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

Data management plan

Deliverable information	
Deliverable number	D1.1
WP number and title	WP1 - Project Management
Lead beneficiary	UNIBO
Dissemination level	Public
Due date	30 June 2021
Actual date of delivery	30 June 2021
Author(s)	Michela Milano, Roberta Calegari (UNIBO)
Contributors	All partners
Deliverable reviewers	-

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	4 June 2021	First draft	UNIBO
0.2	9 June 2021	Second draft	UNIBO/ BCA / FBA / TIL / UCC
0.3	23 June 2021	Revised draft by the internal reviewer and circulated to partners	UNIBO
1.0	30 June 2021	Final version including feedback from partners	All partners

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

Table of contents

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

1. The Data Management Plan (DMP)

The DMP is a document that provides details regarding all the research data collected and generated within StairwAI project. In particular, it explains the way research data are handled, organized, licensed and made openly available to the public, and how they will be preserved after the project is completed. The DMP also provides motivations when versions or parts of the project research data cannot be openly shared on account of third-party copyright issues, confidentiality, or personal data protection requirements or when open dissemination could jeopardize the project achievements.
This DMP reflects the current state of the art of the StairwAI project. More details on the datasets can be found in D3.1 where the information described contains the requirements and needs of these with respect to the different components and algorithms to be developed in WPs 4, 5 and 6.
An important consideration is that the final number of the project datasets may vary during the course of research. The variations will be recorded in updated versions of this DMP.

2. Data Summary

StairwAI is a project that brings together knowledge, algorithms, tools, and resources available in the European AI landscape providing a Stairway to AI for low-tech users (i.e., low-tech industries and SMEs). The aim of the project is to enable easy and intuitive interaction with the AI-on-demand (AI4EU[1]) platform, guiding the low-tech users in the discovery of the relevant tools, datasets, experts, and employees, boosting the adoption of the AI-on-demand platform and guaranteeing its long-term sustainability. StairwAI will enrich the AI-on-demand platform through a service layer that enables natural multi-language interaction and performs horizontal matchmaking, namely an automatic mapping between user requirements into assets of the AI-on-demand platform to meet users' business needs. In addition, StairwAI will develop an automatic mechanism for hardware resource dimensioning and provider discovery that, given an application or service corresponding to a set of AI algorithms and given end-user preferences – time, cost, quality of service – identifies the best algorithms to be deployed and the hardware resource provider to satisfy end-user needs.

This deliverable provides an initial version of the data management plan (DMP) for the project. It will act as a guideline for all stakeholders, which contribute to the platform providing and exploiting data. Given the complexity of the project, at this point, many details with respect to data collection are still to be defined, thus the DMP is considered a living document and will be subject to constant change, adaptation, and expansion throughout the project's lifetime.

The remainder of this document is structured according to the DMP requirements specified by the European Union (EU) in the H2020 Programme. Additionally, we provide an initial and hypothetical list of project datasets (that are then described in *Annex I*), based on the information provided by the partners. In *Annex II* is showed the questionnaire exploited to collect information from the partners about their data and the data management procedures they will implement.

[1] AI4EU platform, https://www.ai4europe.eu/

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

During the project, different types of data will be collected and/or generated. Research teams have agreed to convert research data from proprietary formats to well-known and documented standard (open) formats, in order to facilitate accessibility and reusability (Table 1).

Table 1 - Summary of data types and formats

Type of data	Formats used during data processing	Formats for sharing reuse and preservation
Numerical	.xls/.xlsx	.csv, .ods
Textual (free text or categorical variables)	.csv, .ppt, .pdf, .json	.csv, .ppt, .pdf, .txt, .json
Video	.mp4	.mp4
Code	.py	.py

The overall size of the project data is still uncertain at this early stage of the project. However, for some expected dataset it has been possible to estimate the volume of the data, more information about it can be found the descriptive tables in *Annex I*.

We distinguish between three kinds of datasets: I) datasets provided by the members of the consortium, II) external data sources, which might be integrated into the StairwAI platform (such as data from AI4EU platform), III) data originating from applicants participating in the open calls (see Section 2.2).

2.1. StairwAI collected/generated data not from open call

According to the project's goals the datasets involved into the project will be the following.

Industrial Use Cases. (textual data, csv). This dataset will contain descriptions of industrial use cases that have been addressed or could conceivably be addressed via AI techniques, plus annotations about the AI techniques applied to the problem (when applicable). Use case descriptions are in semi-structured form (a few paragraphs in natural language), while annotations are structured. After the collection is finalized, the dataset is expected to have several examples ranging from a few hundred to slightly more than one thousand. Data will be collected by sending a form (questionnaire) to companies (both AI-savvy and no AI-savvy) possibly via intermediaries, with an emphasis on SMEs. The dataset may be of interest for the definition of AI tools aimed at classifying AI use cases. (Corresponding tasks: Task 2.1- Requirements for the horizontal matchmaking layer and reputation mechanisms and task 3.4–Data collection and datasets generation for Horizontal and Vertical Matchmaking. See list of dataset: Dataset 1).

Chatbot data. (textual data, csv). This is the collection of datasets required for the development of the StairwAI Chatbot (Corresponding Task 3.3– Data collected and datasets generation for Multi-lingual NLP. See list of dataset: Dataset 2). The datasets will include:

1. Data for user stories and dialogue scenarios reflecting the content of the eventual conversations with the end-users of the solution (e.g., topical questions, themes, issues etc.)
2. Data for development and domain adaptation of the multilingual Natural Language Understanding modules and other Natural Language Processing tools
3. parallel data and respective terminology for development and domain adaptation of the Machine Translation engines for the respective language combinations.

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

AI-on-demand platform data. (textual data, csv). This dataset will contain available job offers - industrial, academic, expressions of interest - listed on Europe's AI-on-demand platform (AI4EU). The data is expected to be collected by an online form (same data of industrial use cases), which results in having semi-structured data including some categorical data such as location, key skills, etc.; text blocks that require natural language processing; and numerical data such as the years of experience. The finished dataset is expected to contain a few hundred job listings; to the best of the collectors' knowledge, no similar dataset is available. The dataset may be of interest to people with an intention to develop AI algorithms that match job-seekers to available jobs. Furthermore, anonymized user profiles (researchers, developers, domain experts, citizens) extracted from the AI4EU platform with user/experts' preferences on AI topics will be necessary to conduct adequately this task. (Corresponding Task 5.2–Matchmaking between job offers and experts/consultants. See list of dataset: Dataset 3).

Vertical matchmaking benchmarking. (Textual data json, csv, and numerical data). The vertical matchmaking benchmark will provide a set of metrics on efficiency, latency, memory, and power consumption to help the vertical matchmaking engine identify the best solution for a given AI challenge, i.e., models vs. hardware platforms. Metrics such as accuracy need a validation dataset against which the accuracy of the AI model can be computed. The envision datasets will mostly image-based open-source datasets including elements for image classification, face landmarks detection, in-car object detection, emotion recognition, etc. The scope of the vertical matchmaking benchmarking can be also extended to signal processing datasets for surface electromyography gesture recogniton and heart-rate classification or audio -based datasets for keywords spotting.

Vertical matchmaking engine. (textual data, csv, and numerical data). The dataset will consist of benchmarks for a range of AI applications (e.g., ML models, online and offline algorithms, optimization models) executed on a variety of heterogeneous resources. The collected data (e.g., performance metrics) will enable the construction of ML models to estimate the behavior of an AI application on different hardware architecture, thus enabling the creation of the matchmaking engine (hardware dimensioning). To the best of our knowledge, no public or very partial data is currently available for the proposed task. The dataset might be of interest to other AI experts for the creation of ML and optimization models for similar tasks. (Corresponding Task 6.3–Vertical matchmaking optimization engine. See list of dataset: Dataset 5).

This data can be of interest to different potential users. They may include researchers and companies in the field.

2.2. StairwAI data from open call

Besides data being generated directly or indirectly by the partners within the consortium, there will also be data generated by partners and organizations outside the project consortium, i.e., data of the applicants participating in the project's open calls, through an online form within the FundingBox (FBOX) platform.

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

FSTP data of the third party calls and recipients is of administrative nature and as such should not be part of the DMP. However, for the sake of completeness, we report the dataset description as well as the data treatment.

The information gathered will serve to evaluate and select the most promising SME to involve in the project. Therefore, it is necessary to collect, store and process data from the online forms submitted by project applicants. The datasets to be collected will be provided by applicants in the application forms.

