Data Pitch

H2020-ICT-2016-1

Project number: 732506

D3.2 Data catalogue and documentation V1

Coordinators: Jérémy Decis and Anthoine Dusselier (Dawex)

Contributor: Stefano Modafferi (university of Southampton)

Quality reviewer: Elena Simperl (University of Southampton)

Deliverable nature:	Other
Dissemination level: (Confidentiality)	Public
Work package	WP3
Contractual delivery date:	31 August 2017
Actual delivery date:	07 September 2017
Version:	V1.0
Keywords:	Data catalogue, documentation, data repository, data-sharing platform

Table of Contents

1. ABSTRACT

The DNA of Data Pitch is to allow experimenters and providers to extract value from closed data. This document describes the how the Data Pitch catalogue and document repository is utilised at each of the fives stages of this process, from initial discussions with data providers to the creation of innovation data solutions.

2. EXECUTIVE SUMMARY

This deliverable describes the part of the Data Pitch sharing platform related to the data catalogue and the document repository.

The substitution of a former partner with Dawex at the beginning of the project has produced some delays for the delivery and integration of the Data Pitch version of the marketplace. For this reason, for the first call, a complete, yet temporary, solution has been created to describe and present to the applicants information on the datasets associated with the challenges.

The document describes both the temporary and the long-term solution.

The temporary solution is a set of static HTML pages associated to a Wordpress template. The information on the dataset descriptions has been collected offline by consortium members and included by them in the web pages.

The long term solution is an ad-hoc version of the existing Dawex Marketplace platform. This is the main technical asset brought by Dawex to the consortium and it includes lots of features and functionality that are required by Data Pitch. The ad-hoc version is going to be available in M9 (i.e. September 2017) for internal testing and will be fully operational for the second round of the call.

3. INTRODUCTION

Data Pitch provides a bridge for experimenters and data providers to identify new instruments for extracting value from private datasets. The support the consortium offers to the applicants is global, covering all the legal, technical and business model aspects.

The Data Pitch goal is fulfilled via a process divided in stages.

Stage 1 of the process includes the discussions with data providers and the signing of all the legal contracts for including their data in one of the Data Pitch challenges. Stage 1 is concluded by the publication of relevant information and possibly a sample of datasets related to an identified challenge (i.e. the creation of a data catalogue).

Stage 2 includes the publication of the challenges for the benefit of the applicants and the collection of their applications.

Stage 3 includes the evaluation of the received proposals.

Stage 4 includes the negotiation phase where the fine detail of the experiment is defined in conjunction with the successful applicant(s) and the data provider.

Stage 5 is the actual experimentation period where data provider and the successful applicant work together to solve a relevant business problem.

Within the general expression of 'data sharing platform', several instruments are identified and available to support the different stages:

- The legal toolkit. The purpose of the legal toolkit is to make the data providers aware of the implication of the legal aspects of the data sharing (Cf. Del. D3.1). This instrument is relevant in Stage 1 and Stage 5.
- The application manager. This is an off the shelf service called F6S that is well-known in the startup world and provides a means for managing programme application and review. This instrument is relevant in Stage 2[1] and Stage 3[2] (https://www.f6s.com/datapitchaccelerator/apply).
- The data catalogue and document repository. This instrument is an ad-hoc version of the Dawex marketplace and it is relevant in Stage 1 to publish the dataset description, Stage 2 to provide access for the applicants to the dataset description, Stage 4 when contracts will be again one of the main topics and possibly Stage 5 when the access to the dataset is made via API.
  The substitution of a former partner with Dawex at the beginning of the project has produced some delays for the delivery and integration of the Data Pitch version of the marketplace. For this reason a complete, yet temporary solution, has been created to describe and present information on the datasets associated with the challenges to potential applicants.
- The experimental platform (cf D2.1). This instrument is used when either the data providers or the successful applicant requires an infrastructural support. It is relevant in Stage 5.

This deliverable focuses on the presentation of the data catalogue and document repository instrument.

