A customisable analytical platform for public statistics

Aldona Tomczyńska[1*], Emil Podwysocki[1] and Sylwia Ostrowska[1]
[1] The National Information Processing Institute, Poland
aldona.tomczynska@opi.org.pl, emil.podwysocki@opi.org.pl,
sylwia.ostrowska@opi.org.pl

Abstract

In this article, we showcase an analytical platform developed by the National Information Processing Institute (OPI PIB) and share the rationale behind its design. We also draw attention more broadly to the benefits of analytical platforms for the science and higher education sector. We demonstrate how analytical tools other than popular licensed business intelligence systems can be utilised by research organisations and decision-makers. We offer a brief comparison between licensed business intelligence tools and on-premises analytical platforms. By including an extensive overview of the technical architecture of our own software solution, we provide readers with a ready-to-use guide to the organisation of different layers of analytical platforms that can be integrated easily in any IT environment. Finally, we describe an implementation of the OPI PIB analytical platform for the purpose of the Genderaction project, which was financed as part of the Horizon 2020 programme of the European Union. By presenting data on women in science and higher education, the platform promotes progress towards the implementation of gender equality in research and innovation.

1 Introduction

Analytical platforms for data analysis and visualisation are becoming increasingly popular. We define an analytical platform as a unified solution that combines technologies to meet specific business needs across the end-to-end analytics life cycle. Such platforms also incorporate data storage, management, and preparation, as well as other data analytics processes.

As new digital data is generated, analytical platforms have assumed a deeper role in informing decision-making processes (Williamson 2018). In the context of research and education, analytical platforms are used to support research communities and policy makers by gathering research data or statistics, by introducing interoperability standards for benchmarking purposes (such as the evaluation

[*] https://orcid.org/0000-0002-0832-8081

of scientific achievements) locally, nationally, and internationally, and by making it quicker and more efficient for end-users to draw conclusions and make data-driven decisions.

The HESA analytical platform of the Higher Education Statistics Agency in the United Kingdom (HESA)[†] is notable, as it offers information on all aspects of the country's higher education landscape. Eurostat offers databases and dashboards[‡] that present public statistics on research, innovation, and tertiary education. In Poland, the RAD-on platform[§]—an information system comprising reports, analyses, and data on higher education and science (Michajłowicz et al. 2018; Protasiewicz et al. 2019; Protasiewicz et al. 2021)—offers advanced solutions. It incorporates a dashboard that presents statistics on higher education and research in Poland as interactive tables and graphs.

In this article, we showcase an analytical platform developed on-premises by the National Information Processing Institute (OPI PIB) and share our experiences in designing its architecture. We also discuss the benefits of such analytical platforms for science and higher education. We intend to demonstrate how analytical tools other than popular licensed business intelligence systems can be designed and used by research organisations and policy-makers.

The contributions of this research paper are as follows:

- A comparison between licensed business intelligence (BI) tools and on-premises analytical platforms.
- A case study of the analytical platform built by OPI PIB that focuses on its architecture and key functionalities.

First, in section 2, we explore the differences between BI tools and on-premises analytical platforms. In the *Technical description of an analytical platform* section 3, we describe the platform built by the team at OPI PIB in detail, demonstrating its scalability and suitability for various data processing challenges. Then, in the *Implementation example* section 4, we demonstrate the platform's functionalities as they were used in an international project on gender equality in science and innovation. Finally, the *Conclusion* section 5 offers a summary of cases in which analytical platforms are preferable to licensed BI software.

2 Differences between business intelligence tools and analytical platforms

Business intelligence platforms are applications that are used to analyse data and gain new insights into business. The term, *business intelligence* was introduced by Howard Dresner of the Gartner Group in 1989 (Power, 2007). According to the Gartner Group (Richardson et al. 2020), the leading software vendors in the field are currently Microsoft, Tableau, and Qlik. The potential advantages and disadvantages of different BI solutions are summarised in the Gartner Group report and in the Gartner Magic Quadrant (see Figure 1).

[†] Website: https://www.hesa.ac.uk.

[‡] Website: https://ec.europa.eu/eurostat/web/main/data.

[§] Website: https://radon.nauka.gov.pl/ (English version available).

