DELIVERABLE

BIGPROD data platform operation report

Deliverable Information

Deliverable number and name:	**Final platform operation report**
Due Date:	March 31, 2022
Deliverable:	D9
Work Package:	WP3
Lead Partner for the Deliverable:	PPMI
Author:	Lukas Pukelis
Reviewers:	Arho Suominen, Hugo Hollanders, and Scott Cunningham
Approved by:	Arho Suominen
Dissemination Level:	Public
Version	Apr 22, 2022

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

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 the case that 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 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 program under grant agreement No 870822.

Table of Contents

BIG PROD

Summary

This report provides an overview of the implementation of a data platform for the BIGPROD project. We have created a data platform composed of the following principal areas:

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

This design was chosen because it best meets the requirements of the project -- it allows for efficient, large-scale data collection and effective data exchange between project partners. This approach also allows for sharing a sub-set of the collected data with the general public to disseminate the project results. All this functionality is achieved while also ensuring compliance with personal data protection and General Data Protection Regulation (GDPR) procedures. A detailed overview of the platform is presented within this report. The report is an updated version of the previously published interim report.

Overall, the design and the implementation of the data platform are suitable for the project's needs. The platform facilitates an effective exchange of data between partners and allows for efficiently collaborating on indicator development and analysis tasks. The main shortcoming of the data platform is that it presents the project results to the public using Jupyter notebooks. The notebooks have been very well-received by the members of the public who have some coding experience or knowledge of the Python programming language, but they are difficult for people without these skills. In future projects, the data platform should have a dedicated user interface, such as an PowerBI or Tableau dashboard, that would allow for viewing and interacting with the collected data without having coding skills.

BIG PROD

Introduction

This deliverable presents an overview of the BIGPROD project data platform. The report is an updated version of the first interim version of the report[1] This updated version of the data platform operation report is intended to revise the earlier version with aspects of the platform that have been improved since the interim version. The report covers the functional parts of the platform and describes their functionality.

The intended audience for the report is users of the data resources offered by the project. The BIGPROD project has made data and accompanying code available on the Dataverse repository[2]. This data platform operation report offers users insight on the functions of the data platform that has been used to create the data and variables drawn from the data. The report is not intended for a detailed technical description of each of the functional parts. Additional details are, however, available upon request from the author.

In the following section, the report will cover the overall data platform architecture and the three major Areas the system is constructed on. The overall architecture description is followed by a section giving an overview of the implemented platform features. The features are described in detail within independent sub-sections describing each of the separate functional Areas of the platform.

The data platform operation report is intended to be accompanied with the BIGPROD Personal Data Protection Strategy report[3]. The accompanying deliverable will give a description of how the project has dealt with protecting personal data the data platform can have scraped during its operation. For a complete picture of the data platform operations, readers should consider this report and the BIGPROD Personal Data Protection Strategy report as a single report.

During the project the BIGPROD data platform has successfully web scraped data from approximately 96 000 European Union and United Kingdom firms. The web scraping process completed during the BIGPROD project has resulted in a database, part of which has been made publicly available[4] for interested stakeholders. A descriptive analysis of the database is also available from the project's repository[5] with a video tutorial describing the content[6]. The platform is currently operational and able to produce longitudinal data for the scraped companies.

---

[1] Pukelis, L. (2020) BIGPROD Platform Operation: Interim report. Deliverable 8, BIGPROD project. Available at https://www.bigprod.eu/wp-content/reporting/D8_Interim-Platform-Operation-Report-v3-22-12-2020.pdf

[2] Project Dataverse repository: https://dataverse.nl/dataverse/BIGPROD_Data_Sample

[3] Pukelis, L. (2022) BIGPROD Personal Data Protection Strategy. Deliverable 10, BIGPROD project. Available at https://dataverse.nl/dataverse/BIGPROD_Data_Sample

[4] Ashouri, Sajad; Hajikhani, Arash; Suominen, Arho; Jäger, Angela; Schubert, Torben; Cunningham, Scott; Van Beers, Cees; Türkeli, Serdar, 2022, "Replication Data for: BIGPROD Data Sample - Second Version", https://doi.org/10.34894/BS9XVR, DataverseNL, V1

[5] Ashouri, Sajad; Suominen, Arho; Hajikhani, Arash; Pukelis, Lukas; Schubert, Torben; Türkeli, Serdar; Van Beers, Cees; Cunningham, Scott, 2021, "Data in Brief: Indicators on firm level innovation activities from web scraped data", https://doi.org/10.34894/W3W2JQ, DataverseNL, V1