The anonymized datasets will be exploited through the creation of maps and charts that will be updated at the end of the selection process of each open call. The maps and charts generated will be publicly shown as part of the dissemination activities of the project (e.g., at the public StairwAI events). The full dataset of anonymized data will also be available for third parties that would request access to the information for research purposes.

See list of dataset: Dataset 6-10.

3. FAIR Data

This DMP follows the EU guidelines[2] and describes the data management procedures according to the FAIR principles[3]. The acronym FAIR identifies the main features that the project research data must have in order be findable, accessible, interoperable, and re-useable, allowing thus for maximum knowledge circulation and return of investment.

3.1. Making data findable

StairwAI research data are organized in datasets, which are named collections of data units with the same focus and scope. At the moment of publication of project results, each research teams will deposit and describe the relative underlying datasets in institutional or public data repositories that can attribute persistent unique identifiers to the deposited items using the persistent unique identifiers (DOI or Handle), to cite the datasets as underlying data within their research publications. The project plans to provide an overview on the project website that outlines available datasets, their main characteristics, references to relevant publications (if any) and hosting repositories.
The common rules for dataset naming[4] have been set and shared within the Consortium in order to improve data visibility, discoverability, citation and permanent online tracking (see column “title” in Table 8). Each dataset will be described in English language with a minimum of descriptive metadata including information on the data collection methods and processes, the sensibility of the data, usage and access rights, and information on each field of the content and related keywords.

[2] Guidelines on FAIR Data Management in Horizon 2020 (Version 3.0, 26 July 2016), http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf
[3] The FAIR data principles (Force11 discussion forum), https://www.force11.org/group/fairgroup/fairprinciples
[4] Common rules for dataset naming: *PROJECT ACRONYM. WPnumber. WP title or description specifying WP aims. Tasknumber. Task title or description specifying Task aims. additional information specifying coverage and nature of data (if necessary). version number (optional, in case of revisions to help identifying the updates especially in repositories that do not track versioning automatically)*

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

Furthermore, datasets will be annotated with version numbers and timestamps, where the first number of a version changes with a major change in the data collection method (e.g., new survey, new UI features) and the timestamp indicates the termination of a data collection period.
To support the completeness of metadata, the project will provide a metadata template to all stakeholders (see Table 2). The template will be a living document that might be expanded to fit project specific requirements. The initial template will be in line with the metadata standard "Dublin Core Metadata Schema"[5], that is a flexible and commonly used standard, which is also adopted, by the European OpenAIRE repository.
The data will be properly structured and articulated by using the concepts and terms inherent to the methodology involved in the project. To this end, standard metadata vocabularies from the "ISO 15836-1:2017 Information and documentation - The Dublin Core metadata element set — Part 1: Core elements"[6] will be used to deploy an internal metadata system. Eventually, its implementation will be updated with any custom entry required by the project. When being deposited, the dataset will be identified with proper keywords, commonly used in scientific repositories, and subject headings from the "Library of Congress Subject Headings"[7]. In addition, in case of NLP data (see dataset n°2), Metadata corresponding to "European Language Grid metadata schema"[8] will bel exploited.

Whenever applicable, metadata will be structured as follows:

Table 2 - Structure of metadata.

1	Title	A name given to the resource.
2	Creator	An entity primarily responsible for making the resource.
3	Subject	The topic of the resource.
4	Description	e.g., abstract, table of contents, graphics, ...
5	Publisher	Only for published items.
6	Contributor	Entities that contributed to the making of the resource.
7	Date	The termination of the data collection period.
8	Type	[dataset, article, questionnaire, ...]
9	Format	File format of the resource
10	Identifier	e.g., ISSN if your item has been published
11	Source	Which tools were used to collect the data
12	Language	A language of the resource.

In addition to the dataset's metadata document, dataset providers are required to attach additional documents such as:

- A description of the study:
  - Method of research,
  - Applied questionnaires,
  - Data documentation / usage manual,
- Any other information that might be of interest to a data user.

[5] Dublin Core Metadata Schema, https://dublincore.org/schemas/
[6] ISO 15836-1:2017 Information and documentation - The Dublin Core metadata element set — Part 1: Core elements, https://www.iso.org/standard/71339.html
[7] Library of Congress Subject Headings, https://www.loc.gov/aba/cataloging/subject/
[8] European Language Grid metadata schema, https://european-language-grid.readthedocs.io/en/stable/all/1_Introduction/Introduction.html

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

The chosen data repositories (see next section, table 2) support standard descriptive metadata to ensure datasets indexing and discoverability. In particular, they support Dublin Core[9] and DataCite Metadata Schema[10]. Moreover, they comply with the OpenAIRE requirements for data archives. Consequently, the project datasets will be visible via the OpenAIRE portal; facilitating project reporting procedures (see Table 2 for the list of the chosen data repositories).
Selected institutional repositories use the interoperability protocol Open Archives Initiative (OAI-PMH) to increase the visibility of the documents deposited. This protocol allows other applications to collect metadata items to make other products improving their visibility and impact. To facilitate interoperability, standard codes will be followed when possible. In particular, the data will be registered following internal codifications that will be specified within each file.

3.2. Making data openly accessible

As a guiding principle, StairwAI seeks to make research data openly available, whenever possible, in order to allow dissemination, validation and re-use of research results. To this purpose, all the files will be converted to standard and well-documented (open) formats, whenever possible, and the datasets will be deposited together with all relevant documentation and explanation. *Annex I*, of the DMP, indicates the versions or parts of the datasets that cannot be freely shared providing the specific motivations.

The nature of the project foresees project partners and other project stakeholders to collect and provide a range of datasets. Depending on the characteristics of each dataset (e.g., if the data contained present privacy issues, confidentiality issues due to commercia/industrial exploitation, etc.), different access and usage rights will be assigned, and access modalities will be documented.

The project will distinguish between these different access levels:

1. Open access
   a. permission to access: every user
   b. characteristic of the dataset: dataset does not include any personal information (e.g., anonymized data), unless given permission through informed consent
2. Restricted access
   a. permission to access: granted to involved consortium users (sensitive data will be treated only by UNIBO and FBA)
   b. characteristic of the dataset: dataset containing personal and/or sensitive data that cannot be anonymized or aggregated without losing meaning. The data will be deleted after 20 years, as required by GDPR and indicated in the informed consent
3. Embargo
   a. permission to access: dataset will be available to users after a certain period of time
   b. characteristic of the dataset: dataset contains data that cannot be immediately shared, e.g., to ensure the competitiveness of a product, to allow the researchers to publish, etc.

[9] Dublin Core Metadata Schema, https://dublincore.org/schemas/
[10] DataCite Metadata Schema, https://schema.datacite.org/meta/kernel-4.4/

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

Along these lines, the StairwAI project will be linked to the AI4EU platform providing documentation on all collected datasets and access to them, either directly (as in case of restricted access datasets) or through a host repository (as in case open access and embargoed datasets).

The datasets collected during the Open Calls (datasets n°6-10) will be treated as confidential, since they contain personal data, and they will be available only for authorised users (FBA project members) via authentication. FBA will store the data for 6 years from the end of the year in which the Project ended.
This data mainly consists of individual registers collected within open calls that will be only accessible for evaluation purposes to accredited and authorized evaluators. Each evaluator will be granted limited access to a restricted number of registers from the dataset. Before providing the evaluators with access to the data, they will be requested to sign an 'Experts Evaluators Code of Conduct', a 'Guide for External Evaluators' and a 'Declaration of confidentiality and no conflict of interest' consent via secure authentication.
The data collected during the open calls will facilitate good analysis of proposals and will include (non-exhaustive list): Country; Organization name; Sector; Address; Number of team members; Name of the team members; Use case Abstract; Brief description; Company years of experience, etc. This data will be provided by applicants in the application forms.
The data will be anonymized and then exploited through the creation of maps and charts that will be updated at the end of the selection process of each open call. The maps and charts generated will be publicly shown as part of the dissemination activities of the project (e.g., at the public AI4EU events). The full dataset containing anonymized data will also be available for research purposes.

At the time of publication of results, researchers deposit the project data that can be shared in a data repository in order to guarantee their discoverability, access and preservation beyond the project end. At the end of the project, all shareable data will be deposited in a repository. In this case, if the data are not yet published, an embargo could be applied (the length of the embargo will be appropriate to allow the authors to publish).

The data repositories chosen by partners are both institutional and public repositories. They guarantee long term preservation and attribute persistent unique identifiers to the archived datasets (such as DOI or Handle). They support open licenses and different access levels. Finally, they adopt descriptive metadata standards as required by the OpenAIRE Guidelines and allow cross-linking between publications and the relevant datasets.
Each different dataset is deposited by the team that is responsible for the data collection and management in the repository of their choice.

