DELIVERABLE

BIGPROD Platform operation: intermin report

Summary

This report provides an overview of the implementation of a data platform for the BIGPROD project. To best meet the project needs, the data platform is composed of three areas:

1. Area 1: Cloud hosted PostgreSQL database to facilitate the data exchange between partners;
2. Area 2: NoSQL Mongo Database hosted at PPMI to store data collected from the web;
3. Area 3: Jupyter Hub server to expose some selected project datasets to the general public.

Such design was chosen because it best meets the requirements of the project, while also ensuring compliance for the personal data protection and GDPR. A more detailed overview of the platform design and personal data protection steps is presented within the interim report.

Overall, the progress to date is good and all the platform work is on schedule. Two of the platform Areas – Area I and Area II are fully operational, while the launch of the Area III is scheduled in M19 of the project. This is because Areas I and II are essential for collecting the project data and facilitating the data exchange between the project partners, while Area III is mostly for disseminating project results and engaging the community.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870822

Deliverable Information

Deliverable number and name:	**D8 Interim Platform Operation Report**
Due date:	26th November 2020
Deliverable:	D8
Work Package:	WP3
Lead Partner for the Deliverable:	PPMI
Author:	Lukas Pukelis
Reviewers:	Arho Suominen, Hugo Hollanders, Scott Cunningham
Approved by:	Arho Suominen
Dissemination level:	Public
Version	v. 1.0 26th November 2020 v. 2.0 18th January 2022

Disclaimer

This document contains a description of the **BIGPROD** project findings, work and products. Certain parts of it might be under partner Intellectual Property Right (IPR) rules so, prior to using its content please contact the consortium coordinator for approval.

In case you believe that this document harms in any way IPR held by you as a person or as a representative of an entity, please do notify us immediately.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

This publication has been produced with the assistance of the European Union. The content of this publication is the sole responsibility of BIGPROD consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 27 Member States of the Union. It is based on the European Communities and the member states cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors (http://europa.eu.int/)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870822.

Introduction

This deliverable presents an overview of the current state of the BIGPROD project data platform. In the proposal and the initial deliverables of the project, we have proposed a platform composed of three areas:

1. Area 1: Cloud hosted PostgreSQL database to facilitate the data exchange between partners;
2. Area 2: NoSQL Mongo Database hosted at PPMI to store data collected from the web;
3. Area 3: Jupyter Hub server to expose some selected project datasets to the general public.

The design was chosen because it best meets the requirements of the project, while also ensuring compliance for the personal data protection and GDPR. A more detailed overview of the platform design and personal data protection steps is described within the report.

The structure of the report closely mirrors that of the data platform. The report consists of three main sections corresponding to the areas of the platform. Each section begins by outlining the design of the specific platform area and explaining the logic behind making such design choices. The sections then proceed to outline the work completed on each section during Y1 and set the agenda for Y2.

Two final sections of the report present the work carried out to slightly extend the platform by adding a documentation and code repository to better facilitate data and information exchange between the project partners. The very final section of the report is dedicated to our approach to personal data and GDPR compliance.

The biggest section of the report is dedicated to Area 2, which collects raw text data from company websites, processes it, extracts new insights and merged them with additional data sources. Overall, the progress made in this area is very significant and has already produced very high-quality results. Especially noteworthy is the capability to assign company website texts to the hierarchical topic tree from Microsoft Academic database. This is significant, as it opens new opportunities to link and cluster companies based on the contents of their websites. While previous approaches relied on explicit keyword overlap to do this, our approach represents a significant improvement as it allows the connections to be made via more abstract topics covered in the text.

Planned BIGPROD Data Platform Architecture

In the proposal, we have outlined the following roles for the BIGPROD data platform:

1. It will store the project data assembled from various sources;
2. It will facilitate the data exchange between the consortium partners;
3. It will serve and expose a selected sub-section of the data to the end-users;
4. It will ensure data protection, security and privacy.

The BIGPROD project utilises Big Data and unstructured data from a variety of sources. We project that by the project's end, we will have gathered and processed well over 1 TB of data. Furthermore, the project has multiple partners in diverse geographical locations performing different tasks, which require constant and quick updates/ changes to the data. Additionally, while we do not expect to generate very intense user traffic (compared to the popular platforms on the internet, such as news portals), we expect our visitors to interact with the data collected and analysed during the project, which means that each platform user will create demands for the platform resources. Finally, each component of the platform and the system must fully meet the security and privacy requirements.

