EPiC Series in Computing

Volume 95, 2023, Pages 79–88

Proceedings of European University Information Systems Congress 2023

Data Governance at the National Information Processing Institute in Poland

Maria Bylina[1*], Emil Podwysocki[1], and Marek Michajłowicz[1†]
[1] The National Information Processing Institute, Poland
maria.bylina@opi.org.pl, emil.podwysocki@opi.org.pl,
marek.michalowicz@opi.org.pl

Abstract

The collection and exchange of data in distributed systems are challenges for many institutions that deliver IT systems. These challenges are rooted not only in the technical layer, but above all, in proper management and communication. In Poland, data collection on science and higher education is conducted in large centralised repositories that are part of the POL-on ecosystem. These systems were developed during different periods by various teams and subcontractors. This has created a set of uniformed expectations regarding the integrity and interoperability of information processed by them. Good practices related to the data governance framework were used for the proper implementation of this goal. This article presents our conclusions from experience in the implementation of this approach at the National Information Processing Institute (OPI PIB) in Poland. It describes not only the outcomes and final results, but also the benefits of developing similar solutions in science and higher education.

1 Introduction

This article focuses on the implementation of data governance at the National Information Processing Institute (OPI PIB) in Poland. It describes the key aspects of the process and how we have adapted the data governance framework defined in DAMA-DMBOK to OPI PIB, the only national research institute in Poland that is responsible for collecting data about higher education. The institute's IT systems, including POL-on, PBN, SEDN, and RAD-on, collect data and support the Polish Ministry of Education and Science and other institutions in the decision-making processes (Michajłowicz et al. 2018; Protasiewicz et al. 2019; Michajłowicz et al. 2022).

At present, every organisation's most valuable assets are people and data. Every day, OPI PIB processes huge amounts of data, and the institute understands its importance. Every day it must ensure

* https://orcid.org/0000-0003-3551-4709
† https://orcid.org/0000-0003-2096-5005

the security of its data, extract as much information as possible from that data, and continue to build a data-driven organisational culture. The institute also needs well-organised data management processes to simplify its work, including that in the development of intelligent IT services.

The information systems developed by OPI PIB use a variety of data storage technologies. The dominant one is a relational database, which is used frequently to store JSON or XML type documents.

The following data sources have been integrated into the data warehouse:

1. Relational databases:
    a. Oracle
    b. PostgreSQL
    c. MySQL
    d. MSSQL
2. MongoDB object-oriented database
3. Apache Kafka message broker
4. REST API's
5. Static files and documents
    a. XML
    b. XLSX, CSV, TXT files, etc.

OPI PIB holds approximately 12 TB of data. The institute provides answers to questions from journalists, agencies, and other stakeholders, as well as clarifying issues related to the definition of data, and the possibility of reprocessing it or making it available to the public.

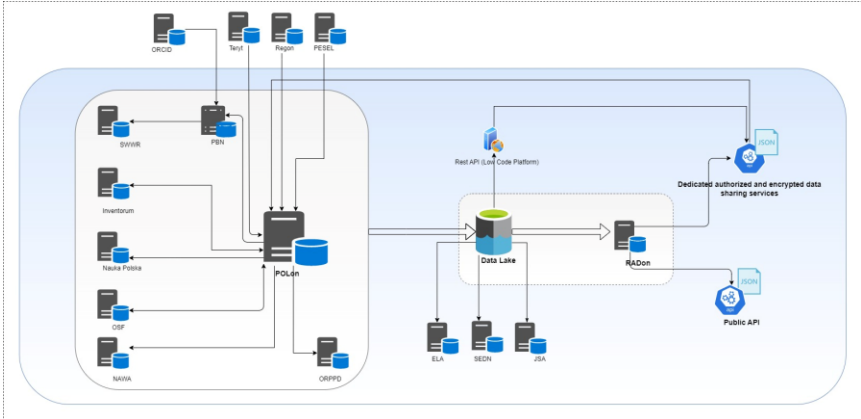

Figure 1: Data flow between systems at OPI PIB

The data collected by OPI PIB is scattered across different systems, which sometimes causes redundancy. Data is provided by external users, and the institute cannot always ensure its highest quality. To improve the quality of data, as well as its proper processing and sharing, a data governance policy should be implemented. The institute emphasises data inventory and consistent understanding at the business level.

