Managing dataflows in a research hospital

Managed Research Data Management with iRODS

NETHERLANDS
CANCER
INSTITUTE
ANTONI VAN LEEUWENHOEK

The Netherlands Cancer Institute

The Netherlands Cancer Institute comprises an internationally acclaimed research institute as well as a dedicated cancer clinic. This combination ensures rapid translation of basic research into clinical applications: today’s research for tomorrow’s cure.

NETHERLANDS CANCER INSTITUTE
ANTONI VAN LEEUWENHOEK

Today’s research for tomorrow’s cure

We believe that cancer does not need to be a deadly disease. Our researchers and doctors are highly motivated to make this vision a reality by unraveling the biology of cancer and using this knowledge to improve the prospects of cancer patients. They conduct innovative and excellent fundamental, translational, and clinical research. Through close collaboration between lab and clinic, we create maximum impact for cancer patients.

The challenge

- At the Antoni van Leeuwenhoek hospital and Netherlands Cancer Institute (NKI-AVL) the demands for data storage and data management are rapidly evolving. **Departments in our institute increasingly integrate their data acquisition and analysis, sparking interdisciplinary research projects**. Furthermore, national and international regulations require researchers to **make their data FAIR** (Findable, Accessible, Interoperable, Reusable). Also, the development of the "-omic" techniques, such as genomic and proteomics, massively **increases the size of acquired data**. After analysis, this data should be **archived for longer periods of time** (>10y). Storing this data on rapid and available storage is **a waste of resources and money**. All these developments necessitate **meta-data driven data management**.

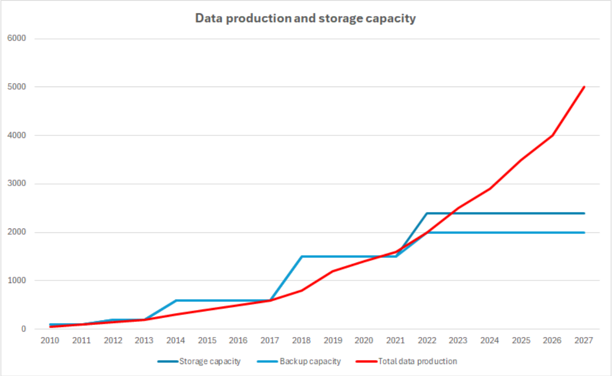
RDM tools available

My Documents

Address: My Documents

Name	Size	Type	Date Modified
Camtasia Studio		File Folder	4/24/2006 12:04 PM
comps		File Folder	3/29/2006 10:35 AM
My Captivate Projects		File Folder	4/12/2006 9:42 AM
My Music		File Folder	2/8/2006 2:18 PM
My Notebook Content		File Folder	4/4/2006 4:33 PM
My Pictures		File Folder	4/18/2006 7:20 PM
My Videos		File Folder	4/11/2005 7:58 PM
New Folder		File Folder	4/4/2006 7:31 PM
SnagIt Catalog		File Folder	3/28/2006 10:17 AM
te_win		File Folder	4/12/2006 8:37 AM
Trials		File Folder	4/24/2006 9:21 AM
USF Preschool		File Folder	3/28/2006 1:31 PM
Aware Final	4 KB	Easy CD Creator Pr...	4/22/2005 10:08 AM
bookmark	5 KB	HTM File	3/30/2006 10:23 AM
comps	343 KB	Compressed (zippe...	3/29/2006 10:35 AM
CursorXPPlus131	3,819 KB	Application	4/18/2006 9:54 AM
Kirkpatrick	21 KB	Microsoft Word Doc...	3/30/2005 6:44 PM
part6	2,468 KB	Compressed (zippe...	1/20/2006 9:56 AM

Data production vs capacity



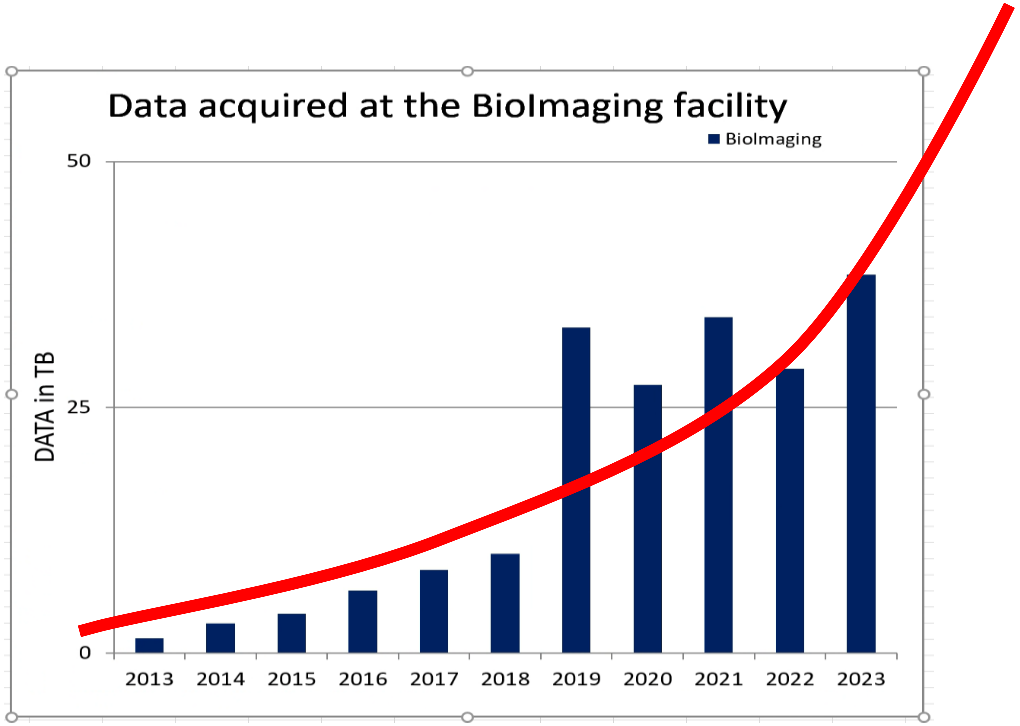
Cost and budget

Hardware and financial investment.

