STAIRWAI

Stairway to AI: Ease the Engagement of Low-Tech users to the AI-on-Demand platform through AI, H2020

D2.1 Requirements for AI-on-demand platform

Deliverable information	
Deliverable number	D2.1
WP number and title	WP2 Technical Requirements, service layer design and integration with the Platform
Lead beneficiary	UNIBO
Dissemination level	Public
Due date	30.6.2021
Actual date of delivery	30/06/2021
Author(s)	Ville Tenhunen, Daniele Cesini, Jean-Marc Bonnefous, Michele Lombardi, Michela Milano
Contributors	
Deliverable reviewers	Javier Vázquez Salceda

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

Document Control Sheet

Version	Date	Summary of changes	Author(s)
0.1	2.6.2021	First draft	Ville Tenhunen
0.2	13.6.2021	Revised draft by the internal reviewer and circulated to partners	Daniele Cesini, Jean-Marc Bonnefous, Michele Lombardi, Michela Milano
0.3	29.6.2021	Final version including feedback from partners	Ville Tenhunen

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

Table of contents

1. Executive summary 5
2. Introduction 5
2.1. Purpose of the document 5
2.2. Scope of the document 5
2.3. Structure of the document 5
3. Requirements for the horizontal matchmaking layer and reputation mechanisms 6
3.1. Horizontal Matchmaking and Reputation Mechanism 6
3.2. Use Case Collection and Analysis 7
3.3. Requirement collection methodology 8
3.4. Functional and Non-Functional Requirements for horizontal matchmaking 9
3.4.1. Functional Requirements 9
3.4.2. Non-Functional Requirements 10
3.5. First proof of concept on Planning 11
4. Requirements for the vertical matchmaking layer 12
4.1. Introduction to the vertical matchmaking 12
4.1.1. Vertical matchmaking 12
4.1.2. User needs and constraints 13
4.1.3. Neural Processing Units 13
4.1.4. Data flow of a machine learning algorithm with an NPU 15
4.2. Architecture definitions 16
4.2.1. Hardware resource providers 16
4.2.2. Used standards for hardware and software 16
4.2.3. Interface protocols 18
4.3. Functional and non-functional requirements for the vertical matchmaking 18
4.3.1. Introduction to the functional and non-functional requirements 18
4.3.2. Functional requirements for the vertical matchmaking 19
4.3.3. Non-functional requirements for the vertical matchmaking 21
5. Conclusions 23
Annex 1: Questionnaire Outline 25

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

List of figures

- Figure 1. Horizontal matchmaking
- Figure 2. Vertical Matchmaking detailed conceptual blocks and inputs –outputs
- Figure 3. NPU principles
- Figure 4. The data flow of a ML algorithm with the NPU
- Figure 5. Layered presentation of the AI-on-demand platform
- Figure 6. Natural language interactions, horizontal matchmaking and vertical matchmaking

Acronyms

Acronym	Explanation
AI	Artificial Intelligence
API	Application Programming Interface
ASIC	Application-Specific Integrated Circuit
CPU	Central Processing Unit
DDA	Data-Driven Algorithm
EC2	Elastic Compute Cloud
FPGA	Field-Programmable Gate Array
eduroam	Education roaming
FTP	File Transfer Protocol
GPU	Graphics Processing Unit
HPC	High Performance Computing
HW	Hardware
IaaS	Infrastructure as a Service
LPDNN	Low-Power Deep Neural Network
LSF	Load Sharing Facility
ML	Machine Learning
NLP	Natural Language Processing
NNM	Neural network model
NPU	Neural Processing Unit
PaaS	Platform as a Service
POSIX	The Portable Operating System Interface
QoS	Quality of Service
SME	Small and Mid-size Enterprise
SoC	System on Chip
SLURM	Simple Linux Utility for Resource Management
SSD	Solid State Drive
SSH	Secure Shell Protocol
WebDAV	Web-based Distributed Authoring and Versioning
WP	Work Package

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

1. Executive summary

The StairwAI project aims to create a bridge between users in a low-tech level to the higher-level AI resources. The project will do this by facilitating low-tech users' engagement on the AI on-demand Platform. This will be achieved through a new service layer enriching the functionalities of the on-demand platform and containing:

(1) a multi-lingual interaction layer enabling conversations with the Platform in the user's own language,

(2) a horizontal matchmaking service for the automatic discovery of AI assets (tools, data sets, AI experts, consultants, papers, courses etc.) meeting the user requirements and,

(3) a vertical matchmaking service that will dimension and provision hardware resources through a proper hardware provider (HPC, Cloud and Edge infrastructures).

This means that it is needed to define user requirements on several level among low-tech users, resource provides and those who develop or integrate these elements as a one service layer. This document presents the first version of requirements and architecture component for horizontal matchmaking and vertical matchmaking.

2. Introduction

2.1. Purpose of the document

This deliverable reports the tentative requirements for the multi-lingual interaction, the vertical and horizontal matchmaking layers. Purpose of this deliverable is to be the first version of high level requirements for an AI-on-demand platform. Additionally, the document contains definitions for hardware resource providers.

This deliverable will be followed by D2.2 "Requirements for the AI-on-demand platform-2nd version" which reports the requirements for the multi-lingual interaction, the vertical and horizontal matchmaking layers on detailed level for Open Call selective pilots.

2.2. Scope of the document

Deliverable D2.1 is produced within the WP2 and it affects specifically on WP4 (Multi-lingual interaction with the platform), WP5 (Horizontal matchmaking), WP6 (Vertical matchmaking and integrations) and later WP7 (Open calls and especially T7.4 Call to Identify Experts and HW Resources Providers). D2.1 will be followed by a second version of the requirement specifications on deliverable D2.2.

