Grant Agreement Number: 825225

Safe-DEED

www.safe-deed.eu

D4.4 Big Data Valuation Component v.1

Deliverable number	*D4.4*
Dissemination level	*Consortium*
Delivery date	*30 November 2020*
Status	*Final*
Author(s)	*Mihnea Tufiș*

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 825225.

Changes Summary

Date	Author	Summary	Version
10.11.2020	Mihnea Tufiș	First draft	0.1
11.11.2020	Ioannis Markopoulos	Comments from FNET	0.2
11.11.2020	Mihnea Tufiș	Integrate FNET review	1.0
14.11.2020	Evangelos Kotsifakos	Comments from LST	1.0
16.11.2020	Mihnea Tufiș	Integrate LST review	1.1
26.11.2020	Mihnea Tufiș	Minor corrections	1.2

Executive summary

This document supports deliverable D4.4 – Release of the Big Data Valuation Component version 1. This is the second release of the Data Valuation Component (DVC) following that in deliverable D4.2 [15]. It is supported by an updated architecture of the one presented in D4.1 [13] and includes advanced implementations of context processing and data quality assessment, following the results from the report on the context aware and context unaware valuation in deliverable D4.3 [16]. The context processing is also considering the role of privacy to data value, following collaboration with our colleagues from KUL (WP3). The functionality of the demonstrator has been tested as part of FNET data trials documented in deliverable D6.2 [12].

Table of Contents

List of Figures

Notation

Abbreviations

ADAS	Automatic Data Analysis and Scoring Component
CPL	Communication and Presentation Layer
DFD	Data Flow Diagram
DIL	Data Ingestion Layer
DQA	Data Quality Assessment
DVC	Data Valuation Component
EUT	Fundacio Eurecat
FNET	Forthnet S.A.
GUI	Graphical User Interface
IFX	Infineon Technologies AG
KUL	KU Leuven
QDSC	Qualitative Information Extraction and Data Scoring
S2VM	Score-to-Value Mapping
WP	Work Package

1 Introduction

This document accompanies the software release of version 1 of the big data valuation component, in deliverable D4.4. The code is available in the same archive as this document, is stored on the Safe-DEED git repository, the Safe-DEED Next Cloud and is available on request.

The rest of the document is structured as follows: Section 2 describes the implementation details of the Data Valuation Component (DVC), including a description of the sub-components, the class diagram of the solution, the data flow between sub-components. Section 3 describes the structure, the dependencies and how to run the demonstrator package. Section 4 concludes the document and discusses the next steps in the development of the DVC.

This deliverable represents the development step following D4.2 – Baseline prototype for data valuation released in November 2019 [15]. The new features (see Section 2.1) are based on the results of the extensive state of the art review of methods for data valuation – deliverable D4.3 [16], and the contributions about legal, ethical and privacy perspectives from our colleagues at KUL (WP3) (see Section 2.1.2.2). The deliverable was included in the FNET data trials documented in D6.2 [12].

2 Implementation of the Data Valuation Component

The goal of the DVC is to perform the valuation of a data set over three aspects: data quality, data exploitability and economic value. The current version of the DVC receives a structured data set, together with a context and a set of rules for evaluating data quality and returns three scores which describe the value of the data:

1. a score based on the contextual information provided;
2. a score based on the quality rules provided;
3. the aggregate value of the data set currently computed as a mean of the 2 previous scores.

2.1 Updates from previous version

This is a list of the improvements to the Data Valuation Component (DVC), compared to deliverable D4.2 [15]. We remind our reader that the respective version represented only a baseline implementation.

1. Improvement of the architecture for the sub-components and the data flow; subsequent code refactoring.
2. Implementation of an improved GUI, based on HTML and Bootstrap[1], to replace the less user-friendly console interactions.
3. Data Ingestion Layer. Development of a friendly method for ingesting the data set.
4. Context Processing. Development of a 5-steps process to establish the context under which data valuation is performed. This replaces the previously tedious process of preparing a custom JSON file with the context representation.
5. Data Quality Requirements. Development of a multi-step process for creating data quality rules that will be applied during the data quality assessment (DQA). Supported by an intuitive GUI.

[1] https://getbootstrap.com/

6. Scoring components. Development of methods for computing the contextual (QDSC) and data quality (ADAS) scores. These are then aggregated in a final data value (S2VM).
7. Communication and Presentation Layer. Development of a user-friendly method for presenting the three valuation scores (context, quality, aggregate), enabled by HTML and Bootstrap[1]. The values of the scores are highlighted by a colour scheme (red, yellow, green) to promote the interpretability of the results.

2.2 Architecture

The DVC comprises of the following modules, as illustrated in the data flow diagram in Figure 9.

1. Data Ingestion Layer (DIL);
2. Qualitative information extraction and Data Scoring Sub-Component (QDSC);
3. Automatic Data Analysis and Scoring (ADAS);
4. Score-to-Value Mapping (S2VM);
5. Communication and Presentation Layer (CPL).

2.1.1 Data Ingestion Layer (DIL)

DIL represents the entry-point of the data to the platform. The data is currently ingested via a GUI, where the path to the data set is specified. The layer detects the data format and performs the suitable operations for ingesting it. Currently the only supported formats are CSV and XLS(X), with the capacity of choosing a specific datasheet for the latter.

