

QUALITY

DIGITAL MANUFACTURING PLATFORMS FOR CONNECTED SMART FACTORIES

D5.10 Open, Secure Industrial Data Space for ZDM

Deliverable Id:	5.10
Deliverable Name:	Open, Secure Industrial Data Space for ZDM
Status:	Final
Dissemination Level:	CO
Due date of deliverable:	30/09/2021
Actual submission date:	08/10/2021
Work Package:	WP5
Organization name of lead contractor for this deliverable:	Fraunhofer ISST
Author(s):	Marcel Altendeitering Markus Spiekermann Jan Jürjens
Partner(s) contributing:	Lukas Schulte (TUDO), Xiaochen Zheng (EPFL), Javier Hitado (ATOS), Giulia Giussani (IDSA) Sarah Vetter (UKL) Shayan Ahmadian (UKL) Abdulrahman Moussa (PACE)

Abstract: Report describing the developments within Task 5.5, which aim to realize the data exchange between different processes by leveraging the International Data Space (IDS). Moreover the autonomous data quality management is realized using an IDS Data App.



QU4LITY	Project	QU4LITY - Digital Reality in Zero Defect Manufacturing		
	Title	Open, Secure Industrial Data Space for ZDM	Date	30/09/2021
	Del. Code	D5.10	Diss. Level	CO

Contents

HISTORY	3
EXECUTIVE SUMMARY	4
1. INTRODUCTION	5
2. FUNCTIONING OF THE DATA APP	6
IDS COMMUNICATION	6
DATA QUALITY ANALYSIS	7
3. USER GUIDE	9
QUICK START GUIDE.....	9
USING THE APPLICATION	9
4. EVALUATION	11
PRELIMINARY EVALUATION WITH REAL-WORLD DATASETS	11
EVALUATION IN WP7	11
5. CONCLUSION	13
LIST OF FIGURES	14
LIST OF ABBREVIATIONS	15
REFERENCES	16
PARTNERS:	17

QU4LITY	Project	QU4LITY - Digital Reality in Zero Defect Manufacturing		
	Title	Open, Secure Industrial Data Space for ZDM	Date	30/09/2021
	Del. Code	D5.10	Diss. Level	CO

HISTORY

Version	Date	Modification reason	Modified by
1.0	30/09/2021	Initial Document	Marcel Altendeitering
1.1	06/10/2021	Included First Review Feedback	Marcel Altendeitering
1.2	08/10/2021	Small Corrections	Abdulrahman Moussa
2.0	08/10/2021	Final version	Marcel Altendeitering

QU4LITY	Project	QU4LITY - Digital Reality in Zero Defect Manufacturing		
	Title	Open, Secure Industrial Data Space for ZDM	Date	30/09/2021
	Del. Code	D5.10	Diss. Level	CO

Executive Summary

As it becomes easier to collect large amounts of data it is more important than ever to make validated decisions over what data to use for automated decision making and sustain data sovereignty at the same time. Towards this end, we leveraged the International Data Spaces (IDS) reference architecture and developed a novel Data App, which ensures a high-quality standard in data transfers. To accomplish this, we realized a data quality analysis that determines the quality of a data set along several dimensions and established a connection with the Data Space Connector. This results in an easy-to-use tool for data quality analysis available to IDS users which leads to high-quality data sharing.

QU4LITY	Project	QU4LITY - Digital Reality in Zero Defect Manufacturing		
	Title	Open, Secure Industrial Data Space for ZDM	Date	30/09/2021
	Del. Code	D5.10	Diss. Level	CO

1. Introduction

With the goal of Zero-Defect Manufacturing (ZDM) and Autonomous Quality (AQ) in mind collecting relevant data from factories is crucial, but even the largest amounts of data are only as good as its quality. If the collected data is of poor quality, e.g., if it is inaccurate, it is not suitable for deriving actions to improve manufacturing efficiency. To reach the goals of ZDM and AQ in light of complex supply chains and information networks we intended to leverage the standards of the International Data Space (IDS) and extend it with a component for autonomous data management. Specifically, we aimed to implement two IDS entities in a concrete usage scenario: the IDS Dataspace Connector (DSC) and the Data App (see also Deliverable 5.9). The Data App realizes autonomous data management by ensuring a high-quality standard in data transfers. We, hereby, intended an application that measures the quality of a data set and reports it as an easy-to-understand rating while offering compatibility with the IDS ecosystem. By reducing the complexity of this issue down to a concrete ration between zero and one the end user gains a tool to support decision making in inter-organizational data exchange. For example, the IDS compatibility allows to outsource the data analysis tasks to a partnering firm or research institute while sustaining data sovereignty and ensuring high quality data. Within task 5.5 of WP5 we focused on the development of the Data App and relied on an open-source implementation of the DSC for integration with the IDS.