2.3. Structure of the document

The deliverable consists of 3 main sections. The section 3 defines and describes requirements for the horizontal matchmaking layer and the reputation mechanisms. Section 4 defines requirements for the vertical matchmaking layer and finally section 5 describes integrations between the StairwAI matchmaking modules and the AI-on-demand platform in general.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

3. Requirements for the horizontal matchmaking layer and reputation mechanisms

3.1. Horizontal Matchmaking and Reputation Mechanism

Horizontal matchmaking is a service that enables an easy access and discovery of the AI assets on the AI-on-demand platform. It takes as input the result of the natural language component, namely a set of user's needs and maps them into proper ontology categories where AI assets and resources are classified thanks to the Structured information system.

The mapping (classification) will be performed via machine learning techniques possibly mixed with rules designed by experts. StairwAI will create a data set of use cases that enables proper training of the machine learning models.

The Horizontal matchmaking engine will receive as input the outcome of the NLP engine that provides user needs. The horizonal matchmaking service aims to find the combination of techniques that satisfy the case study. Figure 3 shows the blocks that describe the conceptual components of the Horizontal matchmaking and the inputs and outputs of this block.

To enable matchmaking, it is essential to construct a common language that, from the NLP input, produces a set of key components that are used to identify into the Ontology the AI4EU components as AI-assets: tools, people, data sets, benchmarks, courses that are then proposed to the user as alternative results for the matchmaking.

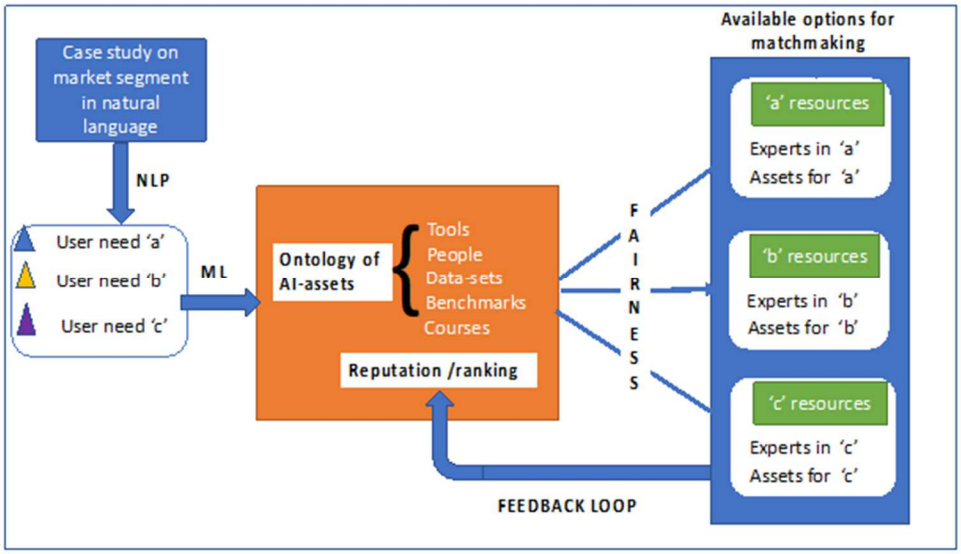

Figure 1. Horizontal matchmaking

Fairness is an essential property of the matchmaking and it should be guaranteed for every result. Therefore, StairwAI will provide a reputation mechanism that will rank resources available on the

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

platform. This mechanism can be enforced and fueled via a feedback loop from users toward the platform. Since the reputation mechanism has a strong dependence on the design of the matchmaking system, its specific requirements will be discussed in the forthcoming D2.2.

There are two other matching alternatives that can benefit from the Horizontal matchmaking, the matching of: *job offers with curricula, and training requests with courses, papers and experts*.

These two additional services will be developed in the second part of the project. Therefore, in this first deliverable, we consider only the mapping of use cases to AI categories containing AI assets. The two additional services will be reported in D2.2.

In the following we outline the process we are pursuing for collecting requirements and data needed to develop horizontal matchmaking. Then we describe the functional and non-functional requirements collected.

3.2. Use Case Collection and Analysis

To train a model for mapping a use case into the appropriate AI category that might provide a solution, it is important to create a data set of use cases. They should be multi-lingual, and then translated in English. The collection of use cases will be performed by using an on-line survey.

The survey consists of a multi-part form, designed to offer a compromise between simplicity and level of details of the questions. On the one hand fewer, simpler questions allow one to fill a form quickly, increasing the odds of receiving a large number of submission; however, they also lend themself to vague answers and lack of consistency. On the other hand, many detailed questions make it easier to process the survey result via automated approaches, but they may discourage users from replying.

Specifically, the form consists of three main parts:

1. The first part of the form concerns personal details about the user and its affiliation
2. The second part is aimed at collecting information about a real or hypothetical use case that the user company values and that has required (or may require) AI techniques. This section consists of five, free-form, questions about A) the use case context; B) the use case motivation; C) Data availability/provisioning; D) Use case objective; E) Additional requirements (e.g., fairness or explainability, when not part of the use case core).
3. The third part needs to be filled only for use cases that have already been addressed using AI techniques, at least to some degree. It consists of twelve multiple-choice questions structured according to the AI Watch taxonomy1: collectively, the questions allow an analyst to characterize the employed AI techniques.

The document contains a notice about GDPR compliance, and users are asked whether they consent to make their survey entry public in anonymized format (including the company details). Users may also opt to be contacted for additional feedback (e.g., requirement collection

[1] https://publications.jrc.ec.europa.eu/repository/handle/JRC118163

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

interviews, similar to those discussed in Section 3.3), by the project coordinator and (if consent is given) by other project partners.