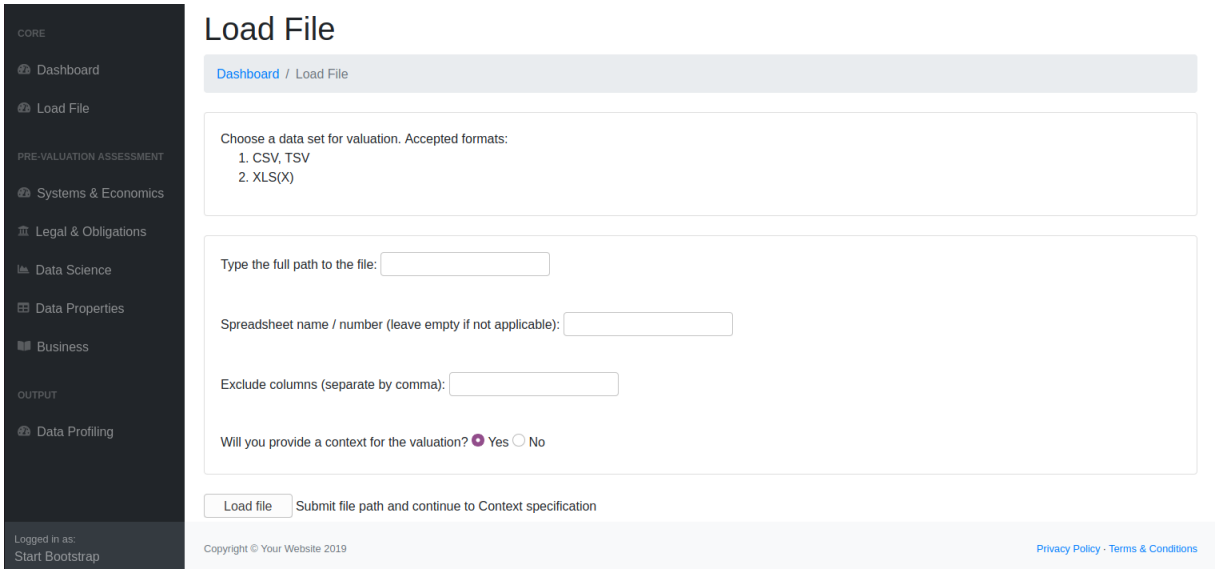

Figure 1: DVC - Load the data file

2.1.2 Qualitative Information Extraction and Data Scoring Component (QDSC)

The first stage of the data valuation process directly involves the user, who is required to provide information about the context in which to valuate an input dataset. This is achieved by a multi-step questionnaire focusing on the following aspects:

1. Systems & Economics: Availability & access, Purpose
2. Legal & Obligations: Data protection, Legal-Terms-Obligations

3. Data Science: Tools, Format
4. Data Properties: Data velocity, Data transformations, Data quality, Data age
5. Business: Frequency of use, Benefits

2.1.2.1 Systems & Economics

This is the first section of the QDSC context establishing process and collects information related to two main aspects surrounding the system and economics of the input data set. It collects information about the availability and access of the data set, as well as a declaration of the purpose of its usage (see Figure 2).