Table 3 shows the repositories for datasets publication and preservation chosen by the partners, their features and their compatibility with OpenAIRE. Table 4 shows the repositories chosen for the preservation of research products/publications.

Table 3 – Summary of data repositories selected by each partner.

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

Repository name	Type	Permanent ID	OpenAIRE compatibility	Catalogued in R3data?	Partner
AMS Acta	Institutional	DOI	OpenAIRE Data (funded, referenced datasets)	https://www.re3data.org/repository/r3d100012604	**UNIBO**
ZENODO	Multi-disciplinary	DOI	OpenAIRE Basic (DRIVER OA)	https://www.re3data.org/repository/r3d100010468	**UNIBO, BCA, EGI, FBA, HBA, HUA, UCC, THA**
Open Access Repository (OAR)	Institutional	DOI	OpenAIRE 3.0 (OA, funding)	-	**INFN**
GitHub	Multi-disciplinary	DOI (via Zenodo)	via Zenodo	-	**TIL[11], TUE, UCC[12]**
UPCommons – Research Data	Institutional	DOI, Handle	OpenAIRE Data (funded, referenced datasets)	https://www.re3data.org/repository/r3d100012607	**UPC**

During the project BCA team will use for data collection GitLab[13], however BCA will deposit data and code for long-term preservation also in ZENODO. Similarly, FBA team will use FundingBox[14] Google Workspace[15] for data collection during the project, also in this case the long-term preservation will be accomplished depositing the data in ZENODO. Instead, TIL team will deposit its data in the European Language Grid (ELG)[16] platform other than in ZENODO.

Table 4 – Summary of repositories selected by each partner to preserve research outputs/publications

Repository name	Type	Permanent ID	OpenAIRE compatibility?	Partner
IRIS UNIBO	Institutional	Handle	OpenAIRE 3.0 (OA, funding)	**UNIBO**
arXiv	Disciplinary	DOI	OpenAIRE Basic (DRIVER OA)	**TUE, TIL**
Cork Open Research Archive (CORA)	Institutional	Handle	OpenAIRE 3.0 (OA, funding)	**UCC**
Open Access Repository (OAR)	Institutional	DOI	OpenAIRE 3.0 (OA, funding)	**INFN**
UPCommons	Institutional	DOI, Handle	OpenAIRE 3.0 (OA, funding)	**UPC**
Eindhoven University of	Institutional	DOI	OpenAIRE Basic (DRIVER OA)	

[11] TIL GitHub repository, https://github.com/tilde-nlp
[12] UCC GitHub repository, https://github.com/ai4eu
[13] BCA GitLab repository, https://gitlab.com/bonseyes
[14] FundingBox, https://fundingbox.com/
[15] Google Workspace, https://workspace.google.com/
[16] European Language Grid (ELG) platform, https://live.european-language-grid.eu/

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

Repository name	Type	Permanent ID	OpenAIRE compatibility?	Partner
Technology research portal				
ZENODO	Multi-disciplinary	DOI	OpenAIRE 3.0 (OA, funding)	**BCA, EGI, HBA, HUA, TUE, THA, TIL**

TUE team will use both arXiv and ZENODO: arXiv for the pre-print versions of its publications, ZENODO for the post-print versions and publisher's versions (in case of gold/hybrid open access publication).

For each deposited dataset, all relevant documentation explaining data collection procedures and analysis (such as codebooks, methodologies, etc.) will be made available along with the data, in order to guarantee intelligibility, reproducibility and the validation of the project findings. Moreover, the deposited documentation specifies the tools and software recommended to reproduce and reuse the data, when necessary. (See Table 5 for examples of tools and software enabling reuse of the dataset).

3.3. Making data interoperable

When possible, all datasets will be described using standard descriptive metadata in order to ensure metadata interoperability for indexing and discoverability. All relevant documentation explaining codebooks, users' manuals, data collection procedures and analysis will be made available along with the data in order to guarantee intelligibility, reproducibility and the validation of the project findings.
In addition, when applicable, codebooks (Jupyter notebooks[17], python scripts) and readme files will be provided to act as metadata.

Whenever applicable, in order to ensure the interoperability of the data, the Dublin Core metadata standard[18] will be adopted. Furthermore, to facilitate interoperability, standard codes will be followed when possible. In particular, the data will be registered following internal codifications that will be specified within each file.
The data will be properly structured and articulated by using the concepts and terms inherent to the methodology involved in the project. To this end, standard metadata vocabularies from the "ISO 15836-1:2017 Information and documentation - The Dublin Core metadata element set — Part 1: Core elements"[19] will be used to deploy an internal metadata system. Eventually, its implementation will be updated with any custom entry required by the project.

To allow data exchange and re-use among researchers, institutions, organisations, countries, etc., partners will convert all shareable data from proprietary formats and will made them available in well-known and documented standard (open) formats (see Table 1, and the dataset descriptive

[17] Jupiter notebooks, https://jupyter.org/
[18] Dublin Core Metadata Schema, https://dublincore.org/schemas/
[19] ISO 15836-1:2017 Information and documentation - The Dublin Core metadata element set — Part 1: Core elements, https://www.iso.org/standard/71339.html

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

tables in *Annex I* for details), as much as possible compliant with available (open) software applications. In case a particular software is used in data processing, full explanation and instructions will be included in the deposited documentation (a summary of the tools and software necessary to reuse of datasets is described in Tab.5).

Table 5 – Summary of tools and software for enabling re-use of the datasets.

Tools/software	Type of data
free/open source document and spreadsheet editors (e.g. LibreOffice suite[20])	Numerical and textual
variety of SW tools (mostly based on Python)	
python scripts	
open-source code (managed through git repositories)	
Jupyter notebooks[16]	code
free/open source audio/video players (e.g. VLC[21])	video

3.4. Increase data re-use

StairwAI distributes the shareable data by adopting licenses that allow re-use of the data and of the datasets in their entirety by other scholars and stakeholders.

The datasets will be made available, unless otherwise stated in the individual data set description sheets in *Annex I*, under the license Creative Commons Attribution 4.0 International (CC- BY 4.0)[22].

Copyright and IP issues are managed in the Consortium Agreement for all involved partners. Information with respect to the terms-of-use of the StairwAI platform will follow in the next version of the DMP.

In general, data are made openly available as underlying data necessary to validate the research results immediately at the time of publication of public reports and scientific papers. Data are cited in the official project publications and website, and they are made available through institutional or public data repositories compliant with OpenAIRE requirements[23]. (See Table 3)
It is possible that an embargo period may be applied to some datasets to allow full exploitation of research results by the partners (see access level type 3 described above). Possible embargoes applied to the datasets will be specified in any updated versions of this DMP.
The research data that are made openly available are deposited in open formats in AMS Acta, Open Access Repository (OAR), UPCommons - Research Data and ZENODO, repositories that guarantee long term preservation to archived items, therefore they will be re-usable by third parties after the end of the project.
The research data that cannot be shared because there is no way to anonymize or aggregate the data without losing its meaning, will not be deposited in repository but will be stored and made available only to authorized team members in charge of collecting them using FundingBox Google Workspace. This data will be conserved for 20 and then deleted.

[20] Libreoffice, https://www.libreoffice.org/
[21] VLC media player, https://www.videolan.org/
[22] Creative Commons Attribution 4.0 International (CC BY 4.0), https://creativecommons.org/licenses/by/4.0/
[23] OpenAIRE, *For Data Providers* https://www.openaire.eu/intro-data-providers

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

The quality of the data will be carefully assured using different approaches. For example, T3.3 Chatbot data (monolingual and multilingual textual data) will be cleaned and preprocessed in order to be used in NLP models; regarding T5.2. matchmaking data, computation and tests will be carried out to make sure the dataset is not biased, and de-biasing techniques will be applied, if necessary. With respect to open calls, the quality of the data will be carefully ensured using different approaches:

1. Data will be collected via platforms dedicated to data collection: e.g., FundingBox platform, and/or dedicated event platforms
2. Data will be stored as files e.g. (.PDFs, .DOCs, .XLSs) in storage clouds such as Google Drive
3. Data will be stored on FundingBox's internal Google Drive (which allows a back-up of each document) according to FundingBox's internal folders structure and naming convention, and the latest versions will be uploaded on the project's Microsoft shared folder.
4. The standard procedures for processing personal data will be implemented, e.g., verification of identity, eligibility check, legal check, data subject check.