[1] See page 9
[2] See page 11

Within the document, the acceleration stage is also called the experimentation stage and the successful applicants are also referred to as challenge winners or experimenters.

4. The Data Catalogue and Document Repository

One of the Data Pitch goals is to provide a secure and transparent environment for the data providers to describe and share their data, but also for the applicants, to help them consult these data or, at least, their description, in order to write their application.

Access to data is a sensitive topic for any data provider aware of the value hidden within it. The data catalogue must provide the right level of security and privacy while also providing information to the other stakeholders of the project (e.g. the applicants). For this reason the catalogue supports different level of access, from completely public to limited to named person only.

4.1 Current Catalogue Implementation

4.1.1 Metadata description and collection process

Each data provider has received a template document to describe their own data. Some of the providers grant API access and have a separate, linked, documentation. The completed templates have been included in the web page catalogue.

The following template has been used to collect information on the metadata:

The information provided in the following table should help someone submitting a proposal to your challenge to understand what data they will get access to if accepted into the programme. The information is meant to be a starting point for the applicants. The challenge winners will be provided with more detailed information about the datasets once they have been selected to join Data Pitch.

Table 1: Metadata description template

Dataset title *E.g., customer transactions from July 2012 till December 2012*
Dataset description *~100 words; please include the business sector, as well as key attributes and mention if this is a static or stream-like dataset.*
Dataset keywords *Up to 5 keywords to help us structure and search through the Data Pitch catalog.*
Data provider name:
Data provider country:
Dataset update frequency: *You might want to provide updated, fresh data to the SMEs during their 6 months acceleration. If this is the case, please give an indication of the expected frequency (e.g., once an hour, once a week, not at all etc.)*
Dataset size: *Rough estimate of the space (e.g., how many Terabytes) and number of records (e.g., 20k products, 30k transactions; 20 stores)*
If applicable, please give an indication of how many attributes are in your dataset[3]:
Data format and storage: *E.g. csv files, SQL database, text files in a file system etc.*
Data attributes: *Please provide the most relevant information present in the dataset to address the challenge. For example, if the challenge is about the optimization of a distribution plan the following could be a good list:* ● *Stock level in company warehouses (SQL table containing integers values calculated daily);* ● *Alerts related to low level of specific goods (text alerts generated by an existing system);*

[3]e.g. in a SQL database the number of columns. In some context (e.g. when a RDF representation is used) this information may be highly variable and therefore not really significant. In this latter case, please state n/a.

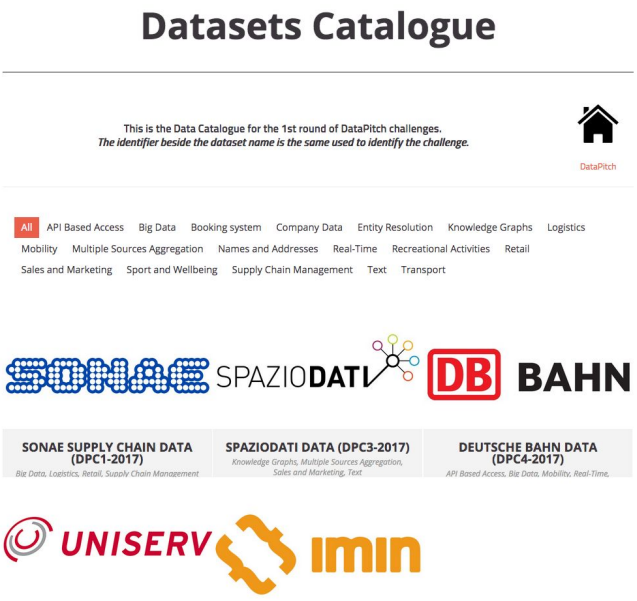
• *Distribution plan for goods delivery (logistics data);* • *Current and future promotional plans across branches (#itemID, discount percentage, timeline)* • *Goods database (core structure)*
The dataset contains (tick all that apply): o Anonymized or pseudo anonymized data derived from personal data o Personal data o No data relating to persons
Is any synthetic[4] data present in the dataset that is going to be used during the experiment? *List the name of the attributes for which the use of a synthetic version is or will be made available*
Geographic coverage: *If applicable, please provide the dataset geographic coverage (e.g., information about all schools in the UK). Note that this question does not refer to where the data are stored.*
Dataset timespan and production period o Timespan: o Start: o End/No end: o Production period: o Start: o End/No end *The timespan refers to the time covered by the data (e.g., student records from 1950 to 1975). The production period refers to when the dataset is created (e.g., from 2000 to 2005, for example as part of a digitalization project). This information might not be relevant for all datasets and may be omitted.*
Level of aggregation: *Does your dataset contain raw data or is it subject to any preliminary aggregation? If possible, please express in a simple way the level of granularity expected (e.g. raw data, low/medium/high level of aggregation, and/or an indication of the compression ratio final data/raw data)*
Data access (tick as applicable): o Via API o Bulk download o Subject to negotiation

