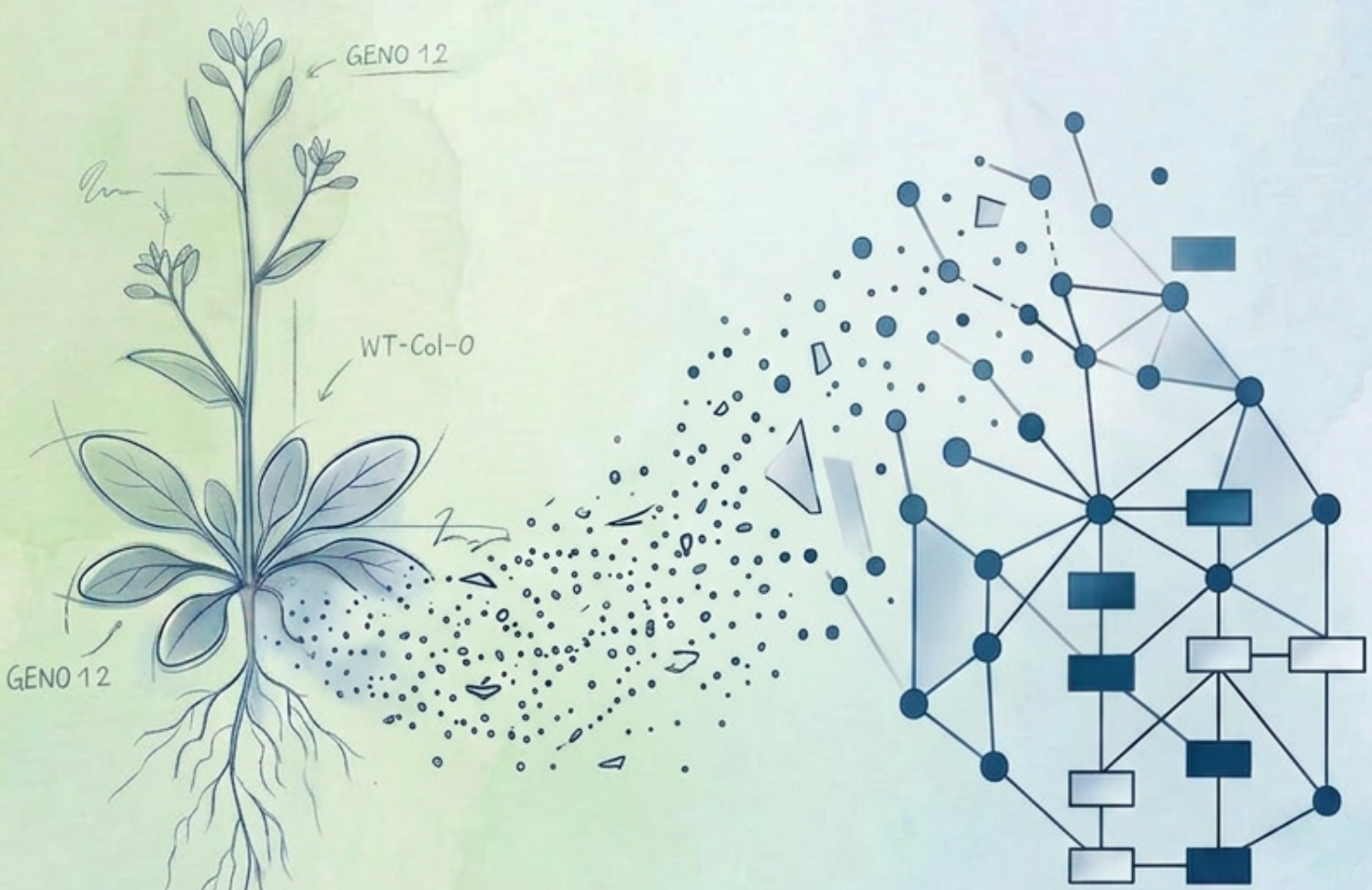
Metadata Extraction from Scientific Lab Notebooks

Design and Evaluation of a Multi-Agent System

Master thesis

Suzanne Backer

Delft University of Technology



TU Delft

Metadata Extraction from Scientific Lab Notebooks

Design and Evaluation of a Multi-Agent System

by

Suzanne Backer

to obtain the degree of Master of Science in Computer Science
at the Delft University of Technology,
to be defended publicly on Tuesday May 19, 2026 at 14:00 PM.

Student Number	5091632	
Project Duration	September 2025 - May 2026	
Thesis Committee	Dr. C. Lofi	Supervisor, Associate Prof. TU Delft
	Dr. M. Reinders	Full Prof. TU Delft
Daily supervisor	Eva Eleonora Ferradosa	PhD Candidate, TU Delft

TU Delft

Abstract

High-quality, interpretable metadata is essential for enabling data reuse and reproducibility in the life sciences. Frameworks such as the Investigation–Study–Assay (ISA) model provide a structured way to describe experimental workflows, yet in practice metadata remains incomplete, inconsistent, and costly to produce. Lab notebooks, which record experimental procedures, offer a promising source to extract metadata from, but their unstructured nature makes extraction challenging.

This thesis investigates extracting ISA metadata from lab notebooks using LLM-based multi-agent systems. It also addresses the lack of suitable methods for evaluating the performance of such tasks. To this end, this thesis contributes two artefacts: (1) a prototype multi-agent system that extracts ISA metadata from lab notebooks, and (2) a rubric-based evaluation framework which formalises and assesses extraction performance.

A feasibility study is conducted on a set of real-world lab notebooks to evaluate the proposed multi-agent system. The results show that partial extraction is achievable with the current prototype, with 41% of ISA entities represented well, and another 27% sufficiently. Therefore, the prototype can provide a practical starting point for researchers in their ISA metadata creation task. While fully automated extraction remains challenging with the current prototype due to missing and insufficiently represented ISA entities, this work provides recommendations for future system design based on the performed feasibility study. Specifically, to operationalise multi-agent ISA extraction in practical workflows, this thesis recommends building a human-in-the-loop system where researchers and agents collaborate to construct metadata.

Preface

The completion of this master's thesis is the result of months of work towards improving metadata practices for scientific research. I hope my work inspires others to continue working on this challenge. The completion of this thesis also marks the end of my student experience in Delft, a phase which has in many ways been transformative for me. I would like to express my gratitude to all those who have helped me along the way of my master's thesis and stood beside me during my studies.

First, thank you to Eva, for helping me define such a relevant challenge to work on, for helping me navigate my struggles every week and for providing me with useful feedback on my thesis drafts. Besides our discussions on the thesis, I have truly enjoyed our coffees and chats. Something I will not soon forget is the time we left Delft before dawn to harvest potatoes and returning with such heavy sacks of potatoes that we could barely carry them.

Also, thank you to Christoph. Our discussions have helped me greatly in reflecting realistically on my work, providing me with the courage and trust to continue. Thank you for helping me steer my work towards a meaningful direction. Thank you to Marcel as well for taking the time to review my work and being a member of my thesis committee. I would also like to express my appreciation for the CropXR consortium and their researchers. Your work provided the inspiration for this thesis. I will follow the rest of CropXR's work in the coming years with great interest. Thank you as well to all the researchers who provided me their lab notebooks to work with.

Finally, thank you to my family and friends who have been there alongside me all the way. Thank you for the moments you fill me with pure joy or provide a listening ear. Your support means more to me than I can express. Zoë, I am grateful that we got to experience so much of our thesis journey side by side.

Suzanne Backer
Delft, May 2026

Contents

1

Introduction

Scientific progress relies on researchers' ability to build on previous work. This ability requires that experimental data, methods, and findings are not only accessible but also interpretable and reusable. To support this, the scientific community has increasingly emphasised data management practices. This emphasis is reflected in the priorities of consortiums such as CropXR [1], aimed at developing crop varieties that are better adjusted to climate change. The consortium actively contributes to research into improving sharing and reuse of experimental data in order to accelerate research.

A central example of emphasis on data management practices is provided by the FAIR (Findable, Accessible, Interoperable, Reusable) principles [30], which have become widely adopted to guide the organisation and stewardship of scientific data. In parallel, domain-specific metadata frameworks and standards are used to formalise how experimental information should be recorded. One prominent example in the life sciences is the ISA (Investigation-Study-Assay) framework, which provides a general structure to describe experimental workflows [22].

Despite the availability of such standards, scientific data often remains insufficiently documented for effective interpretability. In practice, metadata is often incomplete, inconsistent or inaccurate [7, 29]. For example, discrepancies in reported clinical phenotype data in transcriptomics studies have been shown to result in a 35% data loss between scientific publications and their corresponding public repositories [20]. Such incorrect or incomplete metadata does not come without risk, as it can compromise data reuse and the reliability of downstream analyses. This issue became particularly evident during the COVID-19 pandemic, where insufficient adherence to metadata standards limited the usability of genomic data [25]. For example, of approximately twelve thousand SARS-CoV-2 samples submitted to public repositories, 68% had no information on the geographic location where the sample was taken. This reduced the potential to analyse the spread of different variants of the disease, potentially slowing broader understanding.

The lack of interpretability of metadata can, in part, be explained by the observation that researchers face multiple barriers in creating metadata. They often experience metadata creation as a labour-intensive and expensive activity which distracts from their primary work [19]. Additionally, researchers might lack specific knowledge and skills required for metadata creation, while finding little incentive for the task [8].

Given these barriers experienced by data producers, there is increasing interest, including within the CropXR consortium, in automating metadata creation through extraction. A natural source for such extraction is lab notebooks. In the life sciences, lab notebooks are routinely used to document experimental procedures, materials, and observations at the time experiments are conducted. As such, they capture much of the contextual information that metadata frameworks aim to formalise. Leveraging lab notebooks as an extraction source is therefore a logical choice: their use is already embedded in scientific practice. Consequently, extracting metadata from them allows researchers to create metadata

[1] https://cropxr.org/

without requiring changes to their existing workflows.

At the same time, recent work has shown that large language models (LLMs) can reliably convert unstructured text into structured representations [2]. Additionally, multi-agent systems have been shown to outperform single LLM calls on complex tasks requiring decomposition [12]. This suggests an opportunity to leverage such systems for the extraction of structured metadata from lab notebooks.

This thesis, conducted in collaboration with CropXR, explores this opportunity. Specifically, it explores extracting metadata from lab notebooks that is compliant with the ISA-framework. The choice of ISA as the target framework is motivated by its widespread recognition and use in the life sciences, as shown by its adoption by major life science data initiatives such as the FAIRDOMHub [31] and ELIXIR [5] communities.

Despite the presented opportunity, extracting metadata in accordance with the ISA framework comes with significant challenges. ISA requires the construction of a schema-constrained representation of an experiment, defining ISA entities and the relationship between them. Furthermore, the flexibility of the ISA framework introduces a degree of subjectivity, as different researchers may construct different, yet still acceptable, representations of the same experiment. This complicates the definition of a single correct output. Moreover, lab notebooks are typically written in unstructured natural language and often include informal notes, abbreviations and might miss crucial context [24, 14], making them a challenging extraction source. These characteristics distinguish ISA extraction from lab notebooks from conventional information extraction tasks.

Existing work on metadata extraction has explored the use of natural language processing techniques to extract structured information from scientific text. However, existing methods either rely on annotated input data or do not enforce compliance with a complex schema such as ISA [16, 15]. These limitations restrict practical applicability in real-world settings, as such methods require manual annotation or fail to capture the full complexity of metadata required by real-world frameworks. As a result, fully automated extraction of metadata in a format recognised by the life science community (in our case, ISA) from unstructured lab notebooks remains a largely unexplored challenge. In addition, there is a lack of established methods to evaluate the performance of a system performing this task, particularly given that multiple valid metadata representations may exist for the same experiment.

This thesis addresses these challenges by exploring the use of multi-agent systems for ISA metadata extraction from unstructured lab notebooks. In addition to a prototype multi-agent extraction system, this thesis proposes a structured evaluation framework to assess the extraction performance of such a system, accounting for the inherent flexibility of the ISA framework. The main research question addressed in this thesis is:

How can a multi-agent system be designed and evaluated for ISA metadata extraction from unstructured lab notebooks, and what can a prototype reveal about the feasibility of this approach in practice?

To answer this question, the following sub-research questions are defined:

1. **RQ1**: How do challenges encountered in prototyping ISA extraction systems using multi-agent approaches reveal design considerations, and how can these inform the design of a prototype system?
2. **RQ2**: How can the extraction performance of an ISA extraction system be formalised, taking into account the flexible nature of the ISA framework?
3. **RQ3:** To what extent does a prototype implementing these design considerations (RQ1) demonstrate the feasibility of ISA extraction in practice?

To address these questions, this thesis adopts an artefact prototyping approach, inspired by design science research [6], in which two artefacts are developed in parallel: a multi-agent ISA extraction system (addressing RQ1 & RQ3) and a rubric-based evaluation framework (addressing RQ2 & RQ3). Both artefacts are available in the GitHub repository of this project [2].

[2] https://github.com/Suzanne108/Metadata-Extraction-from-Lab-Notebooks

To address RQ3, a feasibility study is conducted using the prototype multi-agent system. This study aims to establish a lower bound on the extraction performance that can be achieved using the prototype, to facilitate reasoning about its usability in a real-world setting. The study also aims to analyse the prototype's failure modes, to identify opportunities for improvement to inform future system iterations.

Together, the contributions of this thesis lay the groundwork for the development and evaluation of a large-scale ISA extraction system to support researchers in the construction of structured metadata. The proposed rubric-based evaluation framework offers a tool for assessing extraction performance in future system iterations. In addition, the implemented prototype establishes a lower bound for achievable extraction performance, while the insights derived from the feasibility study highlight failure modes and provide direction for future system development.

The remainder of this thesis is structured as follows. Chapter 2 provides background on the ISA framework, lab notebooks as an extraction source and multi-agent system design. Chapter 3 describes the artefact prototyping methodology and the feasibility study setup. Chapter 4 details the design considerations found during system prototyping, the workings of the developed rubric-based evaluation framework, and the results of the feasibility study. Finally, Chapter 5 synthesises these findings and Chapter 6 suggests directions for future work.

2

Background

This chapter provides contextual background necessary to understand the problem of ISA metadata extraction from lab notebooks. Section 2.1 introduces the ISA framework, outlining its structure for representing experimental metadata. Section 2.2 describes the challenges accompanied with extracting ISA metadata from lab notebooks, specifically going into the difficulties which can be encountered in lab notebooks as extraction source. Finally, Section 2.3 outlines key concepts in multi-agent system design that inform the approach taken in this thesis for the development of the multi-agent extraction prototype.

2.1. The ISA framework

ISA is a widely used metadata framework that captures descriptions of life science experiments, making these experiments and the data they produce reproducible and reusable [22]. ISA provides a general structure for organising experimental metadata and functions as an overarching framework to coordinate the use of different reporting standards within the bioscience community. The framework is frequently implemented with domain-specific standards, such as MIAPPE [17] for phenotyping experiments, which define the specific metadata field required within the ISA structure. However, some public data repositories, for example FAIRDOMHub [31] and ELIXIR [5], also accept metadata in ISA format directly. While ISA has multiple versions, this work targets extracting ISA version 1.0 [1].

A conceptual overview of the ISA framework can be found in Figure 2.1. It defines the concepts of an *Investigation*, *Study* and *Assay*. An Investigation contains information to understand the project context and goals. It defines entities such as *Person*, describing people involved in the project and *Publication*, describing publications resulting from the project.

An Investigation has one or more Study entities, which describe the materials used and procedures followed during the experiment. Specifically, the Study entity defines a *Study Design* and study *Factors*, which are independent variables manipulated by the experimentalist to affect biological systems in a measurable way. It also defines *Sources*, *Samples* and *Materials*. Materials describe the materials consumed or produced during an experimental workflow. Sources are a special kind of Material which are considered the starting biological material used in a study. Samples are also a special kind of Material that represent major outputs resulting from a *Protocol* application. Essentially, Sample materials represent Source materials which have undergone some kind of manipulation. Any other materials consumed or produced during the experiment are represented as Materials.