3.5. Allocation of resources

Making data FAIR requires an investment of money and researchers' time. In StairwAI case, cost of data preservation in data repository after the project end are null because the chosen repositories do not apply fees for archiving and data curation.
This section gives an overview of the main aspect in relation to costs and responsibilities:

- Cost
  - Costs are included in the project budget.
- Responsibilities
  - The consortium project leader has the main responsibility to communicate to all project members Horizon 2020 FAIR data management and open access obligations, and to coordinate the activities for the preparation of the DMP, that will involve all the partners.
  - Each project member is responsible for implementing the DMP within his respective work packages (WP).
  - All the data collected in the StairwAI project will be integrated as part of the AI4EU platform (or linked to it).
  - Each consortium member is responsible for the data that will collect/generate and process.

Responsible for data management are the datasets creators who are generally the team leaders directly involved in research data organization and collection (see Table 6).

Table 6 – Summary and contacts of the research team leaders.

Team	Leader	ORCID ID (if available)	mail
UNIBO	Lombardi, Michele	0000-0003-4709-8888	michele.lombardi2@unibo.it
	Borghesi, Andrea	0000-0002-2298-2944	andrea.borghesi3@unibo.it
BCA	Llewellynn, Tim		tim@bonseyes.com
	Bonnefou, Jean-Marc		jean-marc@bonseyes.com
FBA	Milcent, Lucie		lucie.milcent@fundingbox.com
TIL	Skadina, Inguna		Inguna.Skadina@tilde.lv

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

Team	Leader	ORCID ID (if available)	mail
UCC	Gonzales-Castañè, Gabriel	0000-0003-0486-1492	gabriel.castane@insight-centre.org

Moreover, partners are encouraged to identify and cite all contributors participating in data management activities, specifying their roles according to a given standard vocabulary (DataCite Metadata Schema[24]). The already known contributors are listed in table 7 (see *Annex I* for details about data management responsibilities related to each project dataset); other contributors are likely to be involved in the future activities, for instance in tasks 2.1 and 3.4 where the collection of data from SMEs could benefit from the involvement of intermediaries relevant to the sector. The table will be updated accordingly.

Table 7 – Summary of team members involved in the datasets collection and management.

Team	Member	ORCID ID (if available)	Role
UNIBO	Milano, Michela	0000-0001-7379-1411	Project Member
	Boscarino Andrea		Researcher
	Calegari, Roberta	0000-0003-3794-2942	Researcher
	Ciatto Giovanni	0000-0002-1841-8996	Researcher
	De Filippo, Allegra	0000-0002-1954-7271	Researcher
	Sabbatini Federico	0000-0002-0532-6777	Researcher
	Tagliavini Giuseppe		Researcher
UCC	Genc, Begum	0000-0003-0116-6005	Researcher

Keys for "Role" column: Data Collector (such as survey conductors, interviewers…), Producer (person responsible for the form of a media product), Project Member (a researcher indicated in the GA), Researcher (an assistant to one of the authors who helped with research, data collection, processing and analysis but is not part of team indicated in the GA), Research Group (the name of a research institution or group that contributed to the dataset).

3.6. Data security

At each institution, research data will be stored in intranets or hard-drives accessible through institutional password periodically modified according to national law provisions for data security and protected by regularly updated antiviruses. None of the project data will be left inadvertently available. All the research materials stored in computers are subject to regular backup in order to safeguard them from accidental losses.

Backups will be conducted in regular intervals. Currently, weekly full backups are performed.
Each project partner gets individual access to the datasets. The research partners will get unrestricted read-access to the data. Commercial partners will get access based on project needs. Data access will be given to named users. Relevant transactions will be logged within the system. All the data collected in the StairwAI project will be integrated as part of the AI4EU platform (or linked to it), utilizing the security measurements in place.

Long term preservation of shareable data is ensured by the chosen data repositories that have specific preservation policies. UNIBO AMS Acta guarantees long term preservation to the archived materials also thanks to a deposit agreement with the National Central Library in Florence. Zenodo policy ensures that the items will be retained for the lifetime of the repository and in case of closure,

[24] DataCite Metadata Schema, https://schema.datacite.org/meta/kernel-4.4/

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

best efforts will be made to integrate all content into suitable alternative institutional and/or subject based repositories. The storage of the datasets within the UPCommons repository will provide support for its correct duplication and preservation. All the responsibilities of data recovering, and secure storage, are of the repository storing the dataset.
(See Table 3 for details).

3.7. Ethical aspects

Partners in charge of data collecting will follow the protocols described for each task and will comply with this Data Management Plan and with the Ethics Deliverable of Work Package 9 “Ethics Requirements”. Recruited participants will be informed in an unambiguous way about what is collected, why it is collected, how it is collected, who is the data controller and who is the jointly controller (if any). Informed consent will be requested where necessary each time it will be needed during the life of the project and depending on the activities carried out. Data acquisition will be restricted only to information essential to the success of the project. Data collection activities will follow rigorous confidentiality rules to ensure the participants’ privacy and only aggregated or anonymized data will be shared between partners when possible.
During the life of the project, six Deliverables will be submitted within the Work Package 9. The Consortium has already submitted a deliverable in which a detailed information on the informed consent procedures about data processing has been presented (DL9.4-POPD, Requirement no.7) Furthermore, at the same time of the submission of this DMP, the Consortium will submit three other deliverables in which the information presented is the following:

- DL 9.2 OEI – Requirement No.2: In this deliverable a justification and criteria for choosing specific European Language for the development of natural language services has been provided.
- DL 9.3 POPD – Requirement No.6: In this derivable a description of the anonymization and pseudonymization techniques that will be implemented by partners who will deal with personal data (UNIBO, TILDE, FBA) has been detailed.
- DL 9.6 POPD – Requirement No.9: In this deliverable a detailed description of responsibilities of each partner and Consortium as a whole with regards to data processing has been further provided and explained. Specifically, the Deliverable consists of an analysis of all the Work Packages and subsequent subtasks in which personal data will be collected and the data controllers and any joint controllers are then defined. From the analysis carried out it emerged that there will be three partners who will collect personal data and who will manage the data in accordance with the rules of the GDPR (UNIBO, TILDE and FBA).

At the end of the first year of the project (M12), the DL9.1 - POPD Requirement No.1 will also be submitted, here the Consortium will provide explanation on how the data subjects will be informed of the existence of the profiling, its possible consequences and how their fundamental rights will be safeguarded. Finally, at the end of the project (M36) all the templates of the informed consent forms and information sheet will be presented as a deliverable.
Further detailed information on how all the data will be managed during and after the project is available in the above-mentioned deliverable of Work Package 9 “Ethics requirements”.

3.8. Other issues

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

The STAIRWAI project does not adopt any other procedures for data management than the ones mentioned in the previous sections of this DMP.

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

4. Datasets overview

The following table (Table 8) offers an overview of the datasets expected from the project which are described more in detail in *Annex I*. It will be updated according to DMP changes and variations.

Table 8 – Datasets list.
Table acronyms and abbreviations: n°= dataset progressive number, LB = WP lead beneficiary, CT = creator team in charge of curating the dataset, C=collected, G=generated, A=available, IP=in progress, NYA=not yet available.

n°	WP	LB	TASK or SUBTASK	CT	DATASET (tentative title)	SOURCE	STATUS
1	**WP2 WP3**	UNIBO EGI	**Task 2.1 Task 3.4**	**UNIBO**	StairwAI. WP3. Knowledge.T3.4. Matchmaking Data. Use Cases and Solutions. v1.0	**C**	**NYA**
2	**WP3**	TIL	**Task 3.3**	**TIL**	StairwAI. WP3. Knowledge Task3.3. Datasets for the StairwAI Chatbot v1.0	**C**	**NYA**
3	**WP5**	UCC	**Task 5.2**	**UCC**	StairwAI. WP5. Platform knowledge, job offers and experts/consultants. T5.2. Matchmaking Data. Job matching. v1.0	**C**	**NYA**
4	**WP6**	BCA	**Task 6.1 Task 6.2**	**BCA**	StairwAI WP6 - Vertical Matchmaking T6.1 Cross-platform benchmarking T6.2 LPDNN benchmarking and benchmarking results v1.0	**G**	**IP**
5	**WP6**	UNIBO	**Task 6.3**	**UNIBO**	StairwAI. WP6. Vertical Matchmaking.T6.3 Optimization Engine. v1.0	**G**	**IP**
6	**WP3**	FBA	**Task 3.5 Task 7.1 Task 7.2 Task 7.3**	**FBA**	StairwAI. WP3. Knowledge.T3.5. Applications. v1.0	**C**	**NYA**
7	**WP7**	FBA	**Task 7.2**	**FBA**	StairwAI. WP7. Open Call. T7.2. Event Attendees. v1.0	**C**	**NYA**
8	**WP7**	FBA	**Task 7.3**	**FBA**	StairwAI. WP7. Open Call. T7.3. Beneficiary List Fstp. v1.0	**C**	**NYA**
9	**WP7**	FBA	**Task 7.3**	**FBA**	StairwAI. WP7. Open Call. T7.3. Evaluation expert and mentors. v1.0	**C**	**NYA**
10	**WP8**	FBA	**Task 8.2**	**FBA**	StairwAI. WP7. Open Call. T7.3. Stakeholders. v1.0	**C**	**NYA**

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

Annex I: Datasets tables

The analytic descriptions of the expected datasets of StairwAI project are reported in this Annex organized by work-packages.
For collecting further information from the project partners about the data management procedures, the following questionnaire has been used, the template is showed in Annex II.