(Data) Interaction between clinic and research

Research data (storage) challenges

- The digital era is well under way
- More and more data is being generated
  - Size and number of the data is increasing
  - Data is producing new data (fueling itself)
- How to keep data findable and reusable data (FAIR)?
- Were do we put the data at what cost
- Is data safe and can we find the data?
- How do we share data (safe and controlled)



Data production is increasing rapidly

Looking at the data; a research perspective

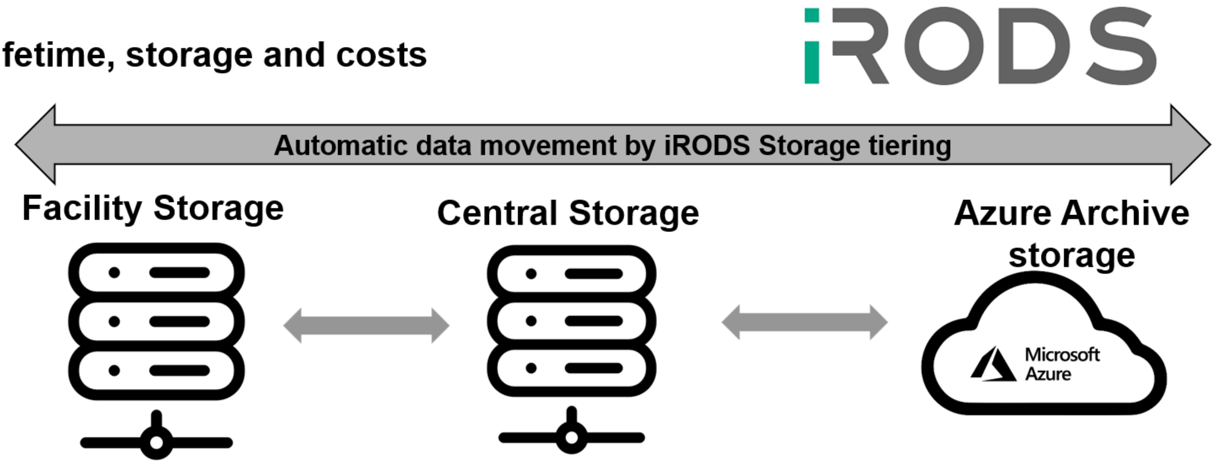
- RDM not a priority for researcher (just wants to store data)
  - (Manual) RDM distracts from research
- Data is not/poorly structured & described
  - Finding the data is hard/data gets lost
  - Do I have the correct data?
  - Poor to no insight in data linage
- Research outcome and reproducibility?
- Sharing data?
- Data automation is hard

Looking at facts of the data

- Data production and growth is (becoming) exponential
- Data is never deleted
- Big differences in data between research groups and facilities
  - Facilities are data producers, groups are data users
- 80 % of data is not being used any more (cold data)
  - Data becomes cold after ~6 months
- Why do we keep old/cold data on expensive storage (online)
  - and backup old/cold data in operational backup?
- Data is lost (not to be found)

iRODS solution for storage

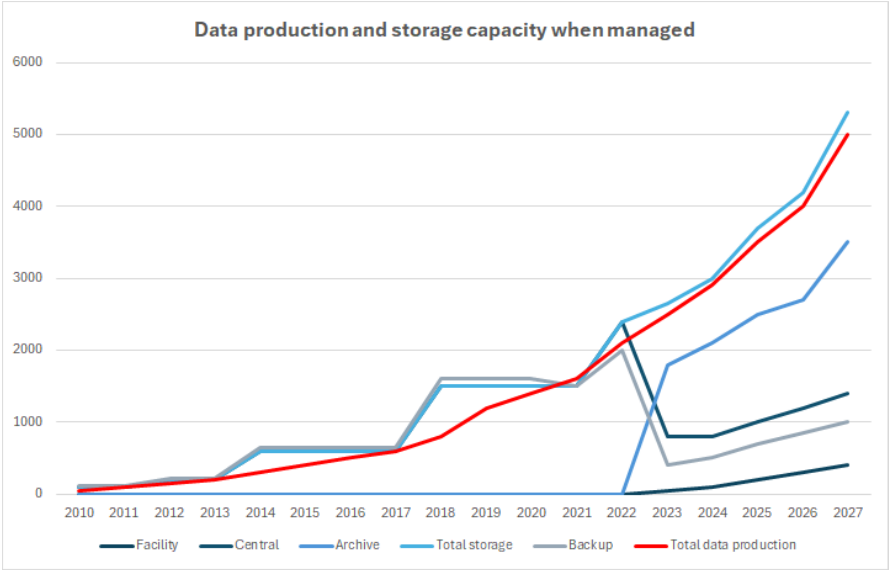
Data lifetime, storage and costs



Automatic data movement by iRODS Storage tiering

Facility Storage	Central Storage	Azure Archive storage
Acquisition Data	**Project Data**	**Archive Data**
6M-1YR (Fast storage)	6M-4YR (Slow storage)	Indefinite (offline storage)
Cost TB: **€€€**	Cost TB: **€€**	Cost TB: **€**
Available TB: **+**	Available TB: **++**	Available TB: **+++**

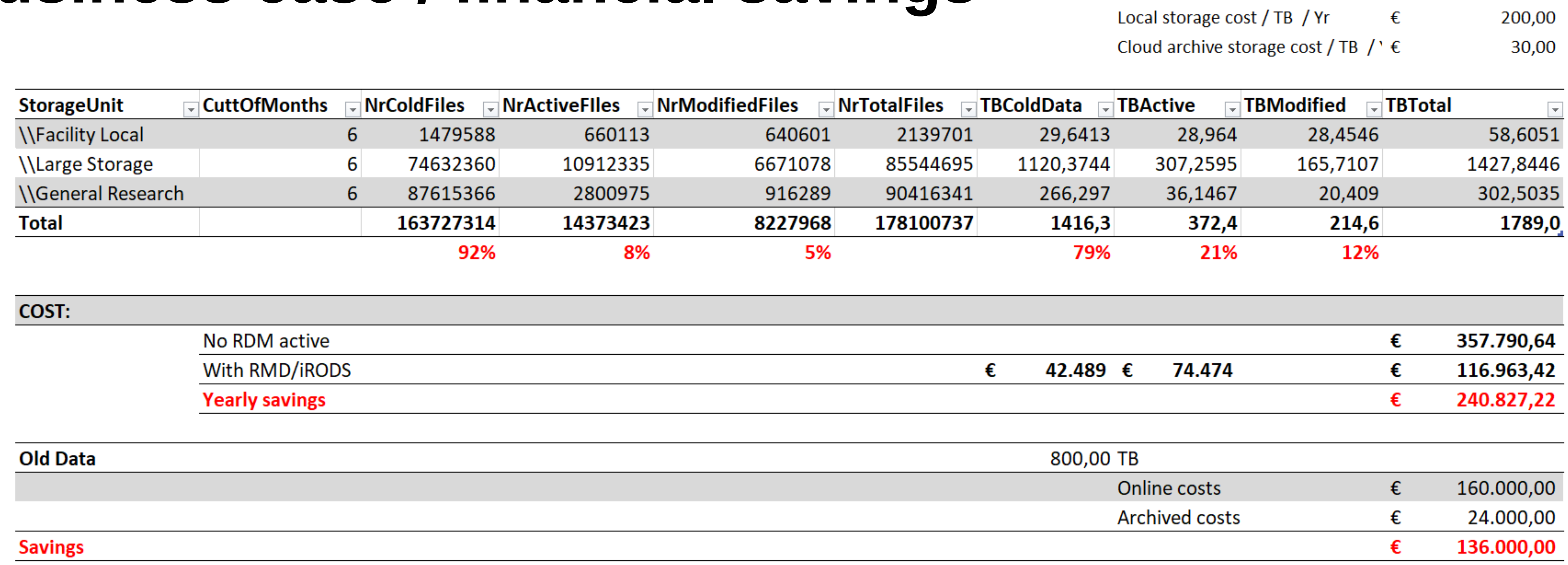
** Data policies meta data driven and flexible*