In other words, it is required that the platform facilitates different levels of access and performs diverse functions: from data manipulation to data visualisation and storytelling. Given this diversity of requirements we have decided to segment the platform into three autonomous areas:

1. Area 1: Cloud hosted SQL database;
2. Area 2: On premises hosted NoSQL database operating in PPMI;
3. Area 3: Jupyter Notebook Server

This way, we can select a sub-system best meeting the demands for that area and connect them together in such manner that benefits the needs of the project the most. Using such an approach, we ensure that the data is clearly separated, i.e. that no sensitive data is exposed to the end user or that no proprietary data is exposed to the project partners that would violate the usage agreements with the third-party providers (e.g. Bureau van Dijk). The schema for the platform high level architecture is shown in Figure 1.

The centrepiece of the platform is Area 1, which hosts the main datasets for the BIGPROD project:

1. Company descriptive data;
2. Indicators calculated from the data scraped from company websites;
3. Indicators calculated from review and other websites;
4. Indicators calculated by matching company records with other databases (PATSTAT and EUIPO);
5. Results from the econometric modelling;

6. Other indicators.

These datasets are in a cloud-hosted SQL database. The main database with all the project datasets will not be exposed to the public, instead datasets approved for public use by the consortium will be transferred to another smaller database, which will be made publicly available.

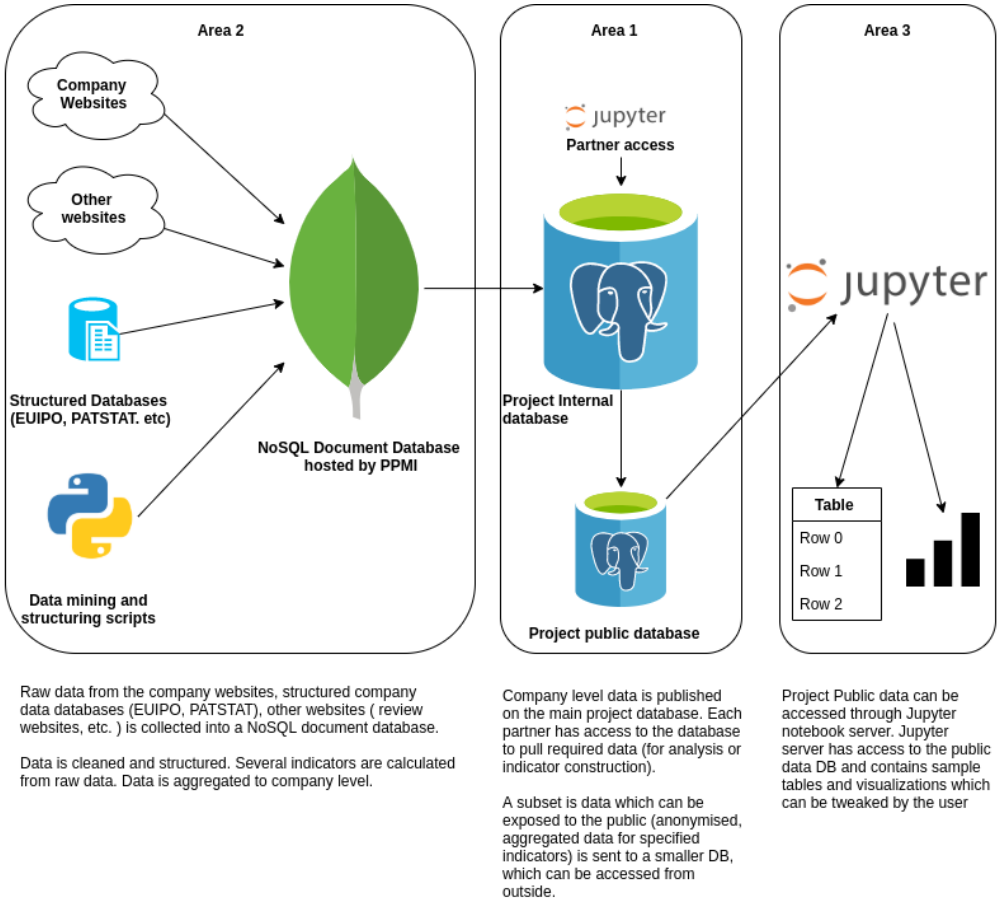

Figure 1 BIGPROD Platform High Level Architecture (Source: BIGPROD Project)

Area 2, meanwhile, houses the data scraped from company and other websites as well as interim data needed for indicator construction. Finally, Area 3 allows users to view and interact with the project data via Jupyter Hub platform. Jupyter Hub server fetches the data from the smaller public database in the Area 1 and allows users to perform various data analysis steps and do data visualisation.

The areas of the platform are clearly separated and the data flows between them strongly regulated. This is done to ensure data security and to prevent possible data breaches, at the same time ensuring that the project team has free, full, and easy access to the project data.

It is our understanding that such platform design can meet the requirements of the project best. Sections of the report below provide more details on the progress made implementing the planned features in the three areas.