4.1.2 Implementation

The catalogue has been implemented via a set of web pages on a wordpress template. It is available here: http://wordpress.it-innovation.soton.ac.uk/datapitch-datasets/.

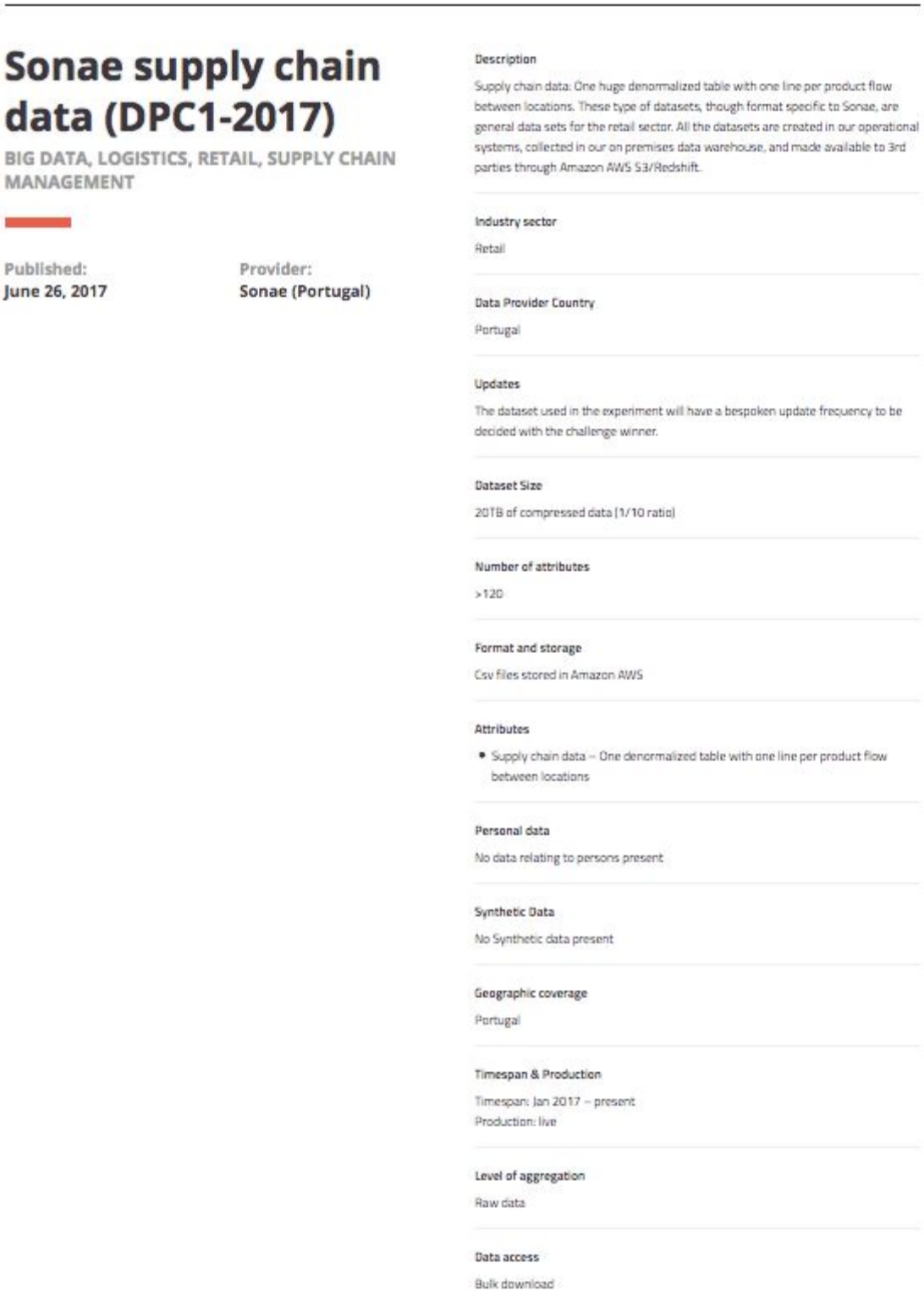
[4] Synthetic means dummy. An anonymised or pseudo-anonymised data is not considered synthetic in this context.

