A UML Activity Diagram Extension and Template for Bioinformatics Workflows: A Design Science Study

Bachelor of Science Thesis in Software Engineering and Management

Laiz Heckmann Barbalho de Figueroa
Rema Salman

Department of Computer Science and Engineering
UNIVERSITY OF GOTHENBURG
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2019

This paper presents a UML extension with its concrete syntax and written template to document bioinformatics workflows. The produced artefacts were evaluated and validated in cooperation with three facilities and seven bioinformaticians following a data science methodology. The results of this work are our contribution to the bioinformatics domain and the ongoing research: Optimized Bioinformatics Workflows from Requirements Engineering of Solution Specifications.

Supervisor: Jennifer Horkoff
Examiner: Richard Berntsson Svensson

University of Gothenburg
Chalmers University of Technology
Department of Computer Science and Engineering
SE-412 96 Göteborg
Sweden
Telephone + 46 (0)31-772 1000

Department of Computer Science and Engineering
UNIVERSITY OF GOTHENBURG
CHALMERS UNIVERSITY OF TECHNOLOGY
Gothenburg, Sweden 2019

A UML Activity Diagram Extension and Template for Bioinformatics Workflows: A Design Science Study

Laiz Heckmann Barbalho de Figueroa
Software Engineering and Management, BSc
University of Gothenburg
Gothenburg, Sweden

Rema Salman
Software Engineering and Management, BSc
University of Gothenburg
Gothenburg, Sweden

***Abstract*—Bioinformaticians execute daily scripted workflows, also known as pipelines, to process data. There are many tools to manage and conduct these workflows, but there is no domain-specific way to textually and diagrammatically document them. Consequently, this thesis, part of ongoing research, aims to extend the Unified Modelling Language (UML) activity diagram (AD) to the bioinformatics domain by including domain-specific and understandable concepts and notations. Additionally, a template was created to document the same concepts in a written format. A design science methodology was followed, where three iterations with seven domain experts tailored the artefacts, by extending the concepts and improving the artefacts' usage, terminology, and notations. The extended UML AD and its concrete syntax received positive validations because of their simplicity. On the contrary, the written documentation template was rejected due to its amount of text and complexity. Another finding was that the domain experts requested a different software to do the modelling with the possibility to have the written documentation automatically generated from the drawn diagram to save time.**

***Index Terms*—UML, activity diagram, workflow, bioinformatics, documentation.**

I. Introduction

Bioinformatics is a branch of biology, which is connected to computational methods [1]. One of its core competencies is to describe the understanding of the technologies for biological data generation [1]. The data generation for biological analysis, such as DNA sequencing, requires several connected tools in a pipeline or workflow [2]. The latter is defined as a sequence of tasks that cover the steps of a process from initialisation to producing final results [3]. Similarly, a pipeline is a flow, where files are shepherded through a series of transformations [2]. Bioinformaticians create workflows that need to be followed precisely to achieve satisfactory results [4]. To design and manage these workflows, bioinformaticians use a mixture of tools, frameworks, and requirements from various online sources [3]–[8], which were tailor-made for a specific organisation or process resulting in partial solutions. Work in [7] reported the usability challenges faced by bioinformaticians when using the available tools, including the limitations of attempting to visualise data and patterns for workflows. Additionally, [9] describes the lack of features, notations and or concepts, such as the absence of loops, support of control-flow operations, and connections between pipelines modules. These limitations hinder bioinformaticians, researchers, and practitioners to visualise, share, and identify workflows problems as well as replicate the analysis.

The literature reports several languages and approaches, which can be used to describe bioinformatics workflows. One of these approaches is the Domain-Specific Language (DSL), which is a famous research area in the Software Engineering (SE) field that helps tailoring languages to specific domains [10]. Furthermore, Unified Modelling Language (UML) is widely adopted and holds an extended range of diagrammatic notations, construction and systemic documentation, being used to design, capture, describe and specify any complex system and processes [11], [12]. Additionally, the authors in [11] stated that UML could capture biological systems, requiring specific-domain extensions, corroborated by [5], [9], [12]. However, there is no evidence that these languages and approaches can solve the problems identified in this thesis.

After an attempt of using several modelling languages, Horkoff et al. identified in [5] that UML Activity Diagram (AD), one of the UML behavioural modelling approaches [13], was the most suitable to capture biological perspectives effectively and represent bioinformaticians' workflows. On top of that, some reported gaps in the UML AD concepts are lack of motivation, sources, thresholds, files, etc. [5]. Further work identified several problems, for example, the high level of abstraction and the misuse of modelling elements, in the mechanical engineering domain, due to the lack of knowledge in the modelling specification and time limitation [14]. These problems hinder mechanical engineers from identifying issues and missing steps in workflows [14]. Additionally, the deficit of standard workflows documentation among bioinformaticians and facilities leads to sub-optimal documentations and personalised workflows creation. Consequently, Horkoff et al. suggested further studies to evaluate and extend their UML AD proposal and offered a draft for the workflow elicitation process, which adopted requirement engineering solutions for bioinformatics domain [5]. As known in the SE field, requirements elicitation is essential to system success, but sometimes it is underestimated, ending with incomplete and inexact requirements [15].

The purpose of this study is to extend the UML AD meta-model, create new concrete syntax, and generate a Workflow Documentation Specification Template (WDST). These artefacts intend to: increase efficiency to manage bioinformatics workflow; establish a shared understanding and consistency between the activities and tasks of the involved parties; create a sharable documentation set to provide a clear vision of the process; support training new bioinformaticians; identify problems during any step of the workflow design; reduce the bioinformaticians reliance on individual interpretations; increase the replication precision of the analysis; lessen the involvement of knowledgeable people to perform the workflow [5], [14]; and increase the decisions traceability, when using a mixture of conceptual workflow systems or scripting languages [16].

If the artefacts, UML AD meta-model extension, concrete syntax, and WDST, were used individually, they would be unable to capture the complexity of the bioinformatics workflows and the domain needs. The meta-model extension introduces new concepts to the language; however, without their concrete syntax, the modelling part of the language is absent. Considering individuals' graphical and written documentation preferences, the WDST provides a written format of the workflow diagram, acting as a complement to the graph. In addition, the WDST relates directly to the meta-model concepts by containing all of the attributes.

This thesis is part of ongoing research reported in [5], and its findings were used as the starting point of this research. The aim is to collect qualitative data, from bioinformaticians in Gothenburg, following a design science methodology to answer the main research question and its three sub-questions:

RQ1: How can we extend the UML activity diagram and use a template for workflow documentation to understand and improve bioinformatics workflows?

- *RQ1.1: What are the defining and unique characteristics of bioinformatics workflows compared to standard workflows?*
- *RQ1.2: How should workflows, including the concepts discovered in RQ1.1 be visualised to be understandable by the bioinformaticians?*
- *RQ1.3: How can we design a useful and understandable template to document the concepts from RQ1.1 from the bioinformaticians viewpoint?*

RQ1.1 aims to identify the needs for notations and concepts while creating and documenting bioinformatics workflows. RQ1.2 aims to understand if the existing or proposed notations are understandable by the bioinformaticians. Lastly, RQ1.3 is related to the design of a written template to document the workflow based on the requirement specification from the SE field and its evaluation by the domain experts.

The rest of this document is structured as follows: section II brings the related work covering UML AD and stereotype profiles, visual notations design, requirements engineering, DSL versus UML, UML usage problems, and the found UML extensions. Section III describes how the design science methodology was used for raising the data and how they were analysed. In section IV, the results for each iteration are presented together with the artefacts, while section V discusses the findings and the study limitations, and section VI concludes the paper.

II. Related Work

A. Requirements Engineering and Documentation

Horkoff et al. in [5] introduced a draft for requirements elicitation to bridge between the bioinformatics and SE domains and capture bioinformatics workflows. This draft will be used as a base for creating the WDST in this thesis since requirements elicitation is a common practice in the SE field. In the requirements elicitation process, raw data are collected from the stakeholders and end-users to produce final results, and these data are used to verify and validate these results [17]. Work in [3] reported a similar process to this practice, which was followed to document a workflow specification for a genomics data analysis. Furthermore, several templates for requirement elicitation of SE products exist in the literature, as the proposed semi-formal template by Gallina and Guelfi in [18]. This template comprises commonalities and variabilities that are formulated based on use case scenarios and domain specification [18]. Nevertheless, these templates and the one in [18] do not meet the bioinformaticians needs, because they are created for SE product lines instead of documenting bioinformatics workflows. Moreover, the WDST will contain the hierarchical content structure of requirement artefacts for the SE domain, as mentioned by the authors in [19], and the extended meta-model will mirror the concepts and methods. This content is composed of concepts, syntax, and methods, where concepts describe domain-specification elements and their relationships [19]. Each concept has a specific representation or concrete syntax, and the methods describe the procedure of approaching these concepts [19]. However, the works in [3] and [17] are not particularspecific to the bioinformatics domain, while the work in [19] is in the SE field and does not cover any aspects of bioinformatics requirement artefacts generation.

B. UML Activity Diagram and Stereotype Profiles

UML AD is characterised by the behavioural semantics of the UML semantic categories [13], making the AD graphical notations appropriate to model system-level behaviour, workflows, and business processes [3]. AD behavioural characteristics describe dynamic aspects of the systems, being it a flowchart that consists of control flow from an activity to another [13], [20], [21]. This flow is composed of different elements, such as join, fork, decision [13], [20], initial, and final nodes [21]. The activity, in the diagram, models a task that the system must perform, and the connecting arrows represent transitions or activity edges [13], [20], [22]. Additionally, AD has swimlanes or activity partitions that divide the activities based on their common characteristics or according to the actors who execute the activities [13], [20], [22]. UML AD includes visual modularity and hierarchical structure, which make it capable of representing complex systems [23].

The abstract syntax is visualised as a UML meta-class that constructs, exchanges, and represents each element of UML diagrams, while the concrete syntax represents the graphical notations of the elements. Meanwhile, the UML semantics is the meaning of concepts and relations between the elements, which is usually used for code generation, model execution, or semantic model analysis [13]. Furthermore, the creation of UML packages, stereotype profiles, allow UML meta-models extension and adaptation [13], while keeping the existing UML syntax and semantics of the elements. These stereotypes can have a different representation and extend either a meta-model class or another profile [24], [25]. This approach of UML meta-model tailoring is a light-weight extension [25]. However, there is still no specific profile found for bioinformatics domain, except the outlined extension provided by Horkoff's et al. that does not include syntax, semantics, or a meta-model [5].

C. General Visual Notation Design

Visual notations for bioinformatics workflows will be produced in this study as one of its final artefacts. Visual notations have a wide role in Requirements Engineering (RE) [23], where RE Modelling and Data Flow Diagrams (DFDs) are the most used modelling techniques [26]. Experimental studies show that visual notation design can improve end-user understanding of RE diagrams by more than 50% [27]. The visual alphabet theory includes the planar and retinal variables that influence the human ability to span or judge symbols, such as the number of different shapes, colours, textural encodings [23]. Due to the lack of design rationale in i* modelling language and most RE notations, examples of different visual representations and geometric shapes were defined [23]. Moody described the Physics of Notations theory, containing the design principles to achieve effective visual notations that are semiotic clarity, visual expressiveness, perceptual discriminability, complexity management, perceptual directness, and graphic economy [26]. Further, the theory has a hybrid representation, combined symbols, which helps to enforce the used text to the meaning of the graphics [26]. The bioinformaticians currently overuse this hybrid theory because their models hold notations overload. For instance, all of their concepts are represented by boxes and arrows, which needs to be differentiated by text. Thus, this thesis introduces SE approaches, RE, and modelling languages to the bioinformatics domain and uses these design principles to create the concrete syntax of the proposed UML extension to control the notations overload in the current bioinformatics workflows.

D. DSL vs UML

A new DSL would require the creation and specification of all its concepts as well as notations, while a UML extension would require only detailed information about the new additions. Moreover, the usage of a well-known language, such as UML, increases the number of programs to model diagrams. These reasons were recognised by the researchers for selecting UML extension over DSL.

Fitting a language to a specific use or domain can be done either by developing a new DSL or tailoring a general-purpose modelling language, such as UML. According to Gray and Rumpe, the decision on which one of these paths to select is an interesting point [14]. UML profiles can be used either for defining DSL or tailoring it to a domain viewpoint. Selic [28] provided a guide to establish a systematic UML profile that consists of three main methods: 1) defining the domain meta-model through the use of OMG Meta-Object Facility (MOF) language [13], in other words creating a stereotype, for adding new model elements [29]; 2) mapping the domain model to a profile [13], tagged values, in order to add new features and extend characteristics [29] to a suitable UML base-concept; 3) formally specifying the semantic constraint restrictions [13], [29]. In this thesis, the researchers decided to tailor a UML AD to fit the bioinformatics domain, using methods 1, 2 and 3 to provide a valid UML extension with the reported gaps in [5].

E. UML Usage Problems

Gray and Rumpe listed problems when domain experts from a mechanical engineering company used "explicit modelling languages", where the produced models contained several issues because the participants lack fundamentals in SE [14]. Selic's work in [28] aligns with [14] since both stated that domain expertise familiarity with the UML meta-model is important when establishing a profile. Some of the problems found in the models were their high-level depiction or information overload; the difficulties to have an overview; the notations misusage, leading to self-interpretation models; that notes represented the most important information, lacking behavioural information; its focus on a specific situation absenting reusability [14]. Additionally, the authors in [14] stated that domain experts usually had done modelling as a tiny part of their routines. The work in [5] identified some of these issues in the bioinformatics field. Therefore, this study will provide a solution that might help to mitigate most of these problems by introducing artefacts that contain a standardised modelling language for the bioinformatics domain. However, to be able to use this solution, the bioinformaticians will be required to learn the language.

F. UML Extensions Found in the Literature

The literature covers several attempts to extend the UML AD meta-model. One of the efforts found was created to ease project management, for which the extended meta-model contains artefacts specifications associated with the project management tasks for allocating resources, estimating the costs, and planning [30]. Another attempt was a profile establishment for representing different levels of an activity execution for business process and enterprise activity, the function entities needed for the executions, and the relationship between an activity and its associated software application [22]. In addition to this profile, Ricardo and Duncan covered the representation of possible transition paths alternatives of the process flow, such as iteration, single thread, or-join, etc. [22]. Further, Störrle in [31] proposed the arrow representation of a 'LoopNode' based on coloured Petri nets, one of the mathematical modelling languages, since it is not provided in the standard UML meta-model.

The authors in [32] used the UML AD to capture the action of entities populations, by creating diagrammatic notations for arrows to represent their three relationships, propagation, interruptible, and contributory, without extending the UML AD. The work in [20] proposed a UML extension to capture context-awareness requirements of context-aware systems, where the authors provided new concrete syntax for context objects, context constraints, meta-swimlane separation, and adaptation activities. These concrete syntax differentiate between the system objects and decision making. Further, the authors in [24] described their UML AD profile, which is used to represent a business process by introducing several concepts to specify the process relationships, such as, data repositories, data objects, and presentation objects.

None of these UML extensions solves completely the problem stated in this research neither covers any specifications for the bioinformatics domain, where the authors study specific topics like biology, business process or project management and their needs. However, some of these concepts and notations are useful and align with this research modelling-extensions necessities. Therefore, the study will use them precisely or as inspiration, as explained in the design and development step along the results section in this thesis.

III. Methodology

A. Facilities Description

The research was conducted with participants from three different facilities. The first is the Bioinformatics Core Facility, a bioinformaticians and statisticians consultation and data analysis agency, which makes part of the Sahlgrenska Academy Core Facilities at the University of Gothenburg. The second is the Genomic Medicine Sweden (GMS), a collaborative initiative, which focuses on improving healthcare, innovation, and collaboration in Sweden. The third is the Translational Genomics Platform, a research platform created between Wallenberg Centre for Molecular and Translational Medicine and Västra Götalandsregionen, that aims to bring innovation into the healthcare system.

B. Selection of Participants

The head of the Bioinformatics Core Facility used purposive sampling technique to select the participants for this research. This sampling technique aims to diminish the accidental sampling bias [33] since the participants' selection is based on the researchers' belief [34] that they fulfil stipulated criteria [35], [36]. The facility head chose bioinformaticians with workflows knowledge, which were the criteria for this study. The reason for adopting this method is to reach a meaningful result by including participants who can provide information that others cannot [34], [35]. The seven participants of this research were identified by the letter 'P' with a random number from 1 to 7, being consistent between the iterations.

C. Research Approach

This thesis uses the Design Science Research Methodology (DSRM) due to its pragmatic nature [37] and strength to solve real-world issues [38]. The DSRM aims to reach the Pasteur quadrant [39], being the joint between basic and applied research. The former relates to studies that search for fundamental knowledge, while the latter aims to solve problems. Therefore, DSRM contributes to the applied domain with rigour [37], depending on how it is executed. Peffers et al. proposed a procedure to perform a DSRM in [40], which was adapted to the needs of this research. Their method had six steps: Identify Problem and Motivate, Define Objectives and Solutions, Design and Development, Demonstration, Evaluation, and Communication [40]. The alteration was done in the purpose and outcome of the Identify Problem and Motivate step since the problems were identified by Horkoff et al. in [5]. However, the acknowledgement of the problems was iteratively expanded and refined throughout the study.

Figure 1 portrays the performed steps on this thesis and its three iterations. Based on [5], three artefacts were created, the UML AD meta-model extension, its corresponding concrete syntax, and the WDST; they were used, improved, and validated along the process.

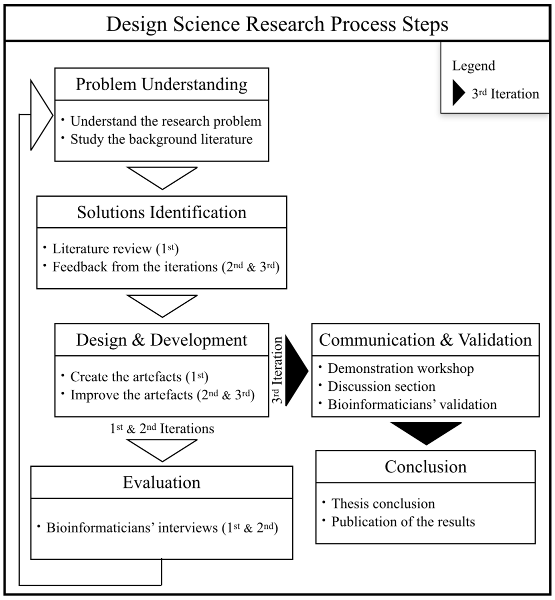

Design Science Research Process Steps
Legend
3rd Iteration
Problem Understanding
• Understand the research problem
• Study the background literature
Solutions Identification
• Literature review (1st)
• Feedback from the iterations (2nd & 3rd)
Design & Development
• Create the artefacts (1st)
• Improve the artefacts (2nd & 3rd)
3rd Iteration
Communication & Validation
• Demonstration workshop
• Discussion section
• Bioinformaticians' validation
1st & 2nd Iterations
Evaluation
• Bioinformaticians' interviews (1st & 2nd)
Conclusion
• Thesis conclusion
• Publication of the results

Figure 1. The DSRM process followed in this research.

Each one of the steps shown in Figure 1 is described below:

1) Problem Understanding: This step was created to understand the identified problem and findings from [5]. It was a step for the researchers to get acquainted with the problem and learn iteratively from the participants' feedback.

2) Solutions Identification: In this step, the researchers acknowledged the problem and studied it to propose a solution. For the first round, this step was done based on [5] and literature review, while for the next two iterations, the received feedback during the interviews guided the identification of the solution.

3) Design and Development: During this part of the process, the artefacts were improved by the researchers based on the identified solutions from the previous step.

4) Evaluation: During this step, the interaction with Bioinformaticians was through semi-structured interviews and a usability test, where the domain experts assessed the understandability, completion, and fulfilment of the artefacts. This step objective was to receive support from the domain-experts to develop and improve the artefacts.

5) Communication and Validation: This part of the process was performed with a focus group, during a workshop, where all the participants from the first two iterations and one more were invited. The goal was to show, explain, and demonstrate the usage of the final artefact.

6) Conclusion: In this step, this research reaches its end with a validated solution to the problems in the bioinformaticians' workflow design procedure, and the publication of the results was done.

The three iterations are described on the following subsections, each one containing the specific methodology used and the data analysis procedure.

D. First Iteration

During this iteration, the researchers understood the identified problems in [5], studied their background, created the three artefacts, and evaluated them with the bioinformaticians.

1) Artefacts Creation: The three artefacts creation, UML AD meta-model extension, concrete syntax, and WDST, started by incorporating the concepts as suggested in [5]. The standard UML AD [13], UML ADs extended stereotypes profiles in [20], [22], [24], [30], [31], the usage of the goal concept from i* visual notations in [23], [26], and the researchers' creativity were the foundation for creating these artefacts. The new notations were invented or reused to align with the bioinformaticians' needs for the missing graphical concepts representations. For each concept, in this iteration, two notations were provided to the domain experts aiming a selection based on familiarity or understandability against a normative approach. Moreover, the WDST followed a basic content structure to a domain-specific requirements artefact containing the extended meta-model concepts, being its written version.

2) Interview Design: A semi-structured interview was conducted with five bioinformaticians, allowing active participation to obtain nuanced, deeply contextualised, and authentic answers [41] to gather qualitative data. Additionally, the researchers can clarify questions to eliminate duality and inconsistency by formulating freely new questions [42]. Appendix A contains the followed predefined questionnaire, which was ordered from general to elaborated questions [43].

The interviews were time-boxed to 30 minutes to avoid participant fatigue, and the anonymity of the participants' ideas and answers were assured. Furthermore, the interviews were hosted at the laboratory's facility and recorded upon interviewees' agreement to allow a natural flow of the conversation.

During the interviews, the created WDST was explained to the participants and shown in a paper format for 5 minutes, see Appendix B. After that, the written template predefined questions were asked for around 8 minutes. Following that, the concrete syntax, in Appendix C, was given to the participants in a paper as a diagram legend that contains the concepts and the shapes, see Appendix D. They glanced at the concepts while being asked their associated questions, for 5 minutes. Afterwards, 10 minutes was used to show the two examples of diagrams to the participants, so they could select one of the two notations, explaining their reasoning, see Appendix E for the examples.

3) Data Analysis Procedure: The recorded data from the interviews were transcribed using Temi[1]. After the transcript was ready for analysis, the thematic analysis method was used to identify patterns of meanings that are significant [44] and group them into themes [45]. Subsequently, one researcher used the transcribed data to create a codebook containing the codes, definitions, and the participants' statements. Afterwards, its content was verified by the other researcher, for reliability purpose [44]. The theme for the codes originated inductively from the gathered data [46]. The suggestions and problems were approached during the *Solutions Identification* and *Design and Development* steps for the second iteration.

E. Second Iteration

The raised issues and suggestions from the first iteration were used together with the literature to find enhanced solutions that lead the researchers to improve the artefacts to be tested and evaluated in a second evaluation.

1) Artefacts Improvement: The artefacts were improved by deleting, adding or modifying the areas mentioned by the participants during the first round of interviews.

2) Interview Design: Five bioinformaticians were invited to be part of this round of the study, where only one of them did not participate in the previous iteration. The tasks required the usage of the participants' computers since familiarity with the keyboard, language, and other settings could speed up the process and reduce the annoyance.

The interview was time-boxed to 1 hour and composed of six sections: section 1, the participants were asked two questions about their understanding of workflow and step definitions for 5 minutes, see Appendix F for this interview predefined questions. In section 2, the participants were requested to use the notations library to draw a bioinformatics workflow for 15 minutes, provided as an XML file for draw.io[2], see Appendix G for the notations legend. Additionally, the updated DNA sequencing workflow example was given to them, see Appendix I. In section 3, the participants answered the semi-structured interview questions about the modelling language usability, which lasted 10 minutes and was inspired by the System Usability Scale (SUS). The SUS questionnaire is one of the most widely used [47] because it is a simple ten-item survey to assess usability [48] and learnability [49]. However, SUS itself does not identify the usability flaws of the system [49]; therefore, it was used as an inspiration to create the open-ended interview questions, allowing further questioning. Section 4 was a WDST filling test, where each participant was asked to fill, for 15 minutes, the Google sheets[3]

[1] https://www.temi.com/

[2] https://www.draw.io/

[3] https://www.google.com/sheets

template with a workflow description and one of its steps, see Appendix J. In section 5, the participants answered the semi-structured interview for 10 minutes, which was inspired again by SUS. Section 6 had three questions related to the usage and impression of the artefact, taking around 5 minutes.

The reason for leaving the participants select the workflow scenario, in the interview section 2 and 4, was that the familiarity could decrease the spent time and different workflows could identify more problems or gaps in the artefacts. During these two sections, the participants were asked to follow the think-aloud protocol, which is used in many fields to collect insightful data, while participants perform tasks [50]. This protocol enriches the data by observing the users' behaviour with the possibility to walk-through their mind [51] with a drawback that most people do not work naturally while explaining what they are doing [52]. The researcher responsible for the log of the interview used a template to document these observations, see Appendix K. All the precautions were taken to provide a comfortable and unbiased environment for the participants. The meetings were recorded upon interviewees' agreement.

3) Data Analysis Procedure: This iteration used the same data analysis methods as the first round. This study did not utilise the produced diagrams and documents as data inputs, due to the participants' statement regarding their inaccurate and incomplete content; however, the observation and think aloud methods collected valuable insights of the bioinformaticians' struggles with the templates and concepts. Further, the purpose of letting the experts try the artefacts was to get more in-depth feedback, without requiring grammar and graphical precision. The collected data were used to improve the artefacts during the *Solutions Identification* and *Design and Development* steps in the third iteration.

F. Third Iteration

The feedback from the second iteration guided to improve the artefacts during this final iteration, resulting in an updated version to be validated at the end.

1) Artefacts Improvement: A very similar procedure from the second iteration was performed, differing by the existence of only one notation for each concept. The third iteration artefacts are the end-result of the second iteration interviews.

2) Workshop Design: All participants from previous iterations, six, plus the head of the Bioinformatics Core Facility were invited to the one-hour time-boxed workshop. Moreover, workshops are described as an event in which a group of people solve domain-specific problems creatively or innovatively [47], through observations and interactions [48]. Thus, the goal of this workshop was to evaluate and validate the artefacts through collaborative groups' discussions. The workshop was recorded upon the participants' unanimous approval.

The workshop was divided into nine sections, which lasted accordingly to the time in Figure 2. During these sections, the performed activities were: 1) the workshop agenda was explained; 2) the notations and concepts were displayed and described through examples, see Appendix L; 3) the participants were paired to the closest person to discuss the usability and understandability of the notations and concepts; 4) each pair exposed their thoughts; 5) the participants individually and anonymously validated the notations and concepts using Mentimeter[4]; 6) WDST was displayed and described in details; 7) the pairs discussed the usability and understandability of the WDST; 8) each pair introduced their opinions; 9) the participants were asked to validate the WDST using Mentimeter, which is a user-friendly platform that allows participants to engage anonymously to presentations via the internet by using any device. It provides the results in a graphical and or text format, depending on the type of question. Appendix M contains the asked questions during this workshop discussion and validation sections.

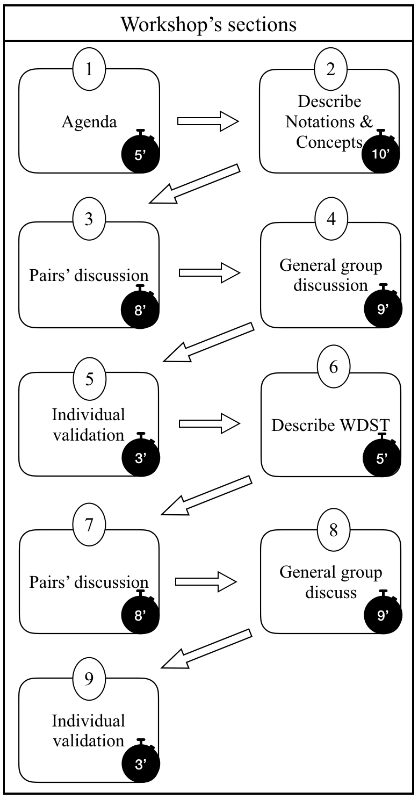


Figure 2. The workshop structure and the planned duration for each section, in minutes.

