

Phuse EU Connect 2025

DH06

FAIRchive

**Cloud-Based and GxP-Compliant
Long-Term Archiving for Clinical Trial Data**

Alexander Lüders | Frank Stoinski

Agenda

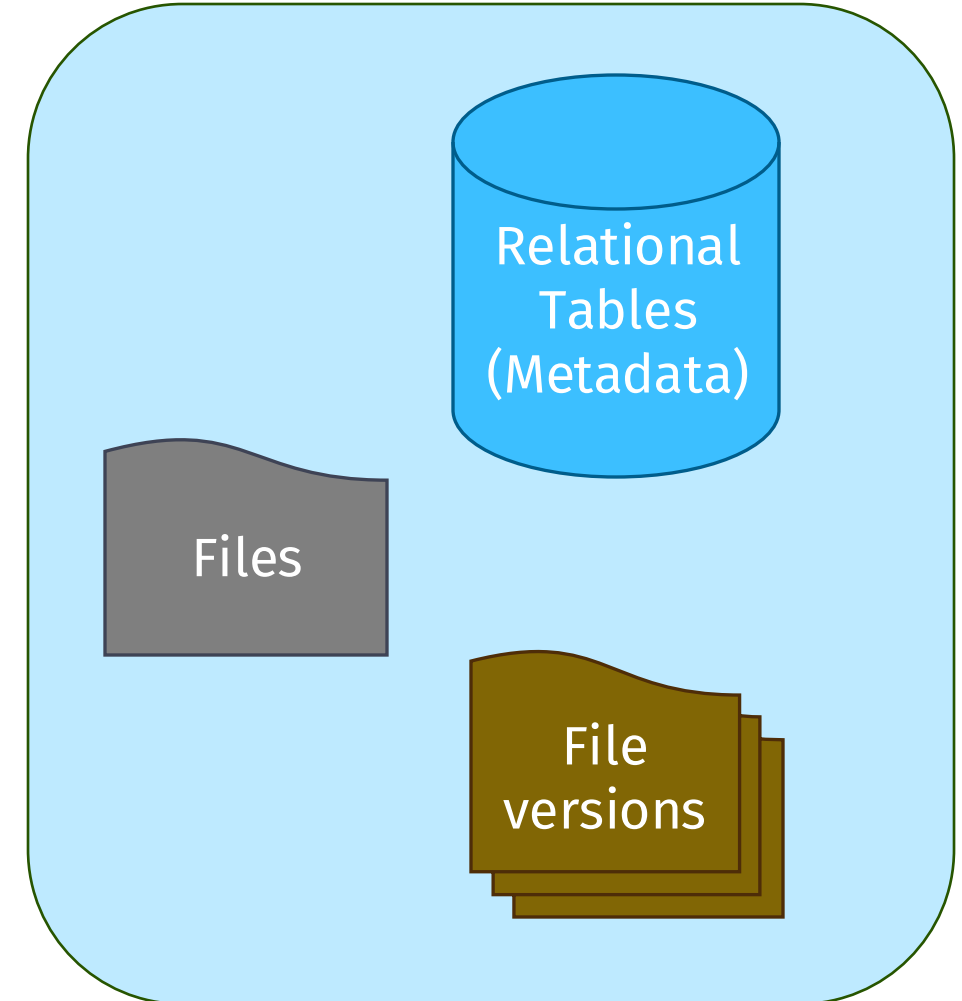
- Why FAIR Archiving?
- Overview Source System
- The Naive Approach
- Optimization and Tuning
- Is it FAIR?
- Accessing the Archive
- Summary and Conclusion

Why FAIR Archiving?