Clearly, for horizontal matchmaking the survey should collect not only the use case, but also the AI technique that has been used for solving the company problem. For this reason, the data collected by companies will be divided into two sets: in the first we have use case description collected by low-tech companies that have not yet adopted AI but would like to. These use cases will have the natural language description of the problem the company has to solve, but not the AI technique with which it is indeed solved. This set will serve for any form of unsupervised learning, aimed at analyzing text structure, cluster use cases and will be made available publicly upon agreement of the use case owner. The second set instead is complete, namely it contains the use case description and the corresponding AI technique that has been used for solving it. Therefore, we aim to target low tech SMEs that have already adopted AI technology in their business. This second set will be used to classification purposes and represent the main corpus of data used for our horizontal matchmaking. Also, the second set will be made available publicly upon agreement of the use case owner.

3.3. Requirement collection methodology

For the requirement collection we have opted for a set of direct interviews with SMEs and companies, including both consumers and producers of AI resources. The purpose of such interviews are to understand which have been their main barriers and difficulties in adopting AI, which have been the advantages and impacts of adopting AI, and what are the main features they expect from a platform to support AI adoption or promote the distribution of AI assets.

An important task to carry before the workshop/interviews take place consists in defining which classes of companies we need to interview. We need to make sure that we target both potential "consumers" and "producers" or AI related resources. We have devised a simple primary characterization for the main classes of companies to be included in the process:

Group	Subgroup	Notes
AI Consumers	AI-savvy	Companies whose main business is not AI, but know about AI techniques and have possibly employed them (or are employing them)
	Non-AI savvy	Companies whose main business is not AI, have no experience with AI techniques (or very limited experience), and may or may not have some degree of awareness of AI techniques
AI Producers	--	Any company, institution or group, that is responsible for one or more AI resources.

As a secondary subdivision, we distinguish between Small-Medium Enterprises and large ones. While the focus of the project will be on SMEs (especially on the consumer side), some larger companies may be consulted during the interview process to ensure maximum generality of the matchmaking system.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

Drafts for the questions to be asked during the interviews have been prepared, targeted to the main classes of companies identified above. All interviewed personnel will be briefed to the project objective when the interview starts.

Questions for AI savvy Consumer Users:

1. Describe a use case that has benefited or may benefit from AI techniques
2. List the main AI techniques that the company knows about
3. List the main AI techniques that the company has used in the past
4. List the main AI techniques that have been employed in the described use case
5. Identify the major obstacles encountered when trying to apply AI techniques (at all levels, including both technical difficulties, acceptance issues, regulation and ethical concerns)
6. Identify which features would be critical for a matchmaking service to be useful in your case

Questions for Non-AI savvy Consumer Users

1. Describe a use case that has benefited or may benefit from AI techniques
2. List AI techniques that the company knows or has heard about (if any)
3. Attempt to identify the main barriers encountered when approaching AI techniques
4. Identify tools and services that would make it simpler to approach AI techniques
5. Identify which features would be critical for a matchmaking service to be useful in your case

Questions for AI Producers

1. Characterize the kind of assets or resources that the company/institution/group produces
2. Define the kind of problems that the assets/resources are meant to address
3. Describe how the asset/resource has been used in the past
4. Identify which features would be critical for a matchmaking service to be useful in your case

As a rule of thumb, questions will be designed so that they start from the individual experience of the interviewed person, and end up with any direct suggestions about requirements and potential features of the matchmaking system.

3.4. Functional and Non-Functional Requirements for horizontal matchmaking

In this section we outline a preliminary list of functional and non-functional requirements, with the aim to provide a better characterization of the horizontal matchmaking service and its design priorities. This preliminary list is meant to be integrated and expanded in D2.2, thanks to the outcome of the interview process.

3.4.1. Functional Requirements

Functional requirements define the input output behavior of the system, so as to define its behavior.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

In the case of the Horizontal Matchmaking, the system input consists of a natural language description of a request, which in the first stages of the project will always be the description of an industrial use case. The description will follow a semi-structured format (i.e. a few text boxes), in an effort to make sure that important pieces of information are not omitted. The format of the input will be the same employed in the data collection questionnaire (see section 2.2 Use Case Collection and Analysis) and it is described in the following table:

Field	**Definition/Help Text**
Use Case Context	The physical or digital system that benefited/may benefit from the use of AI
Use Case Motivation	How the problem arises (or has arisen) in the company business, what makes it challenging
Data Availability/Provisioning	Which data are (or were) available, what needs to be still collected?
Use Case Objective	What the AI system is supposed did/is supposed to do and which benefits you got/plan to get
Additional requirements	Any property of the AI system that are not immediately tied to its function, but are still needed for the application (e.g., fairness, explainability, energy or power efficiency, latency)

In Tasks 5.2 and 5.3, additional request types will be considered, including professional profiles (e.g., to be matched to open positions) or request for courses.

The system output can be characterized at two distinct stages:

- At the first stage, the output will consists of a labeling of the input requests in terms of a reference ontology, representing the applicable classes of AI content. Both single-class (e.g., a single AI category) and multi-class mappings (multiple categories, with different weights or intensities) will be considered.
- As a second stage, the above mapping will be used to retrieve relevant resources and rank them, also characterized in terms of a mapping over the same ontology. The resources may include tools, similar use cases (possibly solved), datasets, papers, contact details for experts, and courses. The resource will be ranked primarily on the basis of the degree of matching between their mapping and that of the request.