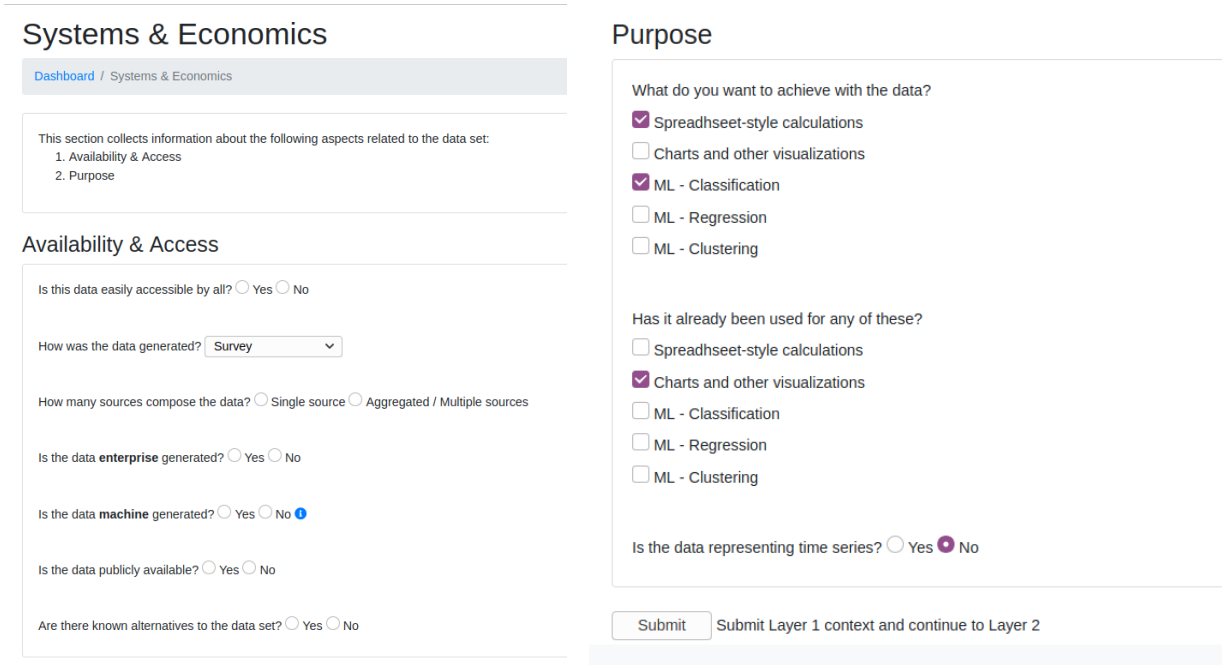

Figure 2: Context establishing process – Step 1 – Systems & Economics

2.1.2.2 Legal & Obligations

The content of this section of the QDSC context establishing process is the result of the cooperation with our colleagues from WP3. Its objective is to collect the main information with respect to the data processing and legal practices concerning the input data set (see Figure 3).

2.1.2.3 Data Science

This section aims to collect information about the format of the original data and tools necessary to perform the data science tasks required to achieve the purpose declared in the previous step (see Figure 4).

2.1.2.4 Data Properties

This section collects information about relevant properties of the data set, such as velocity, transformations, quality, age (see Figure 5).

2.1.2.5 Business

Finally, the goal of this QDSC section is to establish the business context in which the data set will be used, by answering questions about the frequency of use of the data set and the expected benefits resulting from its use.

Data protection

Is the purpose of the processing activity clearly defined? (•) Yes ( ) No

Does the dataset contain personal data? (•) Yes ( ) No

Does the dataset contain any of the following?

Business confidential information:

- [ ] Revenue or financial data
- [ ] Data covered by copyright
- [ ] Trade secrets

Personal data:

- [ ] Sensitive data (e.g., medical)
- [x] Digital traces (e.g., cookies, IP address, MAC address, search or browsing history etc.)
- [x] Any other personal data

Was the processing of personal data carried out respecting privacy and data protection principles?

(•) Yes ( ) No ( ) Not applicable (no personal data)

Is the controller able to demonstrate compliance with the EU Data Protection Principles

(•) Yes ( ) No ( ) Not applicable (no personal data)

Have data subjects been given all the necessary information about processing their data in order to exercise their rights?

(•) Yes ( ) No ( ) Not applicable

Have appropriate technical and organisational measures been implemented to ensure a level of security appropriate to the risk?