[6] Ashouri, Sajad; Suominen, Arho; Hajikhani, Arash; Pukelis, Lukas; Schubert, Torben; Türkeli, Serdar; Van Beers, Cees; Cunningham, Scott, 2021, "Data in Brief: Indicators on firm level innovation activities from web scraped data", https://doi.org/10.34894/W3W2JQ, DataverseNL, V1

BIGPROD Data Platform Architecture

We have outlined the following roles for the BIGPROD data platform:

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

The BIGPROD project utilizes Big Data and unstructured data from a variety of sources. During the course of the project, we have processed over 200 TB of company website data, and our main database, which stores aggregated and processed data only, has expanded to reach around 300 GB in size. The challenges in handling such large amounts of data were further compounded by the fact that project partners are based in many countries and must exchange data or make updates to the project database over significant distances. Anticipating these challenges, we have implemented a data platform consisting of the three distinct areas:

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

In this way, we can select a sub-system that best meets the demands of the area and connect them in such a manner that benefits the needs of the project most. Using this approach, we ensure that the data are clearly separated, i.e., that no sensitive data are exposed to the end user and that no proprietary data are 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's high-level architecture is shown in Figure 1.

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

- Company descriptive data;
- Indicators calculated from the data scraped from company websites;
- Indicators calculated from review and other websites;
- Indicators calculated by matching company records with other databases (PATSTAT and EUIPO);
- Results from econometric modelling;

- 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; rather, datasets approved for public use by the consortium will be transferred to another smaller database, which will be made publicly available.

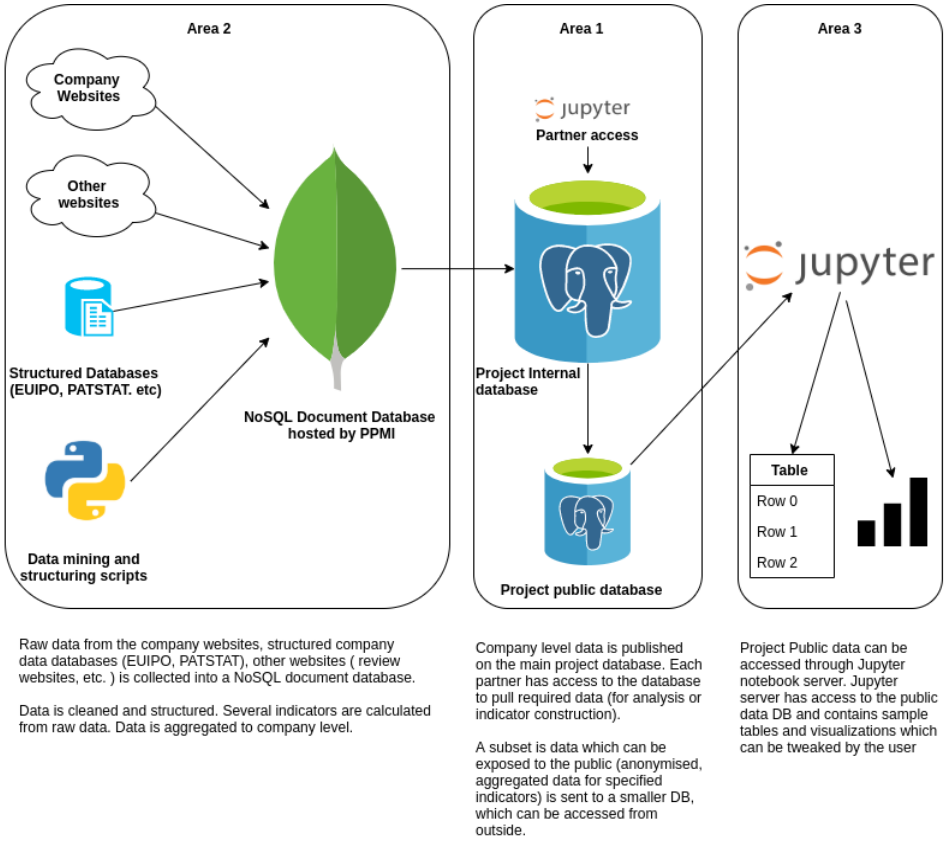

Figure 1. BIGPROD Platform High Level Architecture (Source: BIGPROD Project)

Area 2 houses the data scraped from the 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 the Jupyter Hub platform. The Jupyter Hub server fetches data from the smaller public database in Area 1 and allows users to perform various data analysis steps as well as data visualization.

