Rediscover chromatography data – Enabling data-science through creation of a chromatography results database

S.A. Pfeiffer[a], L. Berstler[b]

Agilent Technologies

[a] Hewlett-Packard-Str. 8, 76337 Waldbronn, Germany

[b] 2850 Centerville Rd, Wilmington, DE 19808, USA

HPLC 2023
P-MISC06

Agilent OpenLab

Introduction

Chromatography data an underutilized resource

Chromatography data play a major role in chemical, food, and pharmaceutical product development and are collected at various points throughout a product's lifecycle. These data are often gathered for specific needs, such as purity or yield calculations. They are then filed as an electronic document and are often not used again. This is unfortunate because chromatography data contain much more information than what they are typically reduced to. The chromatograms themselves, unknown peaks, mass spectrometry (MS) spectral data, to name a few, constitute a vast resource that could be used to generate more value in data science projects.

Obstacles – Extract transform load (ETL)

We found that the main challenge to utilize chromatography data lies in the process of making the data available and searchable to tools such as data analysis/visualization packages. In a typical situation a lab might find itself implementing a complex pipeline as shown in figure 1.

Challenges in ETL of chromatography data

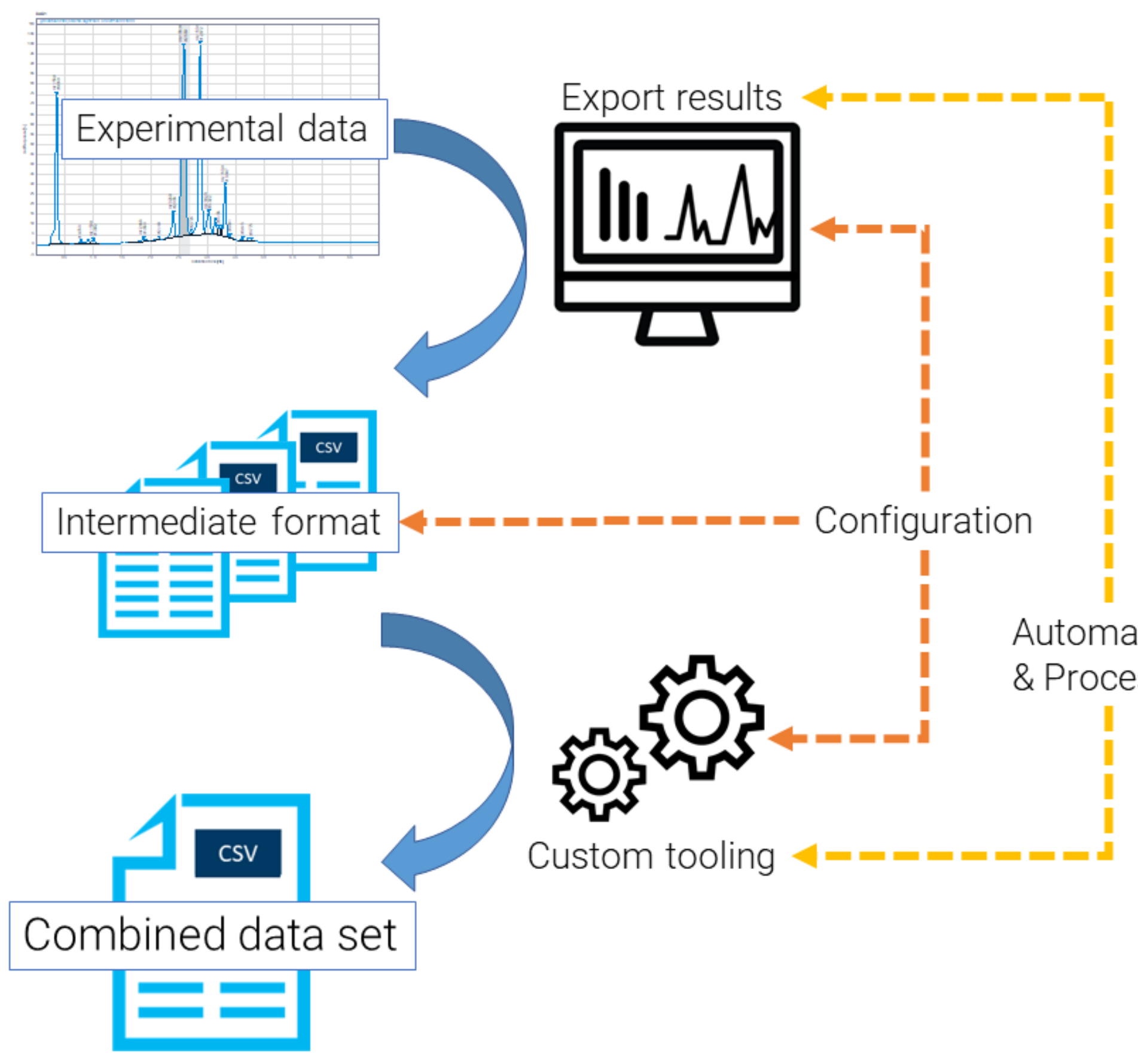


Figure 1. A typical generic approach to data extraction from laboratory instruments via built-in reporting capabilities to an intermediary format and aggregation with custom tooling.