Managed data

- Manage storage on lifetime and usage
- Implement storage at the correct location and store data where it’s needed (onlive vs archive)
- Data archiving is automated, flexible and transparent for end user
- Annotate data with meta data so it can be found
- Multiple way’s/API’s to access data

Business case / financial savings



Local storage cost / TB / Yr	€	200,00
Cloud archive storage cost / TB / '	€	30,00

StorageUnit	CuttOfMonths	NrColdFiles	NrActiveFIles	NrModifiedFiles	NrTotalFiles	TBColdData	TBActive	TBModified	TBTotal
\\Facility Local	6	1479588	660113	640601	2139701	29,6413	28,964	28,4546	58,6051
\\Large Storage	6	74632360	10912335	6671078	85544695	1120,3744	307,2595	165,7107	1427,8446
\\General Research	6	87615366	2800975	916289	90416341	266,297	36,1467	20,409	302,5035
Total		**163727314**	**14373423**	**8227968**	**178100737**	**1416,3**	**372,4**	**214,6**	**1789,0**
		92%	8%	5%		79%	21%	12%	

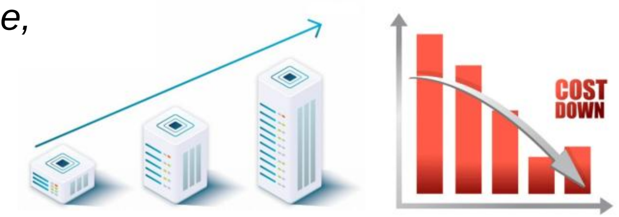
COST:					
	No RDM active			€	**357.790,64**
	With RMD/iRODS	€ **42.489**	€ **74.474**	€	**116.963,42**
	Yearly savings			€	**240.827,22**

Old Data		800,00 TB		
			Online costs	€	160.000,00
			Archived costs	€	24.000,00
Savings				€	**136.000,00**

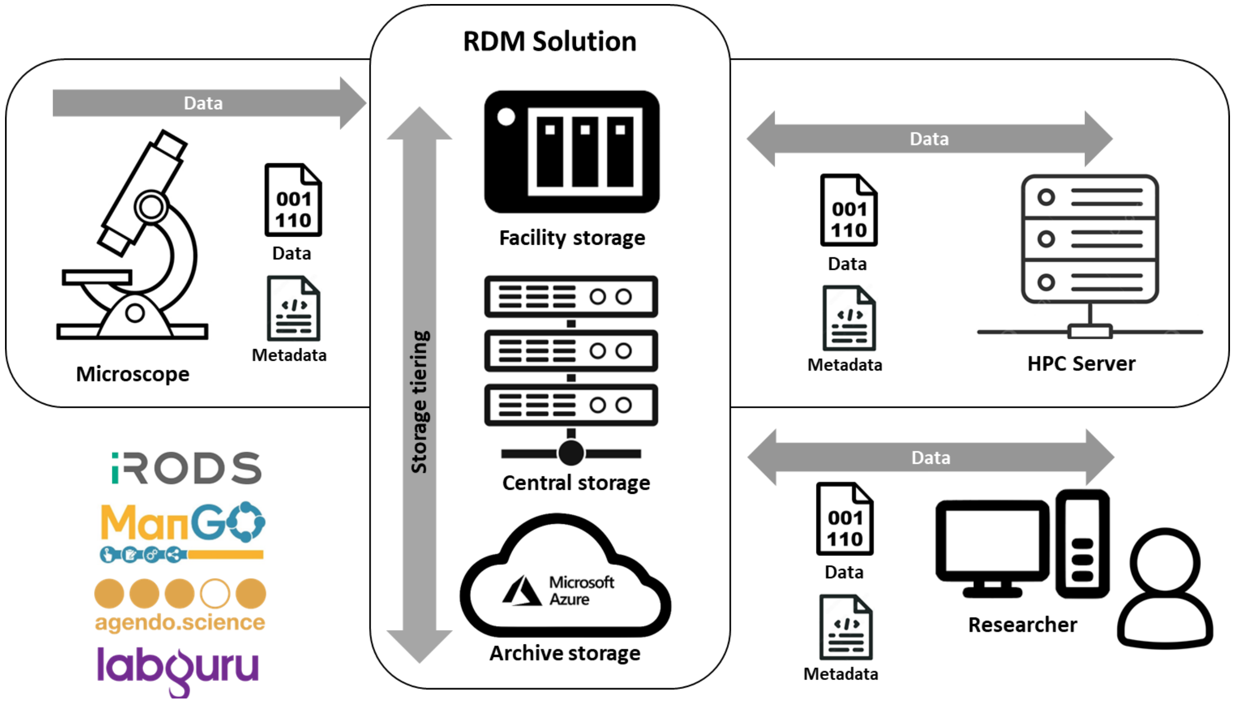
Total yearly savings € 376.827,22

Business-case to support RDM/iRODS:

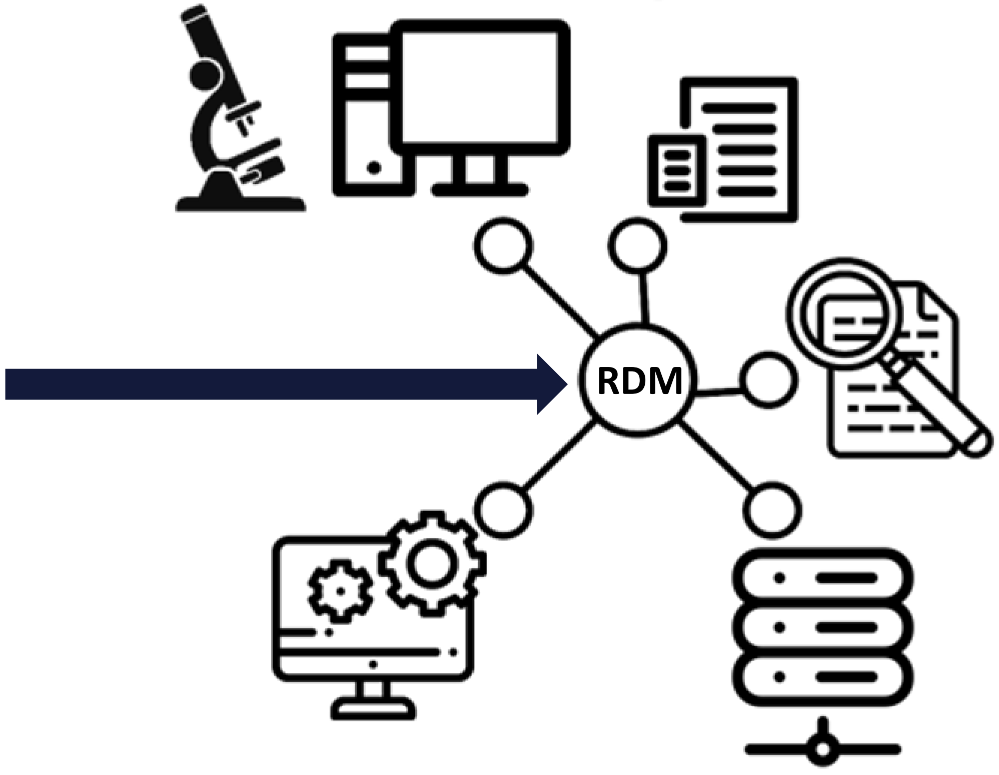
By correctly scaling and managing storage, financial investments in IT are reduced and research can be kept going