Figure 1: screenshot of the current version of data catalogue



The pages are linked by the challenge description page and are hosted in an IT Innovation department server at the University of Southampton.

Figure 2: screenshot of current version of the data catalogue (provider page)



Sonae supply chain data (DPC1-2017)

BIG DATA, LOGISTICS, RETAIL, SUPPLY CHAIN MANAGEMENT

Published: June 26, 2017

Provider: Sonae (Portugal)

Description

Supply chain data: One huge denormalized table with one line per product flow between locations. These type of datasets, though format specific to Sonae, are general data sets for the retail sector. All the datasets are created in our operational systems, collected in our on premises data warehouse, and made available to 3rd parties through Amazon AWS S3/Redshift.

Industry sector

Retail

Data Provider Country

Portugal

Updates

The dataset used in the experiment will have a bespoken update frequency to be decided with the challenge winner.

Dataset Size

20TB of compressed data (1/10 ratio)

Number of attributes

>120

Format and storage

Csv files stored in Amazon AWS

Attributes

- Supply chain data – One denormalized table with one line per product flow between locations

Personal data

No data relating to persons present

Synthetic Data

No Synthetic data present

Geographic coverage

Portugal

Timespan & Production

Timespan: Jan 2017 – present
Production: live

Level of aggregation

Raw data

Data access

Bulk download

4.2 Data catalogue and document repository for the second call

As explained in the introduction the data catalogue and document repository will be used in Stage 1 to publish the dataset description, in Stage 2 to provide access for the applicants to the dataset description, in Stage 4 to onboard and vet the experimenters on the platform, and possibly in Stage 5 where the Dawex platform will support the access to data for the experimentation when the data are streamed or accessed via API. In the following paragraphs the catalogue functionalities and the Data Pitch requirements will be presented according to the different stages.

4.2.1 Stage 1: Contact with providers and contract signing

During Stage 1 of the project, the consortium will present the functionalities of the platform to the data providers and onboard them once the contracts are signed.

The early outcome of this stage is the definition of the challenge and the signing of the contract.

4.2.2 Stage 2: Challenge definition

Stage 2 includes the publication of the information related to the datasets. Ideally this step should be done by the data provider, who registers and joins the catalogue autonomously. To speed up the process an “acting on behalf” function will allow selected members of the consortium to input the data themselves if they are received in a different way. In the case where the data providers give us a link to an existing documentation, a member of the consortium will integrate it manually in the catalogue.

Once registered, each provider has the choice to create an “offering” or a “theme”, based on their data sets and depending on the level of sensitivity of their data.

Metadata both from offerings and thematics will thereafter be used in the public catalogue to describe the data provided.

4.2.2.1 Offering:

An “offering” is for a data provider who already has the data and is willing to upload it on the platform. By creating an offering, a data provider will enable the Data Pitch platform to create a sample. This sample will be generated with random algorithms. The sample will be available on the public catalogue during the Stage 3.