- System replacement or decommissioning
 - Original data not directly reusable without source system
 - Data needs to be accessible according to FAIR principle
 - Regulatory requests
 - Exploratory use cases
- We need cost-efficient, source-system-agnostic archive with
 - Few technology dependencies
 - Stable and open technologies

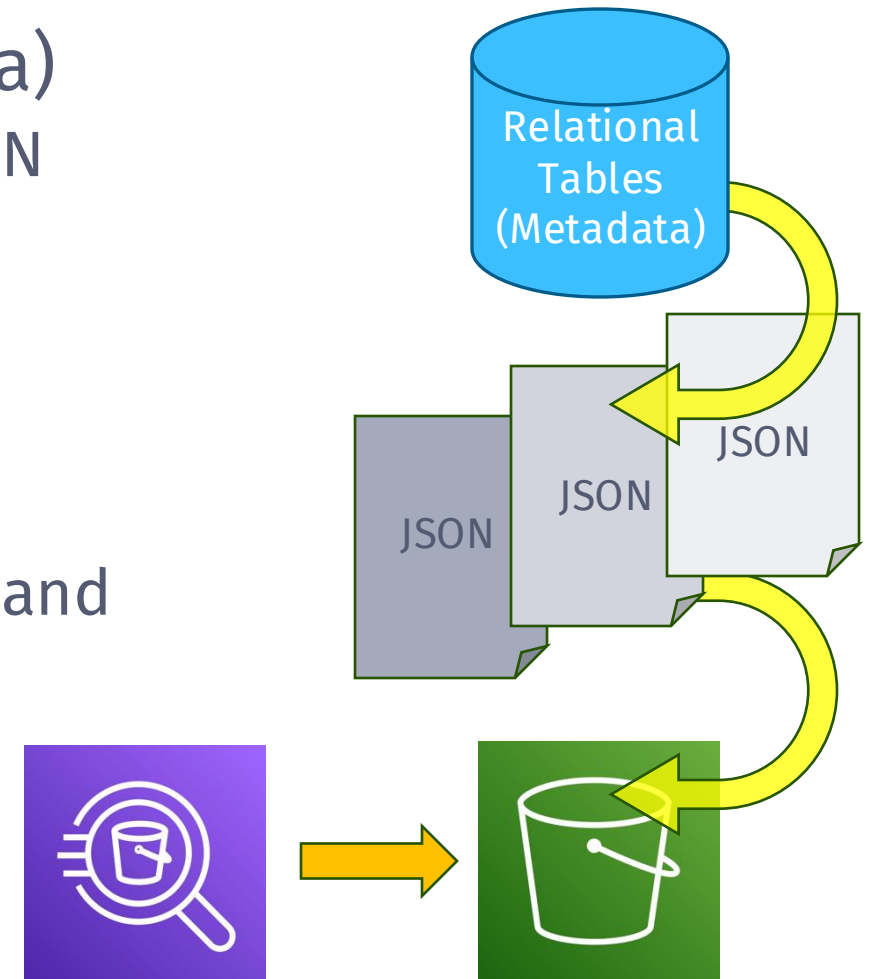
Overview Source System

- Parameters:
 - 1000 studies
 - 1 million files, 7 Tbyte filesystem, > 50 Tbyte file versions
 - 100 Gbyte relational data
- Data domains:
 - Object Search Domain
 - Permissions Domain
 - Audit Trail Domain
 - Action Log Domain
 - Execution Trace Domain



