



Validation Challenges and Insights: Data Migration during Integration

Sean Lacey & Jing Yu

Boston Development Group, Novo Nordisk A/S

07Sep2023

Views and opinions expressed are those of the speaker and not necessarily Novo Nordisk

Agenda

1. Introduction
2. Objectives
3. Data Migration: Key Challenges
4. Migration tool - Filigator
5. Initial migration Plan
6. Strategies for Overcoming Data Migration Challenges
7. Conclusion

Introduction

- Dicerna Pharmaceuticals was acquired by Novo Nordisk at the end of 2021.
- Novo Nordisk is a global healthcare company specializing in developing, producing, and distributing pharmaceutical products and services. Their primary Therapeutical Areas (TAs) are diabetes, obesity, cardiovascular, rare blood, endocrine disorders and rare disease.
- The acquisition process posed various challenges for Dicerna during the integration phase. One of the specific challenges, explored here, is the data migration process.

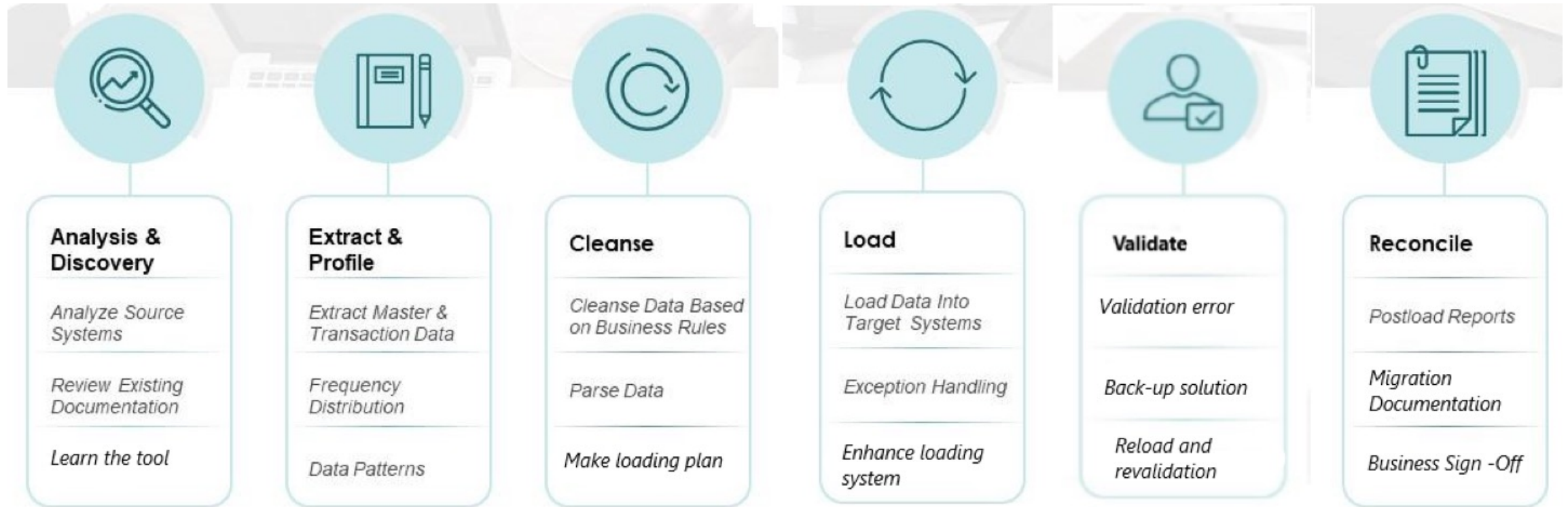
Introduction

— the necessity of data migration

- Integration of System
- Data Consolidation
- Data Quality and standardization
- Compliance and Reporting
- Cost Efficiency
- Efficiency and Productivity
- Scalability