While running such an export and aggregate workflow is universal and can be applied to a wide range of instrumentation it has a few drawbacks. First, it relies on precise prior knowledge of which parameters are of interest and requires configuration changes along the whole pipeline if these requirements ever change (e.g. additional peak parameter). Depending on the automation approach taken, handling manually processed data can be a challenge and requires operational rigor to prevent data duplication. Lastly, the workflow is prone to disruptions specifically in multi-instrument scenarios where the intermediate exports must be sent over the network.

Herein, we present a prototype application addressing the pain points we found in the generic, and similar, approaches to data extraction. We demonstrate the versatility of the approach by implementing different use cases and evaluated the database with regards to its ability to supply off-the-shelf data analytics tools directly.

Experimental

Data set

The data set used to outline the use cases consisted of a mix of chromatography data acquired for different purposes on a variety of instruments. The table below gives an overview over these data.

Data set overview	
Total # injections	17733
Total # of peaks	84242
Instrument technique split	59% LC, 40% GC, 1% Other
Acquired between	2008 - 2023

Software

All data were acquired with or reprocessed in OpenLab CDS 2.6/2.7 before storage in a OpenLab ECM XT data management system. Third party commercial visualization and business intelligence packages used to implement the uses cases were Spotfire (TIBCO) and PowerBI (Microsoft).

Experimental

Chromatography results database - prototype

The chromatography results database prototype was implemented based on PostgreSQL with accompanying services to connect to ECM XT, ingest the data, and maintain the most recent results. The operation of the results database is shown schematically in figure 2a. Data are collected from different instruments using a distributed setup of OpenLab CDS and the data are stored in the ECM XT server. The chromatography results database fetches the result sets and stores the data in a relational schema (Fig 2b). Finally, the consuming applications get access to the database via a structured query language (SQL) interface.

