dHealth 2024
D. Hayn et al. (Eds.)

doi:10.3233/SHTI240037

Exploring Opportunities for Clinical Data Warehouse Enhancement Through Data Catalog Integration

Andreas WALKER[a,1], Werner O. HACKL[a] and Bernhard PFEIFER[a,b]

[a] *Division for Digital Health and Telemedicine, UMIT TIROL - Private University for Health Sciences and Health Technology, Hall in Tirol, Austria*
[b] *Department of Clinical Epidemiology, Tirol Kliniken GmbH, Innsbruck, Austria*

Abstract. Secondary use of clinical health data implies a prior integration of mostly heterogenous and multidimensional data sets. A clinical data warehouse addresses the technological and organizational framework conditions required for this, by making any data available for analysis. However, users of a data warehouse often do not have a comprehensive overview of all available data and only know about their own data in their own systems - a situation which is also referred to as 'data siloed state'. This problem can be addressed and ultimately solved by implementation of a data catalog. Its core function is a search engine, which allows for searching the metadata collected from different data sources and thereby accessing all data there is. With this in mind, we conducted an explorative online market survey followed by vendor comparison as a pre-requisite for system selection of a data catalog. Assessment of vendor performance was based on seven predetermined and weighted selection criteria. Although three vendors achieved the highest score, results were lying closely together. Detailed investigations and test installations are needed for further narrowing down the selection process.

Keywords. Metadata management, Data Catalog, Clinical Data Warehouse, Secondary use, Health data

1. Introduction

The increasing digitization of health information goes hand in hand with great potential for secondary use. However, this indirect use of electronic health data, also known as reuse, implies a prior integration of usually very heterogeneous and multidimensional data sets [1, 2]. A clinical data warehouse (CDW) addresses the technological and organizational framework conditions required for this, by making any data available for analysis [3].

However, data warehouse users often do not have a comprehensive overview of all the data contained in the CDW. Furthermore, they do not know the organizational contexts in which this data was created. Hindered by the so called 'data siloed state', practitioners often have limited access to comprehensive knowledge of clinical health

[1] Corresponding Author: Andreas Walker, UMIT TIROL, Hall in Tirol, Austria, E-Mail: Andreas.Walker@umit-tirol.at

data beyond their domain or specific documentation systems. Overcoming this challenge is crucial for unlocking the full potential of a CDW. To achieve this, there is a need for detailed metadata to dismantle data silos and enable efficient secondary use of health data.

Effectively managing metadata is essential for transforming existing information into a valuable resource for data analytics. Achieving systematic metadata management is made possible through the implementation of a data catalog. Such a catalog serves as a centralized database, containing metadata from various data sources within the application system landscape. At its core, a data catalog functions as a search engine, enabling users to explore metadata collected from diverse data sources and gain access to a comprehensive dataset. In the realm of data analysis, organizations utilize data catalogs for data-driven innovations [4].

This transformative technology is not limited to corporate advantages; when integrated with a CDW, data analysts and medical IT specialists are empowered to work beyond their familiar datasets, broadening their access to valuable information. The global view of the CDW enables them to use the most suitable data for their analyses and thus efficient secondary use. In this context, the selection and implementation of a data catalog are underway for the CDW at an Austrian university hospital.

Our goal is to enhance the secondary utilization of clinical health data and elevate user interaction with CDW. This paper outlines the planned methodology as a pre-requisite for system selection, providing insights into the approach adopted for this optimization process.

2. Methods

Before specifying the concrete requirements for such a data catalog system, we conducted an initial online market survey. This survey focused on seven predetermined requirements based on a literature analysis, aiming to narrow down the numerous potential vendors. Out of these, two were established as mandatory criteria, while three were designated as target criteria and two as optional criteria. Table 1 gives an overview of the seven prioritized requirements.

Table 1. Prioritized requirements for pre-selection of potential vendors. 1 = optional criterion, 2 = target criterion and 3 = mandatory criterion. The first five criteria were taken from Olesen-Bagneaux's book "The Enterprise Data Catalog" [4]

Requirement	Description	Weighting (Scale: 1 – 3)
Data lineage	Central component for the automated documentation of data movements. Data lineage shows how data moves within the IT landscape (from system to system) and, ideally, is transformed in the process (in other words: the organizational context in which this data was created and what it is used for)	3
On premises	Server-based application that is installed and used locally (as opposed to cloud computing or software as a service)	2
Data intelligence	Use of the data catalog for data governance and data analytics purpose	2
Data Governance	Supports governance end users in the management of confidential and sensitive data	3

Knowledge graph-powered	Improves data search by modeling the knowledge universe as an ontology	1
Open Source	Free software and source code open	1
Health Data	Vendors of data catalogs that specialize in or have experience processing healthcare data to meet the needs of a healthcare organization	2

For vendor comparison we examined whether vendors meet the criteria completely, partially, or not at all. To assess vendor performance, we calculated the total score of the criteria met. For this purpose, the number of fulfilled mandatory, target and optional criteria was multiplied by their respective weighting value for each vendor. If a criterion was partially met, only half of the weighting value was considered. The total score was then calculated by adding those results together.