3) Data Analysis Procedure: The gathered data were transcribed and then grouped into a codebook, using the thematic analysis method. Additionally, the results from the Likert scale evaluation were presented as graphs with the average. However, the artefacts will not be further refined; instead, the changes will be suggested for future work.

IV. RESULTS

A. Final Artefacts

The final version of the developed artefacts along this research is a UML AD meta-model extension for bioinformatics domain, see Figure 3. Subsequently, Table I shows the final concrete syntax based on the UML AD extension. Lastly, Figure 4 depicts the final version of the WDST.

[4] https://www.mentimeter.com

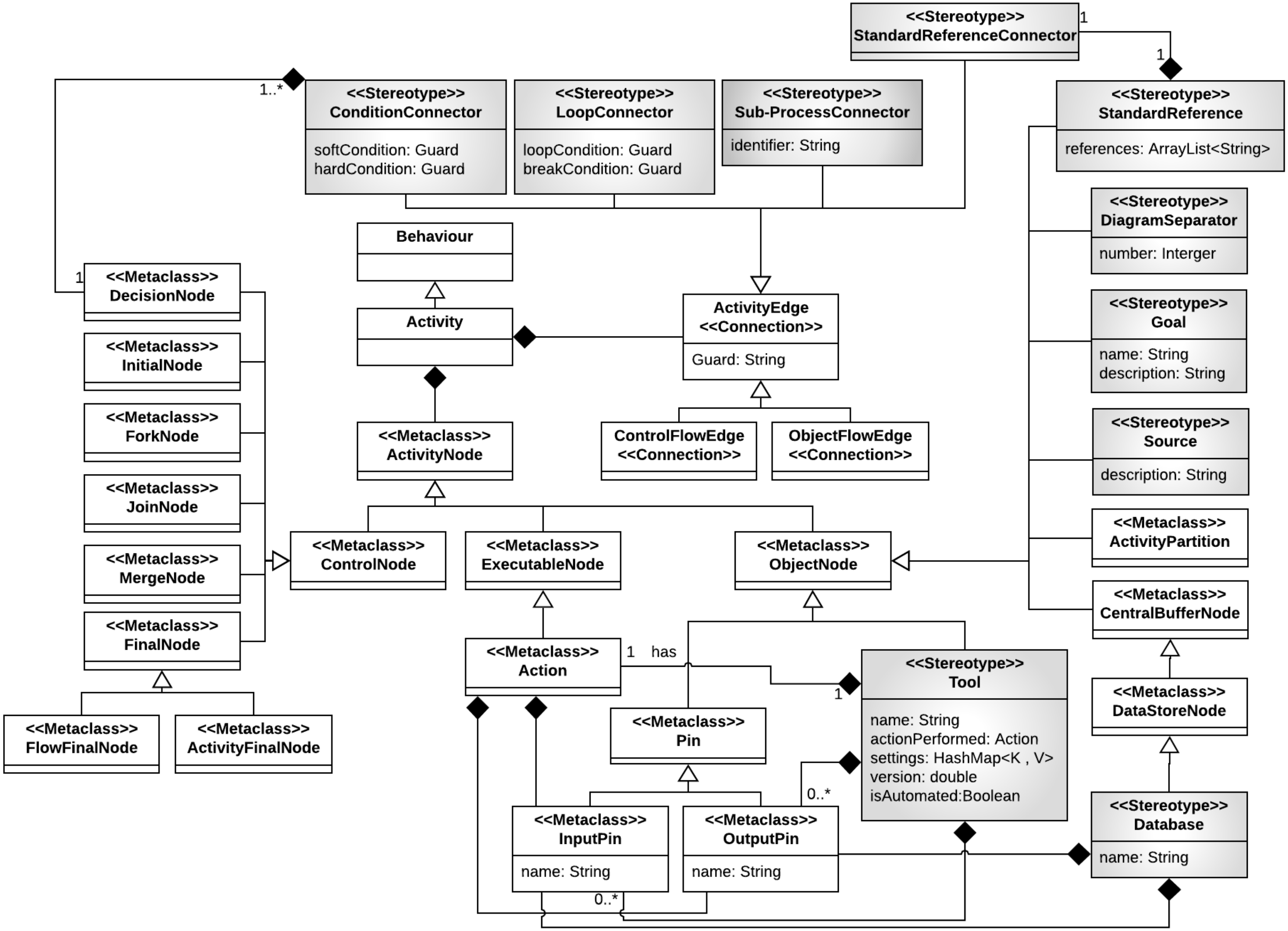


Figure 3. The final version of the extended UML AD meta-model (white classes are from UML AD [13], and grey ones were extended in this work).

Table I
THE FINAL VERSION OF THE CONCRETE SYNTAX EXTENSION

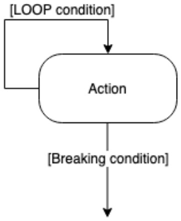
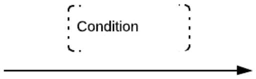
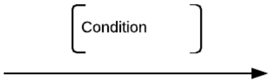
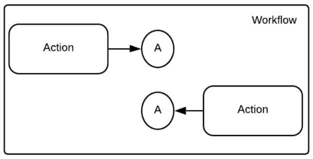
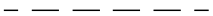
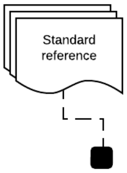
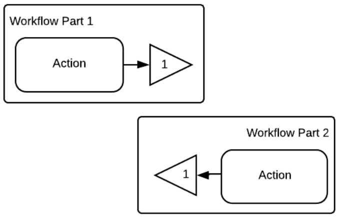
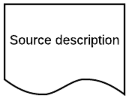
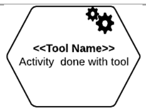
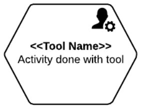
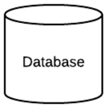
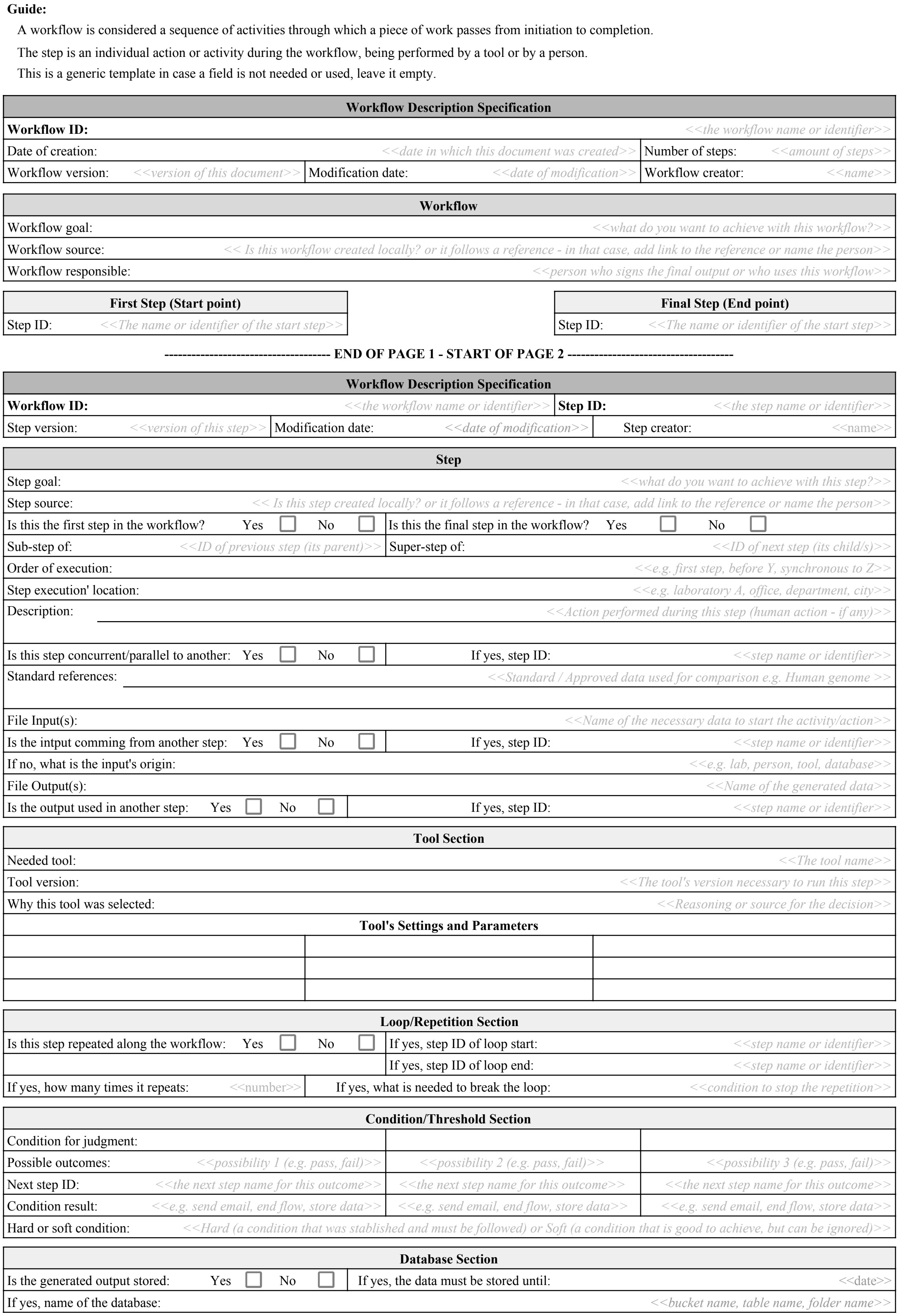
Name	Base Class	Description	Notation
Loop	ActivityEdge	An iterative set of activities and actions until reaching the defined condition.	
SoftCondition	ActivityEdge	A condition with a limited soft-condition value, which is used for test outcomes. The condition is predefined within dashed guards on the outgoing edges.	
HardCondition	ActivityEdge	A condition with a limited hard-condition value, which is used for test outcomes. The condition is predefined within solid guards on the outgoing edges.	
Sub-processConnector	ActivityEdge	A connector between the sub-processes parts within the same diagram.	
StandardReferenceConnector	Activity Edge	A connector between the dark input and the standard reference notation (multiple documents).	
StandardReference	ObjectNode	Data, usually a standard, that are used for comparisons, such as the human genome.	
DiagramSeparator	ObjectNode	A labelled triangle that represents the connection point with another part of the diagram from another page.	 Workflow Part 1 Action 1 Workflow Part 2 1 Action
Source	ObjectNode	A link, document title, or person's name, which is the source for a specific set of actions.	
Tool	ObjectNode	An automatically operated tool or software used to perform an activity with its description.	
	ObjectNode	A manually operated tool or software used to perform an activity with its description.	
Database	DataStoreNode	A structured set of data that is accessible in various ways.	

Figure 4. The final version of the Workflow Description Specification Template - WDST.



Guide:

A workflow is considered a sequence of activities through which a piece of work passes from initiation to completion.

The step is an individual action or activity during the workflow, being performed by a tool or by a person.

This is a generic template in case a field is not needed or used, leave it empty.

Workflow Description Specification					
Workflow ID:	<<the workflow name or identifier>>				
Date of creation:	<<date in which this document was created>>			Number of steps:	<<amount of steps>>
Workflow version:	<<version of this document>>	Modification date:	<<date of modification>>	Workflow creator:	<<name>>

Workflow	
Workflow goal:	<<what do you want to achieve with this workflow?>>
Workflow source:	<< Is this workflow created locally? or it follows a reference - in that case, add link to the reference or name the person>>
Workflow responsible:	<<person who signs the final output or who uses this workflow>>

First Step (Start point)	
Step ID:	<<The name or identifier of the start step>>

Final Step (End point)	
Step ID:	<<The name or identifier of the start step>>

------------------------------------- END OF PAGE 1 - START OF PAGE 2 -------------------------------------

Workflow Description Specification					
Workflow ID:	<<the workflow name or identifier>>			**Step ID:**	<<the step name or identifier>>
Step version:	<<version of this step>>	Modification date:	<<date of modification>>	Step creator:	<<name>>

Step									
Step goal:	<<what do you want to achieve with this step?>>								
Step source:	<< Is this step created locally? or it follows a reference - in that case, add link to the reference or name the person>>								
Is this the first step in the workflow?	Yes	□	No	□	Is this the final step in the workflow?	Yes	□	No	□
Sub-step of:	<<ID of previous step (its parent)>>				Super-step of:	<<ID of next step (its child/s)>>			
Order of execution:	<<e.g. first step, before Y, synchronous to Z>>								
Step execution' location:	<<e.g. laboratory A, office, department, city>>								
Description:	<<Action performed during this step (human action - if any)>>								
Is this step concurrent/parallel to another:	Yes	□	No	□	If yes, step ID:	<<step name or identifier>>			
Standard references:	<<Standard / Approved data used for comparison e.g. Human genome >>								
File Input(s):	<<Name of the necessary data to start the activity/action>>								
Is the intput comming from another step:	Yes	□	No	□	If yes, step ID:	<<step name or identifier>>			
If no, what is the input's origin:	<<e.g. lab, person, tool, database>>								
File Output(s):	<<Name of the generated data>>								
Is the output used in another step:	Yes	□	No	□	If yes, step ID:	<<step name or identifier>>			

Tool Section		
Needed tool:	<<The tool name>>	
Tool version:	<<The tool's version necessary to run this step>>	
Why this tool was selected:	<<Reasoning or source for the decision>>	
Tool's Settings and Parameters		

Loop/Repetition Section						
Is this step repeated along the workflow:	Yes	□	No	□	If yes, step ID of loop start:	<<step name or identifier>>
					If yes, step ID of loop end:	<<step name or identifier>>
If yes, how many times it repeats:	<<number>>				If yes, what is needed to break the loop:	<<condition to stop the repetition>>

Condition/Threshold Section			
Condition for judgment:			
Possible outcomes:	<<possibility 1 (e.g. pass, fail)>>	<<possibility 2 (e.g. pass, fail)>>	<<possibility 3 (e.g. pass, fail)>>
Next step ID:	<<the next step name for this outcome>>	<<the next step name for this outcome>>	<<the next step name for this outcome>>
Condition result:	<<e.g. send email, end flow, store data>>	<<e.g. send email, end flow, store data>>	<<e.g. send email, end flow, store data>>
Hard or soft condition:	<<Hard (a condition that was stablished and must be followed) or Soft (a condition that is good to achieve, but can be ignored)>>		

Database Section						
Is the generated output stored:	Yes	□	No	□	If yes, the data must be stored until:	<<date>>
If yes, name of the database:	<<bucket name, table name, folder name>>					

B. First Iteration

Figure 1, in the methodology section, depicts the iterations steps followed during this thesis.

1) Solutions Identification: Horkoff et al. in [5] identified *thresholds*, *source*, differentiation of *files*, *goals*, *sub-process*, and *repeated iterations* as needed by bioinformaticians while creating their workflows. Consequently, these concepts were included in the UML AD extension since [5] was the only found work for the bioinformatics domain that explicitly mentioned bioinformatics concepts. Moreover, the provided draft in [5], for the bioinformatics workflow elicitation, was used as an initial content and extended correspondingly to the modelling language extensions to maintain consistency between the artefacts.

2) Design and Development: The researchers avoided using different colours or texture to define the visual syntax of the concepts to follow the UML AD patterns, and to provide an inclusive language that can be used by colour blind people or any person with visual disabilities.

Based on the nature of UML profile, all the UML AD syntax and semantics were kept, which are: *action*, *decision*, *merge*, *forks*, *join*, *initial node*, *flow final*, and *activity final* [13]. Additionally, the *activity edge connector* from the UML AD [13] maintained the same syntax with an additional utilisation to represent sub-processes for the bioinformatics domain. The same was done for *swimlanes*, which had its usage based on [53].

a) ***UML AD meta-model extension****:* The UML AD meta-model in [13] was used as a starting point for this extension, where the original meta-classes are represented with a white colour, while the modified or extended profiles are visualized with a grey colour, as shown in Appendix N. See Table II for a summary of the implemented extensions and Table III for the reasoning and source of each concept depending on its concrete syntax.

Table II
FIRST ITERATION UML AD META-MODEL EXTENSION SUMMARY

Concept	Extension
Tool, *DiagramSeparator*, *Source* & *Goal*	Added as stereotypes of the inheriting classifier *ObjectNode*
Tool	Added a composition relationship with the metaclasses *Action*, *InputPin* & *OutputPin*
Datastore	Had been added as a stereotype due to some changes on the notation
LoopConnector	Inherited from the super-class *ActivityEdge*, containing *loopCondition* & *breakCondition* guards
ThresholdConnector	Inherited from the super-class *ActivityEdge*, containing the specified guards *softThreshold* & *hardThreshold*
DecisionNode	Composites at least one *ThresholdConnector*

b) ***Concrete syntax****:* The design decisions for both concrete syntax, 1 and 2, considered the principles for cognitive effectiveness of the visual notations, which are symbols deficit, redundancy, overload, and excess. These principles ensure the correspondence between semantics and graphical shapes of notations [23]. Table III lists the extended concepts and the sources of the two concrete syntax with their explanations. Appendix C contains the concrete syntax notations, their names, base classes and definitions.

c) ***WDST****:* Its purpose is to help bioinformaticians during the workflow elicitation process and keep the information documented. Therefore, the researchers added the basics of documentation traceability in the WDST, such as *workflow* and *step ID*, *name*, *creator*, *version number*, and *date of creation*. Firstly, the workflow information was separated from the step because a workflow might contain several steps, allowing steps multiplication without details repetition. After that, the missing concepts, identified in [5], were added as fields and sections to the WDST to maintain the correlation between the artefacts, see Appendix B for its first version.

3) Evaluation: Five bioinformaticians from the three facilities were interviewed on March 27th. The time limit was exceeded in about 20 minutes in the first interview and 10 minutes in the second, due to the researchers' inexperience and participant's long answers respectively. A further interview was very fast, taking around 16 minutes. The recordings were transcribed, and a thematic analysis was done to create a codebook, which contains 13 codes, see Appendix O.

a) ***WDST****:* This subsection contains an overview of the participants' evaluation of the WDST, which was collected during the first iteration, see Table IV.

b) ***Concrete Syntax****:* This subsection consists of the participants' evaluation of the concrete syntax from the first iteration, see Table V for an overview of the collected data and findings.

The participants were asked about their ***preferences*** related to the two provided notations for each concept. The selected notations are in Table VI, based on the highest number of answers.

c) ***WDST and concrete syntax****:* Related to the *order* of ***artefacts usage***, all participants said that they would draw the workflow first and then fill the WDST. However, P2 stated that "I think like there's so much here (WDST) that's, that would be redundant when you're using this (both artefacts)".

C. Second Iteration

The second iteration started with identifying solutions to the issues raised by the participants, leading to the creation of the artefacts in the design and development step. The artefacts were re-evaluated by the participants, producing new data that were analysed. On the methodology section, Figure 1 depicts these steps.

1) Identify Solutions: The solutions to the participants' issues and missing points in the artefacts were, in most cases, provided by the participants during the interviews. Concerning the concrete syntax, one notation was selected by the participants. Meanwhile, *standard references* was mentioned as the only missing concept by two participants. The WDST had several missing fields, such as guidance for the template usage; the required input and output data for each tool; parallelise

Table III
THE NEW META-MODEL CONCEPTS AND THE CREATED CONCRETE SYNTAX IN THE FIRST ITERATION

Meta-model concepts	Concrete syntax 1 source	Explanation	Concrete syntax 2 source	Explanation
Tool	Flowchart notations & i* visual syntax [23]	Hexagon notation provides visual differentiation since UML AD lacks it. Tools perform a task, linked to i* visual syntax concept.	Flowchart notations and i* visual syntax [23]	Identical to the concrete syntax 1, with an additional gear icon on its corner to allow a faster visualisation of the tools.
DiagramSeparator	UML AD [13]	The semantic and syntax are inspired by *ActivityEdgeConnector* with a graphical modification, a triangle with a number instead of circles with letters.	UML AD [13] & from the loop notation in [30]	The same inspiration as concrete syntax 1, where a rectangle with a letter inside and fence around the compressed diagram part.
Source	Flowchart notations & i* visual syntax [23]	Concept identical to *Resource* in i*, using the document notation from the flowchart notations.	Flowchart notations & i* visual syntax [23]	Identical to the *source* in concrete syntax 1.
Goal	i* visual syntax [23]	Identical to *Goal* in i*.	i* visual syntax [23]	Identical to the *goal* in concrete syntax 1.
Datastore	UML AD [13]	Follow exactly the standard notations and usage in UML AD.	UML AD extensions in [24]	Concept identical to UML datastore, but with the flowchart cylinder shape, database notation.
InputPin & OutputPin	UML AD [13]	Follow exactly the standard notations and usage in UML AD.	UML AD [13] and flowchart notations	Concept identical to stand alone pin from UML, but using the parallelogram shape from flowchart.
SoftThreshold	UML AD [13] & i* visual syntax [23]	Visual syntax was a graphically encoded connector with 2 dashed-lines and the semantic based on standard UML guards.	UML AD [13] and different line styles from [20]	Follow exactly the standard UML AD semantics and usage, where the guards syntax were modified to dashed lines.
HardThreshold	UML AD [13] & i* visual syntax [23]	Visual syntax was a graphically encoded connector with 2 solid-lines, and the semantic based on standard UML guards.	UML AD [13] and different line styles from [20]	Follow notation and concept of the guards in the standard UML AD.
LoopConnector	UML sequence diagram [13]	Identical to the loop semantics and syntax in UML sequence diagram.	UML structured nodes [31]	Follow the loop semantics and syntax suggested for UML, where using arrows with guards lead to the activity repetition.

Table IV
THE FINDINGS FROM THE FIRST ITERATION WDST EVALUATION

Code	Findings with Illustrative Statements
WDST improvements	*Field deletion*: P2 said "step responsible and who, who conducts the step and where does that happens? I feel like a lot of the times it's going to be the same".
	Understandability: where the participants asked for more explanation, as P4 stated "what do you mean with threshold here?"
Missing in the WDST	*Tool settings and parameters*: four participants mentioned WDST importance for knowledge sharing, where P4 said "we used this first X tool kit with the parameters this, this, this".
WDST usage	*Knowledge sharing*, *structuralisation* of thoughts, the process *formalisation*, and for hospitals *system documentation*. However, a concern about WDST was raised that was shared by the other participants, P1 said "I mean we have to write a bunch of stuff that I don't think anyone ever reads it. It just needs to be there in case of someone needing to read it".
Test of the WDST	Two of five participants stated that they would provide better feedback on missing, understandability and usability, if they could try to fill it.
WDST users	*Stakeholders* involved in the process.
WDST current state	Bioinformaticians write *free text*s without any standards, which they believe are understandable to everyone.

Table V
THE FINDINGS FROM THE FIRST ITERATION CONCRETE SYNTAX EVALUATION

Code	Findings with Illustrative Statements.
Notations & Concepts improvement	Two issues mentioned for *understandability* relating to the use of *swimlanes* and *loop*s inclusion and exclusion factors.
Missing Notations & Concepts	Only an *addition* was explained by two participants, one of them, P2 said "I don't know if there's some workflows have a ton of like references it could be like 15 or something; like data inputs, it could be like the human genome or, and some like database software".
Notations & Concepts usage	*System documentation*, as P2 mentioned "if we have to document our workflows for like the hospital to put it into there like documents system. Then we have to design these things".
	Thoughts *structuralisation* and process overview were cited by three participants.
Diagram users	*Bioinformaticians* and *stakeholders* were mentioned by the participants.
Notations & Concepts current state	Depends on workflow creators' own way of drawing, as described by P1 "I'm usually just drawing like each program has a box and then an arrow and then the name like file on the arrow".

steps; and information corresponding to the tools, like version and settings. The solution for the missing points was to add fields holding the required information. The requested im-

Table VI
THE PARTICIPANTS' CONCRETE SYNTAX SELECTIONS IN THE FIRST ITERATION

Notation	№of answers	Selection	Reasoning
Loop	4/5	2a	P3 said "You can actually see and follow, where it breaks and where it starts again, where the loop goes".
Threshold	3/5	2a	P1 selected an option, but stated "I don't mind either way".
Input & Output pins	3/5	1b	P2 explained that "it makes it less cluttery".
Datastore	5/5	2b	The participants' familiarity with the notation.
Tools	4/5	2c	P5 referred to the two gears icon as "it's quickly seen".
Diagram Separators	3/5	1a	P4 said "it's like a different symbol than the other ones, so it's clear".

provements were regarding the repetition of information; thus, fields such as the step responsible, where the step happens, and the several fields of tools were removed during the step design and development. In addition to these improvements, the used word 'thresholds' in the WDST required more in-depth explanation; therefore, a more comprehensible synonym replaced it.

2) Design and Development:

a) ***UML AD meta-model extension****:* Table VII contains the meta-model changes with the names of its classes, their types and attributes, and the relationships and modifications. The updated version of the meta-model was produced by including the participants' suggested changes in the evaluation during the first iteration, see Appendix P.

Table VII
THE ENHANCEMENTS FOR THE UML AD META-MODEL DURING THE SECOND ITERATION

Meta-model Classes	Class Type	Added Attributes	Relationships & Modifications
Tool	stereotype	tool version & settings	
StandardReference	stereotype	list of references	Inheriting classifier of *ObjectNode*
StandardReference-Connector	stereotype		Inheriting classifier of *ActivityEdge*
Threshold	stereotype	softCondition & hardCondition	Changed to *ConditionConnector*
ActivityEdge-Connector	meta-class		Changed to *SubprocessConnector* as a *Stereotype*
Datastore	meta-class		Changed to a *Database Stereotype*

b) ***Concrete syntax****:* Its unique modification was the addition of a sub-concept for input, the *standard references*. Its notation originates in a black input pin with a dashed connector, inspired by the note connector representation from UML [13]. At the end of this connector, a notation that represents multiple documents were used with the possibility to expand to write several references, which was inspired by a figure in [23]. Appendix H consists of the concrete syntax of this iteration.

c) ***WDST****:* Its applied changes are listed on Table VIII, which had its usage changed, based on the provided feedback, from a helper during the workflow elicitation process to a standardised way to document workflows for stakeholders and share knowledge. Accordingly, its name was changed from workflow requirement specification document to WDST. Appendix J contains the WDST with all these modifications.

Table VIII
WDST CHANGES IN THE SECOND ITERATION

Concepts	Change	Reasoning
Step responsible	Removed	Based on participants' feedback
Two out of three *tool* fields	Removed	To decrease redundancy
Threshold	Nomenclature changed to *condition*	To increase familiarity
Initial step	Nomenclature changed to *first step(start point)*	To increase familiarity
Step description	Added	Based on participants' feedback
Concurrent step option	Added	Based on participants' feedback
Tool settings and parameters	Added	Based on participants' feedback
Output rows	Moved to the *step* section instead of the *tool* subsection	Steps can perform and produce outputs without tools involvement

3) Evaluation: The five bioinformaticians were interviewed on April 27th; all of them agreed to be recorded. These recordings were transcribed, followed by the codebook creation, which contains the 19 codes with its subcodes and explanations. Seven of these codes were identical to the previous iteration, see Appendix Q for the codebook.

Related to ***workflow definition***, P3 and P4 said that it is a conversion process from an input to an output passing through steps. P6 defined it as the project process, while P1 described it as an overview of what is running in which order and P5 said that it is an overview of how to run the program.