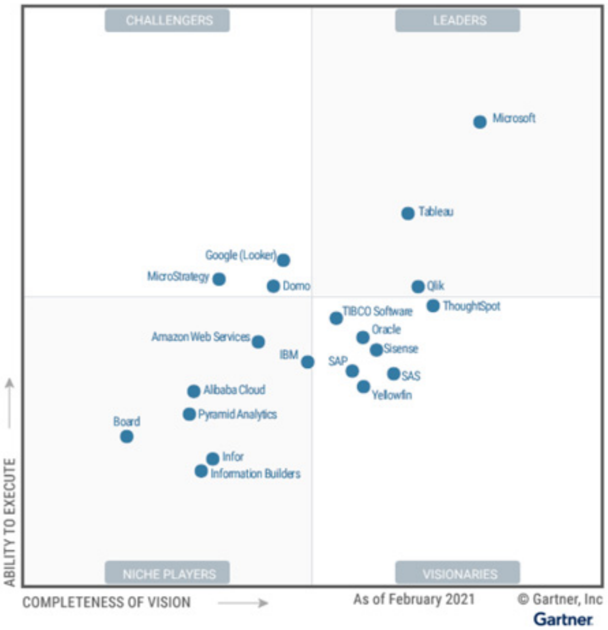


Figure 1. Gartner Magic Quadrant for Analytics and BI Platforms

Although almost all BI tools offer data extraction, storage, and analysis, they are often differentiated by their data visualisation capabilities. Their main customers are large enterprises that use BI reporting functionalities to analyse their own businesses. The majority of user of such tool are employees of the those organisations.

From the perspective of public institutions, the primary disadvantages of BI software are prohibitive licensing costs and the inaccessibility of its source code. Licensing costs serve to make data less accessible to the public, which works against the open government data policies; the lack of access to source code creates difficulties in tailoring services to the specific needs of end-users.

BI tools are typically priced on one of two bases: per user and as licenses for hardware (CPU licenses). According to the official price lists of the top BI software providers, the average environment cost is two million dollars; however, the total cost of ownership could grow by up to 20% with each year of use. Similarly, the costs of cloud BI solutions are high and difficult to forecast.

OPI PIB, a research institute that delivers IT services to other institutions in the science and higher education sector, has developed an alternative to BI software. The institute's analytical platform can operate free of licensing costs and on any website. The system has been used in the RAD-on platform to generate reports on the higher education and science sector in Poland, and in the Genderaction Data Dashboard, which presents statistics on gender equality in Europe.

Initially, the software integrated only static files; its creators were later able to integrate it with a corporate data warehouse (compare Figure 2 and Figure 3).

1. Basic integration (static files only)

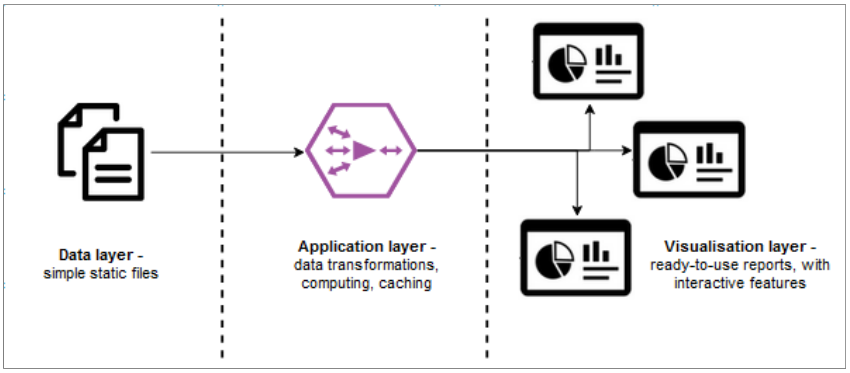

Data layer - simple static files
Application layer - data transformations, computing, caching
Visualisation layer - ready-to-use reports, with interactive features

Figure 2. Basic integration

2. Enterprise integration (full integration with corporate data warehouse)

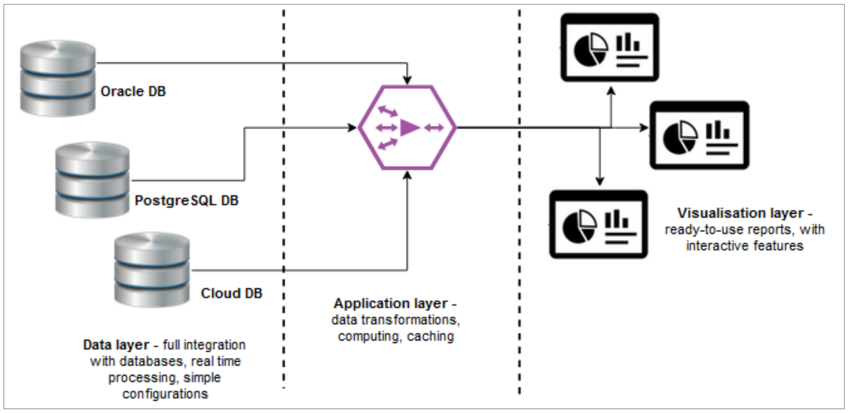

Oracle DB
PostgreSQL DB
Cloud DB
Data layer - full integration with databases, real time processing, simple configurations
Application layer - data transformations, computing, caching
Visualisation layer - ready-to-use reports, with interactive features