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

Overview of the Implemented Platform Features by Area

Overall, we have managed to implement and run the data platform smoothly throughout the project without any major disturbances or accidents. The implementation of the data platform Areas 1 & 2 mostly took place during Y1, and work on Area 3 mostly occurred in Y2 of the project.

Area 1: Cloud-Hosted SQL Database

Description

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

- "Orbis" database;
- Indicators derived from unstructured company website/ review website data;
- Indicators derived from semi-structured data sources (PATSTAT, EUIPO);
- Indicators derived for and from econometric modelling.

We have chosen to store the data in this area in a SQL (Postgres) database. This format was chosen because the data in this area follows a uniform schema[7] and because SQL is a widely familiar and established database standard, making it easily accessible to the project partners. We have chosen PostgreSQL as the database server due to its well-developed functionality and performance. As with all the components of 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 based on several concerns:

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

[7] We expect to have the same variables for all the companies.

Implementation

The database is functional and can be accessed via an IP address[8]. As 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 changed after consulting the project partners and evaluating their preferences and needs. We have chosen the MS Azure from the existing cloud infrastructure providers because it has features that allow for ensuring that the data are stored and backed-up in the EU and that 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 it curates. This is done either independently or by passing on the prepared datasets to PPMI to handle the upload.

Because no sensitive data are collected during this project and the personal data that are collected are part of the public domain, the BIGPROD project does not require any highly advanced measures to ensure GDPR compliance; however, each project partner, especially PPMI, will remain GDPR-aware and GDPR-cautious at all the infrastructure design and set-up steps, and 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 200,000 companies in the database, with minor updates and revisions being carried out periodically.

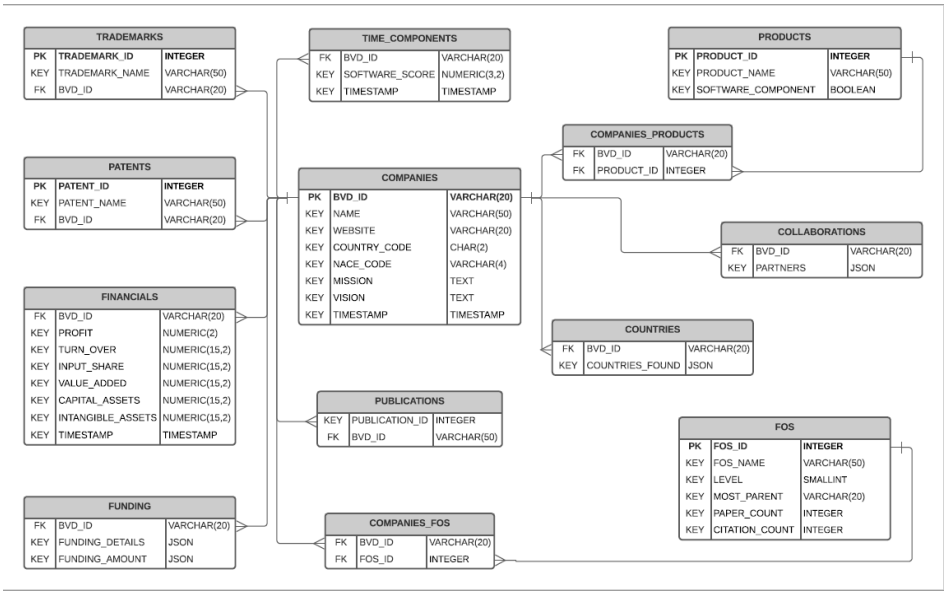

Figure 2. BIGPROD Platform Area I Database Schema

[8] http://40.113.156.93:1815

Area 2: No-SQL Database Hosted by PPMI

Area 2 is responsible for collecting the data from company websites and processing and enriching them through text-mining and text classification processes. Additionally, it facilitates the link 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.

Due to the need to store large amounts of text data, we chose to implement this database as a NoSQL document database. Among the many NoSQL options, we chose MongoDB due to its reliability and performance. Because the data collected in Area 2 generally are not to be shared with or accessed by the other consortium partners, we chose to host this database on PPMI premises and stay 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 data-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 are pooled together. Then, various data mining, information extraction, text classification, and text fragment matching algorithms are run to:

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

As a result, from the disaggregated dataset, where a unit of analysis is a single URL, we have built 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.