RDM infrastructure for Bio Imaging facility



Local Facility Environment

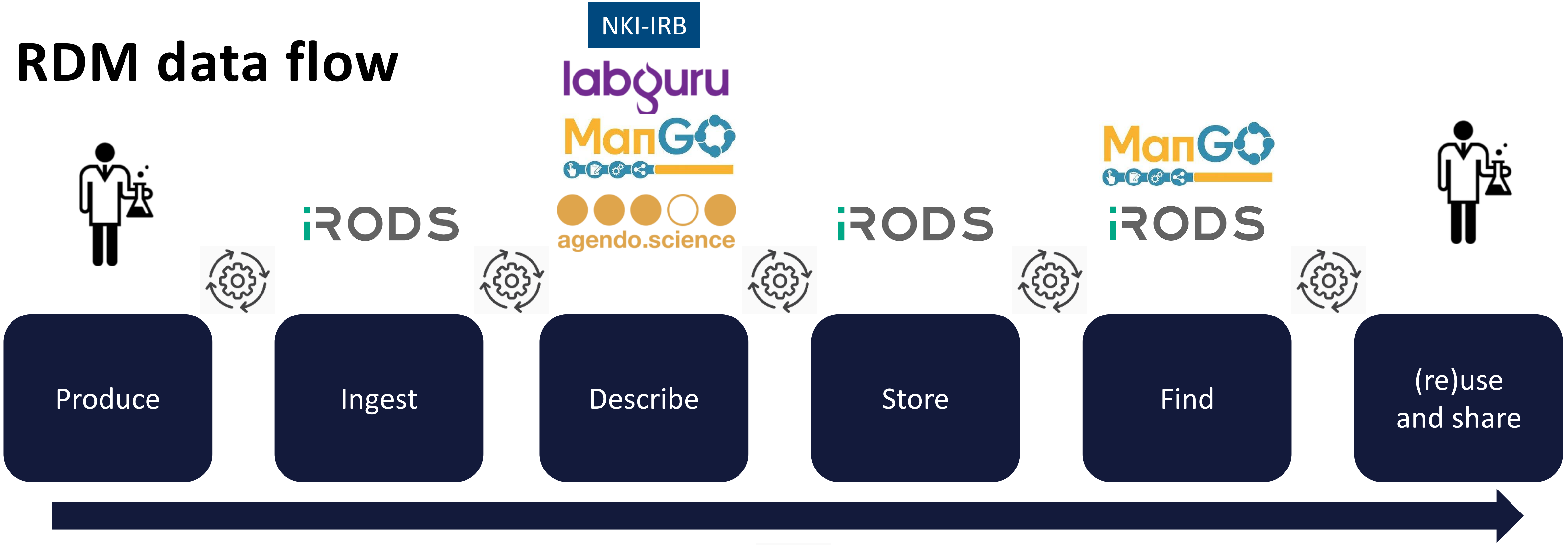


Local

High speed network & local storage

Local compute

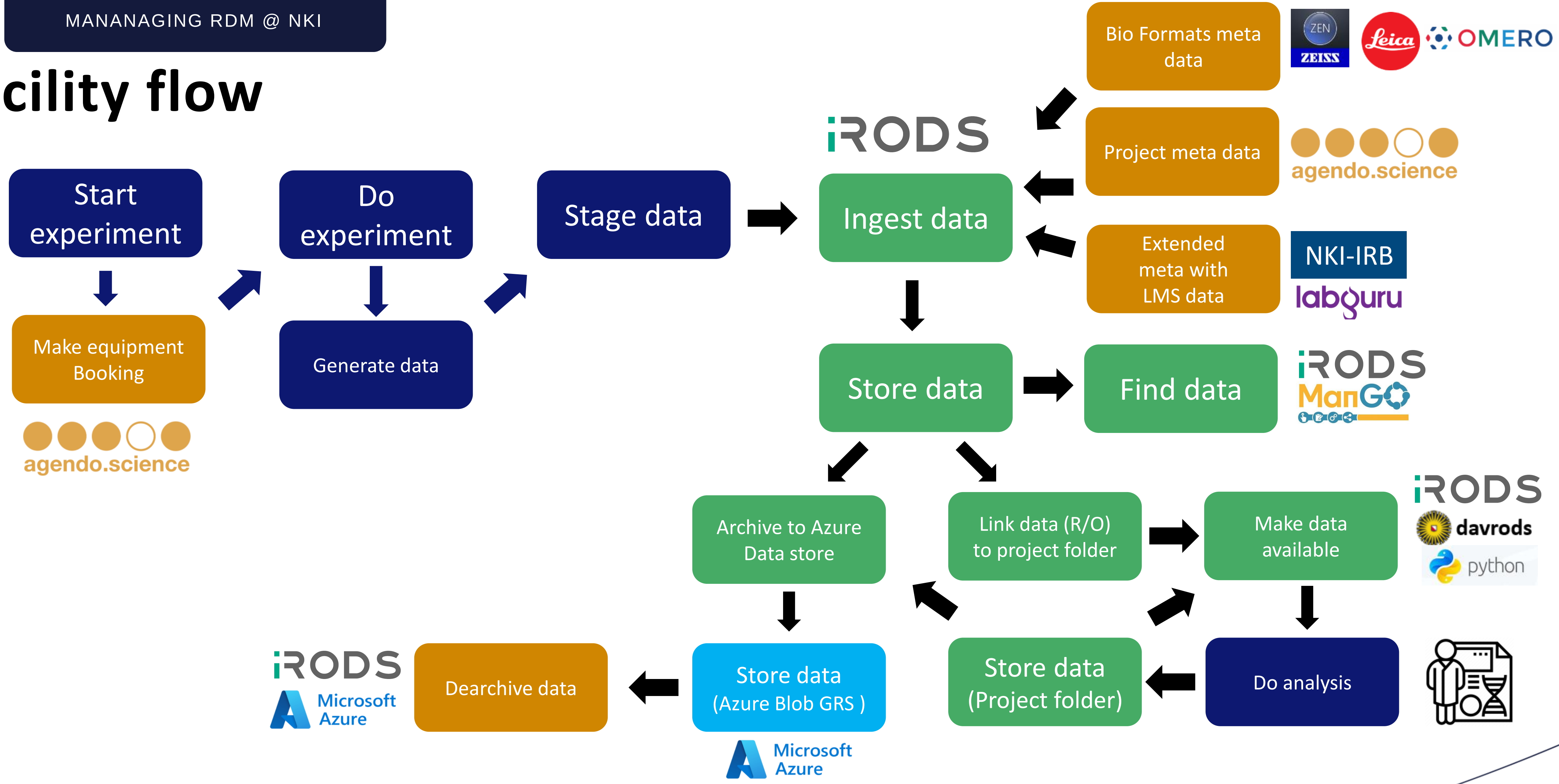
RDM data flow



Produce → Ingest → Describe → Store → Find → (re)use and share

agendo.science
globus
labguru
Research Manager
NKI-IRB
SLIDESCORE
python
arivis
MathWorks
AIVIA
CellProfiler
cell image analysis software
indica labs
OracleBio
OMERO
davrods
ZEN
ZEISS
Leica