This feature will be used when the data are not sensitive. The offerings allow to share data by APIs

4.2.2.2 Metadata collected for an offering for data files (csv, json, pdf, txt, xml, xls, geojson):

- Title of the data set
- Detailed description of the data set
    - free text description of the data set
    - free text description of each header present in the data
- Territories covered by the data
- Business sectors covered by the data
- Type(s) of data available
- The data produced contains...

    - Anonymized data derived from personal data
    - Personal data
    - No personal data
- Keywords that describe the data set
- History of your data offering
- Level of data structure for this theme
    - Not structured (never applied the resources to structure it)
    - Not structured (by nature)
    - Partially structured
    - Highly structured
    - I don't how to evaluate it

4.2.2.3 Metadata collected from Data provided via API's[5]:

One of the most challenging and also interesting case is the execution of analytic functions on data streams. The platform has the ability to manage the streaming live of data. The platform works as a proxy : it will play the role of a bridge between the data provider, and the data user, who wants to acquire the data. In the case of Data Pitch, the source will be the data provider and the destination will be the experimenter's infrastructure.

Using this functionality of the Dawex marketplace, we will ensure the consistency of the data and allows for API from IoT, Business Intelligence software, DPM and app.

Name of the data set

- Summary (250 characters)
- Data (The types of data making up your offering. Several choices are possible.)
    - Financial and administrative data
    - Sales and marketing data
    - Production and technical data
    - Etc.
- Indicate whether or not personal data is present
    - Anonymized data derived from personal data
    - Personal data
    - No personal data

Business sector (The business sectors covered by the data of your offering.)

- Geographical areas covered
- Production period (A date or period during which data is produced.)
- Description of the API
    - New URL (Fill in all the information needed to describe the URL for your API.)
        - Name
        - URL
        - Request (GET/POST)

[5] https://www.dawex.com/en/product/#api

- Output data format
- Error management protocol
- Does this URL contain parameters? (Yes/No)
- Details of your offering
    - Description – Keywords

4.2.2.4 Theme:

If the data provider does not have the data ready, or if they do not want to upload it on the platform, they will create a “theme”. A “theme” is used only to describe the data that the provider will provide later to the experimenters. It is not possible to create a sample in this case. The metadata collected during the creation process will be used in the public catalogue during the Stage 3.

This feature will be used when the data are sensitive.

4.2.2.5 Metadata collected for a theme

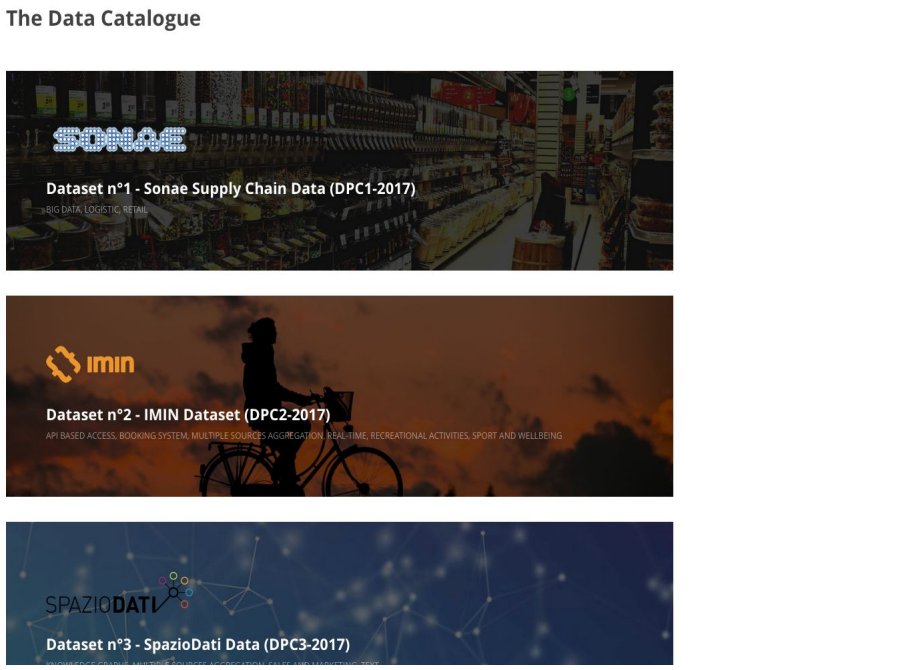
- Title of the theme
- Detailed description of the theme and of the data (This includes the free text description of each header present in the data)
- Territories covered by the data
- Business sectors covered by this theme’s data
- Type(s) of data available for this theme
- The data produced for this theme contains
    - Anonymized data derived from personal data
    - Personal data
    - No personal data
- Keywords that describe this theme
- History of the theme’s data
- Level of data structure for this theme
    - Not structured (never applied the resources to structure it)
    - Not structured (by nature)
    - Partially structured
    - Highly structured
    - I don’t how to evaluate it