Implementation

During the project, we have managed to achieve all the goals we set for the Area of the Data Platform, with some aspects improved beyond our initial goals and estimates. The subsequent sections break down our progress by sphere and outline the main developments.

Web scraping

Prior to starting the BIGPROD project, PPMI had already developed a powerful web-scraper that can fully traverse company domains and extract text from various elements, including dynamic Java Scrip sections of the page. As such, the PPMI scraper could scrape 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 so and increased the capacities of the scraper from about 1M webpages per day initially to about 10 M at the end of the project. Having the capacity to process so much information allows us to carry out periodic updates for our data.

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 several aspect described in detail in the following.

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, such as “we are introducing a new NOUN” or “we are happy to announce a NOUN,” in the pages that 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 for the 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 to minimize false-positive instances. For collaborations, we also employ simple classification on the extracted entities. Using a list of keywords, we classify the entities into education sector entities and the rest.

Text Mining: Funding

In the same vein, we extract statements related to funding received by a company. 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[9] (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)[10]. We have taken this data and developed a tool to assign the FOS categories to any text input. The tool works by vectorizing 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 that are the most like its vector (have the 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. This is a highly original and novel approach that allows for linking company data to the wider body of scholarly literature.[11]

Text Classification: SDG

In the BIGPROD project, we also seek to identify companies that are active in the areas linked to the UN's SDGs. To this end, we employ OSDG[12], a text classification tool that PPMI has designed in the past to classify EU-funded research projects to SDGs and that has since been open-sourced and developed jointly with the United Nations Development Program (UNDP) and selected academic partners. The OSDG is essentially a mapping between FOS categories and the SDGs. By utilizing the FOS tags from the previous section, we can easily map texts to SDGs.[13]

Automation

During the project, we also focused on automating the data processing process, which includes data cleaning and text mining tasks. 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).

[9] Microsoft Academic: https://academic.microsoft.com/home (discontinued)
[10] Microsoft Academic FAQ : https://academic.microsoft.com/topics; Paper: https://arxiv.org/pdf/1805.12216.pdf
[11] Hajikhani, A., Pukelis, L., Suominen, A., Ashouri, S., Schubert, T., Notten, A. and Cunningham, S.W., 2022. Connecting firm's web scraped textual content to body of science: Utilizing Microsoft academic graph hierarchical topic modelling. *MethodsX*, *9*, p.101650.
[12] OSDG OSDG.ai
[13] 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*.