3.4.2. Non-Functional Requirements

Non-functional requirements define properties of the system that, while being critical to ensure proper behavior, are less directly tied to its input and output. In the first stage of the problem we have identified the following list of non-functional requirements:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

- In an effort to maximize the system impact across Europe, especially on so-called low tech industries and SMEs, the horizontal matchmaking system should accept use case description in multiple languages. Only a limited set of languages will be supported in the project timeline, the precise list to be defined based on the availability of training data and technology for Natural Language Processing, on the expertise of the consortium partners, and on the population size.
- The ontology used for the characterization of both use case description and AI resource should be sufficiently general and flexible to ensure longevity and wide applicability for the matchmaking system. A preliminary analysis has identified the AI Watch taxonomy as a promising candidate.
- The system should ensure fairness in the sense that, in absence of strong motivation, no AI resource should receive excessively low exposure. Defining an adequate fairness criterion is not trivial, since natural bias in the requests (e.g., an abundance of requests in a specific AI context) will inevitably translate in a bias on the proposed resources. Some effort will need to be dedicated in the course of the project to identify a suitable fairness metric and define reasonable thresholds.
- The system will need to feature a reputation mechanism, designed to collect feedback from the users and make sure that it reflects on the resource ranking, so that higher-quality or more reliable resources will be given some degree of prominence. As a consequence, this reputation mechanism will affect how fairness is treated and enforced (both aspects will be investigated in detail in Task T5.4).

3.5. First proof of concept on Planning

Together with the ICT-49 project AIPlan4EU, we have set a first proof of concept that has the double aim of defining requirements for the StairwAI services, and set a connection between two related European projects.

AIPlan4EU has the objective to make modern planning technology applicable for everyone by developing a uniform, user-centered framework to access the existing planning technology and by devising concrete guidelines for innovators and practitioners on how to use this technology.

To do so, the project will consider use-cases from diverse application areas, and include several available planning systems as assets on the AI-on-demand platform that can be selected to solve practical problems. A selected subset of the AIPlan4EU use-cases and planning systems will be used to enrich the set of use cases and solutions that StairwAI plans to collect.

Common initiatives of StairwAI with AIPlan4EU are the following:

1. AIPlan4EU will distribute the questionnaire for collecting use cases related to planning.
2. AIPlan4EU will put StairwAI in contact with selected SMEs and companies using planning for interviews.
3. Using selected AIPlan4EU use cases, StairwAI will define a model for horizontal matchmaking on planning assets.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

4. StairwAI will glue this model with planning experts' domain knowledge to enrich and refine the data driven model.
5. AIPlan4EU will help refining the ontology section on planning in StairwAI, to better reflect the state-of-the-art planning technologies and tools.
6. If possible, AIPlan4EU will help providing algorithms for benchmarking applications in vertical matchmaking.
7. StairwAI will made available the developed services to the AIPlan4EU companies and partners at an early stage, to allow them to test and validate those services.

4. Requirements for the vertical matchmaking layer

4.1. Introduction to the vertical matchmaking

4.1.1. Vertical matchmaking

The dimensioning of hardware resources for some specific AI algorithm can depend on the inputs, making it unpredictable. However, in StairwAI we will benchmark and profile those in the AI on-demand Platform that do not have big variability with respect to the inputs.

Moreover, for those that they do, we will select the type of hardware and resource provider that will suit best the end user preferences. For this aim, machine learning tools performing regression will be used are able to extract the running time, memory and hardware resources required by a given algorithm.

The data-set of algorithm resource demand needs on specific hardware platforms will be built during the project lifetime enabling proper training of the machine learning model.

In instances where the development of the case study needs to deploy AI assets on physical resources, a Vertical Matchmaking process will be triggered. The process consists on given a set of requirements from the end user – time constraints, costs, privacy, or any other – to find the appropriate set of resources that satisfy a given algorithm and tool in the hardware resource provider marketplace.

The workflow of the Vertical Matchmaking is depicted in figure 2, in which the optimization system for the user request receives as input as set of selected algorithms or services that need to be deployed for the success of the case study. Then, the engine also receives as inputs the user constraints — time and cost, but can be extended to any other parameters that can be identified during the development of StairwAI.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

