The Annotation Guideline

1. Statement of Purpose

This document provides a guideline for the annotators/expert reviewers (AER) to aid associating cases of technical debt (TD) with relevant ISO/IEC 5338 or ISO/IEC 12207 processes that can prevent such cases. The main goal of this task is to create a systematic mapping of TD. The future use-cases of this mapping could be the development of a Retrieval-Augmented Generation (RAG) or Cache-Augmented Generation (CAG)-based artifact.

2. Task Definition

A document including project details, a list of cases of TDs, its category and their root causes is given to an AER. Then, she is expected to decide for each TD whether one or more processes from ISO/IEC 5338 or ISO/IEC 12207 standards might prevent it and choose the appropriate process type (A or B) considering if the relevant process addresses the TD case directly or indirectly. The process types' aim is to aid AER when they provide reasoning for the mapping process.

3. Annotation types

Each case of TD should be assigned one or more processes with the necessary related excerpt from the ISO standard document that address the case itself or its root cause.

4. Data preparation

The AER should keep the unique project IDs to the cases of TD by concatenating two columns in the original dataset that contain Company IDs and Project IDs. She should also keep necessary information for the process, such as the main category of TD, its sub-category, the root cause and the ideal solution (if it is present). This ensures that sensitive and private information is not used in the mapping process. Besides, a reference table has been created to assign unique IDs to each main TD category and sub-TD category, with ID assignments made according to this table. Only the Data Debt main TD category has been included in this study (TD ID=1), and the sub-TDs under this category have also been structured as outlined below, incorporating the TD ID to reflect the relationship with the main TD.

Sub-TD ID = TD ID.X (e.g., 1.1, 1.2, etc.)

5. Annotation format

Each row in the mapping document consists of a case of TD followed by a process that can prevent it. Each attribute of the mapping is defined in the following table:

Attributes	**Definitions**
Project ID	Unique project identifier associated with the case of TD
TD ID	Unique identifier associated with the main category of TD
Main TD Category	Name of the main category of TD
Sub-TD ID	Unique identifier associated with the sub-category of TD
Sub-TD Category	Unique identifier associated with the sub-category of TD

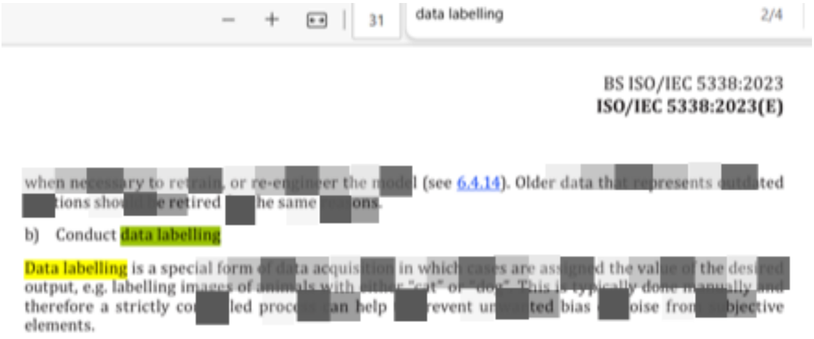
Abstract Case Description	Description of the case
Root Cause of TD	Description of the root cause of the case
ISO5338/ISO12207 Processes	Name of the relevant process associated with the case of TD
Source	Source of the process: ISO5338 or ISO12207
Activities and Tasks	Relevant activities and tasks that address the TD
Details	Details of the relevant process that address the TD
Reason for Association	Detailed description of the reason for associating the relevant process with the case of TD
Applied or Ideal Solution	Ideal solution to the case of TD

The ones highlighted as red come as empty cells and should be filled by annotators

6. Annotation steps

- **Step 0:** Preparation for information retrieval and information extraction
    - Study ISO documents by reading the document thoroughly, reviewing their structure with process names and taking notes of important sections.
    - Study the list of TD cases by taking notes of significant keywords and main TD categories or sub-categories.
    - Review key processes from the ISO documents that are most likely to be related to the given main TD categories or sub-categories.
        - For example, if the main TD category is "Project Management", such cases will be most likely mapped to processes such as Project Planning Process (6.3.1) or Project Assessment and Control Process (6.3.2).
- **Step 1:** Information retrieval
    - Scan the processes that cases may be related to, based on knowledge. Use your notes if necessary.
    - Search ISO documents for candidate processes that can prevent the given TD case by using keywords, in order to ensure that the potential relationship is not overlooked.
    - Select processes with relevance.
        - For example, the given case of Resource Management Debt is "No previously labelled data". Its root cause is "Limited human resources for labelling" and its TD sub-category is "Human Resource Management Debt".
        - Keywords identified from this case include "data labelling" and "human resources". The search for "data labelling" in ISO/IEC 5338 produced the following results:

Some areas in the screenshots have been blurred to prevent copyright violation.



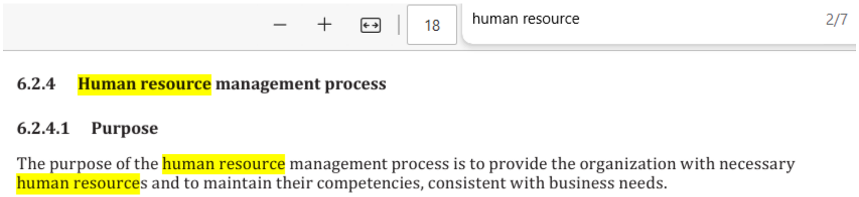
BS ISO/IEC 5338:2023
ISO/IEC 5338:2023(E)

when ne[redacted]ry to re[redacted], or re-e[redacted]er the [redacted]l (see 6.4.14). Older data th[redacted]resents [redacted]ted [redacted]tions sho[redacted]e retired [redacted]he same [redacted]ons.

b) Conduct data labelling

Data labelling is a special form [redacted]a acquis[redacted] in which[redacted]s are as[redacted]d the valu[redacted] the des[redacted] output, e.g. labelling im[redacted] therefore a strictly co[redacted]led proc[redacted]an help [redacted]revent un[redacted]ted bias [redacted]oise fro[redacted]bjective elements.