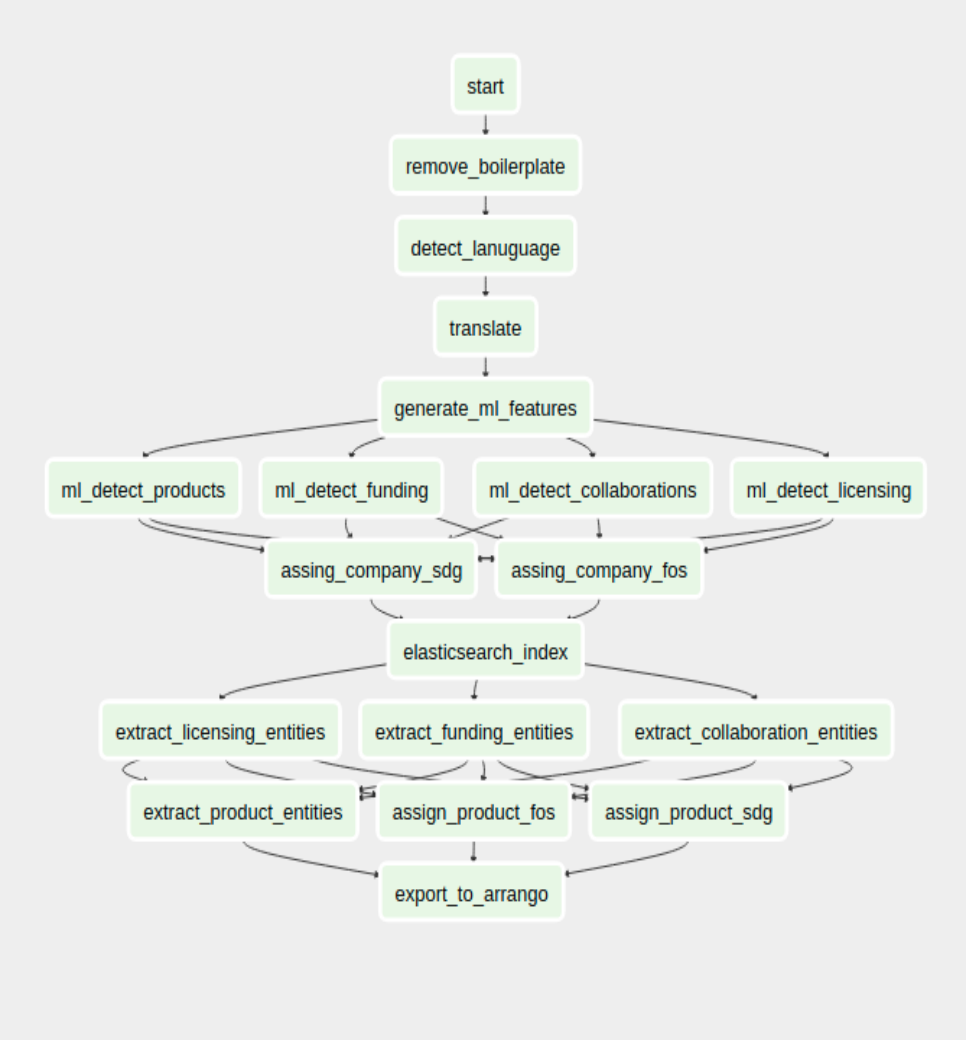

Figure 1. Automated Data Processing Pipeline for the BIGPROD project

Company matching

PATSTAT

To date, we have developed an algorithm that uses company meta-data to match companies in the BIGPROD sample to the PATSTAT database. To perform the matching, we use the company name, location information, and 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, the algorithm has matched roughly 20% of companies to entries in PATSTAT, which is consistent with the expectations.

Additional Workstreams

Furthermore, an additional workstream was the development of new indicators derived from company website text. First, we have developed a “Company Digitalization Score” indicator, which will showcase which share of company’s products have a software component. 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 this is done, we will calculate the indicator by counting which share of company products have a software component.

Additionally, we sought to develop indicators based on the contents of company “Mission and Vision” statements; however, we succeeded in identifying these statements with a high degree of accuracy only for a small number of companies (often larger, publicly traded companies), which did not allow us to perform a comprehensive analysis. We still maintain that these statements are a promising data source, but more research and experimentation are needed to produce quality insights from this data.

Area 3: Jupyter Notebooks

Overview

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

- Scholars from academia;
- Policy professionals;
- Analysts from research organizations.

In other words, we have foreseen 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 specialized area where advanced users will be able to perform queries of their own design for the public BIGPROD datasets. For an example of what Jupiter notebook looks like, see Figure 4.

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

Jupyter platform is excellent for this exercise because they allow for combining several different aspects:

- They facilitate storytelling–Jupyter notebooks work efficiently in combining text narratives, code commands, and data visualizations. In other words, they allow us to prepare report-like texts and illustrate them with interactive graphs and visualizations to facilitate data exploration by the user. All the data are 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 visualizations to user preferences;

- They facilitate an interactive exploration of the data—Jupyter notebooks also allow users to perform data exploration and analysis on their own. In other words, users can explore how the BIGPROD data can help them answer their own research questions, which were not asked by the project. In this way, Area 3 will allow users to co-create knowledge together with the BIGPROD consortium.

- Jupyter notebooks facilitate several programming languages—all the main languages used form data analyses, such as R, python, Scala, and Julia. This ensures that they can be used by many people from different backgrounds.

Originally, we envisioned launching a Jupyter server on the cloud on a separate VM than the main database but connected to the main database via an ETL pipeline; however, we had to deviate from this original idea. Early in the project, we made a decision to use company identifiers from the "Orbis" database as the key identifiers (primary keys) that connect different sections of our database together; however, during the project implementation, it came to light that these IDs are proprietary and thus cannot be shared without breaching our agreement with the Bureau van Dijk.

Consequentially, to open up the data, we had to eliminate all the proprietary data from our datasets and come up with new IDs that would link different sections of the dataset. We succeeded in this effort, and as a result, we were able to open up a smaller subset of our data than we originally envisioned. Though this meant downsizing our initial ambitions, this also had some positive effects. The smaller dataset size meant that we could now share it using various dissemination channels and not be limited to the project data platform. We have opted to store project related data, notebooks/scripts, and recordings and introductory presentations on the Dataverse platform.[14]

Additional Work to Supplement the Data Platform

During Y1, it became obvious that an 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.[15]

In addition to the abovementioned features, we are also developing a dedicated wiki, where we will 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 a way, if some aspect of project implementation requires additional clarification, they are all added in one place, so they can easily be looked up again (see Figure 5).

