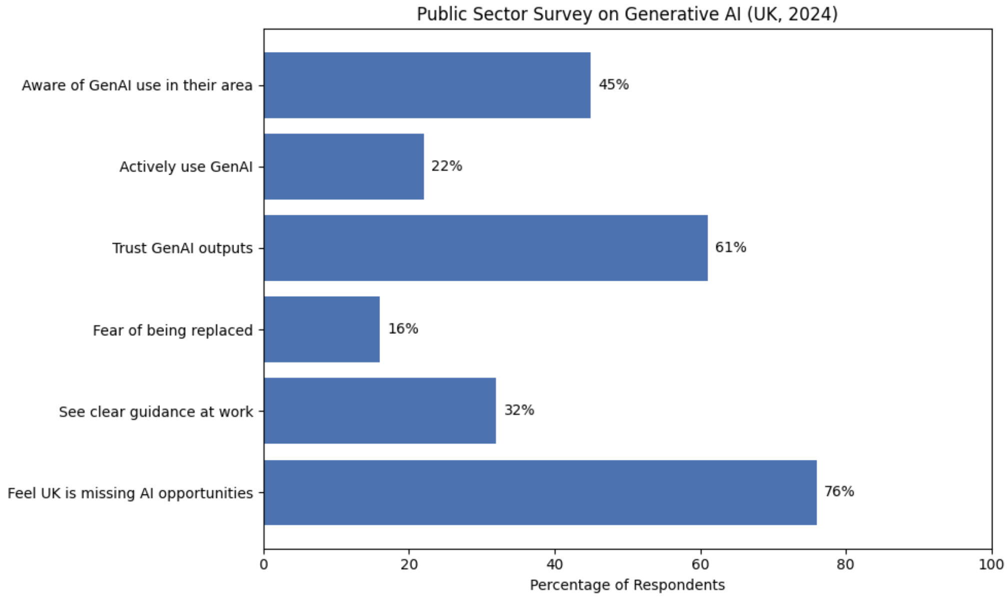

Figure 1.1: Public sector professionals' attitudes toward generative AI (Bright et al., 2024)

Traditional impact measurement frameworks—while widely used—are often too rigid for the dynamic and experimental nature of many public sector initiatives (see Figure 1.2) These frameworks may not accommodate evolving goals, emergent outcomes, or context-specific indicators. Moreover, despite the variety of available frameworks, organizations tend to rely on a single predefined model, often because it is mandated or institutionally recognized. This one-size-fits-all approach can limit flexibility and hinder meaningful evaluation. There is, therefore, a growing need for more adaptive, intelligent systems that can integrate multiple perspectives and evolve alongside the initiatives they aim to assess.

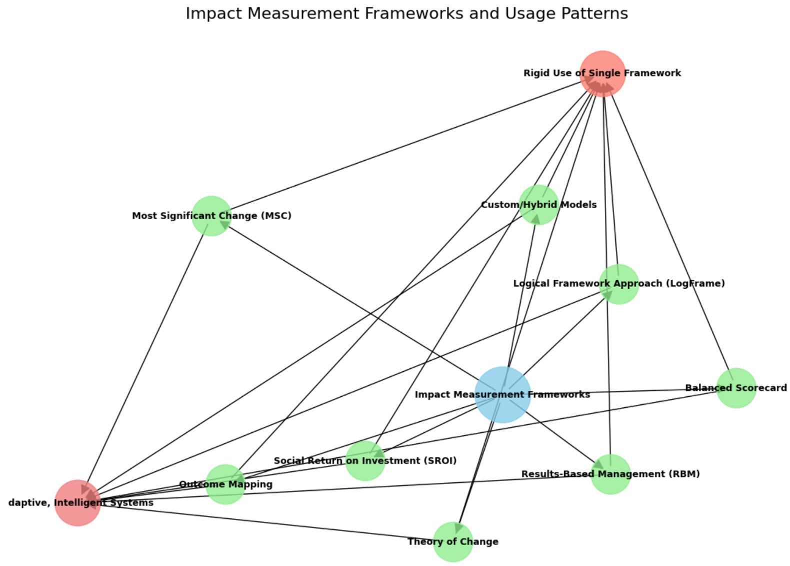

Figure 1.2: Rigid use of single frameworks

This thesis is embedded in a real-world initiative from the **Public Value Hub in Leipzig** and aligns with the goals of the **Public Value Academy**, an emerging digital platform aimed at fostering innovation literacy and sustainable impact measurement in the public sector.

1.2 Problem statement and research gap

Despite the growing use of AI across sectors, there remains a significant gap in how AI can support **meaningful, qualitative impact assessment** — particularly in the public domain. Existing tools often rely on rigid indicators and retrospective analysis, failing to capture complexity, learning, or long-term public value creation (Ebrahim & Rangan, 2014a; Patton, 2011). Moreover, public sector organizations often lack the resources, expertise, or institutional capacity to adopt and adapt AI tools effectively (Mikhaylov & Esteve, 2018).

There is a need to explore **how AI technologies can be applied to support dynamic, context-sensitive, and participatory impact measurement**, integrating frameworks such as those developed by **PHINEO** and supported by **UnternehmerTUM's educational content**.

1.3 Objectives

The aim of this thesis is to develop a **conceptual and technical artefact** for AI-supported impact measurement. By combining theory, stakeholder insights, and prototyping with Python-based methods, the goal is to investigate how such a system could function in practice as part of the Public Value Academy's software platform.

1.4 Research questions

This thesis investigates the potential of artificial intelligence to enhance impact measurement practices. Given the complexity and evolving nature of public value creation, the research is guided by the following questions:

- **How can artificial intelligence contribute to improved impact measurement in public sector innovation?**
  This question explores the capabilities of AI to support more nuanced, dynamic, and qualitative assessments beyond traditional rigid indicators.
- **What are the challenges and opportunities of integrating AI with existing impact measurement frameworks?**
  Here, the focus is on identifying barriers, enablers, and practical considerations when combining AI tools with established impact measurement methodologies.
- **What would a prototype AI-supported measurement tool look like in practice?**
  This question aims to conceptualize and design a practical application that demonstrates how AI can be embedded in impact measurement workflows.

1.5 Scope and limitations

This thesis focuses on the **conceptual design and development** of an AI-supported measurement framework. The implementation centers on a **Python-based Minimum Viable Product (MVP)** that demonstrates core functionalities but stops short of a full-scale deployment. While informed by existing frameworks and stakeholder input, it does not include extensive empirical validation.

Note: The focus is on **public innovation projects** in the German context, though the framework has broader applicability.

1.6 Methodology overview

The research combines:

- A literature review on impact measurement and AI in the public sector,
- Exploration of frameworks (such as PHINEO's IMM),
- Qualitative insights from relevant stakeholders (e.g., Public Value Hub),
- And the development of basic Python-based prototypes to test technical feasibility and application logic.

1.7 Structure of the Thesis

This thesis is structured as follows:

- **Chapter 2** provides the theoretical and conceptual foundation, reviewing relevant literature on Artificial Intelligence, Impact Measurement and Management (IMM), and public value creation.
- **Chapter 3** describes the research methodology and design process applied in the study.
- **Chapter 4** presents the development and demonstration of the prototype, including key stakeholder insights.
- **Chapter 5** discusses the findings in relation to existing frameworks and reflects on implications, challenges, and opportunities.
- **Chapter 6** concludes the thesis with a summary of contributions, limitations, and recommendations for future research.

Chapter 2

Theoretical Background

2.1 Introduction

This chapter reviews existing literature across three interconnected areas: **Impact Measurement and Management (IMM)**, **public sector innovation (PSI)**, and the application of **Artificial Intelligence (AI)** in these domains. The objective is to establish a conceptual foundation for AI-supported, values-driven impact evaluation in public sector innovation ecosystems, and to identify gaps that the thesis artefact implemented in *Inluma* will address.

2.2 Impact Measurement and Management (IMM)

The measurement of impact, particularly in social and public sector contexts, has evolved significantly over the past decades. Scholars such as Ebrahim and Rangan (2014b) emphasize the importance of aligning measurement approaches with a theory of change and organizational strategy. Organizations often struggle to balance accountability and learning, particularly when the expected impact is diffuse or long-term.

Nicholls et al. (2012) highlight tensions between standardized, quantitative measurement systems and the qualitative, context-specific nature of many social interventions. Their work formalizes a typology of impact logic models, demonstrating that one-size-fits-all approaches are rarely effective.

In the German context, intermediaries such as Phineo and UnternehmerTUM provide practical IMM frameworks tailored to social enterprises and innovation labs. These frameworks integrate stakeholder mapping, output-outcome mapping, and logic modelling to clarify how public interventions generate value.

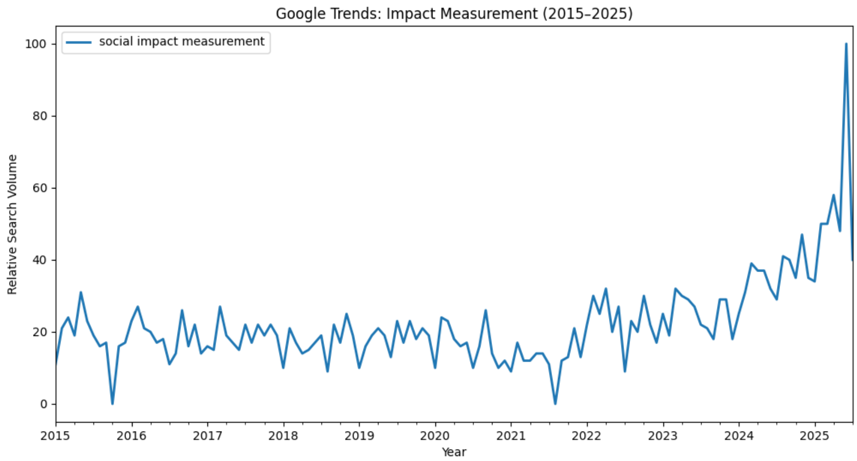

Figure 2.1: Google Trends for "impact measurement".

2.3 Public Sector Innovation and Value Creation

Public sector innovation requires institutions to not only introduce new tools or practices but also foster legitimacy, collaboration, and accountability (Sun & Medaglia, 2019). The OECD has documented challenges and opportunities associated with innovation in government, including an increasing emphasis on public value creation, citizen co-production, and agile experimentation (for Economic Co-operation & Development, 2020a).

Wirtz et al. (2020) propose a conceptual model for digital transformation in public services, emphasizing that data-driven approaches can enhance or erode trust depending on their transparency, inclusiveness, and fairness. The concept of **public value**—first introduced by Moore (1995) and later expanded—serves as a central reference for evaluating the outcomes of public innovation. Initiatives such as Project Athena and CityLAB Berlin exemplify stakeholder-driven innovation aligned with public value frameworks.

2.4 Artificial Intelligence Methods for Qualitative and Quantitative Data Analysis

AI has become increasingly prevalent in public governance, ranging from algorithmic decision-making to NLP-based policy analysis. Devlin et al. (2018) introduced BERT, a transformer-based model foundational for text classification, topic modeling, and semantic similarity analysis. Such methods can be applied to IMM to analyze unstructured stakeholder data, such as survey responses or social media feedback.

Scholars caution that AI must be embedded within deliberative governance structures to ensure its use complements rather than replaces human judgment (Sun & Medaglia, 2019). Similarly, Brown et al.

(2020) highlight that while AI can improve monitoring and accountability, it carries risks such as value misalignment, opacity, and exclusion. In this thesis, AI is employed within the IMM tool *Inluma* to augment human interpretation, particularly in the analysis of complex qualitative narratives.

2.5 Synthesis and Gaps

IMM frameworks, public sector innovation, and AI-supported decision-making offer complementary approaches to tackle complex societal challenges. However, a coherent framework that systematically integrates these domains remains largely absent. Traditional IMM approaches often rely on structured metrics and overlook unstructured qualitative data (Epstein & Yuthas, 2014; Institute, 2023). Public sector innovation initiatives emphasize stakeholder engagement and legitimacy but underutilize AI to scale qualitative data analysis (Berlin, 2024). AI applications, while powerful, often prioritize efficiency over social complexity and normative commitments such as transparency, equity, and public value (Benington & Moore, 2011; Moore, 1995; Union, 2024).

This thesis addresses these gaps by proposing a framework where AI in *Inluma* augments human interpretation, integrates stakeholder input, and aligns with public value principles. For example, a municipal digital inclusion initiative in Hamburg could be analyzed using NLP tools to identify barriers such as affordability, with stakeholders validating results and refining impact metrics. This framework bridges IMM's technical limitations and AI's normative shortcomings, offering an inclusive, transparent, and effective approach to public sector impact measurement.