The potential benefits of implementing data governance in an organisation are as follows:

- increased awareness of the value of data processed in the organisation and confidence in that data; knowledgeable decision-making based on very high-quality data
- increased quality of data in IT systems
- increased security of sensitive data
- conscious use of data to achieve the organisation's business goals
- increased potential for ongoing research and analysis through greater awareness of the roles and uses of data processed in source systems
- increased interoperability of data by its linking and exchange with reference systems and sources.

Figure 2: The DAMA-DMBOK2 Data Management Framework

Promoting the responsible use of high-quality data and making it available to the academic community and beyond is equally important. Implementing data governance can create opportunities for better decision-making locally and nationally, and can encourage other communities, including universities, to implement the process in their environments.

To achieve these benefits, we spread awareness of data governance, conduct thematic workshops, identify individuals who possess business knowledge about data, define data areas, analyse the tools available to enable data governance, as well as attempting to build a data business model with the preparation of a business dictionary and metrics for measuring data quality.

2 Data Governance

Data governance is a series of processes, specific roles, internal policies, and standards. Standardised metrics can be used to measure the success rate of the implementation of the data governance framework in organisations. This enables organisations to achieve their goals and to ensure the most efficient use of information. It establishes processes and responsibilities that ensure the quality and security of the data used across an organisation. Data governance identifies which methods and actions can be used by defined roles and what relevant data can be accessed.

The implementation of any new initiative in an organisation, such as data governance, is associated with the establishment of an appropriate organisational model that is tailored to the needs of the organisation. In the case of data governance, it is advisable that new roles, such as data owner, data steward, chief data officer, data architect, or data quality analyst be created. Due to the specific nature of the systems produced, domain knowledge about the data is accumulated within teams that develop systems and teams or individuals involved in the creation of analytical reports. The high degree of information dispersion in the systems does not easily facilitate data analysis and reasoning. Extracting cross-cutting information requires the involvement of many individuals. When analysing the personnel structure of OPI PIB, the data management model described in the literature had to be simplified. As a scientific institute, OPI PIB has a very flat organisational structure, with only one head, three deputy heads, and several laboratory managers. Ultimately, the institute decided to limit its data governance team to three basic roles: data owner, data quality manager, and chief data officer. In the future, if necessary, the institute will expand the team to include more roles.

The goal of data governance is to enable an organisation to manage data as an asset. Data governance utilises the principles, policy, processes, framework, metrics, and oversight to manage data as an asset and to guide data management activities at all levels (Earley S. et al. 2017).

Data governance refers to the full lifecycle of data (collection, storage, sharing, archiving, and deletion) in an organisation. It is a key component of data management that binds together ten other disciplines, including data quality, master data management, data architecture or data warehousing, and BI.

2.1 Data governance in higher education and other sectors

Although data governance exists across many industries, its interpretation varies. In Poland, it is used primarily in the banking sector. In science and higher education in Poland, only OPI PIB can undertake such tasks. Within the European Union, the CERIF (Jörg, 2010) and HERM models (Nauwerck et al., 2022) have been implemented; however, they do not cover the entire thematic range of data collected at OPI PIB.

A review of the literature indicates that some universities in the United States have created new data governance units using different labels (e.g. data governance, institutional research, or data management/analytics), while some universities have extended IT governance or information governance to data governance. (Jim & Chang, 2018). Such institutional data governance programmes (such as that of the University of Toronto in Canada) have been implemented at particular universities, but not centrally.

Work is underway in Poland on the development of an information architecture that aims to simplify the country's IT by indicating reference systems and datasets. One of the reference systems is the Integrated System of Information on Science and Higher Education in Poland (POL-on)[‡]. To fulfil this task, the data model developed for the purposes of data governance should provide details of the model developed at the national level. It should be emphasised, however, that the type and scope of data collected by OPI PIB systems are regulated by law. Some of them may be publicly available, and some may only be accessed by a small group of institutions and OPI PIB's programme should account for this.

Data governance is a cornerstone of the state policy for IT system architecture; the data governance framework implemented at OPI PIB should be consistent with its assumptions.