Overview of the currently implemented platform features by Area

Overall, during the first year of the project, the progress with implementing the platform was steady and we are on schedule. Naturally, during the first year of the project most attention was paid to Areas I and II which facilitate data gathering and exchange between the project partners. These two areas are fully operational at the time of writing and will change only incrementally as new variables crafted from the unstructured data (such as “Mission and Vision” statements of companies) are added to the existing data. Meanwhile, the platform development in Y2 of the project will mostly focus on Area III, which will serve to disseminate project results and to engage the wider community.

Area 1: Cloud-hosted SQL Database

Description

This is the main area in which all the data created during the project resides. This database brings together the company-level data from:

1. “Orbis” database;
2. Indicators derived from unstructured company website/ review website data;
3. Indicators derived from semi-structured data sources (PATSTAT, EUIPO);
4. Indicators derived from the econometric modelling;

We have chosen to store the data in this Area in a SQL (Postgre) database. This format was chosen because the data in this area follows a uniform schema and because SQL is a widely familiar and established database standard, making it easily accessible by the project partners. We have chosen PostgreSQL as the database server because of its well-developed functionality and performance. As with all the components for the platform, we only chose open source tools.

We have decided to host this database in the Microsoft Cloud. The choice to host the database in the cloud was prompted for several concerns:

1. We wanted to minimise additional load on the PPMI infrastructure (which is hosting Area 2 of the platform);
2. We wanted to ensure high reliability and availability of the database;
3. We wanted to simplify networking and security aspects of the infrastructure deployment.

Progress during Y1

The database is live at **http://40.113.156.93:1815.** Because it contains sensitive data for project internal use, it does not facilitate anonymous access. Only registered users with passwords can access the data. Our initial intent was not to directly expose the database to the internet, but this was changed after consulting the project partners and evaluating their preferences and needs.

We have chosen the MS Azure out of the existing Cloud infrastructure providers because it has features which allow to ensure that the data is stored and backed-up in the EU and no data related to the functioning of the Area 1 will leave the EU.

The database is being iteratively populated by the consortium partners. Company data collected and processed by PPMI is uploaded in batches and each partner is responsible for uploading the data they curate. That is done either independently or by passing on the prepared datasets to PPMI to handle the upload.

Since no sensitive data is collected during this project and the personal data that is collected is part of the public domain, the BIGPROD project does not need to take any highly advanced measures to ensure the GDPR compliance, however, each project partner and especially PPMI will remain GDPR-aware and GDPR-cautious at all the infrastructure design and set-up steps, we will ensure that the provisions of the Data Management Plan (D33) and Personal Data Protection Strategy (D7 and D10) are followed in each step of the process.

The BIGPROD database follows the schema presented in Figure 2. Currently, we have around 20 K companies in the database, which constitutes around 10% of the company sample. Additional companies are being added to the database in batches as PPMI finishes scraping and processing their data.