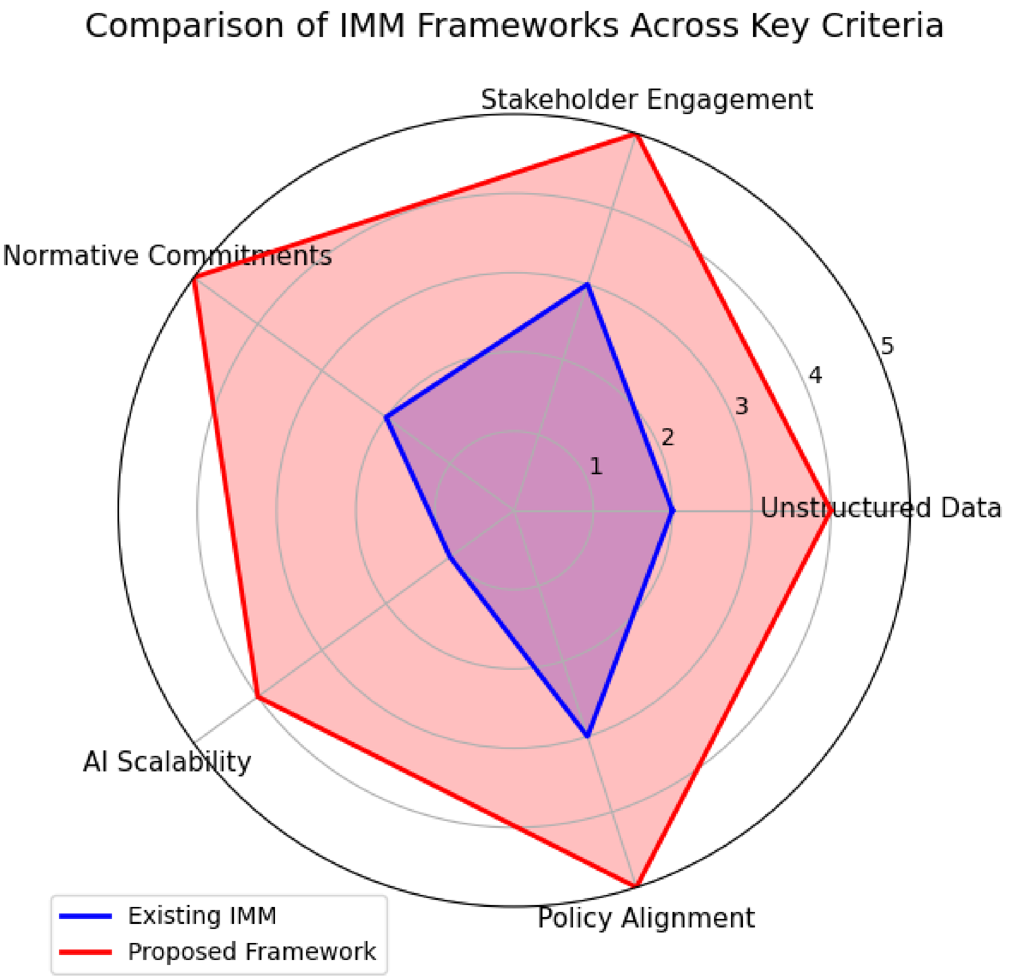

Figure 2.2: Comparison of IMM frameworks.

2.6 Conclusion and Research Direction

This chapter has established the theoretical foundation for the thesis, integrating literature on IMM, AI methods, and public sector innovation. It highlights the research gap that motivates the design, implementation, and evaluation of an AI-supported IMM artefact in *Inluma*. The following chapter presents the methodology used to develop and assess this framework.

Chapter 3

Methodology

This chapter outlines the methodology guiding this research. Building on the principles of **Design Science Research (DSR)**, it describes the process through which an AI-enabled Impact Measurement and Management (IMM) artefact was designed, developed, demonstrated, and evaluated within the context of *Inluma* and the Public Value Hub in Leipzig. The chapter first introduces the methodological foundation, then explains the research context, followed by the stages of artefact creation and evaluation, and concludes with reflections on contributions and ethical considerations.

The process structure primarily follows the six-step model proposed by Peffers et al. (Peffers et al., 2007), while drawing on Hevner et al. (Hevner et al., 2004) for overarching design principles and evaluation criteria.

3.1 Research Methodology

This research applies the **Design Science Research (DSR)** methodology, which provides a structured process for developing and evaluating innovative artefacts in information systems research (Hevner et al., 2004; Peffers et al., 2007). DSR is particularly suited to this thesis, as the objective is not only to analyze existing IMM practices but to design, implement, and evaluate a novel artefact that integrates Artificial Intelligence (AI) into impact measurement and management.

The artefact is implemented as a **prototypical instantiation**—a proof of concept designed to explore feasibility and generate insights for future development. The evaluation therefore focuses on usability, interpretability, and improvement potential rather than generalizability or market readiness.

Following the DSR framework, the research proceeds through six iterative stages (Figure 3.1): problem identification, knowledge base grounding, artefact design and development, demonstration, evaluation, and reflection and contribution.

In this thesis, the artefact takes the form of an AI-enabled IMM framework instantiated as a prototypical software component within the *Inluma* platform.

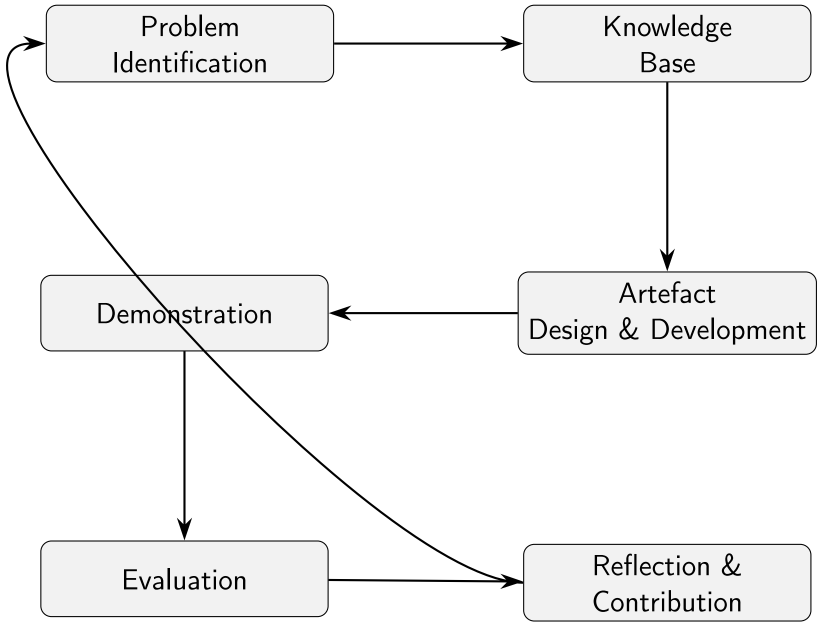

Figure 3.1: Design Science Research (DSR) process cycle (based on Hevner et al., 2004).

3.2 Research Context: Inluma and the Public Value Hub

The *Inluma* initiative, developed within the Public Value Hub in Leipzig, provides a practical setting for the design and demonstration of the artefact. The Public Value Hub connects researchers, practitioners, and public sector innovators through the *Public Value Academy*, which facilitates reflection and learning on public value creation. This environment enables a participatory design process in which academic insights and practitioner experiences inform one another—aligning with DSR's principle of *relevance through engagement*.

Inluma functions as both a conceptual framework and a digital platform for exploring AI-supported learning and reflection processes. It is therefore well suited for the iterative development and evaluation of a proof-of-concept artefact within a real-world innovation ecosystem.

3.3 Problem Identification and Knowledge Base

The first stages of the DSR process involve identifying the practical problem and grounding it in a solid theoretical and empirical knowledge base. In this research, qualitative inquiry was employed to understand existing challenges in impact measurement and management and to identify opportunities for AI integration.

Semi-structured interviews and participatory workshops were conducted with public sector innovators and researchers affiliated with the Public Value Hub and the Public Value Academy. These engagements focused on:

- Limitations in current impact measurement and reporting practices,
- Approaches to operationalizing concepts such as **public value** and **social impact**,
- Stakeholder needs for learning, reflection, and transparency in evaluation processes.

A thematic analysis of the qualitative data informed the artefact's design requirements. Key insights emphasized the need for interpretability, adaptability, and the ability to integrate both quantitative and narrative dimensions of impact. The theoretical grounding draws on literature from impact measurement, artificial intelligence, and public sector innovation, providing the knowledge base that guides artefact development.

3.4 Artefact Design and Development

The central outcome of the DSR process is the design and development of an artefact that addresses the identified problem. In this case, the artefact is an **AI-enabled Impact Measurement and Management (IMM) framework** instantiated within the *Inluma* environment. It aims to support sense-making in impact assessment through natural language processing (NLP), semantic search, and automated knowledge organization.

The artefact consists of four interconnected modules:

3.4.1 Narrative Analysis of Pitch Decks

This module uses large language models (LLMs) to analyze qualitative project materials such as pitch decks or reports. It extracts key entities, identifies value propositions, and translates narrative inputs into structured representations.

3.4.2 Semantic Similarity Search Across Frameworks

An embedding-based search mechanism allows comparison between project narratives and reference frameworks such as the Sustainable Development Goals (SDGs) or public value dimensions. This enables contextual mapping of activities and outcomes.

3.4.3 Clustering and Thematic Grouping of Narratives

Using vector embeddings, thematically related concepts are grouped together to reveal emergent impact patterns and shared priorities across projects. These clusters serve as a foundation for reflection and learning rather than automated judgment.

3.4.4 Automated KPI Derivation via LangGraph Pipelines

An experimental module applies the `LangGraph` orchestration framework to derive candidate indicators and measurable outcomes from qualitative inputs. This step illustrates how AI can support, rather than replace, expert-driven evaluation design.

3.4.5 Text Analysis and Topic Modeling Pipeline

To derive thematic insights and improve indicator recommendations, narrative inputs (such as problem statements, vision, and impact descriptions) are processed through a structured text analysis workflow. This enables clustering of projects with similar focus areas and enhances automated KPI suggestions.

- **Preprocessing:** Tokenization, stopword removal, and lemmatization prepare textual data for analysis.
- **Vectorization:** Both TF–IDF and Bag-of-Words representations are computed for interpretability.
- **Topic Modeling (LDA):** Latent Dirichlet Allocation identifies thematic structures within project narratives. In the scope of this thesis, topic modeling was explored at a prototypical level to assess feasibility and interpretability, rather than to optimize model performance or exhaustively validate topic coherence.
- **Clustering:** Projects are grouped based on topic distributions or semantic embeddings to reveal recurring social and environmental domains.
- **Similarity Search:** Cosine similarity enables retrieval of similar projects or indicators, supporting recommendation logic.

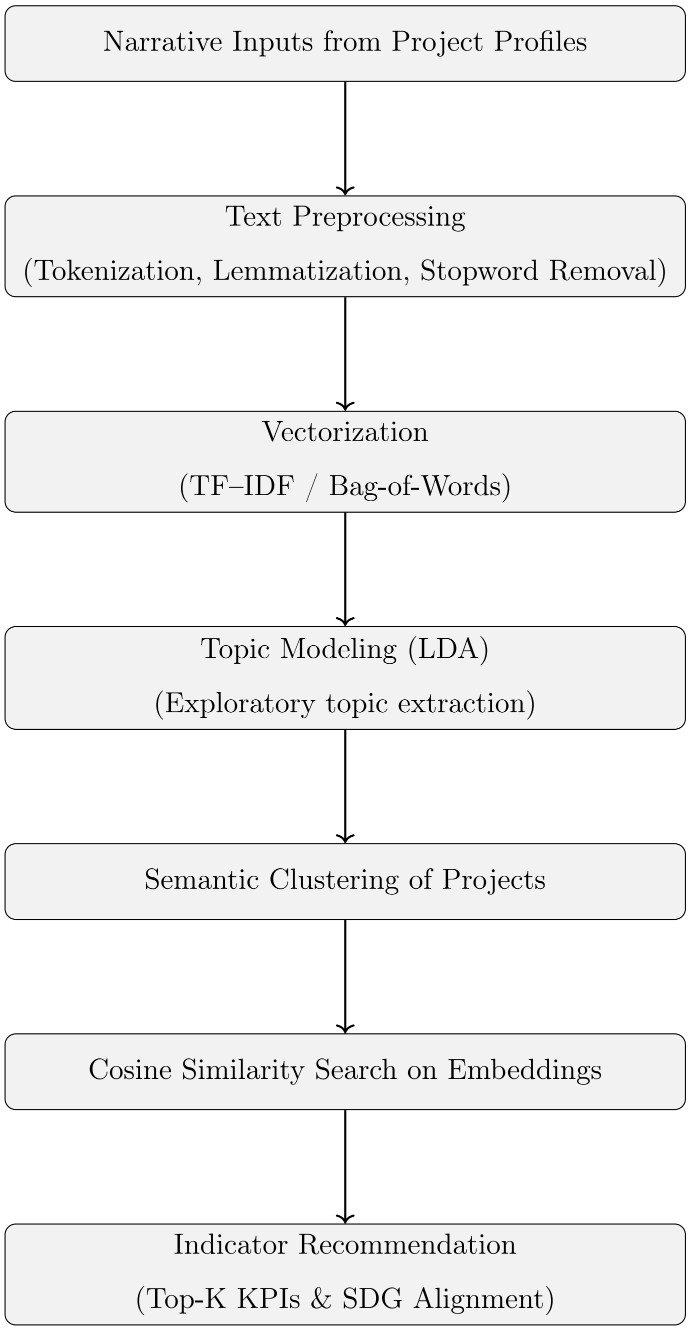

Figure 3.2: Vertical Workflow for Text Analysis, Topic Modeling, and Indicator Recommendation

This pipeline not only structures unstructured text but also provides a data-driven foundation for impact assessment by identifying recurring themes and mapping them to relevant KPIs.

3.5 Demonstration and Evaluation

The demonstration and evaluation stages assess the artefact's utility, usability, and relevance in its intended context. The prototype was integrated into the digital platform of the Public Value Academy, allowing practical demonstration during workshops and learning sessions on impact and innovation.

A formative evaluation approach was adopted. The artefact was tested with anonymized project materials and synthetic inputs to ensure data protection. Practitioner feedback was collected through guided user walkthroughs and structured reflection sessions, documented using qualitative notes and thematic synthesis.

Evaluation criteria included:

- **Usefulness** — the extent to which AI-generated outputs supported reflection and learning,
- **Transparency** — the clarity of AI reasoning and output explainability,
- **Alignment** — consistency of generated insights with stakeholder expectations and value frameworks,
- **Usability** — ease of interaction and perceived integration potential within existing workflows.

Findings from the evaluation informed iterative refinement of the artefact, consistent with DSR's cyclical nature of design, demonstration, and assessment.

3.6 Reflection and Contribution

The reflection stage consolidates theoretical and practical insights from the artefact's design and evaluation. From a theoretical perspective, this research extends the application of DSR into the emerging field of AI-supported impact measurement and management. Practically, it provides a transparent, participatory, and adaptable framework for integrating AI methods into public sector innovation and learning processes.