3 Implementing Data Governance at OPI PIB

3.1 The data governance maturity model and business case

The first exercise involved assessing OPI PIB's awareness and maturity in data management. According to the Gartner Data Governance maturity model, we evaluated the institute's awareness between levels 1 and 2. This indicated that OPI PIB needed to work on implementing a data governance framework.

[‡] More information on this subject can be found on the dedicated national portal about the Model of the Information Architecture of the State (https://www.gov.pl/web/ia/model-aip).

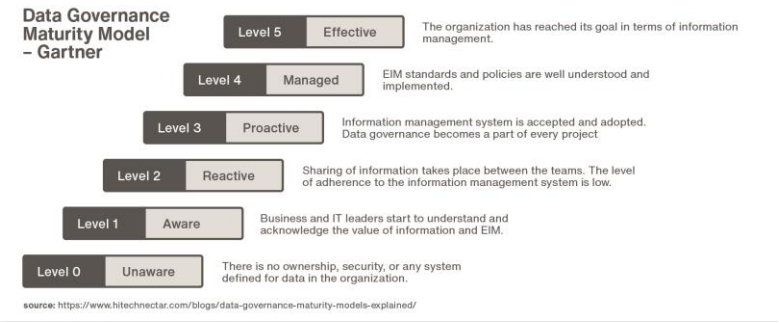

Figure 3: The Gartner Data Governance maturity model

We first prepared a business case to convince the head of OPI PIB that such a programme was necessary. Establishing the business case by identifying critical business drivers to justify the effort and investment associated with data governance (Earley et al. 2017) was crucial for securing approval from the institute's head. Data governance is a sizeable investment for any organisation, both financially and personally. For that reason, it is extremely important to have the full support of an organisation's management team.

3.2 Building the team

When we were certain that the management of OPI PIB understood and supported the implementation of a data governance policy fully, we started building a team. Data governance is created by people for people, and people are the most valuable part of the process. According to DAMA International, the data owner should be placed at a high level in the organisational structure e.g. a department manager. In the flat structure of OPI PIB, this would be inadvisable; instead, we opted to nominate a product owner as the data owner. We believe that the product owner possesses sufficient decision-making power and extensive knowledge about the domain systems in OPI-PIB's IT infrastructure. Under the previous model of the institute's operation, the functions that resulted from the agile SCRUM methodology were key. The role of the product owner chiefly incorporated skills related to decision-making and shaping the development of IT systems in the data management layer. We recognised that this role would be crucial in the further implementation of data governance. The first decision was the nomination of the chief data officer (CDO), who is responsible for the entire project. Next, we selected twelve data owners (DOs) and one data quality manager (DQM). The key challenge was to establish an interdepartmental team and delegate management powers to the CDO. The main difficulty in selection of the data owners was identifying individuals who wished to assume responsibility for the entire area and create a policy for the entire organisation.

3.3 Data inventory

The first visible success of the implementation of data governance at OPI PIB was the creation of a data inventory. We analysed our business processes, data flow, and their transformation, and reminded ourselves that the main goal was to provide the highest quality data for business users. We divided our data into business areas and sub-areas.

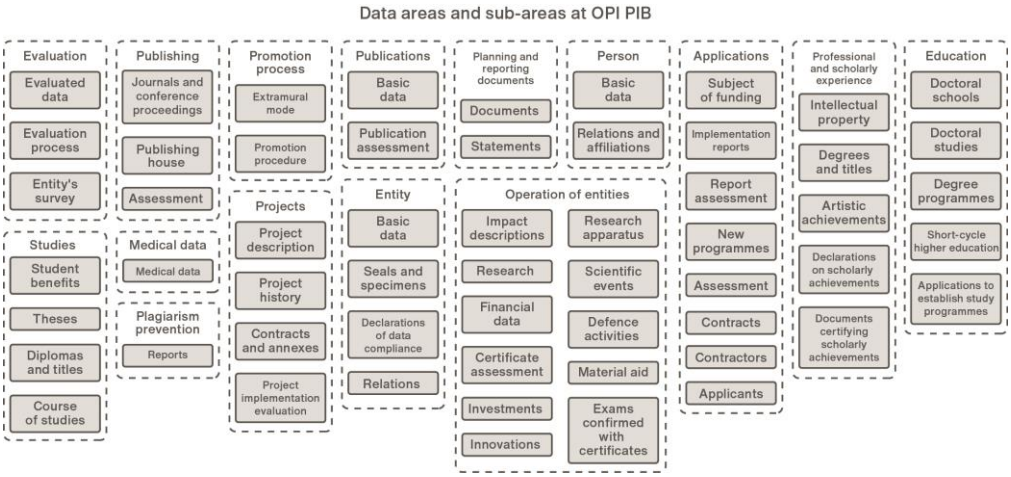

Figure 4: Business areas and sub-areas at OPI PIB

The following methods of grouping data according to its types have been adopted:

- category information: data used to classify and assign types, e.g. entities classified by type
- information about resources: basic profiles of resources needed to conduct operational processes, e.g. publications, scientific achievements, or teaching activities
- information about business events: data created during operational processes, e.g. commencing studies, commencing employment, or obtaining a scientific degree

For each of the business sub-areas, an information scope was defined, which considers the characteristics of the data and contains all information objects, then creates entities of the logical model that reflect not only the data structure, but also its connections and relationships.

3.4 Enterprise data model

An enterprise data model is a type of integration model that covers all (in practice, most) of the data of an enterprise. An enterprise architecture may include enterprise-wide data models that are also conceptual, logical, or physical data models. (West M. 2011). The approach of OPI PIB attempts to correlate the institute's enterprise architecture with an enterprise data model to achieve a full consistent information architecture. The institute's key goal is to create an information architecture that is highly compatibile with the Polish National Information Architecture (https://www.gov.pl/web/krmc/aip). Below is a fragment of OPI PIB's model related to the promotion process.

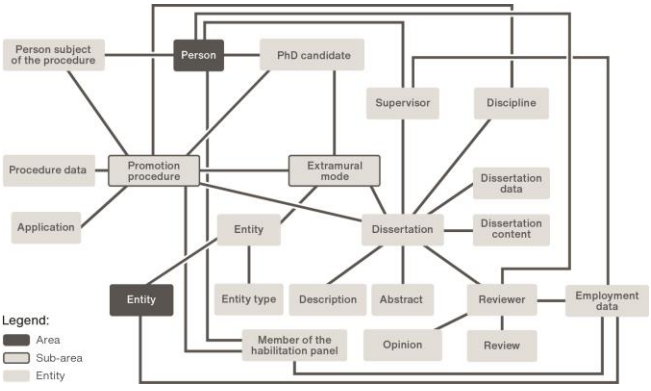

Figure 5: Business enterprise model

Additionally, for each business domain, we created a metric that contains essential information, such as the one below:

AREA	HIGHER EDUCATION
AREA DESCRIPTION	Data related to universities, research institutes, and educational and higher education institutions
INFORMATION SCOPE	The area covers identification, registration and basic data, addresses of the registered office and correspondence addresses, contact details, and details of the head of the unit.
DOMAIN	POL-on, OSF, NAWA
GROUPING LEVEL	Category
STAKEHOLDERS	Data Owner, Head of …

Table 1: Area metric

3.5 Proof of concept with data catalog

After several months working with data governance, we began to search for an IT tool that could help OPI PIB in data management processes. We conducted research and compared five of the most prominent tools:

- Oracle Data Catalog (ODC) - https://docs.oracle.com/en-us/iaas/data-catalog
- AtaccamaOne (ATA) - https://www.ataccama.com
- Atlan (ATL) - https://docs.atlan.com
- Collibra (COL) - https://marketplace.collibra.com
- Talend (TAL) - https://help.talend.com

Available (selected) data connectors:

	ODC	ATA	ATL	COL	TAL
ORACLE DB	✓	✓	✓	✓	✓
MYSQL	✓	✓	✓	✓	✓
POSTGRESQL	✓	✓	✓	✓	✓
MICROSOFT SQL SERVER	✓	✓		✓	✓
MONGO DB		✓		✓	✓
CSV	✓	✓		✓	✓
XLSX	✓	✓		✓	✓
XML	✓	✓		✓	✓
JSON	✓	✓		✓	
APACHE KAFKA	✓	✓		✓	✓
HIVE	✓	✓	✓	✓	✓
ELASTIC		✓			✓
JIRA			✓		
AZURE	✓	✓	✓	✓	✓

Table 2: Available data connectors in data catalog tools

Available (selected) functionalites:

	ODC	ATA	ATL	COL	TAL
DATA MODELLING CONCEPTUALISATION	✓	✓	✓	✓	✓
DICTIONARIES	✓	✓	✓	✓	✓
CLASSIFICATIONS	✓	✓	✓	✓	✓
DATA QUALITY		✓		✓	✓
DATA LINEAGE		✓	✓	✓	✓
EMBEDDED AI (CLASSIFICATIONS, PROFILING, RELATIONS...)	✓	✓		✓	✓

Table 3: Available funcionalities in data catalog tools

Although the comparison above contains empty cells, all tools available on the market are similar. When functionalities are unavailable 'out of the box', we can find add-ons from the same vendor (although they require additional payment). For instance, Oracle and Atlan have separate tools related to data quality and data lineage. No data catalog tools have data modelling functionalities included; separate tools are necessary. Ultimately, we decided to use Oracle Data Catalog, because it is included in Oracle Cloud Infrastracture and does not require an additional fee.

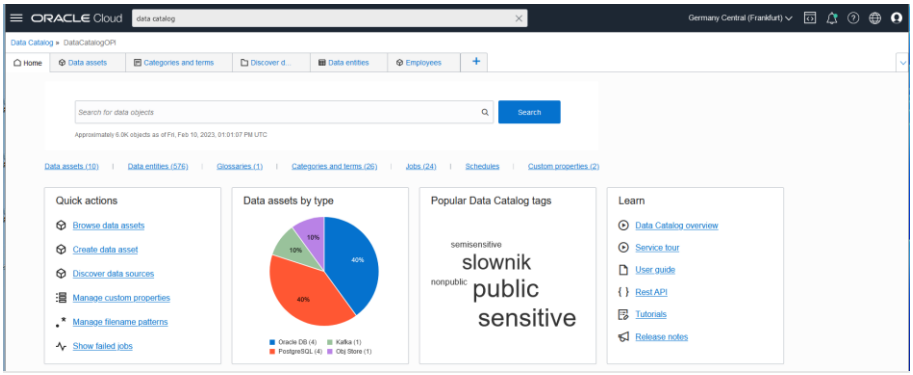

Figure 6: Oracle Data Catalog at OPI PIB

Currently, OPI PIB has registered only a few of its databases, for the purpose of ensuring proof of concept. At the institute's current stage of evolution and implementation of data governance, Oracle Data Catalog is sufficient and meets most of its requirements.

4 Conclusion

The work involved in implementing elements of the data governance programme at OPI PIB has led to deeper understanding of how crucial proper data management is for organisations. Developing an organisation's awareness and understanding of data facilitates day-to-day work, improves security, and enables the organisation to deliver superior products faster.

We attempted to evangelise about data governance. We conducted several workshops and meetings that allowed the organisation's members to explore the subject more deeply. We identified data areas and assigned business owners to them. During meetings with software developers, however, we encountered objections to change and to imposed standards for working with data.

Nevertheless, challenges remain. These include the development of a business dictionary that will enable consistent understanding of business terminology and improve collaboration between business

and technology. Implementing the business dictionary into the data catalog and linking its terms to physical data will enable easier searching and management of metadata. Data security issues can be incorporated into the data governance process and continuous implementation can be maintained while ensuring the highest possible data quality.

Data governance has now been incorporated in OPI PIB's strategic goals. To measure the effectiveness of this approach, a set of descriptive and quantifiable KPIs and OKRs has been defined since early 2023. The primary objective is to increase the organisation's maturity level to at least level 4 in accordance with the Data Governance maturity model.

From the perspective of OPI PIB, the implementation of data governance has contributed to increasing the quality of data, but above all to increasing the awareness of those responsible for the analysis and design of IT systems. This solution is highly demanding from the perspective of the change management process, and it is important not to focus solely on measurable effects and indicators, but also to be aware of the strategic goal and processable nature of this approach. Implementation of specific solutions is merely the beginning of the road; even more important is their consolidation and inclusion in the process of continuous improvement.

The implementation of data governance has already delivered measurable benefits in shortening the time necessary for analysis and obtainment of information for external stakeholders. A particularly positive role was played in these achievements by the data catalog, which shortened and unified the collection of metadata within the organisation. Plans related to the digitalisation of science and higher education in Poland are moving towards greater integration and increased interoperability with models and systems nationally and internationally. Coordination of the data governance policy between OPI PIB and other IT agencies at government level will be another challenge and a confirmation of the correctness of the institute's decision.