Facility flow



Start experiment

Make equipment Booking

Do experiment

Generate data

Stage data

Ingest data

Bio Formats meta data

Project meta data

Extended meta with LMS data

Store data

Find data

Archive to Azure Data store

Link data (R/O) to project folder

Make data available

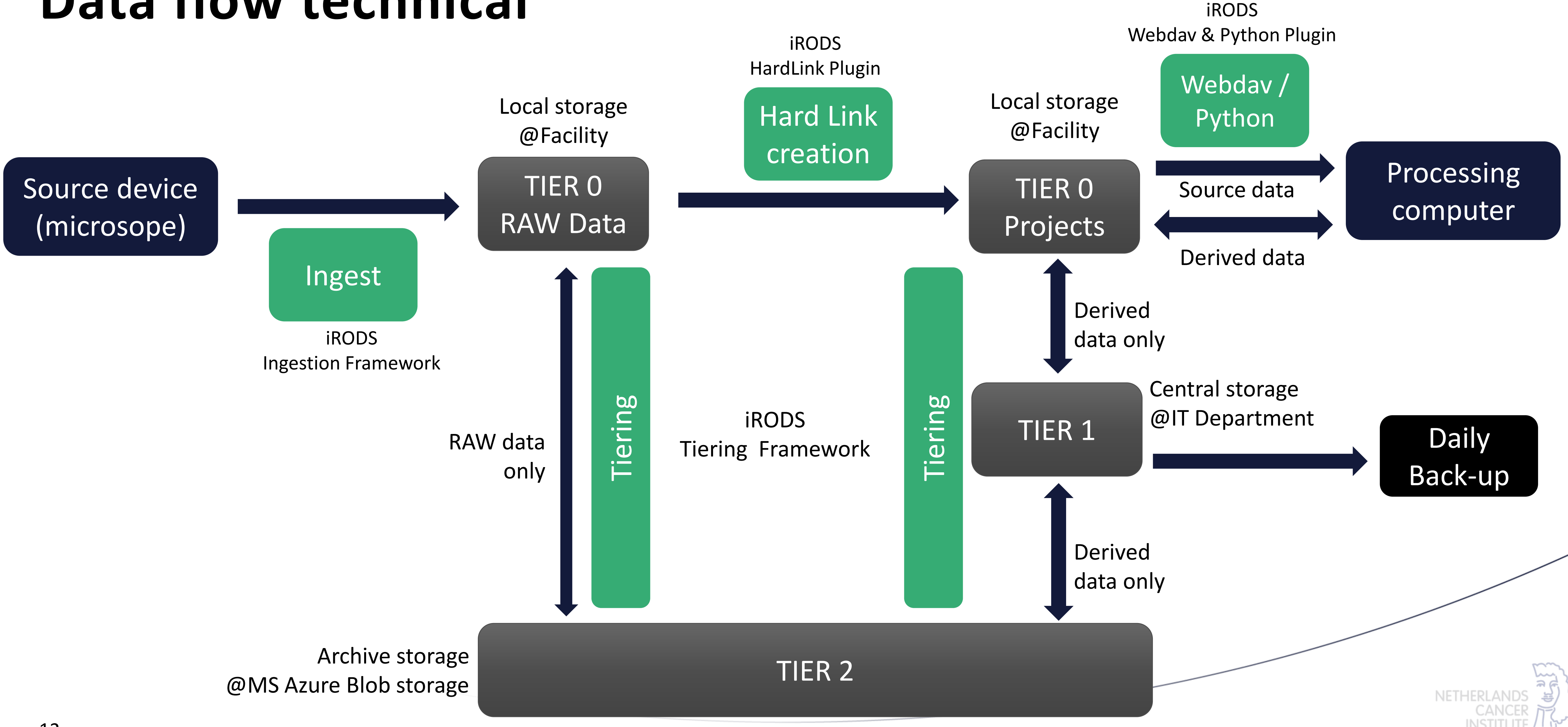
Store data (Azure Blob GRS)

Store data (Project folder)