DATASET 1 - WP2-WP3

- Task 2.1– Requirements for the horizontal matchmaking layer and reputation mechanisms
- Task 3.4– Data collection and datasets generation for Horizontal and Vertical Matchmaking

1	**Not yet available (expected in autumn 2021)**	***StairwAI. WP3. Knowledge.T3.4. Matchmaking Data. Use Cases and Solutions. v1.0***
DOI		Not yet available
Version		v01
Team in charge		UNIBO
Creator/s		**Lombardi, Michele** [UNIBO]
Contributor/s		**Gonzales- Castañe, Gabriel** [UCC]; **Calegari, Roberta** [UNIBO]; **Ciatto, Giovanni** [UNIBO]; **Sabbatini, Federico** [UNIBO]; **Milano, Michela** [UNIBO]
Contact Person/s		**Lombardi, Michele** [UNIBO, michele.lombardi2@unibo.it]
Contents		This dataset will contain description of industrial use cases that have been addressed or could conceivably be addressed via AI techniques, plus annotations about the AI techniques applied to the problem (when applicable). Use case description are in semi-structured form (a few paragraphs in natural language, aimed at different aspects of the use case), while annotations about the AI solution that may have been used to addressed are structured (via multiple choice questions). After collection is finalized, the dataset is expected to have several examples ranging from a few hundreds to slightly more than one thousand, and the data will be analyzed via statistical and Artificial Intelligence methods. Data will be collected by sending a form (questionnaire) to both AI-savvy and non AI-savvy companies, with an emphasis on SMEs. To the best of the collectors' knowledge, no similar dataset is available. The dataset may be of interest for the definition of AI tools aimed at classifying AI use cases.
Data format		Text based format (e.g. csv)
Data volume		5-10 MB
Accessibility		Raw data containing personal information will be available only to the project coordinator (UNIBO). It may be shared with follow-up projects sharing similar objectives, only in those case where explicit consent was given. In the StairwAI project, personal data will be occasionally used to obtain clarifications, and to contact potential beta-testers for the developed system, in those case where explicit consent was given.

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

1	**Not yet available (expected in autumn 2021)**	***StairwAI. WP3. Knowledge.T3.4. Matchmaking Data. Use Cases and Solutions. v1.0***
		Non-personal data related to industrial use cases and their solution will be shared with all partners for the duration of the project via private repository on AMS acta or Zenodo. All non-personal data will also be made public on AMS acta or Zenodo under the CC0 license at the end of the project. Part of the data may be published earlier (in the same repositories and under the same license) in case it is needed to validate scientific publications.
Repository		AMS Acta
Related publication/s		Not yet available

DATASET 2 - WP3

- Task 3.3– Data collected and datasets generation for Multi-lingual NLP

2	**Not yet available**	***StairwAI.WP3. Task3.3. Datasets for the StairwAI Chatbot v1.0***
DOI (via ZENODO)		Not yet available
Version		v01
Team in charge		TIL
Creator/s		**Skadina, Inguna** [TIL]
Contributor/s		Not yet available
Contact Person/s		**Skadina, Inguna** [TIL]
Contents		This is the collection of monolingual and multilingual textual data, necessary for the development of StairwAI Chatbot. The dataset, that will facilitate research in human-computer interaction and could be used in research and development activities related to this topic, will include: - data for user stories and dialogue scenarios reflecting content of the eventual conversations with the end users of the solution (e.g., topical questions, themes, issues etc.); - data for development and domain adaptation of the multilingual Natural Language Understanding (NLU) modules and other Natural Language Processing (NLP) tools; - parallel data and respective terminology for development and domain adaptation of the Machine Translation (MT) engines for the respective language combinations.
Data format		Textual (csv, txt)
Data volume		Not yet available

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

2	Not yet available	***StairwAI.WP3. Task3.3. Datasets for the StairwAI Chatbot v1.0***
Accessibility		Data will be available under Creative Commons Attribution 4.0 International (CC BY 4.0) license. Access to the part of data needed to validate the results presented in scientific manuscripts will be given immediately at the time of publication.
Repository		GitHub & ZENODO (a copy of the data will be deposited also in ELG[25] platform)
Related publication/s		Not yet available

DATASET 3 - WP5

- Task 5.2– Matchmaking between job offers and experts/consultants

3	Not yet available (expected in autumn 2021)	***StairwAI. WP5. Platform knowledge, job offers and experts/consultants. T5.2. Matchmaking Data. Job matching. v1.0***
DOI		Not yet available
Version		v01
Team in charge		**UCC**
Creator/s		**Gonzalez-Castañé, Gabriel** [UCC];
Contributor/s		**Genc, Begum** [UCC]; **Lombardi, Michele** [UNIBO]; **Calegari, Roberta** [UNIBO]; **Ciatto, Giovanni** [UNIBO]; **Sabbatini, Federico** [UNIBO]; **Milano, Michela** [UNIBO]
Contact Person/s		**Genc, Begum** [UCC, begum.genc@insight-centre.org]
Contents		This dataset will contain available job offers - industrial, academic, expressions of interest - listed on Europe's AI-on-demand platform (AI4EU). The data is expected to be collected by an online form, which results in having semi-structured data including some categorical data such as location, key skills, etc.; text blocks that requires natural language processing; and numerical data such as the years of experience. The finished dataset is expected to contain a few hundred job listings, to the best of the collector's knowledge, no similar dataset is available. The dataset may be of interest to people with an intention to develop AI algorithms that match jobseekers to available jobs. Furthermore, anonymised user profiles (researchers, developers, domain experts, citizens) extracted from the AI4EU platform with user / experts' preferences on AI topics will be necessary to conduct adequately this task. However, we understand that this dataset will be a horizontal need required by all the tasks in WP5. The data will be analyzed via statistical and Artificial Intelligence methods. Necessary computation and tests will be carried out to make sure the dataset is not biased, and de-biasing techniques will be applied, if necessary.

[25] European Language Grid (ELG) platform, https://live.european-language-grid.eu

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

3	**Not yet available (expected in autumn 2021)**	***StairwAI. WP5. Platform knowledge, job offers and experts/consultants. T5.2. Matchmaking Data. Job matching. v1.0***
Data format		Text based format. Job description datasets stored in csv file format, where each row corresponds to a distinct job description, and each feature of the job is separated by a comma
Data volume		< 25 MB
Accessibility		Participants' publication consent will be explicitly asked for in the collection questionnaire: authorized data will be anonymized (by removing personal information) in order to be openly shared. Shareable data will be deposited in repository and will be available under CC BY license. Access to the part of shareable data needed to validate the results presented in scientific manuscripts will be given immediately at the time of publication. An extended version of the job description dataset will be available only to those project partners that need the full set, in order to progress the development of the tasks they are leading: this version will contain data about use cases that were authorized by participants to be used only for the StairwAI project purpose. The task of the collection of user profiles is horizontal to all the tasks in WP5. If any additional need for user profile collection emerges to complete the T5.2, profile data collection will be supported to be made available to the project coordinator team and the project partners who may work further on the anonymisation of user profiles (if consent is given by the data owner). In this case, the personal data may be collected and used for ensuring the fairness of the matching algorithm (e.g. elimination of negative discrimination).
Repository		ZENODO
Related publication/s		Not yet available

DATASET 4 - WP6

- Task 6.1– Development of cross-platform benchmarking for AI applications on heterogeneous hardware
- Task 6.2– Deployment and verification of LPDNN benchmarking framework on hardware platforms and generation of benchmarking results

4	**In progress (expected in autumn 2021)**	***StairwAI WP6 - Vertical Matchmaking T6.1 Cross-platform benchmarking T6.2 LPDNN benchmarking and benchmarking results v1.0***
DOI		Not yet available
Version		v01

