D3M: Automated Data-Driven Decision Making

Carles Farré, Javier Flores, Sergi Nadal, Alejandra Volkova

Universitat Politècnica de Catalunya, Barcelona, Catalonia, Spain

Abstract

Data has an undoubtedly impact on society. Storing, processing and analyzing large amounts of available data is currently one of the key success factors for an organization. Nonetheless, we are recently witnessing a change represented by huge and heterogeneous amounts of data. Thus, in order to carry on these data exploitation tasks, organizations must first perform data integration combining data from multiple sources to yield a unified view over them. In this paper, we report on the Automated Data-Driven Decision Making (D3M) project, whose main objective is to provide a mature software solution for automatic data integration with advanced decision making capabilities.

Keywords

Data-driven software engineering, Data integration, Decision making

1. Introduction

The importance of data in today's society is unquestionable. A large share of organizations base their business model on the collection, storage, and analysis of any data relevant to their business. This vision implies a radical change in the management of organizations' operations, where the collected data can be analyzed to generate relevant information for making informed decisions. This paper presents the **D3M** project, an acronym for Automated Data-Driven Decision Making. D3M[1] is a 2-year project started in December 2021, funded by the Spanish research agency, under the National Spanish Program for Research Aimed at the Challenges of Society 2020 (PdC 2021), which aims to address the current challenge of democratizing access to independent data sources to gain deeper analytical insights via *automatic data integration* and *domain-specific decision making*.

D3M is run by the integrated Software, Services, Information and Data Engineering (**inSSIDE**[2]) research group at the Universitat Politècnica de Catalunya (**UPC**). inSSIDE is composed of two subgroups: *(i)* the Software and Service Engineering (**GESSI**) group[3] and *(ii)* the Database Technologies and Information Management (**DTIM**) group[4]. These two subgroups together cover the relevant aspects related to software engineering and data engineering that lay the foundations for D3M. Altogether, the D3M research team is composed of 7 senior researchers, 4 post docs, 1 PhD student, and 1 MSc student. Moreover, use cases for D3M include external support from end-users such as epidemiologists and junior software developers to validate our proposal.

Joint Proceedings of RCIS 2022 Workshops and Research Projects Track, May 17-20, 2022, Barcelona, Spain
EMAIL: farre@essi.upc.edu (A.1); jflores@essi.upc.edu (A.2); snadal@essi.upc.edu (A.3); alejandra.volkova@upc.edu (A.4)
ORCID: 0000-0001-5814-3782 (A.1); 0000-0002-2998-9962 (A.2); 0000-0002-8565-952X (A.3); 0000-0003-4315-7684 (A.4)

© 2020 Copyright for this paper by its authors.
Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

[1] https://d3m.upc.edu/en
[2] https://insside.upc.edu
[3] https://gessi.upc.edu/en
[4] https://www.essi.upc.edu/dtim

1.1 Context

D3M is grounded on research carried out in the lines of automated data integration and domain-specific decision making. Here, we provide an overview of each.

Automating data integration tasks. Data integration is a well-studied area aimed at facilitating transparent access to various heterogeneous data sources [1]. A prominent approach to data integration is exposing a knowledge graph conceptualizing the domain of interest to offer a uniform query interface over the sources. Queries over the knowledge graph are rewritten over the sources via schema mappings. The maintenance of such constructs (e.g., evolving the knowledge graph, adding new sources and mappings) is an arduous and manually-intensive task that hinders the ability of such systems to flexibly adapt and provide right-time integration [2]. This limitation has been coined as the data variety challenge, which refers to the complexity of providing on-demand integration over a vast and evolving set of data sources. Dataspaces, which are data integration systems embracing a pay-as-you-go approach by gradually integrating data sources as needed, represent a significant step toward tackling the variety challenge. With the vision of reducing the usual upfront and maintenance costs, dataspaces claim for the adoption of a flexible and dynamic approach where different integration tasks are automated. One of them, known as *bootstrapping* [3], is the process of automatically generating the knowledge graph driven by the data sources, with the goal of incrementally building the query interface and mappings to query such heterogeneous data sources in an integrated manner.

Domain-specific decision making. Organizations require facilitating access to informed decision making based on Key Performance Indicators (KPIs) relevant to their business. However, creating decision making support systems is expensive, time-consuming, and error-prone. The use of domain-specific, operationalized quality models offering actionable analytics from heterogeneous sources has been successful in multiple domains (e.g., software analytics [4]). It enables plenty of analytics scenarios, from current situation assessment to prediction and what-if analysis. In a recent systematic review, data integration and final data aggregation were reported as part of the remaining challenges in Big Data analytics [5]. At the same time, current approaches shall analyze more than one artifact and focus on integrating data from different sources and getting a holistic view [6]. Thus, to enable domain-specific strategic indicators and data-driven decision making, it becomes necessary to facilitate the integration of data sources driven by the real information needs of end users.

1.2 Background

This project builds upon two research assets reported in the project Generation and Evolution of Smart APIs (**GENESIS**), funded by the National Spanish Program for Research Aimed at the Challenges of Society 2016: a dataspace management system (hereafter referred to as ODIN), and a software analytics tool (hereafter referred to as Strategic Dashboard). These products have been successfully validated as a prototype in pilot projects.

ODIN. ODIN (short for *On-demand Data INtegration*) is a dataspace management system grounded on knowledge graphs [7]. ODIN is conceived to overcome the limitations of traditional virtual data integration in large-scale scenarios where data variety plays a key role [8]. Figure 1, depicts how ODIN supports the dataspaces' complete lifecycle. ODIN automatically extracts the schemata from structured (e.g., relational) and semi-structured (e.g., JSON) data sources and translates them into a canonical data model. To that end, a set of production rules parse their metadata and generate *source graphs*. These are aligned while considering user feedback throughout this process. As a result, ODIN generates *provenance graphs* (PG) tracing the results of the previous stages. A PG is a target-agnostic metadata construct describing the integration of a particular set of data sources. It captures the results of bootstrapping the sources and aligning their schemata, and guarantees we can generate target-specific metadata from them. Thus, PGs are used to generate specific constructs of a given integration tool, such as *conjunctive query (CQ)-oriented graphs*, which expose the sources' schemata in first-normal form, and are then linked via *local-as-view* (LAV) schema mappings that connects elements of the sources' schemas to the *global graph*. LAV mappings characterize the sources in terms of a query over the

knowledge graph, making them inherently more suitable in data variety settings, where new sources may be added or outdated sources removed dynamically.

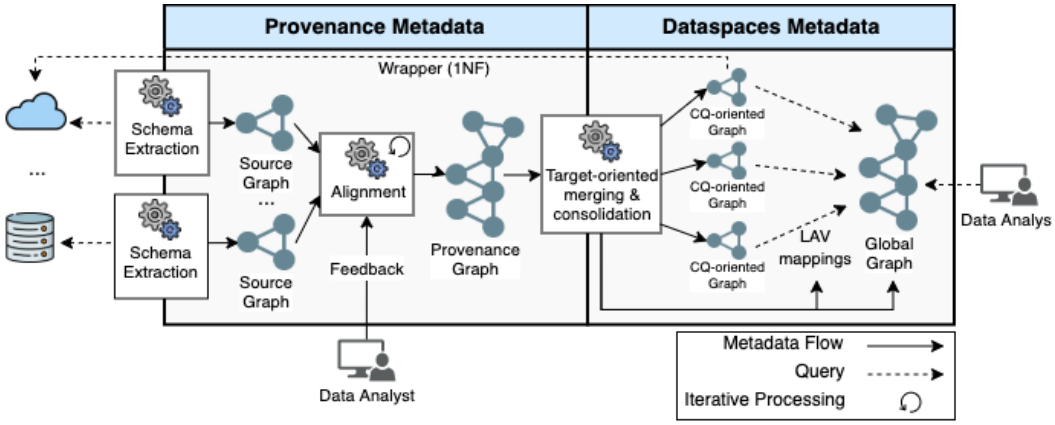

Figure 1: High-level overview of ODIN

The Strategic Dashboard. The Strategic Dashboard is a modular, configurable, and extensible software analytics tool used in Agile Software Development projects to improve the software development process and the quality of the software produced. The Strategic Dashboard (Figure 2) enables decision makers to define their own *Quality Model*, which is composed of quality-related *Strategic Indicators* (e.g., customer satisfaction, process performance, risk level) [4], decomposed in their turn into *Quality Factors* related to system development and usage (e.g., development speed, software performance). Quality factors are defined over different *Quality Metrics* (e.g., commits per day, duplicated lines of code, software response time). The Strategic Dashboard automatically performs a quality assessment of the whole quality model defined. Raw data is collected from multiple sources of information, such as *development* tools used by the software development team (e.g., JIRA, Github) and *software usage* from end-users (e.g. software logs). All the information is collected through *Connectors* that feed a Distributed Data Sink from which the quality metrics, quality factors, and strategic indicators are computed bottom-up. The quality assessment enables the strategic dashboard to perform several analyses that are provided to the Decision Maker:

- *Visualization* of the current (and historical) status of software products and development processes through an easy-to-use interface with advanced navigational capabilities.
- *What-if analysis techniques* enable decision makers to evaluate different scenarios based on the impact of metrics on quality factors and, further on, on the strategic indicators.
- *Forecasting techniques* estimate the values of the strategic indicators and quality factors in a time frame, to predict and anticipate future issues in the software development process.
- Semi-automatic *generation of new requirements* in response to alerts when a quality model element (typically, a strategic indicator) drops below unsatisfactory levels of quality [9].

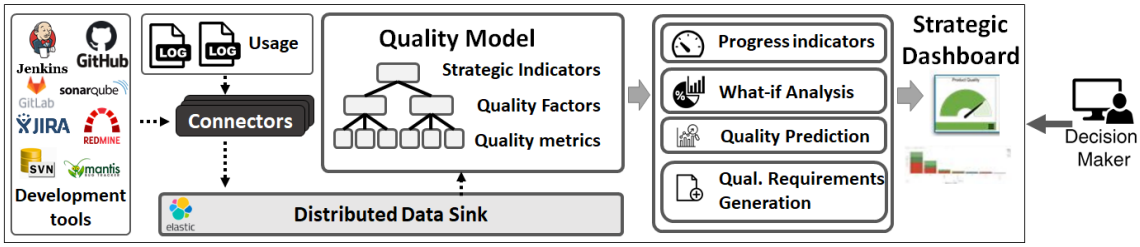

Figure 2: Strategic Dashboard architecture

Despite the benefits ODIN provides in terms of data integration, its query interface is limited to technical users familiar with semantic web technologies. Thus, there is a gap between such a low-level interface and the advanced capabilities that decision makers need in their organizations (e.g., progress indicators, what-if analysis). Additionally, the Strategic Dashboard is tightly coupled to the Distributed

Data Sink, built ad-hoc for a specific domain. This difficults the integration of new data sources and the calculation of new types of metrics, quality factors, and strategic indicators.

2. Objectives

The main objective of D3M is **to adapt and integrate these two independent tools, ODIN and the Strategic Dashboard, into a unified product** bringing together the benefits of them both: **(i)** enabling the integration of disparate data sources in an incremental manner and **(ii)** provide advanced support on top of them for decision making processes via advanced dashboard interfaces. The project's main objective further decomposes into four specific objectives:

- **O1**: *Data-driven semi-automatic bootstrapping*. Provide means to enable an incremental semi-automatic extraction of the knowledge graph from a set of heterogeneous and independent data sources. This objective starting point is ODIN's core and will endorse it with new support for its enrichment with day-to-day concept vocabulary, and its enrichment with the domain-specific quality models.
- **O2**: *Integrated data exploration interface*. Enable data wrangling tasks (navigational queries on tabular and semantic data) from heterogeneous data sources federated through a knowledge graph. This refers to a new exploitation feature required by our industrial partners (as an alternative to traditional decision making support).
- **O3**: *Customized decision making support*. Enable the creation of an advanced dashboard that spans heterogeneous data sources applying domain-specific quality models to assist decision makers. This objective generalizes the available Strategic Dashboard to provide support in any domain, with improved techniques.
- **O4**: *Unified product to support the end-to-end decision making process over heterogeneous data sources*. O4 integrates the results of O1-O3; i.e., it features incremental bootstrapping of the knowledge graph from the data sources of interest, and two kinds of exploitation: decision making support based on strategic indicators and data exploration based on data wrangling.

To achieve such objectives, D3M proposes the architecture presented in Figure 3. D3M serves two types of data consumers: *data wranglers* (for data exploration) and *decision makers* (for advanced analytics and data-driven quality models). Besides, it requires interaction with other users for managing the system metadata, such as *domain experts* (for enriching the bootstrapped knowledge graph with day-to-day concepts and a domain-specific quality model) and *data stewards* (for assisting in the configuration of the alignments among heterogeneous data sources). While the integrated architecture proposed in Figure 3 offers the benefits of both ODIN and the Strategic Dashboard, it as well offers innovation by boosting the automated decision making process by means of linking heterogeneous data sources to the defined quality model via knowledge graphs, hence facilitating and mainly automating the calculation of the strategic indicators and their visualization for the decision makers. Besides the aforementioned objectives, D3M also aims to attain the following ones:

- **O5**: *Incremental technology transfer of the proof of concept*. Execute a technology transfer plan to assure an incremental evolution of the maturity level of the developed software components for D3M via validation and demonstration of the proposed proof-of-concept.
- **O6**: *Assessment of the viability of the proof of concept*. Perform a market analysis to assess the technical, commercial, and social viability of the proposed product, and uncover evolutionary paths for D3M becoming a product adapted to current industry needs.
- **O7**: *Long-term sustainability of the proof of concept*. Cultivate a broad network of industry and public sector contacts to create awareness and attract prospective customers.
- **O8**: *Intellectual property right assurance*. Develop a strategy for managing the intellectual and industrial property rights of the developed proof of concept.
- **O9**: *Endorsing the project team with entrepreneurship skills*. Define a training plan with a list of entrepreneurship courses and monitor its execution.

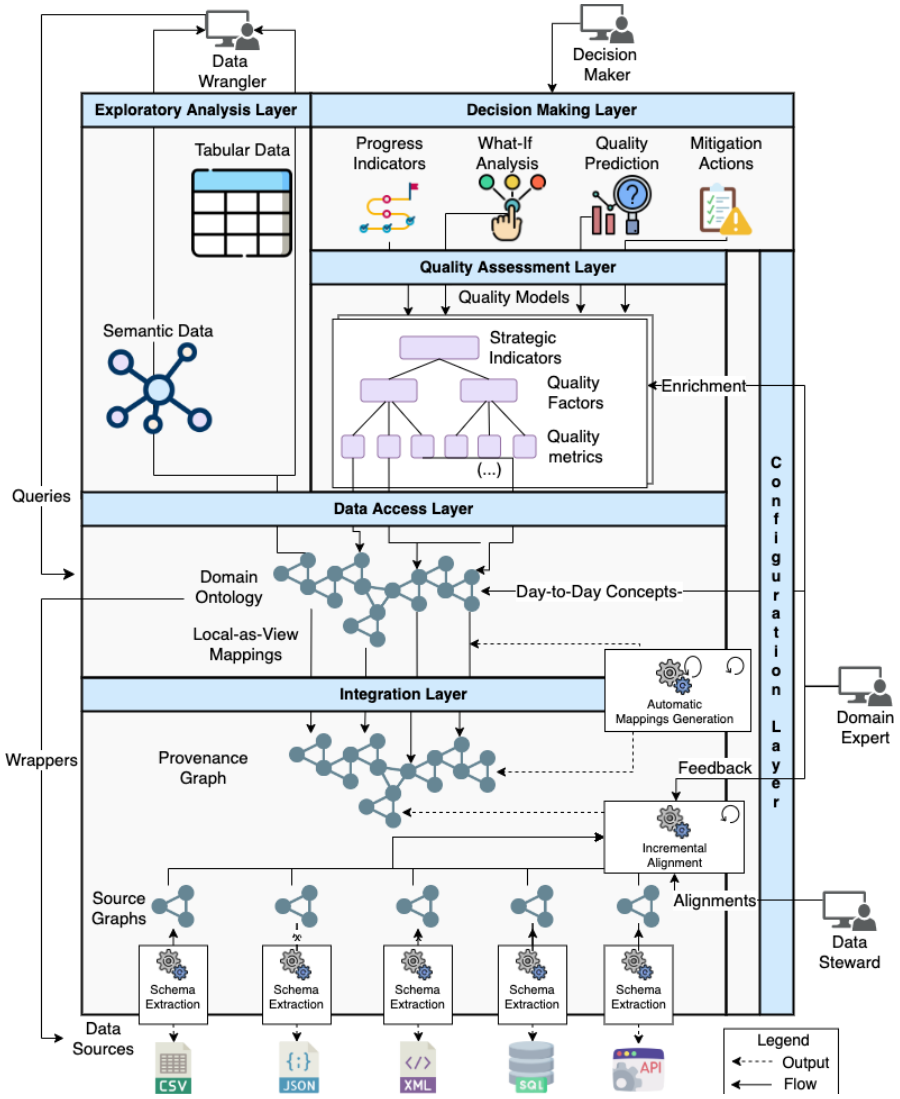

Figure 3: Overview of the D3M concept

3. Use cases

Here, we depict two industrial projects that serve as use cases for D3M. Currently, each one is evolving ODIN and Strategic Dashboard in parallel, so that it is possible to apply the improvements achieved to the overall D3M project. For each, we describe the use case context and how adopting D3M can aid in the organization's decision making needs.

3.1. Development of an imaging platform development for Malaria and Neglected Tropical Diseases (NTDs)

A recent study by the World Health Organization shows that in 2018 an estimated 228 million cases of malaria occurred worldwide, the majority in the African region[6]. The SDG targets 3.3 and 3.8 call for an end to such kind of epidemics by 2030. The main goal of this project is to develop an imaging platform by using artificial intelligence techniques for automated diagnosis of Malaria, Tuberculosis, and NTDs. The specific objectives are: *(i)* create an open source image bank and database; *(ii)* develop an image diagnostic system by image analysis using artificial intelligence techniques; *(iii)* develop software for Android-phones to move the microscopy slides, images acquisition, image analysis, and

[6] https://www.who.int/publications/i/item/9789241565721

diagnosis; *(iv)* model the laboratory management software to be able to import the microscopic images and resend them to the general microscopy image bank; *(vi)* establish a quality control of the slides preparation, digital microscopic images and image diagnosis; *(vii)* validate the imaging platform in the field.

The role of D3M in this use case is to empower epidemiologists to cross-analyze diagnosis data predicted automatically by the imaging platform with other contextual data collected from the available data sources. For instance, the analysis of comorbidity, or coinfections, represents a paradigm change in how health diseases are treated. Traditionally, individual diagnoses were performed for each analyzed disease. However, major disease outbreaks have shown that previous conditions can impact the diagnosis. Similarly, many countries (un)intentionally omit to report on new infection cases, either due to limited resources or political issues. To get a complete picture of the situation, cross country-reported data with other sources may indicate the prevalence (e.g., amount of medicine requested). However, data integration needed to calculate these indicators is far from being trivial, especially in the case of NTDs that lack systematic data collection and in developing countries with minimal resources. To that end, as depicted in Figure 4, D3M presents the user with a knowledge graph conceptualizing all domain elements of interest which are further linked to the different available data sources. With D3M, epidemiologists will be able to cross different data sources guided by relevant strategic indicators from the analytical dashboard, thus obtaining a more realistic and complete picture of the situation, and making a paradigm shift from a disease-centric to a patient-centric analysis.

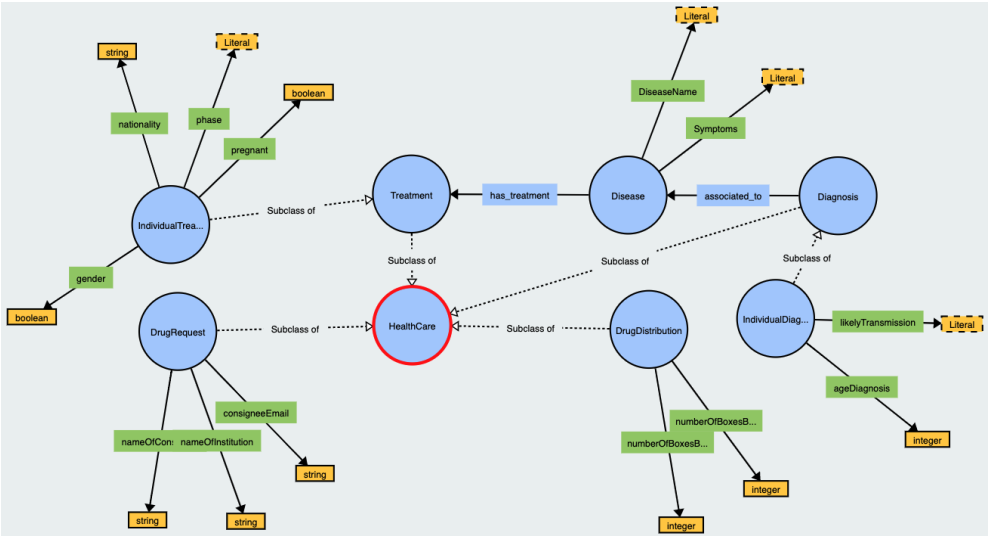

Figure 4: Visualization of the health use case knowledge graph

3.2. Software Analytics

The domain of Software Analytics is broad and can be applied to various environments. This use case focuses on the higher education (i.e., universities) domain via the project *Implementation of a dashboard for monitoring the progress of software projects developed by student teams*, which can be easily extrapolated to scenarios with teams of junior software developers. The project aims to allow both students and professors to receive accurate and objective feedback on the individual and team learning process. Informed decisions can be made about prioritizing, planning, and evaluating their actions throughout the project. To this aim, an onboarding step will be beneficial for training juniors to learn how to extract insights and take data driven decisions from the information generated by the dashboard. D3M comes into play by using the Strategic Dashboard, which was adapted to the specific domain by creating new *Connectors*, and defining specific *Quality Metrics,* and customizing several *Visualizations*, so as part of future work it would be helpful to incorporate ODIN that can give support for *Connectors'* and *Quality Model* dataspace management systems.

The first connector that we created for *GitHub*, a provider of Internet hosting for software development and version control using *Git*, allows us to collect information about commits, modified

lines, and issues. Another connector was made for *Taiga*, a free and open-source project management system for startups and agile developers, this one supplies us data about the *Scrum* methodology' resources, as user stories and tasks to deal with in each *Sprint*. Based on the information provided by these connectors, we defined different metrics: (*i*) the percentage of commits of each member of the team and its corresponding modified lines; (*ii*) the percentage of tasks assigned to each developer; (*iii*) the percentage of closed task by assignee; (*iv*) number of tasks without assignee and (*v*) standard deviation on numbers of commits or tasks. In addition to the previous metrics, we decided to focus on the quality and correctness of team members's information to connectors when they use *Taiga* or GitHub. For instance, we check (*vi*) if acceptance criteria are used when a user story is created or (*vii*) if a standard user story pattern is applied, also it is interesting to see (*viii*) if commits contain real task reference. In conclusion, all of them help to monitor the progress of software projects, some from the point of view of project management and others from the point of view of code development.

For team project metrics visualizations (see Figure 5), there is a display of the current evaluation, which is calculated according to the formula settings for each metric and from the data collected by the connectors for this particular day. With the following representation, we can see the exact value of the metric rounded to the hundredth, through a half circle graph with different color categories. The categories are customizable, that is, the number of colors and the limits of each color can be defined in a way that best suits the metrics. Apart from the current evaluation, it is possible to visualize the historical data of the metrics through a line graph, that is, their evolution over time, to monitor progress as the course progresses.

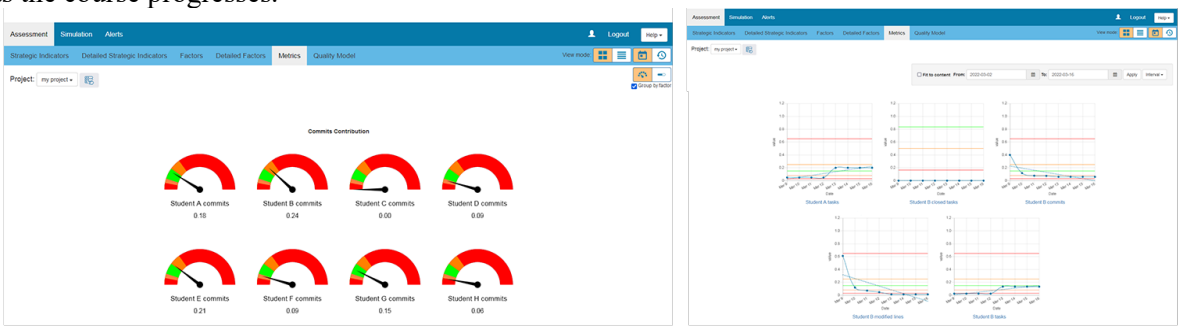

Figure 5: Visualization of the current state (on the left) and historical evolution (on the right) of the metrics.