Figure 3. Enterprise integration

3 A technical description of the analytical platform

In this section, we focus on the backend of the OPI PIB analytical platform. Its core comprises an HTML page or an output file that contains a graphical representation of data that is edited by data analysts to produce individual reports. The data presented in a single report can be sourced from relational databases, CSV, or XLSX files. These reports can be then displayed on public or private web pages.

To control the structure of the reports' dependencies, the architecture of the analytical platform has been divided into four abstract layers: connectors, cache, query, and reports. The relationships between the layers are presented in Figure 4.

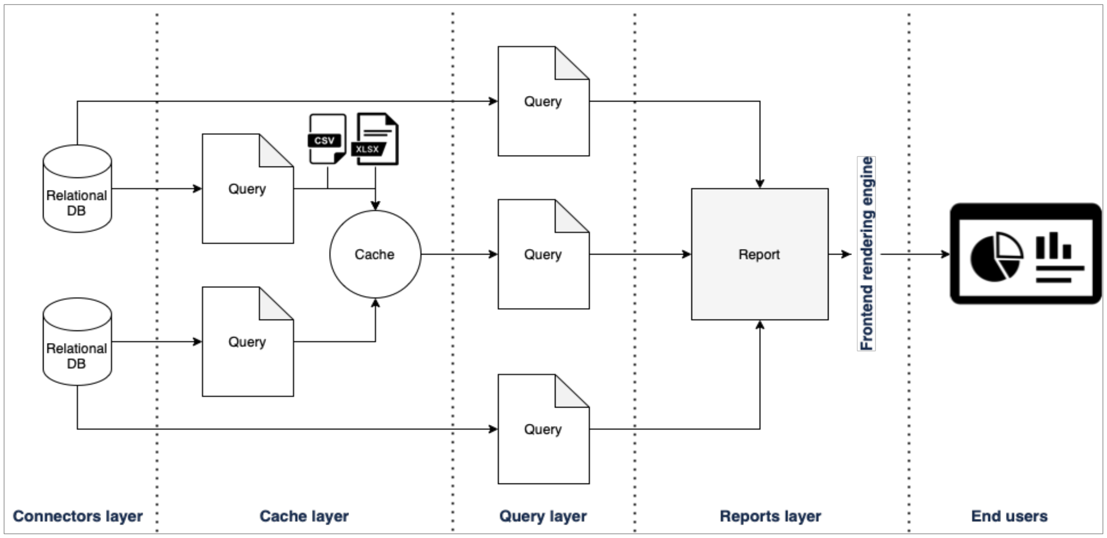


Figure 4. Analytical platform architecture

3.1 Architecture layers

Each layer is responsible for managing one type of object. The responsibilities of the individual layers are as follows:

1. Connectors layer: enables the definition and management of connections to external data sources. In the case of a relational database, the connector is the pool of connections to any database, such as Oracle or MySQL.
2. Cache layer: the analytical platform incorporates a local H2 database that can store copies of data from external sources. This provides faster access to data, as it is unnecessary for such systems to continuously poll the source servers. The use of the cache is optional, as the system also enables the execution of reports and queries directly from the source databases.
3. Query layer: enables the definition and execution of queries. A query object comprises a connector object reference, an SQL query, and an optional set of named parameters. If the query is parameterised, the values of the parameters defined should be passed when running the query. This enables the creation of filters to be used in each unique report. Any request to execute a query should also contain one or more definitions of so-called 'outputs', which instruct the system how to process a query result. The most common outputs involve returning data in an HTTP response, saving data to a file, and returning query statistics. Each output allows direct interaction with the query layer. The system also uses the transfer of data to the next layer (the report that uses a query's results) internally.
4. Reports layer: contains the report definitions. A report specification is a complex object for which a set is defined that comprises a list of queries to be used in the report, a list of

report parameters that coincides with the parameters of previously-declared queries, and a report section list. In the process of generating the report, specification objects are transformed into result objects that contain data to be presented to the end-user. A key element of this layer is the frontend rendering engine, which converts result objects into ready pages that present the corresponding report to end-users.