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

4	**In progress (expected in autumn 2021)**	***StairwAI WP6 - Vertical Matchmaking T6.1 Cross-platform benchmarking T6.2 LPDNN benchmarking and benchmarking results v1.0***
Team in charge		**BCA**
Creator/s		**Llewellynn, Tim** [BCA]; **Bonnefous, Jean-Marc** [BCA]; **Lombardi, Michele** [UNIBO]
Contributor/s		**Tenuhen, Ville** [EGI]; **Milano, Michela** [UNIBO]; **Brasche, Goetz** [HUA]; **Cesini, Daniele** [INFN]
Contact Person/s		**Bonnefous, Jean-Marc** [BCA, jean-marc@bonseyes.com]
Contents		The vertical matchmaking will employ a set of metrics on efficiency, model description and platform specifications to identify the best solution for a given AI challenge. The metrics that will be considered can be hardware agnostic, e.g., number of parameters and operations of an AI model, or hardware dependent, e.g., latency, memory, accuracy, power consumption. Metrics such as accuracy need a validation dataset against which the accuracy of the AI model can be computed. In that regard, the dataset required for assessing the accuracy are mostly open-source datasets. By contrast, metrics such as latency and memory do not need a complete dataset, they only need the format, e.g., image format, of those data samples used for assessing the accuracy.
Data format		Numerical and textual
Data volume		25 MB
Accessibility		The dataset will be anonymized (by removing personal information). Shareable anonymized data will be deposited in repository and will be available under Creative Commons Attribution 4.0 International (CC BY 4.0) license. Access to the part of shareable data needed to validate the results presented in scientific manuscripts will be given immediately at the time of publication.
Repository		GitHub & ZENODO (a copy of the data will be deposited also in BCA GitLab and will be available at least for the duration of the StairwAI project for the development of the matchmaking service layer for the AI on demand platform)
Related publication/s		Not yet available

DATASET 5 - WP6

- Task 6.3– Vertical matchmaking optimization engine

5	**In progress (expected in autumn 2021)**	***StairwAI. WP6. Vertical Matchmaking.T6.3 Optimization Engine. v1.0***
DOI		Not yet available
Version		v01
Team in charge		**UNIBO**

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

5	**In progress (expected in autumn 2021)**	***StairwAI. WP6. Vertical Matchmaking.T6.3 Optimization Engine. v1.0***
Creator/s		**De Filippo, Allegra** [UNIBO]; **Borghesi, Andrea** [UNIBO]
Contributor/s		**Milano, Michela** [UNIBO]; **Boscarino, Andrea** [UNIBO]; **Tagliavini, Giuseppe** [UNIBO]
Contact Person/s		**Borghesi, Andrea** [UNIBO, andrea.borghesi3@unibo.it]
Contents		The main sources of data for T6.3 will be the T6.1 and T6.2, as T6.3 is mostly focused on the creation of the vertical matchmaking engine. The dataset will consist of benchmarks for a range of AI applications (e.g., ML models, online and offline algorithms, optimization models) executed on a variety of heterogeneous resources (i.e. hardware platforms). AI algorithms will be run and performance metrics will be collected using a variety of SW tools (mostly based on Python). The collected data (e.g., performance metrics) will enable the construction of ML models to estimate the behavior of an AI application on different hardware architecture, thus enabling the creation of the matchmaking engine (hardware dimensioning). To the best of our knowledge, no public or very partial data is currently available for the proposed task. The dataset might be of interest to other AI experts for the creation of ML and optimization models for similar tasks.
Data format		.csv
Data volume		Not yet available
Accessibility		The data contained in this dataset are artificially generated and will be openly shared. The data will be available under Creative Commons Attribution 4.0 International (CC BY 4.0) license. Access to the part of data needed to validate the results presented in scientific manuscripts will be given immediately at the time of publication.
Repository		AMS Acta. All the data will be shared will be shared through public repositories (e.g. ZENODO). In general, open-source code to handle the data (e.g. python scripts, Jupyter notebooks, etc.) will be published using git repositories, such as GitHub (also in this case the code will be long-term preserved in ZENODO, through the GitHub-ZENODO plugin).
Related publication/s		Not yet available

DATASET 6-11 - WP3-WP8

- Task 3.5–Data Management of data produced in the Open Calls and Tasks 7.1-7.4–Open Call
- Task 8.2–DIH Connection, Regional Interactions and Industry Clustering

6	**Not yet available**	***StairwAI. WP3. Knowledge.T3.5. Applications. v1.0***
ID [ID type]		Not yet available
Version		v01

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

6	**Not yet available**	***StairwAI. WP3. Knowledge.T3.5. Applications. v1.0***
Team in charge		**FBA**
Creator/s		[FBA]
Contributor/s		**Milcent, Lucie** [FBA]
Contact Person/s		**Milcent, Lucie** [FBA, lucie.milcent@fundingbox.com]
Contents		The dataset contains data related to StairwAI Open Calls, mainly data provided through electronic applications on FundingBox Platform. Dataset name: **APPLICATIONS** (drafts, submitted, 3 Open Calls, 1 Call for Adopters and 2 Pilot Calls, Experts Open Call and HW Call and Validation) The dataset will contain all the information related to applicants. **WP3 - Platform knowledge and community organisation** **Task 3.5** - Data Management of data produced in the Open Calls This task is devoted to the collection of the data produced by the applicants in the open calls (Task 7.2). This data will be collected in deliverable D3.6 by FBA and it will be then used in Tsk 3.6 to refine the knowledge models, thus the participation of UPC. It might be used also to enhance the models and components in WP4, WP5 and WP6. **WP7. Open Call Management** **Task 7.1**. Open Calls Preparatory Tasks This task will start 2 months before launching the Open Calls and will include the following tasks: 1). Challenges Definition. 2). Open Call Package of Documents. 3). The Open call management tool will be setting up and customized in FundingBox Platform **Task 7.2. Open Calls launch, management and dissemination**. StarwAI will launch 3 Open Calls, 1 Call for Adopters and 2 Pilot Calls. Once the Open Call is launched, and during the 2 months until the deadline, FBA will coordinate the following tasks: 1). Help desk to support applicants regarding the formal aspects of the application process and IT support for the Open Call Tool. 2). Open Call dissemination 3). Open Call monitoring. **Task 7.3 Evaluation of proposals & FSTP Agreements**. The evaluation of the proposals will be done in the following steps: 1). Proposals reception exclusively through FundingBox Platform. 2). Eligibility checks. 3) External evaluation. 4) Consensus Meeting. 5) Jury Day FBA will organise an online Jury Day 6). Communication of results. 7). Legal Validation. (8). FSTP Agreement signature. (9). Ethical Review
Data format		.ppt, .pdf, .xls, mp4
Data volume		Not yet available (the whole volume of the data related to StairwAI Open Calls managed by FBA is estimated to be 8 GB)
Accessibility		Raw data cannot be openly shared because they contain personal information, they will be locally archived on the FundingBox Enterprise platform, and on the FundingBox Google Drive, in accordance with the retention policy of the data controller, which is FundingBox Accelerator Sp. z o.o. Data will be anonymized and aggregated and exploited through the creation of maps and charts that will be publicly shown as part of the dissemination activities of the

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

6	Not yet available	***StairwAI. WP3. Knowledge.T3.5. Applications. v1.0***
		project (e.g., at the public StairwAI events). The anonymised data, archived in the FundingBox Google Drive, will also be available for third parties for research purposes.
Repository		• FundingBox Enterprise platform, • FundingBox Google Drive
Related publication/s		Not yet available

7	Not yet available	***StairwAI. WP7. Open Call. T7.2. Event Attendees. v1.0***
ID [ID type]		Not yet available
Version		v01
Team in charge		**FBA**
Creator/s		[FBA]
Contributor/s		**Milcent, Lucie** [FBA]
Contact Person/s		**Milcent, Lucie** [FBA, lucie.milcent@fundingbox.com]
Contents		The dataset contains data related to StairwAI Open Calls, mainly data provided through electronic applications on FundingBox Platform. Dataset name: **EVENT ATTENDEES** (webinars, info days, jury day, consensus meeting) The event attendees' names, surnames and emails will be collected by FBA through the registration form. These data will be used to contact participants only about matters related to the event (materials, video recording etc.). The data will not be shared.
Data format		.ppt, .pdf, .xls, mp4
Data volume		Not yet available (the whole volume of the data related to StairwAI Open Calls managed by FBA is estimated to be 8 GB)
Accessibility		Raw data cannot be openly shared because they contain personal information, they will be locally archived on the FundingBox Enterprise platform, and on the FundingBox Google Drive, in accordance with the retention policy of the data controller, which is FundingBox Accelerator Sp. z o.o.
Repository		• FundingBox Enterprise platform • FundingBox Google Drive
Related publication/s		Not yet available