P3 said the different used tools are the ***step definition***, while P5 was not sure, but noted that "input, output or tool" would be the steps. P4 stated that a step is "something that takes some files or something as an input and produce something as an output". P6 mentioned that it is an action performed at a certain point, and the P1 said that it is an action but involving a "file changing shape, or being transferred to another computer."

a) ***Concrete syntax****:* This subsection contains an overview of the findings based on the participants' evaluation of the concrete syntax, which was collected during the second iteration, see Table IX.

Table IX
CONCRETE SYNTAX EVALUATION FINDINGS IN THE SECOND ITERATION

Code	Findings with Illustrative Statements
Notations & Concepts improvement	*Dislike*: P5 described the pins location on the tool shape that caused a diagonal gradient in the diagrammatic flow.
	Understandability: four participants misunderstood the difference between the concepts *action* and *tool*, where P1 asked "what's the difference between tool and action?".
Unnecessary notations	*Goal* notation was mentioned by two participants as *unneeded*, while two others pointed *vertical* and *horizontal join/fork* as *unfamiliar*, and P1 stated that "I would probably just do many arrows pointing to one tool or something like that".
Missing Notations & Concepts	Two participants mentioned *additions*; P5 stated the missing parallelogram shape of the pins and P1 mentioned "there's no file database with a box" and "I would like different kind of arrows"
Notations & Concepts usage	Four participants stated that the provided library would definitely be used. However, P5 stated that they "usually don't write the workflows. I mean if we need to, we do it for publications, but usually, it's just text". Additionally, P6 said "the work we do, it is quite standard, so we have kind of the workflow in our mind", thus "we don't use it that often". In opposition, P3 said "maybe" for its usage because "it takes time to do it".
Notations complexity	Four participants stated that the graphical shapes are not complex. In contrast, P6 stated that the notations are "a little bit" complex because of the number of shapes, but "if somebody learns this quite well, I would say it's quite straightforward".
Notations & Concepts tutorial necessity	P3 and P5 said that a descriptive *manual* would be enough, while P1 and P6 stated that *training* is a necessity. Contrarily, P4 said "I will learn just by using it".
Confidence to use the Notations & Concepts	Was felt by four participants. However, P1, P3 and P5 stated that it was challenging to use draw.io as a modelling tool.

b) ***WDST:*** This subsection has the findings and participants' evaluation to the WDST, during the second iteration, see Table X.

c) ***WDST and concrete syntax:*** The ***artefacts usage*** was related to process *overview, traceability and learnability, publication*, and *validation*. According to the participants, the ***artefacts users*** are *bioinformaticians, researchers*, and *tools developers*. Moreover, the participants' answers regarding the ***artefacts general impression*** were that the diagram is good, useful, and provides a clear overview while the WDST requires time and holds a lot of information. In addition to that, P4 stated that both artefacts "complement each other."

d) ***Observations:*** Table XI contains the observations collected during the tests of the artefacts, while participants followed the think-aloud protocol. The data in this table originated from using the log keeper's template, where six points were tracked and described in the log categories column. However, none of the participants had failed a task; therefore, it was recorded but unreported.

Table X
WDST EVALUATION FINDINGS IN THE SECOND ITERATION

Code	Findings with Illustrative Statements
WDST improvement	*Annoyance*: P3 stated "this took long, it just keep going".
	Understandability: P6 asked "what does this mean?".
	P1 spotted two non-matching text in the template fields.
	Format: P1 asked "Is it meant to be like in an excel?".
WDST missing fields	P1 said "we have a condition, but we don't say what is done as a result of that condition". Thus, that is missing in the condition section.
WDST content flow	The participants declared that it was good; P3 said "you have the right things in the beginning, and you're going through the steps in a nice order", and P5 complemented "I don't think you can change the order of things".
WDST usage	P1 and P3 said that they would use, if they were asked to. P1 mentioned that WDST "makes it easier than just writing free text". Three other participants stated that they would not use it, due to its complexity and time consummation; instead they would do "the scripting directly".
WDST complexity	Four participants ensured its high complexity, where two of them linked the complexity with the amount of information to be written. In opposition, the document was not complex, as P4 said "I think it was clear".
WDST tutorial necessity	P1 and P5 said that a *manual or example* would be enough, and P6 stated that *training* is a good idea. In opposite, P3 said that text is sufficient "with the light grey" and P4 stated that it is "self-explained".
Confidence to fill the WDST	Was felt by three participants unlike the others.

D. Third Iteration

The third and final iteration also started with identifying solutions for the participants' raised issues, creating a new version of the artefacts in the design and development step. The artefacts were re-evaluated and validated by the participants, producing new data that were analysed. See Figure 1, on the methodology section, for this iteration steps.

1) Identify Solutions: The UML AD meta-model required an association between the pins and the database represented with a new composition arrow. Additionally, based on feedback, an attribute was added to identify if the tool is automatically or manually operated. This addition to the tool stereotype-class made it necessary to include a new icon to differentiate these states. Further, the alterations on the concrete syntax were identified when participants expressed annoyance, confusion or issues faced.

Furthermore, the WDST annoyed the participants because of its documentation traceability fields and its descriptive nature, which was unfamiliar to the participants. For example, the *workflow ID* and *workflow name* cells, which were used interchangeably, occasioning the removal of the *workflow name* and *step name* fields. The participants suggested linking the *workflow ID* on the first page to the second page to avoid typing the same information twice. Thus, this suggestion was considered, and a basic excel formula created to solve this

Table XI
THE COLLECTED RESULTS FROM THE PARTICIPANTS' THINK-ALOUD LOGS IN THE SECOND ITERATION

Log Categories	Concrete Syntax Library	WDST
Missing points	P1 - *Database* with *input pin* notation	
Annoying points	P3 - draw.io	P1, P3, P5 & P6 - Fields repetition - *ID/Name*
	P3 - Small font size	P5 - A lot of typing
Medium problem (unclear usage)	P1, P3, P4, P5 & P6 - The difference between *action* and *tool* is confusing	P1 - *step ID* is confusing
	P3 - Where do I write the condition? (*soft* and *hard conditions*)	P3 - *step name* & *step ID* are being interchanged
	P4 - The description of the performed activity, in the *tool* shape was removed	P4, P5 & P6 - Why: *where it happens?*
	P4 - *Input/output pins* were not used as intended, instead they were used as the *standalone pin*	P5 - Why: all these *IDs*?
	P5 - The description of the performed activity, in the tool shape was removed	P5 - Why: *creation day* in step?
	P6 - What is the difference between *hard* and *soft condition*?	
Minor Problem (unclear language)	P3 - The *input* and *output pin* description text box are far away from the *action* and *tool*	P1 - What do you mean by *concurrent*?
	P5 - *Input* and *output pins* location on the *tool* shape makes it go diagonally	P1 - *Conditions section* needs an explanation
	P6 - What is *flow final?*	P3 - What is *workflow source*?
		P3 & P5 - What is *super process of?*
		P4 - What is *concurrent*?
		P6 - What are *first* and *last step*?
		P6 - What is *order of execution*?
Others		P1 - *Process step ID/Name* on page 2 is different than *Step ID/Name* on page 1
		P3 - Template issue on *tool settings & parameters*
		P5 - Too much detail
		P5 - The grey text needs to be black
		P6 - What happens if there is no *output*?

issue; it copies the fields content to other sheets in the same document. *Creation day* in the *step* section was identified as unnecessary since the creation of a step is most likely to be the same as the workflow. However, this field was replaced with the *modification date* to increase the traceability of decisions. Several fields generated questions related to their meaning or purpose; thus, a better explanation as a light grey text was provided for these cells. The sentence *process step* was confusing; therefore, the word 'process' was removed from WDST. Additionally, two participants had issues to understand the word 'concurrent'. Thus, to help users, it was decided to aggregate the word 'parallel' to this sentence. Another issue was a badly formatted cell, *tool settings and parameters*, that had its main text deleted when the participants were typing; to fix this, the cell was merged and given more space below. Furthermore, a participant mentioned that the grey text makes reading a challenge, leading to the change of text colour by implementing conditional formatting. The grey text fields held the explanation to help new users and were thus kept since they are vital to the WDST understandability. Hence, the grey colour was used to allow the users to differentiate visually between the template's main and guidance texts.

2) Design and Development:

a) ***UML AD meta-model extension****:* The composition between the *database* stereotype class and *input* and *output pins* meta-classes was added as a new relationship. The *goal* was removed from the concrete syntax based on the participants' request, but it was kept on the meta-model since the WDST covers its description. Additionally, an attribute was added to the *tool* stereotype-class to differentiate between the automated and manual operations. Figure 3 contains the updated meta-model.

b) ***Concrete syntax****:* This section lists the concrete syntax applied improvements and the reasoning, in Table XII, while Table I shows the new version of the artefact.

Table XII
THE THIRD ITERATION CONCRETE SYNTAX IMPROVEMENTS

Concept	Improvement	Reasoning
InputPin	Location on *tool*	To ensure the vertical gradient of the diagram.
Database	The *input* and *output pins* were attached	To represent the data flow and to maintain the consistency between shapes in the XML notations library.
Tool & *Action*	Their descriptions were improved	To decrease confusion.
Tool	A separate text field was added for the performed activity	To remove the issue of deleting the name or performed activity when writing them.
Tool	Added new notation for the manually operated *tool*	To increase the process and steps visibility and transparency related to the automation level.
Goal	Removed	Based on the users' request.
Standalone pin	Added to the XML notations library	To include familiar notations to the bioinformaticians.

c) ***WDST****:* The researchers made several changes to the WDST based on the participants' feedback, Table XIII lists all of them.

The formula and conditional formatting are notable only on the excel template; therefore, they are not visible on Figure 4, which contains the updated version of the WDST.

Table XIII
THE THIRD ITERATION WDST IMPROVEMENTS

WDST	Improvement
Workflow ID & Workflow name	Grouped into *workflow ID*
Step ID & Step name	Grouped into *step ID*
Workflow ID	Added formula for copying sheet 1 content to sheet 2
All grey text fields	Added conditional formatting: if "value is not equal to" grey text, the font is set to black
In Sheet 1 and 2	Removal of the word 'process'
Workflow version, *Workflow source*, *Step version*, *Order of execution*, *File Input(s)*, *If no, what is the input's origin & File Output(s)*	Explanations rewording
Modification date	Added a field with its explanation field
Sub-process of Super-process of	Renamed to *sub-step of* & *super-step of*
Is this step concurrent to another	Changed to: *Is this step concurrent/parallel to another*
Tool settings and parameters	Field formatting was fixed
Condition/Threshold section	Modified section body, where most of its cells were reorganised
Condition result	Added a field with its explanation
DataStorage	Reworded to *database*
Where it happens	Reworded to *step execution location*

3) *Evaluation:* Six participants, including the head of the Bioinformatics Core Facility, joined the workshop on May 14th to validate the final version of the artefacts, one less than planned. During the workshop, these artefacts were explained using examples, see Appendix L, then the participants had paired-discussions, and after revealed their feedback in a group discussion. The generated codebook of these discussions, covering both artefacts, are presented in Appendix R.

a) ***Concrete syntax:*** This section contains an overview of the findings based on the participants' feedback to the concrete syntax during the third iteration, see Table XIV.

b) ***WDST:*** This section has a summary of the collected data regarding the WDST during the third iteration, see Table XV.

4) *Communication and Validation:* The participants individually answered Likert scales and an open-ended question using Mentimeter to validate each artefact. The four Likert scale questions and their results are depicted in Figure 5 and 6, where the selected answers for both artefacts are shown with their average numbers and graphical representations with the range of provided values. These scales had values from 1 to 5, where 1 is very unlikely, incomprehensible, and arduous, while 5 is very likely, understandable, and easy. The answers to the Mentimeter open-ended questions for both artefacts are presented in Appendix S.

Table XIV
CONCRETE SYNTAX EVALUATION FINDINGS IN THE THIRD ITERATION

Code	Findings with Illustrative Statements
Notations & Concepts Overview	P7 said "I think it looks pretty neat, simple. At least this is the first time that I'm seeing it and I do understand what you're talking about".
Notations & Concepts improvements	P2 did not understand two concepts, *soft-conditions* and *source*, where P1 explained the former.
	At least 3 of the participants requested better software to draw the diagrams, not requiring to be web-based.
	P3 wanted the diagram to be automatically generated, as in Snakemake[5].
Unnecessary notations	The participants mentioned: *Fork* and *join nodes*, *swimlanes*, and *standard references*.
Diagrams current state	Overloaded and overused with *boxes* and *notes* symbols.
Test of the library	The two participants that did not participate in it, during the second iteration, stated that they would have provided better feedback, if they have done it.
Notations and concepts usage	P1 and P7 said that it would be for final and standard documentation, after sketching.
Library usage effect on the current state	P1 and P7 said that it would increase the time spent to draw the workflows.

Table XV
WDST EVALUATION FINDINGS IN THIRD ITERATION

Code	Findings with Illustrative Statements
WDST Improvement	*Disliked* points: high amount of typing, traceability issues, and possible confusion that the text could generate were indicated by three participants.
	Automation: P2 mentioned "I mean now when we document workflows for the hospital, we have to present like a table of the tools, and the parameters used and stuff. So if that could be automated as well and done from the graphics, that would be good. Would save a little bit of time".
WDST complexity	All participants stated that the stakeholders would have trouble understanding it.
WDST usage	None was identified & P3 stated that WDST is "more complicated".
WDST usage effect on the current state	Relates to the time spent to fill the document.

a) ***Concrete syntax:*** Figure 5 shows that the participants find the concepts and notations of the library understandable with an average of 4.3, where 3.7 was the result of how easy it was to use them. Meanwhile, the participants would likely use the concepts and notations, with an average of 3, and 2.8 is their average belief on stakeholders' understandability.

The open-ended question had similar results from the discussion described above. However, one participant requested a further improvement to, "make it easier to add several outputs". Moreover, a participant proposed renaming the *soft-condition* to "manual-inspection or manual evaluation" and changing its concrete syntax to differentiate it even more from the *hard-condition*. A participant abstained from answering.

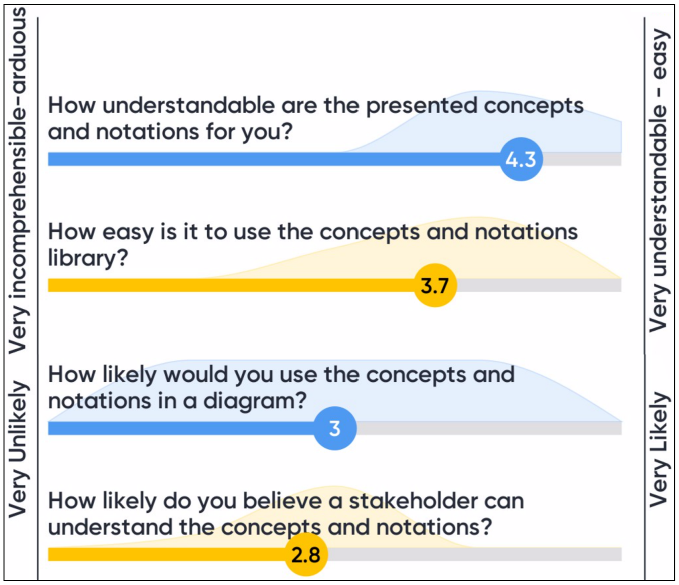


Figure 5. The concrete syntax four validation Likert scale questions with their average results and range of values in a graphical representation (Provided by Mentimeter).

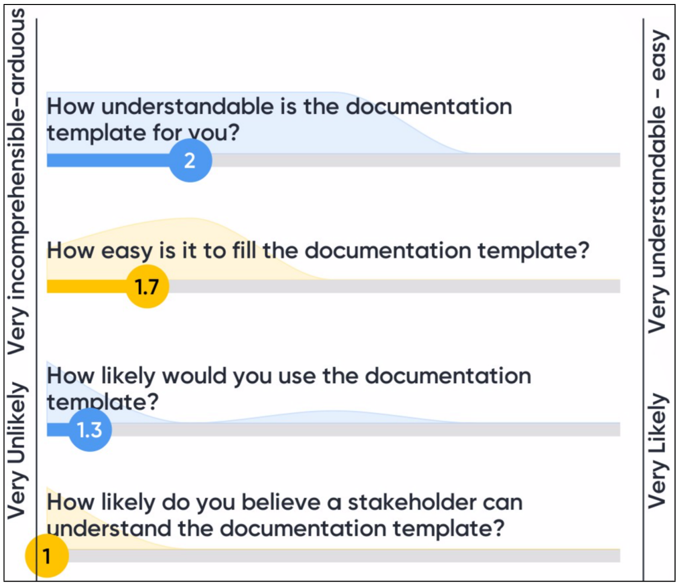


Figure 6. The WDST four validation Likert scale questions with their average results and range of values in a graphical representation (Provided by Mentimeter).

b) ***WDST:*** Figure 6 depicts that the WDST was incomprehensible by most of the participants, with an average of 2 and 1.7 regarding the ease of filling it. The participants would very unlikely use the WDST, 1.3, and they do not believe that the stakeholders would understand it, resulting in an average of 1.

The Mentimeter individual and anonymous open-ended question had more direct answers than the discussion, where five participants agreed that it is complicated. Thus, they suggested to simplify it by removing most of its content, keeping only the *tool section*, and adding a place to input the command line commands. One participant left the question unanswered.

V. Discussion

A. *RQ1.1*

The answers to the question *what are the defining and unique characteristics of bioinformatics workflows compared to standard workflows?* were found mainly on the first iteration, being expanded and improved on the second and third iterations. As a starting point, the concepts used for this extension were: from the standard UML AD [13]; the identified gaps in [5], such as *goals*, *source*, *loop*, *soft*, and *hard-conditions*; the researchers' proposals for *tool* and *diagram separators* in the first iteration; and the second iteration addition based on the participants' suggestions, such as *standard reference* concept and the attributes *concurrent steps* as well as *tool settings and parameters*. However, these concepts and attributes are highly used characteristics by bioinformaticians, but not all of the concepts can be considered unique to the bioinformatics field because their origins are from different domains or modelling languages. Three of these concepts, *diagram separators*, *standard references*, and *tool* with its attributes, were not encountered in any related work, but they were requested by the domain expertise to fulfil their needs, and are thus considered unique to this field. The concepts used for extending UML AD notations bridge between the standard-workflow modelling, flowcharts, and standard UML AD, adding the data flow behaviour to the AD. This study also found that modifications and simplifications of the UML AD concrete syntax and semantics of the concepts are necessary to make them more understandable to the domain experts. Some participants requested removing the *standard reference* from the final UML AD extension, and three of the standard UML AD concepts, see Table XIV. These requests were based on language misunderstanding or confusion, and the participants' preference for the current way of modelling. Thus, the researchers agree with the authors in [14] that the lack of knowledge in the modelling languages and the minimum time employed to create the diagram results in misuse of the language, as identified by the bioinformaticians preference for their current modelling status of using boxes, texts, and arrows. To summarise, some of the participants' requests were considered, and concept simplification was introduced, but nothing was removed from the standard UML AD due to the profile nature. Meanwhile, other suggestions were rejected during the iterations, based on the provided counter-arguments given by the researchers and, in some cases, by the other participants.

B. *RQ1.2*

At least two theories and the feedback provided by the participants were used to answer the research question *how should workflows, including the concepts discovered in RQ1.1 be visualised to be understandable by the bioinformaticians?* The theories, Visual Alphabet and Physics of Notations, covered in [23] and [26] were employed to create the concrete syntax.

The Visual Alphabet has eight variables, which are separated into two groups, planar and retinal [23]. The former was applied then changed, based on the received feedback, to avoid mixing the planar variables as in the SE field, while four out of six variables from the latter were used to diagrammatically encode information: distinct shapes, size, orientation, and brightness [23]. The other two, colour and texture, were unutilised in this study, even though the careful usage of colour increases the visual expressiveness [23].

Related to using the Physics of Notations theory in this research, a suggestion for symbol overload was proposed, for the *tool* and *actions* to have the same shape with different labels, that decreases the language complexity. However, the suggestion would diminish visual distance by lowering their distinguishability [23], [26], as well as conflict with the used theories. Therefore, the researchers decided to keep the two distinct shapes. Further, written labels, a UML common practice, were used to distinguish the overloaded *control flow* arrow since it represents almost all relationships. In the second iteration, icons were introduced to increase perceptual discriminability and directness for the automated and manually operated *tool*s since the icons suggest a meaning [23]. The latter was a significant new concept added to the language since it can help to identify bottlenecks in the process for the full automation, as desired by the participants.

The concrete syntax passed through three evaluations. Regarding the participants' preferences in the first iteration, the feedback received was compatible with the theories from [23], [26], which were used by the researchers as a base for designing and refining the concrete syntax. Nevertheless, the first and second iterations did not result in any deletion, while in the last iteration, four concepts and notations were seen as unneeded. The researchers believe that the change of heart was due to the provided discussion possibility in a group, resulting in the participants' confidence to use specific concepts over others.

Finally, the UML AD extension in this thesis has a high graphical complexity, measured by the size of its visual vocabulary, containing 14 standards and 9 extended notations, totalising 23 shapes, see Table I for the final graphical syntax extension and [13] for the standard. To decrease its complexity, symbol deficit was introduced [23] for the *goal* concept. Even though the complexity is high, the participants mentioned an understandability average of 4.3 because they probably learned the notations and their usage along with this thesis, while the average decreased to 2.8 out of 5 for stakeholders' understandability. Based on the participants' comments, they prefer their current graphical representations than the provided notations from this study because using the former requires less knowledge about the modelling language and more about the context. Additionally, some participants mentioned that the shapes were not intuitive when validating the concrete syntax; thus, a label would be needed for presenting a workflow diagram to people unfamiliar with the language, which highlights its low perceptual directness. Therefore, the researchers believe that to diminish the current faced problems and the solution to this research question would be the usage of the concrete syntax created in this research, together with labels, to represent the bioinformaticians' workflows in an understandable way diagrammatically.

C. RQ1.3

At the beginning of this study, the WDST was envisioned for elicitation, where the bioinformaticians would use it to gather all the information from their stakeholders to draw and create the workflow. However, during the first iteration, none of the participants said that they use it this way, rather they would draw a diagram first and then fill the documentation. Therefore, the researchers changed the WDST purpose to document the workflow, holding all its information. Even after this change, the participants preferred the diagrams over the WDST, but during the three iterations, the participants described for what and why they would use it. This inclination is probably related to the belief that graphics can deliver information to non-technical people in a more tangible way than text [54].

The attempt to answer the research question *how can we design a useful and understandable template to document the concepts from RQ1.1 from the bioinformaticians viewpoint?* resulted in a unanimously disliked template. The WDST artefact had only negative average scales ranging from 1 to 2 during its validation, resulting in a failed attempt of providing an easy, understandable, and usable template, due to its high complexity, and their preference to keep the actual state of writing subjective free texts. However, three important findings were distinguishable: first, the participants want an automatically generated documentation; second, it must contain the tools settings and parameters; and third, the amount of text and technicality should be as low as possible. The researchers agree with the first and second; however, lowering the amount of text will not solve the lack of traceability, details, and written documentation of the workflow identified in the domain [5]. The researchers believe that an automatically generated documentation after drawing the workflow is the best solution to this research question.

D. Comparison with Related Work

The previous work in [20], [23], [24], and [30] had not addressed any extension for bioinformatics workflows nor bioinformatics domain-specifications, which makes their work incomparable to this study. However, the authors in [18] evaluated their written template and found that it is useful for small or academic software product lines, suggesting to improve its automated filling for multiple fields. Likewise, the findings of this research showed that the produced WDST would also be used in specific cases, such as when several facilities are involved or for standard and repeatable projects. Additionally, we have tackled that automated filling issue in WSDT excel sheets, but the participants were still requesting its full generation from the diagrams automatically.

E. Further Work

The participants described using Powerpoint[6] and some non-modelling tools to draw the workflows, which provides

[6] https://products.office.com/en/powerpoint

simplistic notations, boxes and arrows, that are used for all concepts. However, the used modelling tool, draw.io, was described by the participants as not being the right choice either, although the software was chosen based on four requirements: free, online, shareable, and easy to use. During the final iteration, the participants stated that an online tool is unnecessary; instead, the requirements should be higher precision when positioning the shapes, the possibility to input the tool settings and parameters in the shapes fields, and to generate documentation from it. Thus, the researchers would suggest finding and evaluating alternative software that meets the bioinformaticians' requirements.

The researchers would recommend validating the concepts with a broader bioinformatics community to eradicate individual preferences and subjectiveness. Additionally, a suggestion to improve this study would be to reduce the overloaded *control flow* shape by using different sizes, brightness, and arrowheads, which might increase the language complexity and be even more unfamiliar to the bioinformaticians. Another research suggestion could be to verify if the proposed UML AD extension and its documentation would improve shareability and understandability among facilities and bioinformaticians. Lastly, it would be interesting to measure, in a new study, how many problems can be identified in the bioinformatics workflows or to identify the number of manual operations that were thought automated.

F. Validity Threats

1) External: This thesis aimed to solve a domain specific problem following a DSRM with a small sample sizes [35]. Thus, purposive sampling is convenient with low-cost and minimal time consumption. This sampling method generates non-generalised results when a particular set of people or organisation are involved [34]. Aiming to address that, three facilities took part during this study, and the participants work with different workflows or different ways of designing workflows.

2) Internal: The lack of bioinformaticians resulted in the availability of only seven participants. Thus, they were repeated along with the iterations based on their schedules, where a new participant was invited for each iteration. One of the disadvantages of having new participants in the middle of the process is the time demanded to explain the purpose, artefacts, and process. On the contrary, the artefacts being seen with fresh eyes can increase the number of constructive feedback. Besides, the bioinformaticians, that participated in previous iterations may get annoyed not to see their suggested improvements applied.

One of the drawbacks of using a workshop is because its success has a close relationship to the researcher's ability to engage, e.g., the usage of a common language can straighten this relationship [55]. The researchers had not done workshops before, but they tried to use the bioinformaticians' language and workflow examples to communicate with the domain experts. Additionally, the authors in [56] stated that the effectiveness of workshops relies on the participant's experience and or desire to participate. Therefore, the participants from previous iterations were invited to attend the workshop, so the content was not new to them, and their willingness to engage was checked during the repeated iterations in this research. Another disadvantage of group activities is the possibility for individuals to avoid taking part in the discussions and follow the crowd. To mitigate that, the seven participants were paired during the discussions to stimulate participation and prevent inhibition.

The researchers observed that the participants were avoiding answering the questions related to the WDST usage, addition and removal of fields, by providing evasive and polite answers. As a mitigation to this occurrence, the validation question in the final iteration was performed entirely anonymous, even to the researchers, using Mentimeter. This approach revealed the participants' real thoughts about WDST.

As mentioned in the methodology section, none of the participants neither the researchers were native English speakers neither share the same expertise domain. Thus, language barriers could be a problem along with this study since they affect knowledge transference and understandability. Another observation made was the divergence of concepts usage among the three involved facilities. Hence, the concepts were familiar to some and unfamiliar to others; therefore, some participants needed clarifications, such as some of the participants used thresholds while others not, while some used parallel steps others not. However, the researchers adopted a simple language while interacting with the participants, created discussions sections, asked questions to follow up, and provided clarifications to mitigate any misunderstanding and miscommunication.

3) Construct: Interviews are known to be intrusive for the participant, time-consuming and being susceptible to bias [42]. Therefore, the interviews were short and time-boxed to prevent participants' fatigue and were in a familiar environment. Additionally, two tactics were employed to avoid biased answers, first, the guaranteed anonymity of the participants' ideas and responses, and second, the questions had no correct answers since they asked for opinions.