The Study entity further defines *Protocols* and *Processes* to describe the experimental workflow. A Protocol captures a coherent description of steps taken during the experimental workflow. These steps are described in a general way such that they may be replicated. Additionally, the Protocol should define parameters, which can later be given a value by a Process which applies the Protocol. A Process,

[1] https://github.com/ISA-tools/isa-api/tree/master/isatools/resources/schemas/isa_model_version_1_0_schemas/core

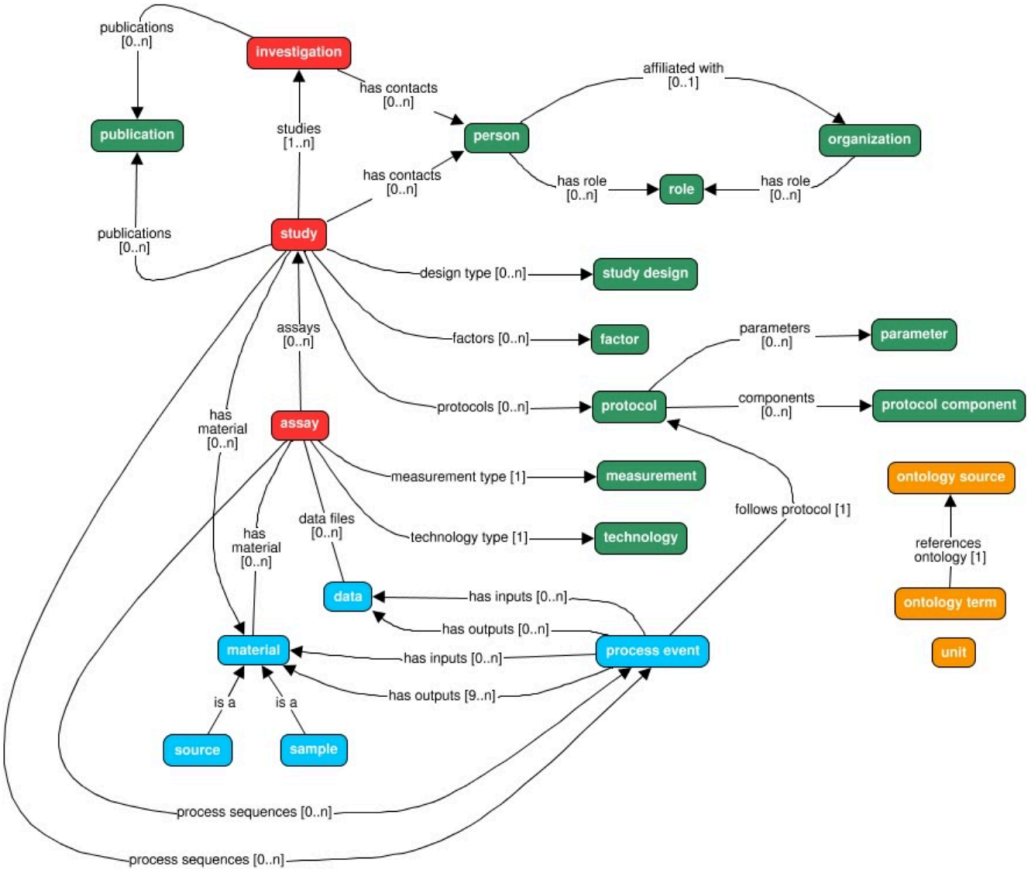


Figure 2.1: A conceptual overview of the ISA framework, detailing the main entity types and the relationships between them [2].

defines a Protocol application to some input Material to produce some output. Multiple Processes can apply the same Protocol. For example, a Protocol might define a step to grow plants in a growth chamber. The settings of the growth chamber (temperature, light cycle, humidity) are defined as parameters of the Protocol. Two different Processes might implement the same Protocol, but with different values for temperature, light cycle and humidity.

An Assay represents a test performed on a Material, producing a measurement. An Assay records the type of measurement that was taken and the technology that was used to do so. It also records which Processes, Materials and Samples were involved. Furthermore, an Assay references one or more *Data* entities, which represent raw data files produced by an Assay. Instances of all ISA entities can be represented as nodes. They reference each other and, in this way, are conceptually linked in a graph to represent the experimental workflow performed [21].

An important characteristic of ISA is that it is flexible by design. The same experiment can be represented in multiple valid ways depending on modelling choices. For example, a researcher may choose to model a sequence of experimental steps as separate Process nodes linked together, while another may represent the same sequence as a single Process associated with a more detailed Protocol. Both representations describe the same underlying experiment, yet differ structurally in how the workflow is captured. This implies that ISA metadata creation is not fully deterministic, but involves a degree of subjective interpretation and preferred style. This aligns with the general understanding that metadata creation is non-deterministic, as researchers choose what to document based on the goal of the research and the needs of future research tasks [11, 13]. Also as a general notion, interpreting metadata requires researchers to use their existing knowledge to understand what others have done. This means that two different researchers might require different metadata to achieve the same understanding of an experiment or dataset. This further emphasizes the inherent ambiguity of the task of metadata creation.

2.2. Challenges of extracting ISA metadata from lab notebooks

In this section, the difficulties inherent to extracting ISA metadata from lab notebooks are discussed. Section 2.2.1 explains why ISA extraction and its evaluation are inherently complex, while Section 2.2.2 highlights the interpretability issues associated with lab notebooks.

2.2.1. ISA as extraction challenge

ISA extraction from unstructured lab notebooks differs from standard information extraction tasks. Information extraction tasks range from named entity recognition, relation or event extraction tasks, to downstream knowledge graph construction tasks [32]. Many of these approaches focus on identifying local pieces of information, which are independently evaluated using benchmark datasets with a single ground truth. These approaches typically do not capture a full experimental workflow structure required by metadata frameworks such as ISA.

Alternatively, metadata authoring platforms support the creation and validation of structured metadata, but assume that the relevant information is already available in a structured form [3]. As a result, existing approaches either focus on local information extraction or indirectly rely on manual metadata creation by requiring structured input formats.

In contrast to local information extraction, extracting metadata in accordance with the ISA framework requires constructing a schema-constrained graphical representation of an experiment. ISA defines a set of typed entities (Investigation, Study, Assay, Protocol, Material) and requires these to be connected in a way that reflects the underlying experimental workflow. As a result, ISA extraction is not limited to identifying relevant information in local text, but also involves ensuring that the resulting representation is globally consistent. This positions ISA extraction as a non-standard type of information extraction problem.

Recent work has explored the use of multi-agent systems for metadata extraction for similar tasks. For example, Mondal et al. [15] propose a system in which specialised agents collaboratively extract metadata fields from biomedical datasets, achieving high-quality metadata curation across a predefined set of attributes. While such approaches demonstrate the potential of multi-agent systems to improve metadata quality and scalability, they primarily focus on extracting independent metadata fields and do not construct schema-constrained, workflow-oriented representations such as those required by the ISA framework.

One of the few systems that explicitly targets ISA metadata is LISTER [16]. This system performs semi-automatic extraction of ISA metadata from laboratory records. However, it requires lab notebooks to be pre-structured and annotated according to ISA-specific conventions, effectively shifting part of the metadata creation effort to the data producer, rather than being fully derived from unstructured text. Consequently, fully automated extraction of ISA-compliant metadata from unstructured lab notebooks remains largely unexplored.

Additionally, there is a lack of established methods for evaluating extraction performance in unstructured settings where multiple valid representations may be possible. Other extraction challenges are typically evaluated against predefined benchmark datasets that constrain the space of valid outputs, using metrics such as precision and recall, which operate on isolated components [1]. This evaluation methodology is not suitable for ISA extraction, as ISA representations may be partially correct, suitable or appropriate. For example, correctly describing a performed protocol, but omitting the required parameters could still be considered valuable metadata. Combined with a degree of subjectivity due to different representations potentially being equally valid, this complicates the evaluation of extraction performance using standard metrics such as precision and recall.

In summary, ISA extraction from unstructured lab notebooks constitutes a non-standard information extraction problem. Existing approaches either extract independent metadata fields instead of a representative experimental workflow or rely on pre-structured inputs, leaving a clear gap for ISA-compliant extraction from unstructured lab notebooks, as well as in evaluation frameworks to assess extraction performance.

[2] https://isa-tools.org/isa-api/content/isamodel.html

2.2.2. Lab notebooks

Lab notebooks serve as records of followed experimental procedures in the bioscience domain. One of the main goals of keeping a lab notebook is to ensure interpretability and reproducibility of an experiment. Moreover, they can serve as a personal organisational tool and aid for critical reflections on scientific work [24]. They may contain protocols used, raw results, experimental parameters, or settings for lab equipment [14]. Therefore, lab notebooks form a relevant potential candidate to extract experimental metadata from.

However, while notebook-keeping guidelines exist, in practice, the interpretability of lab notebooks varies significantly. They might contain language specifically used in a particular research group [14], use personal and inconsistent style or lack important context. It is important to note, however, that lab notebooks traditionally are not shared between people and are considered personal artefacts. The personal notes are usually later transformed into metadata or publications intended for sharing. Due to the personal nature of lab notebooks, extracting interpretable metadata from them can be considered challenging.

To illustrate the challenge of extracting ISA metadata from lab notebooks, Appendix A provides an example of what a lab notebook might look like. The notebook documents an experiment that studies the development of roots and phenotyping characteristics of a certain plant species under different drought conditions, either through a drought plate or through different watering regimes. Although this example is not an actual notebook used in this thesis, it is inspired by those that were analysed and reflects the types of challenges encountered in them.

The extraction issues present in the example notebook are categorized and summarized in Table 2.1. As a first challenge, lab notebooks often introduce ambiguity, either regarding which steps were performed or how relevant certain statements are for ISA documentation. For instance, when a step is described as being performed “preferably while it is still 90 °C”, it is unclear how essential this condition is and how important it is to formally capture it in an ISA protocol. Similarly, the note “MaxSWC 155% (apparently 140%)” is ambiguous, as it remains uncertain whether the maximum soil water content is 155% or 140%.

Second, essential information may be omitted from lab notebooks. For example, a notebook may mention that an autoclaving process is applied (a sterilization step using steam) without specifying key parameters such as temperature or duration. In other cases, notebooks refer to external documents where relevant details are supposedly recorded, making it difficult to obtain a complete description of the experiment.

Issue Type	**Example (Appendix A)**
Ambiguousness	“nutrients were added regularly” “preferably while it is still >90°C” “MaxSWC = 155% [apparently 140%]!”
Omission of relevant details	“Autoclave” “[SENSOR DATA]” “See lichtmetingen.docx” “Drying them for ?? nights at 80°C”
Protocol deviations	“First few days 80% were watered to 70%” “From day 10 onwards, pots are not weighted and watered separately but per tray to minimize workload” “middle rack –> later more close to door”
Undefined terminology	“Mickey” “Dryzotron”
Abbreviations	“DAG”
Usage of multiple languages	“(Trays placed in middle rack – later meer richting deur)”
Noise	“Hi <name>, so you start with....” “13-M2-links 45”

Table 2.1: Categorisation of issues found in the example lab notebook presented in Appendix A, including examples.

Third, notebooks may contain protocol deviations, even though they are not explicitly labelled as such. For example, a note such as “From day 10 onwards, pots are not weighted and watered separately but per tray to minimize workload” indicates a change in procedure. It is unclear whether such a change substantially affects the applied process and should therefore be formally documented, or whether it is merely a practical adjustment that can be omitted. Protocol deviations complicate metadata extraction because they challenge the assumption of a single, fixed protocol underlying the experiment, as assumed by frameworks such as the ISA framework. In practice, procedures often evolve during execution. Lab notebooks describe these changes informally. This creates a representation dilemma: a deviation could be modelled as a new protocol, a parameter change within an existing process, or simply as a comment. Because the text rarely clarifies the importance or scope of such deviations, an automated system must interpret the situation rather than merely extract information, making the extraction task difficult.

Finally, lab notebooks may use undefined or inconsistent terminology, further complicating interpretation. Terms such as “dryzotron” (likely a combination of “dry” and “rhizotron”) may not be formally defined. In addition, notebooks may mix multiple languages, use abbreviations, or include seemingly unrelated information such as informal messages from colleagues. These factors introduce noise and increase the difficulty of reliably extracting structured metadata.

2.3. Multi-agent system design

LLM-based multi-agent systems (MAS) are systems composed of multiple interacting LLM components (agents), often using decentralized decision-making and information sharing to accomplish tasks [12]. Multi-agent systems are particularly suitable for complex tasks that require decomposition into specialised subtasks. As described in Section 2.2, ISA metadata extraction from unstructured lab notebooks can be considered a complex task because of the multiple ways ISA can be instantiated and interpretability issues accompanying lab notebooks. As such, a multi-agent approach is deemed suitable, as it allows for individual but collaborative reasoning, decision-making and task execution components.

To guide the design of a multi-agent system for ISA extraction, several design principles have been identified in prior work. First, proper task decomposition and topology is essential for addressing complex problems, allowing agents to operate on well-defined subtasks rather than solving problems monolithically. Hence, some works explore if proper task decomposition and topology construction can be achieved dynamically [9, 33]. Second, agent role specialisation for the different subtasks is widely adopted, allowing individual agents to work on a specific task [4]. Third, equipping agents with external tools [23], such as retrieval-augmented generation (RAG) mechanisms [27], can enhance their capabilities. Fourth, reflection and feedback mechanisms are often built-in components of MAS architectures to improve performance [18, 26]. These design principles were considered and experimented with during the prototyping methodology of the multi-agent extraction system developed in this work. The principles of task decomposition, agent profiling and usage of external tools were also included in the final version of the system, as described in Section 4.1.2.

3

Methodology

This chapter describes the research methodology used in this study. Section 3.1 introduces the artefact prototyping approach used to develop an ISA extraction evaluation framework and a multi-agent extraction system. Section 3.2 presents the feasibility study performed to investigate the viability of multi-agent ISA extraction from unstructured lab notebooks using a prototype system.

3.1. Artefact development

As discussed in Section 2.2.1, ISA extraction from lab notebooks is a task which is characterised by the absence of a single ground truth of what a good metadata representation is. As such, the task itself is ill-defined and requires clear understanding. Additionally, because extraction outputs may be partially correct and cannot be meaningfully evaluated using standard metrics, an appropriate evaluation method is required to assess task performance. Artefact prototyping was used as an instrument to explore the task requirements, evaluation methodology and to take the first steps toward an ISA extraction system. This methodology is inspired by principles from design science research [6], which emphasises the creation and study of artefacts such as methods, models, systems, and frameworks, to solve real-world problems and generate knowledge about how and why solutions succeed or fail. Prototyping also serves as an effective means to surface assumptions and latent requirements, as argued by Sommerville and Sawyer [28].

The prototyping methodology was employed with two goals in parallel. The first goal was to gain insights into the development of a multi-agent ISA extraction system, addressing RQ1. This prototyping phase was guided by design principles for LLM-based multi-agent systems (Section 2.3). The insights gained were implemented in a final prototype used during the feasibility study (Section 3.2), addressing RQ3. The final prototype was implemented using LangChain [1] and LangGraph [2], and is presented in Section 4.1.2.