Do analysis

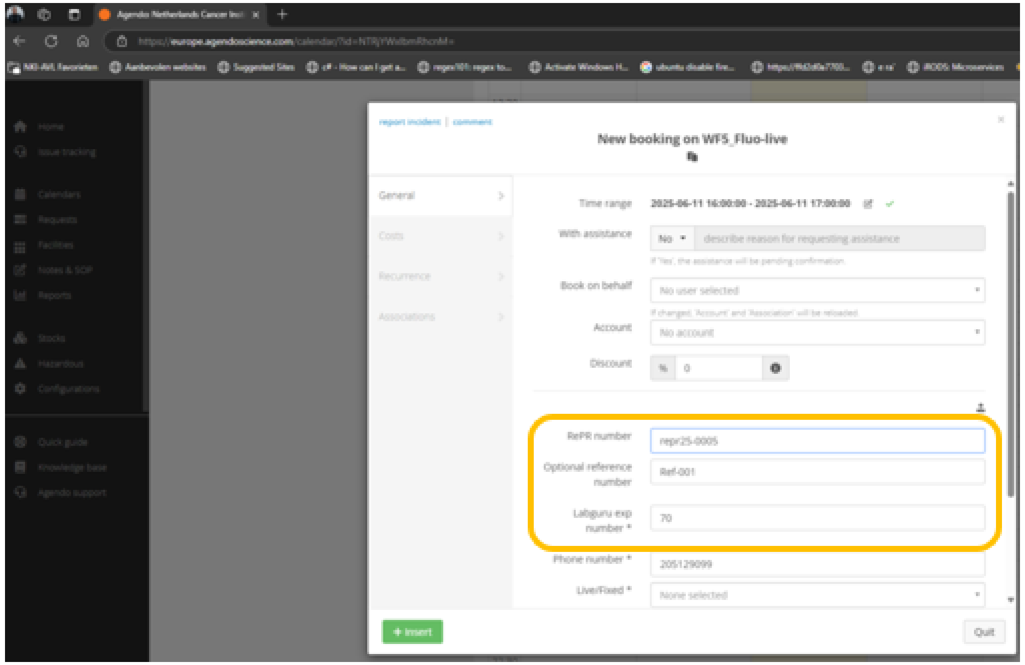
Dearchive data

Data flow technical



RDM Web User interfaces

agendo.science



New booking on WFS_Fluo-live

Time range: 2025-06-11 16:00:00 - 2025-06-11 17:00:00

With assistance: No — describe reason for requesting assistance

If 'Yes', the assistance will be pending confirmation

Book on behalf: No user selected

If changed, 'Account' and 'Association' will be reloaded

Account: No account

Discount: % 0

RePR number: repr25-0005

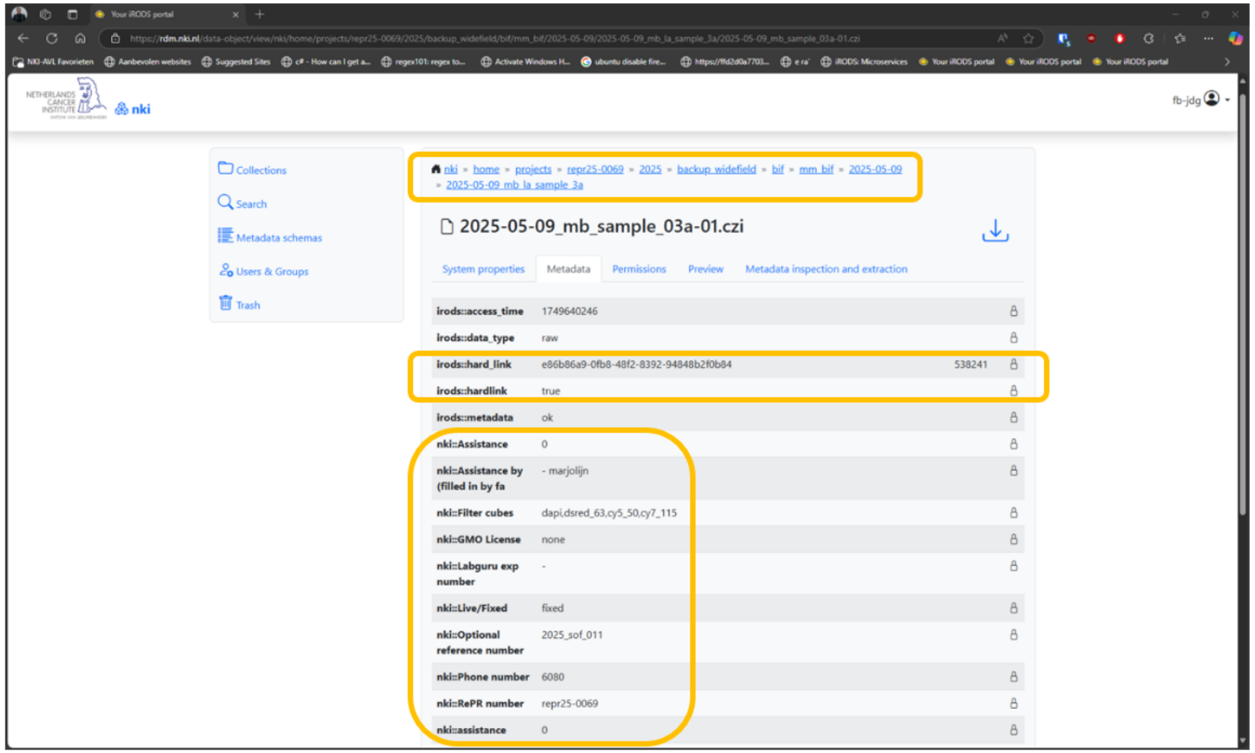
Optional reference number: Ref-001

Labguru exp number *: 70

Phone number *: 205129099

Live/Fixed *: None selected

ManGO



nki » home » projects » repr25-0069 » 2025 » backup_widefield » bif » mm_bif » 2025-05-09 » 2025-05-09_mb_la_sample_3a

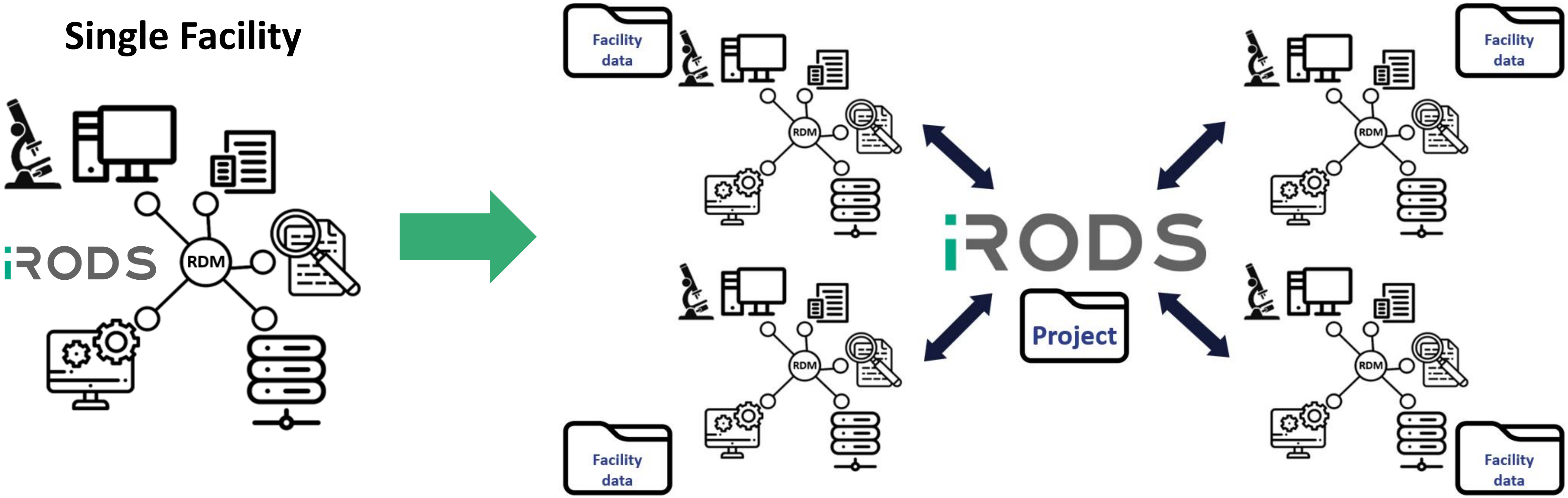
2025-05-09_mb_sample_03a-01.czi

irods::access_time	1749640246	
irods::data_type	raw	
irods::hard_link	e86b86a9-0fb8-48f2-8392-94848b2f0b84	538241
irods::hardlink	true	
irods::metadata	ok	
nki::Assistance	0	
nki::Assistance by (filled in by fa	- marjolijn	
nki::Filter cubes	dapi,dsred_63,cy5_50,cy7_115	
nki::GMO License	none	
nki::Labguru exp number	-	
nki::Live/Fixed	fixed	
nki::Optional reference number	2025_sof_011	
nki::Phone number	6080	
nki::RePR number	repr25-0069	
nki::assistance	0	

NKI Facilities RDM/iRODS Expansion

Single Facility

Multi Facility



Clinical Pathology and research with AI/ML

Current situation

- Pathology slides are scanned at Clinical Pathology department for diagnostic
- Pathology slides are scanned at research facility for research purposes
- Pathology slides are scanned and stored twice, no data carry over/data re-use
- Manual labor involved for older slides (>6month). Must be retrieved out of physical archive.
- No digital archive for clinical pathology department.
- Complexity: clinical network is separated from research network.