3. Results

A total of six different vendors were compared as part of an online market survey. Table 2 shows vendors meeting the predetermined criteria completely (C), partially (P), or not at all (N). For assessment of vendor performance, their total score was calculated based on the fully or partially fulfilled criteria.

In the result, three vendors achieved the highest score with ten points, two vendors took the second place (nine points) and one vendor scored lowest with eight points.

Table 2. Vendor comparison. C = Complete, P = Partial and N = None. Total score = value of vendor performance.

Criteria	**Compendium** [5]	**Informatica** [6]	**Amundsen** [7]	**Alation** [8]	**Manta** [9]	**data.world** [10]
Data lineage	C	C	C	C	C	C
On premises	C	C	C	N	C	P
Data intelligence	N	C	N	C	N	C
Data Governance	C	C	C	C	C	C
Knowledge graph-powered	N	N	N	N	N	C
Open Source	N	N	C	N	N	N
Health data	C	N	N	N	P	N
Total score	**10**	**10**	**9**	**8**	**9**	**10**

4. Discussion

Implementing a data catalog for the Clinical Data Warehouse (CDW) can hold significant potential for enhancing data science efficiency, particularly in the context of healthcare. The fundamental purpose of a data catalog is to facilitate data discovery, offering data analysts and IT specialists a comprehensive overview of available datasets.

Vendor comparison showed comparable results for all vendors. With the highest score of ten points and the lowest score of eight points, all vendors lie very closely together. Although every vendor fulfills the mandatory criteria “Data lineage” and “Data Governance”, they vary greatly among each other regarding the other requirements.

Presumably, every vendor specializes in one or more capabilities. For example, 'Compendium' has been developed especially for the healthcare sector, whereas 'Alation' is a key-player in the field of data intelligence. Thus, detailed investigations and, where appropriate, test installations are necessary to narrow the selection further down before making a final decision.

Given the multitude of commercially available data catalogs with varying capabilities, a crucial step in our approach was to conduct an explorative market survey. This survey not only served as a preliminary exploration but also laid the foundation for the subsequent system specification of the data catalog. The survey involved a meticulous comparison of vendors, considering their offerings against a predefined set of criteria.

Before initiating the market survey, it was imperative to define core capabilities for the data catalog. Recognizing that no single data catalog fulfills every criterion, we selected seven requirements based on a literature analysis outlined in Table 1. As the system selection process unfolds, these criteria will be further complemented by additional requirements to ensure a comprehensive evaluation.

It is essential to emphasize that the preliminary market survey and vendor pre-selection were instrumental aids for the subsequent system specification. Their primary purpose was to provide an initial market overview and facilitate the identification of potential candidates. However, a comprehensive and systematic market analysis is reserved for the post-conceptualization phase.

In the subsequent, more in-depth analysis, vendors that may have scored lower in the initial survey are subject to reconsideration. This approach allows for more nuanced evaluation, considering factors that might not have been immediately evident in the first survey. For instance, clarity on whether individual vendors exclusively operate on a cloud basis or also offer on-premises solutions might become apparent during a more detailed examination.

In summary, our decision to proceed with a data catalog for the CDW involves a methodical and phased approach, ensuring that system specification is informed by a thorough understanding of the market and tailored to the specific needs of our healthcare context. The iterative nature of our process acknowledges the dynamic nature of the market and the evolving requirements of our data management goals.

References

[1] S. M. Meystre et al., Clinical Data Reuse or Secondary Use: Current Status and Potential Future Progress I Introduction, (2017),

[2] C. Safran et al., JAMIA Perspectives on Informatics Toward a National Framework for the Secondary Use of Health Data: An American Medical Informatics Association White Paper, *Journal of the American Medical Informatics Association* **14**(1) (2007), 1–9

[3] E. Tute et al., Modeling of ETL-Processes and Processed Information in Clinical Data Warehousing., *books.google.com* **248** (2018), 204–211

[4] O. Olesen-Bagneux, *The Enterprise Data Catalog*. O'Reilly Media, (2023).

[5] The Data Catalog for Healthcare | Compendium. https://compendiumdatacatalog.com/ (accessed Mar. 12, 2024).

[6] Data Catalog – Tools und Lösungen | Informatica Deutschland. https://www.informatica.com/de/products/data-catalog.html (accessed Mar. 12, 2024).
[7] Amundsen, the leading open source data catalog. https://www.amundsen.io/ (accessed Mar. 12, 2024).
[8] Alation Data Catalog | Alation. https://www.alation.com/product/data-catalog/ (accessed Mar. 12, 2024).
[9] Data Lineage Done Right | Data Lineage Tool | Manta. https://manta.io/ (accessed Mar. 12, 2024).
[10] Data Catalog | data. world. https://data.world/product/data-catalog/ (accessed Mar. 12, 2024).