Usually, in a process of preparing a report, a data analyst is responsible for providing an SQL query or xlsx file containing relevant data. As a result, a cache is created. Another query of the cached data transforms a simple data table into a dynamic report with filters.

3.2 Reports layer

A report is an object that integrates the elements defined in the previous application layers. Interdependencies are created between these objects and those defined inside the report, such as parameters or sections. To understand how the reports layer works, it is necessary to understand these interdependencies and their implications; changing one element of a report often requires that other elements that depend on it be changed accordingly. Broadly, the report specification comprises the following parts:

1. List of queries: the data that each report presents must be retrieved from a defined data source. This part of the report declares all queries from the query layer that the report uses to present the data. The result of running a single query can simultaneously be consumed by various report elements—for example, in the presentation of charts or tables, or in complete dictionaries of report parameters.
2. List of report parameters: queries can be parameterised. This also applies to reports that are based on the execution of such queries. In this part, a global list of all parameters that the report uses is declared. Such a list should be the sum of all parameters of the queries declared.
3. Report section list: this is visible to the end-user who views a report. Each section of the report has its own specification (i.e. a description of how it is to be presented and what data it should use) and the resulting representation (i.e. a generated set of information on the basis of which the section can be presented to the end-user). For example, a chart section specification might include information that the chart contains two series, and data for them should be retrieved from queries A and B. The resulting representation will return a series list with the data already entered, without information on the queries from which this data has come, as it is irrelevant to the end-user.

In order to create a ready-to-publish report, a data analyst uses this layer to add different functionalities of the report, such as filters or charts. Sections that can be used to present data are described below.

3.3 Types of section

Data analysts decide how to structure each report viewed by end-users. As different types of section perform different functions, each section contains an individual set of fields that control its behaviour and presentation. The most important sections are as follows:

- Text section: accepts HTML code, so any text formatting is possible.

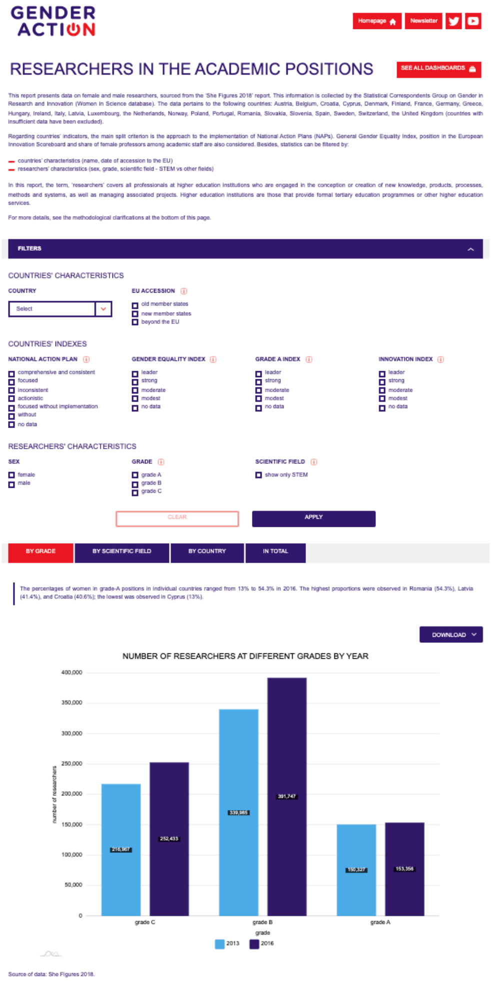
- Filters section: used to interact with the report by filtering its parameters. The section specification contains information on what types of component should be displayed (e.g. multiple checkboxes, text boxes, or selections) and to what report parameters they connect.
- Chart section: used to present various charts or maps that contain data from the queries declared in the queries part of a report. A chart can have one or more data series and each series can come from a different query.
- Table section: used to present the data from a query in tabular form.
- Summary tiles section: displays tiles that present numeric values that summarise a report.

To summarise, the process of report creation is fairly simple. A data analyst uses SQL queries or even simple csv files to prepare datasets. The most important part is the parametrization of queries which enables users to interact with data in each final report. A data analyst can build many different reports using sections of the reports layer. As a result, a ready-to-publish dashboard can be implemented on the website.

4 Implementation example

The OPI PIB software has been used to develop reports on the higher education and science sector in Poland on the RAD-on platform and to implement a data dashboard that presents international statistics on women in science, higher education, and the research and development sector on the Genderaction project website. Genderaction is financed as part of the European Union's Horizon 2020 programme for promoting progress towards implementation of gender equality in research and innovation.