The artefact demonstrates that AI can act as a *cognitive partner* in impact assessment—facilitating sense-making, comparison, and interpretation without displacing human judgment. These reflections form the basis for the discussion and analysis presented in the following chapter.

3.7 Ethical Considerations

Ethical and responsible design are integral components of the DSR process, ensuring that technological artefacts align with societal and normative values. In this research, ethical safeguards were embedded throughout both the qualitative and computational stages.

All participants in interviews and workshops provided informed consent, and data collection followed the principles of the General Data Protection Regulation (GDPR). Anonymized datasets were used for prototype testing. From a technical perspective, explainability and transparency were addressed by designing the artefact with interpretability in mind, including the consideration of model interpretation techniques such as SHAP (SHapley Additive exPlanations) and systematic logging of AI interactions.

Additionally, the design process considered potential risks of bias, over-automation, and the ethical use of public sector data. Mitigation strategies included human-in-the-loop validation, traceability of model outputs, and clear boundaries between automated analysis and human interpretation.

Chapter 4

Artefact Development

This chapter describes the design and implementation of the AI-supported Impact Measurement and Management (IMM) artefact for *Inluma*. It details the workflow from onboarding new projects, parsing and structuring pitch decks, AI-assisted KPI generation, and integration with the Public Value Academy platform.

4.1 Project Onboarding and Pitch Deck Parsing

To reduce early-stage assessment pain points, a **Pitch Deck Parsing** function was developed:

- PDF documents are processed using `PyPDF` to extract text and graphic information.
- AI models correct scrambled text, OCR errors, or formatting inconsistencies.
- Extracted data is structured with `Pydantic` classes for downstream processing.

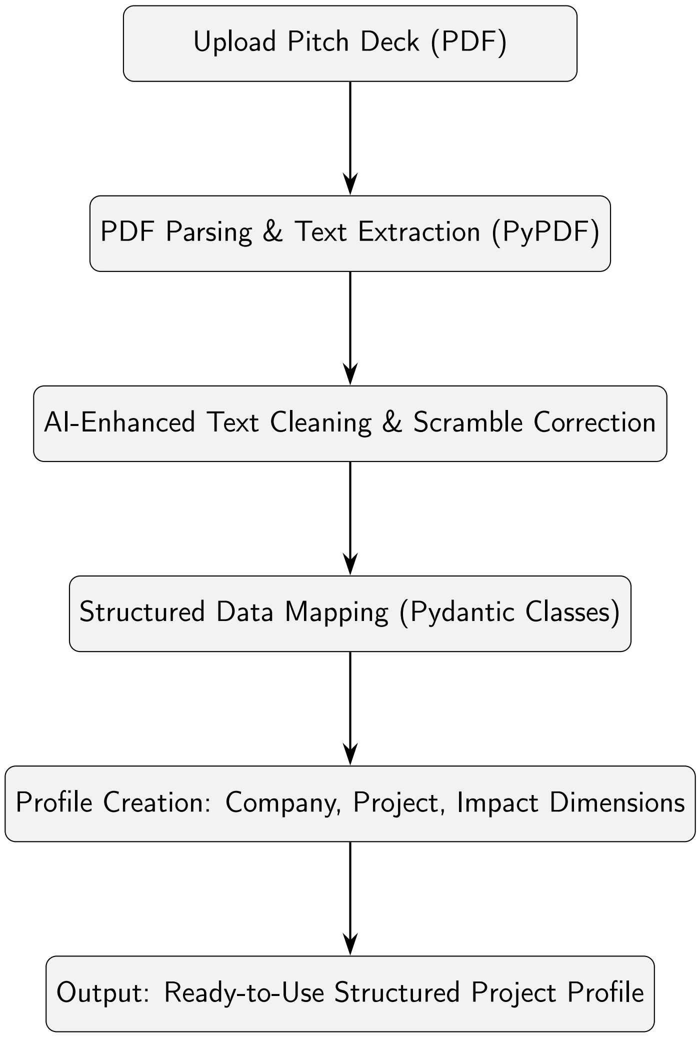

Figure 4.1: Automated Pitch Deck Parsing and AI-Enhanced Extraction Workflow (vertical layout).

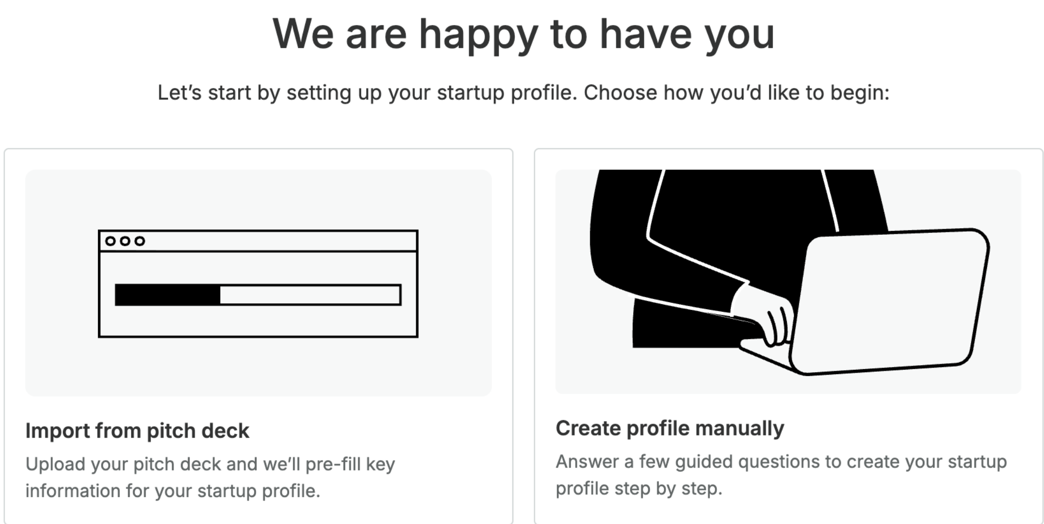

Figure 4.2: Screenshot of the pitch deck upload and parsing interface in the prototype.

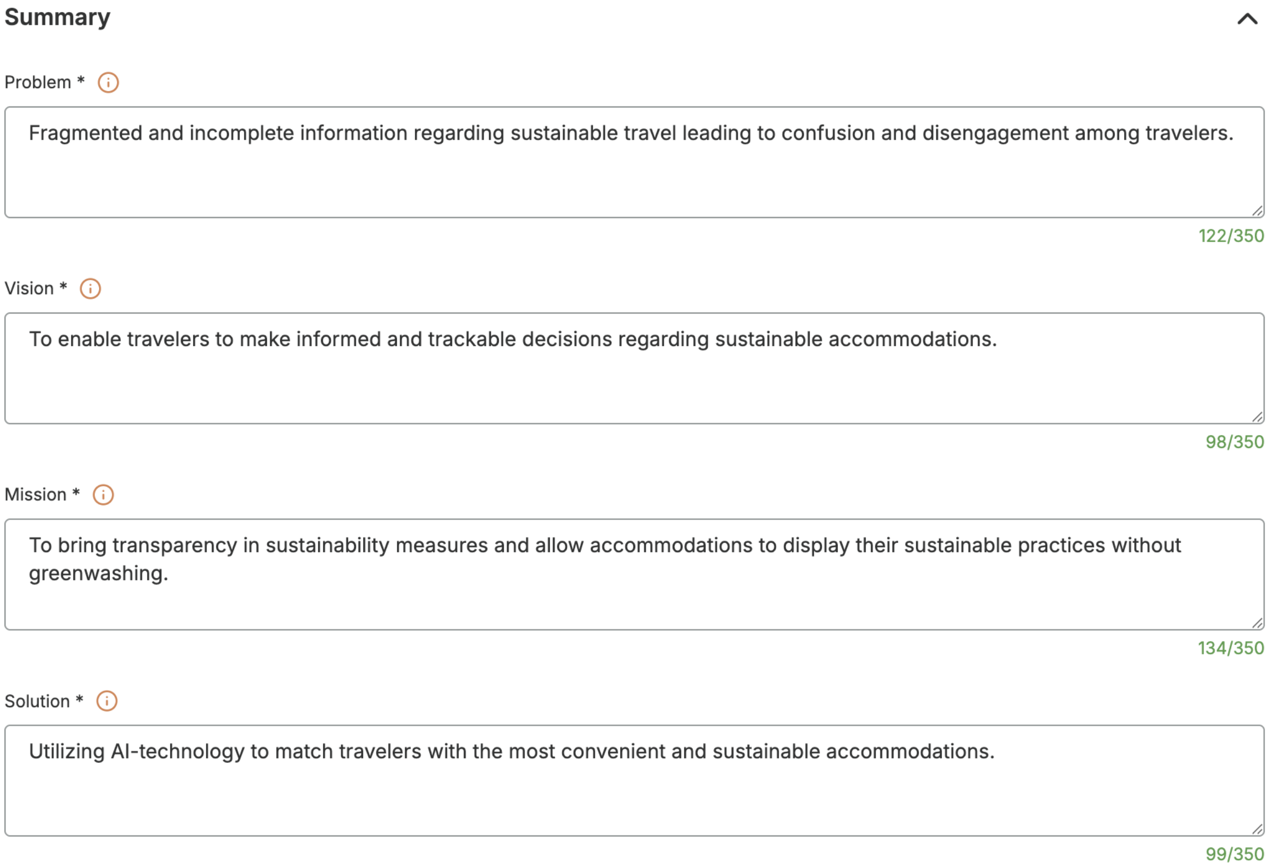

Figure 4.3: Screenshot of the parsed info being populated into the profile

4.1.1 Structured Project Profile

The `Profile` Pydantic model captures essential company, founder, and project information:

```
class IndicatorKPI(BaseModel):
    category: str = Field(description="The area of interest.")
    sub_category: str = Field(description="The sub-category of interest.")

    goal: List[str] = Field(
        description="The goal os goals of the indicator")

    short_term_goal_1: str = Field(description="The short term goal of the
        indicator")
    short_term_indicator_1: str = Field(description="The indicator.")
    short_term_question_1: str = Field(description="The short term
        question.")
    type_of_short_term_question_1: QuestionType = Field(description="Type
        of question.")
```

```
answer_options_short_term_question_1: List[str] =
   Field(description="List of options for the question.")
measurement_method_short_term_question_1: str =
   Field(description="Measurement method for the indicator or its
   title/label.")
unit_method_short_term_question_1: str = Field(description="Unit for
   the indicator")
justification_method_short_term_question_1: str =
   Field(description="Justification for the indicator")
source_method_short_term_question_1: str = Field(description="The
   source (framework) of the indicator")

long_term_goal_1: List[str] = Field(description="The long term goal of
   the indicator")
long_term_indicator_1: List[str]= Field(description="The indicator.")
long_term_question_1: str = Field(description="Question text for the
   indicator")
type_of_long_term_question_1: QuestionType = Field(description="Type of
   question.")
answer_options_long_term_question_1: List[str] =
   Field(description="List of options for the question.")
measurement_method_long_term_question_1: str =
   Field(description="Measurement method for the indicator or its
   title/label.")
unit_method_long_term_question_1: str = Field(description="Unit for the
   indicator")
justification_method_long_term_question_1: str =
   Field(description="Justification for the indicator")
source_method_long_term_question_1: str = Field(description="The source
   (framework) of the indicator")

short_term_goal_2: List[str] = Field(description="The short term goal
   of the indicator")
short_term_indicator_2: List[str] = Field(description="The indicator.")
short_term_question_2: str = Field(description="The short term
   question.")
type_of_short_term_question_2: QuestionType = Field(description="Type
   of question.")
answer_options_short_term_question_2: List[str] =
   Field(description="List of options for the question.")
```

```
measurement_method_short_term_question_2: str =
    Field(description="Measurement method for the indicator or its
    title/label.")
unit_method_short_term_question_2: str = Field(description="Unit for
    the indicator")
justification_method_short_term_question_2: str =
    Field(description="Justification for the indicator")
source_method_short_term_question_2: str = Field(description="The
    source (framework) of the indicator")

long_term_goal_2: List[str] = Field(description="The long term goal of
    the indicator")
long_term_indicator_2: List[str] = Field(description="The indicator.")
long_term_question_2: str = Field(description="Question text for the
    indicator")
type_of_long_term_question_2: QuestionType = Field(description="Type of
    question.")
answer_options_long_term_question_2: List[str] =
    Field(description="List of options for the question.")
measurement_method_long_term_question_2: str =
    Field(description="Measurement method for the indicator or its
    title/label.")
unit_method_long_term_question_2: str = Field(description="Unit for the
    indicator")
justification_method_long_term_question_2: str =
    Field(description="Justification for the indicator")
source_method_long_term_question_2: str = Field(description="The source
    (framework) of the indicator")

sdg_target_1:str = Field(description="The best matching SDG target.")
sdg_target_2: str = Field(description="1. Optional sub SDG target. ")
sdg_target_3: str = Field(description="2. Optional sub SDG target.")

question:str = Field(description="Question text for the indicator")
type_of_question:QuestionType = Field(description="Type of question.")
answer_options:List[str] = Field(description="List of options for the
    question.")
measurement_method:str = Field(description="Measurement method for the
    indicator or its title/label.")
unit:str = Field(description="Unit for the indicator")
```

```
    justification: str = Field(
        description="Justification for the indicator")
    source:str = Field(description="The source (framework) of the
        indicator")
```