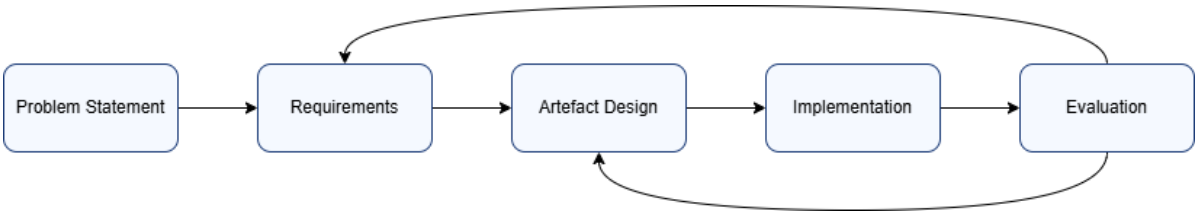


Figure 3.1: A conceptual overview of the iterative prototyping methodology used in this study.

The second goal was to develop an evaluation method that can assess the performance of a system extracting ISA metadata from unstructured lab notebooks, addressing RQ2. The systematic evaluation of ISA extraction performance is a central challenge to ISA extraction. Binary metrics are inappropriate

[1] https://www.langchain.com/
[2] https://www.langchain.com/langgraph

for this task because ISA modelling is inherently flexible (Section 2.1), making it impossible to define a single authoritative ground truth. In addition, systems may produce outputs that are partially useful. Such partially correct metadata may still be valuable to users, yet binary metrics such as precision and recall cannot meaningfully capture these graduations of correctness.

A rubric-based evaluation framework was selected as a suitable method to address these challenges. Scoring rubrics are widely used in education, where open-ended tasks require the assessment of complex competencies [10]. They allow evaluators to distinguish between fully correct, partially correct, and insufficient representations, and to articulate why an output meets or fails to meet expectations. This transparency is particularly important for ISA, where semantic adequacy, whether the metadata is interpretable, meaningful, and accurate according to the described experiment, requires complex judgement.

Prototyping the evaluation framework and the multi-agent extraction system followed the procedure as shown in Figure 3.1. From a problem statement initial design objectives for both the evaluation framework and the system were established. From this choice, cycles of artefact development for both the framework and the system followed. After an iteration of artefact development, the system was evaluated by the evaluation framework. After behaviour inspection in the evaluation phase, either the objective setting or the artefact development phase was circled back towards.

The development of the evaluation framework was guided by six design principles. These criteria reflect practical considerations that emerged during early iterations and were used to evaluate whether each version of the framework supported meaningful, consistent, and fair assessment of ISA extraction performance. The design principles are the following:

1. **Fairness** - the evaluation framework should be fair to the system it evaluates and refrain from imposing impossible expectations. This is especially important since extraction of ISA metadata not only depends on the extraction capabilities of an extraction system, but also on the complexity of the described experiment and the clarity with which a reference document describes the document. It would be unfair to expect an extraction system to extract perfect ISA metadata from a lab notebook that is judged uninterpretable by an evaluator.
2. **Granularity of assessment** - The evaluation framework should both support a clear overview of the overall extraction performance of the evaluated system, while also capturing nuanced differences in metadata quality across parts of the extracted metadata.
3. **Comparability across systems** - The framework should enable meaningful comparison between different extraction systems for the same lab notebooks.
4. **Simplicity of use** - The framework should be straightforward to apply for evaluators who are familiar with the ISA framework.
5. **Inter-rater consistency** - Different evaluators should arrive at similar scores when assessing the same metadata
6. **Transparency** - The rationale behind scoring decisions should be clear to both evaluators and interpreters of the evaluation process.

3.2. Prototype-based ISA extraction feasibility study

This section describes the feasibility study conducted to assess the viability of operationalising ISA metadata extraction using a prototype multi-agent system. Figure 3.2 details the workflow of the performed study. ISA metadata was extracted from seven lab notebooks, as can be seen in the first phase of the figure. More information about the lab notebooks is provided in Section 3.2.1. In the first phase, the multi-agent system prototype as presented in Section 4.1.2 extracts ISA metadata from the lab notebook. Due to resource constraints of the project, two different large language models were used by the prototype system. For some lab notebooks this was gpt-5-nano, and for others it was gpt-5-mini. Specifically, gpt-5-mini was used for lab notebooks N1 - N4 and gpt-5-nano was used for lab notebooks N5 - N7, as presented in Table 3.1. The respective LLM was used for all components of the multi-agent prototype which required an LLM. The use of relatively small models reflects practical deployment constraints that are likely relevant for ISA extraction workflows. Scientific practitioners

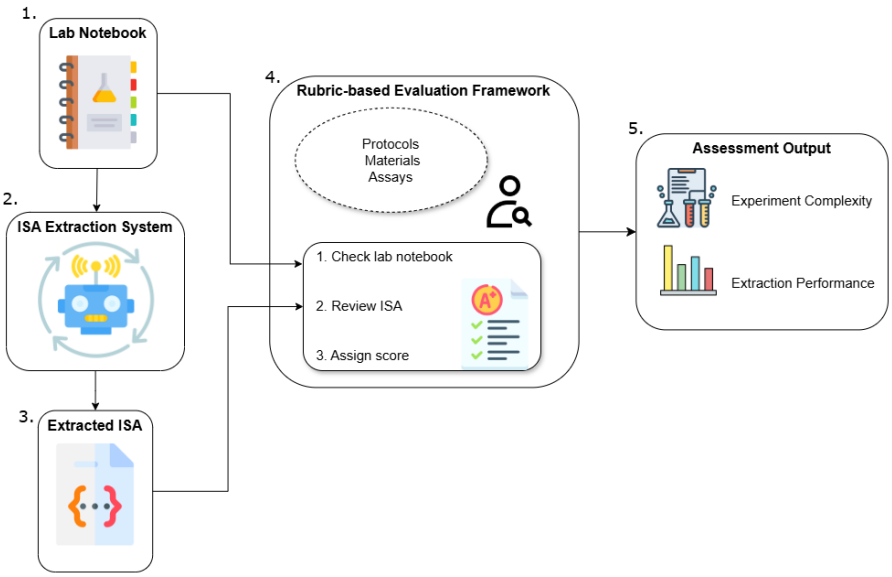


Figure 3.2: A conceptual overview of the workflow of the ISA extraction feasibility study.

may prefer privacy-preserving or locally hosted LLM deployments, where computational resources and operational costs can limit the feasibility of larger models.

During the second phase, a single annotator used the rubric-based evaluation framework, as presented in Section 4.2, to evaluate *Experiment Complexity* and *Extraction Performance*, which are identified as two factors influencing the quality of extracted ISA metadata. Experiment complexity represents the inherent complexity of the experiment described by a lab notebook, while extraction performance represents the ability of the system to translate natural language descriptions in a lab notebook into ISA entities that correctly and completely represent the described experiment. Section 4.2 explains how these concepts are formalised by the developed evaluation framework and Section 3.2.2 provides additional information on the evaluation procedure performed.

3.2.1. Lab notebook dataset

The seven lab notebooks used in this experiment were collected from plant science researchers within the CropXR network, who created them during their experiments. Since the notebooks are collected from real-world experimental work, they form a realistic representation of lab notebooks in practice. They, for example, describe plant science experiments investigating plants' growth performance under different conditions, such as the amount of water available. A more detailed description of each notebook can be found in Table 3.1. Note that N1, N2 and N3 described semantically similar experiments and N1 and N2 had some overlap in the content of the documents. While both N6 and N7 describe experiments studying Arabidopsis thaliana under different osmotic stress conditions, the studies had different experimental setups and studied different stress conditions.

The general observation that lab notebooks mainly function as personal documents and can therefore be difficult to interpret for external readers (as discussed in Section 2.2.2), is also reflected in the lab notebooks used for the feasibility study. Both the example lab notebook given in Appendix A and the lab notebook issue categories defined in Table 2.1 were based on the interpretability issues found in the lab notebooks used for the feasibility study. Several illustrative examples of the interpretability difficulties from the notebooks include that notebook N1 refers to lab infrastructure using lab specific terminology (e.g., "Dombo", assumed to refer to a specific growth chamber) without further explanation. N1 also uses abbreviations such as "DAG" (days after germination) without definition. In notebook N2, references are made to materials from prior experiments (e.g., "use soil from experiment 0b2") without providing sufficient context, indicating reliance on external or implicit knowledge. In notebook N5, new terminology is introduced (e.g., "dryzotron") without explanation, reflecting that researchers may use undefined terms. Finally, in notebook N6, the description of experimental conditions is sufficiently un-

clear that determining the amount of experimental conditions used in the experiment is difficult. These examples illustrate that, although the lab notebooks used for this feasibility study contain relevant experimental information, their clarity, completeness, and accessibility can vary substantially.

Notebook	Description	Pages	Content
N1	Studies flowering of Arabidopsis thaliana under 4 different drought conditions. Records phenotype measures.	4	text
N2	Studies flowering of Arabidopsis thaliana under 4 different drought conditions. Records phenotype measures.	4	text
N3	Studies RNA expression levels and flowering of Arabidopsis thaliana under drought conditions.	7	text
N4	Experiment identifies specific genotype given small samples of plant material. Procedures of DNA isolation, PCR and electrophoresis.	15	text, images, tables
N5	Studies the development of Arabidopsis thaliana roots in drought conditions.	1	text
N6	Studies the development of Arabidopsis thaliana under different osmotic stress conditions.	5	text and tables
N7	Studies the development of Arabidopsis thaliana under different osmotic stress conditions.	4	text and tables

Table 3.1: An overview of the lab notebooks used in this study. A short description of the notebook is provided under 'Description'. The column 'Pages' approximates the amount of pages of the lab notebook solely filled with text. The column 'Content' records the type of content which can be found in the notebook.

3.2.2. Evaluation procedure

The evaluation procedure was performed through use of the developed rubric-based evaluation framework, as presented in Section 4.2. The procedure was carried out by the author of this thesis, who is familiar with the ISA framework and has basic knowledge of plant science experimental practices. Two measures were taken to increase the validity of the evaluation procedure. First, the evaluator was familiarised with ISA metadata by studying examples of ISA metadata [3]. Second, a pilot evaluation phase was employed to both uncover ambiguities in the evaluation procedure and to familiarise the evaluator with the task. Based on this pilot, a set of practical guidelines was constructed to support consistent application of the evaluation framework.

First, evaluation was guided by a set of implicit criteria that emerged during the pilot phase. These included whether the extracted metadata captured the core experimental intent, whether key experimental components identified in the lab notebook were present and correctly interpreted in the metadata, whether the representation was sufficiently complete to allow reconstruction of the experiment, whether modelling choices resulted in a coherent and interpretable workflow representation and whether the metadata recorded was relevant and concise. These criteria correspond broadly to notions of accuracy, completeness, modelling logic, relevance and conciseness but were not assessed independently. Instead, they informed a unified judgment of the usefulness and adequacy of each ISA evaluation unit.

To ensure fairness, the evaluator was allowed to adjust the maximum obtainable experiment complexity points for ISA entities that were substantially simpler or more complex than typical instances of the same type. This prevents systems from being penalised or rewarded purely due to modelling choices that fall within the flexibility of the ISA framework. This mechanism of adjusting experiment complexity points is inherent to the used evaluation framework and is further explained in Section 4.2. Additionally, complete descriptions were prioritised over concise descriptions. ISA components that lacked important information which was explicitly provided in the lab notebook were more severely punished than answers that added unnecessary details, as long as these unnecessary details were correct according to the lab notebook.

Finally, because ISA modelling is inherently flexible, the evaluator accepted multiple valid representa-

[3] https://github.com/ISA-tools/ISAdatasets

tions of the same experiment. When several modelling choices were plausible, the evaluation focused on whether the extracted metadata captured the relevant experimental information and adhered to appropriate ISA structure, rather than on enforcing a single preferred representation. In cases where the notebook description was too ambiguous for the evaluator to determine what information should be represented, the corresponding experiment complexity points were excluded from the total.

4

Results

This chapter presents the results of the thesis. Section 4.1 discusses observations from prototyping a multi-agent ISA extraction system and presents the prototype system used for the feasibility study. Section 4.2 describes the rubric-based evaluation framework, detailing how it approximates experiment complexity and extraction performance. Finally, Section 4.3 reports the results of the feasibility study.

4.1. Multi-agent system design: observations and prototype

This section presents the results from the prototyping phase for the development of a multi-agent system for ISA metadata extraction, as described in Section 3.1. First, Section 4.1.1 summarises observations that emerged during early prototype iterations and translates these into design implications for multi-agent ISA extraction systems. Second 4.1.2 presents the final prototype developed for the feasibility study, which operationalises these design implications.

4.1.1. Multi-agent system observations and design implications

The findings presented in this section originate from the exploratory prototyping of a multi-agent ISA extraction system. The observations should not be interpreted as universal properties of ISA extraction systems. Instead, they reflect behaviours observed within the specific system architecture and prototyping setup used in this study. The associated design implications represent reasoned interpretations of these observations and may inform future research on ISA extraction systems.

Output structure

Monolithic generation was observed to be structurally unstable. Earlier prototype versions experimented with interleaved monolithic generation of ISA and correction of structural mistakes, such as omitted object parameters. Fixing structural mistakes in one place induced mistakes in other places, leading to a structural collapse of the metadata. The generated ISA no longer adhered to structural requirements according to the ISA standard. This suggests that ISA extraction systems should enforce ISA-compliant structured output during generation, rather than relying on post-hoc structural corrections.

Agent workflow control

Unconstrained agent autonomy produced drift in task execution. Earlier prototypes contained agents with relative freedom in the execution of the task at hand. For example, an agent had the freedom to drive the task with no clear guidelines on how to break the task into smaller subtasks. It was given the freedom to determine what step should be executed next to create ISA, like reading from the lab notebook, invoking an evaluation module or creating part of the ISA structure. This resulted in drift in the task execution as the agent made unreasonable choices given the current task progress state. This suggests that an ISA extraction system requires proper task decomposition and agent profiling, limiting the freedom of agents and providing structured prompts on how to execute their task.

Extraction strategy

ISA extraction is not a flat entity extraction task. Earlier prototypes attempted to directly convert natural

language descriptions in lab notebooks to corresponding ISA entities. This led to shortcomings in the accuracy with which the extracted ISA reflected the experiment in the lab notebook. ISA entities were not logically related to each other in a coherent experiment workflow. This indicates that ISA extraction might benefit from the use of an intermediate conceptual representation that captures experimental workflows prior to constructing ISA-compliant entities. This step would globally model the experimental graph and conceptually define which natural language concepts map to which ISA entities and how they should relate to each other.

ISA conceptual stability
Conceptual modelling of ISA is not an objective task. ISA is positioned as a relatively flexible framework and thus multiple valid ISA configurations could be created for the same experiment. This is reflected in the behaviour of prototype versions implementing a conceptual modelling step, as the model created for the same notebook is not stable over different runs of the prototype. Whether this behaviour is undesirable depends on the goals of the system user. While some users might be satisfied by any correct ISA configuration, others might prefer a particular one. If the latter situation applies, the observation suggests that steps should be taken to stabilise the conceptual modelling phase.

4.1.2. Multi-agent system design

The implemented prototype extracts ISA metadata from lab notebooks by operationalising the design insights identified during the prototyping phase as presented in Section 4.1.1. At a high level, the system first interprets the lab notebook at a conceptual level, then decomposes the extraction task into smaller subtasks, and finally constructs ISA entities and their relationships.

The prototype operationalises ISA entity extraction, but does not extract production-ready ISA-compliant metadata, because two implementation aspects were intentionally left out of scope. First, extracted metadata is not grounded in existing ontologies. Second, relationships between ISA entities are not properly instantiated, but instead are reasoned about by the system. However, these limitations do not prevent the use of the prototype for the feasibility study conducted in this thesis, where the goal is to assess the viability of agentic ISA extraction rather than to produce production-ready metadata.