The data dashboard comprises individual reports that present data retrieved from official Eurostat databases and XLSX files with that of the *She Figures* report (European Commission 2021). Figure 5 depicts users' view of the data dashboard.



RESEARCHERS IN THE ACADEMIC POSITIONS

This report presents data on female and male researchers, sourced from the 'She Figures 2018' report. This information is collected by the Statistical Correspondents Group on Gender in Research and Innovation (Women in Science database). The data pertains to the following countries: Austria, Belgium, Croatia, Cyprus, Denmark, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Luxembourg, the Netherlands, Norway, Poland, Portugal, Romania, Slovakia, Slovenia, Spain, Sweden, Switzerland, the United Kingdom (countries with insufficient data have been excluded).

Regarding countries' indicators, the main split criterion is the approach to the implementation of National Action Plans (NAPs). General Gender Equality Index, position in the European Innovation Scoreboard and share of female professors among academic staff are also considered. Besides, statistics can be filtered by:

- countries' characteristics (name, date of accession to the EU)
- researchers' characteristics (sex, grade, scientific field - STEM vs other fields)

In this report, the term, 'researchers' covers all professionals at higher education institutions who are engaged in the conception or creation of new knowledge, products, processes, methods and systems, as well as managing associated projects. Higher education institutions are those that provide formal tertiary education programmes or other higher education services.

For more details, see the methodological clarifications at the bottom of this page.

COUNTRIES' CHARACTERISTICS

COUNTRY: Select

EU ACCESSION
□ old member states
□ new member states
□ beyond the EU

COUNTRIES' INDEXES

NATIONAL ACTION PLAN
□ comprehensive and consistent
□ focused
□ inconsistent
□ actionistic
□ focused without implementation
□ without
□ no data

GENDER EQUALITY INDEX
□ leader
□ strong
□ moderate
□ modest
□ no data

GRADE A INDEX
□ leader
□ strong
□ moderate
□ modest
□ no data

INNOVATION INDEX
□ leader
□ strong
□ moderate
□ modest
□ no data

RESEARCHERS' CHARACTERISTICS

SEX
□ female
□ male

GRADE
□ grade A
□ grade B
□ grade C

SCIENTIFIC FIELD
□ show only STEM

The percentages of women in grade-A positions in individual countries ranged from 13% to 54.3% in 2016. The highest proportions were observed in Romania (54.3%), Latvia (41.4%), and Croatia (40.6%); the lowest was observed in Cyprus (13%).

Source of data: She Figures 2018.

Figure 5. An example of the Genderaction data dashboard

To provide insight into chronological trends in the data, it was necessary that the OPI software be integrated with another layer in the data flow architecture. After retrieving data from Eurostat's rest API, the software's creators analyse datasets in terms of their completeness. Due to breaks in time series for individual countries, it is necessary to implement estimation algorithms using *R* or Python to transform the initial datasets. The OPI tool enables significantly more flexibility in this regard than many BI tools on the market do.

Each interactive report contains subsections (views) that present different aspects of the data in the form of charts or maps. Users can interact with datasets using filters, as well as downloading visualisations from each view (as PNG files) or sets of data as CSV or XLSX files. Each report contains a brief introduction and methodological clarifications. The development team applied the storytelling approach to data: the resulting data dashboard also offers insights into statistics and their meaning.

Genderaction's openness to the public and accessibility via the project website, as well as the variety of functionalities it incorporates, makes it a truly multipurpose tool. It can be used by a host of stakeholders, including national authorities and policy-makers, researchers, and experts interested in exploring statistics on women in science. The data dashboard is linked to the Genderaction project website. It can be found at: https://genderaction-data-dashboard.opi.org.pl

5 Conclusions

In this article, we have presented an alternative to standard BI tools that can be built and used by individual organisations.

Although the importance of BI tools is increasing, they may, in some cases, be substituted for analytical platforms built by IT teams. It is our opinion that analytical platforms are preferable to BI tools when at least one of the factors outlined below are present.

First, analytical platforms thrive when the potential number of end-users is large and difficult to forecast. This might be the case when a dashboard with open data must be displayed on a public website. Licensed commercial BI tools are much more costly when their user numbers are unlimited.

Second, analytical platforms are preferable when it is necessary to adjust to the developing needs of end-users. Each type of BI software contains a unique set of features, and in most cases, that list is closed. Analytical platforms can be flexible and extensible to accommodate changes easily.