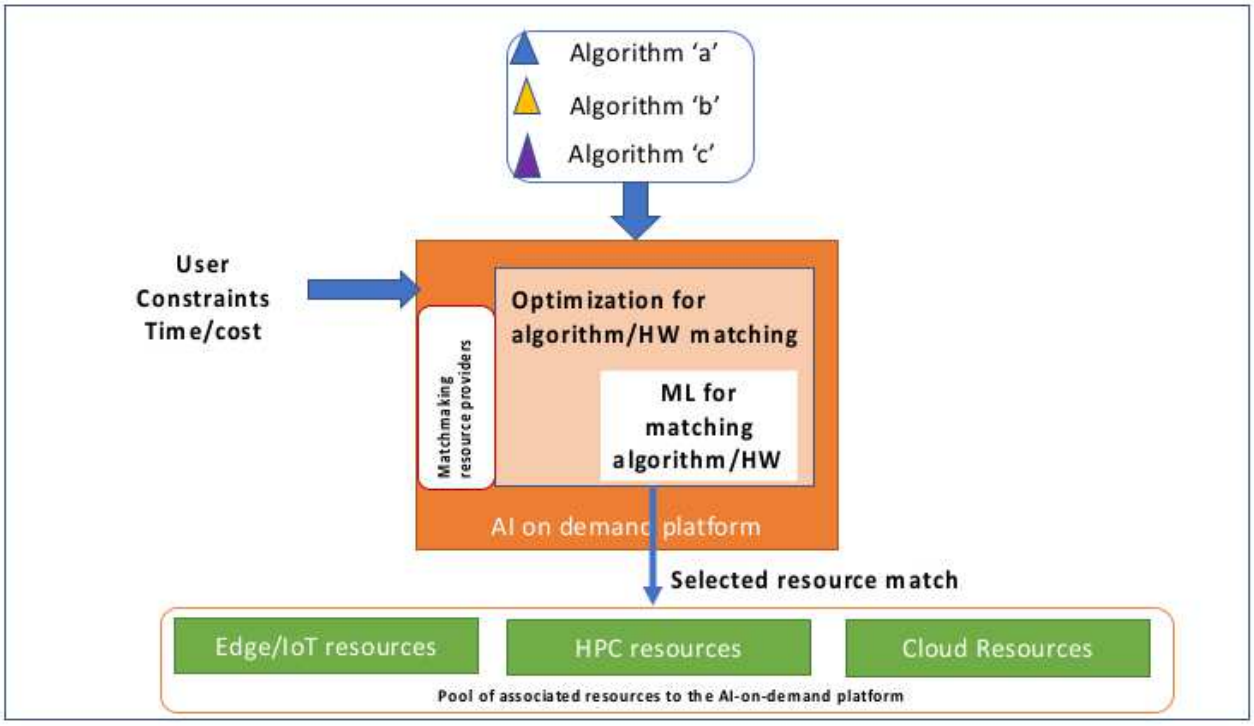

Figure 2. Vertical Matchmaking detailed conceptual blocks and inputs -outputs

The combination of both — as well as the underlying technology — containerization, virtualization, hardware resources — and the implementation available linked to the architecture will prune the multiple options and, by using Machine Learning (ML in the figure) techniques will select the best resources.

The Computing services for supporting the experiments will be provided by those entities selected in the open call for HPC and cloud providers (WP7).

4.1.2. User needs and constraints

Tentative user needs and requirements has described mainly in the previous section of this document. Here has describe some requirements which will be defined detailed in next version of the Requirements for the AI-on-demand platform document (D2.2).

User needs on vertical matchmaking specific needs based on users' choices and algorithms they use. These factors define ML/AI based needs for hardware. StairwAI on demand platform aims to optimize this matching between algorithms and hardware.

Main constraints based on two main factors;

Time consumption and costs for the user and

HW offerings from service and hardware providers.

4.1.3. Neural Processing Units

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

Neural Processing Units (NPUs) are one important part of the modern AI environment. These are one example for hardware usage in this area. Next version of this document will cover deeper and wider insights for this area of infrastructures and architectures.

A neural network model (NNM) is the machine learning component of a data-driven algorithm (DDA). A NNM is represented by its architecture that shows how to transform its input(s) into its output(s). A NPU implements all control and arithmetic logic needed to execute a NNM efficiently on an SoC.

NPUs are specialized AI processors without generic APIs requiring special programming tools.

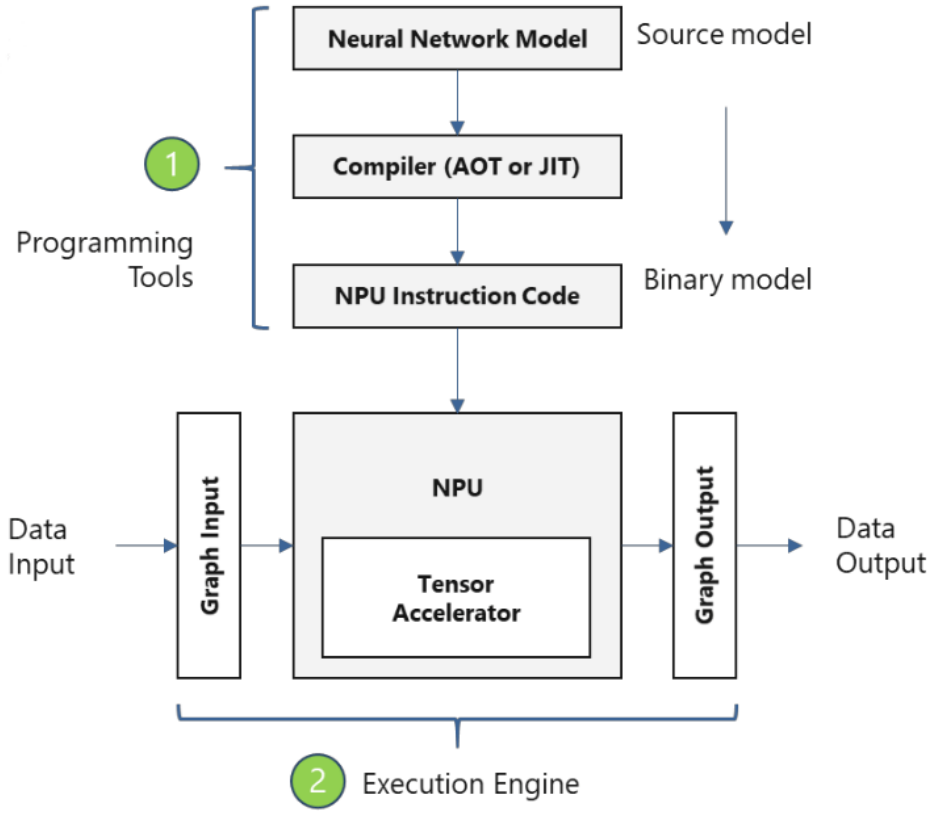

Figure 3. NPU principles[2]

An NPU can be implemented with a variety of computing technologies (CPU, GPU, FPGA, or ASIC) and each have their own advantages and tradeoffs. The best choice of technology depends on the application and use case, which explains the large degree of fragmentation and variety of implementations available in the market. To embrace the hardware diversity, it is important to map the computation to NPU hardware efficiently.