System workflow

The prototype implements a workflow in which a set of role-specialized LLM components collaboratively extract ISA metadata. Figure 4.1 provides a sequence diagram of the workflow. First, the *concept modeller* receives the lab notebook text as input. In addition, it is provided with a natural language description of the workings of the ISA framework. This description is generated by an agent outside the multi-agent system and is derived from the ISA abstract model as specified in 'ISA Model and Serialization Specifications' [21]. Using these inputs the concept modeller produces a natural language mapping between concepts described in the lab notebook and corresponding ISA entities. This intermediate conceptual representation, hereafter referred to as the 'ISA contract', serves as a bridge between unstructured experimental descriptions and the structured ISA schema.

Next, the ISA contract is passed to the *task definer*, which decomposes it into a set of atomic tasks. Each task corresponds to the creation of a specific ISA entity. Task progression is managed by the *Task Manager*, implemented as a deterministic script responsible for maintaining execution order and state. All tasks are subsequently processed by the *task execution planner*, which governs control flow during task execution.

The task execution planner receives both the current task and a limited history of recent agent messages. Based on this context, it selects between two actions. First, it may determine that additional information is required to complete the task. It then generates a query for the *document searcher* module. This module performs a semantic search over the lab notebook and returns the most relevant context to the query. Alternatively, the planner may determine that sufficient information is available and proceed with task execution. In this case, the task execution planner generates an instruction for the *ISA construction coordinator*, which is responsible for delegating the task to the appropriate *ISA specialist*. These specialists are role-specific LLM components, each constrained to generate a particular type of ISA entity (Investigation, Study, Assay, Protocol, Process, Factor, Material, Source, Sample and Data). Given the instruction from the ISA construction coordinator and the latest retrieved context by the workflow, the specialist generates a structured representation of the corresponding ISA entity.

Upon completion of all tasks, the system produces a collection of ISA entities generated by the ISA specialists. These entities represent the components of the experimental workflow and can be interpreted as nodes in an ISA experimental graph. The *assembly advisor* then reasons about the relationships between the generated ISA components. The advisor receives both the constructed ISA entities and the original lab notebook text as input, and produces natural language descriptions of how the objects should be connected. These descriptions specify which samples are derived from which sources, which processes implement which particular protocol and how assays relate to samples, materials and processes. Together, the generated entities and the linking descriptions provide a complete representation of how the ISA experimental graph would be structured.

In a fully implemented system, these inferred relationships could be translated into explicit graph connections by a component that iterates over the entities and establishes links based on the decisions produced by the assembly advisor. This would result in fully automated ISA graph construction. However, for the purposes of the feasibility study, representing these relationships in descriptive form is sufficient to capture the structure of the experimental workflow.

Finally, the *ISA exporter* maps all generated ISA entities to their counterparts in the ISA tools package [1]. Using these representations, the module deterministically constructs the ISA graph and exports the resulting metadata in a standardised format.

Design rationale

A first design aspect is constrained output generation. As highlighted in Section 4.1.1, enforcing output structure improves structural reliability when generating ISA metadata. This aligns with MAS design principles that emphasize constraining agent behaviour to reduce uncertainty. This is reflected in the

[1] https://isatools.readthedocs.io

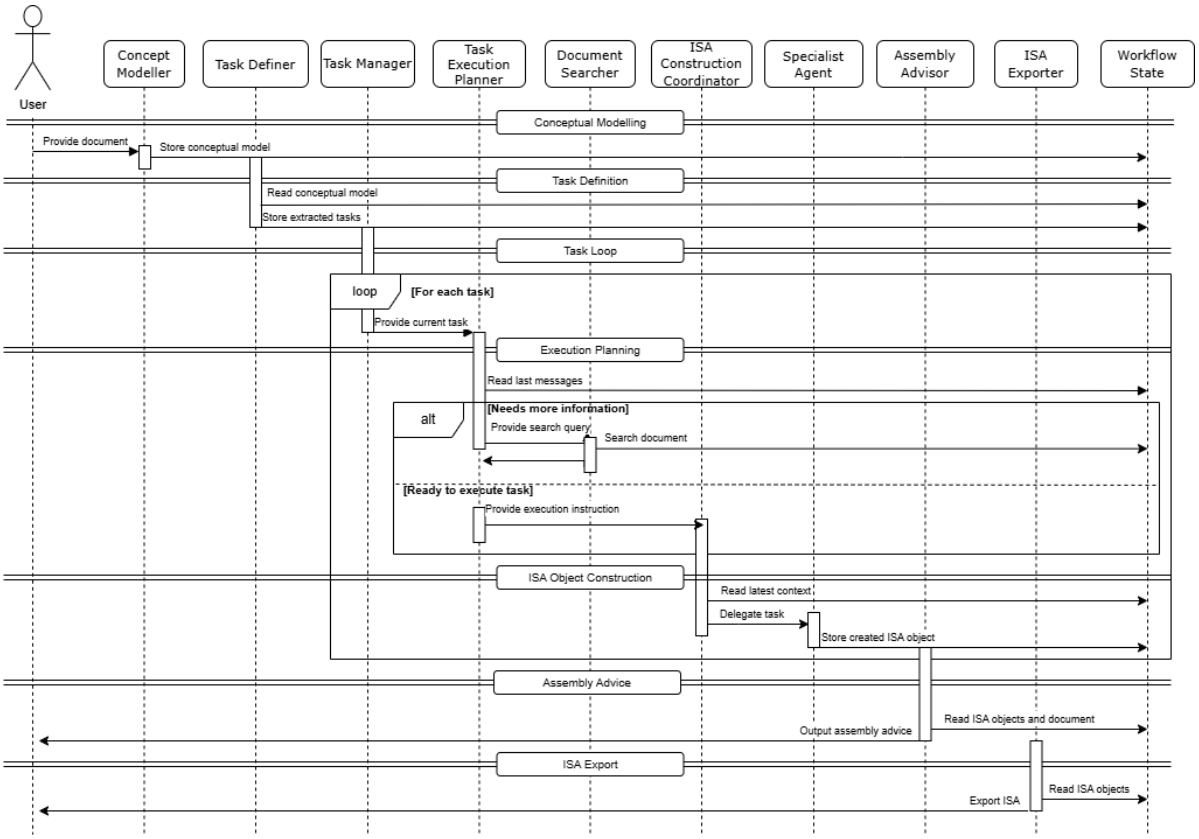


Figure 4.1: A sequence diagram showing the workflow of the prototype multi-agent ISA extraction system. It includes components such as the Concept Modeller, Task Definer, Task Manager, Execution Planner, Document Searcher, ISA Specialists, Assembly Advisor, and ISA Exporter, illustrating how they interact to transform unstructured text into structured ISA metadata.

prototype through the ISA Specialist components, which generate output conforming predefined object structures corresponding to specific ISA entities (Protocol, Material, etc.)

A second design aspect concerns workflow control and agent freedom. Insights from Section 4.1.1 indicate that unconstrained agent behaviour leads to inconsistent results. Accordingly, the prototype employs a clear decomposition of the general task of ISA extraction, in which each component fulfils a specific role. Agent autonomy is deliberately limited through predefined roles and constrained decisions. This is consistent with the usage of agent profiles and a cooperative communication style, as described in Section 2.3.

A third aspect is the execution strategy of the ISA extraction task. Rather than directly extracting structured metadata from the lab notebook, the system first constructs a conceptual representation (the ISA contract) through the concept modeller. This fulfils the identified need for an intermediate representation, as mentioned in Section 4.1.1. This approach is consistent with general decomposition strategies in MAS design, where complex tasks are broken down into intermediate representations and subtasks.

The insight regarding ISA conceptual stability, as identified in Section 4.1.1, is not explicitly addressed in the current implementation and is left as an area for future work.

4.2. Rubric-based evaluation framework

This section presents the rubric-based evaluation framework developed to assess the extraction performance of ISA extraction systems. The framework is a tool that guides experts in providing structured evaluations of the semantic quality of extracted ISA metadata by comparing it against the information present in a reference document. These evaluators should be experts in that they should be familiar with the ISA framework and the evaluation methodology.

The framework measures extraction performance as the proportion of ISA-relevant information in a lab notebook that is correctly and sufficiently represented in the extracted metadata. This is operationalised through the evaluation of evaluation units. These evaluation units either correspond to individual ISA fields, for example the investigation and study title and description, or to composite ISA entities which consist of multiple fields that together form a coherent ISA concept. An example of this is a protocol, which consists of a name, description and a list of parameters.

Each evaluation unit is assigned a complexity score and a performance score. The complexity score represents the amount and structural complexity of information that must be correctly represented in the extracted metadata. Units requiring the representation of a larger number of attributes, relationships, procedural steps, or contextual details are assigned higher complexity scores. The performance score represents a holistic expert judgement of metadata quality. In forming this judgement, evaluators are guided to consider aspects such as accuracy, completeness, modelling logic, relevance and conciseness, but do not assess these dimensions independently. Instead, they are integrated into a single score reflecting the overall adequacy of the representation.

By aggregating the performance scores of all evaluation units relative to their total experiment complexity, the framework produces both an approximation of the inherent complexity of the underlying experiment and the extraction performance of the system. While the interpretability of lab notebooks influences extraction performance, the evaluation framework does not formalise lab notebook quality as a separate evaluation dimension. This is left to future work. However, interpretability can be partly accounted for indirectly through evaluation practices by for example excluding ambiguous information from scoring, as is done in the feasibility study performed in this work (Section 3.2.2).

The remainder of this section provides more detail on how experiment complexity and extraction performance are formalised and how the evaluation framework is used to determine these measures.

4.2.1. Experiment complexity

Experiment complexity is derived solely from the reference document. The evaluation framework guides the evaluator in identifying all ISA-relevant entities (protocols, processes, materials, samples, factors, and assays) based on the experiment description in the document. Each identified ISA entity is an evaluation unit which will be evaluated by the framework. Each unit contributes a fixed number of

complexity points corresponding to the amount of information required to represent it accurately. Table 4.1 provides an overview of the default complexity points awarded to unit. The default complexity points differ per ISA entity, because each entity constitutes different levels of extraction difficulty. For example, material entities typically require relatively limited metadata representation involving the material name and potentially relevant characteristics, whereas protocol descriptions often involve multiple sequential actions and parameters. This is reflected by a higher number of complexity points.

ISA entity	Default complexity points
Experiment Factors	7
Source	3
Sample	5
Material	5
Protocol	10
Process	10
Assay	10

Table 4.1: The default complexity points assigned to each evaluation unit. These points represent the expected information richness and are used to approximate experiment complexity

While most evaluation units will receive a default point value based on their type, the framework encourages evaluators to adjust the number of points for units that capture substantially more or fewer details than typical instances of the same type. This guideline reflects the flexibility of the ISA framework for scoping entities. For example, a soil preparation phase for a plant growth experiment containing roughly five steps could be modelled as five separate protocols or as one larger protocol. Because the rubric-based evaluation framework advises a maximum of 10 points for a protocol entity, a system modelling this as five different protocols could potentially earn or lose more points for the same procedural step as a system modelling this as one larger protocol, while both representations could be correct. To ease this tension, evaluators using the framework are encouraged to increase or decrease the complexity points of entities which were scoped notably larger or smaller compared to a baseline entities of the same type. These baseline entities can be found in Appendix B.

The total experiment complexity of a lab notebook is defined as the sum of the complexity scores of all evaluation units identified in the reference document. Let U denote the set of evaluation units, and let C_u represent the complexity score assigned to unit $u \in U$. The overall experiment complexity is then given by:

$$Complexity = \sum_{u \in U} C_u$$

4.2.2. Extraction performance

Extraction performance measures the ability of the system to construct ISA metadata relative to the information present in the lab notebook. The evaluation framework guides the evaluator in approximating extraction performance by awarding performance points for each evaluation unit. The performance points are a percentage of the complexity points for that unit. This means that a system perfectly extracting ISA from the lab notebook will be awarded the same amount of performance points and complexity points, and the extraction performance will be 100%.

The amount of performance points an evaluation unit is awarded is influenced by how accurate, complete, concise and relevant the provided metadata is and whether logical modelling choices have been made for the unit. Rather than evaluating these aspects separately, the evaluator considers them jointly when assigning points. To guide the evaluator in determining the performance points, for each evaluation unit the categories ‘insufficient’, ‘sufficient’ and ‘good’ are defined. For every evaluation unit, the framework provides guidelines as to under which conditions each category applies.

The guidelines for the different categories are specific to each evaluation unit. However, in general, a insufficient score is assigned when a required ISA object is not represented in the metadata despite being present in the lab notebook, in which case 0% of the points is awarded. A unit may also be categorised as insufficient when it is represented but contains severe deficiencies, such as substantial incompleteness or inaccuracy, such that correct interpretation of the experimental component is not

reliably possible. A sufficient score is assigned when the representation contains the necessary information to describe the experimental component, but may still include omissions or ambiguities. The core meaning of the experimental element is preserved, but the representation is not fully specified. A good score is assigned when the representation is complete and accurately reflects all relevant information available in the lab notebook. Minor omissions or ambiguities may still be present, provided they do not alter the overall correctness or completeness of the representation.

Along with determining the appropriate category, the evaluator determines the exact performance score to the evaluation unit as a percentage of its complexity score. The possible percentages for the categories 'insufficient', 'sufficient' and 'good' are 0-40%, 50-70% or 70-100%, respectively. When all evaluation units are scored, the overall extraction performance is obtained by aggregating the awarded points relative to the total experiment complexity. Let P_u denote the performance points awarded to evaluation unit $u \in U$. The overall extraction performance, expressed as a percentage, is then defined as:

$$Performance = \frac{\sum_{u \in U} P_u}{\sum_{u \in U} C_u} \times 100$$

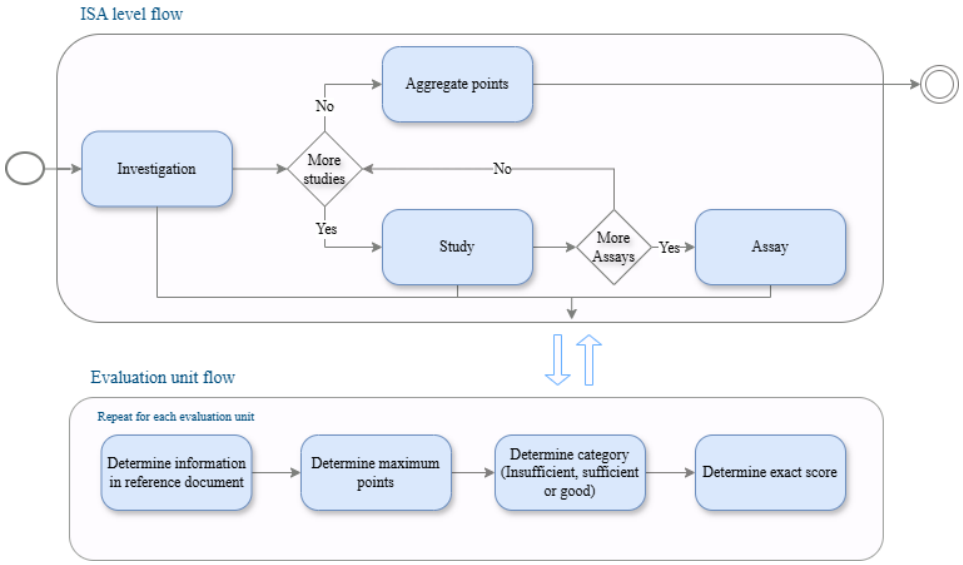


Figure 4.2: A diagram illustrating the evaluation workflow of the rubric-based framework. It shows the workflow from a higher level from Investigation, to Study, to Assay, and the field-level scoring process used to determine complexity and performance points.

4.2.3. Framework workflow

This section describes the step-by-step procedure followed by evaluators when applying the rubric-based evaluation framework. Figure 4.2 provides a conceptual overview of this process. At a high level, the evaluator proceeds through the ISA structure, starting at the Investigation level and continuing with all identified Studies and Assays. For each of these levels, all corresponding evaluation units are assessed individually. The framework defines a row to evaluate each unit individually. Some units defined by the Study object, such as materials, sources, samples, protocols or processes, may occur multiple times when multiple of these entities can be identified within the reference notebook. In such cases, the evaluator records how many distinct entities can be identified in the reference document and adds more rows for the units, so that each unit can be evaluated individually.