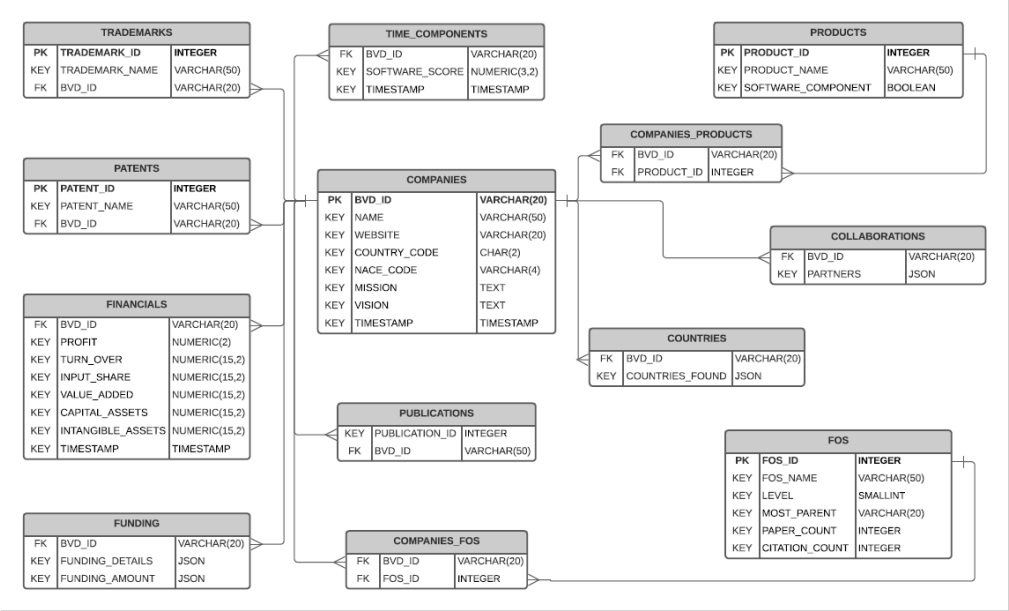

Figure 2- BIGPROD Platform Area I Database Schema

Agenda for Y2

During Y2 we will add the remaining companies to the DB. Since the DB schema is finalised, we do not foresee any difficulties with this step. Additionally, we will isolate the specific data structures in the database (views/tables) and select them to be exposed to the wider public. To do so, we will create a smaller separate database, which will sync these data structures with the main database. This smaller database containing the data for the public use will be exposed to the internet directly as part of Area III of the data platform.

Area 2: No-SQL database hosted by PPMI

This Area is responsible for collecting the data from company websites and processing and enriching it through the text-mining and text classification processes. Additionally, it facilitates the linking between the company website data and other data sources, such as PATSTAT for patents or EUIPO/TM-LINK for other IP. Naturally, this Area received the most development and attention during Y1, as many of the downstream tasks depend on it.

Because of the need to store large amounts of text data, we chose to implement this database as NoSQL document database. Among the many NoSQL options, we chose MongoDB due to its reliability and performance. Since the data collected in the Area 2 generally will not be shared with and accessed by the other consortium partners, we chose to host this database on PPMI premises and keep behind the firewall, unreachable from the internet. The database is populated by a multitude of worker processes which can reach the internet and bring the data to the database. Yet another set of worker processes ensure that the relevant datasets in Area 1 are kept up to date.

In Area 2 the data from the company and other websites as well as semi-structured database data is pooled together. Then various data mining, information extraction, text classification and text fragment matching algorithms are run in order to:

1. Identify and extract valuable pieces of information from the collected raw data;
2. Identify texts with relevant content for further detailed analysis;
3. Match fragments of text, e.g. product names to other records to link and enrich the data;
4. Construct indicators from the collected data.

As a result, from the disaggregated dataset, where a unit of analysis is a single URL, we build a company-level dataset, which contains all the indicators relevant to the BIGPROD project. This dataset is then synced to the main database in Area 1 and made available to all the project partners. We also foresee, uploading highly specific sets of disaggregated data, such as company Mission and Vision statements, to the Area 1 database as well. However, these cases will be limited in number, will not contain any personal data, and will be of high and direct relevance to project aims.

Progress during Y1

During Y1 we have made good progress in Area II and all the activities are on schedule. Since most of the data collection and initial data processing takes place in Area

II, this area has received the most attention and most of PPMI's work in Y1 took place in this area. Sections bellow breaks down our progress by sphere and outlines the main developments.

Web scraping

Prior to starting the BIGPROD project PPMI already had developed a powerful web-scraper which can fully traverse company domain and extract text from various elements, including dynamic Java Scrip sections of the page. As such, PPMI scraper was capable of scraping company websites better than other commonly used scrapers. However, we needed to scale our capacities to scrape considerably to meet the requirements of this project.

We succeeded in doing that and increased the capacities of the scraper from about 1M web-pages per day initially to about 5 M at the time of writing. We will continue to work on and improve the scraper, but the current capacity is more than sufficient for the project needs.

Text mining

After scraping the company websites, we perform various text-mining tasks to retrieve relevant information and to construct indicators from the unstructured data. We start by extracting various artifacts, such as country names, ISO standards and CE marks from the text. We then proceed by doing more elaborate text mining for:

Text Mining: Company Products

We use a combination of linguistic dependency parsing and machine-learning to identify company products. This is done by looking for various phrases like "we are introducing a new NOUN" or "we are happy to announce a NOUN" in the pages which an ML model has labelled as news announcements. We supplement these products by also looking for frequently occurring entities with trademark signs (™) in these announcements.

Text Mining: Collaborations

We employ a similar approach to identification of collaborations and the entities a company collaborates with. We also employ a set of phrase patterns which we apply onto ML per-selected texts to ensure accuracy and minimize false-positive instances. For collaborations we also employ simple classification on the extracted entities. By using a list of key-words, we classify the entities into education sector entities and the rest.

Text Mining: Funding

In the same vein, we also extract statements related to funding received by a company. At the moment, we identify the instance of funding and extract the funding amount. We will also attempt to identify the funder from which funding was received by cross-referencing the extracted entities with a list of known funders.

Text Classification

In addition to text mining, where the aim is to extract valuable data from the text, we also perform text classification with the aim to assign the text to a specific category. We perform two main text classification tasks: assigning texts to fields-of-study/topics from Microsoft Academic and assigning texts to the Sustainable Development Goals (SDGs).