An example of a filled `Profile` instance for a hypothetical early-stage impact startup is shown below:

```
example_profile = Profile(
    startup_name="EcoTrack Analytics",
    legal_form="GmbH",
    founder_first_name="Laura",
    founder_last_name="Schneider",
    founder_gender=Gender.FEMALE,
    startup_email="contact@ecotrack.io",
    startup_phone="+49 176 12345678",
    startup_city="Berlin",
    startup_country="Germany",
    startup_postcode="10115",\chapter{Artefact Development}\label{ch:artefact-development}

This chapter describes the design and implementation of the AI-supported Impact Measurement and
It covers the workflow from onboarding new projects, parsing and structuring pitch decks, AI-ass

\section{Project Onboarding and Pitch Deck Parsing}\label{sec:onboarding}

To reduce early-stage assessment effort, a \textbf{pitch deck parsing} function was developed:

\begin{itemize}
    \item PDF pitch decks are processed using \texttt{PyPDF} to extract text (and, where feasibl
    \item AI-assisted cleaning is applied to mitigate scrambled text, OCR artefacts, and formatt
    \item Extracted content is mapped into structured schemas (Pydantic models) for downstream p
\end{itemize}

\begin{figure}[H]
    \centering
    \begin{tikzpicture}[
        node distance=1.5cm,
        every node/.style={font=\sffamily, align=center},
```

```
        box/.style={rectangle, rounded corners, draw=black, fill=gray!10, minimum width=6cm, min
        arrow/.style={-{Stealth[length=3mm,width=2mm]}, thick}
    ]
    \node[box] (upload) {Upload Pitch Deck (PDF)};
    \node[box, below=of upload] (pdf) {PDF Parsing \& Text Extraction (PyPDF)};
    \node[box, below=of pdf] (ai_clean) {AI-Enhanced Text Cleaning \& Scramble Correction};
    \node[box, below=of ai_clean] (struct) {Structured Data Mapping (Pydantic Schemas)};
    \node[box, below=of struct] (profile) {Profile Creation: Company, Project, Impact Dimensions
    \node[box, below=of profile] (output) {Output: Structured Project Profile};
    \draw[arrow] (upload) -- (pdf);
    \draw[arrow] (pdf) -- (ai_clean);
    \draw[arrow] (ai_clean) -- (struct);
    \draw[arrow] (struct) -- (profile);
    \draw[arrow] (profile) -- (output);
    \end{tikzpicture}
    \caption{Automated pitch deck parsing and AI-enhanced extraction workflow.}
    \label{fig:pitchdeck-parsing}
\end{figure}

% --- FIGMA SLOTS (replace files later) ---
\begin{figure}[H]
    \centering
    \includegraphics[width=\textwidth]{../fig/todo_pitchdeck_upload_ui}
    \caption{Prototype UI: pitch deck upload and parsing view (TODO: replace with Figma export).
    \label{fig:todo_pitchdeck_upload_ui}
\end{figure}

\begin{figure}[H]
    \centering
    \includegraphics[width=\textwidth]{../fig/todo_parsed_profile_autofill}
    \caption{Prototype UI: parsed content mapped into profile fields (TODO: replace with Figma e
    \label{fig:todo_parsed_profile_autofill}
\end{figure}

\subsection{Structured project profile}\label{subsec:profile-model}
```

```
The onboarding stage produces a structured \texttt{Profile} object (company and project metadata
This profile acts as the primary input for indicator retrieval, SDG mapping, and KPI generation.

\begin{verbatim}
example_profile = Profile(
    startup_name="EcoTrack Analytics",
    legal_form="GmbH",
    founder_first_name="Laura",
    founder_last_name="Schneider",
    startup_city="Berlin",
    startup_country="Germany",
    problem="SMEs lack accessible tools to measure and optimize their environmental footprint.",
    mission="Provide affordable, data-driven sustainability analytics for SMEs.",
    solution="SaaS platform for automated carbon footprint analytics and reduction insights.",
    social_impact="Supports emission reductions and improved sustainability reporting.",
    target_group="SMEs in manufacturing and logistics."
)
```

Figure 4.4: Prototype UI: structured project profile view (TODO: replace with Figma export).

4.1.2 IndicatorKPI schema

Generated KPIs are represented as structured `IndicatorKPI` objects. The schema captures short- and long-term goals, indicators, survey questions, measurement methods, units, and SDG alignment.

```
class IndicatorKPI(BaseModel):
    category: str
    sub_category: str
    goal: List[str]

    short_term_goal_1: str
    short_term_indicator_1: str
    short_term_question_1: str
    type_of_short_term_question_1: QuestionType
    answer_options_short_term_question_1: List[str]
    measurement_method_short_term_question_1: str
    unit_method_short_term_question_1: str
    justification_method_short_term_question_1: str
```

```
source_method_short_term_question_1: str

long_term_goal_1: List[str]
long_term_indicator_1: List[str]
long_term_question_1: str
type_of_long_term_question_1: QuestionType
answer_options_long_term_question_1: List[str]
measurement_method_long_term_question_1: str
unit_method_long_term_question_1: str
justification_method_long_term_question_1: str
source_method_long_term_question_1: str

sdg_target_1: str
sdg_target_2: str
sdg_target_3: str
```

4.2 Indicator and KPI Generation

Following onboarding, the IMM phase begins:

- A pre-generated library of over 1,600 indicators serves as a reference base.
- The function `gen_k_measurement_kpi()` generates SMART KPI drafts for selected categories and subcategories.
- Each KPI includes measurement logic (unit, method), survey-ready question text, and a brief justification.
- If multiple outcomes are detected in the input narrative, optional secondary goals can be proposed.

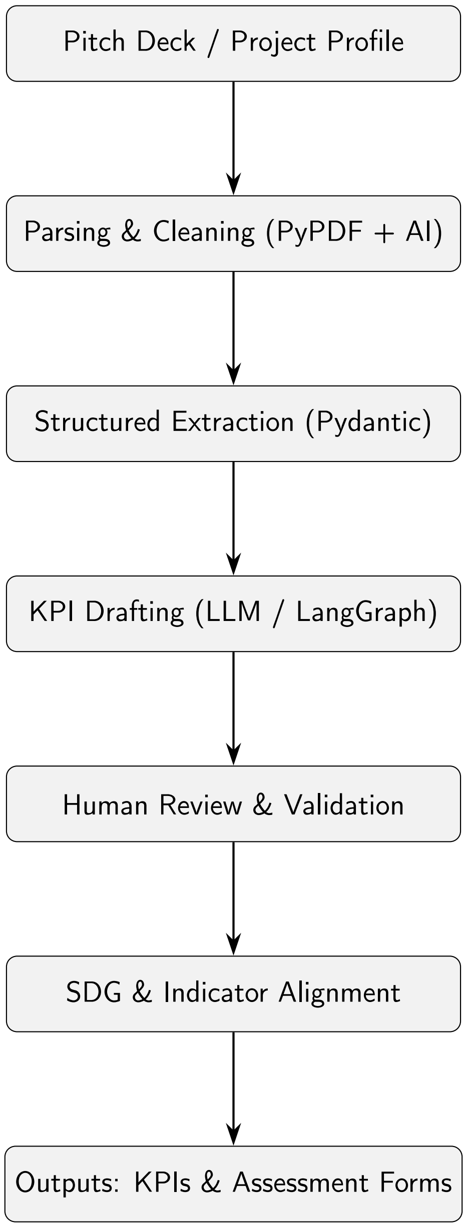

Figure 4.5: AI-assisted KPI generation workflow (compact vertical layout).

Figure 4.6: Prototype UI: KPI generation and editing interface (TODO: replace with Figma export).

```
def gen_k_measurement_kpi(category: str, subcategory: str, k: int = 10):
    """Generate k distinct SMART IndicatorKPI drafts for a given
       category/subcategory."""
    ...
```

4.2.1 Example KPI output

Table 4.1 shows a compact example of a generated KPI instance (excerpt of the full JSON structure). The complete structured object is included in Appendix A to provide full field-level transparency without exceeding page width in the main chapter.

Table 4.1: Compact example of a generated KPI (excerpt)

Field	**Example value (excerpt)**
Category / Subcategory	Agrar & Agrar Tech / "Farm-to-Fork" transparency
Short-term indicator	Consumer scan rate per 100 units sold
Short-term question	"Did you scan the code on your last purchase to view origin data?"
Measurement / Unit	QR/NFC analytics over 30 days; scans per 100 units
Long-term indicator	Share of transactions with at least one origin-data view
SDG alignment	Primary: 12.8; Secondary: 9.c

4.3 Human-in-the-Loop Evaluation

Generated KPIs and assessment forms are embedded in a human-in-the-loop workflow:

- Domain experts review and adjust wording, units, and feasibility of data collection.
- Stakeholders validate relevance and alignment with intended public value contributions.
- Final KPIs are versioned and approved before they are used for data collection.

A simplified example of how expert feedback refines a KPI is shown in Table 4.3.

Table 4.2: Example of human-in-the-loop feedback and KPI refinement

Version	**Description**	**Comment / Rationale**
Original KPI	"Increase the number of platform users."	Too generic; unclear target group and measurement definition.
Expert feedback	"Specify target group, baseline, target, and timeframe; clarify active vs. registered."	Enforces SMART formulation and improves linkage to theory of change.
Refined KPI	"Increase active monthly users among early-stage impact startups from 80 to 160 within 12 months."	Specific, measurable, and time-bound; focuses on meaningful usage.

Figure 4.7: Prototype UI: review and commenting interface for KPI validation (TODO: replace with Figma export).

4.4 Integration with the Public Value Academy Platform

- Supports workshops and structured reflection around public value.
- Embeds expert review and stakeholder feedback into KPI workflows.
- Enables iterative improvement through versioned profiles, KPIs, and dashboards.

Figure 4.8: Platform integration: navigation and embedding of IMM artefact components (TODO: replace with Figma export).

4.5 Ethical and Governance Considerations

- GDPR-oriented handling of participant and project data (data minimisation, consent, access control).
- Transparency through structured outputs, traceable mappings, and auditable revision logs.
- Human oversight enforced in critical steps (profile validation, KPI approval, SDG alignment).

4.6 Next Steps and Data Analysis

To move from prototype to systematic evaluation, the following analysis steps are envisaged:

- **Aggregation and cleaning:** deduplication, missing value handling, plausibility checks, and aggregation by project/cohort/time period.
- **Quantitative analysis:** descriptive statistics and simple trend or pre–post comparisons; normalisation where appropriate (e.g. per beneficiary).
- **Qualitative analysis:** NLP-supported thematic clustering combined with manual coding to identify recurring narratives and unintended effects.
- **Public value integration:** mapping results onto public value dimensions with human-validated narrative summaries.

A scoring rubric can be used to improve interpretability for stakeholders (e.g. underperforming / on-track / exceeding expectations mapped to a 1–5 scale), with aggregation at the level of public value dimensions via weighted averages.

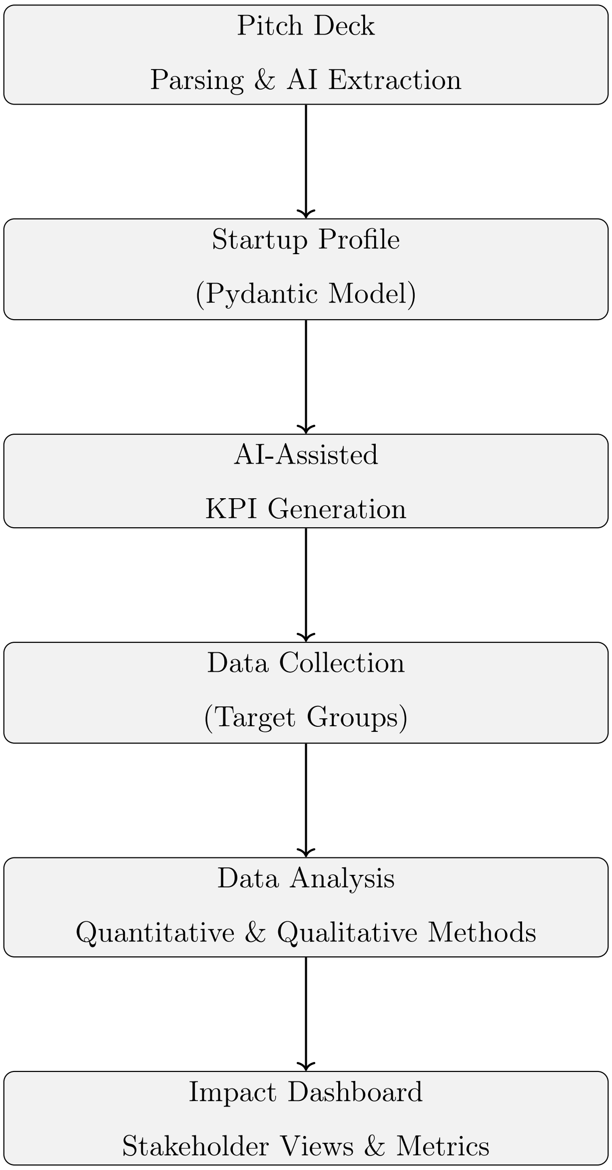

Figure 4.9: End-to-end workflow: from pitch deck parsing to impact dashboard.

Figure 4.10: Dashboard mock-up: summary view (TODO: replace with Figma export).

Figure 4.11: Dashboard mock-up: KPI detail view with trends (TODO: replace with Figma export).

4.7 Summary

This chapter demonstrated that the AI-supported IMM artefact can (1) onboard new projects via automated pitch deck parsing, (2) structure project narratives and metadata into reusable schemas, (3) generate auditable KPI drafts aligned with impact frameworks, (4) embed human validation loops for quality and legitimacy, and (5) support reporting and reflection through dashboard-oriented outputs. website="https://www.ecotrack.io", project$_beginning = "2023-04-15", turnover = 250000, profit =$
$45000, employers = 6, problem = "Smallandmedium-sizedcompanieslackaccessibletoolsto""measureandoptim$
$"Aworldwhereeverybusiness, regardlessofsize, canunderstand""andminimizeitsecologicalimpact.", mission =$
$"Provideaffordable, data-drivensustainabilityanalyticsto""empowercompaniesontheirpathtowardclimateneu$
$"ASaaSplatformcombiningautomateddataingestion, carbon""footprintanalytics, andactionablereductioninsigh$
$"Helpscompaniesreduceemissionsandoperatemore""sustainably, contributingtonationalandEUclimategoals.",$
$"Growingregulatorypressureandmarketdemandfortransparent""sustainabilitystrategies.", value_1 =$
$"Transparency", value_2 = "Sustainability", value_3 = "Innovation", target_group = "Smallandmedium-$
$sizedcompaniesinmanufacturingand""logistics.")$

This example illustrates the level of detail captured during onboarding and demonstrates how unstructured pitch deck content is transformed into a structured project profile.