4) Reliability: The quality of the transcription done with Temi is closely related to the people's accent, free background noise, and closeness of the microphone. The environment noise was avoided by holding the interviews in a closed-door meeting room. Additionally, the microphone was placed in the middle of the table in which the researchers and the participants were sitting. However, none of the recorded people was a native English speaker; thus, the presence of accents was expected to be quite strong. Aiming to increase the reliability of the transcription, the researchers used the tool to correct the transcription manually.

The created codes can be biased or misinterpreted, thus to avoid that, one researcher created the code frame with its description and mentioned statements. After that, the other researcher ascertained its reliability by calculating the correspondence between the applications of the codes to the data [44].

The researchers believe that other authors would create the same concepts of this study but give them different names depending on their origin field and other factors. The validated concepts were: tool, diagram separators, standard reference, standard reference connector, condition connector, goal, source, and loop connector. Some of these additions were justified by the unpublished report in [5] and the participants' validation.

5) Conclusion: By not analyzing the participant's diagrams and written documents, the researchers could have missed relevant feedback to improve the artefacts. However, the researchers mitigated that by using think-aloud protocol and the observation log methods to collect the participants' struggles, ideas, reasoning, etc.

VI. Conclusion

The current state of the bioinformatics workflows diagrammatic and written documentation are subjective and not standardised. This thesis presents a UML AD extension with its concrete syntax and a WDST as one of the first attempts for standardisation, where bioinformaticians validated the proposed concrete syntax as being an understandable and straightforward modelling language. According to the bioinformaticians, this extension will be used to document standard workflows and formal documentation, usually requested by stakeholders for system documentation. The created WDST requires refinement and automation to be used for knowledge sharing and formal documentation by the bioinformaticians, in the future, since it was unsatisfying during the evaluation. In addition to that, these standard documentations would increase management efficiency, understandability, shareability, traceability, and knowledge transference of bioinformatics workflows. Therefore, we suggest a further investigation for a new modelling tool, which allows generating documentation from the diagram with better user experience.

Acknowledgement

We are grateful for Michel Chaudron's guidance while discussing the modelling languages and tools during the early stages of this research. Additionally, we extend our thanks to Marcela Davila and the bioinformaticians that participated for being available and willing to answer our questions. Finally, we would like to thank our supervisor, Jennifer Horkoff, for her incredible patience, effort, and availability to guide us along the research process.

References

[1] Gauthier, J., Vincent, A. T., Charette, S. J., & Derome, N. (2018). A brief history of bioinformatics. *Briefings in Bioinformatics*, 1-16.

[2] Leipzig, J. (2017). A review of bioinformatic pipeline frameworks. Briefings In *Bioinformatics*, 18(3), 530-536.

[3] Kanwal, S., Lonie, A., & Sinnott, R. O. (2017, November). Digital reproducibility requirements of computational genomic workflows. In *2017 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1522-1529). IEEE.

[4] Krishna, R., Elisseev, V., & Antao, S. (2018, August). BaaS: Bioinformatics as a Service. In *European Conference on Parallel Processing* (pp. 601-612). Springer, Cham.

[5] Horkoff, J., de Oliveira Neto, F. G., Schliep, A., & Davila, M. (2018). *Optimized Bioinformatics Workflows from Requirement Engineering of Solution Specifications*. Unpublished report.

[6] Gilbert, D. (2004, September). Bioinformatics software resources. *Briefings in bioinformatics*, 5(3), 300-304.

[7] *Common Workflow Language*. (n.d.). Retrieved March 6, 2019, from https://www.commonwl.org/

[8] Conery, J. S., Catchen, J. M., & Lynch, M. (2005). Rule-based workflow management for bioinformatics. *The VLDB journal*, 14(3), 318-329.

[9] Karim, M. R., Michel, A., Zappa, A., Baranov, P., Sahay, R., & Rebholz-Schuhmann, D. (2017). Improving data workflow systems with cloud services and use of open data for bioinformatics research. *Briefings in bioinformatics*, 19(5), 1035-1050.

[10] Kosar, T., Bohra, S., & Mernik, M. (2016). Domain-specific languages: A systematic mapping study. *Information and Software Technology*, 71, 77-91.

[11] Roux-Rouquié, M., Caritey, N., Gaubert, L., & Rosenthal-Sabroux, C. (2004). Using the Unified Modelling Language (UML) to guide the systemic description of biological processes and systems. *Biosystems*, 75(1-3), 3-14.

[12] Williams, R. A., Timmis, J., & Qwarnstrom, E. E. (2016). Statistical techniques complement UML when developing domain models of complex dynamical biosystems. *PloS one*, 11(8), e0160834.

[13] OMG (2017). *OMG Unified Modeling Language (OMG UML), Superstructure, Version 2.5.1* Object Management Group (Technical report, Object Management Group).

[14] Gray, J., & Rumpe, B. (2018). UML customization versus domain-specific languages. *Software and Systems Modeling (SoSyM)*, 17(3), 713-714.

[15] Pitts, M. G., & Browne, G. J. (2007). Improving requirements elicitation: an empirical investigation of procedural prompts. *Information systems journal*, 17(1), 89-110.

[16] Oinn, T., Addis, M., Ferris, J., Marvin, D., Senger, M., Greenwood, M., Carver, T., Pocock, M. R., Wipat, A., & Li, P. (2004). Taverna: A tool for the composition and enactment of bioinformatics workflows. *Bioinformatics*, 20(17), 3045-3054.

[17] Pandey, D., Suman, U., & Ramani, A. K. (2010, October). An effective requirement engineering process model for software development and requirements management. In *2010 International Conference on Advances in Recent Technologies in Communication and Computing* (pp. 287-291). IEEE.

[18] Gallina, B., & Guelfi, N. (2007, June). A template for requirement elicitation of dependable product lines. In *International Working Conference on Requirements Engineering: Foundation for Software Quality* (pp. 63-77). Springer, Berlin, Heidelberg.

[19] Fernández, D. M., Böhm, W., Vogelsang, A., Mund, J., Broy, M., Kuhrmann, M., & Weyer, T. (2019). Artefacts in software engineering: a fundamental positioning. *Software & Systems Modeling*, 1-10.

[20] Al-alshuhai, A., & Siewe, F. (2015, October). An extension of UML activity diagram to model the behaviour of context-aware systems. In *2015 IEEE International Conference on Computer and Information Technology; Ubiquitous Computing and Communications; Dependable, Autonomic and Secure Computing; Pervasive Intelligence and Computing* (pp. 431-437). IEEE.

[21] Decker, M. (2009, October). Modelling location-aware access control constraints for mobile workflows with UML activity diagrams. In *2009 Third International Conference on Mobile Ubiquitous Computing, Systems, Services and Technologies* (pp. 263-268). IEEE.

[22] Ricardo, M. B., & Duncan, D. A. (2002). Extending UML Activity Diagram for Workflow Modeling in Production Systems. In *The 35th Hawaii International Conference on System Sciences*. Hawaii: IEEE.

[23] Moody, D. L., Heymans, P., & Matulevicius, R. (2009, August). Improving the effectiveness of visual representations in requirements engineering: An evaluation of i* visual syntax. In *2009 17th IEEE International Requirements Engineering Conference* (pp. 171-180). IEEE.

[24] Stefanov, V., List, B., & Korherr, B. (2005, August). Extending UML 2 activity diagrams with business intelligence objects. In *International Conference on Data Warehousing and Knowledge Discovery* (pp. 53-63). Springer, Berlin, Heidelberg.

[25] Korherr, B., & List, B. (2006, November). Extending the UML 2 activity diagram with business process goals and performance measures and the mapping to BPEL. In *International Conference on Conceptual Modeling* (pp. 7-18). Springer, Berlin, Heidelberg.

[26] Moody, D. (2009). The "physics" of notations: toward a scientific basis for constructing visual notations in software engineering. In *IEEE Transactions on software engineering*, 35(6), 756-779.

[27] Moody, D. L. (2002). Complexity effects on end user understanding of data models: An experimental comparison of large data model representation methods. *ECIS 2002 Proceedings*, 10.

[28] Selic, B. (2007). A Systematic Approach to Domain-Specific Language Design Using UML. In *10th IEEE International Symposium on Object and Component-Oriented Real-Time Distributed Computing (ISORC'07)*, 2-9.

[29] Xia, H., Jiao, J., & Dong, J. (2019, January). Extend UML Based Timeliness Modeling Approach for Complex System. In *2018 International Conference on Mathematics, Modeling, Simulation and Statistics Application (MMSSA 2018)*. Atlantis Press.

[30] Syriani, E., & Ergin, H. (2012, September). Operational semantics of UML activity diagram: An application in project management. In *2012 Second IEEE International Workshop on Model-Driven Requirements Engineering (MoDRE)* (pp. 1-8). IEEE.

[31] Störrle, H. (2004, August). Semantics of structured nodes in UML 2.0 activities. In *2nd Nordic Workshop on UML* (pp. 19-32).

[32] Read, M., Andrews, P. S., Timmis, J., & Kumar, V. (2014). Modelling biological behaviours with the unified modelling language: an immunological case study and critique. *Journal of the Royal Society Interface*, 11(99), 20140704.

[33] Linaker, J., Sulaman, S. M., Hst, M., & de Mello, R. M. (2015). *Guidelines for Conducting Surveys in Software Engineering v. 1.1*.

[34] Taherdoost, H. (2016). Sampling methods in research methodology: How to choose a sampling technique for research. In *International Journal of Academic Research in Management (IJARM)*, 5(2), 18-27.

[35] Perumal, T. (2010, July). Topic 10 - Sampling. In *CMRM6103 Research Methodology / GMRM5103 Research Methods in Competitive Intelligence*. (pp.123-131). Seri Kembangan, Selangor Darul Ehsan: Open University Malaysia (OUM).

[36] Fricker Jr, R. D. (2016). *Sampling methods for online surveys*. The SAGE handbook of online research methods, 162-183.

[37] Hevner, A. R. (2007). A three cycle view of design science research. *Scandinavian journal of information systems*, 19(2), 4.

[38] Hevner, A.R., March, S.T., Park, J., & Ram, S. (2004). Design science in information systems. *MIS Q*. 28 (1) (2004) 75-105.

[39] Winter, R. (2008). Design science research in Europe. *European Journal of Information Systems*, 17(5), 470-475.

[40] Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2007). A design science research methodology for information systems research. *Journal of management information systems*, 24(3), 45-77.

[41] Schultze, U., & Avital, M. (2011). Designing interviews to generate rich data for information systems research. *Information and organization*, 21(1), 1-16.

[42] Doody, O., & Noonan, M. (2013). Preparing and conducting interviews to collect data. *Nurse researcher*, 20(5).

[43] Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking questions: the definitive guide to questionnaire design–for market research, political polls, and social and health questionnaires*. John Wiley & Sons.

[44] Joffe, H. (2012). Thematic analysis. *Qualitative research methods in mental health and psychotherapy: A guide for students and practitioners*, 1 , 210-223.

[45] Braun, V., Clarke, V., Hayfield, N., & Terry, G. (2019). *Thematic analysis*. Handbook of Research Methods in Health Social Sciences, 843-860.

[46] Boyatzis, R. E. (1998). Transforming qualitative information: Thematic analysis and code development. *Sage*.

[47] Martins, A. I., Rosa, A. F., Queirós, A., Silva, A., & Rocha, N. P. (2015). European Portuguese validation of the system usability scale (SUS). *Procedia Computer Science*, 67, 293-300.

[48] Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, 189(194), 4-7.

[49] Brooke, J. (2013). SUS: a retrospective. *Journal of usability studies*, 8(2), 29-40.

[50] Güss, C. D. (2018). What Is Going Through Your Mind? Thinking Aloud as a Method in Cross-Cultural Psychology. *Frontiers in psychology*, 9, 1292.

[51] Hertzum, M., & Holmegaard, KD (2015). Thinking aloud influences perceived time. *Human Factors*, 57 (1), 101-109.

[52] Nielsen, J., Clemmensen, T., & Yssing, C. (2002, October). Getting access to what goes on in people's heads?: reflections on the think-aloud technique. In *Proceedings of the second Nordic conference on Human-computer interaction* (pp. 101-110). ACM.

[53] Spyrou, S., Bamidis, P., Pappas, K., & Maglaveras, N. (2005). Extending UML activity diagrams for workflow modelling with clinical documents in regional health information systems. In *Connecting Medical Informatics and Bioinformatics: Proceedings of the 19th Medical Informatics Europe Conference (MIE2005)*. Geneva, Switzerland (pp. 1160-1165).

[54] Avison, D. E. (1996, August). Information systems development methodologies: a broader perspective. In *Working Conference on Method Engineering* (pp. 263-277). Springer, Boston, MA.

[55] Rosner, D. K., Kawas, S., Li, W., Tilly, N., & Sung, Y. C. (2016, February). Out of time, out of place: Reflections on design workshops as a research method. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 1131-1141). ACM.

[56] Orngreen, R., & Levinsen, K. (2017). Workshops as a Research Methodology. *Electronic Journal of E-learning*, 15(1), 70-81.

APPENDIX A
QUESTIONNAIRE FOR THE FIRST ITERATION

Demographic:

1. What is your role?

2. How long have you been working your role?

3. Have you created a workflow for a bioinformatics process?

3.1. If YES, briefly describe what did you do? (Language, software)

Interaction with RE artefact:

(We are going to provide them with the RE template and a short explanation)

4. Do you understand this?

4.1 If NOT, Is there anything that you can't understand?

5. Could this be used? Why, why not?

5.1 If YES, who and how?

6. Is this document missing anything?

6.1 If YES, what is it missing?

7. Did you identify any field that is unnecessary or that you will never need or use?

Interaction with Examples (1 and 2) and concrete syntax:

(We are going to provide them with both examples numbered and the concrete syntax list).

(Give the concrete syntax)

8. Are these concepts useful?

9. Are any of this redundant?

10. Are we missing anything?

(Explain the two different ways)

11. Which ... do you like better? Why?

11.1. Loop (1a or 2a)

11.2. Thresholds (1a or 2a)

11.3. Input/output (1b or 2b)

11.4. Datastore (1b or 2b)

11.5. Tools (1c or 2c)

11.6. Diagrams separators/connectors (1b and 1c or 2b and 2c)

12. Did you think you can draw and use a model like this in your work? Why, why not? Would anyone else be able to use that? Who?

13. Would the template be helpful to draw the diagram or not?

Appendix B
WDST for the First iteration

Workflow Requirement Specification Document			
Workflow Name:	*<<the workflow name or title>>*	**Workflow ID:**	<<workflow identifier>>
Date of creation:	*<<date in which this document was created or the workflow was requested>>*	Number of process step:	<<amount of steps>>
Version number:	*<<number based on modifications (change of tools, insertion of new threshold)>>*	Workflow creator:	<<name>>

Workflow	
Workflow goal:	*<<what do you want to achieve with this workflow?>>*
Workflow source:	*<< Is this workflow created locally, or it follows a reference - in that case link the reference>>*
Workflow responsible:	*<<person who signs the final output or who uses this workflow>>*

Initial Process Step (Start point)	
Process step name:	*<<The name of the start step>>*
Step ID:	

Final Process Step (End point)	
Process step name:	*<<The name of the final step>>*
Step ID:	

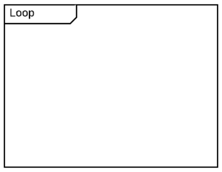
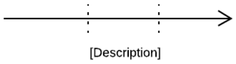
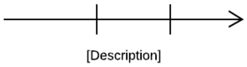
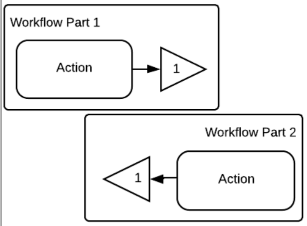
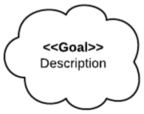
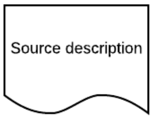
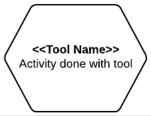
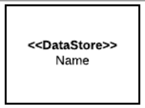
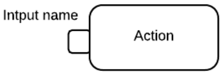
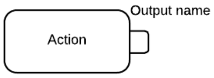
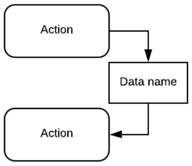
------------------------------------ END OF PAGE 1 - START OF PAGE 2 ------------------------------------

Workflow Requirement Specification Document			
Workflow Name:	*<<the workflow name or title>>*	**Workflow ID:**	<<workflow identifier>>
Process Step Name:	*<<the step name or title>>*	**Process Step ID:**	<<step identifier>>
Date of creation:	*<<date in which this step was added or the step was requested>>*	Step creator:	<<name>>
Version number:	*<<number based on modifications (change of tools, insertion of new threshold)>>*		

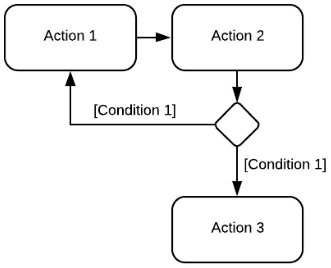
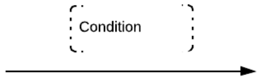
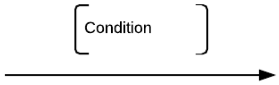
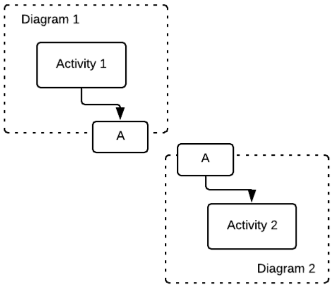
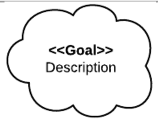
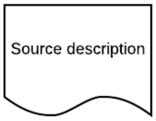
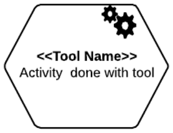
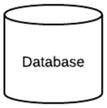
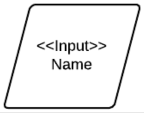
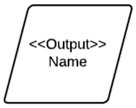
Process Step					
Step goal:	*<<what do you want to achieve with this step?>>*				
Step source:	*<< Is this step created locally, or it follows a reference - in that case link the reference>>*				
Step responsible:	*<<person who signs the final output or who uses this step>>*				
Is this step the initial workflow point:	Yes □ No □	Is this step the final workflow point:	Yes □ No □		
Sub-process of:	*<<ID of a process step>>*	Super-process of:	*<<ID of a process step>>*		
Order of execution:	*<<e.g. before y, synchronous to z>>*				
Who conducts the step:	*<<responsible role or person's name>>*				
Where the step happens:	*<<Lab/office/department - different place than the creator>>*				
File Input(s):	*<<Necessary data to start the process>>*				
Is the intput comming from another step:	Yes □ No □	If yes, step name:	*<<step name>>*	Step ID:	<<identifier>>
If no, what is the input origin:	*<<lab, person, tool >>*				
Tool 1					
Needed tool:	*<<equipment name>>*				
Why this tool was selected:	*<<reasoning or source for the decision>>*				
File Output(s):	*<<Generated data>>*				
Is the output used in another step:	Yes □ No □	If yes, step name:	*<<step name>>*	Step ID:	<<identifier>>
Tool 2					
Needed tool:	*<<equipment name>>*				
Why this tool was selected:	*<<reasoning or source for the decision>>*				
File Output(s):	*<<Generated data>>*				
Is the output used in another step:	Yes □ No □	If yes, step name:	*<<step name>>*	Step ID:	<<identifier>>
Tool 3					
Needed tool:	*<<equipment name>>*				
Why this tool was selected:	*<<reasoning or source for the decision>>*				
File Output(s):	*<<Generated data>>*				
Is the output used in another step:	Yes □ No □	If yes, step name:	*<<step name>>*	Step ID:	<<identifier>>
Loop/Repetion Section					
Is this step repeated during the process:	Yes □ No □	If yes, step name of loop start:	*<<step name>>*	Step ID:	<<identifier>>
		If yes, step name of loop end:	*<<step name>>*	Step ID:	<<identifier>>
If yes, how many times is it repeat:	<<number>>	If yes, what is needed to break the loop:	*<<condition to stop the repetition>>*		
Threshold Section					
Possible outcomes:	*<<possibility 1 (e.g. pass, fail)>>*	*<<possibility 2 (e.g. pass, fail)>>*	*<<possibility 3 (e.g. pass, fail)>>*		
Next step name for each outcome:					
Step ID for each outcomes:					
Threshold for judgment:					
Threshold is hard/soft:					
Data Storage Section					
Is the generated output stored:	Yes □ No □	If yes, the data must be stored until:	<<date>>		
If yes, name of the data storage:	*<<bucket name, table name, folder name>>*				

APPENDIX C
THE TWO CONCRETE SYNTAX FOR THE FIRST ITERATION

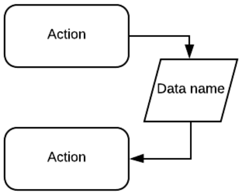
Concrete syntax 1 for First iteration

Name	Base Class	Description	Notation
Loop	ActivityEdge	An iterative set of activities and actions until reaching the defined condition.	
SoftThreshold	ActivityEdge	A set that is good to be reached	
HardThreshold	ActivityEdge	A set that must be reached	
DiagramSeparator	ObjectNode	A labelled triangle that represents the connection point with another part of the diagram from another page.	 Workflow Part 1 Action 1 Workflow Part 2 1 Action
Goal	ObjectNode	The aim of a specific activity.	
Source	ObjectNode	A link, document title, person's name which are the source for a specific set of actions.	
Tool	ObjectNode	The used tool to perform an activity with the activity described.	
DataStore	DataStoreNode	A structured set of data that is accessible in various ways.	
InputPin	Pin	The input values consumed by Actions or Tools.	
OutputPin	Pin	The output values produced by Actions or Tools.	
Standalone Pin	Pin	Optional notations: used when inputs and outputs are identical.	

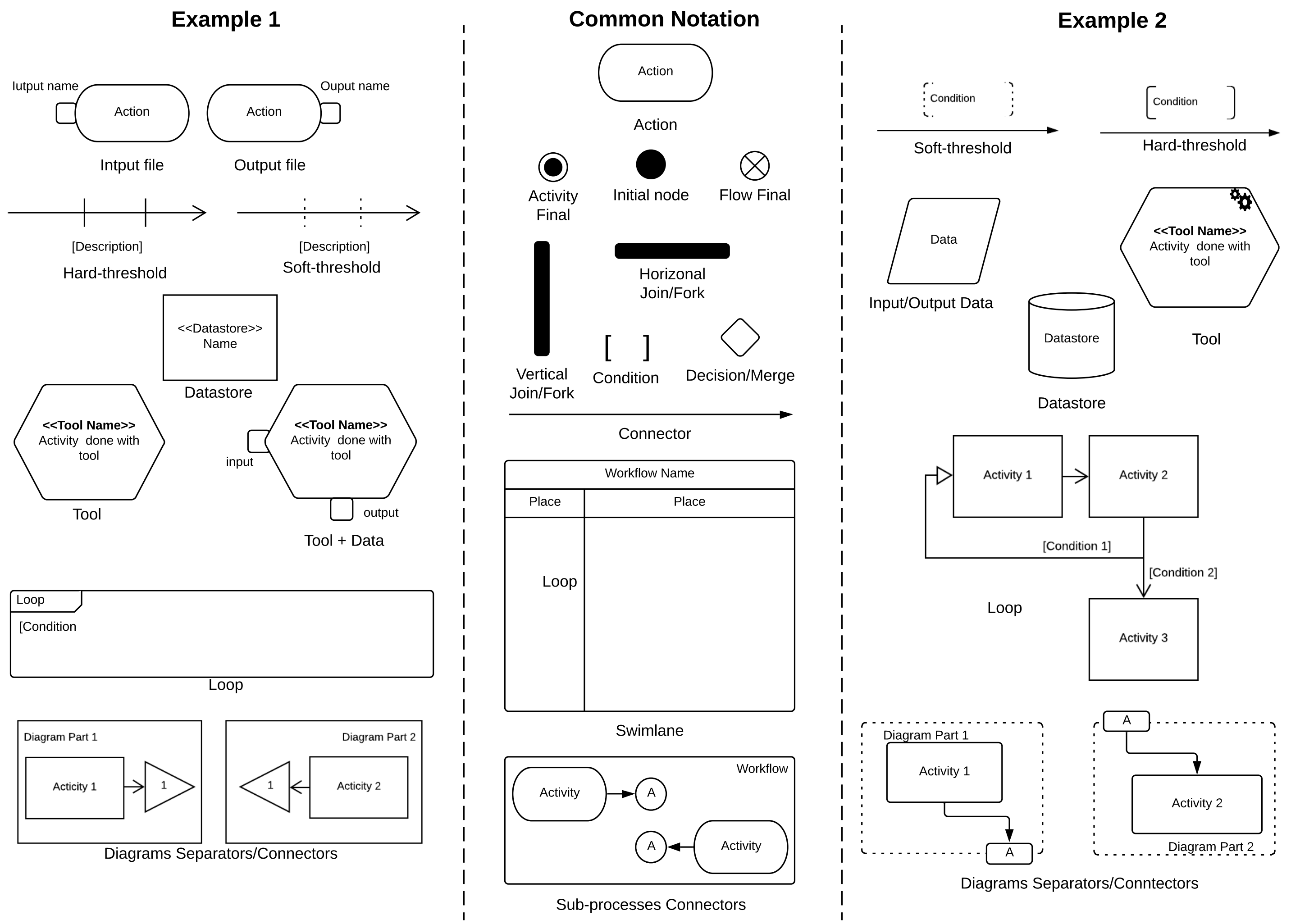
Concrete syntax 2 for First iteration - Part 1

Name	Base Class	Description	Notation
Loop	ActivityEdge	An iterative set of activities and actions until reaching the defined condition.	 Action 1 Action 2 [Condition 1] [Condition 1] Action 3
SoftThreshold	ActivityEdge	A level that is good to be reached.	
HardThreshold	ActivityEdge	A level that must be reached.	
DiagramSeparator	ObjectNode	A labelled square with a dashed fence that represents the connection point with another part of the diagram from another page.	 Diagram 1 Activity 1 A A Activity 2 Diagram 2
Goal	ObjectNode	The aim of a specific activity.	
Source	ObjectNode	A link, document title, or person's name, which is the source for a specific set of actions.	
Tool	ObjectNode	The used tool to perform an activity with the activity described.	
DataStore	DataStoreNode	A structured set of data that is accessible in various ways.	
InputPin	Pin	The input values consumed by Actions or Tools.	
OutputPin	Pin	The output values produced by Actions or Tools.	