In its current form the Data App works with JSON as well as CSV files featuring up to 52 individual sensors and has no specific limits on the number of values per sensor. For JSON files there are also no requirements regarding the data structure, as long as they have sensor names and corresponding values. Everything else can be configured using an environment file (see further in chapter 3).

In the following chapter, we discuss the details of our developments for the Data App and its integration with the DSC. This includes the used technologies as well as an overview of the preliminary evaluation. The third and final chapter is a user guide for the application, which allows for an implementation in a variety of scenarios.

QU4LITY	Project	QU4LITY - Digital Reality in Zero Defect Manufacturing		
	Title	Open, Secure Industrial Data Space for ZDM	Date	30/09/2021
	Del. Code	D5.10	Diss. Level	CO

2. Functioning of the Data App

IDS Communication

For a simple integration in an IDS Environment the Data App can pull data directly from a given DSC. We therefore relied on an open-source version of the DSC¹, which includes a docker compose file for fast provisioning. Thus, it can be quickly connected directly to another DSC at a production facility or the given data source in general. The presented Data App follows the guidelines of the International Data Spaces Association (IDSA) and the IDS Information Model², which means that it structures objects at the DSC as shown in Figure 1. All communication workflows shown are REST-based. To operate the DSC and the Data App need *Resources*. They consist of *Artifact-Representations* and *Contracts*. The first step is to create an *Artifact* via a POST request with the data/result attached. To allow a sovereign communication between connectors further structures need to be created, this happens automatically by the DSC when data is pushed to the connector and functions via GET requests. The procedure begins with creating a *Representation* of the *Artifact* which then gets linked to it, the same happens with a *Resource*, or more specifically, an *Offer*. After that a *Rule* and *Contract* are created and linked to each other as well as to the *Resource*. Finally, a *Catalog* with a link to the created *Resource* makes this structure accessible to partners having permission for it, which is specified in the *Contract*. Now data is stored at the DSC, either provided by a business partner through another Connector or by us (e.g. from backend systems). We can now pull the available data directly to the Data App via a GET request to perform the data quality analysis.

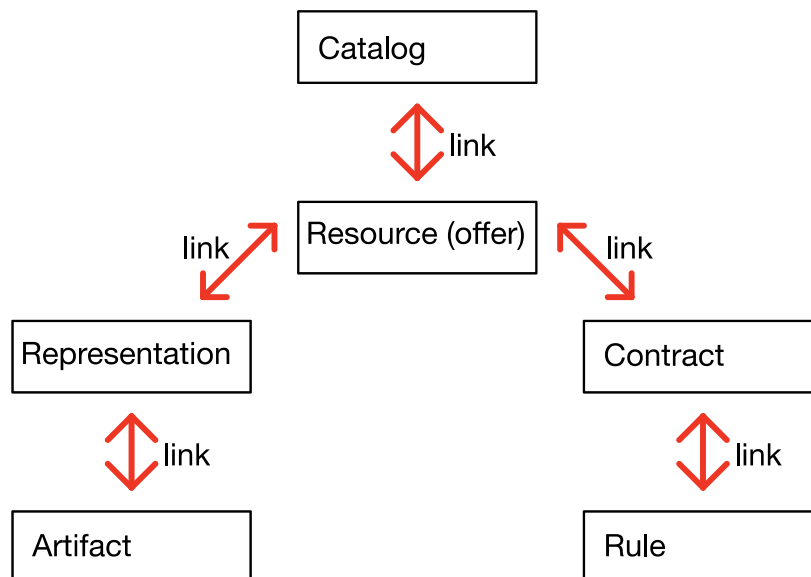



Figure 1 - IDS Connector Configuration

¹ <https://github.com/International-Data-Spaces-Association/DataspaceConnector>

² <https://github.com/International-Data-Spaces-Association/InformationModel>

	Project	QU4LITY - Digital Reality in Zero Defect Manufacturing		
	Title	Open, Secure Industrial Data Space for ZDM	Date	30/09/2021
	Del. Code	D5.10	Diss. Level	CO