Text Classification: Fields of Study

Microsoft Research produces an open-source equivalent of Google Scholar, called Microsoft Academic Graph[1] (MAG). All the MAG data can be downloaded and used for research. One of its most interesting features is the topic categories assigned to research papers called "Fields of Study" (FOS)[2]. We have taken this data and developed a tool to assign the FOS categories to any text input. The tool works by vectorising the input text and creating a TF-IDF vector from it. This vector is then compared to a set of pre-computed TF-IDF vectors for FOS fields. Text is assigned to FOS'es which are the most similar to its vector (have smallest cosine distance). We use this intensively in the BIGPROD project and assign these categories to companies and products. These categories can serve to link unstructured data sources and to calculate similarities between texts.

Text Classification: SDG

In the BIGPROD project we also seek to identify companies which are active in the areas linked to the UN's "Sustainable Development Goals". To this end we employ OSDG[3] a text classification tool which PPMI has designed in the past to classify EU-funded research projects to SDGs and which has since been open-sourced and developed jointly with the United Nations Development Programme (UNDP) and selected academic partners. The OSDG is essentially a mapping between FOS categories and the SDGs. By utilising the FOS tags from the previous section, we can easily map texts to SDGs.[4]

Automation

Additionally, during this period we also focused on automating the data processing process, which includes data cleaning and text mining tasks. On this front we also proceeded as planned and have implemented a robust automatic data processing pipeline using Apache Airflow as the main orchestration framework (see Figure 3).

Company matching

PATSTAT

To date we have developed an algorithm that uses company meta-data to match companies in the BIGPROD sample to PATSTAT database. To perform the matching, we use company name, location information as well as alternative names or aliases used by the company. The algorithmic matches have since been manually validated and have generally proven to be highly precise. We are currently in the process of testing the recall of the algorithm. In its current iteration, we observe that the algorithm has matched roughly 20% of companies to entries in PATSTAT which is consistent with the expectations.

[1] Microsoft Academic: https://academic.microsoft.com/home
[2] Microsoft Academic FAQ : https://academic.microsoft.com/topics; Paper: https://arxiv.org/pdf/1805.12216.pdf
[3]OSDG OSDG.ai
[4] Pukelis, L., Puig, N.B., Skrynik, M. and Stanciauskas, V., 2020. OSDG--Open-Source Approach to Classify Text Data by UN Sustainable Development Goals (SDGs). *arXiv preprint arXiv:2005.14569*.