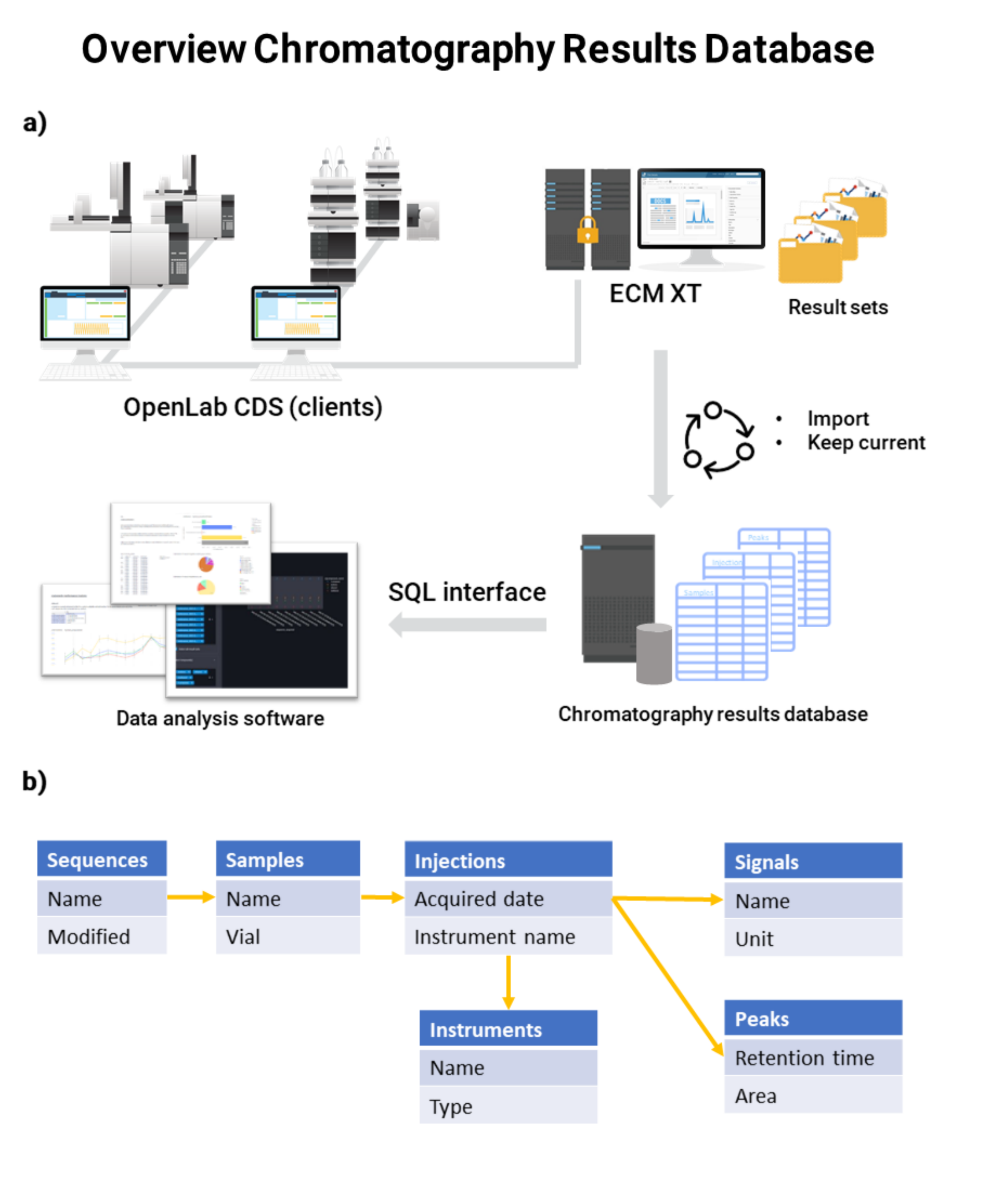


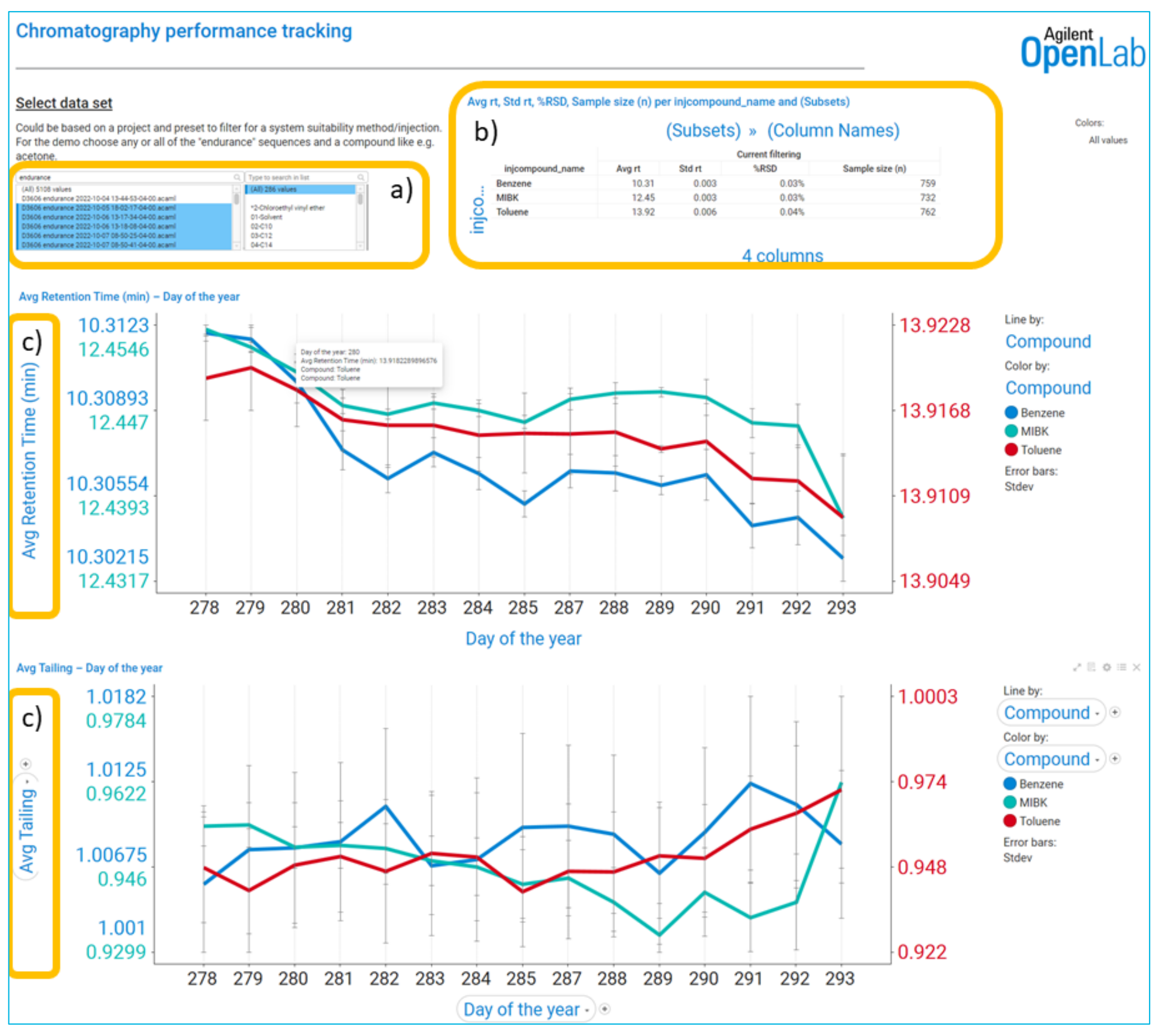
Figure 2. a) Overview of the proposed solution connected to an OpenLab CDS client server installation with ECM XT content management. b) Simplified schema of the chromatography results database with the main entities of interest.