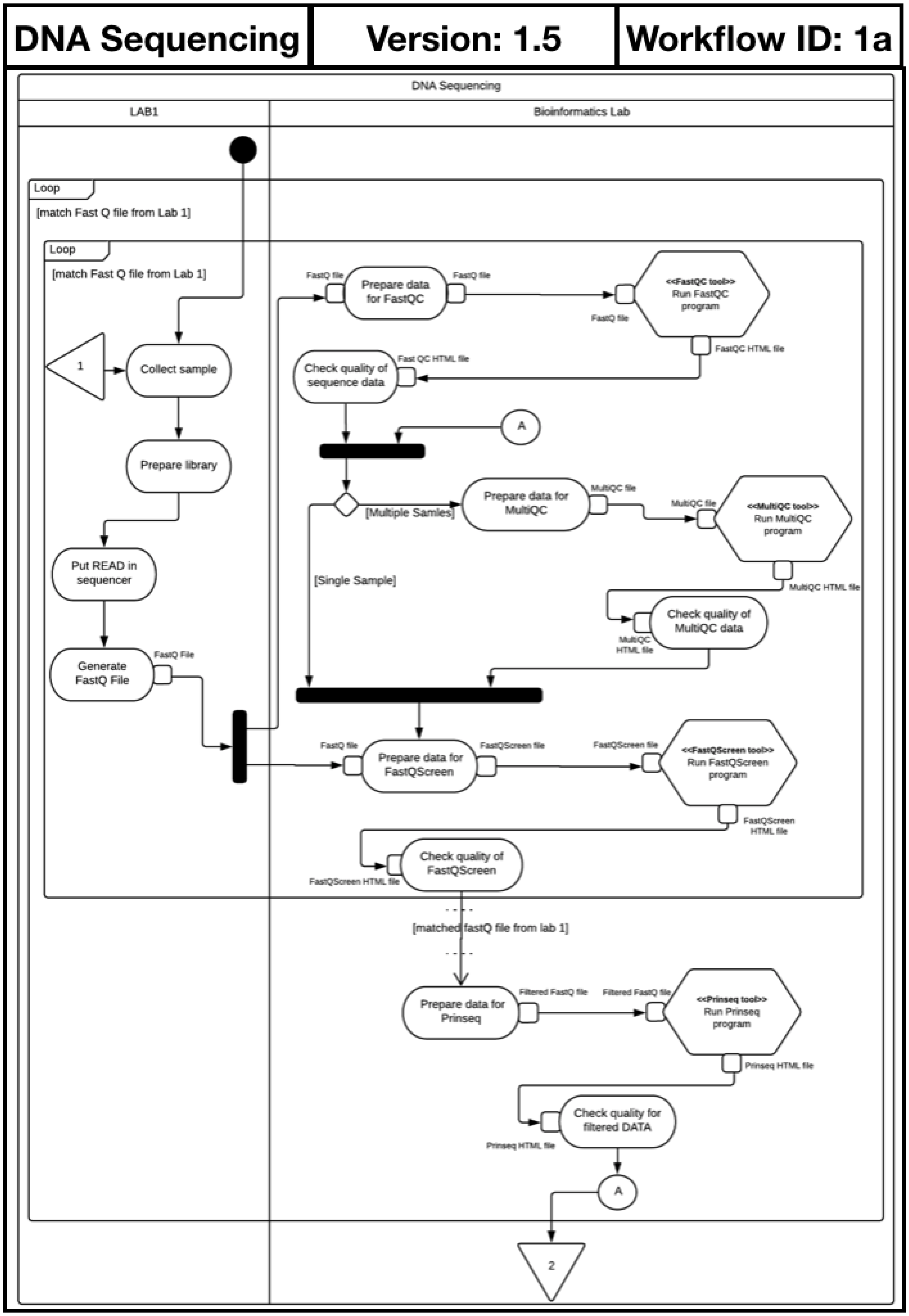
Concrete syntax 2 for First iteration - Part 2

Name	Base Class	Description	Notation
Standalone Pin	Pin	Optional notations: used when inputs and outputs are identical.	

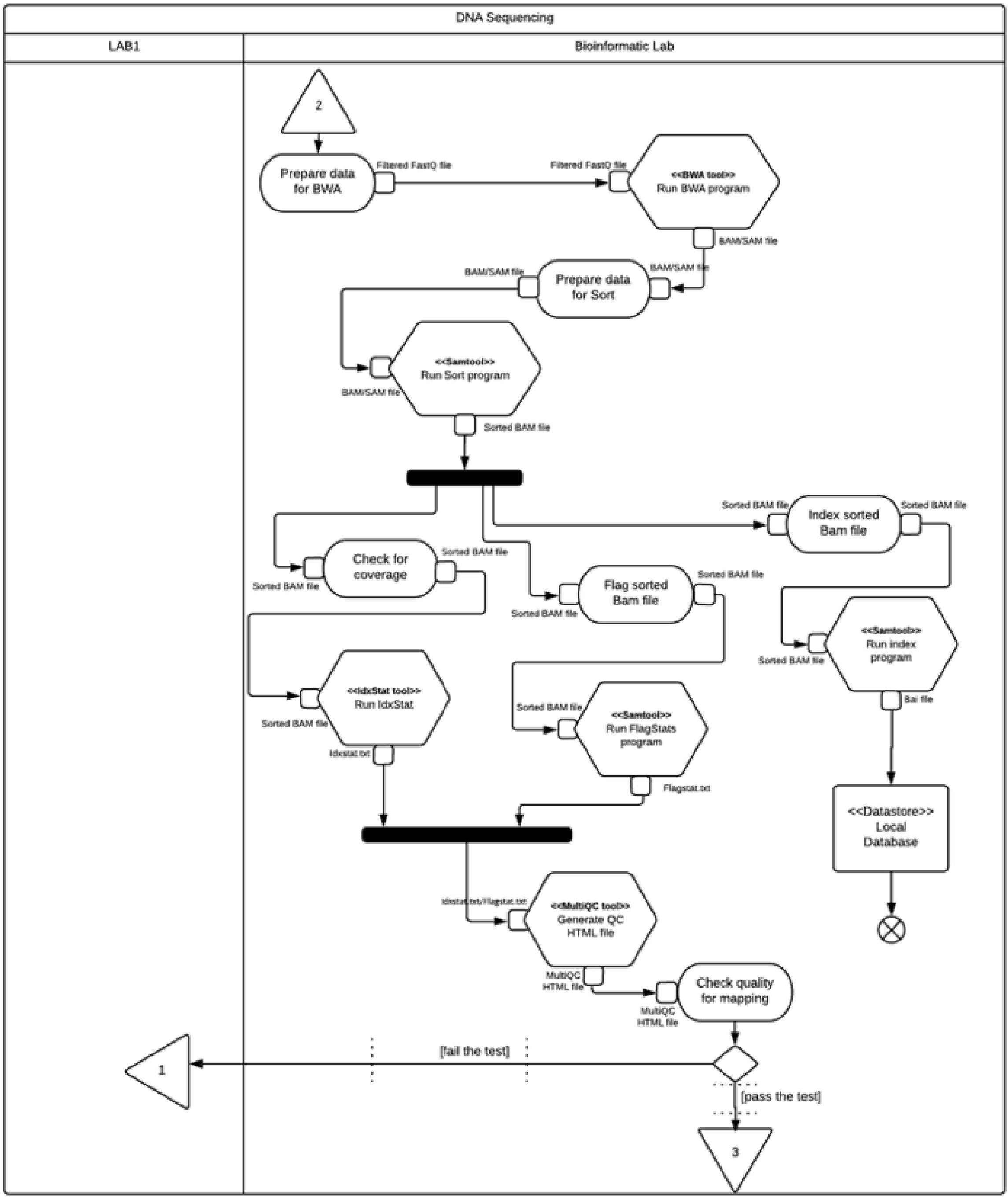
Appendix D
Legend of concrete syntax for the First iteration



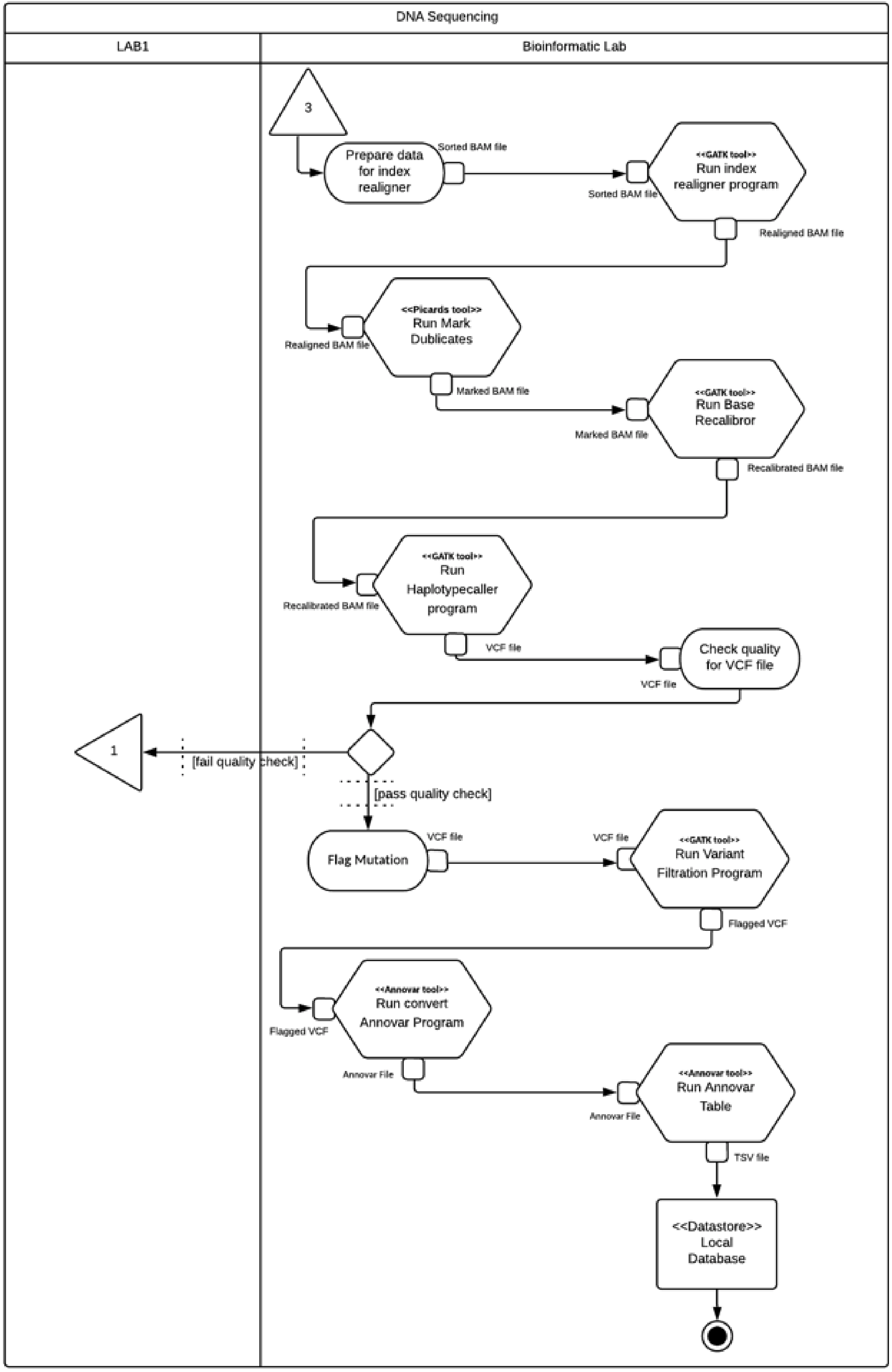
Appendix E
DNA Sequencing examples



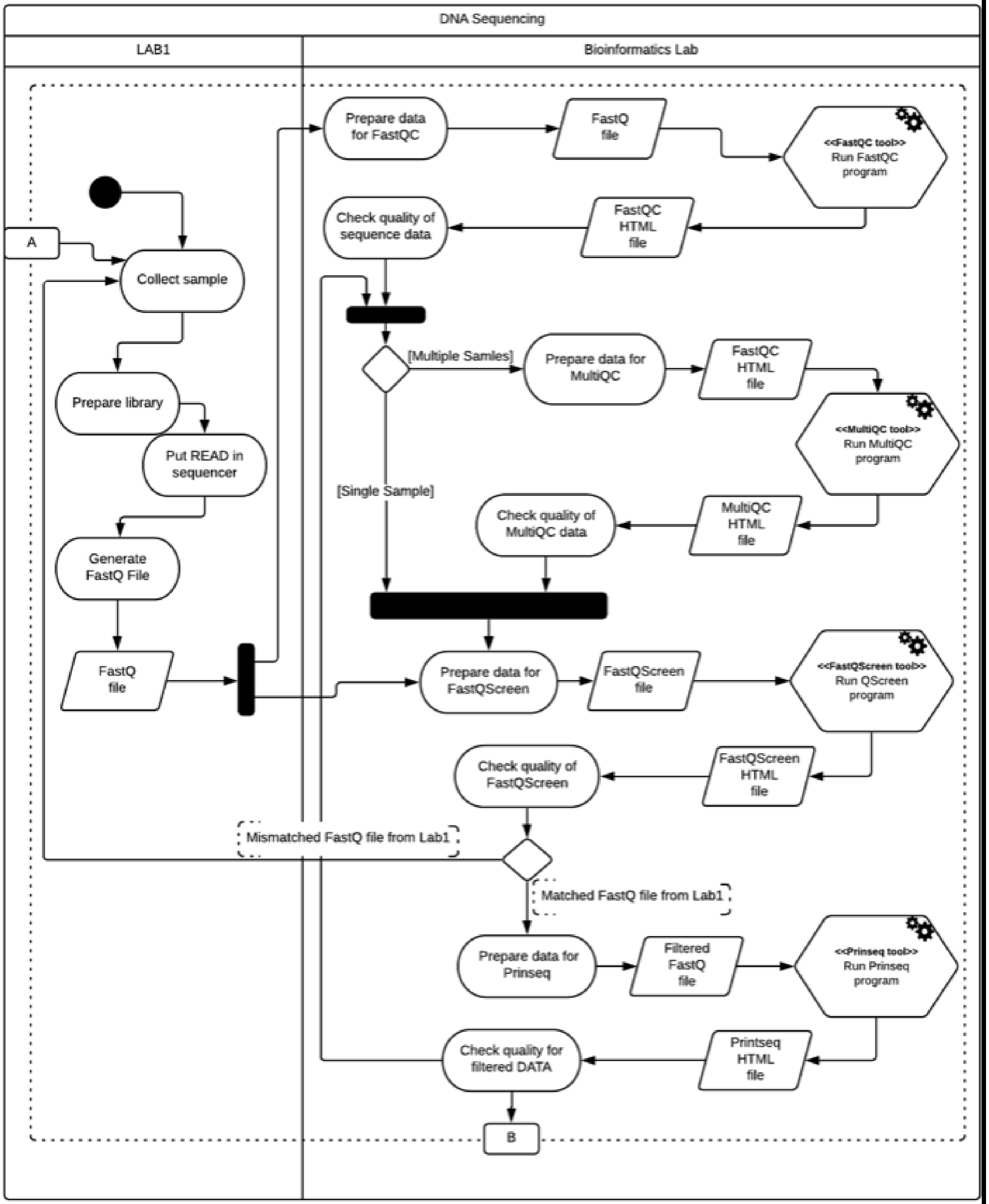
DNA Sequencing	Version: 1.5	Workflow ID: 1b



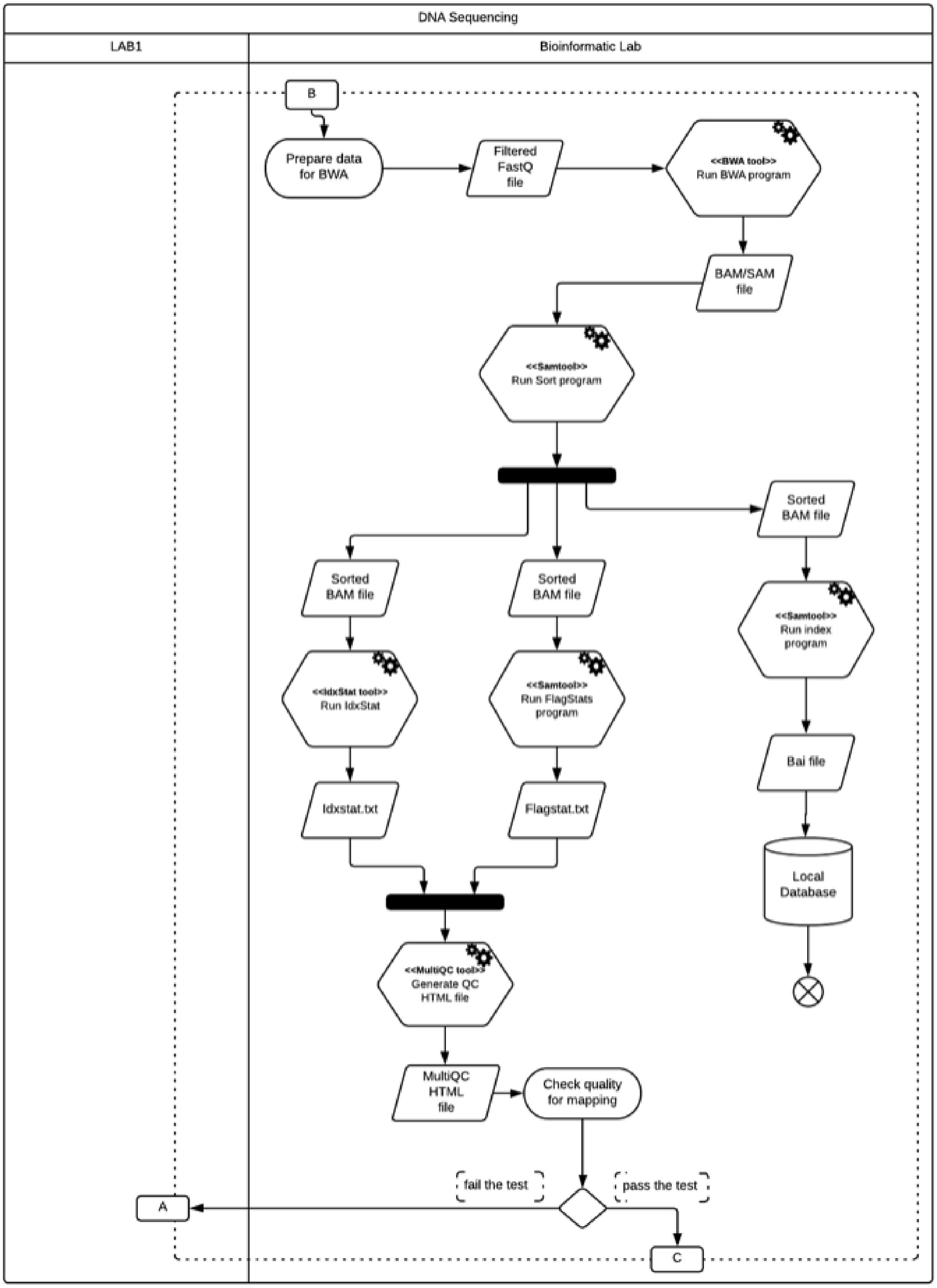
DNA Sequencing	**Version: 1.5**	**Workflow ID: 1c**



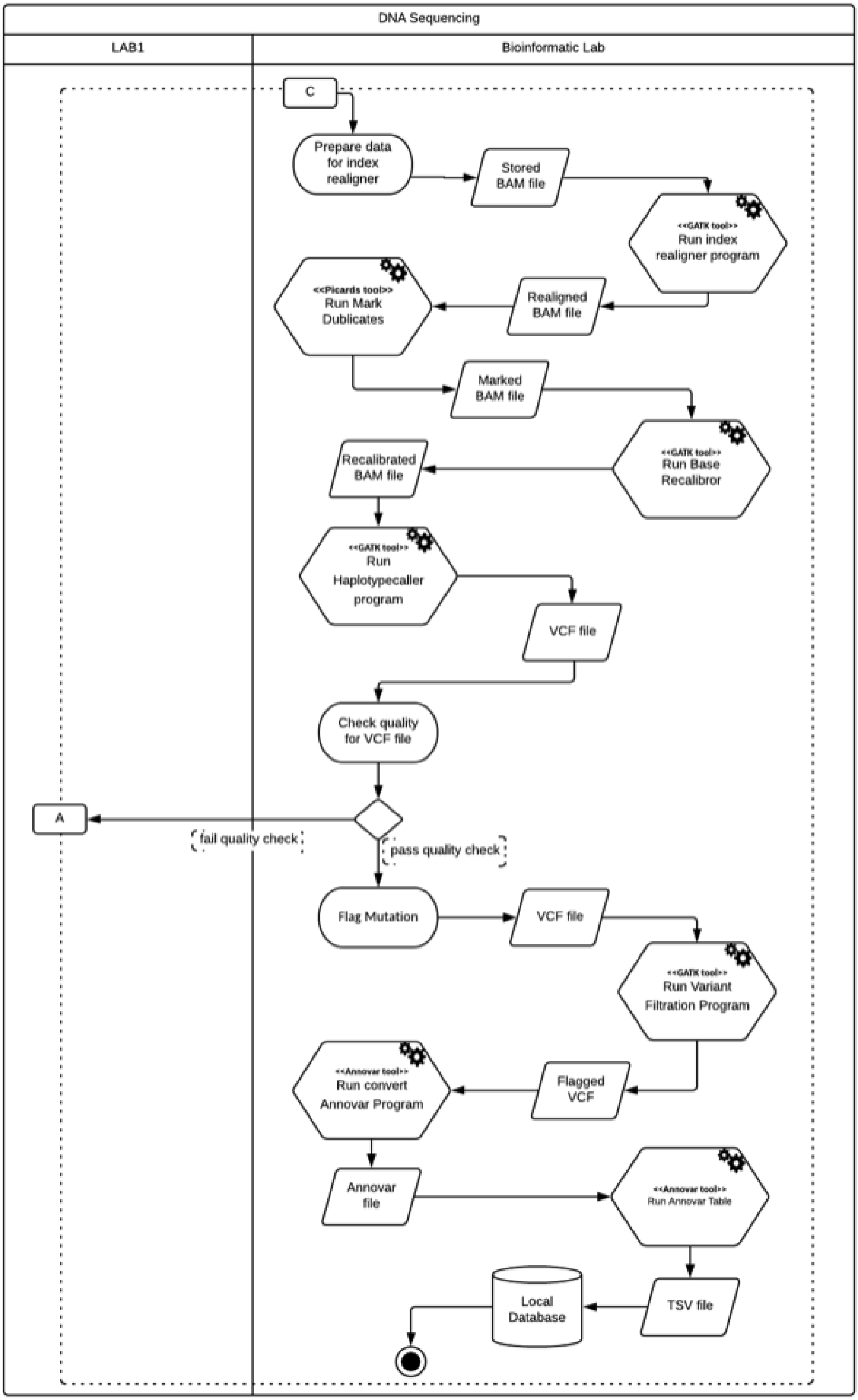
DNA Sequencing	**Version: 1.5**	**Workflow ID: 2a**



DNA Sequencing	Version: 1.5	Workflow ID: 2b



DNA Sequencing	**Version: 1.5**	**Workflow ID: 2c**


DNA Sequencing
LAB1
Bioinformatic Lab
C
Prepare data for index realigner
Stored BAM file
<<GATK tool>> Run index realigner program
Realigned BAM file
<<Picards tool>> Run Mark Dublicates
Marked BAM file
<<GATK tool>> Run Base Recalibror
Recalibrated BAM file
<<GATK tool>> Run Haplotypecaller program
VCF file
Check quality for VCF file
A
fail quality check
pass quality check
Flag Mutation
VCF file
<<GATK tool>> Run Variant Filtration Program
Flagged VCF
<<Annovar tool>> Run convert Annovar Program
Annovar file
<<Annovar tool>> Run Annovar Table
TSV file
Local Database

APPENDIX F
QUESTIONNAIRE FOR THE SECOND ITERATION

Demographic:

(For new participant only)

1. What is your role?

2. How long have you been working your role?

3. Have you created a workflow for a bioinformatics process?

3.1. If YES, briefly describe what did you do? (Language, software)

Demographic:

(For all)

4. What is your definition of workflow?

5. What is your definition of a step in a workflow?

1. We will give the XML and ask to import as a library in draw.io. 2. Draw their workflow scenario for 15 minutes and think aloud. 3. Ask the SUS inspired questions:

6. Would you draw workflows using the shapes in the library?

7. Would you use it frequently? (1 - SUS)

8. Are these concepts useful? If NOT, why?

9. What do you think about these notations complexity? (2 - SUS)

10. Are the notations easy to use? (3 - SUS)

11. Would you need training or tutorial on how to draw workflows using this library? (4, 7 and 10 - SUS)

12. Did you find any inconsistencies? (6 - SUS)

13. Are any of this redundant?

14. Is there any concept missing? If YES, what is it missing?

15. Are these notations understandable? If NOT, which is not?

16. Did you identify any field that is unnecessary or that you will never need or use? Why?

17. Did you find the notations awkward? (8 - SUS)

18. Did you feel confident drawing the workflow? (9 - SUS)

4. Provide the WDST and give them 15 minutes to fill based one of their most complex steps. 5. Ask the SUS inspired questions:

19. Would you make use of this template?

20. Would you use it frequently? (1 - SUS)

21. Do you think that this documentation is useful? Why, why not?

22. What do you think about this document complexity? (2 - SUS)

23. Is the documentation easy to fill? (3 - SUS)

24. Would you need training or tutorial on how to fill this document? (4, 7 and 10 - SUS)

25. Do you think that this documentation has a good flow? (5 - SUS)

26. Did you find any inconsistency? (6 - SUS)

27. Is this documentation missing anything? If YES, what is it missing?

28. Do you understand it? If NOT, Is there anything that you can't understand?

29. Did you identify any field that is unnecessary or that you will never need or use? Why?

30. Did you feel confident using this document? (9 - SUS)

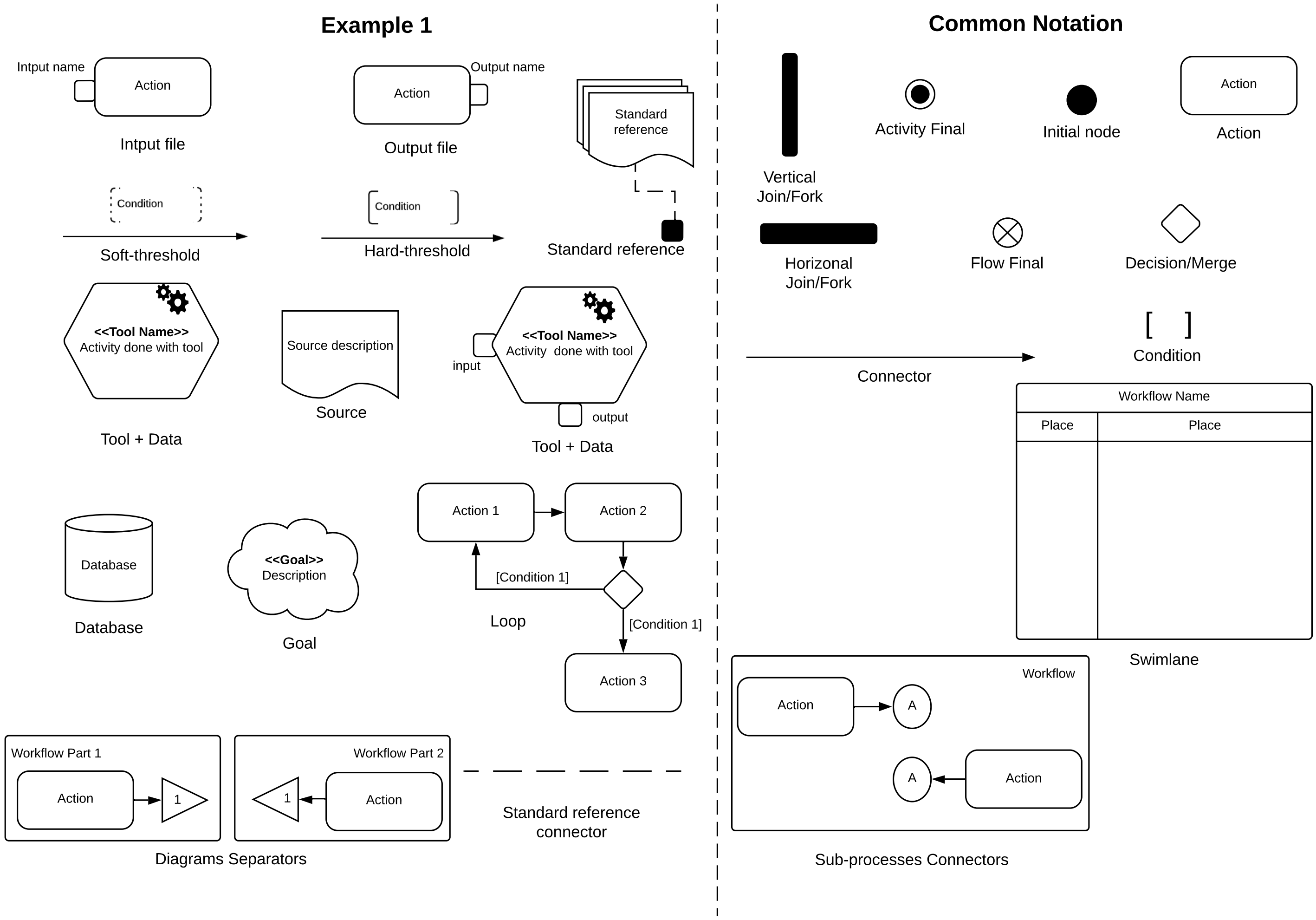
FINAL

29. For what purposes do you think you can use these artefacts (notations + document) at your work?

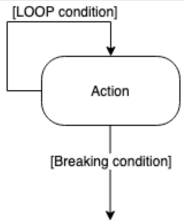
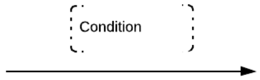
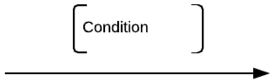
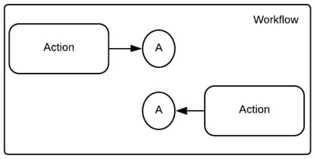
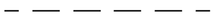
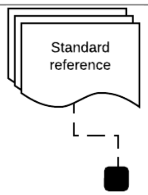
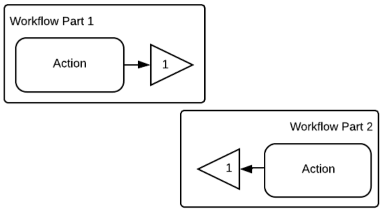
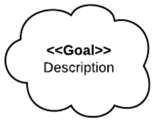
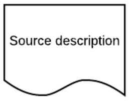
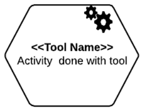
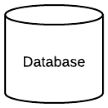
30. Who would use it?