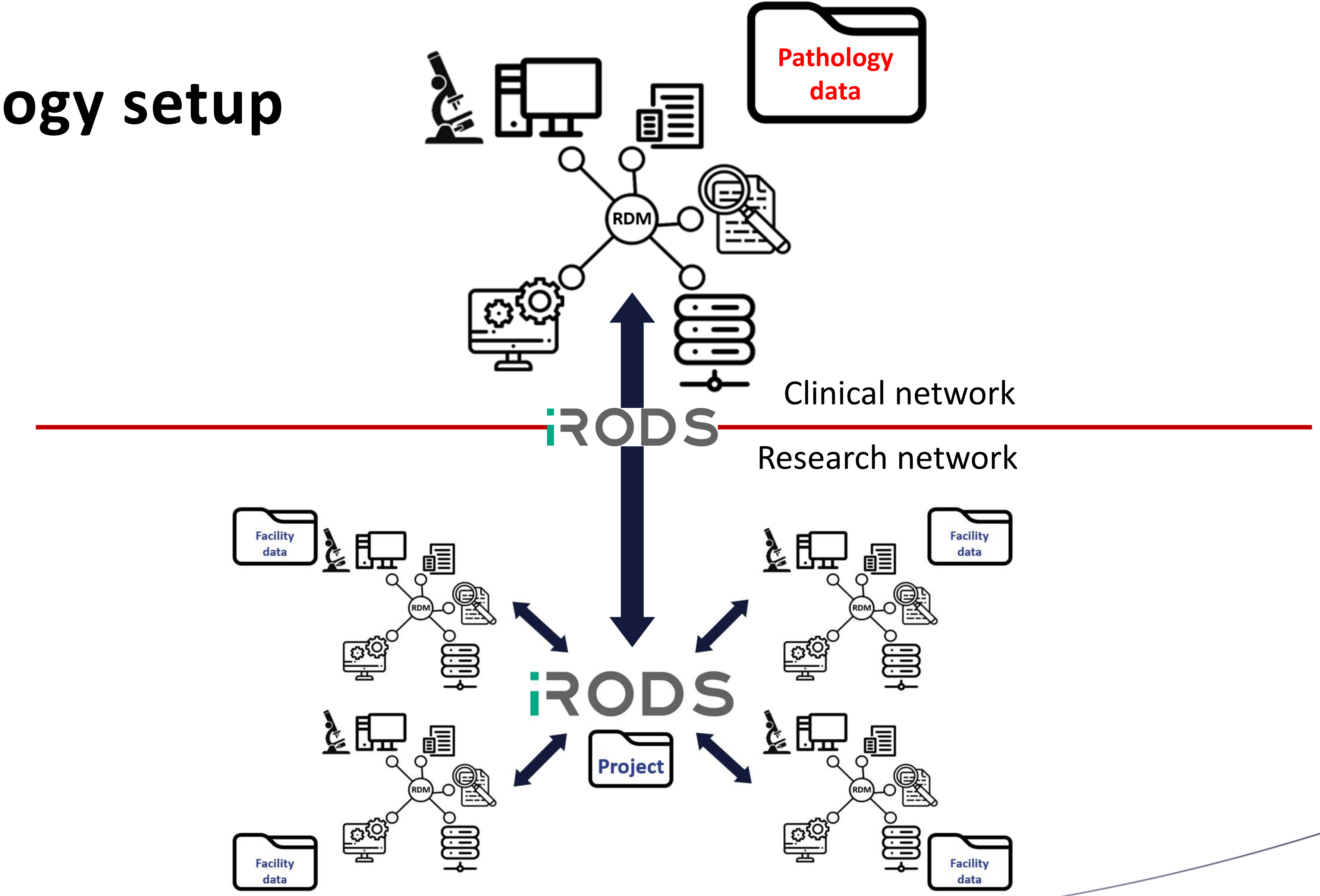
Desired situation

- Pathology slides are scanned at Clinical Pathology department for diagnostic and caried over to research for research purposes.
- Scanned slides are described and stored in long term archive for re-use for both clinic and research.
- Safe data, coordinated and auditable data transfer from clinic to research
- Data (at scale) available for AI/ML research workloads.

Solution

- Use iRODS for safe and auditable data transport, meta data handling, data storage and managed data access

Pathology setup



Pathology data

RDM

iRODS

Clinical network

Research network

Facility data

Facility data

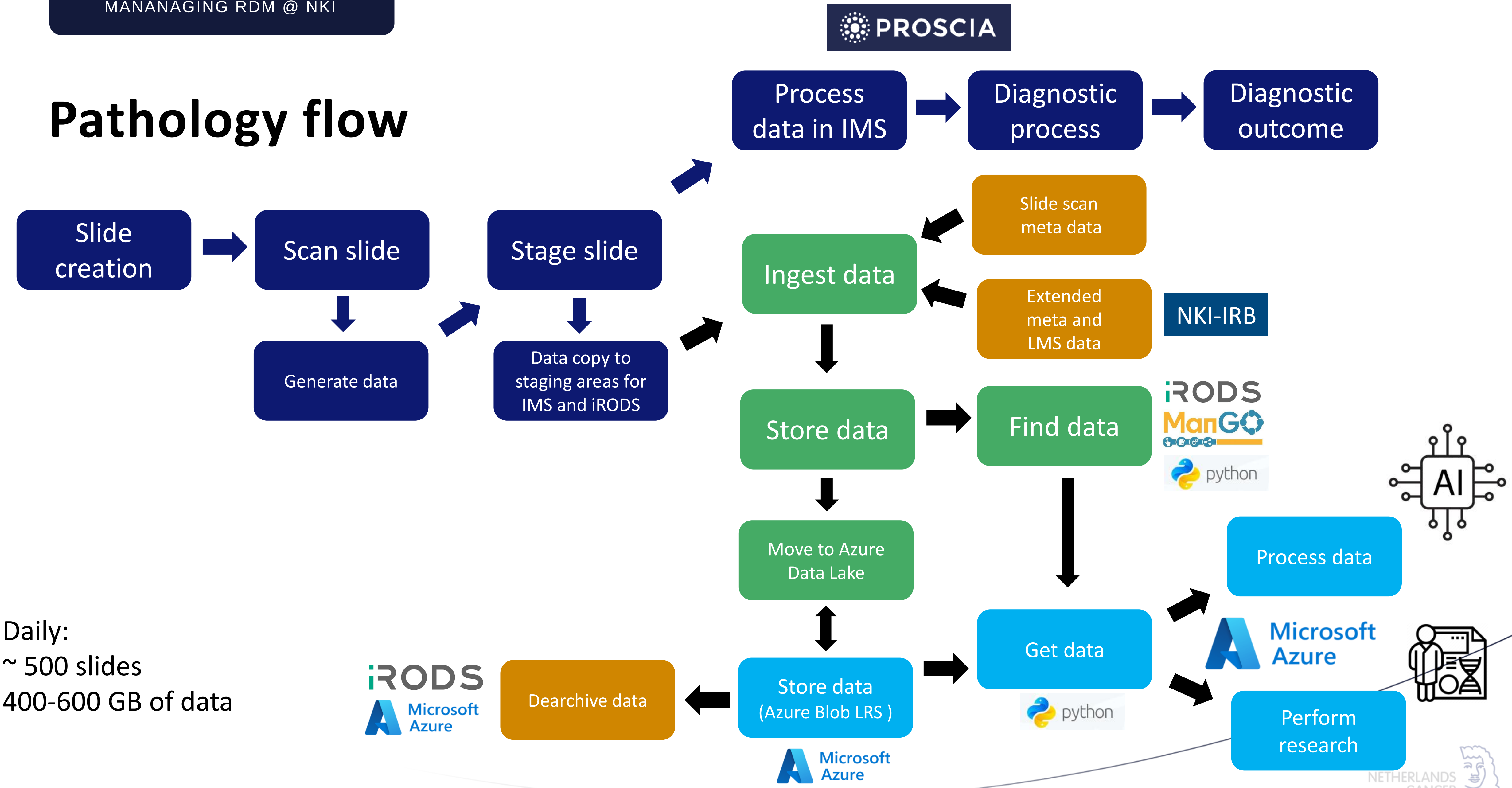
iRODS

Project

Facility data

Facility data

Pathology flow



PROSCIA

Slide creation

Scan slide

Generate data

Stage slide

Data copy to staging areas for IMS and iRODS

Process data in IMS

Diagnostic process

Diagnostic outcome

Slide scan meta data

Extended meta and LMS data

NKI-IRB

Ingest data

Store data

Find data

iRODS

ManGO

python

Move to Azure Data Lake

Store data (Azure Blob LRS)

Microsoft Azure

Dearchive data

iRODS

Microsoft Azure

Get data

python

Process data

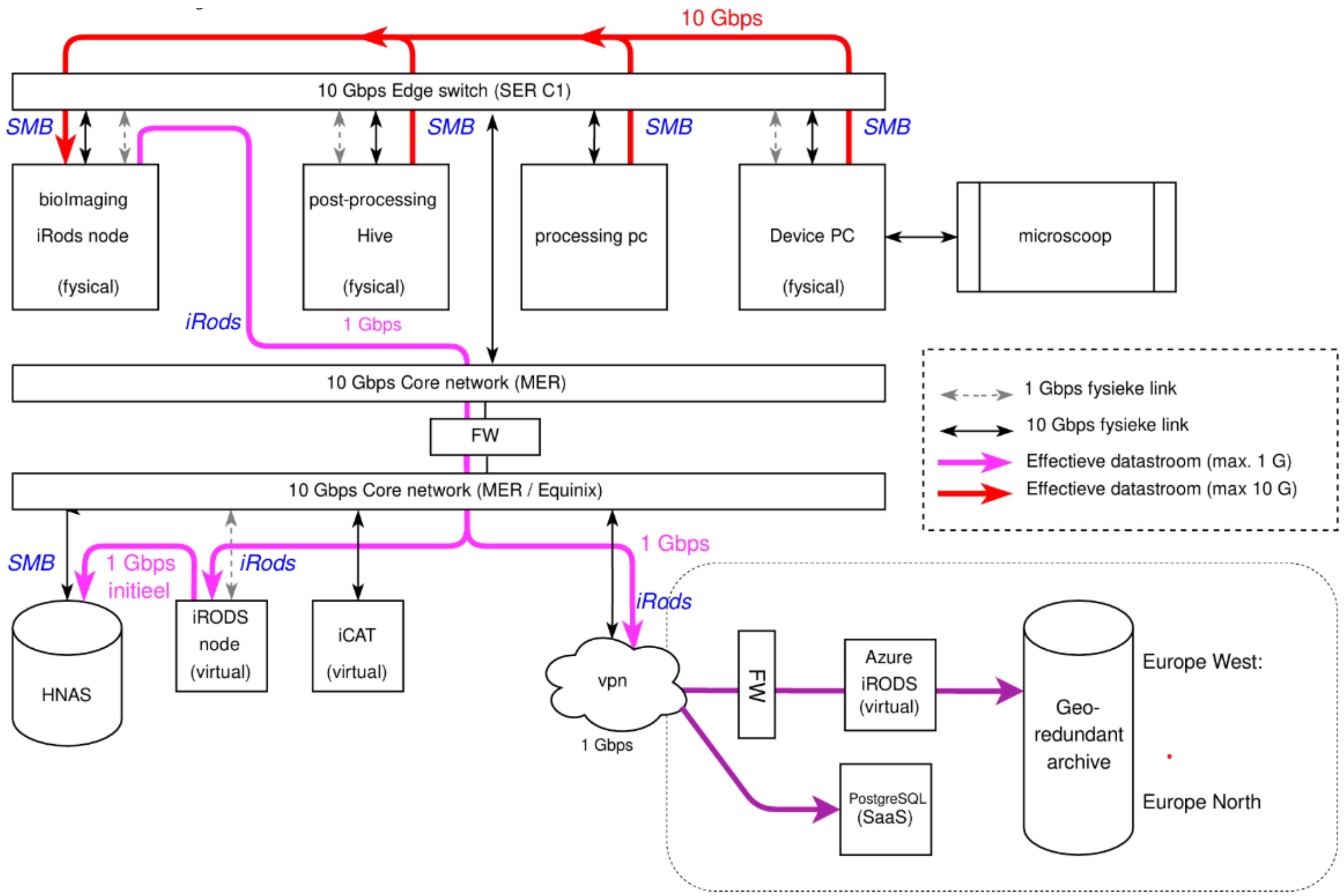
AI

Microsoft Azure

Perform research

Daily:
~ 500 slides
400-600 GB of data

Basic iRODS network architecture



Some numbers

- Total departments: 4
  - 2 operational departments
  - 2 for archive purposes
- Total files: 23.131.950
- Total TB: 270 TB
- Total TB in archive: 160 TB (and counting)
- Total unique meta data points: 52.317.828
- Database size: 150 GB

- servers: 7
  - 1 iCAT, 1 Database,
    4 storage nodes, 1 Web (Mango)
- API interfaces: 4
  - Agendo
  - NKI-IRB
  - LMS
  - Azure
- Data access API's: 3
  - iRODS
  - Python
  - WebDav
  - Web (Mango)