    - This activity belongs to the AI Data Engineering Process (6.4.8). Since this process is relevant to the TD case, it can be mapped to it.
    - The search for "human resource" in ISO/IEC 5338 produced the following results:



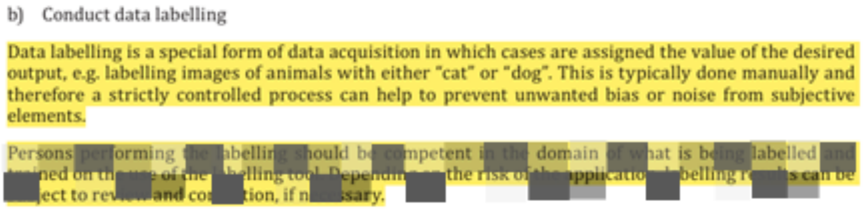
6.2.4 Human resource management process

6.2.4.1 Purpose

The purpose of the human resource management process is to provide the organization with necessary human resources and to maintain their competencies, consistent with business needs.

    - ISO/IEC 5338 does not provide additional tasks in the Human Resource Management Process (6.2.4) and asserts that activities and tasks defined in ISO/IEC 12207 shall apply. Therefore, the annotators can refer to the ISO/IEC 12207 when mapping this process to the given technical debt case.
    - In this example, two processes that can be mapped to the given technical debt case were identified: AI Data Engineering Process (6.4.8) from ISO/IEC 5338 and Human Resource Management Process (6.2.4) from ISO/IEC 12207.
  - Record the ISO standard and the process name (e.g., ISO/IEC 5338, AI Data Engineering Process (6.4.8))
- **Step 2:** Information extraction
  - Identify specific activities and tasks from the Activities and Tasks section of the selected processes that can prevent the given TD. This section is not available in all processes, if AER have inferred this from another item in the standard, they write NA to the Activities and Tasks attribute.
  - Identify specific clauses or paragraphs that address and prevent the given TD case.
  - Record the extracted activities and clauses.
    - For example, for the case above, AI Data Engineering Process (6.4.8) has the activity called "Conduct data labelling". The first two paragraphs of this task explain what constitutes data labelling and how it should be performed. Therefore, the first two paragraphs can be included in the

results:

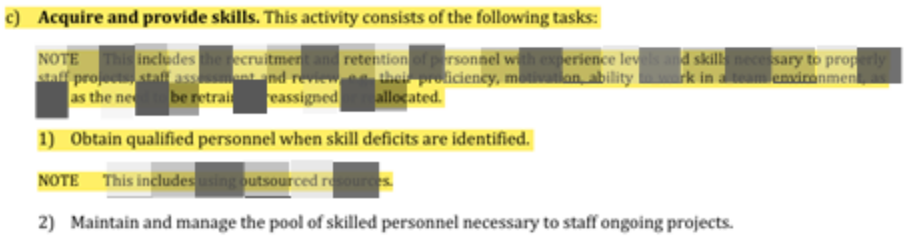


b) Conduct data labelling

Data labelling is a special form of data acquisition in which cases are assigned the value of the desired output, e.g. labelling images of animals with either "cat" or "dog". This is typically done manually and therefore a strictly controlled process can help to prevent unwanted bias or noise from subjective elements.

[redacted]

- The Human Resource Management Process (6.2.4) has three activities: “Identify skills”, “Develop skills” and “Acquire and provide skills”. The root cause of the given case is “limited human resources” (i.e., the case is about having few numbers of staff to do the job and needs more people to do it) which means that the last activity has the highest relevance to the root cause. Furthermore, one of the tasks of this activity include obtaining personnel to properly staff projects; therefore, this portion is the most relevant part of the process for mapping:



c) **Acquire and provide skills.** This activity consists of the following tasks:

[redacted]

1) Obtain qualified personnel when skill deficits are identified.

[redacted]

2) Maintain and manage the pool of skilled personnel necessary to staff ongoing projects.

7. LLM as a judge for verification

Large Language Models (LLMs) such as GPT-4o-mini will be used to validate manual annotations. Prompts from the appendix and the GitHub page[1] will be utilized for verification of annotations. The LLM will be instructed to detect false positive or false negative associations, provide reason for association and the confidence level in percentage.

8. Manual Verification

Two domain experts will be randomly assigned cases, and the established relationships will be manually verified in terms of completeness and accuracy, particularly by evaluating the activities and tasks, as well as the details attributes. The expert annotators will provide feedback which will be used to finalize the mapping tables.

[1] https://github.com/MEdata4/TD

Appendix

False Negative Verification Prompt:

"You are analyzing #technical_debt (TD) in #AI_development_projects. Your goal is to evaluate whether each <Data Debt Sub-TD Category>, <Abstract Case Description> and <Root Cause> relates to specific <ISO/IEC Process> and <ISO/IEC Activities, Tasks & Special Notes>. Complete the following sentence for each input group I will give.

<Data Debt Sub-TD Category> and <Abstract Case Description> are associated with ISO/IEC 5338 <ISO/IEC Process> because …"