For each evaluation unit, the evaluator follows a consistent procedure, which is best explained using Figure 4.3, which provides an excerpt of the rubric used during evaluation. For each evaluation unit, the framework provides an explanation of the type of information that is expected. This explanation is given in column B as seen in Figure 4.3. First, the evaluator uses this explanation and identifies whether relevant information for the unit is present in the reference document. If no relevant information is present, the unit is marked as not applicable and assigned zero complexity points. If relevant information is present, the evaluator records this in column B, overwriting the explanation, and assigns

A	B	C	D	E	F	G
	Present in reference document?	Insufficient (<= 0.5)	Sufficient (0.5 - 0.7)	Good (0.7 - 1.0)	Maximum Points	Points obtained
Protocols	A protocol represents a series of steps taken during an experiment. A protocol description should declare the required parameters, which should be recorded once the protocol is applied. An application of a protocol is a process. Identify which concepts in the document correspond to a protocol and what parameters that protocol should have.	Protocol is not present or misses almost all important contextual information.	Protocol is present and reports at leas some important information. Some relevant parameters are correctly identified.	Protocol is present and reports all important information. Parameters are correctly identified and interpretable.	10 per protocol	
Sources	Sources are a starting biological material used in the study. Sources are unprocessed materials	Source missing or entirely incorrect	Source is present, but interpretation is difficult because of missing properties or incorrect ontology annotations	Source present and intepretable	3 per source	

Figure 4.3: A table excerpt from the rubric used to evaluate ISA extraction quality. It includes descriptions of expected information per field (column B) scoring categories (insufficient, sufficient, good) (Columns C, D, E) and point allocations (column F).

complexity points to the unit, either using the default value or adjusting it. The complexity points are recorded in column F.

Next, the evaluator determines the performance points for the evaluation unit, assessing it on accuracy, completeness, relevancy, conciseness and modelling logic while being guided the guidelines for the categories ‘insufficient’, ‘sufficient’ and ‘good’, which are given in columns C, D and E. The obtained performance points are recorded in column G. Lastly, when all evaluation units are evaluated, the experiment complexity and extraction performance can be calculated according to the formulas given in Sections 4.2.1 and 4.2.2.

4.3. Feasibility study results

As described in Section 3.2, extraction performance and experiment complexity were scored for each lab notebook using the evaluation framework presented in Section 4.2. The results are given in Table 4.2. As a reminder, experiment complexity represents the inherent complexity of the experiment described by a lab notebook, while extraction performance represents the ability of the system to translate natural language descriptions in a lab notebook into ISA objects that correctly and completely represent the described experiment. A higher experiment complexity number thus represents a larger extraction task. It is important to note that the experiment complexity points can vary depending on how the experiment is conceptualised. For this reason, complexity points should be interpreted as an approximate indicator not as an exact measure. The same holds for extraction performance. Because performance scores are based on holistic expert judgement, they should be interpreted as approximate indicators of metadata quality.

Notebook	Extraction Performance	Experiment Complexity
N1	58%	364
N2	56%	347
N3	52%	457
N4	57%	443
N5	46%	278
N6	60%	239*
N7	51%	367

Table 4.2: A table presenting extraction performance and experiment complexity scores for each of the seven lab notebooks. Note that the experiment complexity for N6 in reality is much higher. However, it was not possible to reliable determine this, as it was unclear what the experimental conditions were, and thus the Sample entities for this notebook were considered undetermined.

It is important to note that the experiment complexity of lab notebook N6 should be higher. The notebook did not clearly specify how many experimental conditions were studied, making it difficult to determine how many sample entities were required to model the experiment. Depending on interpretation of the lab notebook, the number of samples could reasonably range from 9 to 26, resulting in a wide range of possible complexity scores. Because of the ambiguity of the notebook, the points associated with sample entities were excluded when evaluating the extraction performance, as the unclarity made it

impossible to determine a fair target for extraction. Consequently, they were also excluded from the experiment complexity score.

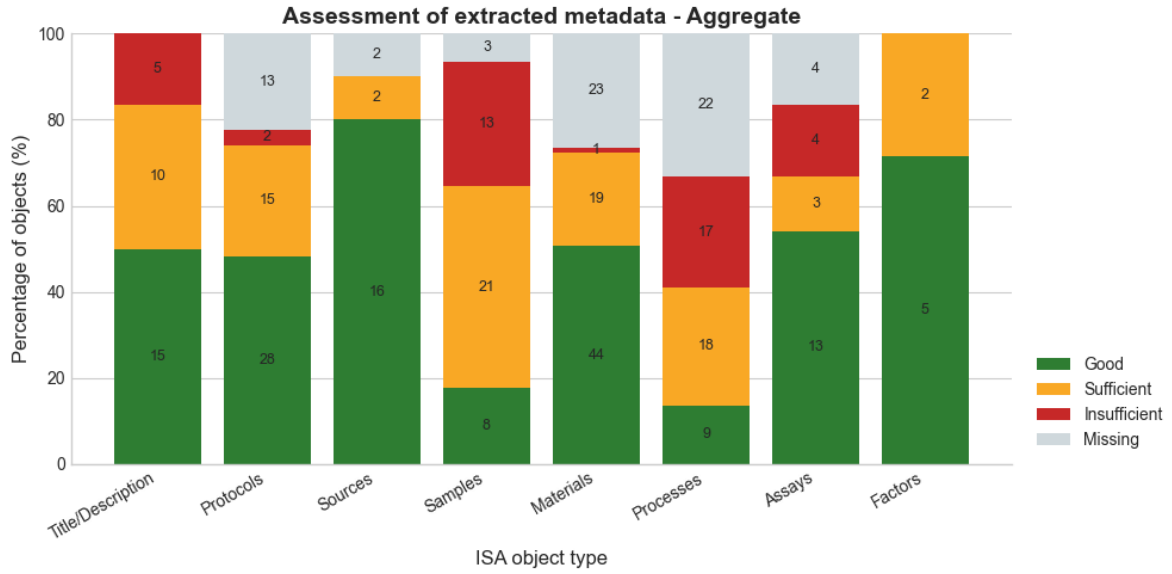


Figure 4.4: A bar chart showing the distribution of quality ratings (good, sufficient, insufficient, missing) across ISA evaluation units aggregated over all notebooks.

The bar chart in Figure 4.4 shows extraction performance per ISA entity type as an aggregated sum across all lab notebooks. Per ISA entity type, it reports how many entities were judged as represented well (awarded between 70% and 100% of points), sufficiently (between 50% and 70%), insufficiently (between 0% and 40%) or were missing from the metadata, while they were identified by the evaluator in the lab notebook (no points awarded). Bar charts showing the same information for each individual lab notebook can be found in Appendix C. Aggregating the results in Figure 4.4 across all ISA entity types, for a total of 337 identified entities, 41% of entities are represented well, 27% of entities are represented sufficiently, 12% are represented insufficiently, and 20% of entities are missing. Because the ISA framework allows multiple valid modelling choices for representing the same experiment, the entity type distribution should not be interpreted as an absolute measure of extraction correctness per ISA category. Instead, it provides an indication of where extraction errors tend to occur, given the conceptual modelling generated by the system.

4.3.1. Qualitative observations

In addition to the quantitative results, several qualitative observations were made regarding the extracted ISA metadata. These observations provide further insight into the strengths and limitations of the used prototyping setup.

First, the system demonstrates a proper ability to segment procedural descriptions in the lab notebooks into coherent protocol units. The extracted protocols generally represent logically consistent steps within the experimental workflow. However, the system does not always capture all relevant protocols, indicating limitations in recall. An example of this are the extracted protocols from notebook N6. The notebook first generally describes a procedure to create drought plates. Either with the medium MS (Murashige and Skoog), mannitol or sorbitol, after which the plates are PEG-infused (polyethylene glycol). After this, the notebook describes sterilising Arabidopsis thaliana seeds and moving them to a growth chamber. After which they are placed on the drought treatment plates. Lastly, a brief scanning procedure is described. The system extracts separate protocols for all of these procedural steps, except for how to produce the plates with medium MS, though some information for this step was captured in the protocol for creating the other plates. An example of a properly extracted protocol from the lab notebook is given in Figure 4.5. While the extracted protocol does not reflect that the seeds were placed in the fridge for three days, this information was reflected in a following protocol, which indicated a waiting time.

Notebook excerpt

Sterilised roughly 1000 Arabidopsis thaliana Col-0 seeds according to the following protocol:
- Add 70% EtOH for 5 minutes
- Add 20% bleach for 15 minutes
- Wash 3 times with sterile MQ
- Plate on 1/2 MS
- Place in fridge for 3 days.

Extracted ISA Protocol

Name: Protocol 1: seed sterilization and plating
Description: This protocol outlines a reusable seed surface sterilization and plating workflow for the current study. Seeds are treated with 70% ethanol for surface sterilization, followed by 20% bleach solution, then rinsed multiple times with sterile deionized water (MQ) to remove residual sterilants, and finally plated on solid 1/2 MS medium under sterile conditions. Exact exposure times and numerical parameters are not defined in this Protocol and will be specified at execution time in corresponding Processes. The protocol is intended for the current Study/Investigation and will be referenced by subsequent Processes."
Parameter list:

1. **Parameter:** Ethanol concentration (%)
2. **Parameter:** Ethanol exposure time (minutes)
3. **Parameter:** Bleach concentration (%)
4. **Parameter:** Bleach exposure time (minutes)
5. **Parameter:** Number of rinse cycles with sterile MQ water (%)
6. **Parameter:** Plating medium (solid 1/2 MS) (%)

Protocol type: seed sterilization and plating protocol

Figure 4.5: Notebook excerpt from N6 and the corresponding extracted ISA protocol.

A second observation is that the system models a one-to-one relationship between processes and protocols, even where this is not a logical modelling choice. As a reminder, a *Protocol* should capture a coherent description of steps taken during the experimental workflow, describing relevant parameters. While a *Process* applies a protocol, giving values for the defined parameters. A one-to-one mapping may be appropriate, because a protocol is applied exactly once in an experiment. However, it may be possible that multiple processes correspond to a single protocol. The system however, does not model this relationship. As a result, the generated representations may lack the expected level of informational details for the experimental workflow. An example of this is given in Figure 4.6

Third, as shown in Figure 4.4, the system is able to identify and instantiate a majority of the expected ISA entities. However, the quality of these representations varies. In many cases, the most essential information is captured correctly, but is accompanied by additional irrelevant details, which for example read like the underlying LLM reasoning about its own execution strategy. This can be seen in Figure 4.5. Part of the description includes content which is irrelevant for the description of the protocol ('Exact exposure times and numerical parameters are not defined in this Protocol and will be specified at execution time in corresponding Processes. The protocol is intended for the current Study/Investigation and will be referenced by subsequent Processes'). Another example of sufficient but clearly suboptimal representation is given by a material entity extracted from notebook N1, given in Figure D.1 in Appendix D. While the material and its function are represented, the characteristics of the material include unnecessary details such as a description, protocol references, completeness flag and provenance notes. These concepts do not need to be modelled according to the ISA framework and although they are not incorrect, could be considered irrelevant.

Fourth, the ISA framework defines experimental factors as independent variables manipulated by the experimentalist to affect biological systems in a measurable way [21]. The system generally succeeds in identifying the key factors under investigation. At the same time, it tends to identify additional factors that, while relevant to the experimental context, do not strictly conform to this definition. For example,

High-level description notebook N7
The lab notebook describes the preparation of treatment plates using four materials. Ten treatments are prepared, each using a different amount for each of the four materials.

Possible ISA modelling
This setup can be represented using a single Protocol defining parameters for the four components, combined with ten separate Process instances, each corresponding to a specific combination of parameter values.

System output
The extracted ISA defines one Protocol for plate preparation, but generates only a single corresponding Process. As a result, the nine additional treatment plate configurations are not represented.

Figure 4.6: Example of relationships modelled between extracted protocols and processes from notebook N7.

lab notebook N2 studies the flowering of Arabidopsis thaliana under four different drought conditions. Plants are grown under identical conditions, except for differences in watering treatments. The system correctly identifies the drought treatment as an experimental factor. However, it additionally identifies as factors to be: the genotype of the plant, the position in the growth cabinet, the nutrients applied and light, temperature and humidity conditions. While these variables influence plant growth, they are kept constant across all treatment groups and therefore do not constitute experimental factors according to the ISA definition. This example illustrates that the system does not fully distinguish between controlled variables and manipulated experimental factors.

Finally, it was observed that the system generates more detailed protocol descriptions than those explicitly present in the lab notebook, requiring the specification of parameters that were previously implicit. For example, in lab notebook N6, the preparation of drought plates for growing Arabidopsis thaliana includes an autoclaving step. While the notebook mentions autoclaving as a sterilisation process, it does not specify the conditions under which it is performed (temperature, duration). In contrast, the extracted protocol introduces standard autoclaving conditions and represents them as a parameter to be filled by a corresponding process application. The full lab notebook excerpt and corresponding extracted protocol can be found in Figure D.2 in Appendix D.

5 Discussion

This chapter discusses the designs and interprets the findings presented in Chapter 4 in relation to the research questions and broader objectives of this thesis. Section 5.1 reflects on the insights obtained from the prototyping process and formulates hypotheses to explain the observed patterns (RQ1). Section 5.2 examines the strengths and limitations of the proposed rubric-based evaluation framework (RQ2). Section 5.3 analyses the extraction performance of the prototype system developed for the feasibility study and discusses its implications for the viability of ISA extraction using multi-agent systems (RQ3). It further identifies failure modes of the prototype and considers how these inform future system design. In addition, it addresses limitations of the feasibility study setup.

5.1. Agentic design insights

Prototyping a multi-agent system for ISA extraction yielded several insights that suggest design implications, as presented in Section 4.1.1. While the primary objective of generating these insights was to inform the development of the prototype used to address RQ3, the insights can be informative for future system iterations as well. This section analyses each insight and formulates hypotheses to explain the observed behaviours.

First, a monolithic approach was observed to be structurally unstable, leading to the design implication that structured output is required. This instability is hypothesised to arise from the difficulty large language models experience in maintaining global structural consistency across iterative modifications and across large outputs. When adjustments are made to one part of the structure, constraints in other parts are not reliably preserved. As a result, ISA generation may constitute a structurally complex task that is difficult to perform without explicit output constraints.

Second, increased agent autonomy was observed to result in drift during task execution, suggesting that explicit task decomposition and agent profiling may improve ISA extraction performance. One possible explanation is that the used large language models have limited capacity for long-term planning and task decomposition. Alternatively, the observed drift may be attributed to insufficient feedback mechanisms, preventing agents from effectively correcting errors during execution. Another explanation might be that the given prompts were too unclear, so that the task was not properly specified, resulting in execution drift by agents.

Third, failures were observed when agents were required to infer ISA structure and instantiate corresponding ISA entities within a single step. This suggests that ISA extraction benefits from decomposition into two stages: conceptual modelling of the ISA experimental workflow and instantiation of entities. A possible explanation is that the reasoning required to perform both steps simultaneously exceeds the capabilities of the models used during prototyping, particularly when consistency with a global conceptual model must be maintained. ISA mapping may require global reasoning prior to entity instantiation, and separating these tasks may better align with the models capabilities.

Finally, instability in the construction of ISA models was observed, with identical notebooks leading to

different conceptual ISA representations. This is hypothesised to result from the inherent flexibility of the ISA framework. Given that human modellers may represent the same experiment in different valid ways, variability across runs of multi-agent systems is not unexpected. If a more standardized modelling approach is desired, future systems may explore incorporating stronger guidelines to constrain how ISA representations are constructed.