(•) Yes ( ) No

Is the data encrypted? ( ) Yes (•) No

Is the data anonymized? (•) Yes ( ) No

Figure 3: Context establishing process – Step 2 – Legal & Obligations

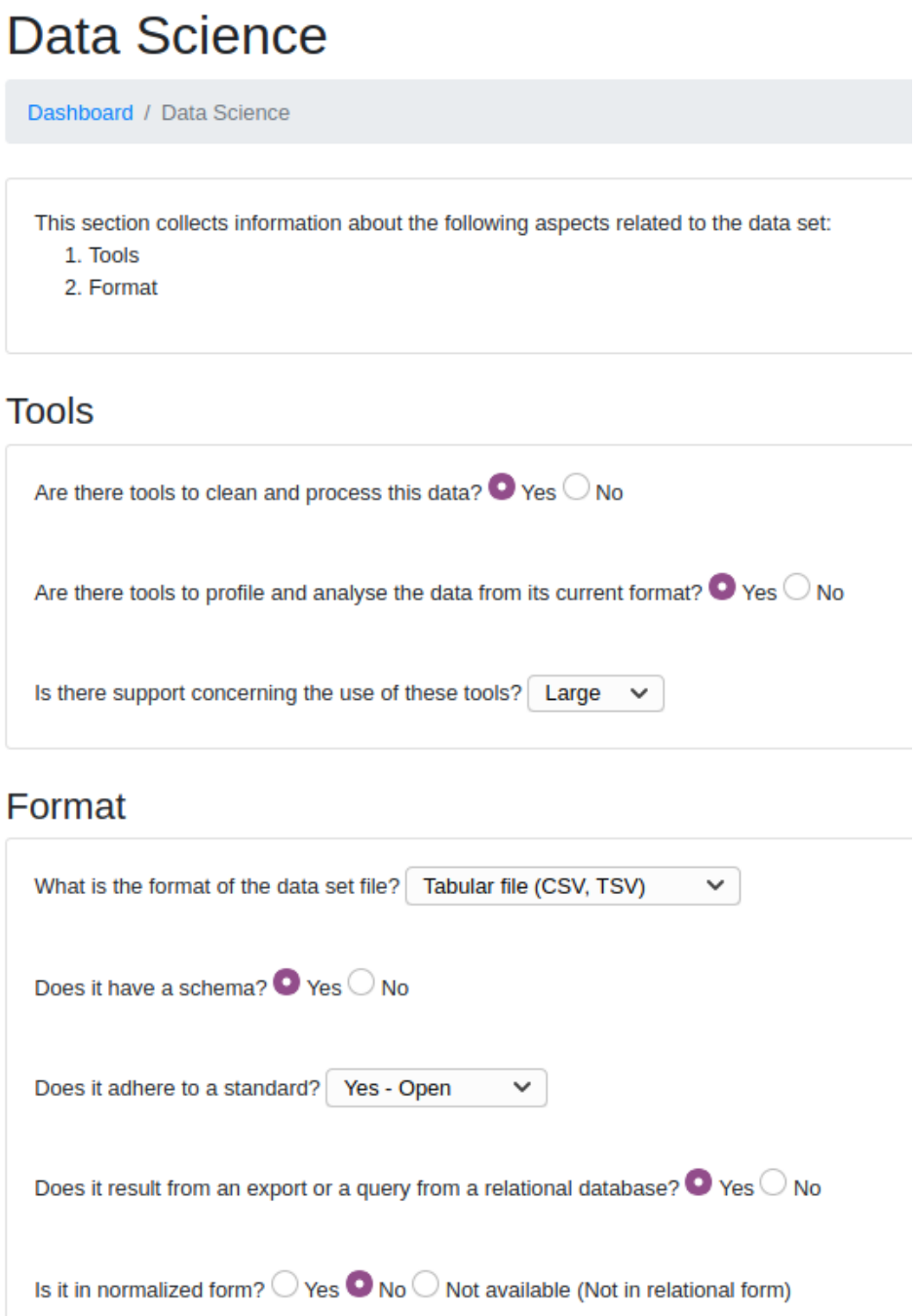

Figure 4: Context establishing process – Step 3 – Data Science

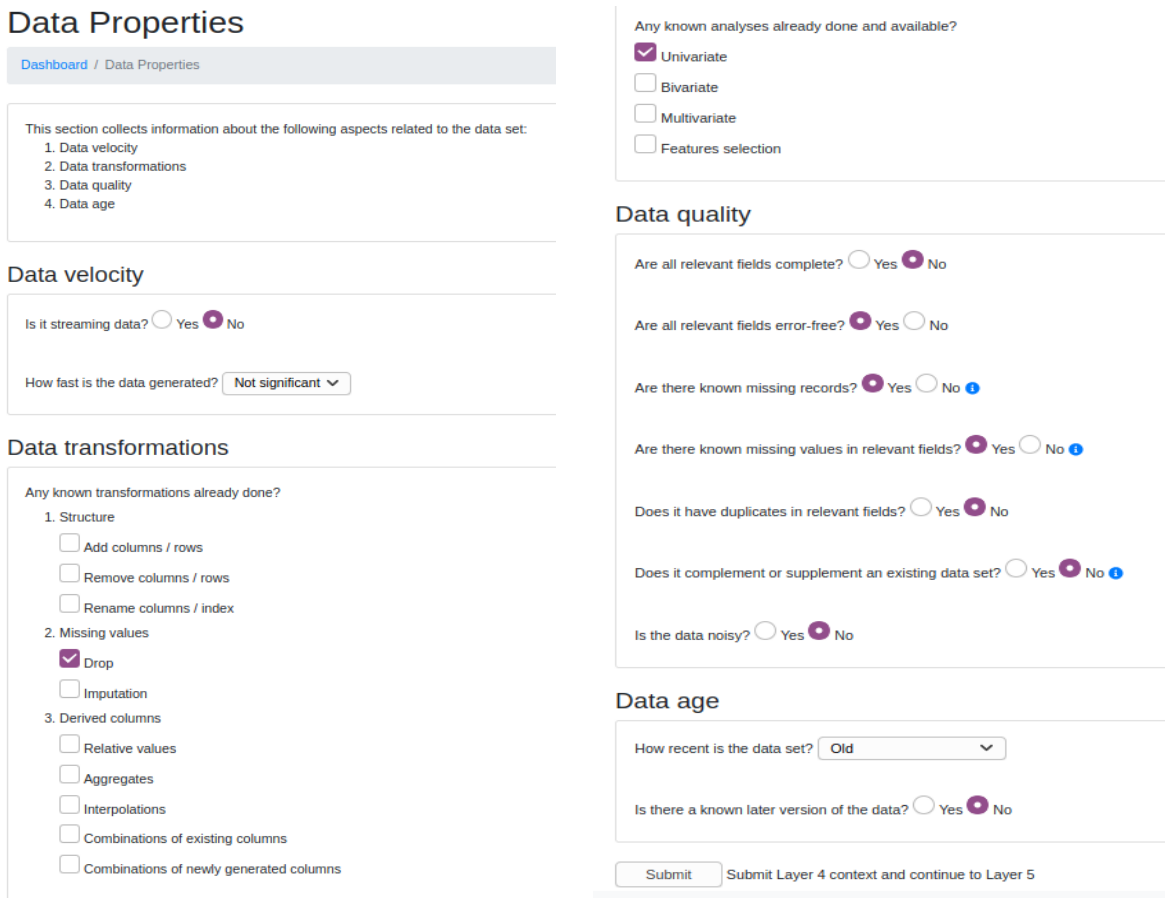

Figure 5: Context establishing process – Step 4 – Data Properties

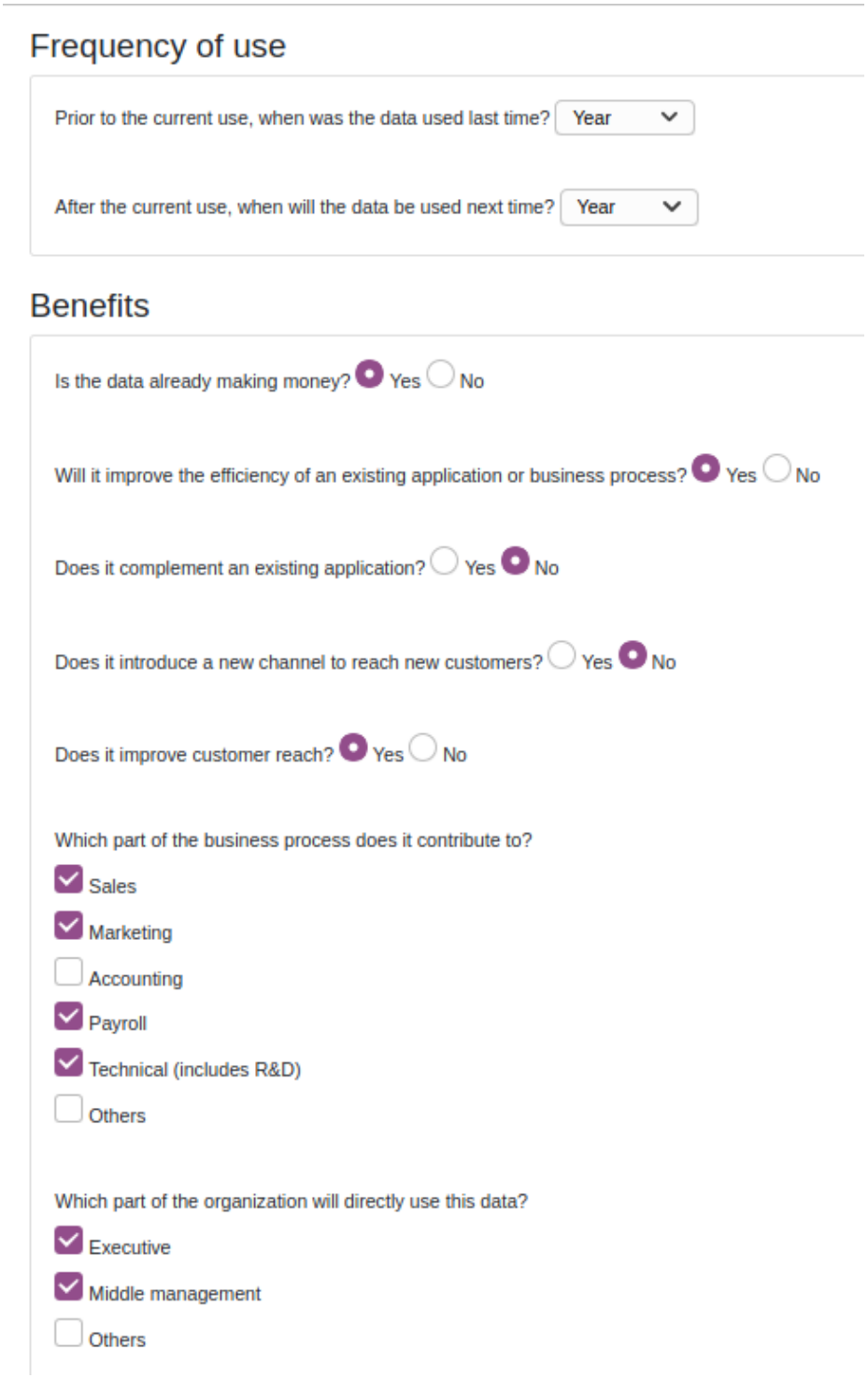

Figure 6: Context establishing process – Step 5 – Business

In the interest of reproducibility of the valuation process, the context is encoded as a JSON file and available for download and reuse. Finally, the component computes and returns a context-based score (QDSC-score).

The method for computing this score is inspired from current research on mapping data properties to data value [8] or data sheets for data sets [11], both of which are discussed in detail in deliverable D4.3 [16]. Essentially, there is a mapping between each potential answer to each question in the five steps before and a score between 0 and 1. At the end of the form, all the answers are summed up and the % score with respect to the maximum possible is computed.

```
"layer_2": {
    "protection": {
        "has_purpose": {
            "yes":1,
            "no":0
        },
        "is_personal": {
            "yes":0,
            "no":1
        },
        "business": {
            "revenue":0.34,
            "copyright":0.33,
            "trade":0.33
        },
        "personal": {
            "sensitive":0.34,
            "traces":0.33,
            "other":0.33
        },
        "is_principles_compliant": {
            "yes":1,
            "no":0,
            "NA":1
        },
```


Figure 7: Snippet of the context-value mappings for Layer 2 - Legal, Terms, Obligations

2.1.3 Automatic Data Analysis and Scoring (ADAS)

It starts by preparing the loaded data and performs a set of analytic operations to extract the intrinsic properties of the dataset:

- data shape (number of lines and columns) and size;
- inference of the most plausible data type for each column (string, integer, float, datetime etc.);
- the profile of each field:
    - missing values (% missing values per column, using default markers – i.e. NULL, NA, nan etc.);
    - distribution of the data from each field (histogram of the distribution of the possible values in each column).