Since the DSC currently saves files encoded as a base64 string the Data App automatically decodes them into a JSON or CSV file depending on the type provided. In case a JSON file is provided the App selects the relevant values from it (configurable through the corresponding environment file) and converts it to a CSV file for simpler data handling. In case of a CSV file no conversion is needed.

Data Quality Analysis

For implementing the Data App, we used the programming language Python as it is well suited for data analytics tasks and offers many useful frameworks. However, traditional Python analytics have its limitations with regards to large volume files, so we used Apache Spark³ as our processing engine to read in the input data. Upon creating a Spark RDD⁴ (resilient distributed dataset) the data gets split up into smaller chunks, which are then analyzed successively. Through this approach, the Data App is more scalable and can handle large amounts of data while utilizing well known python packages like SciKit-learn⁵. The utilized technologies are shown in Figure 2.

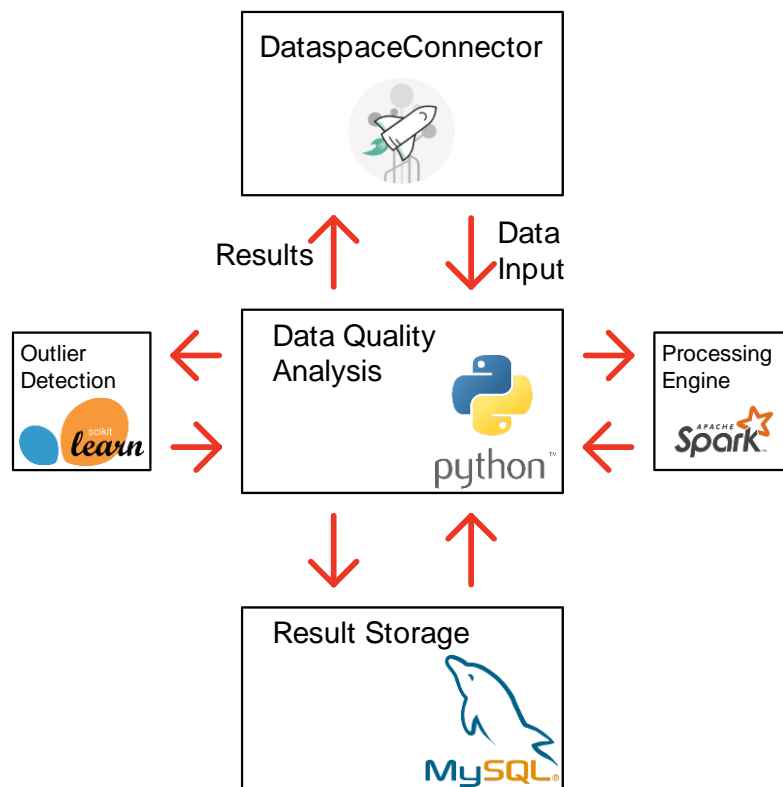



Figure 2 - Architecture Diagram

³ <https://spark.apache.org>

⁴ <https://spark.apache.org/docs/latest/rdd-programming-guide.html>


⁵ <https://scikit-learn.org/stable/>

	Project	QU4LITY - Digital Reality in Zero Defect Manufacturing		
	Title	Open, Secure Industrial Data Space for ZDM	Date	30/09/2021
	Del. Code	D5.10	Diss. Level	CO

To determine the quality of the provided data set, we combined four different data quality measures that are suitable for sensor-based inputs and cover a variety of data quality dimensions. Using an Isolation-Forest algorithm from the Scikit learn package each data block gets analyzed for outliers and based on the commonness of outliers an **outlier measure** is calculated. Furthermore, a potential **concept drift** in the data set is assessed on a per block basis. To store the results a connection to a MySQL database that gets created by the application and is tailored specifically for its needs. First, the results for each individual block are saved in this database so that in a final step the average over all blocks can be calculated leading to an overall measure of the data quality. Secondly, the concept of the current block is saved as an approximation, meaning that the boundaries and averages of each sensor is remembered in storage. Through this knowledge the previous concept can be compared to the next one leading to a measure of Concept Drift. The last two data quality measures that are considered are the **No Value Measure** and the **Constant Measure**. The first one looks for sensors that do not provide data over the whole block and the second one detects sensors staying constant for a long time.