Results and Discussion

We evaluated the interoperability of the database system with data analytics tools by implementing two common use cases and comparing it to the generic approach outlined in figure 1.

Use case – Time series charting

The dashboard depicted in figure 3 was developed using TIBCO Spotfire and the connector for PostgreSQL. To construct the dataset the graphical user interface (GUI) of the software was used and no knowledge of SQL was necessary.


Chromatography performance tracking

Select data set

Could be based on a project and preset to filter for a system suitability method/injection. For the demo choose any or all of the "endurance" sequences and a compound like e.g. acetone.

Avg rt, Std rt, %RSD, Sample size (n) per injcompound_name and (Subsets)

injcompound_name	Avg rt	Std rt	%RSD	Sample size (n)
Benzene	10.31	0.003	0.03%	759
MIBK	12.45	0.003	0.03%	732
Toluene	13.92	0.006	0.04%	762

Figure 3: Example dashboard charting peak properties over time. Yellow highlights interactive elements: a) Filter by result sets and compound names b) summary statistics based on selection c) label to choose desired properties

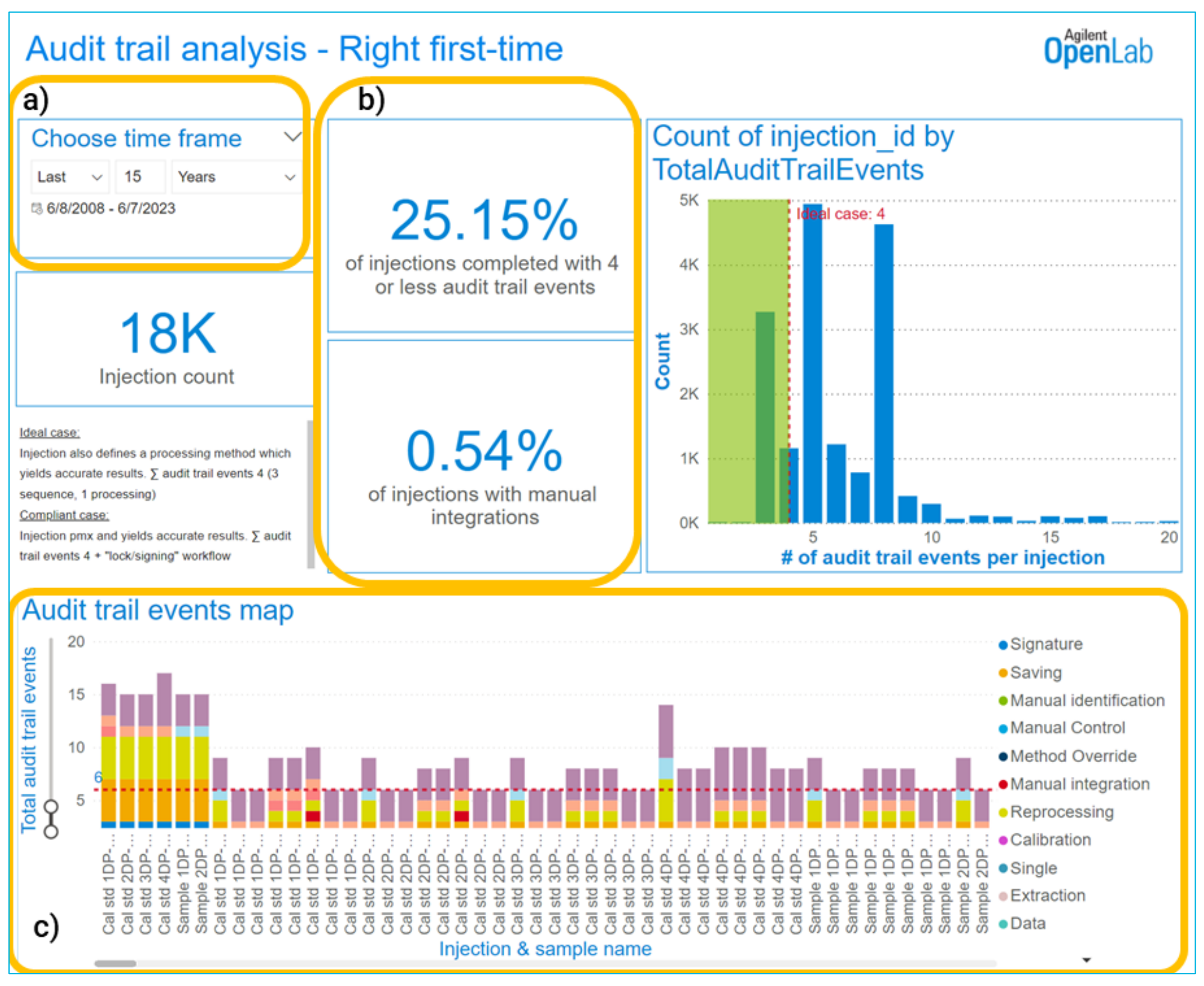
The dashboard allowed quick filtering of the whole database contents by all available properties such as file, compound, operator names, or time frames. It also proved to be very responsive and changes in plot axis refreshed in a few seconds. The biggest advantage we found was that any peak property calculated in OpenLab CDS can be accessed by merely changing the axis label in the dashboard. Via direct SQL connection, the visualization software takes care of requesting the correct data from the database.