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

8	Not yet available	***StairwAI. WP7. Open Call. T7.3. Beneficiary List Fstp. v1.0***
ID [ID type]		Not yet available
Version		v01
Team in charge		**FBA**
Creator/s		[FBA]
Contributor/s		**Milcent, Lucie** [FBA]
Contact Person/s		**Milcent, Lucie** [FBA, lucie.milcent@fundingbox.com]
Contents		The dataset contains data related to StairwAI Open Calls, mainly data provided through electronic applications on FundingBox Platform. Dataset name: **BENEFICIARY LIST FSTP** (SubGrant Agreements) **Task 7.3 Evaluation of proposals & FSTP Agreements**. The list of FSTP beneficiaries will include the following data: names of selected SMEs, countries, description of the selected AI solutions.
Data format		.ppt, .pdf, .xls, mp4
Data volume		Not yet available (the whole volume of the data related to StairwAI Open Calls managed by FBA is estimated to be 8 GB)
Accessibility		The list of FSTP beneficiaries will be publicly shared in the deliverables D7.4, D7.5 and D7.6, and will be published on the website. The list will include the names of the selected SMEs yet will not include personal data.
Repository		FundingBox Google Drive
Related publication/s		Not yet available

9	Not yet available	***StairwAI. WP7. Open Call. T7.3. Evaluation expert and mentors. v1.0***
ID [ID type]		Not yet available
Version		v01
Team in charge		**FBA**
Creator/s		[FBA]
Contributor/s		**Milcent, Lucie** [FBA]
Contact Person/s		**Milcent, Lucie** [FBA, lucie.milcent@fundingbox.com]
Contents		The dataset contains data related to StairwAI Open Calls, mainly data provided through electronic applications on FundingBox Platform. Dataset: **EVALUATION EXPERT AND MENTORS** Concerning the recruitment of external AI experts and evaluators, a Call for Expressions of Interest will be launched using the FBOX Platform. The data collected through this call will be locally archived on the FundingBox Enterprise platform, and on the FundingBox Google Drive, in accordance with the retention policy of the data controller, which is FundingBox Accelerator Sp. z o.o.

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

9	Not yet available	***StairwAI. WP7. Open Call. T7.3. Evaluation expert and mentors. v1.0***
Data format		.ppt, .pdf, .xls, mp4
Data volume		Not yet available (the whole volume of the data related to StairwAI Open Calls managed by FBA is estimated to be 8 GB)
Accessibility		The data will only be shared with the partners involved in T7.4.
Repository		FundingBox Google Drive
Related publication/s		Not yet available

10	Not yet available	***StairwAI. WP7. Open Call. T7.3. Stakeholders. v1.0***
ID [ID type]		Not yet available
Version		v01
Team in charge		**FBA**
Creator/s		[FBA]
Contributor/s		**Milcent, Lucie** [FBA]
Contact Person/s		**Milcent, Lucie** [FBA, lucie.milcent@fundingbox.com]
Contents		The dataset contains data related to StairwAI Open Calls, mainly data provided through electronic applications on FundingBox Platform. Dataset: **STAKEHOLDERS** FBA will engage DIHs to become Supportive Partners of the project, through informative webinars. The data (Name of the DIH and contact details) will be stored on the FundingBox Google Drive.
Data format		.ppt, .pdf, .xls, .mp4
Data volume		Not yet available (the whole volume of the data related to StairwAI Open Calls managed by FBA is estimated to be 8 GB)
Accessibility		Raw data cannot be openly shared because they contain personal information, they will be locally archived on the FundingBox Enterprise platform, and on the FundingBox Google Drive, in accordance with the retention policy of the data controller, which is FundingBox Accelerator Sp. z o.o.
Repository		FundingBox Google Drive
Related publication/s		Not yet available

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

Annex II: Dataset questionnaire

Here is showed the questionnaire used to collect information from project partners about the their data and data management procedures.

1. Data set reference, name and description
1.A.Data set name
Name: *<Insert here>* *[GUIDANCE: Provide a name or a title for the data set. A data set can be defined as a named collection of data units with the same focus and scope.* *For homogeneity reasons, data sets should be named as follows:* ***"PROJECT ACRONYM. WPnumber. WP title or short description specifying WP aims. Tasknumber. Task title or short description specifying Task aims: additional information specifying coverage and nature of data (if necessary). version number (optional, in case of revisions to help identifying the updates, especially when depositing in repositories that do not track versioning automatically)"*** *.* *We recommend to follow similar rules for file naming, adding version number and version date.]*
1.B.Data set ID
ID: *<Insert here>* ID type: *<Insert here>* *[GUIDANCE: Provide an identifier for your data set. Persistent identifiers are generally associated to a data set at the moment of archiving, for dissemination or preservation, in a repository.* ***If the data set has no persistent identifier at the moment, just type "not yet available". You could indicate it later.*** *A permanent unique identifier can be used as a reference to a resource even beyond the resource lifetime. It facilitates online tracking and citation.* *In addition, indicate the identifier type ("ID type"), such as for example: DOI, Handle, PURL, URL, URN.* ***Check the type of persistent identification used by your repository service****.]*
1.C.Data set origin, WP and current status
1.C.1.Indicate if the data set is: □ Collected | □ Generated *[GUIDANCE: Collected data sets are data sets that use existing data sources; generated data sets are based on data created by the project. Some data sets may contain both collected and generated data. In this case, select both the options.]* 1.C.2.Indicate the WP and the task: WP*<Insert here>* | Task*<Insert here>* | Subtask*<Insert here>* 1.C.3.Indicate the data set current status (choose one option from the guidance): *<Insert here>* *[GUIDANCE: To indicate the data set current status use one of the following:* *1)* ***available*** *(i.e. the final version of the data set is completed and ready to be published/has been published)* 2) ***in progress*** *(i.e. at the moment the data are being collected)* 3) **not yet available** *(i.e. the data set is associated to a task/subtask that has not started yet)]*
1.D.Data set creators and contributors
1.D.1.Data set creator/s Creator name: *<family, given name>* ORCID *(if available)*: *<Insert here>* Affiliation: *<Insert here>* e-mail address: *<Insert here>* *(Fields may be repeated if necessary.)* *[GUIDANCE: The data sets creators are identified among WP coordinators and research team leaders responsible for the data collection, organization, processing and management. Partners are encouraged to identify and cite all contributors, specifying their roles and ORCID. ORCID is a unique persistent identifier that can be attributed to any*

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

academic author. Many universities require an ORCID for their researchers. Registration is free of charge for individuals. Further details about ORCID advantages and functions can be obtained at the ORCID web site http://orcid.org.]

1.D.2.Contact Person/s

Contact Person name: *<family, given name>*

Indicated the following details for the Contact Person ***only if not already provided****:*

ORCID *(if available*): *<Insert here>*

Affiliation: *<Insert here>*

e-mail address: *<Insert here>*

(Fields may be repeated if necessary.)

[GUIDANCE: The contact person is a person with knowledge of how to access, troubleshoot, or otherwise address issues related to the data set. Moreover, the contact person must be a person with a stable assignment inside the institution]

1.D.3.Rights Holder/s (*If not yet available, these fields may be filled at later stages.*)

Rights Holder name: *<institution's legal name>*

(Fields may be repeated if necessary.)

[GUIDANCE: The rights holders are usually the public or private institution/s to which belong the data set creator/s OR the institution/s on behalf of which the data set was created. If more than one institution is involved, the rights are shared among them.]

1.D.4.Other Contributor/s (*Involved in the data set production and management.*)

Contributor name: *<family, given name and/or name of organization>*

Contributor type: *<choose an option from the guidance>*

ORCID *(if available and applicable): <Insert here>*

Affiliation *(if applicable)*: *<Insert here>*

e-mail address: *<Insert here>*

(Fields may be repeated if necessary.)

[GUIDANCE:

1) ***Data Collector*** *(such as survey conductors, interviewers…)*
2) ***Producer*** *(person or organization responsible for the form of a media product)*
3) ***Project Member*** *(a researcher indicated in the GA)*
4) ***Researcher*** *(an assistant to one of the authors who helped with research, data collection, processing and analysis but is not part of team indicated in the GA)*
5) ***Research Group*** *(the name of a research institution or group that contributed to the data set)*
6) *Other (****specify****)]*

1.E.Data set Description

Provide an abstract of the data set. Include all relevant information listed in the guidance.