All these measures are saved in the database and then combined to an overall measure that is averaged over all blocks in the dataset. Finally, the result in form of an easy-to-understand ratio between zero and one is sent back to the DSC and filed as its own resource such that it can be read by another connector or a user directly.

Additionally, the created database includes a table that is formatted specifically for visualization with Grafana and always held up to date. This means that with very little effort an illustration of the analyzed data as well as the calculated measurements can be added to the system if needed.

	Project	QU4LITY - Digital Reality in Zero Defect Manufacturing		
	Title	Open, Secure Industrial Data Space for ZDM	Date	30/09/2021
	Del. Code	D5.10	Diss. Level	CO

3. User Guide

Quick Start Guide

First open the 'DataspaceConnector-DataApp-data-structure-configurable-6.0.zip'. Then, if not already done, install Docker and Docker-Compose and unpack the downloaded archive.

To run the application including all necessary components and an instance of the Dataspace Connector open a terminal at this newly created folder und run 'docker-compose build' and 'docker-compose up'.

Using the Application

To allow the application know what data structure to expect we first need to configure it in case it differs from the standard structure found in the 'ValuesIdeko.json' file provided by Mondragon. This can be done by simply editing the 'data-structure.env' file which can be found in the application folder. Upon opening the file, you should find what you can see in the screenshot down below (Figure 2). The first variable specifies whether a JSON or CSV is being used. The following variables are only relevant if the application runs in JSON mode, for a CSV file no further configuration is needed as long as it fulfills the requirements specified in the file.


```

1
2 #This specifies in which format the datafile is. It can be set to 'JSON' or 'CSV'.
3 FORMAT=JSON
4 #If you are using a CSV as Input it needs to be in the specified Format of the sensor.csv file from the Repository.
5 #If there aren't 52 individual sensors available pls fill the rest of the table with 0 and specify how many sensors there actually are.
6 CSV_SENSOR_COUNT=52
7
8 #If you are using a JSON file the following variables are of importance.
9 #These defined sublevels tell the App how many levels deep into the JSON file the relevant Data can be found.
10 # e.g. if your Data looks like
11 # {
12 #   "payload":{
13 #     Data that should be analysed
14 #   },
15 #   {NEXT ITEM}
16 #}
17 #then SUBLVL_A=payload, SUBLVL_B=SUBLVL_C=/ is correct
18
19 SUBLVL_A=Payload
20 SUBLVL_B=/
21 SUBLVL_C=/
22
23 #Where are the names of the sensors written in the file and where are the corresponding values to it?
24 SENSOR_NAMES=IndicatorId
25 SENSOR_VALUES=Value
26
27 # (by default the variables are tuned for the ValuesIdeko.json file provided by Mondragon)
28
29

```

Figure 3 - Data App Configuration

In case a JSON file is what you want to use, the variables 'SUBLVL_A', 'SUBLVL_B' and 'SUBLVL_C' can be used (they are ignored if set to '/') to dig deeper into the file and find values of importance. In each entry in the file the algorithm first looks for the string 'SUBLVL_A' is set to and then proceeds with everything located under this heading. After that it looks for 'SUBLVL_B' in this sublevel we are already in. 'SUBLVL_C' works analogously. Lastly 'SENSOR_NAMES' specifies where the titles of the sensors that need to be analyzed are found and 'SENSOR_VALUES' where the

	Project	QU4LITY - Digital Reality in Zero Defect Manufacturing		
	Title	Open, Secure Industrial Data Space for ZDM	Date	30/09/2021
	Del. Code	D5.10	Diss. Level	CO

corresponding values are located. After configuring the application, we can continue to use it. If you are running the App locally it can be accessed under 'localhost', otherwise replace 'localhost' with the IP Address of your host.