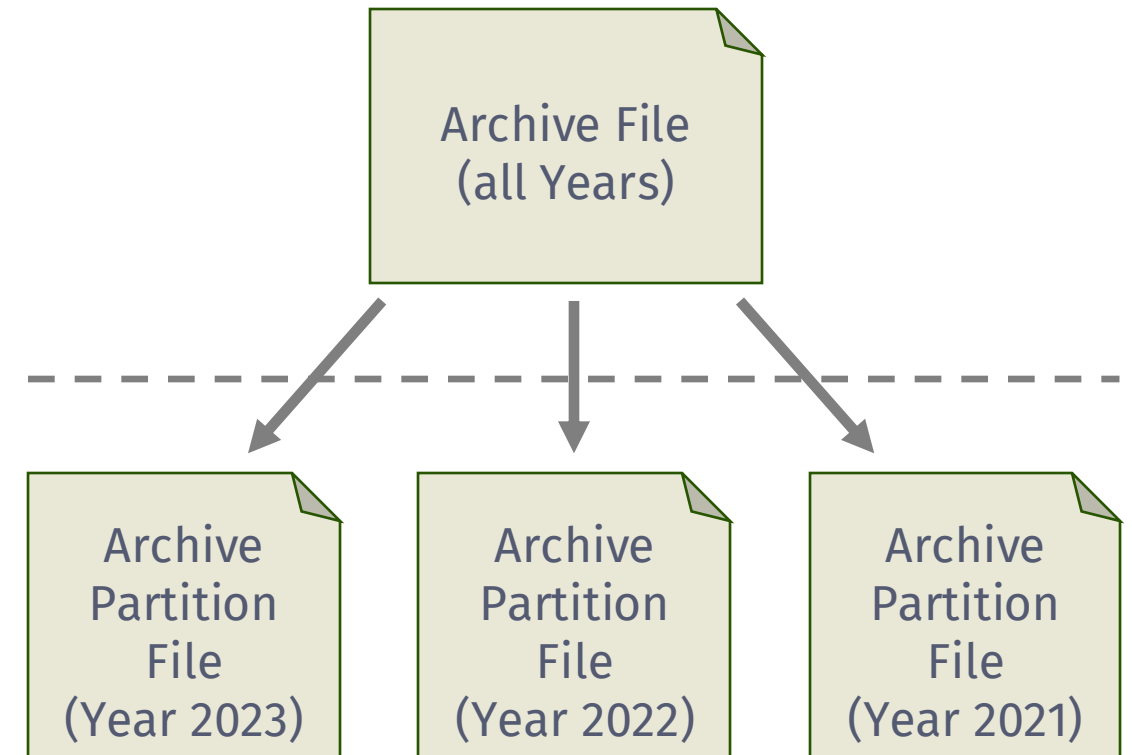
The Naive Approach

- Let's try a Cloud Solution (Amazon Athena)
 - Converting relational database tables to JSON
 - Pushing JSON tables to Amazon S3
 - Pushing binary files and revisions to S3
- Results were sobering
 - Query duration not acceptable
 - Athena is not a database engine (no indices and other query optimizations)
 - Less control over data volume to scan
 - e.g. Joins often result in full JSON table scans



Optimization and Tuning

- Apache Parquet Format
 - Native Support in Amazon Athena
- Partitions
 - Idea: Decrease the amount of scanned data
 - Identify attributes of the archive data that can be used to partition the data, e.g. timestamps
 - Athena can leverage partitions in queries



Accessing the Archive

Query 1 : ✕

✓ Query 2 : ✕

Query 3 : ✕

(+) ▼

1 SELECT path, creationtime FROM object_x_objectpath

2 FETCH FIRST 5 ROWS ONLY;

SQL Ln 1, Col 1

Run again

Explain [\[?\]](#)

Cancel

Clear

Create ▼

Reuse query results

up to 60 minutes ago [\[?\]](#)

Query results

Query stats

✓ Completed

Time in queue: 101 ms

Run time: 5.572 sec

Data scanned: 13.28 KB

Results (5)

[Copy](#)