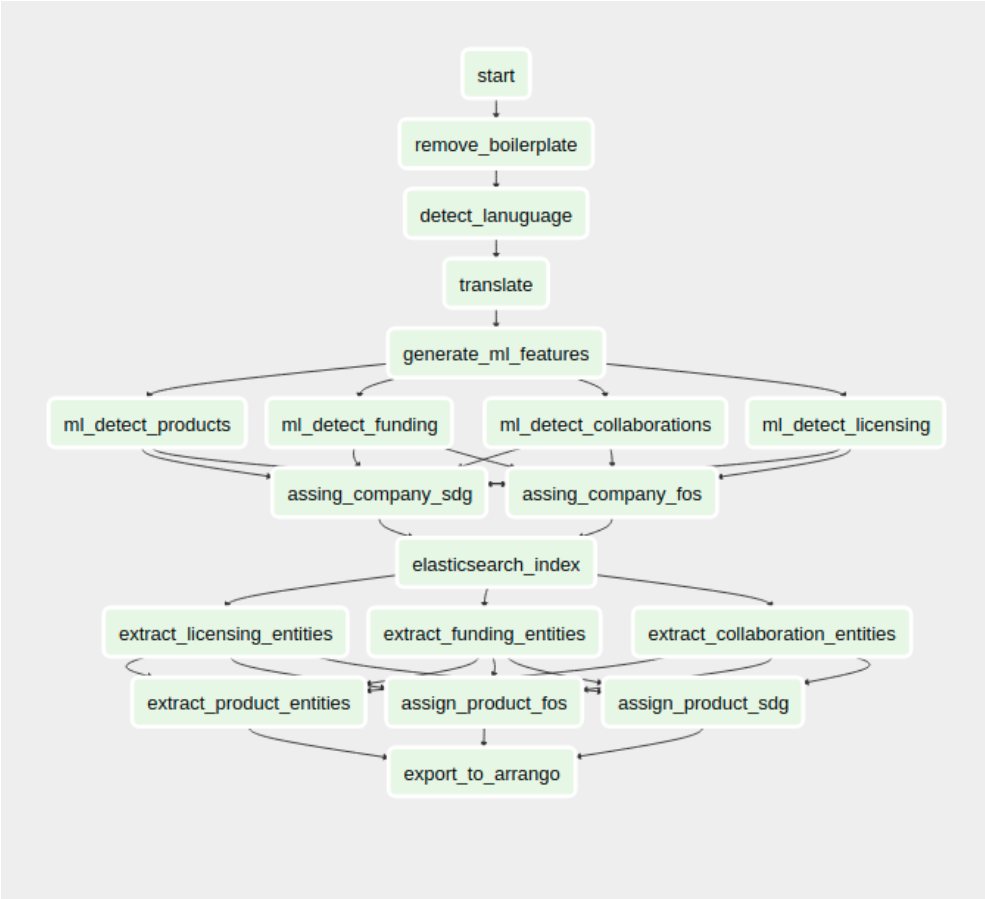

Figure 3 Automated Data Processing Pipeline for BIGPROD project

Agenda for Y2

During Y2 of the project, we will continue to improve various aspects of Area II. We are currently developing next-generation algorithms that will extract more detailed data on products, funding, and collaborations. More specifically, for products we will seek to cross-reference the IPR-protected product names with the respective databases to determine if the company on which website a particular product was mentioned actually owns the copyright to that product or just uses it in its activities or distributes it to consumers. For funding, we will seek to enhance the data by extracting the name of the funding authority and cross-referencing that with a list of known funders. Finally, for collaborations, we will seek to link these names of collaborators with the "Global Research Identifier Database (GRID)".[5] Currently, we have developed prototypes for all these algorithms, but they need further refinements to improve their precision and recall before they can be deployed.

[5] Global Research Identifier Database < https://www.grid.ac/>

Furthermore, an additional workstream for Y2 will be the development of new indicators derived from company website text. First, we aim to develop a “Company Digitalisation Score” indicator which will showcase what share of company’s products have a software component to them. With this indicator we seek to capture not only companies that mainly produce software, but also companies that introduce some software to complement their “hardware” products. A good example of such behaviour is the rise of computer-assisted-manufacturing (CAM) which requires both hardware – e.g. a milling machine – and software to operate it automatically. Another example would be companies producing consumer apps to help consumers to operate/service their products. We are currently prototyping an algorithm to identify whether the product descriptions we extract contain mentions of software. Once that is done, we will calculate the indicator by counting what share of company products have a software component.

Additionally, we will develop indicators based on the contents of company “Mission and Vision” statements. Currently, we are working on a strategy to detect and identify these statements among the scraped texts. After this is done, we will derive new indicators based on what topics/areas these statements cover, what actions/values they mention, etc.

Area 3: Jupyter Notebook Server

The final component of the platform is Area 3, which facilitates the access to the specified BIGPROD datasets to the end-users. We foresee that the main interest in the project results will come from specific groups:

1. Scholars from academia;
2. Policy professionals;
3. Analysts from research organisations.

In other words, we foresee that, due to the highly technical nature of the project, the main groups interested in its results will be the people already working in the general sphere of business productivity. We also expect that they will have some basic skills in working with the data and would like to pose some highly specific and advanced questions to the public BIGPROD dataset. For this reason, in addition to the project website which will provide the basic information about the project and its results, we will also create a specialised area where the advanced users will be able to perform queries of their own design to the public BIGPROD datasets. For an example, how Jupiter notebook looks like, see Figure 4.