31. What is your general impression about the artefacts?

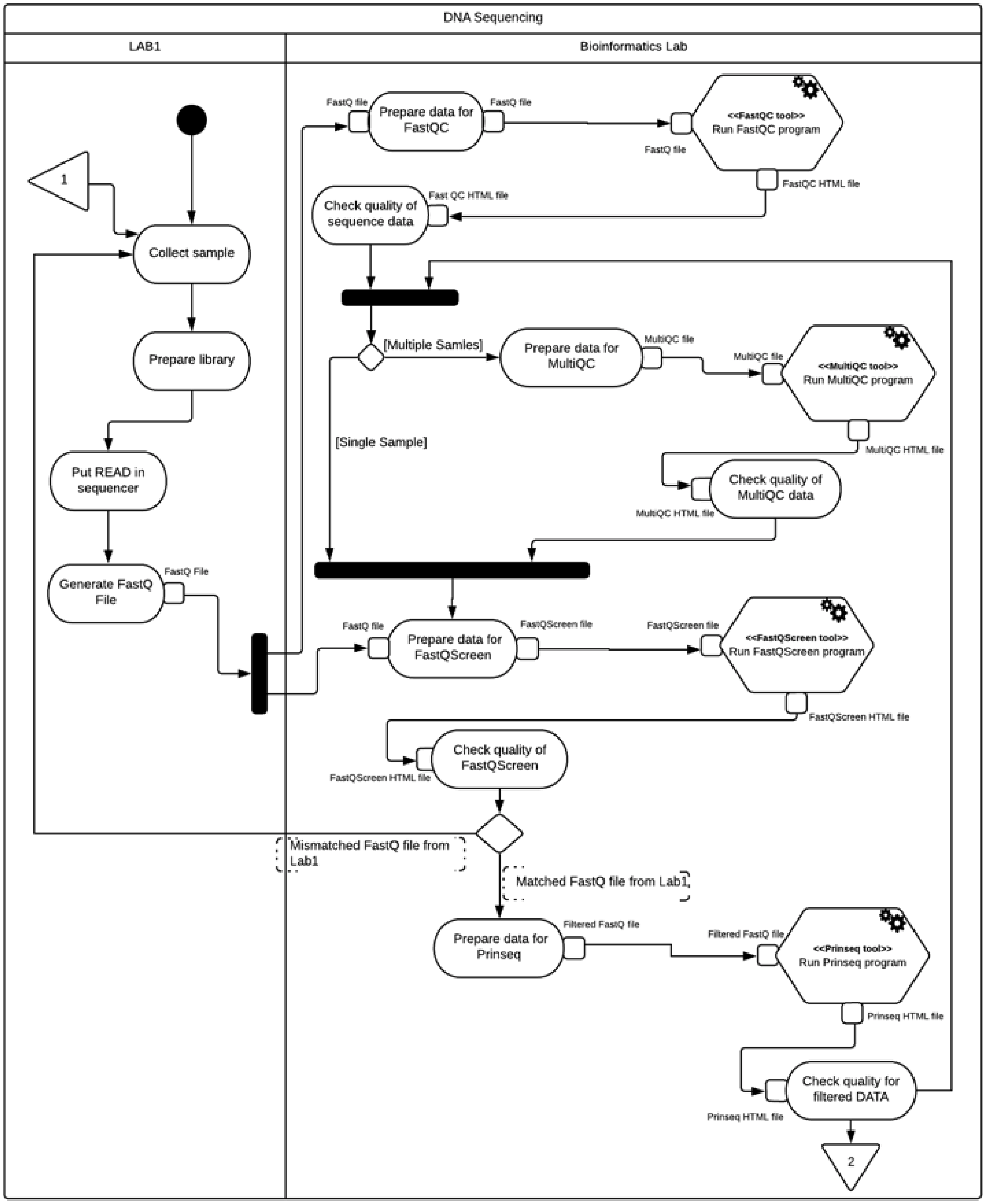
Appendix G
Legend for concrete syntax used in the Second iteration

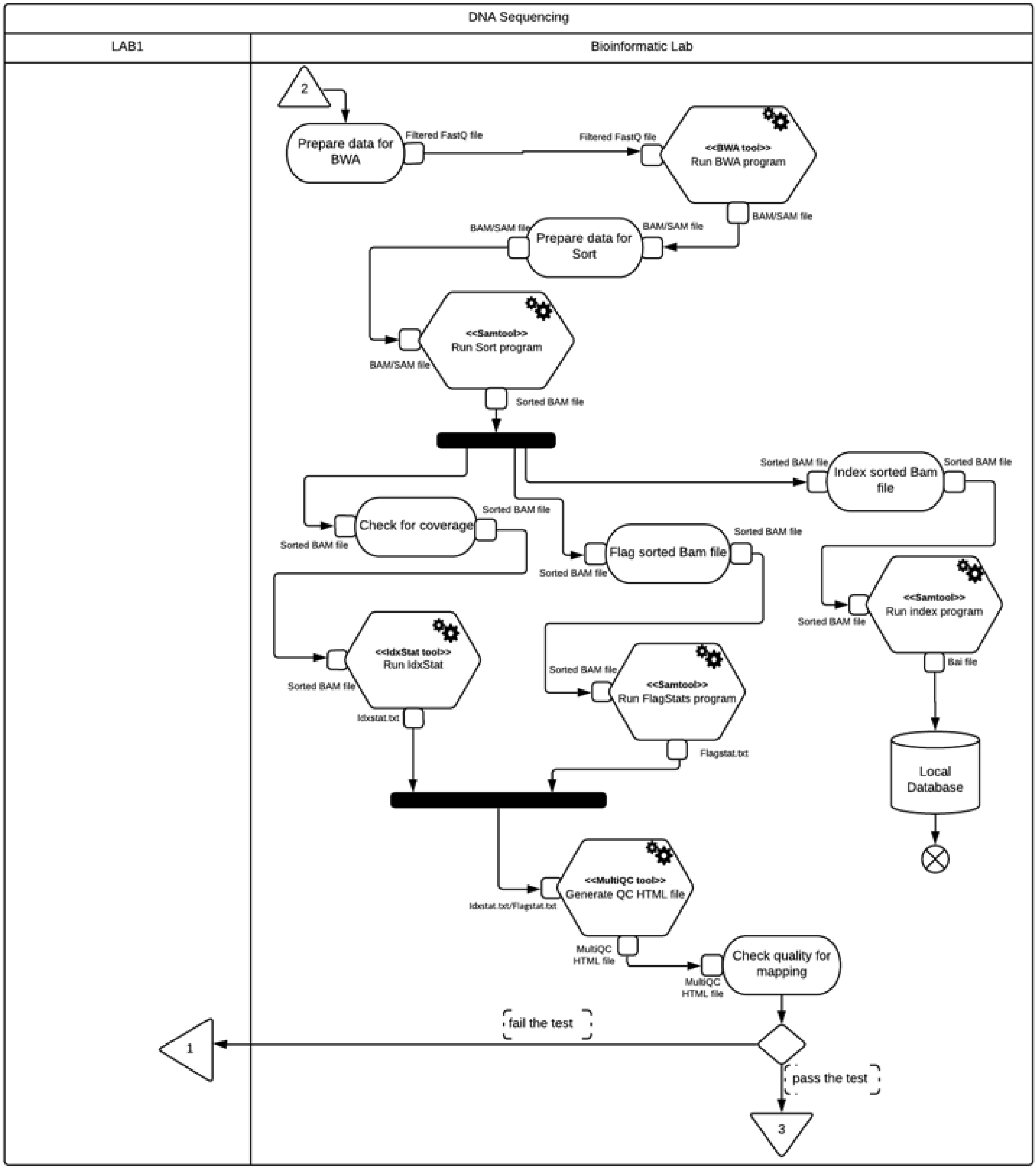


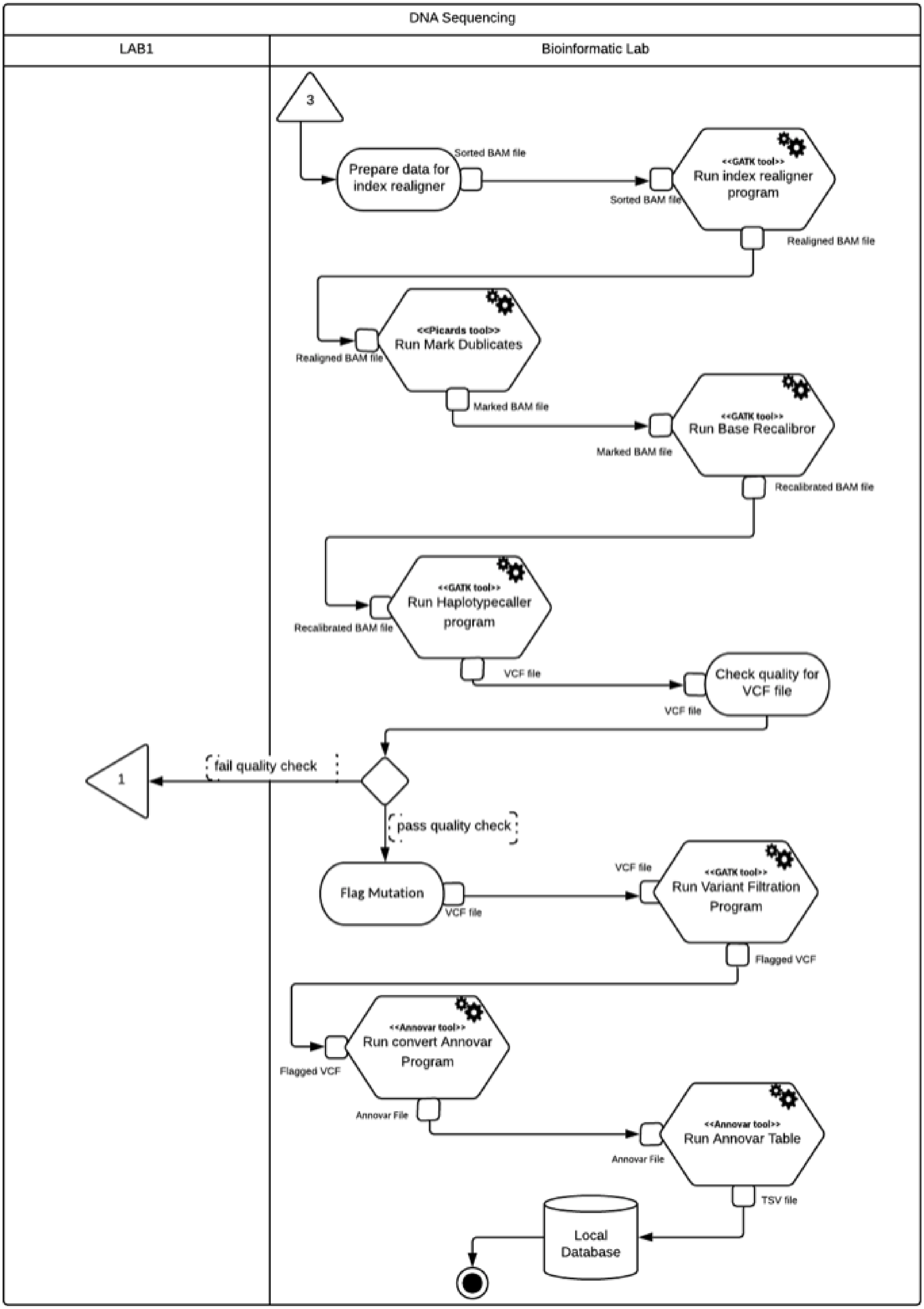
Appendix H
Concrete Syntax for the Second iteration

Name	Base Class	Description	Notation
Loop	ActivityEdge	An iterative set of activities and actions until reaching the defined condition.	
SoftThreshold	ActivityEdge	A condition with a limited soft-condition value, which is used for test outcomes. The condition is predefined within dashed guards on the outgoing edges.	
HardThreshold	ActivityEdge	A condition with a limited hard-condition value, which is used for test outcomes. The condition is predefined within solid guards on the outgoing edges.	
Sub-processConnector	ActivityEdge	A connector between the sub-processes parts within the same diagram.	
StandardReference Connector	Activity Edge	A connector between the dark input and the standard reference notation (multiple documents).	
StandardReference	ObjectNode	Data, usually a standard, that are used for comparisons, such as the human genome.	
DiagramSeparator	ObjectNode	A labelled triangle that represents the connection point with another part of the diagram from another page.	 Workflow Part 1 Action 1 Workflow Part 2 1 Action
Goal	ObjectNode	The aim of a specific activity.	
Source	ObjectNode	A link, document title, or person's name, which is the source for a specific set of actions.	
Tool	ObjectNode	Tool used to perform an activity with its description.	
Database	DataStoreNode	A structured set of data that is accessible in various ways.	

Appendix I
DNA sequencing examples for the Second iteration




DNA Sequencing
LAB1
Bioinformatic Lab
2
Prepare data for BWA
Filtered FastQ file
Filtered FastQ file
<<BWA tool>>
Run BWA program
BAM/SAM file
BAM/SAM file
Prepare data for Sort
BAM/SAM file
BAM/SAM file
<<Samtool>>
Run Sort program
Sorted BAM file
Sorted BAM file
Index sorted Bam file
Sorted BAM file
Check for coverage
Sorted BAM file
Sorted BAM file
Flag sorted Bam file
Sorted BAM file
Sorted BAM file
<<Samtool>>
Run index program
Sorted BAM file
Bai file
<<IdxStat tool>>
Run IdxStat
Sorted BAM file
Idxstat.txt
Sorted BAM file
<<Samtool>>
Run FlagStats program
Flagstat.txt
Local Database
<<MultiQC tool>>
Generate QC HTML file
Idxstat.txt/Flagstat.txt
MultiQC HTML file
MultiQC HTML file
Check quality for mapping
fail the test
1
pass the test
3


DNA Sequencing
LAB1
Bioinformatic Lab
3
Prepare data for index realigner
Sorted BAM file
Sorted BAM file
<<GATK tool>>
Run index realigner program
Realigned BAM file
Realigned BAM file
<<Picards tool>>
Run Mark Dublicates
Marked BAM file
Marked BAM file
<<GATK tool>>
Run Base Recalibror
Recalibrated BAM file
Recalibrated BAM file
<<GATK tool>>
Run Haplotypecaller program
VCF file
VCF file
Check quality for VCF file
1
fail quality check
pass quality check
Flag Mutation
VCF file
VCF file
<<GATK tool>>
Run Variant Filtration Program
Flagged VCF
Flagged VCF
<<Annovar tool>>
Run convert Annovar Program
Annovar File
Annovar File
<<Annovar tool>>
Run Annovar Table
TSV file
Local Database

Appendix J
WDST for the Second iteration

Guide:

A workflow is considered a sequence of activities through which a piece of work passes from initiation to completion.

A process is considered a series of actions or steps taken to achieve a particular end.

Step is an individual action or activity during the process, being performed by a tool or by a person.

This is a general template in case a field is not needed or used, leave it empty.

Workflow Description Specification			
Workflow Name:	<<the workflow name or title>>	**Workflow ID:**	<<workflow identifier>>
Date of creation:	<<date in which this document was created or the workflow was requested>>	Number of process step:	<<amount of steps>>
Version number:	<<number based on modifications (change of tools, insertion of new threshold)>>	Workflow creator:	<<name>>

Workflow	
Workflow goal:	<<what do you want to achieve with this workflow?>>
Workflow source:	<< Is this workflow created locally, or it follows a reference - in that case link the reference>>
Workflow responsible:	<<person who signs the final output or who uses this workflow>>

First Process Step (Start point)	
Process step name:	<<The name of the start step>>
Step ID:	

Final Process Step (End point)	
Process step name:	<<The name of the final step>>
Step ID:	

------------------------------------- END OF PAGE 1 - START OF PAGE 2 -------------------------------------

Workflow Description Specification			
Workflow Name:	<<the workflow name or title>>	**Workflow ID:**	<<workflow identifier>>
Process Step Name:	<<the step name or title>>	**Process Step ID:**	<<step identifier>>
Date of creation:	<<date in which this step was added or the step was requested>>	Step creator:	<<name>>
Version number:	<<number based on workflow step's modification>>		

Process Step					
Step goal:	<<what do you want to achieve with this step?>>				
Step source:	<< Is this step created locally, or it follows a reference - in that case link the reference>>				
Is this the first step in the workflow?	Yes ☐ No ☐	Is this the final step in the workflow?	Yes ☐ No ☐		
Sub-process of:	<<ID of a process step>>	Super-process of:	<<ID of a process step>>		
Order of execution:	<<e.g. before y, synchronous to z>>				
Where the step happens:	<<Lab/office/department - different place than the creator>>				
Description:	<<Action performed during this step (human action - if any)>>				
Is this step concurrent to another:	Yes ☐ No ☐	If yes, step name:	<<step name>>	Step ID:	<<identifier>>
Standard references:	<<Standard / Approved data used for comparison e.g. Human genome >>				
File Input(s):	<<Necessary data to start the process>>				
Is the intput comming from another step:	Yes ☐ No ☐	If yes, step name:	<<step name>>	Step ID:	<<identifier>>
If no, what is the input origin:	<<lab, person, tool >>				
File Output(s):	<<Generated data>>				
Is the output used in another step:	Yes ☐ No ☐	If yes, step name:	<<step name>>	Step ID:	<<identifier>>

Tool		
Needed tool:	<<The tool name>>	
Tool version:	<<The tool's version necessary to run this step>>	
Why this tool was selected:	<<Reasoning or source for the decision>>	
Tool settings and parameters:		

Loop Section (Repetion)					
Is this step repeated during the process:	Yes ☐ No ☐	If yes, step name of loop start:	<<step name>>	Step ID:	<<identifier>>
		If yes, step name of loop end:	<<step name>>	Step ID:	<<identifier>>
If yes, how many times it repeats:	<<number>>	If yes, what is needed to break the loop:	<<condition to stop the repetition>>		

Condition Section (Threshold)		
Possible outcomes: <<possibility 1 (e.g. pass, fail)>>	<<possibility 2 (e.g. pass, fail)>>	<<possibility 3 (e.g. pass, fail)>>
Next step name for each outcome:		
Step ID for each outcomes:		
Condition for judgment:		
Condition is hard/soft:		

Data Storage Section			
Is the generated output stored:	Yes ☐ No ☐	If yes, the data must be stored until:	<<date>>
If yes, name of the data storage:	<<bucket name, table name, folder name>>		

Appendix K
Template for the log keeper

Date: ____/ 04 / 2019 Interview Number: ________

Missing shapes of fields: (What the participant wants to use is not there)

____

____

____

____

Task failure: (The participant feels not able to perform the task)

____

____

____

____

Annoying points:

____

____

____

____

Medium problem: (shape's or field's usage unclear)

____

____

____

____

Minor problem: (Unclear nomenclature or language)

____

____

____

____

Other:

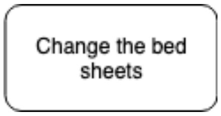
____

____

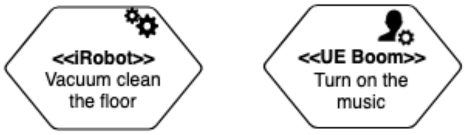
____

Appendix L
Examples for the Workshop

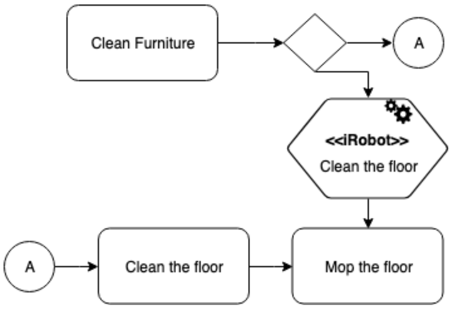
Action



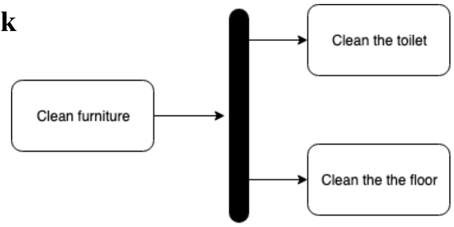
Tools



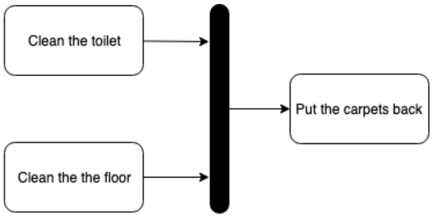
Sub-process connector



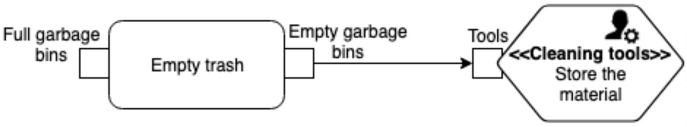
Fork



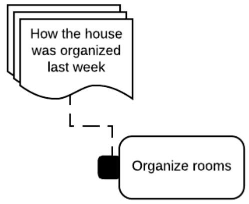
Join



Input/Output example



Standard Reference



Source

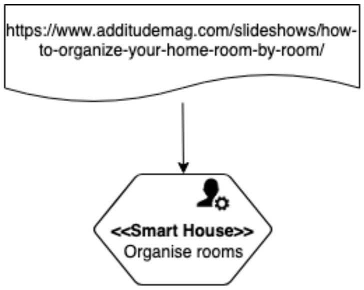
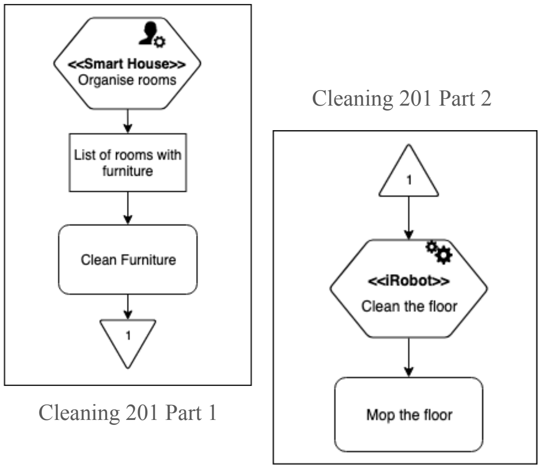
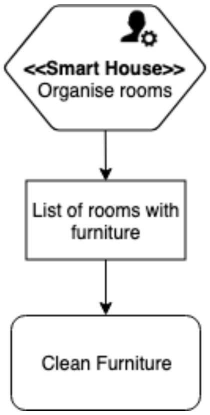

https://www.additudemag.com/slideshows/how-to-organize-your-home-room-by-room/
<<Smart House>>
Organise rooms

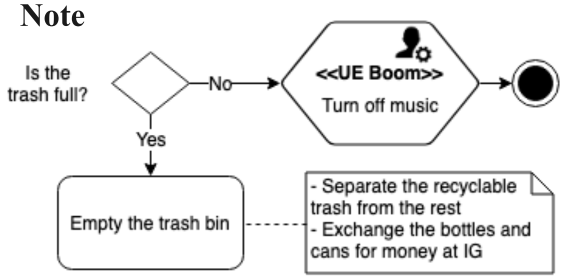
Diagram separators



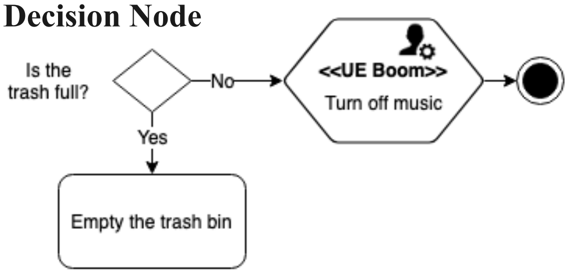
Standalone Pin



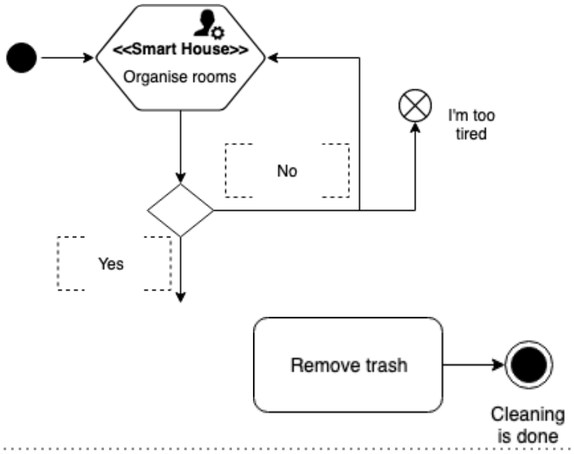
Note


Is the trash full?
No
<<UE Boom>>
Turn off music
Yes
Empty the trash bin
- Separate the recyclable trash from the rest
- Exchange the bottles and cans for money at IG

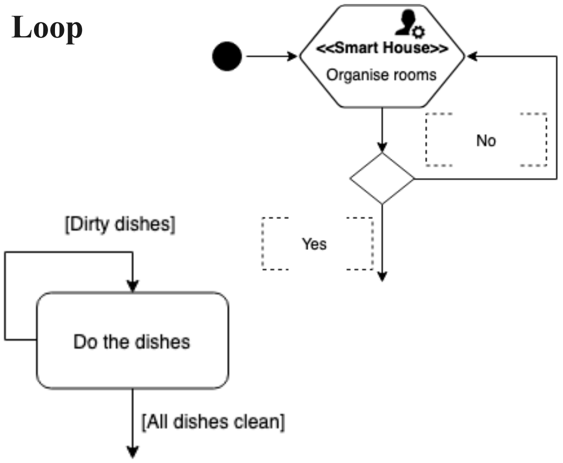
Decision Node



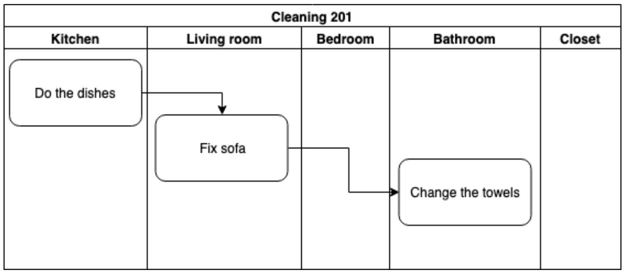
Initial, Flow final and Activity final nodes