4.8 Indicator and KPI Generation

Following onboarding, the IMM phase begins:

- A pre-generated library of over 1,600 indicators serves as reference.
- The function `gen_k_measurement_kpi()` generates SMART KPIs for specific categories and subcategories.
- Each KPI contains short-term and long-term goals, measurement methods, units, survey questions, and justification.
- Optional secondary goals are created if multiple outcomes are detected in the input text.

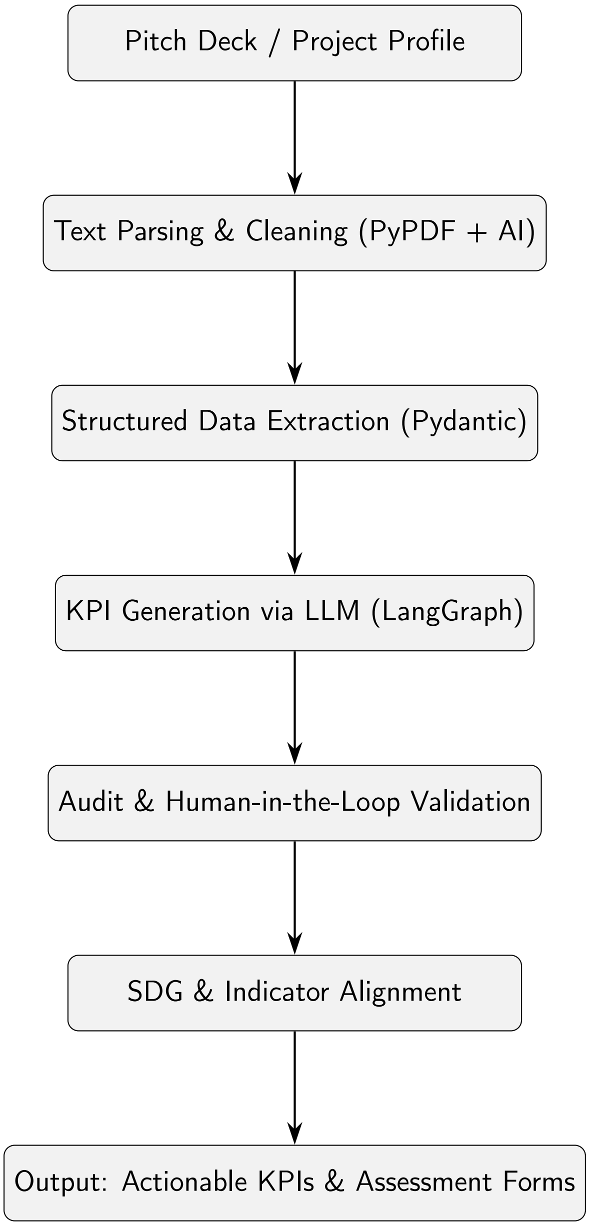

Figure 4.12: AI-Assisted KPI Generation Workflow (vertical layout for compact page fit).

Figure 4.13: Screenshot of the KPI generation and editing interface, showing short-term and long-term fields.

```
def gen_k_measurement_kpi(category: str, subcategory: str, k: int = 10):
    """
    Generate k distinct SMART IndicatorKPIs for a given
        category/subcategory using an LLM.
    """
    # Structured LLM output via API
    ...
```

A complete example of a generated IndicatorKPI, as produced by this function for the category *Environmental Impact* and subcategory *Energy Efficiency*, is shown below. The structure follows the SMART logic and already includes fields for downstream impact reporting:

```
example_kpi = {
            "category": "Agrar & Agrar Tech",
            "sub_category": "\"Farm-to-Fork\" Transparency",
            "goal": [
                "Consumers access product origin data"
            ],
            "short_term_goal_1": "Consumers scan traceability codes",
            "short_term_indicator_1": "Consumer scan rate per 100 units
                sold",
            "short_term_question_1": "Did you scan the code on your last
                purchase to view origin data?",
            "type_of_short_term_question_1": "single_choice",
            "answer_options_short_term_question_1": [
                "Yes",
```

```
    "No"
],
"measurement_method_short_term_question_1": "Count unique
   consumer scan events from QR/NFC analytics over the last 30
   days, divide by units sold for the same SKUs and period,
   multiply by 100.",
"unit_method_short_term_question_1": "scans per 100 units",
"justification_method_short_term_question_1": "Scanning a
   product code indicates access to origin data. Scan rate is a
   direct, low-cost proxy for transparency uptake using
   existing app or GS1 Digital Link analytics.",
"source_method_short_term_question_1": "GS1 Digital Link;
   ISO/IEC 18004 QR Code; IRIS+ Product/Service Users; SDG
   12.8",
"long_term_goal_1": [
    "Consumers view origin data during purchase decisions"
],
"long_term_indicator_1": [
    "Transactions with at least one origin data view share"
],
"long_term_question_1": "How often do you view origin
   information when buying this product?",
"type_of_long_term_question_1": "single_choice",
"answer_options_long_term_question_1": [
    "Never",
    "Rarely",
    "Sometimes",
    "Often",
    "Always"
],
"measurement_method_long_term_question_1": "Link scan events to
   sales by batch ID within a 24 72  hour window to estimate
   the share of transactions with at least one origin data
   view.",
"unit_method_long_term_question_1": "percent of transactions",
"justification_method_long_term_question_1": "Viewing origin
   information during the shopping journey reflects meaningful
   transparency use beyond curiosity scans.",
"source_method_long_term_question_1": "GS1 EPCIS 2.0; ISO 22005
   Traceability in feed and food chain; SDG 12.8",
```

```
    "short_term_goal_2": [],
    "short_term_indicator_2": [],
    "short_term_question_2": "",
    "type_of_short_term_question_2": "single_choice",
    "answer_options_short_term_question_2": [],
    "measurement_method_short_term_question_2": "",
    "unit_method_short_term_question_2": "",
    "justification_method_short_term_question_2": "",
    "source_method_short_term_question_2": "",
    "long_term_goal_2": [],
    "long_term_indicator_2": [],
    "long_term_question_2": "",
    "type_of_long_term_question_2": "single_choice",
    "answer_options_long_term_question_2": [],
    "measurement_method_long_term_question_2": "",
    "unit_method_long_term_question_2": "",
    "justification_method_long_term_question_2": "",
    "source_method_long_term_question_2": "",
    "sdg_target_1": "12.8 Promote information for sustainable
       lifestyles",
    "sdg_target_2": "9.c Access to ICT",
    "sdg_target_3": "",
    "question": "",
    "type_of_question": "single_choice",
    "answer_options": [],
    "measurement_method": "",
    "unit": "",
    "justification": "",
    "source": ""
}
```

In the final thesis, this type of KPI instance can be placed in the appendix as a reference example for readers and practitioners who wish to understand the exact structure produced by the artefact.

4.9 Human-in-the-Loop Evaluation

Generated KPIs and assessment forms are:

- Reviewed by domain experts and stakeholders before deployment.
- Distributed to target groups for feedback and data collection.

- Iteratively refined for alignment with project objectives and public value principles.

A simplified example of how human-in-the-loop feedback affects KPI refinement is shown in Table 4.3. In this example, an initially broad KPI is made more specific, measurable and public-value oriented based on expert and stakeholder input.

Table 4.3: Example of human-in-the-loop feedback and KPI refinement

Version	**Description**	**Comment / Rationale**
Original KPI	"Increase the number of platform users."	Stakeholders considered this too generic and insufficiently connected to the intended public value contribution. No baseline or target group was specified.
Expert feedback	"Specify the target group (e.g. early-stage impact startups), include a numerical baseline and target, and define a timeframe. Clarify whether active or registered users are meant."	The expert feedback emphasised the need for a SMART formulation and clearer linkage to the project's theory of change.
Refined KPI (accepted)	"Increase the number of *active monthly users* among early-stage impact startups from 80 to 160 within 12 months after launch of the platform."	The refined KPI is specific, measurable and time-bound. It focuses on active usage (behavioural change) rather than mere registration and is directly aligned with the intended public value of strengthening the impact startup ecosystem.

This illustrates how the artefact embeds human oversight: the LLM proposes an initial KPI, which is then systematically adapted based on qualitative expert feedback and stakeholder perspectives before it is finalised and used for assessment.

4.10 Integration with the Public Value Academy Platform

- Supports workshops and structured reflection around public value.

- Embeds human-in-the-loop feedback directly into workflows.
- Enables iterative improvement of AI-supported tools.

The IMM artefact is designed as a service component within the Public Value Academy platform. Structured profiles and KPIs can be reused across workshops and learning formats, and dashboards provide a shared basis for discussion between project teams, coaches and investors.

4.11 Ethical and Governance Considerations

- GDPR-compliant handling of participant and project data.
- Explainable AI (XAI) applied throughout parsing, KPI generation, and SDG mapping.
- Human oversight enforced in all critical stages.

In addition to technical measures, the artefact incorporates transparency documentation (e.g. model cards and data flow diagrams) and explicit consent mechanisms for participants whose responses feed into the impact measurement process.

4.12 Next Steps and Data Analysis

To move from prototype to systematic evaluation and learning, the following methodological steps for analysing collected KPI and survey data are envisaged:

- **Aggregation and cleaning of responses from target groups:** Raw responses from online surveys and platform interactions are first cleaned (removal of duplicates, handling of missing values, basic plausibility checks) and aggregated at the level of projects, cohorts and time periods. This ensures that subsequent analyses are based on a consistent and reproducible dataset.
- **Statistical analysis for quantitative indicators:** For numeric KPIs (e.g. energy consumption per workspace, number of active users, share of underrepresented groups), descriptive statistics (means, medians, standard deviations) and simple inferential methods (e.g. pre–post comparisons, trend analysis over time) are applied. Where appropriate, normalisation procedures (e.g. per beneficiary, per euro invested) are used to improve comparability across projects.
- **NLP or thematic analysis for qualitative inputs:** Open-ended survey responses and qualitative feedback from workshops are analysed using basic natural language processing (NLP) tools

(e.g. keyword extraction, clustering of frequently mentioned themes) combined with manual coding. This supports the identification of recurring narratives, perceived benefits and unintended side effects that are not captured by numeric indicators alone.

- **Integration of findings with public value dimensions:** Both quantitative and qualitative findings are mapped onto selected public value dimensions (e.g. equity, sustainability, participation, innovation). For each dimension, a short narrative summary is generated (assisted by the LLM but validated by humans) describing how the project contributes to, or potentially conflicts with, that dimension. This strengthens the alignment between operational KPIs and the broader normative framework.

On top of the analytical procedures, thresholds and a scoring rubric are required in order to make results interpretable for different stakeholders:

- **Thresholds and scoring rubric for KPI performance and public value metrics:** For each KPI, performance bands are defined (e.g. *underperforming*, *on track*, *exceeding expectations*) based on relative improvement compared to baseline (for instance, $< 10\%$, 10–30%, $> 30\%$ improvement). These bands are then translated into a standardised 1–5 score. At the level of public value dimensions, scores from relevant KPIs are aggregated (e.g. via weighted averages) to obtain an overall dimension score. This allows projects to be compared over time and across cohorts while keeping the underlying assumptions explicit.

- **Impact dashboard mock-up:** The impact dashboard is conceived as a web-based interface that displays:

  - a *summary view* with overall public value scores per project (e.g. radar chart or bar chart),
  - a *KPI detail view* showing trends over time (line charts) and numerical targets vs. actuals,
  - filters for time period, cohort, target group and public value dimension,
  - qualitative excerpts (e.g. representative quotes) linked to specific KPIs or dimensions.

  The dashboard is not only a reporting tool but also a starting point for reflection in workshops, enabling users to drill down from aggregate scores to the underlying data and narratives.

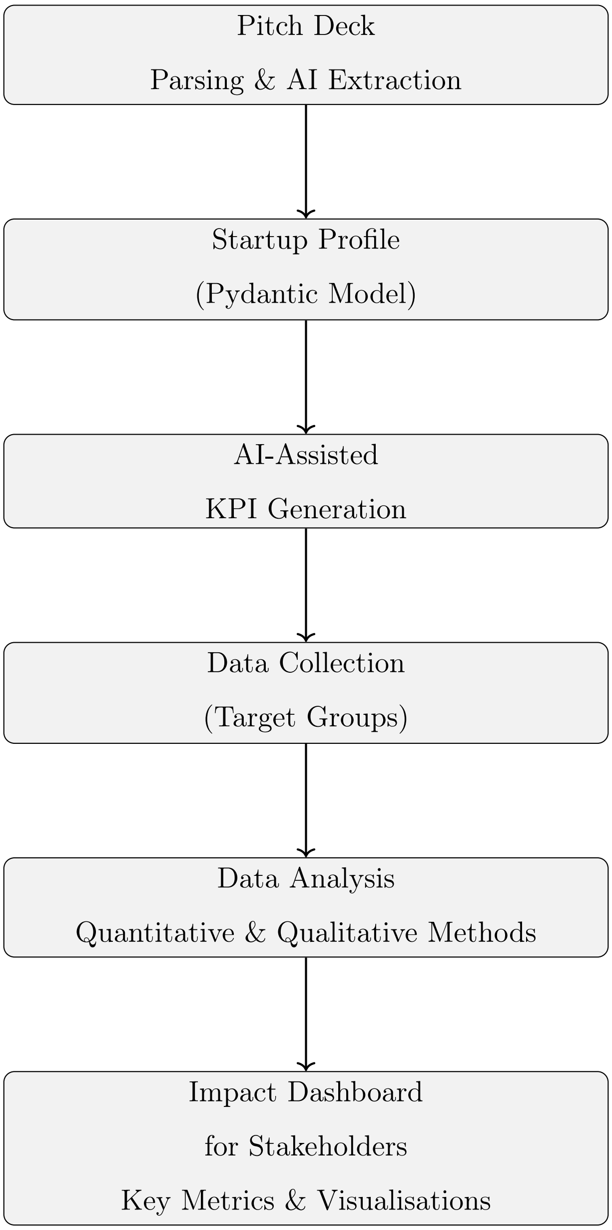

Figure 4.14: End-to-end vertical workflow: from pitch deck parsing to impact dashboard.