First, we need to make sure that the App knows where to get the data from. If there is already an artifact with corresponding structures and a usage contract at the Connector you can use a POST request on 'http://localhost:8081/save_access_from_txt' with the plain text 'https://connector-container:8080/api/artifacts/[UUID_of_your_artifact]/data' in the body to let the App know where to request the data. If you want to pull the data from another DSC "connector-container" needs to be replaced with the IP at which it is hosted.

Alternatively, you can add your own file to the DSC with a POST on 'http://localhost:8081/dsc-config' and the data-file labelled as 'file' in the form-data body. The actual name of the sent file is not of importance, but the structure is, as mentioned earlier.

After that the data can be pulled from the DSC via GET on 'http://localhost:8081/dsc-input'.

And ultimately a GET request on 'http://localhost:8081/dsc-output' starts the Data-Quality analysis and uploads the result encoded in base64 to the DSC, where it can be accessed in its own artifact under the URL provided as a response to the send out request and in the terminal window of the app. A list of the artifacts available at the DSC can be found at 'https://localhost:8080/api/artifacts'.

QU4LITY	Project	QU4LITY - Digital Reality in Zero Defect Manufacturing		
	Title	Open, Secure Industrial Data Space for ZDM	Date	30/09/2021
	Del. Code	D5.10	Diss. Level	CO

4. Evaluation

Preliminary Evaluation with Real-World Datasets

To evaluate the accuracy of the data quality analysis, it has been tested with a two real-world data sets (data sets A and B) provided by the Mondragon pilot. Both datasets came as a JSON file and were converted to a CSV file as described above.

As a result of our quality analysis, we observed that data set A had some missing values for several sensors that usually provide values. The second dataset features consistent and reliable values. The absence of certain measurements is most of the time caused by errors in the installation of sensors e.g., a dead battery. This reduces the data quality of a dataset as well, so it lowers the score. With that in mind it makes sense that dataset A was rated with a 0.769 and dataset B with a 1, on a scale with 0 being the worst and 1 the best result.


This quick evaluation demonstrates that the algorithm is functioning and capable of effectively rating datasets based on their quality characteristics. Deeper and more complex evaluations of our developments as well as the integration with the IDS will take place in WP7 as they require the collaboration of several software components. The desired architecture is described in more detail in the following section.

Evaluation in WP7

As mentioned before, the data sets used for preliminary evaluation and testing have been provided by the Mondragon pilot. Currently, we are working on setting up a workflow between their production facility and our hosting of the Data App at Fraunhofer ISST. This work is part of WP7 and will continue after the end of task 5.5.

The intended architecture and workflow including all components is shown in Figure 3. We hereby aim to enable a sovereign and secure data exchange between the Mondragon pilot and Fraunhofer ISST using two DSCs. To realize the data exchange and quality analysis several steps are necessary.

1. A software component in the backend at a Mondragon production facility creates a resource at their implementation of the DSC via a HTTP-GET request and then adds the file via HTTP-POST request. This is done in regular intervals to provide the DSC with up-to-date data.
2. As soon as the file is available as a resource at the Mondragon DSC the DSC at the Fraunhofer ISST can pull it, so that the data quality analysis can be initiated.
3. The DataApp subscribes to the DSC and pulls the data that was retrieved from the Mondragon DSC. Afterwards, it initiates the data quality analysis. Once the rating is calculated, it is made available again as a new resource at the DSC on the ISST site.

	Project	QU4LITY - Digital Reality in Zero Defect Manufacturing		
	Title	Open, Secure Industrial Data Space for ZDM	Date	30/09/2021
	Del. Code	D5.10	Diss. Level	CO

4. The ISST DSC notifies the DSC at Mondragon via an update message whether new results are available in a regular interval.
5. This makes sure that the Mondragon site knows timely when new results are available which they can pull and use for further decision making.