Loop

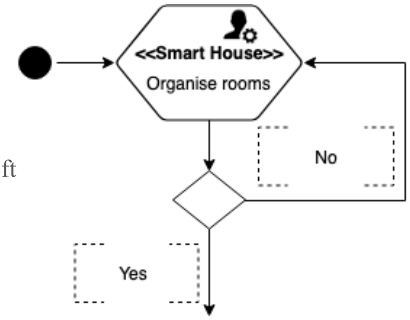


Swimlane

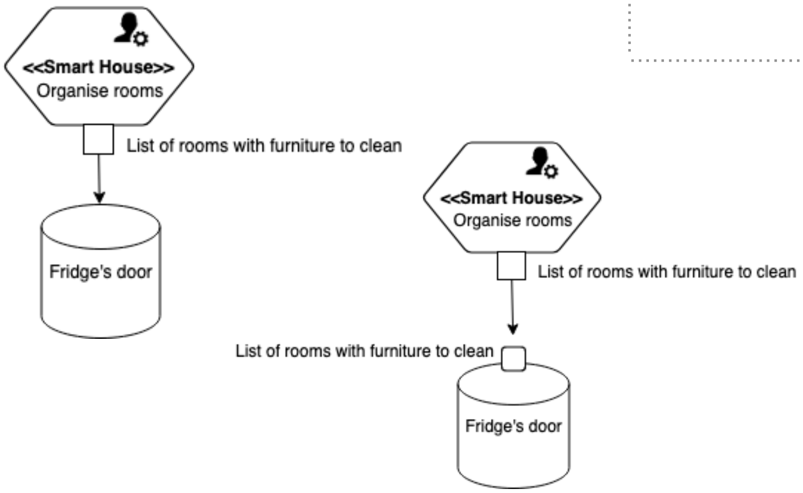


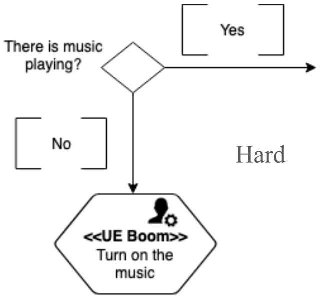
Condition

Soft



Database


<<Smart House>>
Organise rooms
List of rooms with furniture to clean
Fridge's door
<<Smart House>>
Organise rooms
List of rooms with furniture to clean
List of rooms with furniture to clean
Fridge's door



Hard

Workflow Description Specification

Workflow ID:					*Cleaning 201*
Date of creation:			*09/05/2019*	Number of steps:	*10*
Workflow version:	*1*	Modification date:		Workflow creator:	*Laiz*

Workflow

Workflow goal:	*Clean the house*
Workflow source:	*Parents examples, google best practices, tips and tricks found on https://bestlifeonline.com/cleaning-hacks/*
Workflow responsible:	*Laiz and husband*

First Step (Start point)

Step ID:	*Organize rooms*

Final Step (End point)

Step ID:	*Remove trash*

Workflow Description Specification

Workflow ID:		*Cleaning 201*	**Step ID:**		*Organize rooms*
Step version:	*1*	Modification date:		Step creator:	*Laiz*

Step

Step goal:	*Organize the rooms so cleaning can be done properly*
Step source:	*https://www.additudemag.com/slideshows/how-to-organize-your-home-room-by-room/*
Is this the first step in the workflow? Yes ☑ No □	Is this the final step in the workflow? Yes □ No ☑
Sub-step of:	Super-step of: *Clean furniture*
Order of execution:	*It is the first step, and should happen first in any room of the house*
Step execution' location:	*Living room, Bathroom, Kitchen, Bedroom, Closet*
Description:	*We should put all the things in their correct place, including clothes, dishes, mail, etc.*
Is this step concurrent/parallel to another: Yes □ No ☑	If yes, step ID:
Standard references:	*How the house was organized last week*
File Input(s):	*Checklist of the rooms and items*
Is the input comming from another step: Yes □ No ☑	If yes, step ID:
If no, what is the input's origin:	*Laiz' written list on the fridge from the last clean*
File Output(s):	*List of rooms with furniture to clean*
Is the output used in another step: Yes ☑ No □	If yes, step ID: *Clean furniture*

Tool Section

Needed tool:	*Smart House*
Tool version:	*1*
Why this tool was selected:	*It helps to organise the objects spread across the floor, tables and other hard surfaces*

Tool's Settings and Parameters

Speed = 4	*Level of organization = High*	
Power = 78%	*Surface = Floor and furniture*	

Loop/Repetition Section

Is this step repeated along the workflow: Yes ☑ No □	If yes, step ID of loop start: *Organize rooms*
	If yes, step ID of loop end: *Organize rooms*
If yes, how many times it repeats: ?	If yes, what is needed to break the loop: *The rooms are organized following Laiz' standard*

Condition/Threshold Section

Condition for judgment:	*Is it organized?*		
Possible outcomes:	*Yes*	*No*	
Next step ID:	*Clean furniture*	*Organize rooms*	
Condition result:	*House organized ready for next step*	*Organize the remaining rooms*	
Hard or soft condition:			*Soft condition*

Database Section

Is the generated output stored: Yes ☑ No □	If yes, the data must be stored until: *The furniture is cleaned*
If yes, name of the database:	*Fridge's door*

APPENDIX M
QUESTIONS USED IN THE WORKSHOP

Questions for the Concrete syntax, pair discussion:

1. What did you like and dislike in the notation library?

2. If someone would improve it, in the future, what would they change? How would they do it?

3. Who could or would not understand the drawn diagrams using the library?

4. Would the diagrams usage affect the current way of documenting workflows at your facility? If so, why?

Questions for the WDST, pair discussion:

1. What did you like and dislike in the documentation template?

2. If someone would improve it, in the future, what would they change? How would they do it?

3. Who could or would not understand this template?

4. Would the template usage affect the current way of documenting workflows at your facility? If so, why?

Mentimeter questions for the Concrete syntax:

(1 to 4 is a Likert scale from 1 to 5, while 5 is open-ended)

1. How understandable are the presented concepts and notations for you?

2. How easy it is to use the concepts and notations library?

3. How likely would you use the concepts and notations in a diagram?

4. How likely do you believe a stakeholder can understand the concepts and notations?

5. Would you add or remove anything? If yes, please describe.

Mentimeter questions for the WDST:

(1 to 4 is a Likert scale from 1 to 5, while 5 is open-ended)

1. How understandable is the documentation template for you?

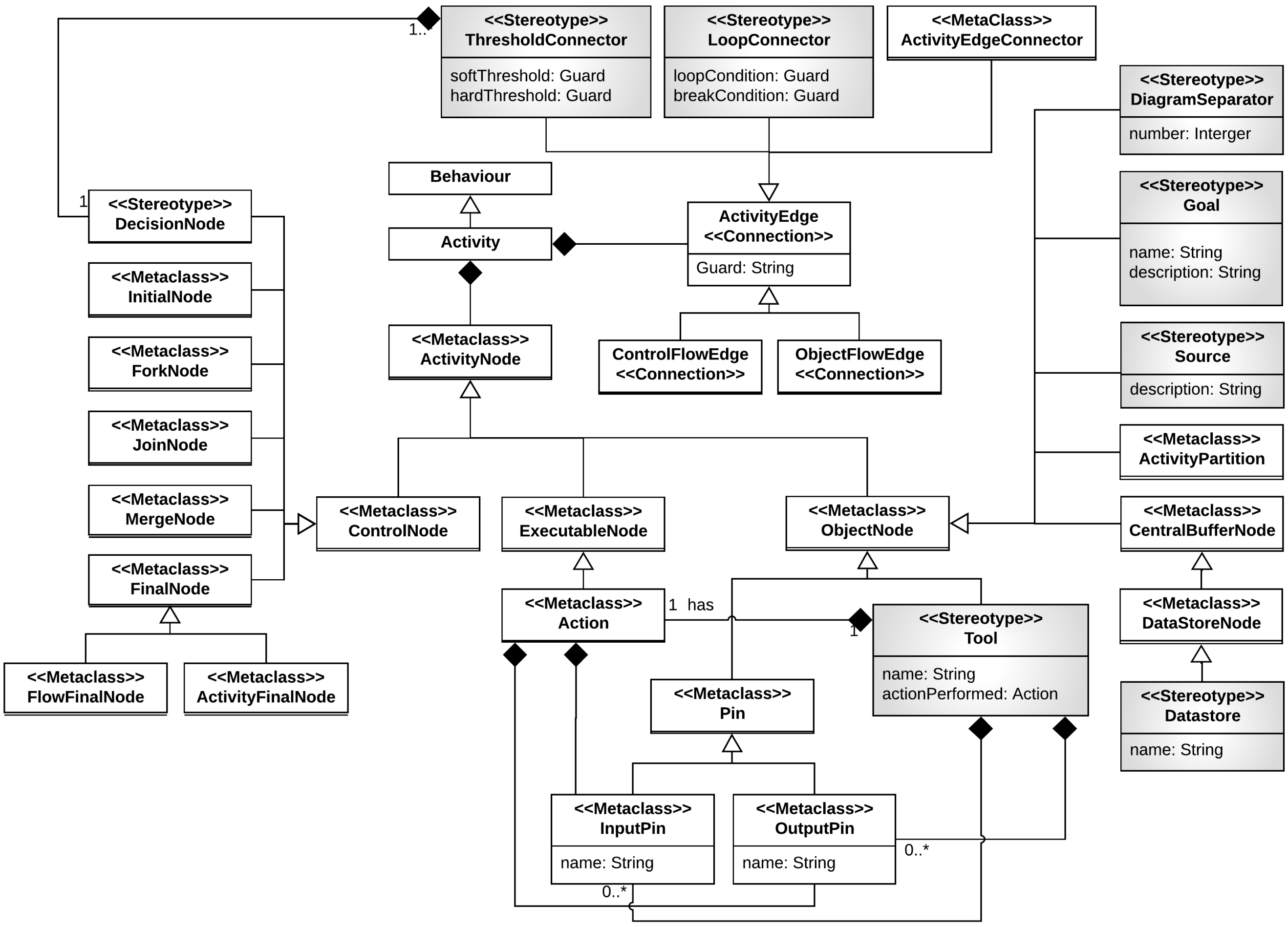
2. How easy it is to fill the documentation template?

3. How likely would you use the documentation template?

4. How likely do you believe a stakeholder can understand the documentation template?

5. Would you add or remove any field? If yes, please describe.

APPENDIX N
FIRST ITERATION EXTENDED UML AD META-MODEL

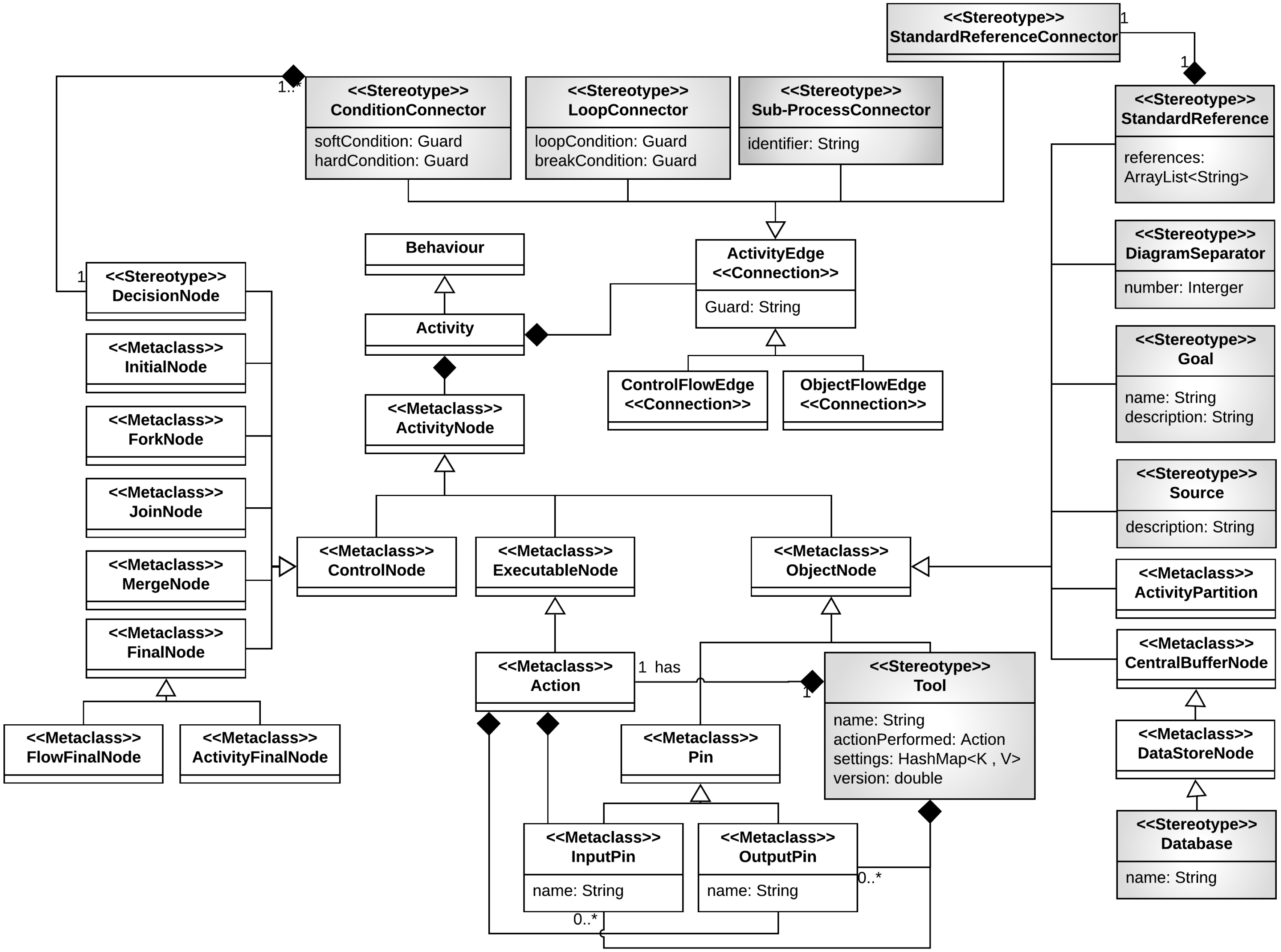


APPENDIX O
CODEBOOK OF THE FIRST ITERATION

Code	Subcodes	Definition	Sentence
WDST Improvement	*Field deletion*	A field that needs to be changed, clarified or organized in a different way since it does not satisfy the domain expert	P2 - “Step responsible and who, who conducts the step and where does that happens? I feel like a lot of the times it's going to be the same I don't if there's often people that are, I mean it was like personally responsible for one step in the workflows. Where does that happens? Would it be like in our cluster? Maybe it's um, if it's, maybe it's different for different facilities.”
	Structure		P3 - “I think this one might be quite difficult to follow. So maybe if you want to loop, do you mean loop through the tools? At first if all in one tool you will have five outputs. You run several of them is this output use in other steps. Yes, but not all of them. Maybe. Maybe just one. Then it's difficult to know which one, if yes step name then they have to. Yes. So let's see, I have several here. so i just put the next one. Um, but if it's for the example you use in tool 2 needed tool file outputs, but it doesn't say what inputs. So if I want to have from them tool one, I want to have input in tool 2” --- “we were running several tools in the same step with the same inputs. So all of them generated like five different output files and of these output files some of them were put into one process or another step. Well, some other ones what do you use as an input into another step, So different kind of steps for 2 different branches so to say, so not in one, the same workflow.”
	Understandability		P1 - “It's the step, the initial workflow point. I don't understand this.”
			P1 - “So with a step you don't mean tools, because they can be multiple tools in a step? So at one step could be one tool. But you have the option to specify more. Yeah. Because it's a part of one step.”
			P4 - “What, what do you mean with threshold here?”
	Lack of instruction		P1 - “Not all of these maybe are applicable in all cases.”
WDST Missing fields	*Field addition*	A field that is not described and the participants felt it is important to have included	P2 - “I don't know if you would want to like specify exact settings of the tools.”
			P2 - “Like maybe here is like specific settings and like if there are things here, like the reference as I mentioned, like if you need to write it need to have a lot of different inputs to the same tool, there'd be like a mess here. But then you could maybe just like have boxes are like references and like arrow there and then like see the table. And there it's like more fully described.”
			P3 - “So one thing is sometimes we do have more tools. you can parallelise your workflow and for example, for variance calling. I don't know what biological knowledge, but for for variance calling you can run several programs in the same time. If you have one file from the beginning with all the raw data and you want to process them through different tools that none one after the other, but one at one time and then merge the results together in the end.”
			P3 - “I don't see here is parameter setting, but that might not be something”
			P4 - “and also sometimes we have to say which, which version of the tool that we use.”
			P4 - “Yeah, Its just like, in some, each step there are several parameters or um, like normally when we write like a publication in the, when, when we want to publish some tool in the methods part either for researcher or for more bioinformatics method, we like say that, okay, we used this first X tool kit with the parameters this, this, this”
WDST Usage	*Knowledge sharing*	The participants' perspective of how the WDST can be used.	P2 - “sharing workflows with other people.”
	Structuralization		P2 - “to help me design it myself / useful to structure, to structure your thoughts.”
			P3 - “it's very good to have something similar to this just to create some structure around it.”
	Formalization		P4 - “we have some kind of structure like this but it's never like formalised.”
	System Documentation		P1 - "I mean it could be used for documentation. Like we have to, when we create something we have to validate it with the hospital people cause we have to make sure everything keeps a certain quality that the hospital requires. And um, yeah, it could be useful to put into their documentation system.”
	WDST Format		P2 - “I mean not in a paper format, but it can definitely be like a, I don't know, like an excel sheet or something”
WDST Users	*Stakeholders*	The people described as users of the documentation	P3 - “I think this is good for everyone that creates workflows. And maybe for the ones that are interested in using them.”
			P4 - “I think that it's useful for people that are developing workflows kind of, because people that use bioinformatics tools, they, they just like, they need to know what, how, how do you run and sometimes they have to know how to run several steps and then maybe it can be useful that they have some documentation or something like that.”
			P5 - “I think the ones that designed it. I think, definitely and there's bioinformaticians if you design it then you can use it of course.”
			P1 - “I don't know if anyone would be like looking at it, but it's, it's, I mean we have to write a bunch of stuff that I don't think anyone ever reads it. It's just needs to be there in case of someone needing to read it. But it's like a hospital bosses and things that actually validate these documents.”
WDST Current State	*Free text*	The participants' description of how workflows are currently documented	P1 - “I mean normally they want us to write like more simple something that anyone can understand it as well, like free text like this does that.”
Test of the WDST	*Test of the WDST*	The participants said that by using the artefacts they could find missing fields and improvements easily	P1 - “I would have to like try to fill it out for one of the workflows I have in order to see like”
			P2 - “I mean I think i would need to, like, try it out. I think and see.”
Notations & Concepts Improvement	*Understandability*	Notations and concepts that required further explanations or that caused confusion	P1 - “when we draw things we use a computer cluster and there are different like networks the things exists on so I like to have like a separate, okay so this is happening on our cluster and this is happening on the external server somewhere and this is like program that you run locally on your machine. So like kind of separate where it happens” (SWIMLANES)
			P1 - “So loops, you mean like, if condition, if the output from this tool does not meet the requirements, you send it back and you do something” --- “Yeah cause usually like when I write the loops I have them like contained in like a tool. So I would have like input and output. But what happens here, I wouldn't really describe loops and things in there. Oh, normally when I do things. But of course it could be. It can be useful to have.”
Missing Notations & Concepts	*Addition*	Lack of notations and concepts, identified by the participants	P1 - (data types) “it can be like some some shapes for the most common ones but they can also be , like what an option to put in if it's some lesser used that doesn't have like a shape assigned to it.”
			P2 - “I don't know if there's some workflows have a ton of like references it could be like 15 or something; like data inputs it could be like the human genome or, and some like database software. There's genetic variation and there's like five different kinds. I imagine that there is a lot of different data boxes or converging on one tool, I don't know if this would be like a data table kind of thing. Have like, uh, input data and then it's like a sort of like a table formats. Where'd you can type in the different, um, different data inputs, maybe.”

Code	Subcodes	Definition	Sentence
Notations & Concepts Usage	*System Documentation*	The participants' perspective of how visual notations and concepts can be used.	P2 - “I think we have need of it sometimes. I mean personally, we don't really use it a lot to help ourselves, but, um, if we have to document our workflows for like the hospital to put it into their like documents system. Then we have to design these things.”
	Structuralization		P2 - “I think some people would like to do with this before they designed the pipeline and use it to help them figure out how to, how to create the pipeline before they even start”
			P3 - “I like the diagrams it’s so much easier to follow. Yeah. And when you put all the inputs and output files here you have an overview in your head like this is how it's actually looks.”
			P4 - “that would probably be useful to structure a bit. Like what, what is the input and output of each step"
			P4 - “sometimes it can be good to see like a diagram also to understand what this”
Diagram Users	*Bioinformaticians*	The people described as users of the modelling language	P2 - “only going to be bioinformaticians.”
			P4 - “Yes, I think so. Yeah. Because it's, it's um, like, because I'm going, I'm working on this pipeline. I'm going to to write the documentation, like what, what you should do in each step, but sometimes it can be good to see like a diagram also to understand what this”
	Stakeholders		P5 - “Everyone that creates workflows, I think they can use it. Definitely.”
Notations & Concepts Current State	*Box and arrows*	The participants' description of how workflows are currently represented	P1 - “I mean we work, we make workflows that are like look like diagrams in a program called CLC where you have different tools and you have like an input and you just draw an arrow to another tool and output from that to another tool.” — “I'm usually just drawing like each program has a box and then an arrow and then the name like file on the Arrow and then to another box.”
Notations Preference	*Loop*	The selected notations and the participants' reasoning	P1 - 2a “because for me, I, I I don't, I wouldn't think of it as like a loop when I hear loop. I think of like on Arrows, I guess. More like a for loop.”
			P2 - 2a
			P3 - 2a “I might like these arrows just to know exactly where the loop ends and where it starts. Maybe it's a bit difficult. I like having these, what's included in the loop and you do know that. Yeah, we get you have it here. It should be, but it's not as easy to follow from exactly from where it starts and where it ends. um, that could be quite confusing here. 1a is more beautiful, But 2a you can actually see and follow, where it breaks and where it starts again, where the loop goes.”
			P4 - 1a “I think this is more clear, this like the inner loop here. Hmm. All or . Yeah, I think, yeah, I think this the left one.”
			P5 - 2a “because it's more familiar. So then that's why I think it's easier because we were used to all these arrows back and forth. Okay, so you're quickly see that then it goes, where it goes.”
	Thresholds		P1 - 1a “I don't mind either way of putting it. Actually. Maybe this one is a bit clear. When you have those, the two like in this case with a hard or soft thresholds. Okay.”
			P2 - 2a
			P3 - 2a “So visually I think this one is better”
			P4 - 2a “Maybe 2a, but I am understand both”
			P5 - 1a “I like the Idea of it like that.”
	Input/Output		P1 - 2b “I wouldn't mind this one. If it takes the inside of this box and I like it when it's on the side here. so, in that case I would like this one or I mean as as long as the actual type, is always in the same shape kind of. So here you have the file name. So BAM SAM. Well the way it looks here otherwise like I don't, I wouldn't mind it if it was connected with this one for example. So I don't think this Arrow is really needed. Maybe, So either if this one was bigger and the text was inside of it or.”
			P2 - 1b “I like these smaller boxes. I mean it's um, it makes it less cluttery”
			P3 - 1b “I like this idea. That's what I thought about when I looked at this one like input and the box i never seen it before and but I think it's good. This is more what I've seen before.”
			P4 - 2b “I prefer this one. What the, yeah, the because here is the same twice, right? Yeah. Yeah, because it did in this one it's more clear that the output from this step is the input to the next step.
			P5 - 1b “Because it doesn't take that much space. I mean I think it would go for this one if people start using them. so you get used to it cause I know how it is when you're fit. This takes up much more than that.”
	Datastore		P1 - 2b “this one, you know, it looks like a stack of disks.”
			P2 - 2b “stands out more compared to the other”
			P3 - 2b “Familiar with this one.” -- “So this one, this is all going to printed in my head as a database.”
			P4 - 2b "This one was just because I more used to it”
			P5 - 2b “because I’m used to it.”
	Tools		P1 - 2c “I like this one”
			P2 - 2c “I preferred the tools. I mean the the gears”
			P3 - 1c “I like this one better but of course it's easier if you just see it quick and wants to know what, what do I need to install, it depends on who you are, who you are, who's going to look at this one. Okay. Because if you are someone that are not going to use, to install and doing things that I think this one is better because it's easier to just see. But, but I like this one better.”
			P4 - 2c “this is more clear with the gear”
			P5 - 2c “I like this one, it's quickly seen.”
	Diagram separators		P1 - 2a “I like this way more these ones look a bit big with the number I also like the dotted lines are like this one is all included in, okay”
			P2 - 1a “I mean these are more clear obviously like the triangles. Just say speed up. What did their different, more different compared to like, I mean this is a box and these are all like box like things whereas this is a triangle, which is the only triangle that's in the graph. So that helps.”
			P3 - 2a “I think maybe I think this one is more is beautiful, but both are equally are good at following”
			P4 - 1a “no it's just, it's because it's like a different symbol that the other ones, so it's clear that it should be, it’s almost an arrow here.”
			P5 - 1a “because it's easy, I think it's easier to follow it because you can see it in the arrows where it goes, cause I don't really understand the fence.”
Artefacts Usage	*Redundancy*	The redundancy between the artefacts	P2 - "And then I think like there's so much here that's, that would be redundant when you're using this.”
	Order	The order of artefacts usage	P1 - “you will draw the diagram and then after fill this, yeah. And I would use the diagram for filling this”
			P2 - “I'd use the diagram first.”
			P3 - “I would definitely go with diagram first and that was writing this one instead of the opposites.”
			P4 - “I think draw the diagram first and then specify first the steps.”
			P5 - “I will do to the diagram to get the overview and then fill it. Yes, I would. then you have visualised it how it looks like and it's easier to fill it I think”

Appendix P
Second iteration extended UML AD Meta-model



Appendix Q
Codebook of the Second iteration

Code	Subcodes	Definition	Sentence
Workflow Definition		The definition provided by the participants for WORKFLOW	P1 - “I mean it's synonymous to a pipeline, or maybe I should actually explain what, what I mean by pipeline. I mean just like a scheme of which software is running in which order. For like a package of software that are online or in parallel or in sequence in different constellations.”
			P3 - “From an input file going through different tools ending, uh, ending up with output file which I'm looking for and this can be like a huge number of different tools and processes.”
			P5 - “Something that you can quickly see how you run the program. That's my definition, I guess a summary of what you have.”
			P4 - “So I would say some computation of workflow that has an inputs and that has an outputs, and can consist of several intermediate steps.”
			P6 - “Aaa it's a process to follow through project.”
Step Definition		The definition provided by the participants for STEP	P1 - “It could be either a program being run or some kind of script or a conversion to another file type or it could be moved over the network. Basically. I guess the step would be like some file changing shape, or being transferred to another computer.”
			P3 - “It depends, but different steps would be the different tools, I want to say eh. And if they are paralysed I would say that this step is still the same, but maybe the workflow goes in different directions but it still the same step, if they do same kind of things.”
			P4 - “It's um, probably as well something that takes some files or something as an input and produce something as an output and either intermediate files or something”
			P5 - “like input, output or tool, i'm not sure”
			P6 - “Um, what you're going to do at certain points.”
Notations & Concepts Improvement	*Dislike*	Notations that the participants commented on how to improve or in what the problems consist of	P3 - “These text boxes are like far away from.”
			P5 - (tool + input/output) “What is the best way when you have it like this? Because it will go really, if you have deiue deiue deiue then it will go really in diagonal.or can I spin it around or something or have it like this docs up.”
			P6 - “Maybe write a text to say this is done by a human or when you say I'm doing some or if you have a tool, you are using and you say the name of that tool. People will know it's a tool. If it's done by a human, people know it's done by humans. I don't think necessarily to come with different shapes.”
	Understandability	Notations and concepts that required further explanations or that caused confusion	P1 - “What's the difference between tool and action?”
			P1 - “So process end and start, is that just like the end and start of the whole thing?”
			P3 - “Where do I write the condition?”
			P3 - “what's an action?”
			P3 - “the workflow name place 1, place2, place 3, I don't about that one either.”
			P3 - “the standard reference, I don't know what that is or how to use it.”
			P4 - “I want 2 input files. How do I do that?”
			P4 - “can choose here if you want to use the tool or action. So I think it's a little bit what, what is the difference?”
			P6 - “What's this one I don't understand?” (end flow)
			P6 - “what is this shape is it like a different tools we will use them?” (tool + input/output)
			P6 - “these 2 are the same?” (hard and soft condition)
			P6 - “If you're using a tool it is also an action, right?”
Missing Notations & Concepts	*Additions*	Lack of notations and concepts identified by the participants	P1 - “There's no file database with a box.” (input/output)
			P1 - “I would like different kind of arrows. Like, cause sometimes it's files that are moved somewhere and sometimes there like files are just in place but they're just used in another software.” --- “I usually want to display kind of how the data moves around on our physical cluster, like different computers and so on.”
			Researcher “Anything missing apart from the parallelogram thing?” / P5 - “No, I don't think so either. Okay. It's usually no.”
	Nothing identified	No missing notations or concepts	P3- “No, not what I see. No, probably not”
			P4 - “Hmm, no.”
			P6 - “I think, for me, it's quite a quite complete.”
Unnecessary Notations	*Unneeded*	The participants do not see usage for these notations	P3 - “Goal, description. I would never use this goal I think.”
			P5 - “I'm not sure. For me it's like the goal and maybe the note, because I would write that outside of the workflow, but it depends because we are doing it for publications then you don't want those. But it might be for others. So that's just for me.”
	Unfamiliar	Due to unfamiliarity with the language, the participants would remove these notations	P1 - “this vertical join/fork thing. I'm not exactly sure. I would probably just do many arrows pointing to one tool or something like that.”
			P6 - “oh, this horizontal join/fork.”
	Nothing identified	The participant did not identify any unnecessary notation	P4 - “Yeah. No, I, I think there's no”
Notations & Concepts Usage		The participants' answers to the query: would they use the notations and concepts at their work?	P1 - “Sure. Yeah why not.”
			P3 - “Maybe, because the why I would use it is I still think it's nice to have the inputs and outputs. Why I don't think I would use it, is it takes time to do it and all the text things are a bit far away from the actual boxes.”
			P4 - “Yes. I. I think that that is good. Um, the, I, I, I, yeah, I, I would use that. I think.”
			P5 - “Yes. I think I definitely liked the tool ones, which I only used almost.” --- “Frequently, Probably not because we usually don't write the workflows. I mean if we need to, we do it for publications, but usually it's just text, like we did this and this and this.”
			P6 - “Yeah. I actually use this website, as well. --- The work we do, its quite standard so we have kind of the workflow in our mind. We actually have, um, uh, when we do scripting, we have the report you can see what kind of workflow we have from the reports. That's why we don't use it that often.”