Based on the information extracted about the structure of the dataset, the ADAS then initiates a multi-step process for collecting any data quality rules applicable to each of the columns of the data set. The current version of the component performs data quality assessment for the following data quality measures:

- completeness (% of missing values or equivalent)
- two types of validity:
    - domain validity (% of values in a defined range / domain)
    - format validity (% of values that respect specific formatting rules)

The current version makes use of a basic textual format for describing the rules, which we intend to enhance (perhaps develop in a more graphical manner) for the final version of the component.

Just like in the case of contexts, in the interest of reproducibility, the data quality rules are encoded as a JSON file and available for download and reuse.

Next, the ADAS uses the loaded data and performs the data quality assessment against the provided quality rules. Quality scores are subsequently computed for each data quality dimension, which are finally aggregated (as a mean) into an ADAS-score.

This employs a classic DQA framework, where each selected quality dimension is supported by one or several quality metrics [1][4]. For more details, please refer to the extended review of Data Quality Assessment methods in Section 4 of deliverable D4.3 [16].

Essentially, after the input of quality rules for each of the data quality measures (DQM) (in this version: completeness, domain validity, format validity), our algorithm looks at the agreement between the content of each column and each of the data quality rule. A score between 0 and 1 is then generated for each DQM, at a column level and then aggregated at data set level [7][9]. Finally, the ADAS-score is obtained by aggregating the three DQMs, with equal weights.

2.1.4 Score-to-Value Mapping (S2VM)

Using the two previous scores (QDSC-score and ADAS-score), this component aggregates them into a final score (as an average) – the data value [7]. While this method is still simplistic, it validates the full data flow through the component and leaves the door open for the addition of economic models for data valuation, which will be integrated for the final version of the DVC.

2.1.5 Communication and Presentation Layer (CPL)

This layer acts as an interface with the user, to report the final results of the data valuation process. It displays in an integrated GUI the following (see Figure 8):

- the QDSC and ADAS scores, together with a 3-colors code;
- the aggregated value of the dataset, together with a 3-colors code;
- a report of the dataset profile, available both as part of the GUI (HTML), as well as in PDF format.

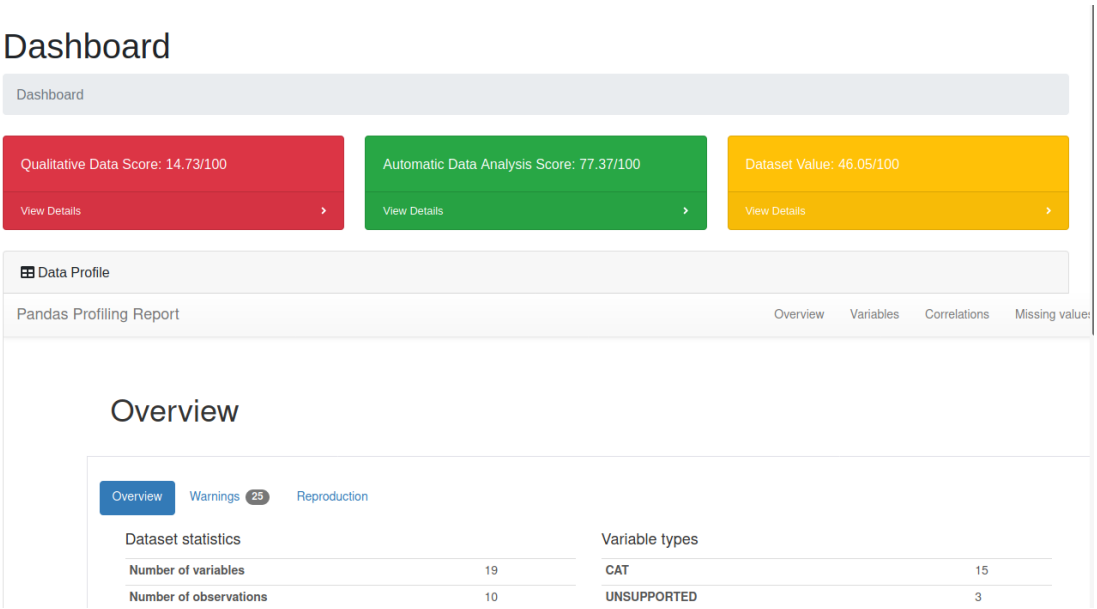

Figure 8: DVC - Output scores and colour codes

2.3 Data flow

The DVC input / output structure is given below, together with the data flow diagram (Figure 9) for the entire component.

Input:

- a dataset (or a snapshot thereof) (only CSV and XLS(X) supported at the moment). Future versions will include support for semi-structured data sets (XML, JSON).
- user provided context;
- user provided data quality rules.

Output:

- a set of scores, which evaluate the input dataset from three perspectives: contextual data valuation, data quality assessment, aggregated data value;
- a set of reports based on the analysis of the intrinsic properties of the dataset (format, shape, data types, missing values, duplicates);
- additionally, both user provided data (context and data quality rules) are persisted in JSON files for subsequent use.

