Structure first, analyze later

Centralized data platforms for managing biomedical research processes

Jonathan Alles, **EVO**BYTE Digital Biology

The three paths of research data

How much time do you spend searching for and reformatting experimental data? If this is a major bottleneck for you and your team you are probably in good company[1].

But it is actually quite simple! A research process consists of three steps: We run an experiment, then we measure what we are interested in, and finally we analyze and interpret the measurements. And then we do it again.

This leaves three paths of process data: There is metadata, which comprises protocol information, annotations or sample descriptions. There is raw data produced by our instruments and measurements, and there is processed or analysis data.

So why not just store everything in a database? Because in practice we are confronted with very different requirements for each data path. Metadata is often stored in separate lab notebooks, raw data is often large in size and difficult to store in a relational database, and processed data needs to be accessed frequently. The result is unstructured data silos which are hindering access and slow down analytics.

Structure first, analyze later

How can we avoid this? We can first connect raw and metadata in a central data store and then perform our (automated) data analysis based on structured inputs. This principle has several advantages: First, a single data store holds all information relevant to the research process and provides a single point of truth. Second, all functionalities can be implemented based on the data store, for instance data export for lab notebooks. Third, a central store can easily be backed-up and migrated, preventing data loss or corruption.

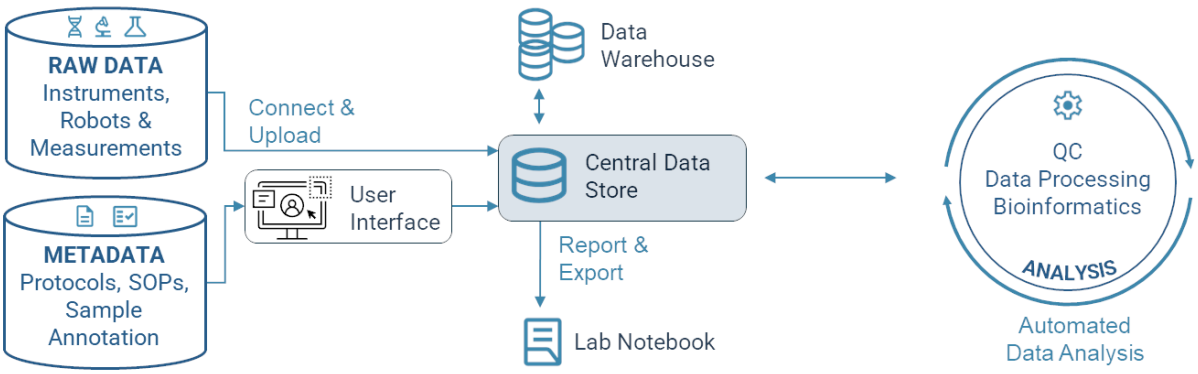

Figure 1 Central Data Store Architecture

[1] https://www.forbes.com/sites/gilpress/2016/03/23/data-preparation-most-time-consuming-least-enjoyable-data-science-task-survey-says/#437d25337f75

The **EVO**DATA Platform translates this concept into a flexible application for managing, storing and analyzing biomedical research data. The core functionality is implemented in the Process Data Layer, which is customized to fit your research process. It provides functionality to ingest instrument raw data and gives users an interface to upload sample annotations or experiment descriptions. The Data Layer then structures the inputs to link raw data with specific experiments or samples. Data processing and analysis steps are also integrated in the data layer and can be automatically run on new data. **EVO**BYTE also provides support in designing bioinformatic data pipelines for analysis of raw data.

Having a Process Data Layer in place enables integration of add-ons like automated reporting functions or interfaces to data warehouses to share data within your organization. All data is hosted on your on-premises or cloud infrastructure, guaranteeing data privacy and compliance.

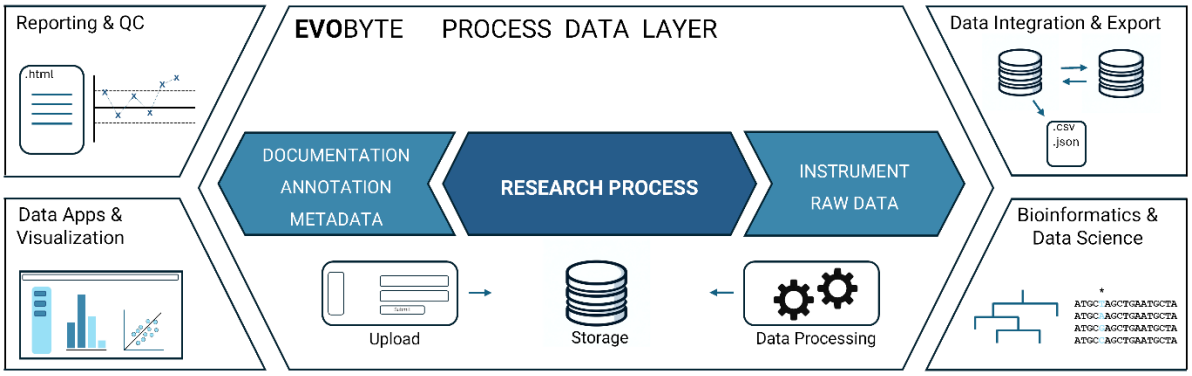

Figure 2 The **EVO**DATA Platform organizes process data and provides interfaces for export and reporting

Designed for individual processes

Seamless integration into your existing research process is most important, this is why each platform is customized to your requirements. As a first step, the research process is carefully analyzed, and all meta and raw data inputs are discovered. User interfaces are designed to enable simple upload of experimental data, while raw data can be automatically uploaded from your fileservers.

Data analysis pipelines are designed based on existing workflows or can be built from scratch. The modules for data analysis are open and flexible, allowing for rapid updates or deployment of new pipelines. Of course, the processed data is stored in the central data store and can be accessed based on metadata.

During a rollout period your research teams can test the platform and define changes to interfaces and functions.

An example from the genomics world

Let's assume we have a simple research process for single-cell RNA sequencing (scRNA-Seq) of blood cells to measure gene expression. Several times a week, blood cells from clinical trial patients are processed and analyzed by scRNA-Seq to investigate biomarkers related to a novel treatment. For each sample, the experiment ID, date, patient identifier, protocol version and trial information are recorded. Those make up the metadata. The process raw data is fastq files produced by next generation sequencing. Fastq files are

automatically uploaded into a data bucket once a sequencing run is completed, and metadata can be uploaded via a user interface.

To save time during analysis, data processing is fully automated. This includes a QC, alignment, quantification of gene expression and an integration step. Processed datasets are synchronized with the central data store and can easily be queried by experiment ID, patients or other metadata features. This allows data analysis to rapidly access the latest version of all data generated and easily manage updates to data pipelines.

A full presentation on the technical details for the genomics case study can be found on the website. If you want more information about the **EVO**DATA Platform please reach out.