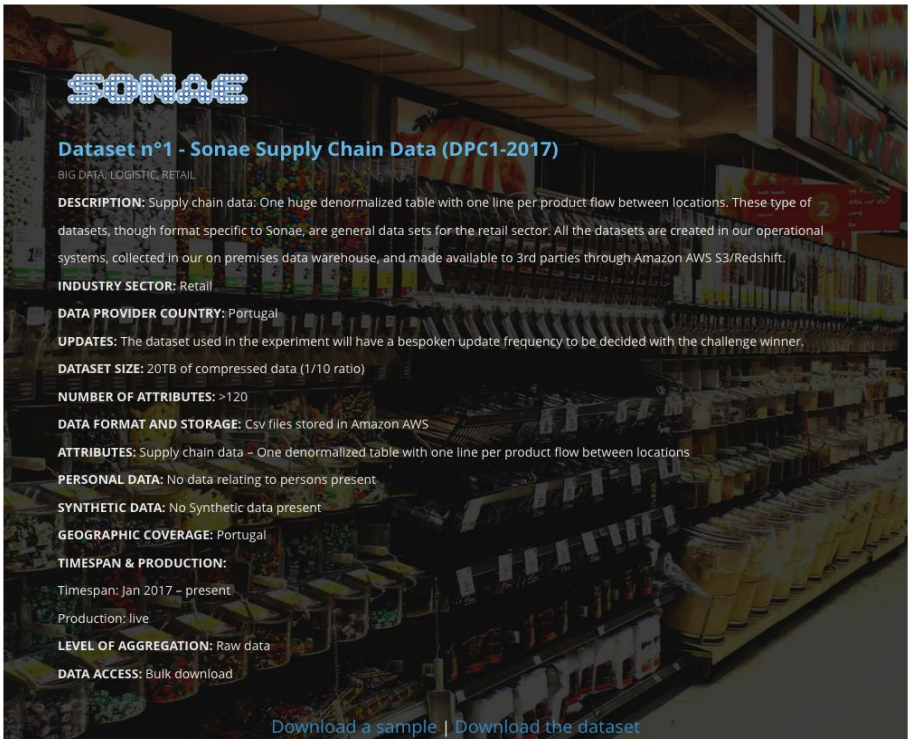
4.2.3 Stage 3: call opens and applications

After the call opens, a public catalogue has to be available to the applicants, to our associates and partners. This catalogue do not include the data itself, but either a sample generated from the data and a description of the data set, or only a description, allowing the applicants to build innovative solutions and to submit their proposals.

For these purposes, a landing page is created, gathering the samples generated from the offerings created on the platform, and the metadata from the themes. A version 1 of this landing page already exists : https://www.dawex.com/en/lp/datapitch-datasets/

Figure 3: screenshots of the data catalogue





This landing page will be the public catalogue available to all during the call.

4.2.4 Stage 4: Negotiations[6]

After the end of the call, the applicant proposals will be evaluated by Data Pitch and the winners will be selected for the acceleration stage.