Objective: The SCE in the Source and Target System

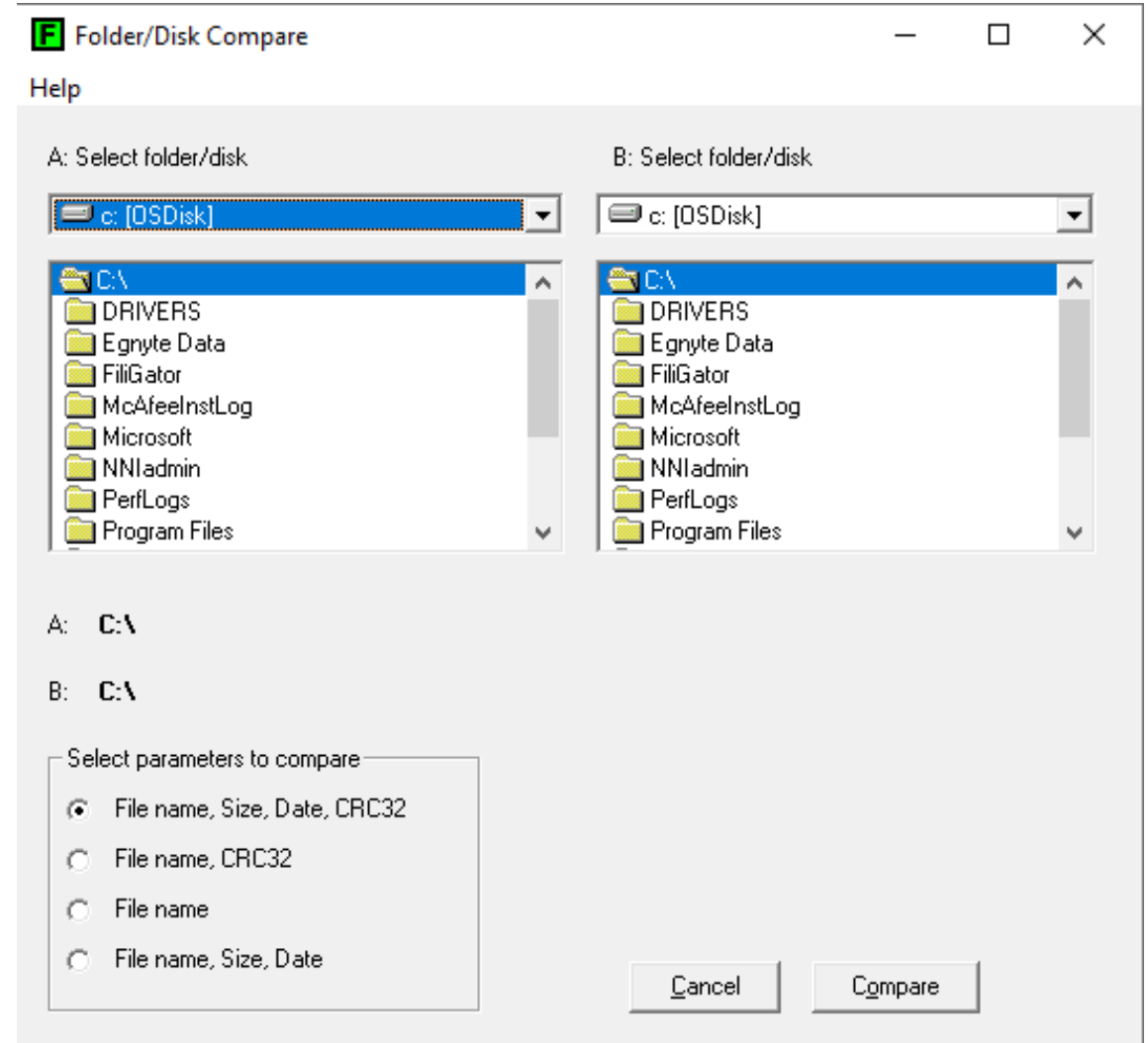
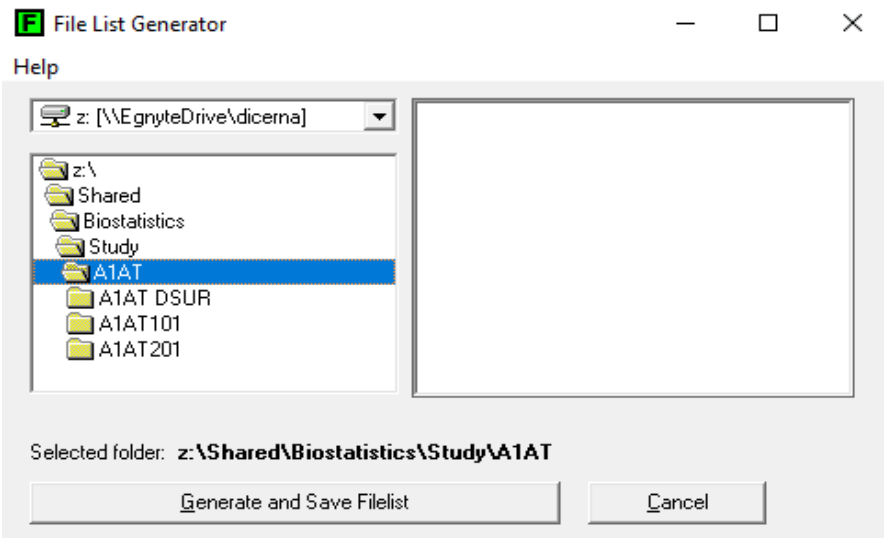
Objective: Data Migration Step Process