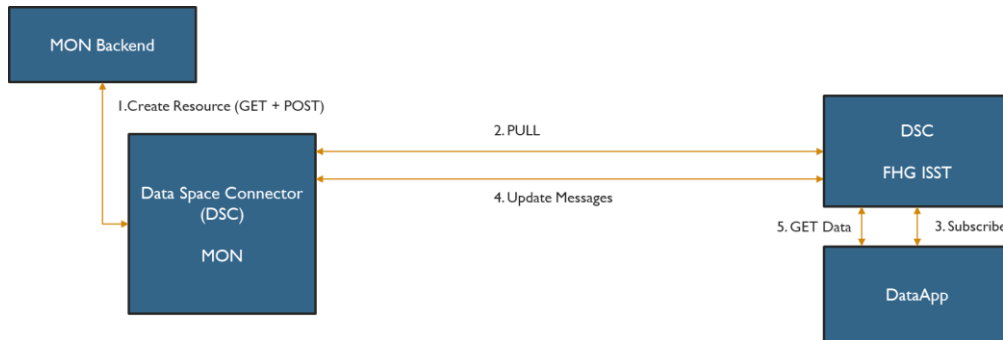


Figure 4 - Evaluation with the Mondragon pilot

Once the workflow is set up, we will perform several evaluations and tests with further data sets. This evaluation will not only cover the Data App but also the orchestration of all components as well as their functionality and usability. This way, we hope to gain further insights in using the IDS ecosystem for inter-organizational data exchanges and what are current obstacles and downsides that should be addressed in future developments. We also hope to gain an understanding of the role the IDS and data quality analysis could play in reaching ZDM and AQ. We will report our findings in the documents provided in WP7.

We would like to thank Mondragon for their participation and collaboration!

QU4LITY	Project	QU4LITY - Digital Reality in Zero Defect Manufacturing		
	Title	Open, Secure Industrial Data Space for ZDM	Date	30/09/2021
	Del. Code	D5.10	Diss. Level	CO

5. Conclusion

In this project, we developed a tool for a data quality analysis, which links directly to the IDS. The goal and the functioning of this application is described in detail and a comprehensive user guide is given. This enables a straightforward integration in a variety of production facilities and gives decision makers an easy-to-use tool for high-quality data sharing. We believe that this can contribute to the pursuit of ZDM and AQ in a meaningful way and lead to a more informed decision-making process.

We successfully evaluated the developed Data App with real-world data sets provided by the Mondragon pilot. This proves the applicability and usefulness of our developments in real-world manufacturing. As part of future developments, we will integrate our developments in the architecture of the Mondragon pilot for a more detailed evaluation.

QU4LITY	Project	QU4LITY - Digital Reality in Zero Defect Manufacturing		
	Title	Open, Secure Industrial Data Space for ZDM	Date	30/09/2021
	Del. Code	D5.10	Diss. Level	CO


List of figures

Figure 1 - IDS Connector Configuration	6
Figure 2 - Architecture Diagram	7
Figure 3 - Data App Configuration.....	9
Figure 4 - Evaluation with the Mondragon pilot	12

QU4LITY	Project	QU4LITY - Digital Reality in Zero Defect Manufacturing		
	Title	Open, Secure Industrial Data Space for ZDM	Date	30/09/2021
	Del. Code	D5.10	Diss. Level	CO

List of Abbreviations

AQ	Autonomous Quality
DSC	Dataspace Connector
IDS	International Data Spaces
IDSA	International Data Spaces Association
Spark RDD	Spark Resilient Distributed Dataset
ZDM	Zero Defect Manufacturing

	Project	QU4LITY - Digital Reality in Zero Defect Manufacturing		
	Title	Open, Secure Industrial Data Space for ZDM	Date	30/09/2021
	Del. Code	D5.10	Diss. Level	CO

References

International Data Spaces Association

DataspaceConnector:

<https://github.com/International-Data-Spaces-Association/DataspaceConnector>

Information Model:

<https://github.com/International-Data-Spaces-Association/InformationModel>

Apache Spark

Main Website:

<https://spark.apache.org>

Spark Resilient Distributed Dataset:

<https://spark.apache.org/docs/latest/rdd-programming-guide.html>

SciKit-learn

Main Website:

<https://scikit-learn.org/stable/>

QU4LITY	Project	QU4LITY - Digital Reality in Zero Defect Manufacturing		
	Title	Open, Secure Industrial Data Space for ZDM	Date	30/09/2021
	Del. Code	D5.10	Diss. Level	CO

Partners:

