Public Health and Informatics
J. Mantas et al. (Eds.)

doi:10.3233/SHTI210201

A Comprehensive Portal for Clinical and Translational Data Warehouses

Marco JOHNS[a,1], Armin MÜLLER[a], Felix Nikolaus WIRTH[a] and Fabian PRASSER[a]
[a] *Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Medical Informatics Group, Charitéplatz 1, 10117 Berlin, Germany*

Abstract. Data-driven methods in biomedical research can help to obtain new insights into the development, progression and therapy of diseases. Clinical and translational data warehouses such as Informatics for Integrating Biology and the Bedside (i2b2) and tranSMART are important solutions for this. From the well-known FAIR data principles, which are used to address the aspects of findability, accessibility, interoperability and reusability. In this paper, we focus on findability. For this purpose, we describe a portal solution that acts as a catalogue for a wide range of data warehouse instances, featuring a central access point and links to training material, such as user manuals and video tutorials. Moreover, the portal provides an overview of the status of multiple warehouses for developers and a set of statistics about the data currently loaded. Due to its modular design and the use of modern web technologies, the portal is easy to extend and customize to reflect different corporate designs and institutional requirements.

Keywords. Translational Research; Data Warehouse; i2b2; tranSMART; Portal

1. Introduction

Data-driven methods in biomedical research can help to obtain new insights into the development, progression, and therapy of diseases. These can then be utilized, for example, in innovative methods of clinical decision support. In order to find relationships between biomedical parameters in large and heterogeneous datasets, these must be integrated. Clinical and translational data warehouses such as Informatics for Integrating Biology and the Bedside (i2b2) [1] and tranSMART [2] are important solutions for this. They are aimed at end users such as medical researchers, and support use cases such as cohort selection, hypothesis generation, and ad hoc data analysis.

To support researchers with the secure and efficient utilization of translational data warehousing solutions, we have established a platform at the Berlin Institute of Health. The infrastructure is based on container technology and supports setting up new instances and terminating old instances in a scalable manner. For this purpose, we have adopted the open source approach presented by Spengler et al. [3], which makes use of the Docker software stack for containerization and management of multiple instances. Another advantage of the approach is that it supports agile processes for data loading based on close feedback cycles between informaticians and medical researchers. Important requirements for data integration solutions can be derived from the FAIR

[1]Corresponding Author, Marco Johns, Berlin Institute of Health at Charité – Universitätsmedizin Berlin, Medical Informatics Group, Charitéplatz 1, 10117 Berlin, Germany; E-mail: marco.johns@charite.de.

principles [4], according to which research data should be Findable, Accessible, Interoperable and Reusable. Data warehousing platforms especially address the aspects of accessibility, interoperability (if according terminology standards are integrated, as is the case with our platform) and reusability. In this paper, we describe our approach for additionally addressing the findability aspect.

2. Objective

Both i2b2 and tranSMART are accessible via web interfaces and our platform is based on the principle of quickly ramping up and shutting down individual instances. This means that different user groups need to access different instances, which are reachable via different web addresses. Moreover, it can be challenging for the development team to maintain an overview of the status of different instances and what exactly they are used for at a current point in time.

In this paper, we describe a portal solution that we have added to our infrastructure which aims to address these challenges. It offers two important functionalities for users: (1) it provides a central access point to all warehouses provided by the infrastructure, (2) it displays links to further information, such as user manuals and video tutorials. Moreover, it is also helpful for developers: (1) it provides a central point at which an overview of the status of all instances can be obtained, (2) it offers access to a set of statistics about the instances, including an overview of the data that is currently loaded.

3. Methods

3.1. Overview of the Data Warehousing Infrastructure

As mentioned above, our platform offers access to i2b2, which is a National Institutes of Health (NIH)-funded platform for clinical data warehouses, and tranSMART, which is based on i2b2 and provides an additional set of data exploration and visualization tools. The current development of both platforms is coordinated by the not-for-profit i2b2-tranSMART Foundation. To date, both platforms are used internationally in hundreds of healthcare institutions [5, 6].

An overview of our infrastructure is shown in Figure 1. As can be seen, the data warehouses are deployed in separate, containerized instances. The instances, containing i2b2 as well as tranSMART, are configured to use a PostgreSQL database, resulting in a setup that solely consists of open-source software. A single web server (i.e. Apache 2) is placed in a dedicated gateway container, which manages ingoing and outgoing communication using proxy and reverse proxy configurations for each instance's i2b2 and tranSMART installation [3].

As a first step, we have extended the gateway container to automatically load the portal's web-resources into a newly configured document root of the web server, thus allowing users to access the portal via a configured domain name.

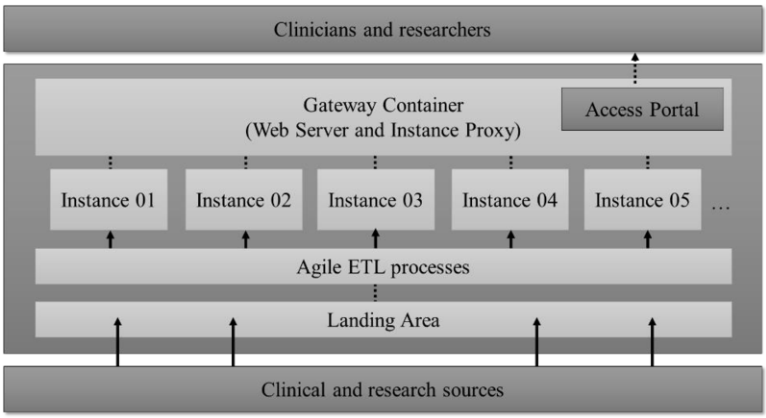

Figure 1. Architecture overview of the warehousing platform.

3.2. Portal-specific Infrastructure Extensions

To provide access to the data displayed in the portal, we extended the infrastructure with an additional statistics service that exposes an overview of the content of each data warehouse via a RESTful interface. These parameters include: (1) a descriptive name of the instance (e.g. a project name), (2) a timestamp of the last loading process and (3) the number of data items contained, broken down by *patients, visits, observations, attributes* and *meta-attributes*. The service was implemented using the Spring Boot framework and can be deployed as a standalone Java archive within a container that has access to the underlying PostgreSQL database. The connection properties can be provided as command line arguments, which are set during the automated configuration of service instances. As is illustrated in Figure 1, we have further configured the gateway container to route requests for statistics for a certain warehouse to the correct service instance.

3.3. Architecture of the Portal

The portal has been implemented as a client-side web application using modern web technologies (HTML5, CSS3, jQuery) and the Materialize CSS framework to get a responsive and clean look and feel. Materialize is a lightweight framework, adding only two additional files, a minified CSS- and JavaScript-file with all non-functional characters removed [7].

The portal features a multi-language approach, currently offering German and English translations accessible via a drop-down menu on every page of the portal. Selecting a different language also ensures that all links embedded in the website also lead to pages in the selected language. Translations are provided by having different language versions of the underlying content pages.

The front end communicates with the back end services via endpoints following the representational state transfer (REST) approach. Therefore, the web server, as well as the extension providing information about the instances, have been configured to allow Cross-Origin Resource Sharing (CORS) only for requests from the Gateway.

To query an instance's status, the front end sends a HTTP GET-request to the back end to check whether the associated service is online. The response contains a HTTP

status code, which is presented in a simplified form to the user. If the request is unsuccessful, the HTTP status code and its message are shown as a tooltip to the user and can be used to communicate problems with the platform administrators. When retrieving data to display on the statistics page, a HTTP GET-request with response type JSON is sent to the back end. On a successful request, the response is then parsed, and the statistics are dynamically loaded into the statistics page. If the request fails, the status of the corresponding service is set accordingly, and no data is displayed.

4. Results

The data warehousing portal is currently used to provide access to a range of data warehouses from different departments and research areas. Since the portal offers multiple perspectives, it can be used by technical professionals as part of the administration process as well as by regular users. An overview of the different perspectives is provided in Figure 2.

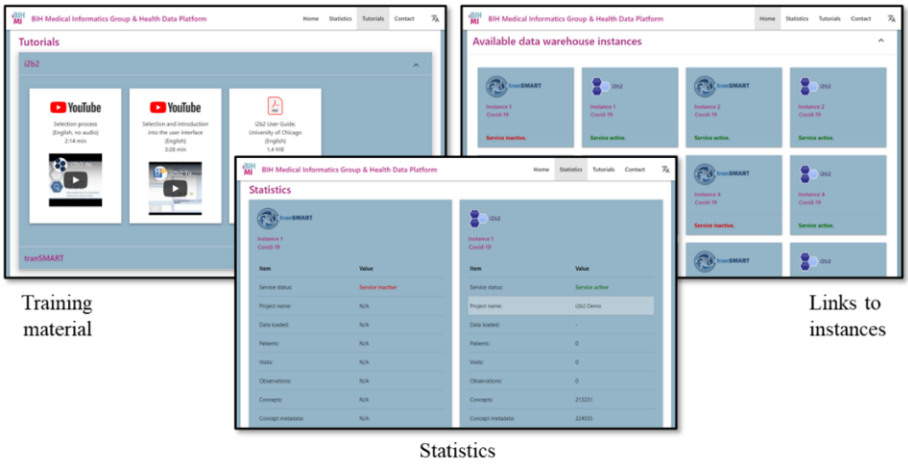

Figure 2. Screenshots of the platform portal.

The portal is built in a modular manner so that new instances or perspectives can easily be added while providing a unified visual representation. Due to the use of modern web technologies, the design can be changed easily to reflect different corporate designs. As can be seen in Figure 2, we chose a design that follows our institutional guidelines. As a “one-stop-shop”, the portal improves the user experience for different types of user groups and allows regular users to access different warehouse instances from the same web resource, contributing to the Findability of data in the context of the FAIR principles within our institution.

5. Discussion and Conclusions

In prior work, portals have often been realized by single instances of i2b2 or tranSMART (see, e.g. [8]) providing integrated access to a common set of selected data elements. This approach is not applicable in our context, however, as we aim to provide

access to a wide range of data warehouses that integrate different types of data for different groups of users. Other groups have created portals by deeply integrating different types of applications managing different types of data (see, e.g., [9]). While this is different from our approach, the underlying mechanisms, e.g. common user authentication and authorization, are also on our roadmap.

In future work, we plan to integrate our approach more closely with local infrastructure resources, in particular with the institutional Active Directory instance. Moreover, we plan to integrate a data provenance and quality tracking pipeline based on the approach presented in [10] and to also present the data quality metrics calculated within our portal. Finally, after a consolidation phase, we plan to publish our platform as open-source software.

References

[1] Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, Kohane I. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010 Mar 1;17(2):124–30.

[2] Athey BD, Braxenthaler M, Haas M, Guo Y. tranSMART: An Open Source and Community-Driven Informatics and Data Sharing Platform for Clinical and Translational Research. AMIA Jt Summits Transl Sci Proc AMIA Jt Summits Transl Sci. 2013;2013:6–8.

[3] Spengler H, Lang C, Mahapatra T, Gatz I, Kuhn KA, Prasser F. Enabling Agile Clinical and Translational Data Warehousing: Platform Development and Evaluation. JMIR Med Inform. 2020 Jul 21;8(7):e15918.

[4] Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten J-W, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJG, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PAC, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone S-A, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016 Dec;3(1).

[5] i2b2 tranSMART Foundation. Our History – i2b2 tranSMART Foundation. [Internet] Available at: https://i2b2transmart.org/our-history/ Accessed 2021 Jan 29.

[6] Wagholikar KB, Mendis M, Dessai P, Sanz J, Law S, Gilson M, Sanders S, Vangala M, Bell DS, Murphy SN. Automating Installation of the Integrating Biology and the Bedside (i2b2) Platform. Biomed Inform Insights. 2018 Jan 1;10.

[7] Prabhu A, Shenoy A. Introducing Materialize. In: Introducing Materialize. Berkeley, CA: Apress; 2016. p. 1–9.

[8] Gainer VS, Cagan A, Castro VM, Duey S, Ghosh B, Goodson AP, Goryachev S, Metta R, Wang TD, Wattanasin N, Murphy SN. The Biobank Portal for Partners Personalized Medicine: A Query Tool for Working with Consented Biobank Samples, Genotypes, and Phenotypes Using i2b2. J Pers Med. 2016 Feb 26;6(1).

[9] Zhang C, Bijlard J, Staiger C, Scollen S, van Enckevort D, Hoogstrate Y, Senf A, Hiltemann S, Repo S, Pipping W, Bierkens M, Payralbe S, Stringer B, Heringa J, Stubbs A, Bonino Da Silva Santos LO, Belien J, Weistra W, Azevedo R, van Bochove K, Meijer G, Boiten J-W, Rambla J, Fijneman R, Spalding JD, Abeln S. Systematically linking tranSMART, Galaxy and EGA for reusing human translational research data. F1000Research. 2017;6.

[10] Spengler H, Gatz I, Kohlmayer F, Kuhn KA, Prasser F. Improving Data Quality in Medical Research: A Monitoring Architecture for Clinical and Translational Data Warehouses. In: 2020 IEEE 33rd International Symposium on Computer-Based Medical Systems (CBMS). Rochester, MN, USA: IEEE; 2020. p. 415–20.