Data Migration: Key Challenges

- **Compatibility and Integration:** Ensuring that the new system is compatible with existing applications and workflows can be difficult. Integration issues may arise, requiring additional development efforts.
- **Data Quality and Integrity:** Data migration requires maintaining data quality and integrity throughout the process. Issues such as incomplete or inconsistent data, duplicate entries, and data corruption can occur during migration, affecting the reliability and accuracy of the transferred data.
- **Downtime and Disruptions:** Data migration often involves downtime during which the systems may be inaccessible or operating in a limited capacity. Minimizing downtime and avoiding disruptions to business operations is a significant challenge.
- **Testing and Validation:** Thoroughly testing the migrated data and validating its accuracy is a time-consuming process. Ensuring that all data dependencies and relationships are preserved is crucial.
- **Data Volume and Complexity:** Large volumes of data and complex data structures can make migration more challenging and time-consuming. It may require specialized tools and technologies to handle the data effectively.
- **Business Process Alignment:** Aligning the new system with existing business processes and workflows may require changes to standard operating procedures, which can be met with resistance.
- **Data Governance and Compliance:** Ensuring that data migration adheres to regulatory requirements and industry standards can be challenging, especially in heavily regulated industries.

Migration Tool - Filigator



Filigator Features

- **File List Generator**

1) Generates .txt file containing the following variables in a comma delimited format:

- Name, Path, Size, Date, CRC32
- CRC32 (Cyclic Redundancy Check) is a checksum algorithm that produces a 32-bit value from input data, commonly used to verify data integrity in files and network communications.

 ZDrive_Filelist_20230526 - Notepad

File Edit Format View Help

"Name","Path","Size","Date","CRC32"

"28 Nov 2022 DCR-A1AT-201 Protocol Deviation Plan.pdf","..\A1AT201\Document\Protocol\","311741","15-05-2023 17:33","D14B0C24"

"DCR-A1AT-201_protocol_V3.0.pdf","..\A1AT201\Document\Protocol\","1781503","04-12-2022 00:20","DB4D3B2"

"Dicerna Protocol DCR-A1AT-201 Version 3.0 24-May-2022.pdf","..\A1AT201\Document\Protocol\","2045344","04-12-2022 00:58","13F8FB7F"

"Dicerna Protocol DCR-A1AT-201 Version 3.0 USA-108-Aug-2022.pdf","..\A1AT201\Document\Protocol\","1236053","04-09-2020 08:46","50DC6F26"

"Protocol Deviation Plan.pdf","..\A1AT201\Document\Protocol\","265681","05-01-2023 19:58","680D603D"

"Randomization Plan.pdf","..\A1AT201\Document\Protocol\","741604","05-01-2023 19:59","6F5B37C"

"28 Nov 2022 DCR-A1AT-201 Protocol Deviation Plan.pdf","..\A1AT201\Document\PD\","311741","15-05-2023 17:33","D14B0C24"

"INVESTIGATOR BROCHURE_04-AUG-2022.pdf","..\A1AT201\Document\IB\","2628921","04-12-2022 01:04","BDF59D0D"

"INVESTIGATOR BROCHURE_13-DEC-2021.pdf","..\A1AT201\Document\IB\","2032700","04-12-2022 01:00","E47BE00E"

"INVESTIGATOR BROCHURE_25-MAY-2022.pdf","..\A1AT201\Document\IB\","2692567","04-12-2022 01:02","C745DB88"

Filigator Features

- **File List Generator**

2) Generates PDF file that summarizes directory information

File list generated for input folder/disk:

```
z:\Shared\Biostatistics\Study\PHXC\PHXC204\  
  
Creation date .....: 03-Jan-2023 11:57:48  
Total number of files ..: 3346  
Folders .....: 244  
Accumulated file size ..: 1.86 GB  
  
Volume label (disk name for z:) .....: dicerna  
Volume serial number (disk id for z:) ..: 12345678  
Share name (network share name for z:) .: \\EgnyteDrive\dicer  
~na
```

File list specifications:

```
Name .....: Filelist.txt  
Path .....: Z:\Shared\Biostatistics\Study\Migra  
~tion file\PHXC\PHXC_204\phxc204_s\  
Size .....: 348.27 KB  
Creation date .....: 04-Jun-2023 23:40:36  
Attributes .....: Read-only  
CRC32 .....: 3821D5B7  
  
Volume label (disk name for Z:) .....: dicerna  
Volume serial number (disk id for Z:) ..: 12345678  
Share name (network share name for Z:) .: \\EgnyteDrive\dicer  
~na
```

```
File list generation finished: 04-Jun-2023 23:40:50  
Elapsed time (hh:mm:ss).....: 00:21:48.8
```

```
File list generated by JYU on LCUSLXPF47S4Y7
```

Date and signature

Filigator Features

- **Folder/disk Compare**
 - Scans two directories and compares the files and folders on one of four parameter lists:
 - File Name, Size, Date CRC32
 - File Name, CRC32
 - File Name
 - File Name, Size, Date
 - If a mismatch is found, will stop processing and report the mismatched file or folder.

SUCCESS - The files in folder/disk A and B are identical.

A: z:\Shared\Biostatistics\Study\PHXC\PHXC104
B: p:\nn7022\general\current\stats\data\migration\PHXC104

File parameters compared:
Names, Sizes, Creation/modification dates and CRC32 values

Folder A:
Creation date: 16-Sep-2020 09:40:26
Volume label (disk name for z:): dicerna
Volume serial number (disk id for z:) ..: 12345678
Share name (network share name for z:) ..: \\EgnyteDrive\dicer
~na

Folder B:
Creation date: 16-Sep-2020 09:40:26
Volume label (disk name for p:): projstat
Volume serial number (disk id for p:) ..: EA8B1E94
Share name (network share name for p:) ..: \\whale.novo.dk\pro
~jstat

Properties for folder A respectively B:

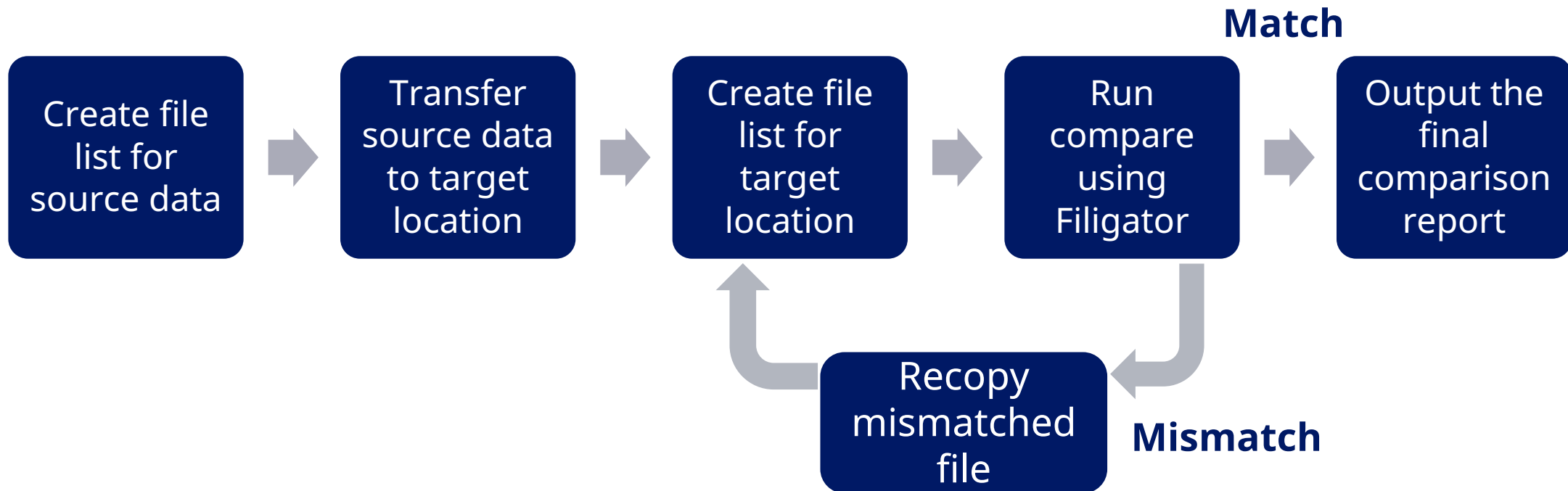
Total number of files ..: 1575
Folders: 128
Accumulated file size ..: 631.93 MB

Folder/disk compare finished: 05-Jun-2023 21:31:22
Elapsed time (hh:mm:ss) ..: 01:10:26.02

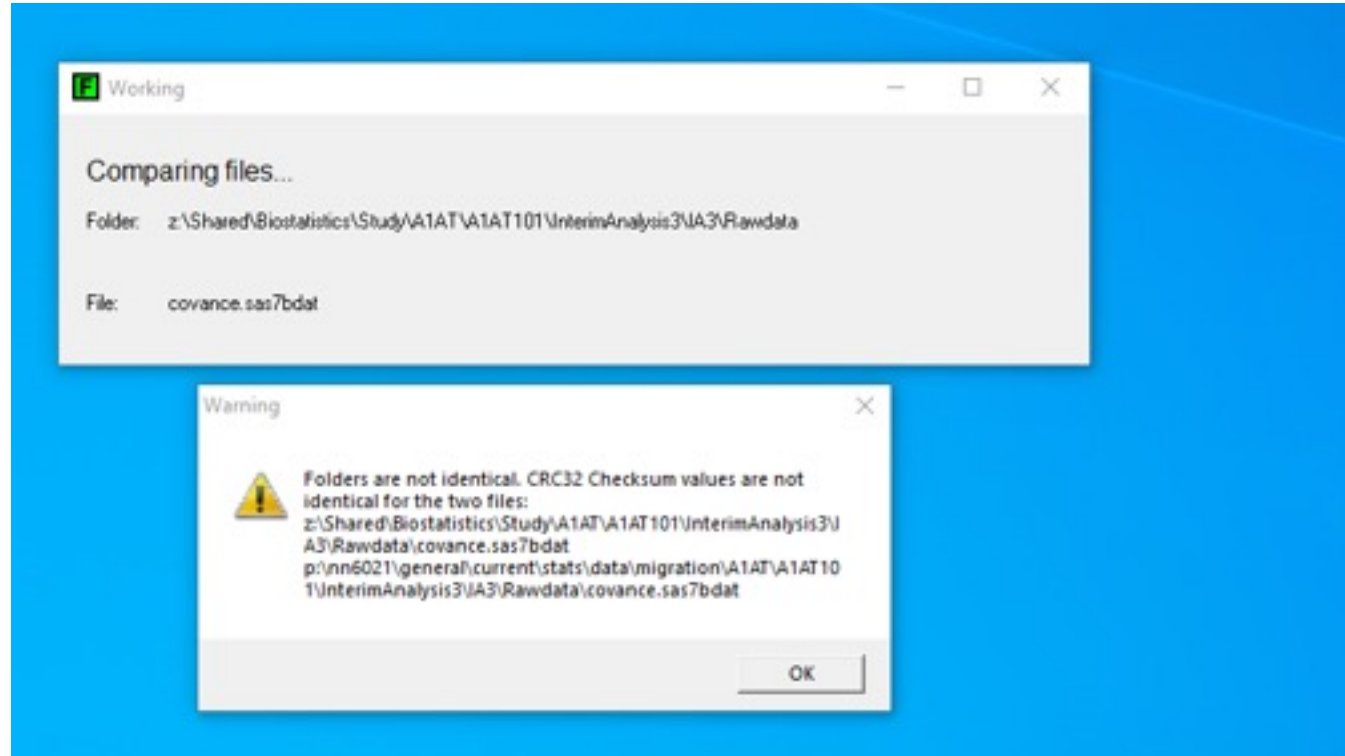
Folders/disks A and B compared by JVIYU on LCUSLXPF47S4Y7

Process and Challenges

Initial Plan



Possible Reasons for Mismatches



- Network connection between Denmark and Boston can be slow.
- Questions about reliability of CRC32 values
- Filigator "File/disk Compare" only reports the first mismatch it finds

Solution 1 : Compare File Lists Manually

- Use the File List Generator lists to identify all mismatches
- Create R code that compares the generated file lists, and R Markdown to output results

Filigator Filelist Compare

Properties of Source Folder

Location	Z:/Shared/Biostatistics/Study/PHXC/PHXC101
Creation Date	11/11/2019 11:44 AM
Volume Label	dicerna
Volume Serial Number	1234-5678
Number of Files	6007
Number of Folders	298

Properties of Destination Folder

Location	P:/nn7022/general/current/stats/data/migration/PHXC101
Creation Date	06/05/2023 01:38 AM
Volume Label	projstat
Volume Serial Number	EA8B-1E94
Number of Files	6007
Number of Folders	298

Solution 2 : Use MD5

- MD5 is a hash function that produces 128-bit hash values.
 - Traditionally used as a cryptographic hash function
 - Can be used as a checksum to verify data
- Hope that by using a more robust algorithm, we would better be able to ensure file integrity.
- R Code updated
 - CRC32 mismatches identified and further compared using MD5
 - Matching on MD5 seen as confirmation of file integrity

Solution 3 : Automated copying

- Additional concern that Dicerna cloud storage provider was adding metadata to files, resulting in mismatches
 - These updates might be happening in time between data copying and running file compare
- R Code updated
 - If files mismatch on MD5, the file is automatically recopied and MD5 is rechecked.
 - This happens one time to avoid any potential infinite loops

File 3 : CRC32 Mismatch

Source File	Z:/Shared/Biostatistics/Study/PHXC/PHXC101/Final folder/eTMF_B/Raw datasets/External Vendor Data/Axo Lab/DCR-PHXC-101 PK Plasma Sent 10Jan2020_Ver01.zip
Destination File	P:/nn7022/general/current/stats/data/migration/PHXC101/Final folder/eTMF_B/Raw datasets/External Vendor Data/Axo Lab/DCR-PHXC-101 PK Plasma Sent 10Jan2020_Ver01.zip
Source MD5	8277e190954814953548a834df66ea5c
Destination MD5	ac65b2270aa48647790ff5793634a91c

The MD5 values do not match. Recopying file...

Source MD5	8277e190954814953548a834df66ea5c
Destination MD5	8277e190954814953548a834df66ea5c

The MD5 values match

Creation of Electronic Record

- After validating data transfer, an electronic record is created using an internal tool
 - Electronic Record: Novo Nordisk term to indicate a folder instance that is set in an archival mode (cannot be changed)
- The record contains:
 - All data transferred from source
 - Migration documentation
 - Any file lists and code used to validate the migration
- Only at that point, is legacy Dicerna code/data available for use in the Novo Nordisk system

Conclusion

- importance of the data migration and its positive impact on the organization. how the migration will add value to the organization.
- Highlight the anticipated benefits of the data migration, such as improved efficiency, cost savings, enhanced data analytics, etc.
- Explain how communication will be managed throughout the migration process.
- Identify key points of contact for any questions or concerns.
- Discuss the training and support plan for end-users and stakeholders who will interact with the new system.
- Mention any resources or documentation that will be provided to assist users during the transition.

The background features a dark, textured surface with a faint grid pattern. A prominent horizontal band of teal color spans the middle of the image. On the left side, there is a large, stylized diamond shape composed of multiple overlapping outlines. Inside this diamond, a photograph of a city skyline is visible, showing buildings and a bright sky. The text "THANK YOU" is centered within the teal band in a white, bold, sans-serif font.

THANK YOU

Appendix 1: R Code Information

- Libraries used
 - tidyverse: Provides updated framework for data manipulation
 - digest: Calculates MD5 values
 - flextable: Used to create tables for output
 - officer: Used along with flextable
 - fs: File transfer library that updates the R default file.copy function
 - RMarkdown: Output report