There will be 3 steps during this stage:

1. On boarding the experimenters on its platform.
2. Vetting the experimenters and due diligence on the status of the company[7] to enhance the security level of the platform and to reassure the data providers. This is a critical point in order for data providers agree to join Data Pitch.
3. When the two processes are finalized, the experimenters have access to the Datapitch platform. Which part of the Data Pitch sharing platform is going to be involved in the experimental phase is going to be part of the negotiation and will result from the analysis of the experiment requirements.

As in Stage 1 contracts and legal aspects in general will play a fundamental role in this stage. The repository part of the Data Pitch sharing platform will offer as many instruments as possible to facilitate and smooth the management of all the legal aspects.

4.2.5 Stage 5: accelerator

In this stage the data catalogue and document repository play a secondary role. Nevertheless other functionalities from the platform are going to be used in Data Pitch as already explained in Deliverable 2.1.

These functions will be activated when possible and convenient. This is most likely when a data stream is going to be involved or the data providers offer an API for accessing her datasets.

When data are going to be accessed via these functionalities, the dataset will remain completely under the control of the data provider that will have their own credentials to perform the task.

4.2.5.1 API technical details

The platform acts as an intermediate server (“proxy”) between the experimenter’s server and the provider’s server (which is providing the API)

- The experimenter makes a request to the server with its identifiers and a reference to the desired API
- The server checks if the subscription is valid, as well as the API reference and query volumes restrictions (daily and per second)
- If the request is valid, the server queries the provider’s server to retrieve the data

[6] Stage 3 is about the evaluation of the proposal and the data catalogue and contract repository does not play a relevant role.

[7] https://drive.google.com/file/d/0B9IIZV_CjqLcOTVIYkh6V0xGNE0/view?pli=1

- The data is then returned to the experimenter via the server.

The server also performs a sampling based on vendor data at regular intervals to ensure that it is always available.

4.2.5.2 Activity reports

At any stage of the process, the data catalogue and the functionality for data access will regularly produce reports from the activities on the platform of both the data providers and of the experimenters.

These reports will include information about:

- searches on the platform made thanks to the search engine
- the pages and the data that the experimenters consulted
- contact requests
- downloads

These reports will allow to monitor the interactions between the providers and the experimenters, in order to evaluate their behaviours, to better understand their needs, but also to ensure that the rules of the project are respected.

5. CONCLUSION

This table summarizes all the functionalities provided by the Data Pitch sharing platform depending the stage of the project.

		Stages				
		Stage1: challenge definition	**Stage2: applications**	**Stage3: Selection process**	**Stage4: negotiations**	**Stage5: acceleration of the experimenters**
Functionalities	**Offering creation**	Collect metadata and generate sample	Provide metadata and sample to the applicants	N/A	N/A	Data access
	Theme creation	Collect metadata	Provide metadata to the applicants	N/A	N/A	N/A
	Data from API	Collect metadata	Provide metadata and description to the applicants	N/A	N/A	Data access
	Sample Production	N/A	Assess data quality and structure	N/A	N/A	N/A
	Landing page	N/A	Gather the metadata and the samples on a single place for the applicants	N/A	N/A	N/A
	Vetting process	N/A	N/A	N/A	Reassure the data providers and enhance the security of the platform.	N/A
	Personalized data management and access	Provide a dedicated environment to the Data providers on the platform.	N/A	N/A	N/A	Provide a dedicated environment both to the data providers and to the experimenters on the platform.
	Activity reports	N/A	Monitor the activity of the applicants on the catalogue	N/A	N/A	Monitor the activity of the data providers and the experimenters.

Table 2: Catalogue function by stage

This document explains the choice made by Data Pitch to provide data to the applicants and, later, to the experiments, by using a data sharing platform allowing to create the data catalogue and the document repository, as well to share data between the providers and the experimenters in some cases (API transfer). Due to the late arrival of Dawex, who provides its platform to the Consortium, two versions of the catalogue have had to be developed : one for the call opened in July 2017, and another one for the next call, scheduled in 2018. The data sharing platform will be available in November, 2017, after internal tests.