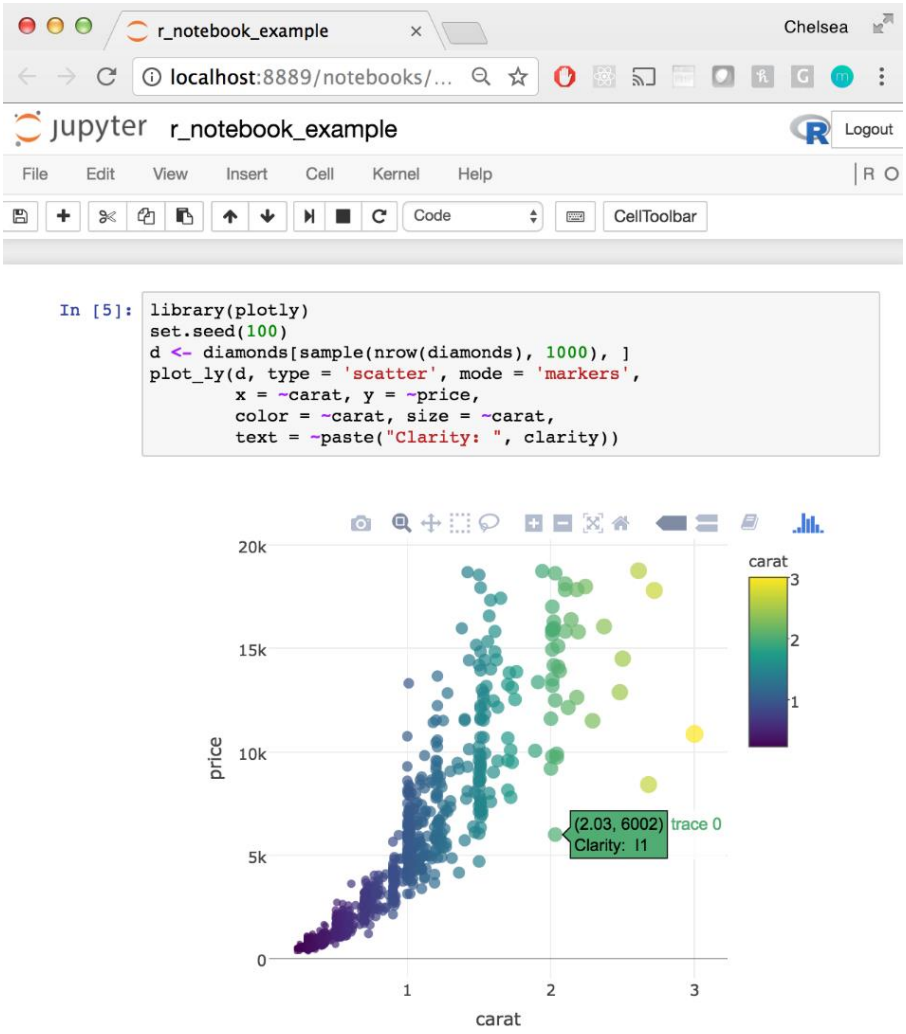

Figure 4 A sample snapshot of Jupyter notebook. (Source: https://plotly.com/python/ipython-notebook-tutorial/)

This will be achieved via the specialised Jupyter Hub server, which will be accessible to the users from the general public upon registration. What makes notebooks on Jupyter hub great for this exercise is that they allow to combine several different aspects:

1. **They allow to facilitate the story telling** – Jupyter notebooks work great in combining text narrative, code commands and data visualisations. In other words, they allow us to prepare report-like texts and illustrate them with interactive graphs and visualisations to facilitate the data exploration by the user. All the data is generated by the code snippets presented alongside the text, to ensure transparency and reproducibility. Furthermore, these snippets can be edited to modify the sample tables and visualisations to user preferences;
2. **They allow to facilitate the interactive exploration of the data** – Jupyter notebooks also allow the users to perform the data exploration and analysis on their own. In other words, users can explore how the BIGPROD data can help them to answer their own research questions, which were not asked by the project. In this way, the Area 3 will allow users to **co-create knowledge** together with the BIGPROD consortium.
3. **Jupyter notebooks facilitate several programming languages -** all the main languages used form data analysis, such as R, python, Scala, and Julia. This ensures that they can be

used by many people from different backgrounds.

We plan to install a dockerised version of the Jupyter Hub (server) on the same virtual machine as the BIGPROD database. However, we would only grant access to the smaller public database to the Jupyter Hub.

The BIGPROD website (http://www.bigprod.eu/) will contain the link to the Jupyter Hub in Area 3 as well as the access request form. Users interested in the BIGPROD data will have to fill out the form, detailing how they are intending to use the data. These requests will be evaluated by the BIGPROD consortium and approved users will be granted a username and password to be able to access and use the Jupyter notebooks.

Progress during Y1

Based on the design of the BIGPROD data platform, Area 3 will house a smaller database and Jupyter Notebook server for members of the wider public to access and explore BIGPROD data. Development of this Area is planned for Y2, when more data is shared on Area 1.

Agenda for Y2

During Y2, we are planning to do the following: first, get the required infrastructure in place. This will require installing Jupyter Hub on the VM and scheduling periodic data exports to the public database. Second, prepare sample data analysis notebooks to facilitate user-interaction with the data. Third, reach-out to a small number of stakeholders and target groups and collect their feedback on the notebooks prior to the full-launch. Upon the incorporation of the feedback and the necessary improvements, the notebooks will be fully opened to the public and disseminated via various channels.

Additional work to supplement the Data Platform

During Y1 it became obvious that additional infrastructure was needed to house data processing scripts, sample queries “how-to” guides and other resources related to the technical data exchange between the project partners. As such, an additional repository on GitLab was created[6].

In addition to the above-mentioned features, we are also developing a dedicated wiki, where we store data documentation, database schemas and other resources. Usually, new wiki pages are added following the consortium calls and relate to the discussions had therein. In such way, if some aspect of project implementation required additional clarification, they are all added in one place, so they can easily be looked up again (see Figure 5).