4.13 Summary

This chapter demonstrates that the AI-supported IMM artefact can:

- efficiently onboard new projects using automated pitch deck parsing,
- generate structured project profiles with AI-assisted data extraction,
- produce actionable KPIs aligned with SDGs and recognised impact frameworks,
- maintain a human-in-the-loop workflow for quality assurance, interpretability and stakeholder validation,

- and feed collected data into a dashboard for actionable insights for impact investors and project teams.

In addition, the chapter has outlined concrete methods for quantitative and qualitative data analysis, proposed a scoring rubric for KPI and public value assessment, provided examples of generated KPIs and expert feedback, and sketched the design of an impact dashboard. Together, these elements show how the artefact can serve as a practical bridge between AI-supported analytics and normative public value deliberation.

Chapter 5

Demonstration and Evaluation

This chapter presents the **demonstration and evaluation** of the AI-supported IMM artefact developed in Chapter 4. The framework was tested using synthetic project data, anonymized pitch materials, and stakeholder walkthroughs, to assess its feasibility, transparency, comparability, and usability.

5.1 Overview of Demonstration

The artefact was applied in the context of *Inluma* to demonstrate its functionality:

- **Semantic clustering**: grouping unstructured narrative inputs into interpretable themes.
- **KPI derivation pipeline**: generating auditable KPIs from structured problem statements.
- **SDG mapping**: aligning project objectives with Sustainable Development Goals and providing transparent justifications.

The demonstration highlights the artefact's capacity to augment human judgment while remaining **transparent and interpretable**.

5.2 Narrative Clustering Results

Narratives from over 20 public innovation cases were embedded using an OpenAI text-embedding model (e.g. `text-embedding-ada-002` or a comparable sentence embedding model), reduced via UMAP, and clustered with HDBSCAN.

Key Observations

- Clusters revealed cross-cutting themes such as citizen participation, data ethics, and local climate action.
- LLM-based summarisation was used to generate interpretable cluster labels for stakeholders.
- Clustering facilitated structured overviews of diverse inputs, supporting reflection and discussion.

5.2.1 UMAP-Based Clustering of Qualitative Responses

To support exploratory analysis of qualitative survey responses, responses can be embedded (e.g. using sentence embeddings) and projected into two dimensions using UMAP. This enables visual inspection of response similarity and provides a basis for clustering. Clusters are then summarised using short labels and representative example statements.

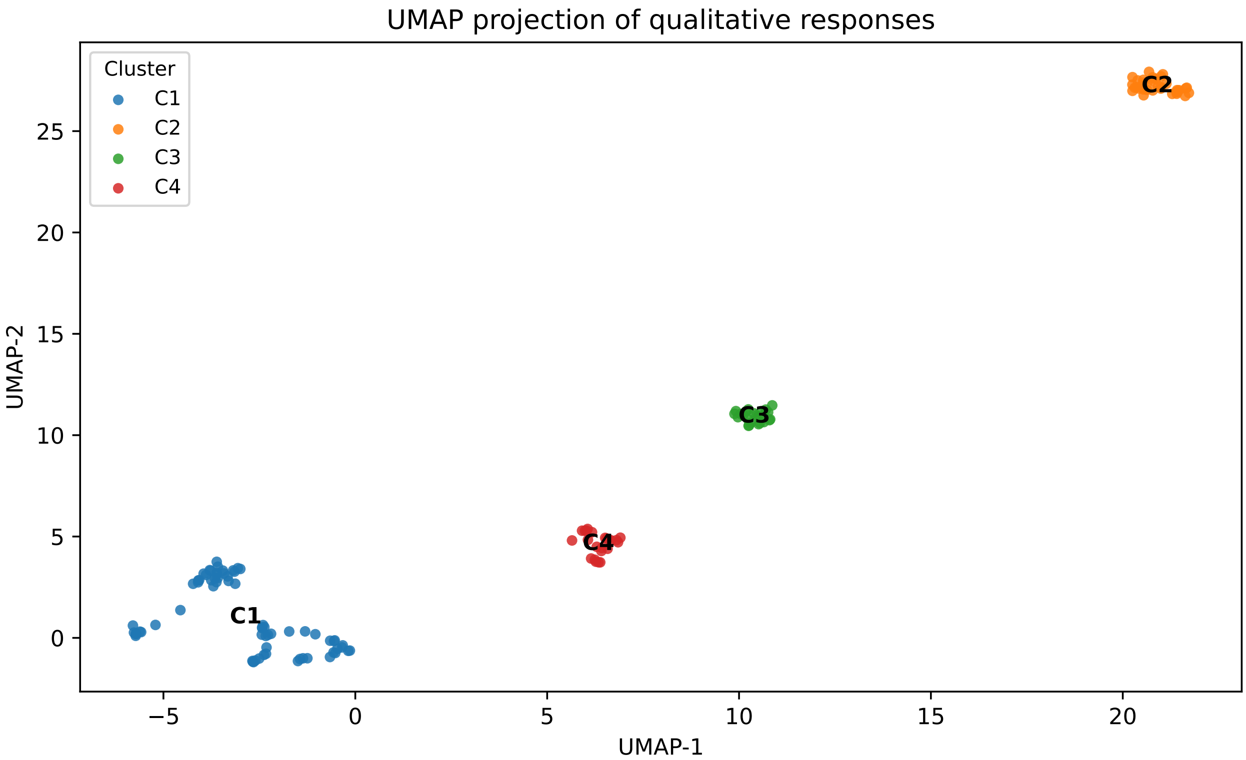

Figure 5.1: UMAP projection of qualitative responses with cluster assignments.

ID	Size	Cluster label/theme (auto)	Representative quotes
C1	69	Top terms: care, scanned, code, check, want, certific	"I scanned the code to check animal welfare." | "I scanned the code to check origi
C2	34	Top terms: information, decision, purchase, influence	"The information could influence in the future my purchase decision." | "The inforr
C3	24	Top terms: didn, scan, notice, hurry, sure, safe, wasn	"I didn't scan because I wasn't sure it was safe." | "I didn't scan because I wasn't s
C4	23	Top terms: scanning, straightforward, needed, confus	"Scanning was frustrating; I gave up." | "Scanning was straightforward; I got more

Figure 5.2: Example cluster summary table derived from UMAP + clustering.

5.3 SDG Mapping Results

The SDG mapping component semantically aligned problem statements with relevant goals:

- In a small manual benchmark, the mapping agreed with expected SDG tags in approximately 85% of cases.
- LLM-based justifications enhanced transparency and trust by making the reasoning behind each SDG assignment explicit.

Example: *"This project addresses SDG 11 (Sustainable Cities and Communities) by increasing civic data accessibility for participatory urban governance."*

KPI / Indicator	Primary SDG target	Secondary SDG target	Justification (short)	Framework reference
Consumer scan rate per 100 units sold	12.8 (Ensure people have	9.c (Access to ICT)	Scanning indicates consumer engagement with origin data. I	SDG 12.8; GS1 Digital Link; IR
Share of transactions with an origin-dat	12.8	12.6 (Encourage compan	Views during shopping reflect practical use of transparency i	SDG 12.8; SDG 12.6; GS1 EPC
Share of products with verified traceabi	12.6	9.4 (Upgrade infrastructu	Batch-level traceability improves accountability in productior	SDG 12.6; SDG 9.4; GS1 EPC
Reduction in customer-reported informa	12.8	16.6 (Effective, accounta	Perceived information sufficiency is a direct user-level outcor	SDG 12.8; SDG 16.6 (transpa
Active monthly users among impact sta	9.3 (Increase access of sı	8.3 (Support productive a	Active usage by early-stage ventures indicates adoption of e	SDG 9.3; SDG 8.3; IRIS+ (Bus
Share of SMEs reporting at least one im	12.6	17.17 (Partnerships for sı	Regular reporting operationalises sustainability management	SDG 12.6; SDG 17.17; impact
Decrease in process energy use per uni	7.3 (Double the global ra	13.2 (Integrate climate m	Energy intensity improvements reduce emissions and operat	SDG 7.3; SDG 13.2; GHG Prot
Share of participants from underreprese	10.2 (Promote social, eco	5.5 (Ensure women's full	Participation rates indicate whether capacity-building format	SDG 10.2; SDG 5.5; DEI repo

Figure 5.3: Example SDG mapping table.

5.4 KPI Derivation Pipeline Results

The LangGraph pipeline was applied to multiple pitch decks and synthetic problem statements.

Example Output

- **Problem:** "Limited mobility access for rural elderly populations."
- **Mapped SDG:** SDG 11
- **KPI:** *"Percentage increase of rural elderly residents with weekly access to on-demand mobility services."*

Audit Loop Observations

- KPIs with quality scores below 80% were regenerated in 42% of test runs.
- Common issues: vague definitions, misalignment with outcomes.
- Audit loops proved essential for maintaining consistency and alignment.

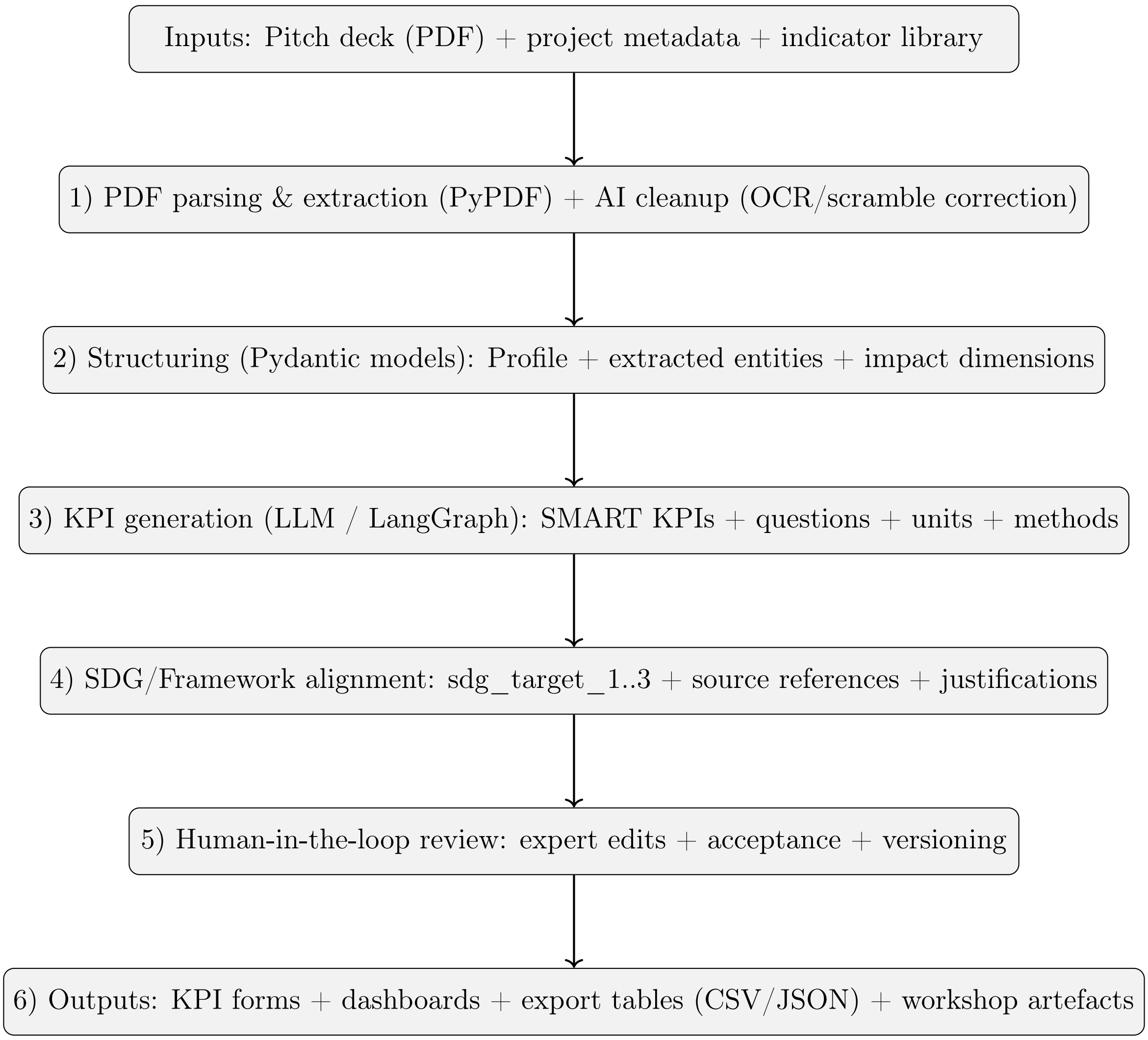

Figure 5.4: Pipeline flow diagram of the AI-supported IMM artefact from pitch deck ingestion to validated KPI outputs.

5.5 Human-in-the-Loop Feedback

Stakeholder walkthroughs confirmed the importance of **human validation**:

- Manual editing of AI-generated problem statements was often needed.
- Feedback loops validated SDG and KPI proposals.
- Alternative perspectives were incorporated through iterative discussion.

This reinforces the artefact's design principle: AI as a **decision-support tool**, not a replacement for human expertise.