4. Relevance to information science

The underlying research carried out in the context of the D3M project addresses a broad spectrum of challenges related to the information science field. Indeed, being D3M a project oriented to the development of a software prototype, it can fall in the area of Information Systems and their Engineering. Additionally, given that data integration is at the heart of D3M, it naturally fits the Data and Information Management area, and Data Science. Furthermore, considering the applicability of the project results via use cases to the industry, D3M is also relevant for the Domain-specific IS Engineering area (e.g., for the health or educational domains).

5. Open lines of research

Numerous open lines of research arise from D3M. A key question to be addressed is *how far can we automate the process of data integration?* In other words, where is the sweet spot that allows automating manual and cumbersome tasks without compromising the quality of the results obtained when the user is involved. It is already known that a fully-automated approach to data integration is not feasible, given that there will always exist some level of uncertainty and ambiguity. Nevertheless, we strive to minimize the efforts required by users to address these cases.

Another scientific challenge that D3M should face is *how to create and assess domain-specific strategic indicators for any domain?* In this regard, we have already met some of the issues that must be addressed in the future: (*i*) enable the on-demand and incremental definition of metrics, factors, and

strategic indicators; (*ii*) define and implement a comprehensive catalog of visualizations for such metrics/factors/indicators; and (*iii*) simplify and automate as much as possible the configuration and deployment of strategic dashboards in new domains.

6. Conclusions

In this paper, we have presented the D3M project, an ongoing two-year project that will combine and extend the efforts accomplished in *ODIN* and *the Strategic Dashboard* into a unified tool. This solution will provide data wranglers with the mechanisms to easily integrate heterogeneous data sources and have the means to extract analytical insights for data-driven decisions. The features of D3M will be used on two industrial projects related to the domains of healthcare and software development. We believe the results of D3M will provide the following achievements: *(i)* scalable and automated data integration life cycle, *(ii)* effectively democratizing data access, *(iii)* advanced analytic models for predicting and optimizing outcomes, *(iv)* a set of user-friendly dashboards to assist non-technical end-users with exploratory and analytical tasks. Therefore, D3M can have a significant impact on the industry.

Acknowledgements

This paper has been funded by the Spanish Agencia Estatal de Investigación (AEI) under project / funding scheme PDC2021-121195-I00.

References

[1] M. Lenzerini. "Data Integration: A Theoretical Perspective". In Proceedings of the 21st ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems (2002), 233-246.
[2] S. Nadal, A. Abelló, O. Romero, S. Vansummeren, and P. Vassiliadis. "Graph-driven Federated Data Management". In IEEE Transactions on Knowledge and Data Engineering (2021). Online (https://ieeexplore.ieee.org/document/9422168)
[3] J. Sequeda, S. H. Tirmizi, O. Corcho et al. "Survey of Directly Mapping SQL Databases to the Semantic Web". In Knowledge Eng. Review (2011), 26.4
[4] S. Martínez-Fernández, A. M. Vollmer, A. Jedlitschka et al. "Continuously assessing and improving software quality with software analytics tools: a case study". IEEE Access 7 (2019), 68219-68239
[5] U. Sivarajah, M. M. Kamal, Z. Irani, V. Weerakkody. "Critical analysis of Big Data challenges and analytical methods". Journal of Business Research, 70 (2017), 263-286.
[6] The Forrester Wave™: Value Stream Management Solutions, Q3 2020, available at https://www.forrester.com/report/The+Forrester+Wave+Value+Stream+Management+Solutio ns +Q3+2020/-/E-RES159825.
[7] S. Nadal, K. Rabbani, O. Romero, S. Tadesse "ODIN: A Dataspace Management System". In International Semantic Web Conference (ISWC 2019) (pp. 185-188).
[8] S. Nadal, O. Romero, A. Abelló, P. Vassiliadis, S. Vansummeren. "An integration-oriented ontology to govern evolution in Big Data ecosystems". Inf. Syst. (2019), 79: 3-19
[9] M. Oriol et al. "Data-driven and Tool-supported Elicitation of Quality Requirements in Agile Companies". Software Quality Journal (2020), 28(3): 931-963