<Insert here>

[GUIDANCE: Provide a description of the data that will be generated or collected, the intended goals and to whom it may be useful. Include the following questions:

- *nature of the data (for example: experimental observations; survey results; interview transcripts; simulation data; models; software; diaries; lab notebooks; …);*
- *the data scale (for example the number of the analyzed interviews or samples …);*
- *give information on the existence (or not) of similar data;*
- *if new data are created, explain why no-suitable existing data are available, analyze the gap between previous or current data and yours; describe limits of previous works and how you will improve results in your project;*

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

- *the data sources (in case it is collected): provide full citation of the data sources;*
- *specify the type of relationship between the data set and its sources (it may be derived, an updated version; it may reuse only a part of the original sources);*
- *describe potential users and how the data can be reused by them.]*

1.F.Related publications

If already available, provide full citation of the project publications that present the dataset. Include publication ID such as DOI, Handle or URL.

<Insert here>

[GUIDANCE: Indicate whether the dataset underpins a scientific publication and provide the full citation. ***If not yet available just indicate "not yet available".****]*

2. Standards and metadata

2.A.Methodologies for data collection or generation, data processing and quality assurance

Provide a description including issues listed in the guidance.

<Insert here>

[GUIDANCE:
- *Describe how the data will be collected or generated and which community data standards (if any) or methodologies will be used.*
- *Describe the unit of analysis and the procedures used to process the data (for example the methods of aggregation of raw data).*
- *Indicate how the data will be managed and organized during the project, mentioning for example back up and security measures, folder structures, file naming conventions, different version control and naming.*
- *Describe procedures for quality assurance that will be carried out on the data at the time of collection or generation, data entry, digitization and revision or validation (e.g. quality control measures can include bias and/or scale of measurement; protocols for recording raw data; controlled vocabularies, code lists and choice lists; computer aided procedures to check data completeness; data cleaning procedures …).]*

2.B.Describe typologies, content and quality of data

Provide a description of data. Indicate whether the data are one or more of the following:

- numerical, textual – graphical, visual or tactile
- created in digital form (born digital) or converted (digitized)
- quantitative o qualitative data
- raw, derived or secondary
- cleaned or processed

<Insert here>

[GUIDANCE:
- *Data may be presented in a tabular, textual, graphical form, or may be audio or video.*
- *Data may be directly created in a digital form or may be later converted, e.g. a digitization of a printable document.*
- *Qualitative data deal with descriptions and cannot be measured; quantitative data is information about quantities that can be measured.*
- *The raw materials are the primary sources of the research process, they represent the records of research or events as first described. Secondary sources are based on primary sources. These sources analyze, describe, and synthesize the primary or original source. 'Derived data' means augmented/enhanced/cooked/adjusted data in analytic data processing.*
- *Cleaned data have undergone a manual process for detecting and removing errors and inconsistencies from data in order to improve their quality. Data processing involves several processes other than simply cleaning data, e.g. validation, sorting, aggregation, analysis, categorizing, etc...]*

2.C.Data format

Indicate formats used for the following, when available and applicable:

Formats used for raw data: *<Insert here>*
Formats used in data processing: *<Insert here>*

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

Formats used for sharing and preservation: *<Insert here>*

[GUIDANCE: Indicate the file formats used in data collection/ generation and processing. Propriety formats should be converted for reuse and long term preservation. The most accepted file formats are listed here http://www.data-archive.ac.uk/create-manage/format/formats-table.]

2.D.Metadata

Indicate which metadata standard will be used to describe the data set and indicate which documentation will be provided along with data files.

<Insert here>

[GUIDANCE: Indicate which metadata standards will be used to describe the data set. (e.g. which controlled vocabulary or thesaurus is used to indicate the subject of the data set or which descriptive domain specific details are provided to represent a data set).
***If no specific metadata standard or vocabulary** (i.e. other than the metadata set provided by repositories for dissemination and preservation) is used, specify which kind of documentation will be provided: e.g. codebooks, readme files etc. To guarantee reusability, the data should be archived along with instructions and documentation (i.e. study-level and variable descriptions). Indicate how this information will be recorded and shared (e.g. a 'readme' text file will be provided along with the files containing the data).]*

3. Data Sharing and dissemination

3.A.Data accessibility

3.A.1.Explain the procedures and the measures adopted to share the data and how the data set will be disseminated:

<Insert here>

[GUIDANCE: Actions for data sharing include anonymizing or aggregating the data; gaining participant consent for data sharing; gaining copyright permissions; converting the file to standard open formats. Also explain how the shareable data are made available (e. g repository for dissemination, project web site, reports, publications…).]

3.A.2.Indicate when the data set will be made openly available and explain why:

<Insert here>

[GUIDANCE: The level of accessibility may be related to all the data set or its separated parts, or a specific version, if appropriate. Access must be given immediately at the time of publication for the data needed to validate the results. For all the other data, it will possible to specify an embargo period.]

3.A.3.Specify which versions or parts of the data set, if any, cannot be openly shared and give the reasons (you can choose one of the options in the guidance):

<Insert here>

If this is your case, please see also to point 4.D

[GUIDANCE: For example, in presence of interviews raw data may be kept confidential because of personal data issues, while processed and analyzed data may be made widely open.
In general, reasons for the impossibility to share data may be:

1) *the data belongs to third party and the consortium or the partner have not the permission for sharing it;*
2) *the data are under confidentiality obligations;*
3) *the data need to be kept confidential due to rules of personal data and ethical reasons;*
4) *other (specify other reasons).]*

3.B.Licensing

If the data are able to be accessible they will be made available under the following conditions. Choose one:

- □ CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/legalcode)
- □ CC BY NC 4.0 (https://creativecommons.org/licenses/by-nc/4.0/legalcode)
- □ ODC – BY (https://opendatacommons.org/licenses/by/1.0/)
- □ Other (*specify terms and conditions*): *<Insert here>*

[GUIDANCE: If the data is able to be shared, the manner in which you license your data can determine its ability to be reused by other scholars. For example, the Creative Commons CC BY is intended to allow users to share, modify, and

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

use the data freely, subject only to full credit to the author/s; while CC BY NC requires full credit and limits reuse for commercial purposes. CC BY NC ND is not suggested here because it limits reuse, which is contrary to the requirements of the H2020 pilot for open data. For further information on Creative Commons licenses (CC) and suggestions for choosing the appropriate license, please, see: https://creativecommons.org/licenses/. Open Data Commons (ODC) is a license specifically drafted for Open Data projects.]

3.C.Repository for the dissemination of the data set

Indicate the repository type: *<provide Name and URL>*

□ Institutional | □ Disciplinary | □ Other (*specify*): *<Insert here>*

[GUIDANCE: Indicate the repository where the data will be stored for dissemination. ***If you have an institutional data repository or use a community public one, we kindly ask you to check its policies using the model letter provided.*** *Recommended repositories must be compliant with OpenAIRE requirements. https://www.openaire.eu/intro-data-providers. At least they must manage project metadata as required by the H2020 funding programme. If your institution does not provide a repository for data set archiving, we suggest to use Zenodo, which is a multidisciplinary repository recommended by EC: https://zenodo.org/.]*

3.D.Data reuse

If applicable, indicate the tools and software for enabling re-use of dataset:

<Insert here>

[GUIDANCE: For example, indicate the software required to access data, which has to be open source (non-proprietary software). If not possible explain why. If ad hoc software or scripts are developed to manage data, indicate them and how they are made available (license and download site). If applicable, describe tools and instruments to validate the results and indicate their availability.]

4. Data set archiving, and preservation

4.A.Long-term preservation of the data set to be made openly available

4.A.1.Indicate how long the data set will be preserved and provide reasons:

<Insert here>

[GUIDANCE: Institutional or subject-based repositories usually guarantee long term preservation to the archived resources well beyond project end date. Check the repository policies.]

4.A.2.If portions or versions of the data set are not archived for preservation specify which and why:

<Insert here>

[GUIDANCE: For example raw data may not be preserved after the end of the project in compliance with the ethical rules as defined by the project.]

4.B.Processes for long-term preservation of the data set to be made openly available

Describe which specific actions are needed to provide long-term preservation:

<Insert here>

[GUIDANCE: Actions needed for preservation include: data files conversion to open formats, data anonymization, preparation of dataset metadata and documentation such as codebooks or instructions.]

4.C.Approximate end volume

Indicate the volume of the data: *<Insert here>* □ MB |□ GB

[GUIDANCE: Indicate a provisional evaluation of the data volume at the end of the project. If not yet available, indicate N/A.]

4.D.Long term preservation of not openly accessible data (OPTIONAL)

If restricted access data are present, describe where they will be stored for long term preservation and which procedures will be adopted to access them.

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

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