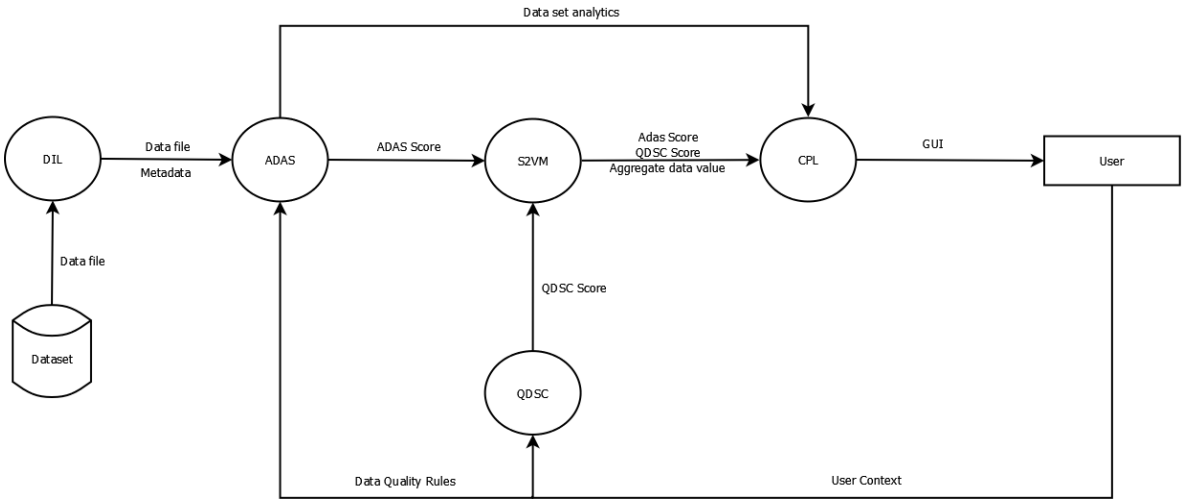

Figure 9: Data flow diagram (DFD) for the Data Valuation Component (DVC)

3 Deliverable details

3.1 Package structure

The structure of the package that forms deliverable D4.4 is the depicted in Figure 10.

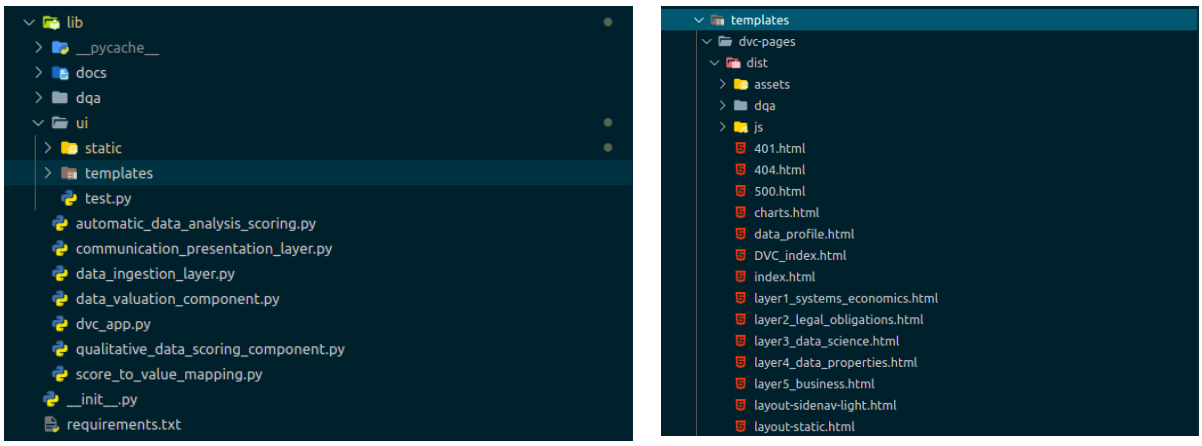

Figure 10: Structure of the repository (left) and of the Flask templates for the GUI (right)

3.2 Library requirements

The current version of the prototype was developed using Python 3.7, runs as a Flask web app and in order to run, it requires the libraries enumerated in the file ***requirements.txt***, included in the deliverable. At this stage, there are no particular requirements with respect to the operating system. Below we give a brief summary of the most important libraries necessary to run the DVC.

flask=1.1.2	matplotlib=3.3.1	pandas=1.1.1
json-c=0.13.1	numpy=1.19.1	pandas-profiling=2.9.0

3.3 Running the DVC

This section describes how to run the current DVC prototype. This is more straightforward and requires less technical expertise than the previous version in D4.2 [15].

0. Make sure that all prerequisite libraries in ***requirements.txt*** are properly installed.
1. Unzip the file in the same archive or download the DVC package from the Safe-DEED Git repository.
2. Open a terminal window
3. Navigate to the *DVC* project folder and run the following command:
   ```
   > python3 lib/dvc_app.py
   ```

4 Conclusions and next steps

Deliverable D4.4 represents version 1 of the Data Valuation Component, a major improvement from the baseline version in D4.2 [15]. It incorporates implementation based on our research on context formalisation (see Section 2.1.2), methods for data valuation [16] and includes a module for assessing legal, ethical and privacy issues (see Section 2.1.2.2).