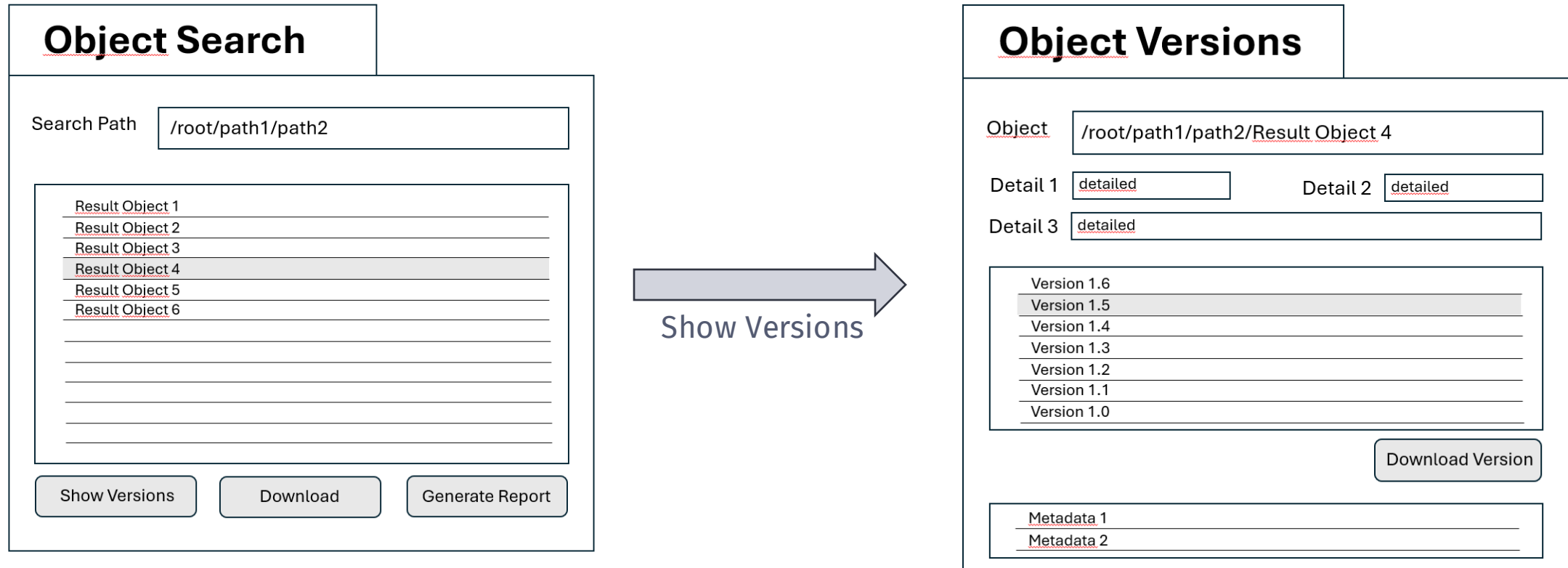
[Download results CSV](#)

🔍 Search rows

< 1 > ⚙️

#	▼	path	▼	creationtime	▼
1		/root/54F0/696744B/662133B6/report/tlf/dev/program/EEE4A6		2018-01-03T03:28:49Z	
2		/root/54F0/696744B/662133B6/report/0AF490277C9/dev/data/data_area/0D		2018-01-03T03:25:54Z	
3		/root/54F0/696744B/662133B6/report/0AF490277C9/dev/data/data_area/62EF19		2018-01-03T03:25:55Z	
4		/root/54F0/696744B/662133B6/report/0AF490277C9/dev/data/data_area/910CA3		2018-01-03T03:26:07Z	
5		/root/54F0/696744B/662133B6/report/0AF490277C9/dev/data/data_area/8D		2018-01-03T03:25:50Z	

Jupyter Notebooks



- Users search for objects based on paths
- Details of selected object shown in subordinate dialogs

Is it FAIR?

- FAIR principle developed in context of scientific archives
 - **F**indable
 - **A**ccessible
 - **I**nteroperable
 - **R**eusable
- **Findable** as metadata of source system will be archived
- **Accessible** and **Reusable** as file and revisions can be located and downloaded from S3 with standard browsers
- **Interoperable** as JSON/Parquet is used as data format for metadata

Summary and Conclusion

- The developed approach for setting up an archive for clinical trial data in the cloud is valid
- Few technology dependencies: serverless implementation
- Stable and open technologies: JSON and Parquet
- GxP compliance with AWS configurations

Summary and Conclusion

- FAIR principles acknowledged
- Solution prepared for more source systems
- System integration enabled through Amazon API
- Migration path out of the cloud to different solution remains open

Questions and Answers

Contact Details

Alexander Lüders
Entimo AG
alu@entimo.de

Frank Stoinski
Entimo AG
sto@entimo.de