[14] BIGPROD on DataVerse NL https://dataverse.nl/dataverse/BIGPROD_Data_Sample
[15] Access point https://gitlab.com/ppmi-data/bigprod

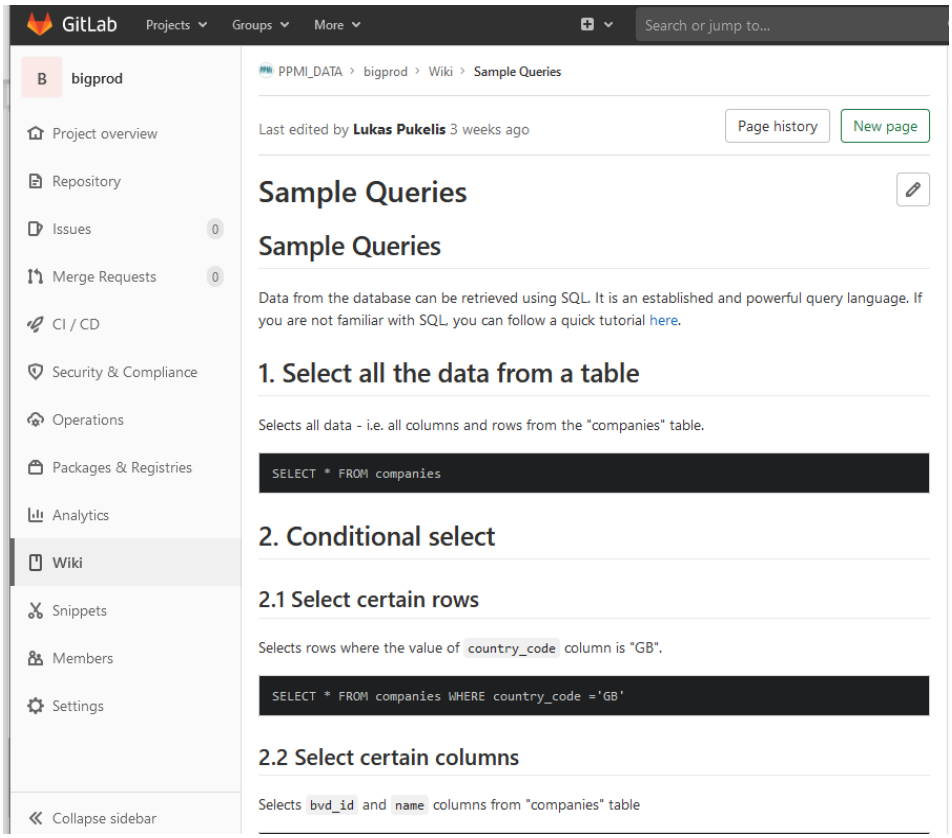

Figure 5. Illustration of GitLab wiki (Source: BIGPROD)

We have also developed specialized software to assign FOS-tags to text, thus enabling other users to replicate project results or to extend the methodologies we have developed to their own data. We placed these software tools in Docker containers so they would be portable between various operating systems and environments. These tools were placed inside “dockerhub”—the world’s largest platform to exchange dockerized software.

Conclusion

In 2,5 years the BIGPROD created and data platform that was able to create a novel dataset based on web scraped big data on approximately 96 000 companies in the European Union and the United Kingdom. The data created with the platform is unique and has offered novel vantage points to measure innovation[16]. This said, the for the platform to meet its potential additional development work is needed. From the stakeholder engagements during the project, it is clear additional work in validating the variables created is needed. In addition, there is a clear need for longitudinal data. This will require the web scraping process to be run multiple times to gain timeseries data on top of the baseline created during the BIGPROD project.

[16] For examples on using the data to measure innovation Suominen, Arho; Hajikani, Arash; Ashouri, Sajad; Cunningham, Scott, 2022, "BIGPROD: Write-up of three pilot cases", https://doi.org/10.34894/R99TRK, DataverseNL, V1

For more information, please contact:

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

About BIGPROD

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

The project partners in the project are the 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, and 7) Faculty of Technology, Policy and Management, Delft University of Technology.

www.bigprod.eu

BIG PROD