5.6 Transparency and Explainability

Each pipeline run logged **decision paths and rationales**, supporting audits and ethical review:

- Justifications captured at SDG mapping, indicator selection, and KPI generation.
- SHAP-inspired and GPT-based explanations were used to support interpretability of selected model outputs.
- Supports accountability and trust in AI-supported evaluation processes.

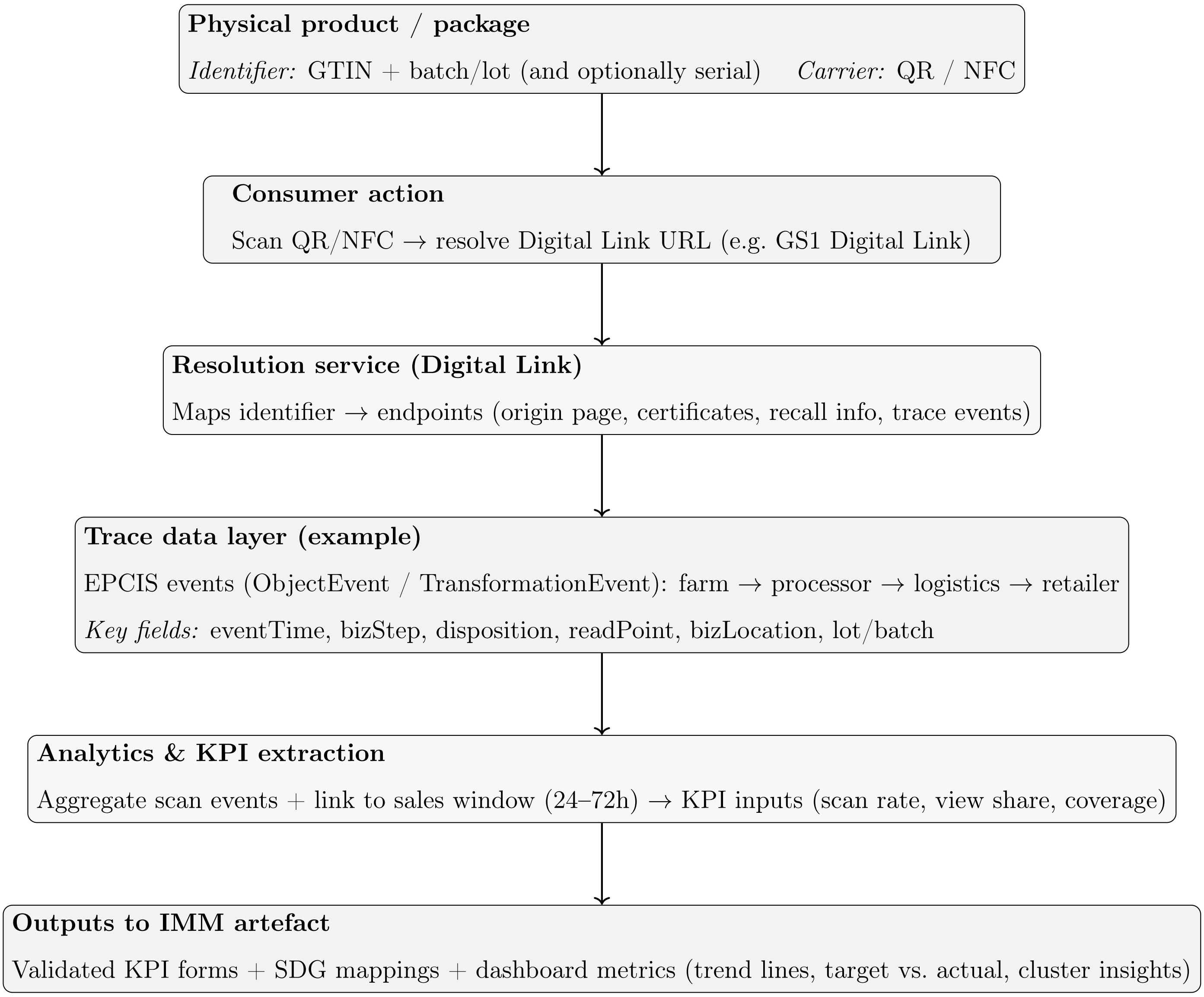

Figure 5.5: Example trace schematic: from a physical product identifier (QR/NFC) to trace events and KPI-ready outputs.

5.7 Evaluation Summary

The artefact was evaluated against pre-defined DSR criteria:

- **Feasibility:** All modules operated successfully on test datasets.
- **Transparency:** Justifications and audit loops increased interpretability.
- **Comparability:** Semantic clustering and KPI derivation facilitated consistent evaluation across cases.
- **Usability:** Stakeholders found outputs informative, with manageable human-in-the-loop requirements.

Key insights:

- AI tools can support reflective, value-aligned impact assessment.
- Human-in-the-loop mechanisms are essential for interpretability and trust.
- Modular design allows adaptation to different data sources and contexts.

The next chapter discusses these results in the context of existing frameworks, reflecting on theoretical and practical implications.

Chapter 6

Conclusion

This chapter summarizes the key findings of the thesis, reflects on the theoretical, practical, and methodological contributions, and outlines directions for further research and development.

6.1 Summary of Findings

The thesis addressed the research question:

> *How can Artificial Intelligence support and improve Impact Measurement and Management in social enterprises and public sector innovation contexts through an artefact developed using the Design Science Research methodology?*

The study demonstrates that AI can meaningfully support Impact Measurement and Management when embedded within a transparent, human-in-the-loop design. Key insights include:

- **AI-Supported IMM:** Natural language processing and semantic analysis enable the structured use of both qualitative and quantitative impact data, addressing limitations of indicator-driven IMM approaches.
- **Human-in-the-Loop Design:** Continuous stakeholder validation is essential to maintain interpretability, legitimacy, and alignment with public value and social impact objectives.
- **Artefact Validation:** The prototypical implementation within *Inluma* demonstrated feasibility, transparency, and practical relevance according to Design Science evaluation criteria.
- **Integration of Frameworks:** Combining IMM principles, AI methods, and public value considerations supports a more holistic and reflective evaluation of social innovation initiatives.

6.2 Theoretical, Practical, and Methodological Contributions

Theoretical Contribution:

- Extends existing work on AI-supported IMM by illustrating how qualitative narratives and quantitative indicators can be integrated through AI-assisted, human-in-the-loop processes.
- Contributes design knowledge on how AI, IMM frameworks, and public value concepts can be coherently linked in social enterprise and public innovation settings.

Practical Contribution:

- Demonstrates a prototypical AI-enabled artefact capable of generating interpretable KPIs, clustering narrative inputs, and mapping initiatives to SDGs.
- Provides social enterprises and innovation intermediaries with a structured, semi-automated approach to enhance transparency, comparability, and evidence-informed decision-making.

Methodological Contribution:

- Shows how Design Science Research can be applied to the development and evaluation of AI-supported artefacts in complex, value-driven domains.
- Highlights the importance of iterative evaluation cycles and human oversight in ensuring relevance and ethical alignment.

6.3 Limitations

- The artefact is prototypical and not intended as a market-ready system; scalability, robustness, and long-term effects remain untested.
- Evaluation relied on synthetic and anonymized project data as well as a limited number of stakeholder walkthroughs.
- The current implementation is tailored to the *Inluma* context and may require adaptation for other organizational or sectoral settings.

6.4 Outlook and Future Work

Future research and development may include:

- Integration with larger datasets and live project pipelines to assess longitudinal impact development.
- Extension of AI capabilities for deeper qualitative analysis, such as narrative change over time, sentiment dynamics, or stakeholder perspective modeling.
- Adaptation of the artefact for broader application in public administration, social entrepreneurship ecosystems, and international development contexts.
- Further refinement of human-in-the-loop workflows to balance automation, transparency, and participatory decision-making.

6.5 Closing Remarks

This thesis demonstrates that AI can act as a supportive cognitive tool in Impact Measurement and Management, enhancing sense-making and comparability while preserving human judgment and ethical oversight. By integrating IMM frameworks, AI methods, and public value considerations, the proposed artefact offers a practical and reflective approach to evaluating social innovation initiatives in complex public and social contexts.

Bibliography

Bamman, D., & Underwood, T. (2020). Large language models for text mining and analysis. *Computational Humanities Research*, *1*, 1–20.

Benington, J., & Moore, M. H. (2011). *Public value: Theory and practice*. Palgrave Macmillan. https://doi.org/10.1007/978-0-230-36931-3

Berlin, C. (2024). *Co-creation for urban innovation: Participatory design in berlin* (tech. rep.). CityLAB Berlin. Retrieved July 3, 2025, from https://www.citylab-berlin.org/en/projects

Braun, V., & Clarke, V. (2019). Thematic analysis: A practical guide. *SAGE Publications*.

Bright, J., Enock, F. E., Esnaashari, S., Francis, J., Hashem, Y., & Morgan, D. (2024). *Generative ai is already widespread in the public sector*. Retrieved July 12, 2025, from https://arxiv.org/abs/2401.01291

Brown, I., Marsden, C., & Binns, R. (2020). Algorithmic accountability: A primer. *Internet Policy Review*, *9*(4), 1–26. https://doi.org/10.14763/2020.4.1504

Campello, R. J. G. B., Moulavi, D., & Sander, J. (2013). Density-based clustering based on hierarchical density estimates. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 160–172.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. Retrieved July 9, 2025, from https://arxiv.org/abs/1810.04805

DIN. (2023). *Din spec 92001-3: Artificial intelligence – ethical standards* (tech. rep.). DIN. Retrieved July 3, 2025, from https://www.din.de/en/innovation-and-research/ai-standards

Dwivedi, R., Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., et al. (2023). Explainable ai (xai): Core ideas, techniques, and solutions. *ACM computing surveys*, *55*(9), 1–33.

Ebrahim, A., & Rangan, V. K. (2014a). What impact? a framework for measuring the scale and scope of social performance. *California Management Review*, *56*(3), 118–141. https://doi.org/10.1525/cmr.2014.56.3.118

Ebrahim, A., & Rangan, V. K. (2014b). What impact? a framework for measuring the scale and scope of social performance. *California Management Review*, *56*(3), 118–141.

Epstein, M. J., & Yuthas, K. (2014). *Measuring and improving social impacts: A guide for nonprofits, companies, and impact investors*. Berrett-Koehler Publishers.

European Commission. (2023). Proposal for a regulation laying down harmonised rules on artificial intelligence (artificial intelligence act) [Accessed: 2025-07-20]. https://artificialintelligenceact.eu

Face, H. (2023). German bert model: Bert-base-german-cased. Retrieved July 3, 2025, from https://huggingface.co/bert-base-german-cased

Feng, V. W., Boyd-Graber, J., & Zhai, C. (2019). Mixed methods for computational social science. *Communications of the ACM*, *62*(4), 72–81.

for Economic Affairs, F. M., & Action, C. (2022). *Germany's digital strategy 2025* (tech. rep.). BMWK. Retrieved July 3, 2025, from https://www.bmwi.de/Redaktion/EN/Dossier/digital-strategy-2025.html

for Economic Co-operation, O., & Development. (2020a). The innovation system of the public sector. Retrieved July 3, 2025, from https://www.oecd.org/publications/the-innovation-system-of-the-public-sector-3c8034f7-en.htm

for Economic Co-operation, O., & Development. (2020b). *Measuring the impact of social investments: Oecd report*. Organisation for Economic Co-operation and Development. Retrieved July 3, 2025, from https://www.oecd.org/social/social-investment.htm

for Economic Co-operation, O., & Development. (2023). *Artificial intelligence in public policy: Opportunities and challenges* (tech. rep.). OECD. Retrieved July 3, 2025, from https://www.oecd.org/governance/ai-in-public-policy

gAG, P. (2022). Wirkung lernen: Grundlagen der wirkungsorientierung. Retrieved July 3, 2025, from https://www.phineo.org/magazin/wirkung-lernen

gAG, P. (2023). Impact measurement and management framework. Retrieved July 3, 2025, from https://www.phineo.org/en/impact-measurement/

(GIIN), G. I. I. N. (2023). *2023 annual impact investor survey*. GIIN. https://thegiin.org/research/publication/impinv-survey-2023

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2004). Design science in information systems research. *MIS Quarterly*, *28*(1), 75–105.

Holzinger, A. (2016). Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics*, *3*(2), 119–131.

Institute, F. (2023). *Impact measurement in smart cities: Tools and applications* (tech. rep.). Fraunhofer IAIS. Retrieved July 3, 2025, from https://www.iais.fraunhofer.de/en/research/impact-measurement.html

LangChain. (2025a). *Building the graph*. Retrieved July 21, 2025, from https://python.langchain.com/docs/versions/migrating_memory/long_term_memory_agent/#build-the-graph

LangChain. (2025b). *Similarity search with euclidean disctance*. Retrieved July 21, 2025, from https://python.langchain.com/docs/integrations/vectorstores/timescalevector/#1-similarity-search-with-euclidean-distance-default

LangChain. (2025c). *Structured outputs*. Retrieved July 21, 2025, from https://python.langchain.com/docs/concepts/structured_outputs/

Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in Neural Information Processing Systems*, *30*.

Marr, B. (2018). *Data strategy: How to profit from a world of big data, analytics and the internet of things*. Kogan Page Publishers.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*. https://arxiv.org/abs/1802.03426

Mikhaylov, S. J., & Esteve, M. (2018). Artificial intelligence for the public sector: Opportunities and challenges of cross-sector collaboration. *Philosophical Transactions of the Royal Society A*, *376*(2128). https://doi.org/10.1098/rsta.2018.0082

Moore, M. H. (1995). *Creating public value: Strategic management in government*. Harvard University Press.

Nicholls, A., Paton, R., & Emerson, J. (2012). *Social finance*. Oxford University Press.