An NPU consists of Programming Tools and an Execution Engine. The Programming Tools consist of a Compressor, Converter, Compiler and Optimizer. SoC vendors do not typically provide all the optimal tool set needed to optimally program an NPU and often have incompatibility issues with

[2] Bonseyes Community Association

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

new standards. From a hardware perspective, the NPU execution engine incorporates multiple elements beyond its tensor accelerator including a scheduler and programmable compute element.

A neural processor requires a high-performance interconnect to access external memory, a host controller, and other complementary accelerators that may be available on SoC. It places special requirements on the MMU and QoS of the Interconnect due to high burst bandwidth requirements and the locality of dense neural networks being different than CPU memory accesses. Additionally, functional safety requirement can place additional diagnostic and debugging requirements on the interconnect.

An NPU provides low latency acceleration of the computation of Machine Learning tasks as compared to GPUs or general-purpose compute engines. They also consume less power and improve resource utilization for Machine Learning tasks as compared to GPUs and CPUs due to low precision and high bandwidth multiply-and-accumulate engines.

4.1.4. Data flow of a machine learning algorithm with an NPU

An NPU executes neural network models from model input to mode! output. Data input from a sensor and pre- and post-processing steps are handled by a host or offloaded to other accelerators. In certain instances, an NPU with integrated vision system can efficiently handle pre- and post-processing steps through the integration of OpenVX APIs.

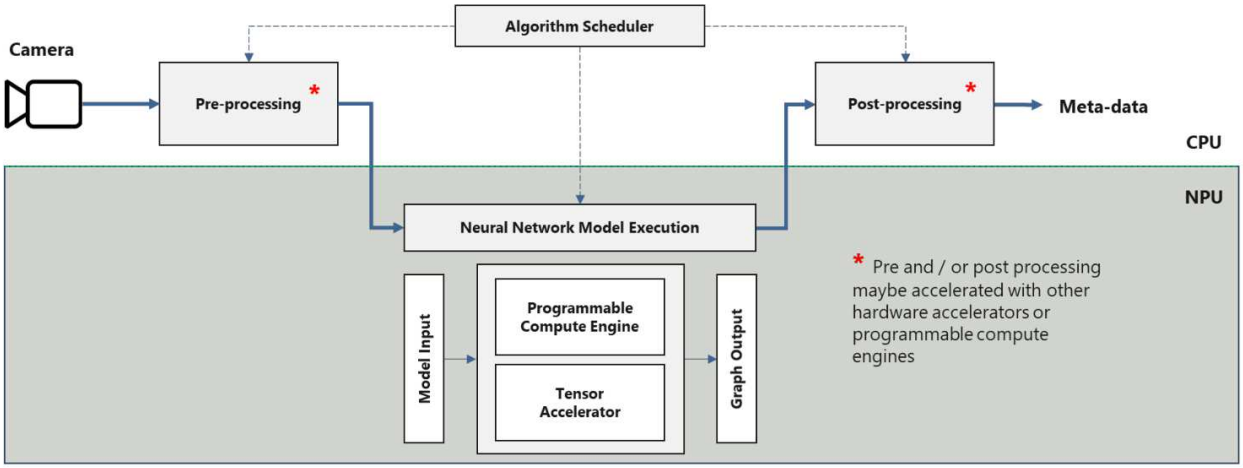

Figure 4. The data flow of a ML algorithm with the NPU[3]

NPUs are lower latency and more power efficient however do not handle pre/post processing steps.

SoC architectures are rather largely supported by major technology providers. NPU profiles range from Sensor NPUs, Head-Unit NPUs and Centralized NPUs based on power consumption profiles and compute performance. All SoC architecture types are supported. Support includes Qualcomm

[3] Bonseyes Community Association

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

Snapdragon, NVIDIA AGX, Renesas R-Car V3U and NXP iMX8. Each SoC platform has its own NPU implementation and particular implementation.

A high level of compliance for most platforms and experience with all major SoC vendors are due to the use Low-Power Deep Neural Network (LPDNN) specialized tooling that has heterogenous support across CPU, GPUs, and NPUs.

4.2. Architecture definitions

Vertical matchmaking based on delivering hardware and software resources. Therefore, there is also some prerequisites which have to take into account when requirements are defined. Here these definitions have described in few main chapters; hardware resource providers, use standards for hardware and software, interface protocols, neural processing units and data flow of a machine learning algorithm with an NPU. Requirement and onboarded resources have to be compliant with these architecture level requirements.

4.2.1. Hardware resource providers

A variety of hardware for StairwAI applications and vertical matchmaking algorithms covering the following categories:

1) High-end resources

- Cloud resources made available at the IaaS level via APIs or Web GUIs
- PaaS deployments on the Cloud infrastructure
- HPC and parallel computing resources including low latency interconnected nodes equipped with computational accelerators

To increase heterogeneity, it should be possible to run on various architectures, i.e., X86, ARM, RISC-V for what concerns CPUs, and AMD or NVIDIA for accelerators.

2) Edge and low-power resources

- Nodes/clusters based on X86 or ARM architecture in the range 4-80Watts, possibly equipped with accelerators, GPUs or NPUs. I.e.
- NVIDIA Jetson-TX1/X2
- Intel low-power system-on-chip (Avoton, Pentium, Apollo Lake, Gemini Lake)
- Intel low-power, server-grade systems (i.e., XeonD processors)
- Development boards powered by SoCs designed for the mobile market (i.e., Exynos processors)
- In a higher power range, motherboard based on the AMD Threadripper processors

Concerning storage resources, it will be possible to preserve data at least in two QoS: spinning disk and low latency devices such as SSD. In addition, multiple replicas for high availability and failover strategy implementation will be possible on the EGI Cloud federated storage.

4.2.2. Used standards for hardware and software

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

To ensure replicability of the performed tests and implementations of the StarwAI project, standards or de-facto standards will be adopted as much as possible. The following tables summarize them in various domains.