Code	Subcodes	Definition	Sentence
Notations Complexity		The participants' answers to the query: are the notations and concepts complex?	P1 - “No, they are not.”
			P3 - “No, they're not complex. I mean, it's just as few as we need they are not more than what we actually do need. Um, so when you get used to knowing, which boxes you will use or the arrows you will use, then it will probably be easier to use them.”
			P4 - “No. They are very nice and clear. --- except for the drawing program not so easy to use, but the the shapes are clear”
			P5 - “No, I mean there's some like the input and output, for instance, I think because that's just a square and usually you have like a form for a file, which I don't think are found or?” -- “but It a parallelogram for a lot of the files usually?” --- “I'm used to it like that and I did that once”
			P6 - “Yeah a little bit.” --- “If somebody learns this quite well, I would say it's quite straightforward, you know. The shape, the different shapes represent different procedures or, yeah, but cause cause the flowchart you want to show also the other people, that's why you have this one, right. when you show it to other people If they don't learn these tools and they don't understand the shapes, It may be difficult for that to follow the flow.” --- “Cause there are too many shapes it makes it a little bit difficult to use.”
Notations & Concepts Tutorial Necessity	*Descriptive Manual*	The participants' answers to the query: would a person need training or tutorial to learn how to use the notations' library and its concepts?	P3 - “No, I think it's good to have like just some kind of uh, paper telling me, okay, so this is is, and this is not this and getting the experience. I mean it says quite clearly on them what it is.”
			P5 - “At least like a manual or something that you can follow, I think”
	Familiarity		P4 - “I will learn just by using it.”
	Training		P1 - “Yeah. yeah.” --- “I don't know which ones to use in which occasions. I mean, usually I would just draw a box. like a box for everything and then try to adobe it after I'm done maybe.”
			P6 - “yeah, training is definitely useful.”
Confidence to use the Notations & Concepts		The participants' answers to the query: did you feel confident using the notations and concepts while drawing?	P1 - “Yeah. Umm, you only look at the shapes and things here. But, I mean, draw.io is a little bit like junk.”
			P3 - “It's new, definitely. Uh, and I have to, but that's, that's draw, probably, takes a lot of time moving the boxes and things coming up in the wrong directions, and I don't know.”
			P4 - “Yeah. That's fine.”
			P5 - “Beside that It's fine, but it's difficult as you say for draw.io I, I think.”
			P6 - “Yeah, I think so.”
WDST Improvement	*Annoyance*	The participants mentioned something on the template that annoyed them	P3 - “This took long, It just keep going”
			P3 - “all these ID needs, it's a lot of them. They are probably consistent, but in my mind, there's a lot of them. And, and it's probably good if you do a really complex workflow then you need them.”
			P5 - “it's a lot of writing the same thing I think there.”
			P6 - “Ahh, you have the workflow and the different steps, you better if like the first, uh, description part, you can automatically link, so you don't have to fill it again.”
	Understandability	Fields that required further explanations or caused confusion	P1 - “I don't know, what to put in here” (process step ID)
			P1 - “Concurrent to another. Hmm. I mean, not in this workflow, but should this be in relation to like the start point cause the start point has to be triggered by something.”
			P3 - “Where should I put that source?” --- “the step source. You will have differences. Sometimes it changes, but if you have the tool and the version, you can always find where you can download it, where you get it.”
			P3 - “Step ID for me it is the same”
			P3 - “Process step name. Process step Id. Its the whole, I mean you said the process was all of it, so then it's Glenn. Process step name, fine step one. This is a bit confusing”
			P3 - “Super Process, ohh Nice, I don't know what it is, but sounds great.”
			P3 - “what was hard and soft conditions?”
			P4 - “where this step, Okay, this I don't understand.”
			P5 - “workflow ID. Ah what is that?”
			P5 - “step ID still not sure what it is.”
			P5 - “Date of creation. It's, it's same as before. or is that suppose to be like when you create that step?”
			P5 - “Super process, Eh, I don't know what that is.”
			P5 - “Is this concurrent to another?”
			P6 - “So you have different steps here or?” (First and second steps)
			P6 - “But these aren't the same?” (the header)
			P6 - “What does this mean?” (Order of execution)
	Fixing	The participants identified a mismatch in the template's pages, a wrong meaning, or a wrong field	P1 - “Here it says process step name and step ID and here process step ID.”
			P5 - “Where the step happens. Office I guess, I don't feel that I'm, that I don't feel why that would be of interest to anyone.”
	Format	The participant said something related to the way the template was provided	P1 - “Is it meant to be like in an excel?”
WDST Missing fields	*Nothing*	The participants think that nothing should be included	P3 - “No, nothing is missing.”
			P4 - “No. Not what I can think of.”
	Field addition	An undescribed field that the participants felt it is important to have included	P1 - “Maybe like a description for what we do if something happens. Like here and step condition for judgment, we have a condition, but we don't say what is done as a result of that condition. So, like here we would collect the standard error and email ourselves or something.”

Code	Subcodes	Definition	Sentence
WDST Content Flow		The participants' answers to the query: do you think that the documentation template has a good flow?	P1 - “Yeah, sure.”
			P3 - “yes, it's actually kind of good flow, it is. You have the right things in the beginning and you're going through the steps in a nice order. Yes. it’s a good flow.”
			P4 - “yes, yeah.”
			P5 - “yeah, yeah, I don't think you can change the order of things.”
WDST Usage		The participants' answers to the query: would they use the documentation template at their work?	P1 - “yeah, we had to, sure. I mean we don't like these kind of documents, but yeah. If someone tell that we have to. Sure. --- I mean, as frequently as i have to.” --- “It makes it easier than just writing free text.”
			P3 - “No, I wouldn't use it I think. I don't think I need it in the workflows I make today, I usually, I only do the scripting directly. I don't draw it or write it down if I do, if I need, if I need to, why I usually draw something, it's because I need to explain it to someone else or if I, it's kind of complex. So in order, for me to make the workflow, I need to write it down to get my head around how it actually works. So using this one with all text, no, because I wouldn't get my head around how it works really.”
			P4 - “Yeah. Yeah. Yes. I think so. Um, but yeah, if it's like, I guess that most people are lazy. So if someone doesn't specifically ask to documenting this way, then people will just documenting in their own way.”
			P5 - “I don't think so.” --- “because it's so much to fill in and I think for reading it, it would also be too much. I think for the workflows. You want to have an overview and quick see aha I recognise this tool and this tool, but if you have, like if I do this for the entire workflow, I think it would be like 10 pages and no one has the energy to do that.”
			P6 - “No, it's so complicated. This may occupy more time than if i just run the script.” --- “Hmm. Yeah. In a way, yes. If you, if your projects you need to check in detail”
WDST Complexity		The participants' answers to the query: is the documentation template complex?	P1 - “um, a little bit, yeah.”
			P3 - “I would say it's complex. Or too much. Too much information. I don't think you need all of this, probably in other types of workflows, but nothing to ones I'm doing.”
			P4 - “No, I don't think so.” --- “I think it was clear.”
			P5 - “Yeah. It's complex , it's a lot.” --- “I think, the largest problem was for the writing, but that's usually because I i'm not used to that. I don't know if anyone is writing.”
			P6 - “Yeah, it's quite complicated, as I can see.”
WDST Tutorial Necessity	*Manual or Example*	The participants' answers to the query: would a person need training or tutorial to learn how to use the documentation template?	P1 - “It's not really needed with training.” --- “I would like to have like a template that describes exactly like examples.”
	Unneeded		P3 - “Yeah, one example sheet maybe, or, but it's kind of a nice draw already in this uhm, with the light grey. So, maybe not.”
			P4 - “No, it's the same self-explained.”
	Manual or Example		P5 - “Yeah, I think so. Or a manual or something.”
	Trainning		P6 - “Yeah definitely, but also training needs time.”
Confidence to fill the WDST		The participants' answers to the query: did you feel confident using the documentation template?	P1 - “Yeah.”
			P3 - “It's also difficult to write these type of pipeline without actually having a pipeline. Then you don't know like this process. This is a step in this process I'm just coming up with something right now. So I don't know, but I don't think it went smoothly.”
			P4 - “Yeah. I think so, this one was even more easy than drawing.”
			P5 - “No, not really, no.”
			P6 - “Yeah. I mean if you get used to it, it's not really hard.”
Artefacts Usage	*Validation*	The participants' description for which is the purpose of the two artefacts presented to them (library for drawing and documentation template)	P1 - “It could be good for us, when we work against the hospital we have to, um, we develop something and then we have to validate it to check that it does what it's supposed to and then we have to write everything down in the hospital documentation system, so it could be useful to fill out and just put in to that system and not have to make, write this free texts, which we have to do now. But everything they do there is like on Word, So it would be good if is compatible with a Word on Windows so we can just paste it and work.”
	Process overview		P3 - “They are for workflows, showing how it, how it is created, how it is running, how people should run it.”
			P4 - “maybe to get a diagram and overview of the workflow, like if you have a workflow that consists of several steps.”
			P6 - “I don't think how we'll use the tables. Yeah. Of course. The Diagram and the flowchart is quite useful for some really complicated and big projects. It's better to have a flowchart.”
	Traceability & Learnability		P3 - “In a couple of years, when I go back and I want to know, what did I do then I can see exactly using this drawings. Like, Okey I did this step, this step, this step very quickly or if another person suddenly gets the same costumer that I had a couple of years ago, um, they can see exactly what we did. What tools did we run.”
	Publication		P5 - “The template i don't know when. the diagram, I def, publication or if you just want a nice picture and a poster and everything, then I think that would be good.”
Artefacts Users	*Researchers & Tools developers*	The described people as users of the documentation and workflow diagram	P3 - “I don't think this one is only for bioinformaticians, everyone building a workflow probably.” --- “it's more for people with the same knowledge or similar knowledge, but sure. If a PHD come here and they have some experience from before or something, they can get a good understanding of we use.”
			P5 - “The researchers, tools' developers, definitely the workflow, that no, I don't, then I don't know about the template, but the diagram.”
			P4 - “I guess in, in my case it would be another researcher, uh, like someone that is maybe not a bioinformatician, but someone that has a produced some data and they want to, to use the workflow for analysing the data. So it's a, it's a PHD student or a postdoc or someone like that that, yeah. That's the the use or is, it's not, it's not the developer of the workflow but it's the user.”
	Bioinformaticians		P1 - “yeah, like me and my colleagues, like bioinformaticians, because I don't think the geneticists want to go into this much of details.”
			P6 - “Bioinformaticians.”

Code	Subcodes	Definition	Sentence
Artefacts General Impression		The participants' answers to the query: what is your general impression of the documentation template and the notations' library?	P1 - “In general it's good, you just, it's just needs a better tool for the draw.io maybe for making the diagrams. But it's form with some like better distinction between like the terms. Maybe so we what each thing means. It's good.”
			P3 - “the diagram I think it's good. I would absolutely use it with these inputs and outputs and everything and especially if I learned how to use it, but that's has nothing to do with your work, yeah, that's me. And about these documents and my general impression is that it would take more time just to fill it out and what it actually gives us as back.”
			P4 - “I think it looks nice and can be a lot clear overview. but. Like, what. One is more overview, and this one's more, the document, is more detailed of them, steps. I think it's good. --- I could say that they complement each other. Here, it's difficult to get the overview of the whole even. Yeah, it's, um, but, yeah, it depends a little bit on the the complexity like here.”
			P5 - “I think, I think it's a good idea that we keep the documentation better, because. Because there's always a problem, especially for us, when you are delivering data or if you're working with the same things. So it's, I think it's a good thing to have it really well documented so you can follow the steps. So the idea is good, but I think it's too much for the, the template. The diagram I think that's really, I think that's good.”
			P6 - “The chart is good, I would like to use it. just maybe decrease that number of shapes. The table, mhhh, I think it can be used for legal usage like If somebody is going to sue you, and this is really a good control and it goes into details so like if you're doing, for example, human data and you are maybe making a drug in the end. and, in the process when you analyse the data, you are using a different, with a wrong tool, I think, that this will be a really useful. If they say this drug, it's unuseful or shouldn't be approved because you used this tool.”

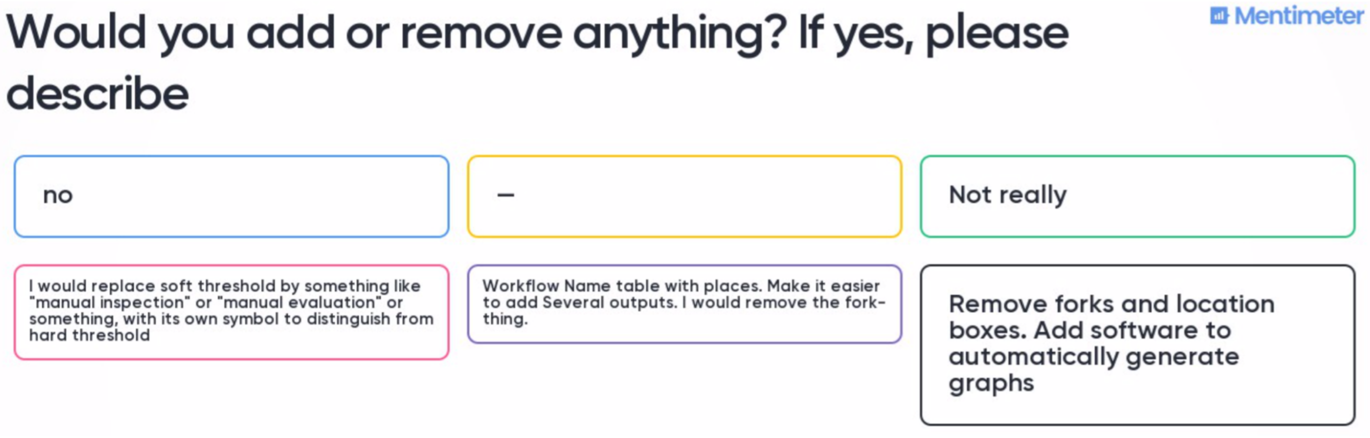
Appendix R
Codebook of the Third iteration

Code	Subcodes	Definition	Sentence
Notations & Concepts Overview	*Neat and Simple*	The participant's first impression of the language	P7 - "I think it looks pretty neat, Simple. At least this is the first time that I'm seeing it and I do understand what you're talking about. So it's simple."
	Missing parameters	The participant pointed out missing attributes in the notation	P2 - "But It wouldn't be for overview because you still don't have parameters and everything that you need to have for running and performing."
Notations & Concepts Improvement	*Understandability*	Notations and concepts that required further explanations or that caused confusion	P2 - "What is source?"
			P2 - "I'm still confused about the thresholds. Like I couldn't really imagine a scenario, where, for me everything is like hard thresholds."
	Label necessity	The participants' answers to the question: Who could or would not understand this template?	(Stakeholders) P1 - "I think they would be able to understand it. But just from the text, kind of, they wouldn't know what the shapes are like intuitively. But uh, as long as you see a few of them, I guess you would kind of make the connection. Okay. This shape is always connected to this function, but I think you would always need to have like an image text underneath to describe the whole workflow."
			P3 - "Yeah. And I also think it's a huge difference. If it's within the groups, I mean, bioinformatics are looking at it, then I think they would understand what it is. But if it's a customer or someone…"
			P7 - "For instance, in my days the one that is called source for me is printing. So things like that, so you still need to have some text or explanation if even if it's not the people that the correct people to look at."
	Provided solution	The explanation provided by a participant to another participant to soft-condition	P1 - "I guess, soft threshold would be if you do it more manual, you make, you can make an interpretation, but if you want to have everything automatically, you have to have."
	Requested software features	The participants requested another software for the implementation of the library	P1 and 2 - "I would like to have a better like program to make them in not, not draw.io or PowerPoint."
			P1 - "Where you can adjust everything by pixel for example and have like most bands in each arrow."
			P1 - "Probably believe not online."
			P7 - "Online is not a necessity."
			P3 - "What if you could automate the workflow, the picture of it, so it's just the output will be the workflow picture. That would be nice."
Unnecessary notations	*Unneeded*	The participants do not see usage for these notations	P1 & P2 - "we don't think we're going to use the forks basically because normally we just have a box and then we draw like arrows from the box and then many arrows coming to another box, so maybe someone can use that sometimes but I couldn't think of when would I use it."
			P1- (swimlanes) "I don't know. I think it's enough with just the workflow and the boxes and everything. Yes. I mean it works okey if you have a few places I guess, but if you have too many places. It can work like an excel sheet, you have to place certain boxes in certain places."
			P5 - "I was thinking of standard reference. I don't think I would use it."
			P3 - "No, neither would I."
			P2- "It's kind of the same thing as inputs file. Only, it's a bit more descriptive and I think that would be very much depending on who you are. Maybe if you want to be very clear, you would use the standard reference, but a lot of people wouldn't I think."
Diagrams current state	*Generic usage of the notes*	How the participants are doing or which notation they are using for a specific concept.	P2 - "So they're like notes?" (Source)
	Boxes for all		(Swimlanes) P1 - "I usually just have one box and I'd just put something, ah, this is like on our server and this is not done or something."
	Documentation without purpose	The described problem of not knowing why they produce documentation and who is interested in it	P1 - "Which I don't know how someone looks at it one time. Maybe."
			P2 - "People don't look at them. They're just supposed to be there."
Test of the Library	*Test of the Library*	The participants said that by using the artefacts, they could provide better feedback	P2- "I think we need to test it, first."
			P7 - "Exactly."
Notations & Concepts Usage	*Secondary for final documentation*	The participants see it as a final documentation step, where they would sketch first and then change the notations	P1 - "I would probably, like, do like a basic one first with just boxes and then try to replace things like stilly to be more proper maybe after, so like that have like a draft and then replace them after."
			P7 - "Yes, I do totally agree with you. So I can see these kind of, whenever you have a project, you do a draft, what are the first steps that you are going to work with, the first step like that and then whenever you're working with it then then you add the different steps. So you can be seen like a, how to say that, overview of the things that you have to do. So it could be close, kind of a checklist of things. And of course whenever, it's documentation right, So whenever other people come instead of reading or looking at the scripts, step by step. This is, at least you know what they're doing and then you focus on specifics lead."

Code	Subcodes	Definition	Sentence
Library usage effect on the current state	*Increase Time-spend*	The participants' answer to the question: Would the template's usage affect the current way of documenting workflows at your facility? If so, why?	P1 - "Would increase the time spent on making them"
			P7 - "It will be a lot of time making them at the beginning until you really get the hang of it."
Notations & Concepts Usability	*Guide*	The participants' beliefs in how they could use the concrete syntax and which purpose	P6 - "It can be helpful if we have like a standard project that we do over and over again and then we put some effort and we make a good flowchart. And if somebody is doing, something else, gets the project or there's a new person coming into the group and do the same product, then it's kind of like a checklist to follow. So you know, then, they will know where to start, what's the next step and then, yeah. So you don't miss anything."
	Document standard projects		P7 - "However, I don't see it to be useful in projects that are not the standard, because since we are switching a lot done and adding a lot of things. I don't think that works only, only if it's the things we do all the time, which is extra work I think."
	Validation documents		P1 - "I mean we could put them in our validation documents."
			P2 - "Yes."
WDST Improvement	*Disliked*	The participants' answer to the question: What did you like and dislike in the documentation template?	P7 - "what I disliked, was the text that you really have to write a lot. Of course."
			P1 - "I think people who look at it would probably be confused by it, would be easier to just have free texts."
			P7 - "Yes. I think so too. if it is a really huge process. This would become really huge, and then just backtrace everything it would be, kind of nightmare, eh. It's too much."
			P5 - "Oh, I agree."
			P5 - "And I wouldn't be a fan of using it."
	Automation	Have the documentation generated by software, using to the diagram	P2 - "At least the parameter parts we have to use. I mean now when we document workflows for the hospital, we have to present like a table of the tools, and the parameters used and stuff. So if that could be automated as well and done from the graphics, that would be good. Would save a little bit of time, I guess."
			P1 - "I mean it would be good. We could just paste it in a document. I mean if it's automatically generated, I mean yeah."
			P2 - “If someone asks for it, then it would be good to generate it automatically. And then it would be cool if you could just right click on the tools and add the parameters."
			P3 - "Exactly or the command, they call, how did you run this."
			P5 - "Because everything is dig digital anyway, right? So It doesn't matter."
WDST complexity	*Incomprehensible*	The participants' answer to the question: Who could or would not understand this template?	All participants - "Everyone."
			P7 - "It is really hard to go through it."
			P2 - "I think it's because it's so thorough."
			P7 - "Yes."
			P2 - "We would, we wouldn't really want to put all of this information in the cells, when creating it. So it would be like, what's this for? And be confused."
WDST Usage	*Not useful*	The participants' perspective of how the WDST can be used.	P7 - "I mean these kinds of templates are nice to have them, but useful? I'm not really sure, at least not for us. Maybe for you two, did you have to have them for legal issues?"
			P1 - "No."
			P2 - "No. The only thing we would need, is like parameters, values and stuff from the tools, because the other thing is basically just describing what the graphics already doing, but instead in text so..."
			P3 - "More complicated."
WDST usage effect on the current state	*Increase Time-spend*	The participants' answer to the question: Would the template's usage affect the current way of documenting workflows at your facility? If so, why?	P7 - "So, we don't know if the time you spent filling this, would really be worth it, because probably the analysis that you do would take minutes and then you would still have to do that. So but not minutes, but you know what I mean, It would take longer to break it down"
			P3 - "So working hours."
			P7 - "If it's only clicking then would be fine. But if not it just increases the time, the working time a lot."
			P5 - "Yes."

Appendix S
Results from Mentimeter open-ended questions

- Concrete Syntax -



Would you add or remove anything? If yes, please describe

no

–

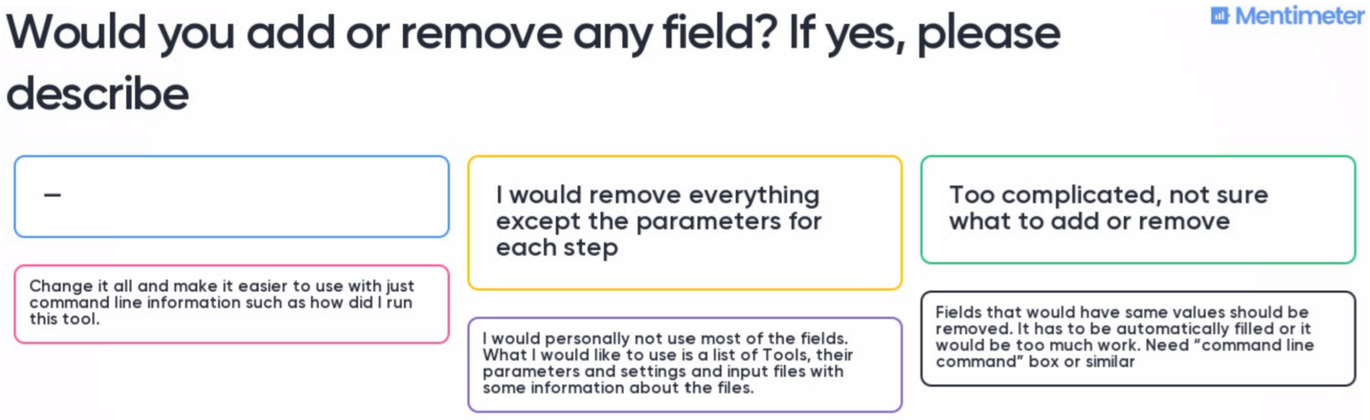
Not really

I would replace soft threshold by something like "manual inspection" or "manual evaluation" or something, with its own symbol to distinguish from hard threshold

Workflow Name table with places. Make it easier to add Several outputs. I would remove the fork-thing.

Remove forks and location boxes. Add software to automatically generate graphs

- WDST -



Would you add or remove any field? If yes, please describe

–

I would remove everything except the parameters for each step

Too complicated, not sure what to add or remove

Change it all and make it easier to use with just command line information such as how did I run this tool.

I would personally not use most of the fields. What I would like to use is a list of Tools, their parameters and settings and input files with some information about the files.

Fields that would have same values should be removed. It has to be automatically filled or it would be too much work. Need “command line command” box or similar