OECD. (2022). Oecd principles on artificial intelligence [Accessed: 2025-07-20]. https://oecd.ai/en/ai-principles

OpenAI. (2023a). Gpt-4 technical report [Accessed: 2025-07-19]. https://openai.com/research/gpt-4

OpenAI. (2023b). Openai text embedding models. https://platform.openai.com/docs/guides/embeddings

Patton, M. Q. (2011). *Developmental evaluation: Applying complexity concepts to enhance innovation and use*. Guilford Press.

Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of Management Information Systems*, *24*(3), 45–77.

PHINEO gAG. (2023). The iooi model: From inputs to impact [Accessed: 2025-07-20]. https://www.phineo.org/magazin/impact-orientation-and-impact-assessment

Reimers, N., & Gurevych, I. (2019). Sentence-bert: Sentence embeddings using siamese bert-networks. *EMNLP 2019*. https://arxiv.org/abs/1908.10084

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). "why should i trust you?": Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 1135–1144.

Russell, S., & Norvig, P. (2016). *Artificial intelligence: A modern approach*. Pearson Education Limited.

SHAP. (2024). *Explainable ai with shap*. Retrieved July 21, 2024, from https://shap.readthedocs.io/en/latest/example_notebooks/overviews/An%20introduction%20to%20explainable%20AI%20with%20Shapley%20values.html

Sun, T., & Medaglia, R. (2019). Mapping the challenges of artificial intelligence in the public sector: Evidence from public healthcare. *Government Information Quarterly*, *36*(2), 368–383. https://doi.org/10.1016/j.giq.2018.09.008

Union, E. (2024). The eu artificial intelligence act: Regulation 2024/1689. Retrieved July 3, 2025, from https://eur-lex.europa.eu/eli/reg/2024/1689

United Nations. (2024). *Sustainable development goals*. Retrieved July 21, 2025, from vahttps://sdgs.un.org/ue

UnternehmerTUM. (2023a). Guide to impact management for startups. Retrieved July 3, 2025, from https://www.unternehmertum.de/impact-management

UnternehmerTUM. (2023b). Impact guide: Wirkungsorientierung für startups. Retrieved July 3, 2025, from https://www.unternehmertum.de/impact-guide

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*, 2579–2605.

Wirtz, B. W., Weyerer, J. C., & Mueller, C. (2020). Public administration in the age of digitalization – from e-government to smart government. *International Journal of Public Administration*, *43*(10), 832–844. https://doi.org/10.1080/01900692.2019.1619810

Appendix A

Additional Data

This appendix contains supplementary materials that support the AI-supported IMM artefact described in Chapters 4 and 5. It includes synthetic project profile examples, generated KPIs, audit log excerpts, and dashboard mock-ups.

Example Project Profiles

Below are structured outputs from the `Profile` Pydantic model for three synthetic sample projects. These illustrate the type of information extracted from pitch decks and structured for KPI generation.

The following profiles are synthetic examples generated for demonstration purposes:


```
{
  "startup_name": "GreenFields AgTech",
  "legal_form": "GmbH",
  "founder_first_name": "Anna",
  "founder_last_name": "Schmidt",
  "founder_gender": "female",
  "startup_email": "anna@greenfields.com",
  "startup_phone": "+49 123 456 789",
  "startup_city": "Berlin",
  "startup_country": "Germany",
  "startup_postcode": "10115",
  "website": "https://greenfields.com",
  "project_beginning": "2023-03-01",
  "turnover": 250000,
  "profit": 30000,
```

```
  "employers": 5,
  "problem": "Excessive synthetic nitrogen usage in small farms",
  "vision": "Reduce fertilizer use while maintaining yields",
  "mission": "Develop sustainable precision farming tools",
  "solution": "IoT soil sensors with AI-driven recommendations",
  "social_impact": "Promote sustainable agriculture and environmental health",
  "reason": "Reduce environmental pollution and farmer costs",
  "value_1": "Sustainability",
  "value_2": "Efficiency",
  "value_3": "Innovation",
  "target_group": "Smallholder farmers in Europe"
}

{
  "startup_name": "CareConnect Health",
  "legal_form": "UG",
  "founder_first_name": "David",
  "founder_last_name": "Müller",
  "founder_gender": "male",
  "startup_email": "contact@careconnect.health",
  "startup_phone": "+49 987 654 321",
  "startup_city": "Hamburg",
  "startup_country": "Germany",
  "startup_postcode": "20095",
  "website": "https://careconnect.health",
  "project_beginning": "2022-10-15",
  "turnover": 180000,
  "profit": -20000,
  "employers": 4,
  "problem": "Limited access to mental health support for young adults",
  "vision": "Affordable mental health support for everyone",
  "mission": "Deliver low-threshold digital mental health services",
  "solution": "Mobile app offering guided self-help and remote coaching",
  "social_impact": "Improved mental well-being and early intervention",
  "reason": "Address growing unmet mental health needs",
  "value_1": "Accessibility",
```

"value_2": "Empathy",
"value_3": "Trust",
"target_group": "Young adults aged 18-30"
}

{
"startup_name": "CarbonSight",
"legal_form": "GmbH",
"founder_first_name": "Sofia",
"founder_last_name": "Lindner",
"founder_gender": "female",
"startup_email": "hello@carbonsight.io",
"startup_phone": "+49 555 123 987",
"startup_city": "Munich",
"startup_country": "Germany",
"startup_postcode": "80331",
"website": "https://carbonsight.io",
"project_beginning": "2021-06-01",
"turnover": 520000,
"profit": 85000,
"employers": 9,
"problem": "SMEs lack tools to measure Scope 1 and 2 emissions",
"vision": "Transparent emissions data for every business",
"mission": "Enable data-driven climate action for SMEs",
"solution": "Carbon accounting software with automated data ingestion",
"social_impact": "Supports emission reductions and climate reporting",
"reason": "Regulatory pressure and sustainability demand",
"value_1": "Transparency",
"value_2": "Accuracy",
"value_3": "Responsibility",
"target_group": "European SMEs"
}

Generated KPIs / Indicators

Example KPI generated from the above project profiles:

```
{
  "category": "Agrar & Agrar Tech",
  "sub_category": "Sustainable Inputs",
  "goal": ["Reduce synthetic nitrogen application rate by 20 percent per hectare within 12 month
  "short_term_goal_1": "Users reduce synthetic nitrogen rate within 6 months",
  "short_term_indicator_1": "Share of users who reduced synthetic nitrogen rate compared to last
  "short_term_question_1": "Did you reduce your synthetic nitrogen application rate per hectare
  "type_of_short_term_question_1": "single_choice",
  "answer_options_short_term_question_1": ["Yes", "No", "Not applicable"],
  "measurement_method_short_term_question_1": "Self-reported comparison to baseline season recor
  "unit_method_short_term_question_1": "percent of users",
  "justification_method_short_term_question_1": "User-level rate reduction is an early signal of
  "source_method_short_term_question_1": "IRIS+ Agrochemical Use intensity; SDG 2.4",
  "long_term_goal_1": ["Users sustain a 20 percent lower synthetic nitrogen rate after 3 seasons
  "long_term_indicator_1": ["Kilograms of synthetic nitrogen applied per hectare per season"],
  "long_term_question_1": "How many kilograms of synthetic nitrogen per hectare did you apply th
  "type_of_long_term_question_1": "open_question",
  "measurement_method_long_term_question_1": "Farmer input logs normalized by field area",
  "unit_method_long_term_question_1": "kg N/ha",
  "justification_method_long_term_question_1": "Rate per area directly measures fertilizer press
  "source_method_long_term_question_1": "IRIS+ Agrochemical Use; FAO fertilizer statistics; SDG
}
```


The following excerpt illustrates how multiple outcomes are detected and represented; non-essential fields are omitted for brevity.


```
{
  "goal": [
    "Increase active platform usage",
    "Improve reporting consistency among startups"
  ],
  "short_term_goal_1": "Startups log in monthly",
  "short_term_goal_2": "Startups submit at least one KPI report per quarter",
  "justification": "Multiple outcomes detected: adoption and learning behaviour"
}
```

Human-in-the-Loop Audit Logs

> *"Short-term indicator was slightly ambiguous; refined wording to ensure farmers understand units and target."*

> *"SDG mapping verified: matches SDG 2.4 (Zero Hunger) and SDG 12.4 (Responsible Consumption)"*

Table A.1: Example anonymized audit log for KPI refinement

Stage	**Actor**	**Comment / Change**
Initial generation	LLM	KPI proposed with generic wording
Expert review	Domain expert	Clarified indicator unit and timeframe
SDG validation	Impact analyst	Confirmed SDG 2.4 and SDG 12.4 alignment
Final approval	Platform admin	KPI approved for deployment

Dashboard Mock-Up

The dashboard mock-up is populated with simulated KPI values based on realistic adoption and impact trajectories. It visualizes KPI performance over time, compares projects within a cohort, and highlights contributions to public value dimensions such as sustainability, equity and participation.

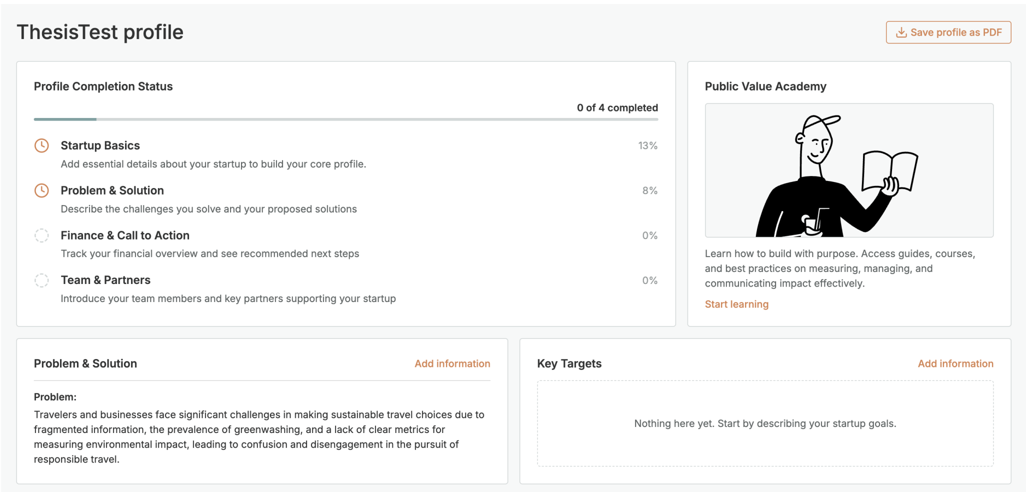

Figure A.1: Prototype Impact Dashboard Showing KPI Performance and Trends

Data Collection Instruments

Example survey question derived from KPI:

Short-term KPI: Share of users who reduced synthetic nitrogen rate **Question:** Did you reduce your synthetic nitrogen application rate per hectare this season compared to last season? **Type:** Single choice **Answer Options:** Yes / No / Not applicable **Measurement Method:** Self-reported comparison to baseline season recorded at onboarding

Long-term KPI Question: How many kilograms of synthetic nitrogen per hectare did you apply during the last growing season?

Open-ended Question: How has the project influenced your understanding or behaviour regarding sustainable farming practices?

Raw Analysis Outputs

Example aggregated results include mean KPI values per cohort, standard deviations, and trend indicators (e.g. improvement vs. baseline). Qualitative responses are clustered thematically (e.g. transparency, usability, trust) using embedding-based methods and manually validated.

Ethical and Governance Documentation

The AI-supported IMM artefact is designed and evaluated in accordance with ethical, legal, and governance principles, with a particular focus on data protection, transparency, and human oversight.

- **GDPR-compliant consent form template (redacted):** All participants whose data is processed within the artefact provide informed consent prior to data collection. The consent form clearly specifies the purpose of data processing, the types of data collected (e.g. survey responses, usage metrics), storage duration, and participants' rights (access, rectification, deletion). Personal identifiers are collected only where strictly necessary and are excluded from analytical outputs. A redacted version of the consent form is used in this appendix to avoid disclosure of sensitive or identifying information.
- **Anonymization and secure storage procedures:** Collected data is anonymized at the earliest possible stage. Direct identifiers (e.g. names, email addresses, phone numbers) are removed or replaced by pseudonymous IDs. Analytical datasets operate exclusively on aggregated or pseudonymized data at project or cohort level. All data is stored on secure servers with access control, and transmission is encrypted using standard protocols. Raw data access is restricted to authorized personnel involved in evaluation and quality assurance. Access to datasets, audit logs, and dashboard views is role-based, ensuring that only authorized analysts or administrators can access sensitive information.

- **Guidelines for human-in-the-loop oversight:** Human oversight is enforced at all critical stages of the IMM pipeline. AI-generated outputs (e.g. extracted profiles, generated KPIs, SDG mappings, narrative summaries) are treated as decision-support artifacts rather than final decisions. Domain experts review, edit, and approve outputs before they are used for assessment or reporting. All revisions and approvals are logged to ensure traceability and accountability. This human-in-the-loop approach mitigates risks of misinterpretation, bias, or inappropriate automation and aligns the artefact with responsible AI principles.

Together, these measures ensure that the artefact complies with GDPR requirements, supports transparency and explainability, and embeds ethical governance mechanisms throughout the AI-supported impact measurement process.

Glossary and Abbreviations

- KPI – Key Performance Indicator
- IMM – Impact Measurement and Management
- SDG – Sustainable Development Goal
- LLM – Large Language Model
- XAI – Explainable Artificial Intelligence
- IRIS+ – Impact Reporting and Investment Standards
- LangGraph – Framework for orchestrating multi-step LLM workflows
- PyPDF – Python library for PDF parsing and text extraction
- Profile Pydantic Model – Structured schema used to store project onboarding data
- EPCIS – Electronic Product Code Information Services (traceability standard)
- GS1 Digital Link – Standard for resolving product identifiers to digital resources

Note: This appendix is intended to provide transparency and reproducibility for the artefact's processing and outputs, while keeping all data anonymized and compliant with privacy standards.