Results and Discussion

In contrast to the approach of data extraction by file exports and custom tooling (Fig. 1), changing the data of interest did not require any reconfiguration of the system or re-export to make the property available.

Use case – Key performance indicators (KPI)

KPIs play a crucial role in analytical laboratories to assess the correct and efficient operation of the facility. Different measures such as time to result and sample throughput are routinely used to evaluate performance. We implemented a right first-time dashboard using the chromatography result database to add additional measures based on audit trail events data.


Audit trail analysis - Right first-time

Ideal case:
Injection also defines a processing method which yields accurate results. ∑ audit trail events 4 (3 sequence, 1 processing)

Compliant case:
Injection pmx and yields accurate results. ∑ audit trail events 4 + "lock/signing" workflow

Figure 4. Audit trail analysis a) A time range filter is used to narrow down the data set b) Key metrics such as injections under a certain threshold of events and important events such as manual integrations are highlighted in separate cards. c) The bar chart gives visual outline of events per injection.

The dashboard was constructed to summarize audit trail events per injection and implemented in Microsoft PowerBI solely using the GUI tools. A threshold was introduced at four events per injection for an ideal case of automatic data processing, the histogram shows the data distribution, and manual actions such as peak integration are highlighted enabling a quick overview over the operation of the lab. In contrast to the generic approach (Fig. 1) no changes to the data import were necessary to enable this dashboard and the same database supplied both use cases (Fig. 3, Fig. 4).

Compatibility with other data analysis tools

Apart from the specific use-cases shown, we also evaluated the compatibility of the results database with other common data analysis and visualization tools. Table 1 shows a list of the tested software that was found to be compatible.

Table 1: Tested compatibility with common data analytics software packages. All trademarks belong to their respective owners.

Type	Software package	Connection mechanism
Commercial	Excel	vendor recommended adaptor
	Spotfire	built-in connector
	PowerBI	
	Qlik	
	Tableau	
Open Source	Grafana	
	Metabase	

Conclusions

In summary, we found that the prototype results database performed well in interfacing with standard data visualization tools. Implementation of the outlined scenarios was possible solely through the user interfaces and did not require specific knowledge of SQL or custom code. More importantly no changes to the data pipeline were necessary to implement the two vastly different use cases. The versatility of the results database is the main advantage over a custom ETL pipeline and a large step forward in making chromatography data more accessible to data-science projects in organizations of any size. Future work with this prototype will entail extended benchmarking with larger data sets, expanded use cases, and deploying the system to cloud environments.

Contact

Simon A. Pfeiffer simon.pfeiffer@agilent.com
Lauren Berstler lauren.berstler@agilent.com

DE45353404