These findings should be interpreted with appropriate caution. The behaviour and outputs of multi-agent systems are influenced by numerous factors, including coordination strategies, tooling, prompting, and specific models used. Consequently, no claims are made regarding the universality of the observed insights. Establishing their general validity would require controlled experimentation in which such factors are systematically varied. Furthermore, as prototyping is inherently exploratory, alternative design choices may have led to different observations. The insights presented here primarily support the development of the prototype for the feasibility study (RQ3), while also contributing an empirical perspective to the literature on multi-agent system design.

5.2. Addressing the evaluation gap

To address the challenge of evaluating the extraction performance of an ISA extraction system, a rubric-based evaluation framework was developed. This section first highlights its contributions and implemented design principles, after which limitations and future directions for ISA extraction evaluation will be discussed.

5.2.1. Contributions and design rationale

Beyond formalising extraction performance, the framework contributes a structured means of characterising inherent experiment complexity. Because experiment complexity directly influences the difficulty of the extraction task, measuring it enables interpretation of system performance relative to task difficulty.

Rather than evaluating quality dimensions independently, the framework adopts a holistic scoring approach at the level of each evaluation unit. Evaluators are guided to consider aspects such as accuracy, completeness, relevance, conciseness and modelling logic jointly when assigning scores. This design reflects the interdependent nature of these aspects in ISA metadata and simplifies the evaluation procedure.

An important advantage of this holistic approach is that it accommodates the inherent flexibility of the ISA model. As discussed in Section 2.1, the same experiment may be represented in multiple valid ways, making it inappropriate to enforce a single gold-standard representation. The rubric takes this into account by evaluating whether the extracted metadata captures all information required to interpret the experiment using appropriate ISA entities in a meaningful manner, rather than whether it matches a specific structural configuration. For example, a system is not penalised for modelling a sequence of experimental steps as a single protocol rather than several smaller ones; instead, the evaluator can adjusts the scope of the protocol and assesses the adequacy of the representation. This allows the framework to evaluate the semantic adequacy of ISA extractions rather than strict structural similarity.

The framework also supports partial correctness, reflecting that partially correctly represented ISA entities may still provide substantial value to users. Even when an object is incomplete or imperfectly modelled, a structured baseline representation can reduce cognitive load for researchers constructing metadata for their experiments compared to reconstructing the experiment from scratch.

Additionally, the framework ensures a fair evaluation through both scoring and weighting of ISA entities relative to the reference lab notebook. Evaluators assess only the information that is present in the lab notebook, and the maximum obtainable score for an ISA component depends on how much relevant information the document provides. This design decision contributes to evaluating the system fairly, while ensuring comparability across systems. When multiple systems are evaluated on the same lab notebook, their scores become directly comparable, as they are judged against an identical baseline.

Furthermore, the rubric allows for fine-grained evaluation of various ISA components by evaluating metadata at the level of individual evaluation units, which could be individual fields or larger entities within the major Investigation, Study, and Assay objects. Entities within the Study that may occur

multiple times, such as protocol, processes and materials are also individually assessed. This supports the design principles of granularity.

Finally, the evaluation procedure is structured as a sequence of small, repeatable decisions, reducing cognitive load for evaluators by structuring evaluation into smaller, localised holistic assessments per evaluation unit. Evaluators first determine whether relevant information is present in the reference notebook and then assess how well it is represented in the extracted metadata. This makes usage of the framework relatively straight forward and supports inter-rater consistency, as breaking the task into smaller decisions reduces ambiguity and variability in interpretation.

5.2.2. Trade-offs and limitations

While the framework offers several advantages, its design choices introduce several methodological trade-offs and limitations. First, although the rubric is designed to support inter-rater consistency, its actual inter-rater reliability has not yet been established. As with any subjective task, evaluators may disagree, particularly when assigning scores within broad percentage ranges (e.g., 70–100% for “good”). A follow-up study is required to quantify inter-annotator agreement. Until such validation is performed, extraction scores should be interpreted cautiously, and only substantial differences should be considered indicative of meaningful performance differences. This limitation affects the comparability of results across systems or studies, as differences in evaluator judgment may introduce variance that is unrelated to system performance.

Second, the framework aggregates multiple aspects of extraction quality, such as accuracy, completeness, conciseness, relevance and modelling logic, into a single performance score per evaluation unit. This design choice is explainable in this development stage, as establishing all relevant quality dimensions might have proven difficult a priori to the feasibility study. Additionally, while this design choice simplifies the evaluation process and supports consistent scoring in the presence of interdependent criteria, it reduces diagnostic granularity, making it difficult to distinguish between different types of extraction errors. As a result, additional qualitative analysis is required to identify specific weaknesses in system behaviour. This suggests that future versions of the framework could benefit from explicitly separating evaluation dimensions.

Third, the evaluation process depends on the evaluator’s ability to interpret the reference notebook and to conceptualise plausible ISA representations of the described experiment. This may be challenging for evaluators with limited familiarity with ISA. Although the rubric includes guidelines for each evaluation unit, some degree of domain knowledge remains necessary. Because the framework relies on expert interpretation of both the reference notebook and the extracted metadata, the evaluation process is inherently interpretive. This introduces a degree of subjectivity into scoring. However, this is not a limitation of the framework alone, but a consequence of the ISA modelling task itself, where no single ground truth can be defined.

Fourth, while the rubric defines standard point allocations for each entity type and allows adjustments when more or less information should be represented than expected, the underlying point distribution remains somewhat arbitrary. Alternative weighting schemes may be equally defensible and could lead to different complexity scores. Because of this, the resulting scores may not be directly comparable across evaluations.

Fifth, the evaluation procedure requires manual assessment at the level of individual ISA entities. Although this enables fine-grained evaluation, it also makes the framework labour-intensive and difficult to scale to larger notebooks. This limits its applicability in settings were large-scale evaluation is required.

Finally, although the framework measures extraction performance and experiment complexity, it does not address the interpretability of the lab notebook. Lab notebook interpretability would represent the extent to which the lab notebook describes an experiment in a clear and interpretable manner. As explained in Section 2.2, lab notebooks may in general not be clearly interpretable. Additionally, this may differ strongly per notebook. This makes notebook interpretability a distinct factor influencing extraction difficulty, and its omission limits the framework’s ability to fully characterise task complexity. Incorporating this factor remains a direction for future work. Several other limitations discussed in this section suggest directions for future work as well. These suggestions will be described in Section 6.2.

5.3. Feasibility analysis

Table 4.2 reports extraction performance scores between 46% and 60% for the evaluated lab notebooks, indicating overall sufficient extraction performance, except for notebook N5. This is observed despite variation in experiment complexity across notebooks. These results suggest that, on average, sufficient ISA representations for the various parts of the experiment are extracted. However, Figure 4.4 suggests large performance differences per ISA evaluation unit. A substantial number (20%) of units were identified in the source lab notebook, but are missing from the extracted metadata entirely, meaning that essential information is not captured by the metadata. While 41% of entities are represented well. Overall, this means that with the current combination of the prototype, LLMs and lab notebooks used, extracting ISA metadata without missing or insufficiently representing substantial parts of the information, is not yet possible without supervision.

The following sections will firstly analyse these results and identify the different failure modes of the prototype in Section 5.3.1. Section 5.3.2 will synthesise the results into broader insights into the task of ISA extraction and its feasibility with multi-agent systems. Finally, Section 5.3.3 will describe the limitations of the performed feasibility study.

5.3.1. Failure modes

Two primary sources of error can be distinguished to contribute to the observed performance. First, as shown in Figure 4.4, extraction performance is significantly reduced by a large number of missing protocols, materials, and processes. These entities are identified as inherent to the lab notebook, but are not included in the metadata extracted from the notebook. These omissions indicate failures not in local extraction but in the system's ability to construct a coherent conceptual model of the experiment. For example, notebook N5, which achieved an extraction score of 46%, fails to model all process entities inherent to the experiment. This modelling failure leads to substantial loss of performance points obtained.

Second, objects evaluated as 'sufficient' or 'insufficient' reflect representation errors occurring during the task execution steps. In these cases, created objects might lack essential information required for interpretation, represent information partially inaccurate or in way which makes it difficult to interpret (see Section 4.3.1). Such representation errors further reduce extraction performance. Additionally, performance may be negatively affected when objects contain excessive irrelevant information. Even when the required information is present and the object is graded as 'good', a small number of points may still be lost due to a lack of conciseness.

Further analysis of Figure 4.4 indicates that the prototype performs worst on sample and process entities. This suggests that both the conceptual modelling and representation of these entity types are particularly challenging. According to the ISA specification, a sample object is a material which is derived from a source material through a sampling process. While a process represents an application of a protocol to some input material to produce some output. A hypothesis would be that these concepts are more difficult to model than other ISA concepts. These concepts require reasoning about derivation, chronology, and protocol application tasks that are more demanding than identifying stand-alone entities.

Given that a substantial portion of the performance loss is attributable to missing entities, improvements to the conceptual modelling step are likely to yield the largest gains in extraction performance. Enhancing how the prototype constructs the initial ISA model, therefore, represents a key direction for future work.

5.3.2. Implications for feasibility

The feasibility study provides the first empirical reference point for fully automated ISA extraction from unstructured lab notebooks. The observed extraction performance of 46-60%, therefore, does not indicate whether the prototype performs well or poorly in absolute terms; instead, it establishes an initial benchmark against which future systems can be compared.

The results highlight that ISA extraction is inherently challenging due to characteristics of both the ISA framework and lab notebooks themselves. ISA is intentionally flexible and allows multiple valid representations of the same experiment. This flexibility introduces conceptual instability: different modelling

choices can lead to different sets of ISA entities, even when based on the same underlying experiment. As a result, determining which entities should be modelled is not a deterministic task.

In addition, lab notebooks are personal and often ambiguous documents. They may omit context, use shorthand, or assume knowledge that is not explicitly stated. In several cases, even the evaluator found it difficult to determine the intended experimental workflow. These characteristics impose an upper bound on the quality with which metadata can be extracted by any automated system. However, if metadata extraction from lab notebooks becomes an integral part of scientists' workflows, scientists have an incentive to produce lab notebooks which are more interpretable by people other than the scientists themselves, and consequently also more interpretable for multi-agent systems.

The failure modes discussed in Section 5.3.1 indicate that fully automated ISA extraction, without parts of the metadata missing or insufficiently represented is not yet feasible with the current combination of prototype, LLMs and lab notebooks used. Omissions and representation errors prevent reliable end-to-end automation, where complete, accurate and relevant metadata is extracted without human intervention. However, 41% of the experimental information is captured well, meaning that the current feasibility setup can provide a meaningful starting point for researchers developing their own metadata. Researchers could use the current setup to extract metadata from their lab notebooks, after which they would need to adjust the extracted metadata to complete omitted information or correct errors. When researchers previously developed metadata for their experiments from scratch, this can already be considered a benefit to them.

Furthermore, the results of the feasibility study reveal where progress towards more reliable ISA extraction systems is most likely to come from: improving conceptual modelling and incorporating mechanisms for resolving ambiguity. More robust multi-agent system designs may prove better performance on this task. However, human-in-the-loop workflows appear to be the most realistic path forward. Given the inherent flexibility of ISA and the ambiguity of lab notebooks, achieving a satisfactory ISA representation will likely require human collaboration in ISA modelling choices. Section 6 explores these ideas further and outlines concrete directions for advancing both the conceptual modelling and system design components of future ISA extraction systems.

5.3.3. Study limitations

The experimental setup used to evaluate the prototype has several limitations. First, the system was executed only once per lab notebook. Due to the inherently non-deterministic nature of large language models, different runs are expected to produce different results. Evaluating performance across multiple runs and reporting average scores would provide a more reliable assessment of extraction performance.

Second, due to resource constraints only seven lab notebooks were evaluated using this study, which is a limited sample size. Additionally, different models were used across the evaluated lab notebooks. And so, differences in model capacity may have influenced extraction quality between notebooks. However, the results show similar feasibility results across the two model versions. As the focus of the study is feasibility and not model benchmarking, this model variation does not undermine the central conclusions of this work. Future work should investigate which model variants are most appropriate for ISA extraction, including considerations of cost, latency, and the practical constraints faced by scientific practitioners.

Third, the evaluation procedure was conducted by a single evaluator who was not an expert in ISA metadata creation. Relying on a single evaluator introduces subjectivity as to what constitutes as a 'good', 'sufficient' or 'insufficient' metadata representation and limits the reliability of the assessment. While consistency was promoted through a pilot study and the use of predefined guidelines, the absence of multiple evaluators prevents the assessment of inter-annotator agreement. The evaluator's lack of expertise was mitigated as much as possible through exposure to example ISA representations.

Finally, the measurement of experiment complexity based on ISA entities introduces inherent ambiguity. The ISA framework allows multiple valid representations of the same experiment, meaning that the number and granularity of entities are not uniquely defined. This makes grounding experiment complexity in ISA entities a suboptimal way to formalise this measure. To mitigate this issue as much as possible, point weights were adjusted when objects deviated significantly in size from typical instances,

ensuring that complexity reflects information content rather than entity count. Moreover, all valid ISA representations must capture the same underlying experimental information, which supports the use of this approach as a consistent approximation of experiment complexity.

6

Future Work

Future work can extend this research in several directions. First, approaches for improving the performance of ISA extraction systems are discussed in Section 6.1. Subsequently, directions are considered to enhance the evaluation of such systems in Section 6.2.

6.1. Towards operational ISA extraction

The implemented prototype exhibits limitations in its ability to construct accurate conceptual representations of ISA metadata from lab notebooks. In particular, errors in conceptual modelling often result in missing entities or entities that lack essential information. A primary direction for future work is therefore the incorporation of more advanced mechanisms for reasoning about the global structure of ISA representations.

Two main approaches can be identified. The first is the exploration of more advanced agentic strategies, in which multiple agents collaboratively construct a conceptual ISA model. For example, agents could engage in iterative refinement or debate-based interactions to converge on an improved representation. The second approach is to introduce a human-in-the-loop component into the conceptual modelling stage. In this setting, the system and the researcher collaboratively define a conceptual ISA contract that guides subsequent extraction steps. The conceptual modelling agent could, for example, generate one or more initial representations, after which the researcher selects and refines the most appropriate option.

Given the inherently subjective nature of ISA modelling, a human-in-the-loop approach may be particularly suitable. Such an approach also enables support for users with limited familiarity with the ISA framework. For instance, a dedicated ISA expert agent could assist users by answering ISA-specific questions, such as clarifying the role of parameters within ISA protocols or explaining the function of different ISA entity types.

A second direction for improving extraction performance lies in strengthening evaluation mechanisms within the system's execution loop. For example, intermediate outputs could be assessed and improved upon by a dedicated evaluation agent after each task is completed. Alternatively, evaluation could be transformed into a human-in-the-loop process as well, in which users are prompted to review and, if necessary, revise ISA entities as they are extracted. Presenting entities incrementally during execution may reduce cognitive load and encourage active user engagement, thereby improving overall metadata quality.

From a technical perspective, the current prototype leaves several implementation components required for a more complete system to future work. In particular, the current prototype lacks an ontology integration module, which would enable the grounding of extracted entities in controlled vocabularies. Furthermore, the process of connecting individual ISA entities into a coherent experimental graph workflow is not yet fully automated and requires a dedicated module for relationship construction. Addressing these limitations would move the prototype closer to a fully operational ISA extraction pipeline

and enhance the structural validity of the extracted metadata.

To illustrate how these improvements could be integrated into a unified system, Figure 6.1 presents a conceptual sequence diagram of an extended prototype system. The figure implements both proposed human-in-the-loop components and the to-be-implemented ontology and experimental graph construction modules.