Silicon Layer:

Type	Name of Body (hyperlinked URL)
CPUs/GPUs	x86_64
	ARM
	NVIDIA
	AMD
	RISC-V
Organizations	RISC-V Foundation

Operating Systems:

Type	Name of Body (hyperlinked URL)
Operating Systems	Linux
	Microsoft Windows
Organizations	Linux Foundation
	Linaro
Standards / Examples	ISO/IEC 23360-x:2006 (LSB)

Virtualization:

Type	Name of Body (hyperlinked URL)
Technologies	Linux Containers
	VM
Organization	OCI
	OCP

Management:

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

Type	Name of Body (hyperlinked URL)
Technologies	Kubernetes
	Docker Swarm
Organization	Apache
SDO	ETSI

Type	Name of Body (hyperlinked URL)
Tools and Framework	TensorFlow
	OpenNN
	Caffe
	Keras
	PyTorch

4.2.3. Interface protocols

Interface protocols will be based on standards, as much as possible, according to the users' needs.

Data Storage interface:

POSIX filesystems will be used for data at rest on local resources while for Cloud Storage access S3 interfaces will be provided to users.

Data transportation:

Mainly three types of data movement interfaces will be provided: ftp, https/WebDAV for Grid and Cloud resources and SSH/rsync based for HPC and low power local resources.

Cloud Resources instantiation

Amazon EC2 compatible APIs and OpenStack APIs will be provided for IaaS Cloud access to the resources.

Batch access to local clusters

No specific requirements received up to know for local/direct access to clusters, if this will change the preferred tools will be SLURM and LSF.

4.3. Functional and non-functional requirements for the vertical matchmaking

4.3.1. Introduction to the functional and non-functional requirements

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

Previous chapters have described general constraints and prerequisites for vertical matchmaking. Based on these it is possible to define some requirements for the StairwAI on-demand platform vertical matchmaking.

Users are low-tech SMEs and their representatives. This has to be taken into account in requirements and deployment of the system.

Functional requirements for specify what the system should do such as:

- Business Rules
- Pricing and payment models (possible methods for users and service providers)
- Authentication (which methods users can use)
- Authorization (which levels users need)
- Deployment mechanism
- Possible licenses
- Ordering

Non-functional requirements for specify how the system performs a certain function such as:

- Usability (regulations and guidelines)
- Performance and QoS (for example response time, throughput, utilization)
- Scalability (component's scalability)
- Capacity (capacity specification on vertical matchmaking)
- Security and data protection – regulations and guidelines
- Interoperability
- Other regulatory requirements

4.3.2. Functional requirements for the vertical matchmaking

Business rules

One objective of the WP7 of the StairwAI is to develop business model/s for the use of the services developed and the work plan of the post-project market penetration of the extended AI-on-demand platform.

R4.1 Business rules
Hardware resources which will be integrated have to follow the StairwAI WP7 definitions and rules. *essential*

Ordering, pricing and payment methods

Ordering process have to be defined and it have to contain also pricing information and suitable payment methods. Typical payment methods are

- Pay per use
- Project contracts
- Billing

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

- Other

R4.2 Ordering
Ordering process *essential*

R4.3 Pricing
Pricing information have to be transparent enough and available for potential users. *essential*

R4.4 Payment methods
Payment methods of services and resources have to be defined clearly also when there is not any method needed. *essential*

Licenses

Here licenses mean possible licenses between end user and resource provider through the StairwAI platform.

R4.5 Licenses
Licenses have to be included to the service or resource information as clearly as possible. Information has to contain link to the exact license description. *essential*

Authentication and authorization

On-demand platforms contain multiple authentication and authorization issues to be solved based on users' needs and requirements.

According to the type of resources, various authentication methods will be made available to the systems. In particular this is used to access cloud resources, it is possible to use federated identity management systems based on standards such as OpenIDConnect tokens.

R4.6 Authentication and authorization
For service provider authentication it is existing federations like eduGAIN, eduroam or EGI Check-In should be used. *essential*

R4.7 Local accounts
Local accounts (i.e username/password) to access HPC or Edge resources will be also needed. *essential*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

On-demand platform

Following on-demand platform functionalities have to be defined during the project with the best practices and based on agile service development methods i.e., detailed use cases and user stories have to be described with product owner(s) to backlogs.

R4.8 On demand platform
On-demand platform cover at least following areas to develop: • Portal functionalities • Resource discoverability and search • Collection and labeling of data • Training models, optimization and benchmarking • Hardware-in-the-Loop Deployment • Solution Integration • Resource onboarding *essential*

Deployment mechanism

The deployment mechanism is the action used to put built application or resource into platform where users can find and use it.

R4.9 Deployment mechanism
StairwAI on demand platform deployment mechanism have to be well described and as open as possible for multiple resource sources. *essential*

4.3.3. Non-functional requirements for the vertical matchmaking

Usability

Standardization organisation ISO defines usability as "The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use."4 I case of web-based services this means merely “a small learning curve, easy content exploration, findability, task efficiency, user satisfaction, and automation”.5

R4.10 Usability
Usability of upcoming services have to follow W3C Guidelines (https://www.w3.org/WAI/test-evaluate/) essential

[4] https://en.wikipedia.org/wiki/Usability
[5] https://en.wikipedia.org/wiki/Web_usability

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 101017142

Accessibility

Web accessibility allows everyone, including people with disabilities, to perceive, understand, navigate and interact with the Internet.

R4.11 Accessibility
StairwAI on-demand platform and vertical matchmaking user interfaces have to take into account regulations like The Web Accessibility Directive (Directive (EU) 2016/2102)[6] *essential*

Performance and QoS

Quality of service (QoS) is the description or measurement of the overall performance of a service, such as response time, throughput, utilization.