Finally, we leave our reader with a detailed list of the work planned or currently in progress, in view of the next release of the DVC (version 2), due at month 36:

1. improve the user interface and user experience of the platform.
2. extend the capabilities of the current dimensions. This is the case of consistency where we need to develop more data quality rules (string regex and datetime format validation are currently supported).
3. add more data quality dimensions.

   a. **Time-related measures** are on top of our priorities (age, currency), together with the underlying decay models [16]. As research has shown [1], besides the DQMs that we have already implemented, these are some of the most relevant from a quality perspective.
   b. **Availability and accessibility** are currently declared by the user as part of the context gathering process and quantified within. We want to study the possibility of automatising this (probably once the platform is deployed in a data market) and considering their scores as independent DQMs.
   c. **Security and privacy**. We are looking forward to the integration with the work in WP5 which could open the possibility of estimating these DQMs and include them in the final score.
   d. **Performance**. This measure will refer to the usability of the data set in intended contexts (e.g., advance analytics, training machine learning models).
4. aggregate measures. Improve the currently basic aggregate measures, by using economic models for data value, which make use of the declared context.
5. validation is something that still raises difficulties, due to a lack of experiments [1] and ground truth valuations. This should be overcome once the component will be deployed and used in a data market. Until then, our strategy will focus on three alternatives:
   i. evaluate personal data sets, for which the review in Deliverable D4.3[16] – Section 2.2.2 provides us with a satisfactory "ground truth";
   ii. evaluate open data sets [1][3][5][6][14], which were previously used in case studies for other valuation methods [8][10];
   iii. work with our colleagues from FNET and IFX (WP6, WP7) and use their in-house data sets and perform a qualitative evaluation of the results.
6. integrate the work on privacy preserving techniques currently developed by our colleagues in WP5.

5 References

[1] Batini, C., Cappiello, C., Francalanci, C., and Maurino, A. (2009). Methodologies for Data Quality Assessment and Improvement. ACM Computing Surveys, 41(3).

[2] Bolukbasi, T., Chang, K.-W., Zou, J.Y., Saligrama, V., Kalai, A.T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. Advances in Neural Information Processing Systems. pp. 4349–4357.

[3] Buolamwini, J., Gebru, T. (2018). Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In: Proceedings of the 1st Conference on Fairness, Accountability and Transparency. Friedler. S.A., Wilson, C. (eds). New York, NY, USA: PMLR; 2018. pp. 77–91.

[4] Cai, L., and Zhu, Y. (2015). The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. Data Science Journal, 14(2), 1–10.

[5] Choi, E., He, H., Iyyer, M., Yatskar, M., Yih, W., Choi, Y., Liang P., and Zettlemoyer, L. (2018). QuAC: Question Answering in Context. CoRR.

[6] Erkut Erdem. 2018. Datasheet for RecipeQA.

[7] Even, A., and Shankaranarayanan, G. (2006, November 10). Value-Driven Data Quality Assessment. Proceedings of the 2005 International Conference on Information Quality. MIT IQ Conference, MIT, Cambridge, MA, USA

[8] Gebru, T., Morgenstern, J., Vecchione, B., Wortman Vaughan, J., Wallach, H., Daumé III, H., and Crawford, K. (2020). Datasheets for Datasets. http://arxiv.org/abs/1803.09010

[9] Görz, Q., and Kaiser, M. (2012). An Indicator Function for Insufficient Data Quality – A Contribution to Data Accuracy. In H. Rahman, A. Mesquita, I. Ramos, and B. Pernici (Eds.),

Knowledge and Technologies in Innovative Information Systems (Vol. 129, pp. 169–184). Springer Berlin Heidelberg.

[10] Holland, S., Hosny, A., Newman, S., Joseph, J., and Chmielinski, K. (2020). The Dataset Nutrition Label: A Framework to Drive Higher Data Quality Standards. In D. Hallinan, R. Leenes, S. Gutwirth, and P. De Hert (Eds.), Data Protection and Privacy: Data Protection and Democracy (pp. 1–26). Oxford: Hart Publishing.

[11] Kannan, K., Ananthanarayanan, R., and Mehta, S. (2018). What is my data worth? From data properties to data value. http://arxiv.org/abs/1811.04665

[12] Markopoulos, I., Kotsifakos, E., Bamboulidis A., Helminger, L., Tufiș, M., Bruni, A. (2020). D6.2 – Second phase personal data trials.

[13] Paraschiv, M., Boratto, L., Kassa, Y., Tufiș M. (2019). D4.1 – Report on requirements and design. H2020 Safe-DEED deliverable.

[14] Seck, I., Dahmane, K., Duthon, P., and Loosli, G. (2018). Baselines and a datasheet for the Cerema AWP dataset. CoRR. http://arxiv.org/abs/1806.04016

[15] Tufiș, M., Boratto, L. (2019). D4.2 – Baseline prototypes for data valuation. H2020 Safe-DEED deliverable.

[16] Tufiș, M., Boratto, L. (2020). D4.3 – Report on the context-aware and context-unaware valuation. H2020 Safe-DEED deliverable.