[6] Access point https://gitlab.com/ppmi-data/bigprod

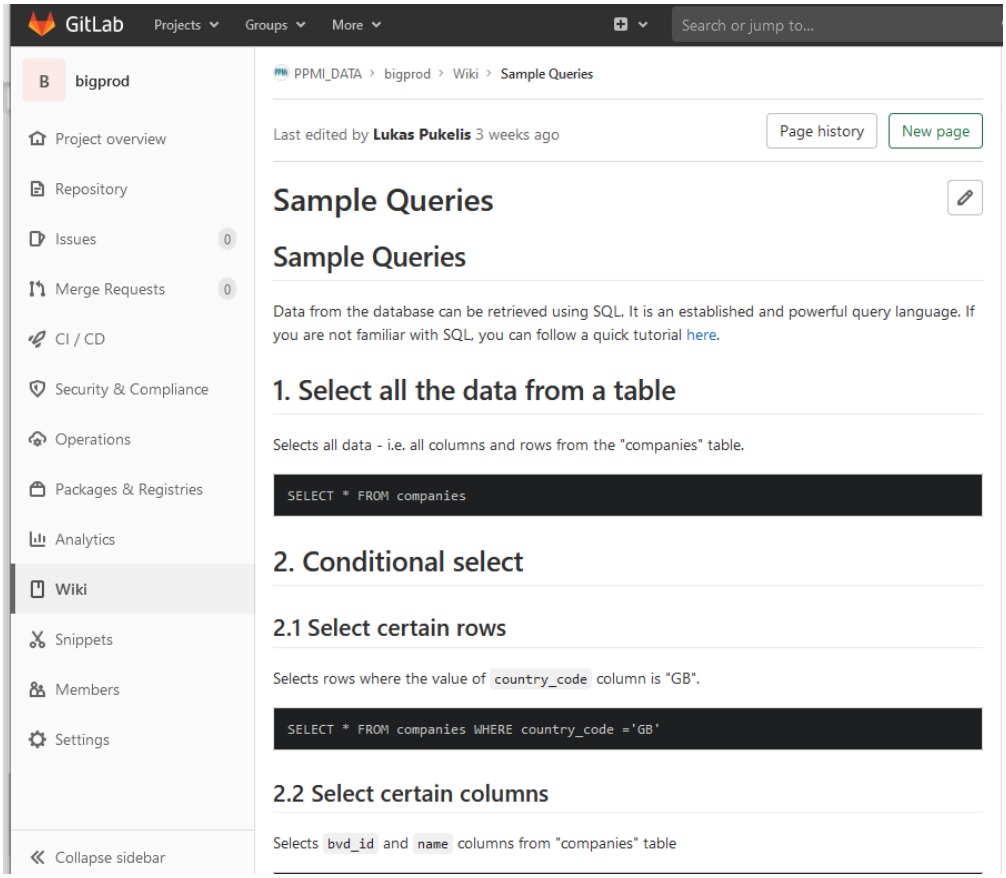

Figure 5 Illustration of GitLab wiki (Source: BIGPROD)

Personal Data Protection and GDPR

We closely follow the measures outlined in the project Data Management Plan (D33) and Personal Data Protection strategy (D7 and D10) to ensure data protection and GDPR compliance. The key points underpinning our actions are:

1. We are collecting data from the web and integrating data from various sources for the purposes of scientific research;
2. We do not intentionally or specifically collect any personal or sensitive data. We do not seek to identify or target individuals in this project.
3. Some personal data from the public domain will be collected during the project as a part of the general data collection effort – inventor names from patents or person names from company websites. However, we take measures to minimize the collection of such data. We do not include such data in indicator construction. We do not circulate such data within the consortium.
4. We employ the platform design, which separates the platform into three functional areas and strictly specify what data can be passed between the areas. In such way we ensure that the data security and prevent possible data breaches.

For more information, please contact

Dr. Arho Suominen (Consortium leader)
Tel. +358 50 5050 354
arho.suominen@vtt.fi

About BIGPROD

BIFPROD is a research project focusing on Big Data based analysis of productivity using webscraped data. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 870822.

The project partners in the project are Quantitative Science and Technology Studies team, Foresight-driven Business Strategies, 1) VTT Technical Research Centre of Finland, Competence Center Innovation and Knowledge Economy (Coordinator), 2) Fraunhofer ISI, Economics of Knowledge and Innovation team, 3) UNU-MERIT, Maastricht University, 4) Public Policy and Management Institute, 5) Economics of Technology and Innovations, Faculty of Technology, Policy and Management, 6) Delft University of Technology, Economics of Technology and Innovations, 7) Faculty of Technology, Policy and Management, Delft University of Technology

VTT PPMi Fraunhofer ISI Maastricht University TU Delft Delft University of Technology University of Strathclyde Glasgow

www.bigprod.eu

BIG PROD