5 References/Citations

Earley, S., Henderson, D., & Data Management Association. (2017). *Dama-dmbok : data management body of knowledge (Second)*. Technics Publications.

Eryurek, E., Gilad, U., Lakshmanan, V., Kibunguchy, A., & Ashdown, J. (2021). *Data governance: the definitive guide : people processes and tools to operationalize data trustworthiness (First).* O'Reilly Media.

Implementation of an Institutional Data Governance Program at the University of Toronto – June 2020 (2020) Retrieved January 23, 2023, from: https://data.utoronto.ca/wp-content/uploads/2021/03/Implementation-of-an-Institutional-DG-Program-Report-2_Final.pdf

Jim, C. K., & Chang, H.-C. (2018). *The current state of data governance in higher education. Proceedings of the Association for Information Science and Technology*, 55(1), 198–206. https://doi.org/10.1002/pra2.2018.14505501022

Jörg, B. (2010). *CERIF: The Common European Research Information Format Model.* Data Science Journal, 9, CRIS24–CRIS31. https://doi.org/10.2481/dsj.CRIS4.

Michajłowicz, M., Idzi, A., Kierzkowski, J., Kucharska, I., Paszkowska, M., Podwysocki, E., Rodzik, P., Sadłowski, A., Sobkowicz, A. (2021). *Systemy informatyczne wspierające naukę i szkolnictwo wyższe : POL-on : System Informacji o Nauce i Szkolnictwie Wyższym : TOM II.* Retrieved February 17, 2023, from: https://opi.org.pl/wp-content/uploads/2021/10/POL-on.TomII-.pdf.

Michajłowicz, M., Kozłowski, M., Kowalski, M., Fijałkowski, S., Drogosz, J., Bylina, M., & Furmankowska-Podniesińska, A. (2022). *Scientific activity evaluation in Poland: the IT ecosystem and the optimal selection of achievements*. Paper presented at the EPiC Series in Computing, , 86 56-65 from: https://easychair.org/publications/paper/Rf1d.

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 February 6, 2023, from: https://www.eunis.org/download/2018/EUNIS_2018_paper_70.pdf.

Nauwerck, G., Maltusch, P., Strat, V. L., & Suominen, E. (2022). *Towards a Sector Specific Higher Education Reference Model–introducing HERM*. Retrieved February 7, 2023, from: https://www.eunis.org/download/2022/EUNIS_2022_paper_39.pdf/.

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 February 7, 2023, from: https://www.eunis.org/download/2019/EUNIS_proceedings_2019.pdf.

West, M., (2011) *3 - Some Types and Uses of Data Models*, Retrieved February 9, 2023, from: https://www.sciencedirect.com/science/article/pii/B9780123751065000038

6 Author biographies

Maria Bylina (MSc) is a software designer and business analyst at the National Information Processing Institute. She obtained a master's degree at the Faculty of Geodesy and Cartography at Warsaw University of Technology. She has also several years' experience as a GIS analyst and service maker. Her areas of interest include agile project management, data governance, ETL, and software design. She holds Professional Scrum Product Owner, Professional Scrum with Kanban and UML Professional 2 Foundation Level certificates.

Email: maria.bylina@opi.org.pl
LinkedIn: www.linkedin.com/in/mariabylina

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 warehousing, and business intelligence. His areas of interest include Oracle technology, big data, business intelligence, low-code/no-code technology, and data visualisation. He is a certified Business Intelligence and APEX specialist. 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

Marek Michajłowicz (MSc) is the Deputy Head of the National Information Processing Institute. He holds a master's degree in sociology from Cardinal Stefan Wyszyński University in Warsaw, a Bachelor of Engineering (B.E.) degree in computer science from Warsaw School of Information Technology under the auspices of the Polish Academy of Sciences, and an MBA from Warsaw University of Technology. He has many years' professional experience in business and systems analysis. His areas of interest include agile project management, software design and development, big data, and warehousing.

Email: marek.michajlowicz@opi.org.pl
LinkedIn: https://www.linkedin.com/in/marek-michajlowicz/