6.2. Evaluation and workflow integration

With respect to evaluation, future work is needed to strengthen both the reliability and scalability of the proposed rubric-based evaluation framework. In particular, future research should establish inter-annotator agreement to assess the consistency and robustness of the evaluation procedure. As discussed in Section 5.2, the current framework relies on interpretative judgments and has not yet been validated in terms of inter-rater reliability, making such validation an important next step.

Another key direction for future work is to improve the scalability of the evaluation process. The current framework relies on manual expert annotation, which limits its applicability for larger notebooks or repeated evaluations. To address this, future work could build upon the current framework and implement it as an LLM-as-a-judge task, in which large language models are used to approximate rubric-based scoring. Such approaches may enable faster and more scalable evaluation. Furthermore, employing multiple independent LLM judges would make it possible to aggregate scores across judges, potentially improving the robustness of the evaluation by reducing reliance on a single annotator. However, the validity of such approaches must be carefully assessed against human judgment, and challenges related to prompt design and quality of model outputs should be taken into account.

Building on improved scalability, another direction for future work is to increase the diagnostic granularity of the evaluation. As discussed in Section 5.2, the current evaluation framework aggregates multiple interdependent aspects of extraction quality into a single performance score per evaluation unit. While this holistic approach supports consistent evaluation in the presence of flexible ISA repre-

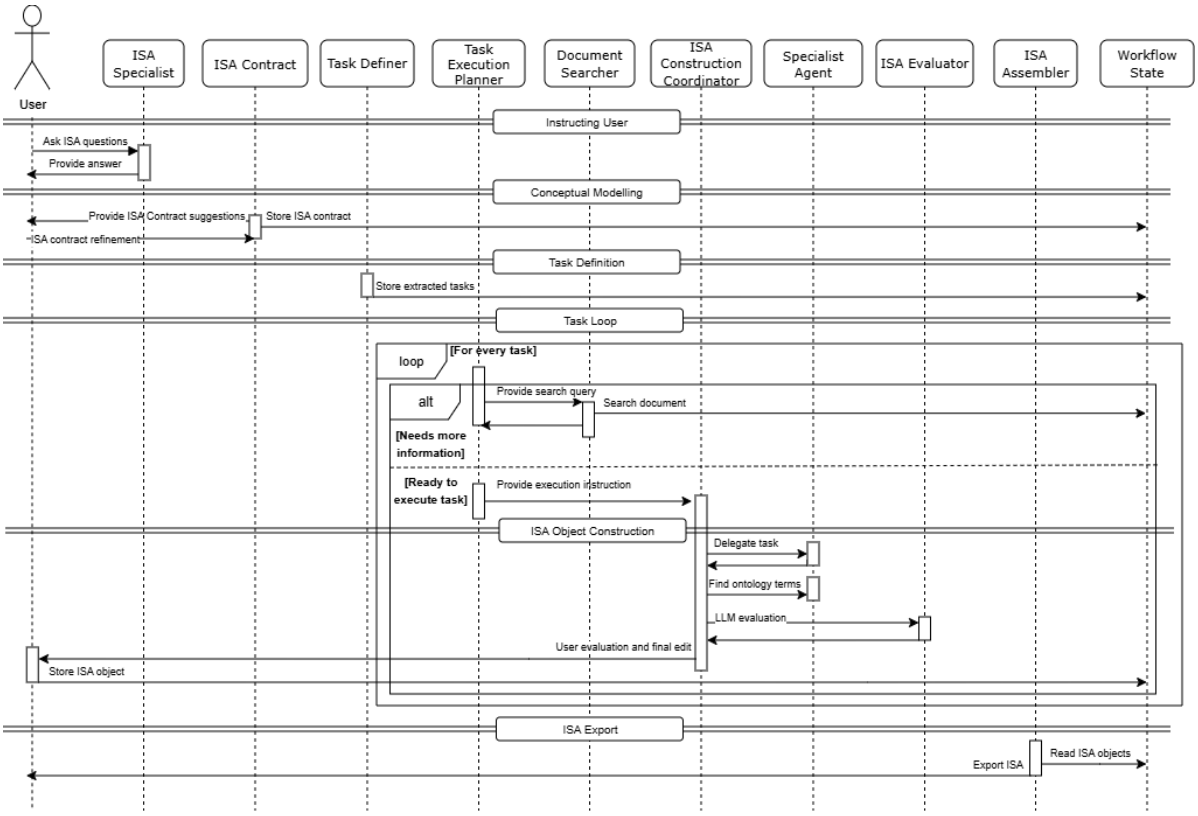


Figure 6.1: A sequence diagram showing the workflow of the proposed future iteration of a multi-agent ISA extraction system.

sentations, it limits the ability to analyse specific system failure modes. Future work could therefore investigate whether a set of evaluation dimensions can be more explicitly operationalised without sacrificing flexibility. As is emergent from the practical experience with the feasibility study, suggestions for these dimensions could be accuracy, completeness, relevancy, conciseness and ISA modelling logic. These dimensions could be complemented by a distinction between high-level modelling quality (e.g., coherence of the workflow and appropriateness of ISA structure) and entity-level correctness (e.g., correctness of individual entities). Such an extension would enable more fine-grained and actionable analysis of system behaviour. However, it could also potentially increase evaluation effort, as evaluators would need to assess multiple criteria per evaluation unit. This reinforces the importance of developing scalable evaluation approaches, such as LLM-assisted judging, to support more detailed evaluation procedures.

Additionally, future work may investigate the integration of ISA extraction systems into real-world research workflows. This includes studying how researchers interact with system-extracted metadata, as well as evaluating whether such systems effectively reduce the burden of metadata creation while maintaining quality and usability. Insights from such studies are essential to determine the practical value and adoption potential of agent-based ISA extraction systems.

7

Conclusion

This thesis addresses the research question, ‘How can a multi-agent system be designed and evaluated for ISA metadata extraction from unstructured lab notebooks and what can a prototype reveal about the feasibility of this approach in practice?’. This main research question is addressed through three contributions, addressing three sub-research questions.

The first research question, focusing on the design of a multi-agent system for ISA extraction, is addressed through a prototyping methodology which revealed design considerations for the ISA extraction task. These findings emphasise that ISA extraction is not merely a local information extraction task, but requires a form of conceptual modelling that requires global reasoning over the notebook as a whole. The developed multi-agent prototype system provides an instantiation of these design requirements.

The second research question, concerning how ISA extraction can be evaluated, is addressed through the development of a rubric-based evaluation framework. The evaluation framework contributes a structured manner for experts to evaluate a performed ISA extraction task, while also providing a manner to formalise the difficulty of the specific task through expressing the complexity of an experiment. It contributes a novel approach to evaluate a task that is inherently subjective and provides a foundation for comparative studies of future iterations of ISA extraction systems. It is encouraged that future work establishes an inter-rater agreement measure for the framework and investigates turning the framework and develops future iterations of the framework. For example by turning it into an LLM-as-a-judge task in order to scale up the evaluation process or by defining explicit quality dimensions instead of using a holistic judgement.

The third research question, concerning the feasibility of automated ISA extraction, is addressed through the evaluation of the prototype ISA extraction system on real-world lab notebooks. The results show that partial automation is achievable, with performance scores ranging from 46% to 60%, and 41% of the expected ISA entities represented well. As such, the extracted metadata may be valuable in providing a starting point for researchers developing ISA metadata for their experiments.

However, fully automated ISA extraction without human oversight is not feasible with the used combination of prototype, LLMs and lab notebooks, without parts of the metadata missing or insufficiently represented. The primary bottleneck lies in the phase of the extraction process where the to be constructed ISA workflow is conceptually modelled. As a result, a fully autonomous system may be unlikely to precisely extract the metadata as desired by the data producer without human oversight. This is why future work is recommended to construct a human-in-the-loop workflow for this part of the extraction process.

Furthermore, the feasibility study revealed that the interpretability issues of the lab notebooks themselves add to the complexity of the task. The ambiguous and personal nature of these documents makes it hard even for other human experts to evaluate what proper ISA metadata for a performed experiment would look like. However, once agentic extraction techniques are improved and used in practice more often to aid in the development of metadata, researchers might have an incentive to

develop more interpretable lab notebooks.

This work also has several limitations. The evaluation was conducted on a relatively small dataset of seven notebooks, which limits the generalizability of the results. In addition, the use of a single annotator introduces potential subjectivity in the scoring process. The prototype system was evaluated under specific configurations and using a single run per notebook, meaning that variability in model output was not explored. These limitations highlight the need for larger-scale and more rigorous evaluations of the ISA extraction task in future research.

Despite these limitations, this thesis makes several important contributions. It introduces a structured evaluation methodology for ISA extraction and provides empirical insights into the challenges of building multi-agent systems to automate this task. The developed prototype demonstrates that partial automation of ISA extraction is feasible, potentially already reducing the efforts researchers have to spend on metadata construction. Although fully automated extraction remains beyond the capabilities of the current prototype, the findings indicate promising directions for future work, indicating that the most effective path forward lies in collaborative systems that combine multi-agent intelligence with human oversight.

Use of AI

In line with TU Delft publishing policies[1], I acknowledge that certain passages of this thesis were rephrased using OpenAI's ChatGPT, Microsoft Copilot, or Perplexity. GitHub Copilot was used to assist with code development and the cover of this thesis was generated using Google's Gemini 3 Flash. All AI-generated text and code was critically reviewed and verified.

[1] https://www.tudelft.nl/library/actuele-themas/open-publishing/about/policies

References

[1] Varvara Arzt and Allan Hanbury. “Beyond the Numbers: Transparency in Relation Extraction Benchmark Creation and Leaderboards”. In: *Proceedings of the 2nd GenBench Workshop on Generalisation (Benchmarking) in NLP*. Ed. by Dieuwke Hupkes et al. Miami, Florida, USA: Association for Computational Linguistics, Nov. 2024, pp. 120–130. DOI: 10.18653/v1/2024.genbench-1.8. URL: https://aclanthology.org/2024.genbench-1.8/.

[2] John Dagdelen et al. “Structured information extraction from scientific text with large language models”. In: *Nature Communications* 15.1 (Feb. 2024), p. 1418. ISSN: 2041-1723. DOI: 10.1038/s41467-024-45563-x. URL: https://doi.org/10.1038/s41467-024-45563-x.

[3] Rafael S Gonçalves et al. “The CEDAR Workbench: An ontology-assisted environment for authoring metadata that describe scientific experiments”. en. In: *Semant. Web ISWC*. Lecture notes in computer science 10588 (Oct. 2017), pp. 103–110.

[4] Taicheng Guo et al. “Large language model based multi-agents: a survey of progress and challenges”. In: *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*. IJCAI ’24. Jeju, Korea, 2024. ISBN: 978-1-956792-04-1. DOI: 10.24963/ijcai.2024/890. URL: https://doi.org/10.24963/ijcai.2024/890.

[5] Jennifer Harrow et al. “ELIXIR: providing a sustainable infrastructure for life science data at European scale”. en. In: *Bioinformatics* 37.16 (Aug. 2021), pp. 2506–2511.

[6] Alan Hevner et al. “Design Science in Information Systems Research”. In: *Management Information Systems Quarterly* 28 (Mar. 2004), pp. 75–.

[7] Yu-Ning Huang et al. “Perceptual and technical barriers in sharing and formatting metadata accompanying omics studies”. en. In: *Cell Genom.* 5.5 (May 2025), p. 100845.

[8] Laura D. Hughes et al. “Addressing barriers in FAIR data practices for biomedical data”. In: *Scientific Data* 10.1 (Feb. 2023), p. 98. ISSN: 2052-4463. DOI: 10.1038/s41597-023-01969-8. URL: https://doi.org/10.1038/s41597-023-01969-8.

[9] Shankar Kumar Jeyakumar, Alaa Alameer Ahmad, and Adrian Garret Gabriel. “Advancing Agentic Systems: Dynamic Task Decomposition, Tool Integration and Evaluation using Novel Metrics and Dataset”. In: *NeurIPS 2024 Workshop on Open-World Agents*. 2024. URL: https://openreview.net/forum?id=kRRLhPp7C0.

[10] Anders Jonsson and Gunilla Svingby. “The use of scoring rubrics: Reliability, validity and educational consequences”. In: *Educational Research Review* 2.2 (Jan. 2007), pp. 130–144. ISSN: 1747-938X. DOI: 10.1016/j.edurev.2007.05.002. URL: https://www.sciencedirect.com/science/article/pii/S1747938X07000188.

[11] Sabina Leonelli. “Integrating data to acquire new knowledge: Three modes of integration in plant science”. In: *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences* 44.4, Part A (2013), pp. 503–514. ISSN: 1369-8486. DOI: https://doi.org/10.1016/j.shpsc.2013.03.020. URL: https://www.sciencedirect.com/science/article/pii/S1369848613000344.

[12] Xinyi Li et al. “A survey on LLM-based multi-agent systems: workflow, infrastructure, and challenges”. en. In: *Vicinagearth* 1.1 (Oct. 2024).

[13] Matthew S Mayernik. “Metadata accounts: Achieving data and evidence in scientific research”. en. In: *Soc. Stud. Sci.* 49.5 (Oct. 2019), pp. 732–757.

[14] J Menzel, P Weil, and S Y Nussbeck. “Metadata capture in an electronic notebook: How to make it as simple as possible?” In: *GMS Med Inform Biom Epidemiol* 11.1 (2015).

[15] Rajdeep Mondal et al. “Multi-agent AI System for High Quality Metadata Curation at Scale”. In: *bioRxiv* (Jan. 2025), p. 2025.06.10.658658. DOI: 10.1101/2025.06.10.658658. URL: http://biorxiv.org/content/early/2025/06/11/2025.06.10.658658.abstract.

[16] Fathoni A. Musyaffa, Kirsten Rapp, and Holger Gohlke. “LISTER: Semiautomatic Metadata Extraction from Annotated Experiment Documentation in eLabFTW”. In: *Journal of Chemical Information and Modeling* 63.20 (Oct. 2023), pp. 6224–6238. ISSN: 1549-9596. DOI: 10.1021/acs.jcim.3c00744. URL: https://doi.org/10.1021/acs.jcim.3c00744.

[17] Evangelia A Papoutsoglou et al. “Enabling reusability of plant phenomic datasets with MIAPPE 1.1”. In: *New Phytol.* 227.1 (July 2020). Publisher: Wiley, pp. 260–273. ISSN: 0028-646X. DOI: 10.1111/nph.16544. URL: http://dx.doi.org/10.1111/nph.16544.

[18] Joon Sung Park et al. “Generative Agents: Interactive Simulacra of Human Behavior”. In: *Proceedings of the 36th Annual ACM Symposium on User Interface Software and Technology*. UIST ’23. San Francisco, CA, USA: Association for Computing Machinery, 2023. ISBN: 9798400701320. DOI: 10.1145/3586183.3606763. URL: https://doi.org/10.1145/3586183.3606763.

[19] Laure Perrier, Erik Blondal, and Heather MacDonald. “The views, perspectives, and experiences of academic researchers with data sharing and reuse: A meta-synthesis”. en. In: *PLoS One* 15.2 (Feb. 2020), e0229182.

[20] Anushka Rajesh et al. “Improving the completeness of public metadata accompanying omics studies”. en. In: *Genome Biol.* 22.1 (Apr. 2021), p. 106.

[21] Susanna-Assunta Sansone et al. *Isa model and serialization specifications 1.0*. 2016.

[22] Susanna-Assunta Sansone et al. “Toward interoperable bioscience data”. In: *Nature Genetics* 44.2 (Feb. 2012), pp. 121–126. ISSN: 1546-1718. DOI: 10.1038/ng.1054. URL: https://doi.org/10.1038/ng.1054.

[23] Timo Schick et al. “Toolformer: language models can teach themselves to use tools”. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS ’23. New Orleans, LA, USA: Curran Associates Inc., 2023.