Third, analytical platforms are much more useful than BI tools when the visual element of a dashboard is crucial—for example, when it must correspond to the design of a larger website. Most BI tools offer limited layouts, whereas analytical platforms can be easier to integrate with existing designs.

The most significant obstacles for the development of analytical platforms are IT teams' availability and their experience of building analytical tools. In the case of the higher education and science sector, this can be overcome when a single central agency is established and made responsible for providing IT services to other research organisations and governmental partners.

References

European Commission, Directorate-General for Research and Innovation (2021), *She figures 2021: gender in research and innovation: statistics and indicators*. Retrieved February 7, 2022, from: https://data.europa.eu/doi/10.2777/06090.

Gartner website (2022). *Magic Quadrant for Analytics and Business Intelligence Platforms*. Retrieved February 7, 2022, from: https://www.gartner.com/doc/reprints?id=1-1YOXON7Q&ct=200330&st=sb.

Michajłowicz, M., Niemczyk, M., Protasiewicz, J., & Mroczkowska, K. (2018). POL-on: The Information System of Science and Higher Education in Poland. EUNIS 2018 Congress Book of Proceedings. Retrieved May 19, 2021, from https://drive.google.com/file/d/1Z-n7ZCJ1FS2r_nF5nCF8-iI4-tQCZbqE/view.

Power, D. J. (2007) *A Brief History of Decision Support Systems*, DSSResources.COM. Retrieved February 7, 2022, from: http://dssresources.com/history/dsshistory.html.

Protasiewicz, J., Podwysocki, E., Ostrowska, S. & Tomczyńska, A. (2021). Open access to data on higher education and science: A case study of the RAD-on platform in Poland. *Proceedings of the European University Information Systems Conference 2021*, *78,* 9-21.

Protasiewicz, J., Rosiak, S., Kucharska, I., Podwysocki, E., Niemczyk, M., Błaszczyk, Ł., & Michajłowicz, M. (2019). RAD-on: An integrated System of Services for Science - Online Elections for the Council of Scientific Excellence in Poland. EUNIS 2019 Congress Book of Proceedings. Retrieved May 19, 2021, from https://www.eunis.org/download/2019/EUNIS_proceedings_2019.pdf.

Richardson, J., Sallam, R., Schlegel, K., Kronz, A., Sun, J. (2020). Magic Quadrant for Analytics and Business Intelligence Platforms. Retrieved February 7, 2022, from: https://bpmtraining.net/wp-content/uploads/2020/10/gartner-magic-quadrant-for-analytics-and-business-intelligence-platforms-feb-2020.pdf.

Williamson, B. (2018). The hidden architecture of higher education: building a big data infrastructure for the 'smarter university'. International Journal of Educational Technology in Higher Education, 15, Article 12. https://doi.org/10.1186/s41239-018-0094-1.

Author biographies

Aldona Tomczyńska (PhD): is a doctor of philosophy in international political economy. She was awarded her doctorate degree with distinction at the University of Warsaw in 2017. She has worked as an assistant professor and data science team leader at the National Information Processing Institute for more than ten years. She conducts research in economy, innovation, sociology, and data science. She has coordinated multiple scientific projects, as well as participating in evaluation studies and research and support actions within the Horizon 2020 programme.

Email: aldona.tomczynska@opi.org.pl

LinkedIn: http://linkedin.com/in/aldona-tomczyńska-phd-8490b718a

Emil Podwysocki (MSc) obtained a master's degree in telecommunications systems at the Technical University of Lodz in 2009. He has over ten years' professional experience related to ETL/ELT, data warehouses, and business intelligence. His areas of interest include Oracle technology, big data, business intelligence, and data visualisation. Currently, he serves as the head of the Laboratory of Databases and Business Analytical Systems at the National Information Processing Institute.

Email: emil.podwysocki@opi.org.pl

LinkedIn: www.linkedin.com/in/emil-podwysocki

Sylwia Ostrowska (MSc): is a business analyst and project manager at the National Information Processing Institute. She obtained a master's degree in Management at the University of Warsaw in 2012 and completed postgraduate studies in information system design at Warsaw University of Technology in 2018. She has professional experience in business analysis and project management in agile environments. Currently, she is the project manager of RAD-on. She holds PRINCE2 Foundation and Professional Scrum Product Owner certificates.

Email: sylwia.ostrowska@opi.org.pl

LinkedIn: https://www.linkedin.com/in/sylwia-ostrowska-06749b23