R4.12 Performance and QoS
During the project, based on users' need and requirements, target level of performance and QoS have to be defined and measures which are needed taken. *essential*

Scalability

Scalability in network-available services, which might be defined as the ability of an application to handle growth efficiently, is typically achieved by making them available on multiple devices.[7]

R4.13 Scalability
StairwAI on-demand platform vertical matchmaking components have to be scalable enough to support uptake resources and services users needs. *essential*

Availability and Capacity

Capacity means in this context that there is sufficient resources available to fulfill users' requests.

R4.14 SLA
During the project service level agreements (SLAs) will be negotiated with resource and service providers. *essential*

R4.15 Capacity and availability
StairwAI project have to identify service performance requirements based on users' needs, plan the resources required to fulfil the requirements and ensure performance monitoring. *essential*

[6] https://digital-strategy.ec.europa.eu/en/policies/web-accessibility

[7] https://www.w3.org/2001/03/WSWS-popa/paper33

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

Security and data protection

Data security and protection are regulated with multiple laws and other regulations.

R4.16 Security and data protection
StairwAI services have to take into account applicable regulations like General Data Protection Regulation (GDPR)[8] and the Network and Information Systems (NIS) Directive[9]. Additionally, guidelines from the StairwAI WP9 Ethics requirements have to be implemented also in these activities. *essential*

Interoperability

R4.17 Interoperability
Interoperability between different service layers, service and resource providers and essential platforms have to be defined to the next version of StairwAI D2.1. Interoperability of the StairwAI vertical matchmaking have to follow New EIF i.e. European Interoperability Framework by ISA2 programme principles[10] *essential*

Standards and architecture framework

Architecture definitions has described in this document (chapter 4.2).

R4.18 Architecture compliance
Vertical matchmaking solutions have to follow standards and architecture definitions described in this document. *essential*

5. Conclusions

StairwAI service based on few major components from the AI4EU architecture and service providers which StairwAI will get from open calls later. Novel parts of the StairwAI are horizontal matchmaking and vertical matchmaking, which makes resource deployment easier for low tech SMEs. Idea of these matchmaking components are use AI to find user requirements and constraints, and also suitable resource provider. Multilanguage NLP techniques will be used to create advance user experience.

[8] https://gdpr.eu/
[9] http://data.europa.eu/eli/dir/2016/1148/oj
[10] https://ec.europa.eu/isa2/eif_en

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

As a first version of the StairwAI requirements and architecture, this document describes merely what have to be done within the project than how it should be done. Next version of the document will focus on this latter part of the architecture.

Next steps to define requirements for AI-on-demand platform and StairwAI services are:

- Use case analysis based on questionnaire and interviews (see chapters 3.2 and 3.3)
- Functional and non-functional requirements based on use case analysis (chapters 3.4 and 3.5)
- Define collaboration across ICT49 projects (such as BonsAPPs) for example in interoperability issues
- Information architecture needs will be discussed with WP3
- Information security and privacy management issues and implementation need in StairwAI
- Based on use cases define possible offerings and requirements for HW (chapters 4.2 and 4.3)
- Technical definitions of the HW components have to be fitted to the use cases (chapter 4.4.) as well as authentication methods and on demand platforms
- Integration requirements with AI4EU services and other possible platforms and marketplaces will be described based on collaboration, information architecture needs and technical descriptions
- Interfaces of the service providers and suitable onboarding technologies for vertical matchmaking have to be described.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

Annex 1: Questionnaire Outline

In this Annex has presented outline of questionnaire described in the section 3.

Section 1: General Information

This section asks for information about the respondent company, plus a few pieces of personal information about the respondent themselves. This section will also include GDPR notices and confirmation requests, not appearing in this list.

- First Name
- Last Name
- Years of experience in Information and Communication Technologies
- Country
- Sector of Operation (public/private/non-profit/academia)
- Industrial sector (According to the International Labour organisation)
- Size of your organisation (in approximate number of employees)
- How many years have the company been operative?
- Company's service (Production services, HR, development of solutions, etc)
- Company URL (optional)

Section 2: Use Case Description

This section contains questions concerning the considered industrial use case. With a few exception, these are all free-form questions that the respondent will address by filling a textbox.

- Application of use case - real or hypothetical
- Use Case Context (The physical or digital system that may benefit/benefited from the use of AI)
- Use Case Motivation (How the problem arises -- or has arisen -- in the company business, what makes it challenging)
- Data Availability/Provisioning (Which data are -- or were -- available, what needs to be still collected)
- Use Case Objective (What the AI system is supposed to do/did and which benefits you plan to get/you got)
- Additional requirements (Any property of the AI system that are not immediately tied to its function, but are still needed for the application -.- e.g. fairness, explainability, energy or power efficiency, latency)
- Have AI components being already employed for the use case? (Yes/No/Maybe)

Section 3: Description of the AI Solution

This section contains questions concerning the AI solutions employed to address the use case (when available). All questions are mostly in checklist form and refer rather closely to the AI Watch ontology for the description of AI topics.

- Approximate date with use case was implemented or launched
- Topics covered on reasoning
- Topics covered on planning and optimization
- Topics covered on learning

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142

- Topics covered on communication
- Topics covered on perception (computer vision)
- Topics covered on perception (audio processing)
- Topics covered on integration and interaction (Multi-agent systems)
- Topics covered on integration and interaction (Robotics and automation)
- Topics covered on integration and interaction (connected and automated vehicles)
- Topics covered on AI Services (These correspond to "analytics perform" in the AI_watch taxonomy)
- Topics covered on AI ethics
- Topics covered on philosophy of AI

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101017142