[24] Santiago Schnell. “Ten simple rules for a computational biologist’s laboratory notebook”. en. In: *PLoS Comput. Biol.* 11.9 (Sept. 2015), e1004385.

[25] Lynn M. Schriml et al. “COVID-19 pandemic reveals the peril of ignoring metadata standards”. In: *Scientific Data* 7.1 (June 2020), p. 188. ISSN: 2052-4463. DOI: 10.1038/s41597-020-0524-5. URL: https://doi.org/10.1038/s41597-020-0524-5.

[26] Noah Shinn et al. “Reflexion: language agents with verbal reinforcement learning”. In: *Proceedings of the 37th International Conference on Neural Information Processing Systems*. NIPS ’23. Red Hook, NY, USA: Curran Associates Inc., 2023.

[27] Aditi Singh et al. *Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG*. 2025. arXiv: 2501.09136 [cs.AI]. URL: https://arxiv.org/abs/2501.09136.

[28] Ian Sommerville and Pete Sawyer. *Requirements Engineering: A Good Practice Guide*. 1st. USA: John Wiley & Sons, Inc., 1997. ISBN: 0471974447.

[29] Hannes Ulrich et al. “Understanding the nature of metadata: Systematic review”. en. In: *J. Med. Internet Res.* 24.1 (Jan. 2022), e25440.

[30] Mark D. Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. In: *Scientific Data* 3.1 (Mar. 2016), p. 160018. ISSN: 2052-4463. DOI: 10.1038/sdata.2016.18. URL: https://doi.org/10.1038/sdata.2016.18.

[31] Katherine Wolstencroft et al. “FAIRDOMHub: a repository and collaboration environment for sharing systems biology research”. In: *Nucleic Acids Research* 45.D1 (Nov. 2016), pp. D404–D407. ISSN: 0305-1048. DOI: 10.1093/nar/gkw1032. URL: https://doi.org/10.1093/nar/gkw1032.

[32] Derong Xu et al. “Large language models for generative information extraction: a survey”. en. In: *Front. Comput. Sci.* 18.6 (Dec. 2024).

[33] Han Zhou et al. “Multi-Agent Design: Optimizing Agents with Better Prompts and Topologies”. In: *CoRR* abs/2502.02533 (Feb. 2025). URL: https://doi.org/10.48550/arXiv.2502.02533.

A

Example lab notebook

Figure A.1 and A.2 together provide an example of what a lab notebook might look like. The notebook documents an experiment that studies the development of roots and phenotyping characteristics of Arabidopsis thaliana under different drought conditions, either through a drought plate or through different watering regimes. Although this example is not an actual notebook used in the feasibility study of this thesis, it is inspired by those that were analysed and reflects the types of challenges encountered in them.

Dryzotron
Effect of different drought treatments on shoot + root growth. Mannitol induced or different water treatments (100% (mannitol), 80%, 60%, 30%). Nutrients were added regularly.

Started: 12/03
Mickey (middle rack → later more close to door)

Make drought plates
Make standard ½ MS medium. Add the following amount of mannitol:

Autoclave

add PEG according to the amounts in excel to your overlay solution (preferably while it is still >90°C to prevent contamination).

14/03

soil:

- Use soil from experiment 4b
- Dryzotron pots filled according to protocol established by <name>
- Max SWC was measured by soaking full pots for 24 hours, dripping them for 24 hours and drying them for ?? nights at 80°C
- MaxSWC = 155% [apparently 140%!]

stratification:
Seeds were sterilised as follows:

- 5 minutes in 80% EtOH
- Rinse 3-4 times with sterile MQ
- Store in fridge for 3 days

Figure A.1: Example lab notebook part 1

Growth conditions:

- long day conditions (16h light: 3.00 till 19.00, 8h dark)
- 21*C [CHECK!]
- 55% hum [MEASURE!]
- Light intensity –> see lichtmetingingen.docx
- Trays placed in middle rack - later meer richting deur

watering

- First few days 80% were watered to 70%.
- Per tray, 5 pots are weighted to determine tray-evaporation; water lost amount with pipet
- Evaporation data > see excel
- watering on Wednesday and Friday
- water around the plant
- From day 10 onwards, pots are not weighted and watered separately but per tray to minimize workload

Hi <name>! so you start with 80% rSWC for all containers right? And then from day 7 you let them dry until you achieved the rSWC you wish to achieve? And then from day 12 to 80% again?

Watering notes

DAG 5

- Pots 1 till ?? have 10 ml too little because decided to water till 80%
- add 5 ml nutrients

DAG 7

- !Less evaporation than thought
- First time pictures

DAG 8

- Dryzotron 1-5 temporarily at shelf next to door

DAG 10

- #322-455 watered at:
  - 100% - 99.4 gr
  - 79% - 77.3 gr
  - 61% - 60.6 gr
  - 30% - 29.5 gr
- watering till 9:45

13-M2-links 4&5

After DAG 15, the dryzotron back was openend. After 24 hours, 90, 80, 50 and 20 % soil capacity was reached and the rhizotrons were closed again.

- Sensors: [SENSOR DATA]
- Dryzotrans were scanned on day 17 and 20
- Score # leaves

Figure A.2: Example lab notebook part 2

B

Baseline ISA entity references for evaluation framework

This appendix defines baseline ISA entity references used to guide evaluators in assessing experiment complexity points in lab notebooks. Each baseline gives an example of the information details which are expected to be found in a lab notebook for a default ISA entity. These baselines serve as reference points to improve inter-rater consistency of the rubric-based evaluation framework, while preserving the flexibility of the ISA framework.

Experimental Factors

> Plants were grown under four drought conditions: 100%, 75%, 50%, and 25% soil water content.

This fragment defines an experimental variable (soil water content) and its levels. A well-represented experimental factor should clearly specify the variable under investigation and the distinct conditions or levels applied. A lab notebook typically only records one experimental factor, though multiple might be possible. Representing all experimental factors correctly corresponds to seven points.

Source

> Arabidopsis thaliana (Col-0) seeds were sown onto the prepared soil.

This fragment introduces the source material for this experiment. Properly extracted metadata represents both the plant species and the specific genotype used. A source detailing the presented information will be worth three points.

Sample

> 100 plants were divided among four treatment groups based on water availability (100%, 75%, 50%, 25%) and labelled as D1–D4.

This text fragment implies that from the source material, four sample materials can be derived, based on the different treatment groups. Properly constructed metadata will represent from what source material each sample is derived (which is the same source material in this case), and what treatment condition is applied. Ideally, it also records how many plants are included in each sample group. Samples detailing the presented information are by default awarded five complexity points per sample.

Material

> A commercial DNA extraction kit (Qiagen DNeasy Plant Mini Kit) was used for isolating genomic DNA.

This fragment defines a material used in the experiment, including its exact version and role. A well-represented material should specify all relevant characteristics about the material. Representing a material entity detailing the presented information correctly corresponds to five points.

Protocol

> Seeds were stratified at 4°C in darkness for 3 days to synchronize germination. After stratification, seeds were sown on soil and transferred to a growth chamber set to 22°C with a 16 h light / 8 h dark photoperiod. Plants were watered daily to maintain consistent soil moisture levels.

This fragment describes a protocol including its purpose, ordered steps, and parameters. A well-represented protocol should capture the exact sequence of steps performed. It will also define parameters for the two different temperature settings, duration and light cycle. Protocols are information-dense and therefore protocols detailing the presented information corresponds to ten points.

Process

> Seeds were stratified at 4°C in darkness for 3 days to synchronize germination. After stratification, seeds were sown on soil and transferred to a growth chamber set to 22°C with a 16 h light / 8 h dark photoperiod. Plants were watered daily to maintain consistent soil moisture levels.

The same fragment describes a process, which is a specific application of a protocol. A correctly represented process correctly refers to the corresponding protocol and to materials, sources or samples used as input or produced as output of the process. It should give the correct values for all parameters defined by the protocol. Representing a process detailing the presented information corresponds to ten points.

Assay

> Root length was measured using scanned images of the plants. Images were analysed using ImageJ software to determine total root length per sample.

This fragment describes a measurement performed on samples, including the method and data generation. A well-represented assay should specify what is measured, on which material, and how the measurement is performed. Representing an assay detailing the presented information corresponds to ten points.

C

Supplementary results

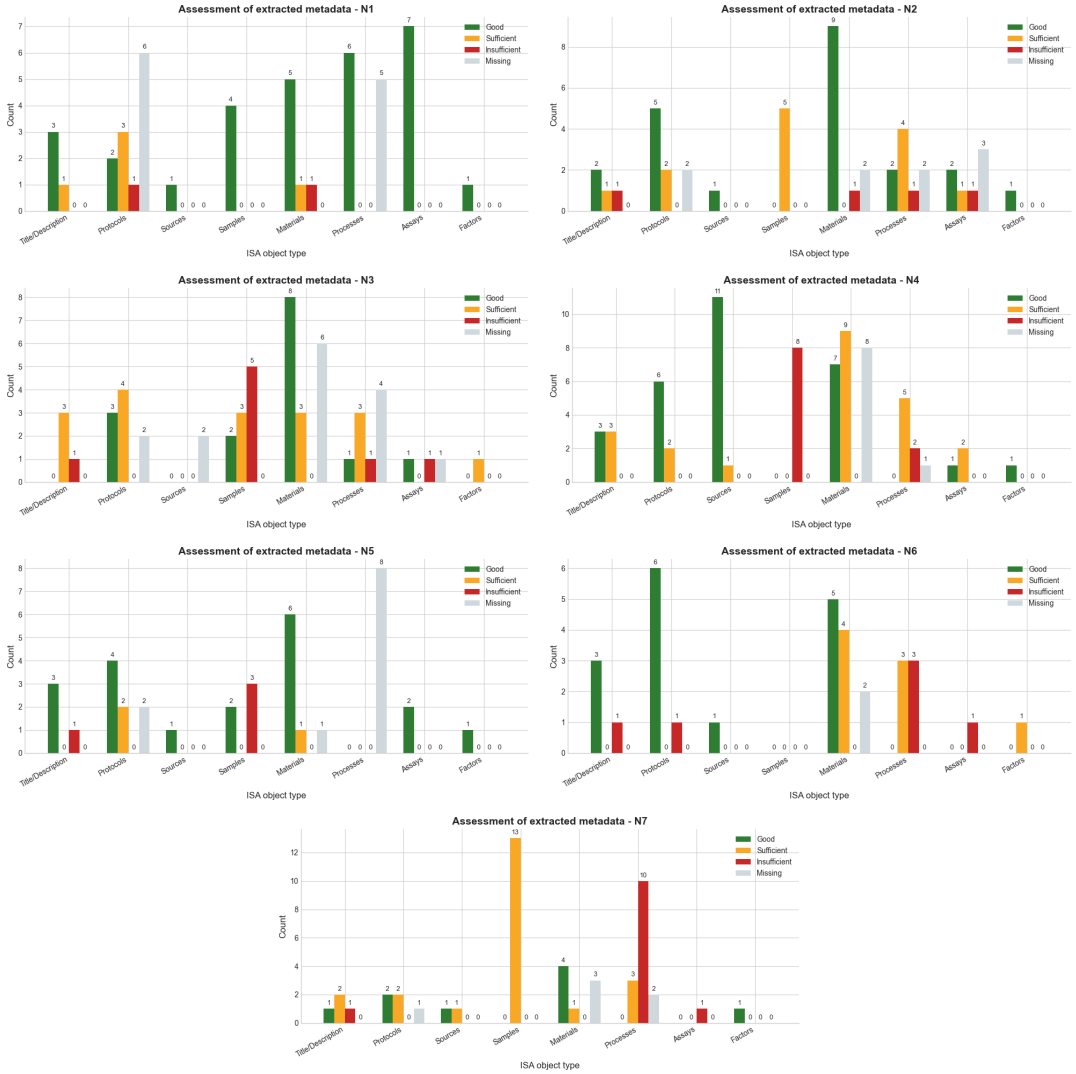


Figure C.1: A set of bar charts showing the quality distribution of extracted ISA evaluation units for each notebook individually

D

Extracted ISA metadata examples

Figure D.1 presents a material extracted from lab notebook N1. It serves as an example of an ISA entity that is judged as 'sufficient', specifically, it was awarded 70% of the obtainable performance points. The material and its function in the experiment (pots and trays used to grow plants in) are sufficiently captured, but an unnecessary protocol reference, description, completeness flag and notes are given as well. Still the material and its purpose can be interpreted.

Figure D.2 provides an example of a protocol extracted from lab notebook N6, accompanied with the relevant notebook fragment. While the notebook mentions autoclaving as a sterilisation process, it does not specify the conditions under which this is performed. The extracted protocol introduces standard autoclaving conditions and presents them as parameters to be filled by a corresponding process application.

Name: pots and trays (growth containers)
Characteristics:

- category: container type
  value: ['pot','tray']
- category: role
  value: growth container
- category: description
  value: Growth containers used in the experiment. These containers are used with the soil mix (sieved potsoil #4) and are handled in Cabinet 'Dombo' (shelf 2). The material is involved in pot filling and weighing steps. Referenced protocol: Protocol #5.
- category: protocol reference
  value: Protocol #5
- category: forward processes
  value: [pot filling, weighing]
- category: completeness flag
  value: partial; pot/tray dimensions, supplier, tray IDs, and marking procedure missing in source
- category: provenance evidence
  value: "<A specific section of the lab notebook is given here> The retrieved evidence lists soil mix and cabinet/shelf (Cabinet = Dombo; shelf 2) and procedures in which these containers are used (pot filling and weighing), but pot/tray specifics (pot volume, pot dimensions, tray IDs), supplier/catalog details, and any labeling/marking procedure were NOT present in the retrieved source"
- category: notes
  value: No pot/tray volume, dimensions, tray ID, supplier, or labeling/marking procedure information found in source.

Figure D.1: A sufficiently represented material extracted from lab notebook N1

Notebook excerpt

1/2 MS standard recipe
To make about 1 L of 1/2 MS
2.2 g of Muraskige and Skoog basal salts including vitamins
10 mL of MES buffer pH5.8
Set pH to 5.8
Note: usually, no adjustments are needed if your buffer is correct
Add 10 grams of plant agar to the bottle.
Autoclave.
Mannitol/sorbitol plates
Make standard 1/2 MS medium. Before autoclaving, add the following quantities of mannitol OR sorbitol (Mw is the same, so concentration is the same too).

Extracted ISA Protocol

Name: Protocol #2: Drought plate preparation (mannitol/sorbitol)
Description: Purpose/Scope: Protocol #2 describing drought plate preparation using 1/2-strength Murashige and Skoog (MS) basal salts with either mannitol or sorbitol as the osmotic agent.
Procedure steps:
1) Prepare 1/2 MS basal salts solution.
2) Add osmotic agent (mannitol or sorbitol) to achieve the target osmolarity (document the osmolarity target as a ProtocolParameter if defined).
3) Add agar to render the medium solid, and perform buffering/pH adjustment to a typical target of pH 5.7.
4) Dispense medium into plates and sterilize by autoclaving under standard conditions (e.g., 121 C for 15-20 minutes).
5) Post-autoclave handling: allow cooling to approximately 50 C before adding any heat-labile components and pouring plates.
Inputs/Outputs: Inputs = MS basal salts, water, osmotic agent, agar, buffering components; Outputs = prepared drought plates ready for pouring.
Notes: This Protocol is standalone and ready to be referenced by a Study/Process without requiring linkage to other Protocols at this time.
Parameter list:

- **Parameter:** MS basal salts strength (half-strength)
- **Parameter:** Osmotic agent type (mannitol or sorbitol)
- **Parameter:** Osmolarity target (mOsm/L)
- **Parameter:** pH target (5.7)
- **Parameter:** Agar concentration (g/L)
- **Parameter:** Autoclave conditions (temperature and duration)

Figure D.2: Notebook excerpt from N6 and corresponding extracted ISA protocol