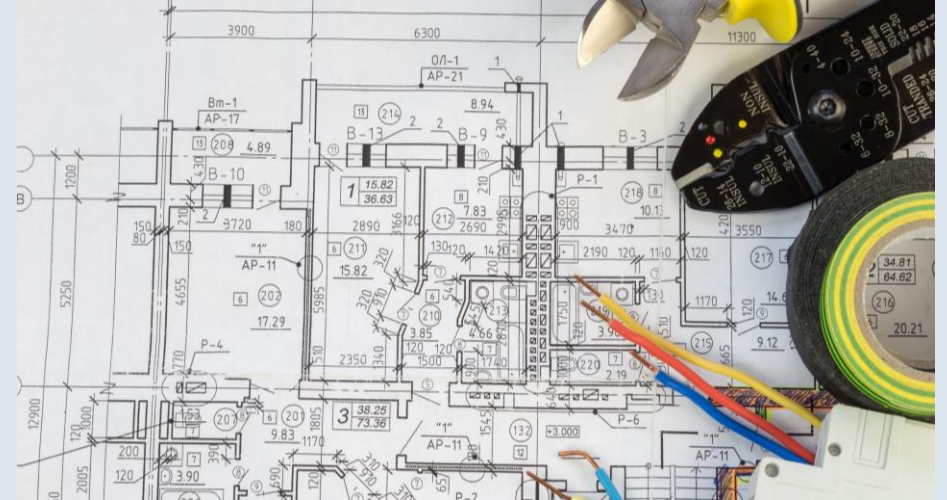


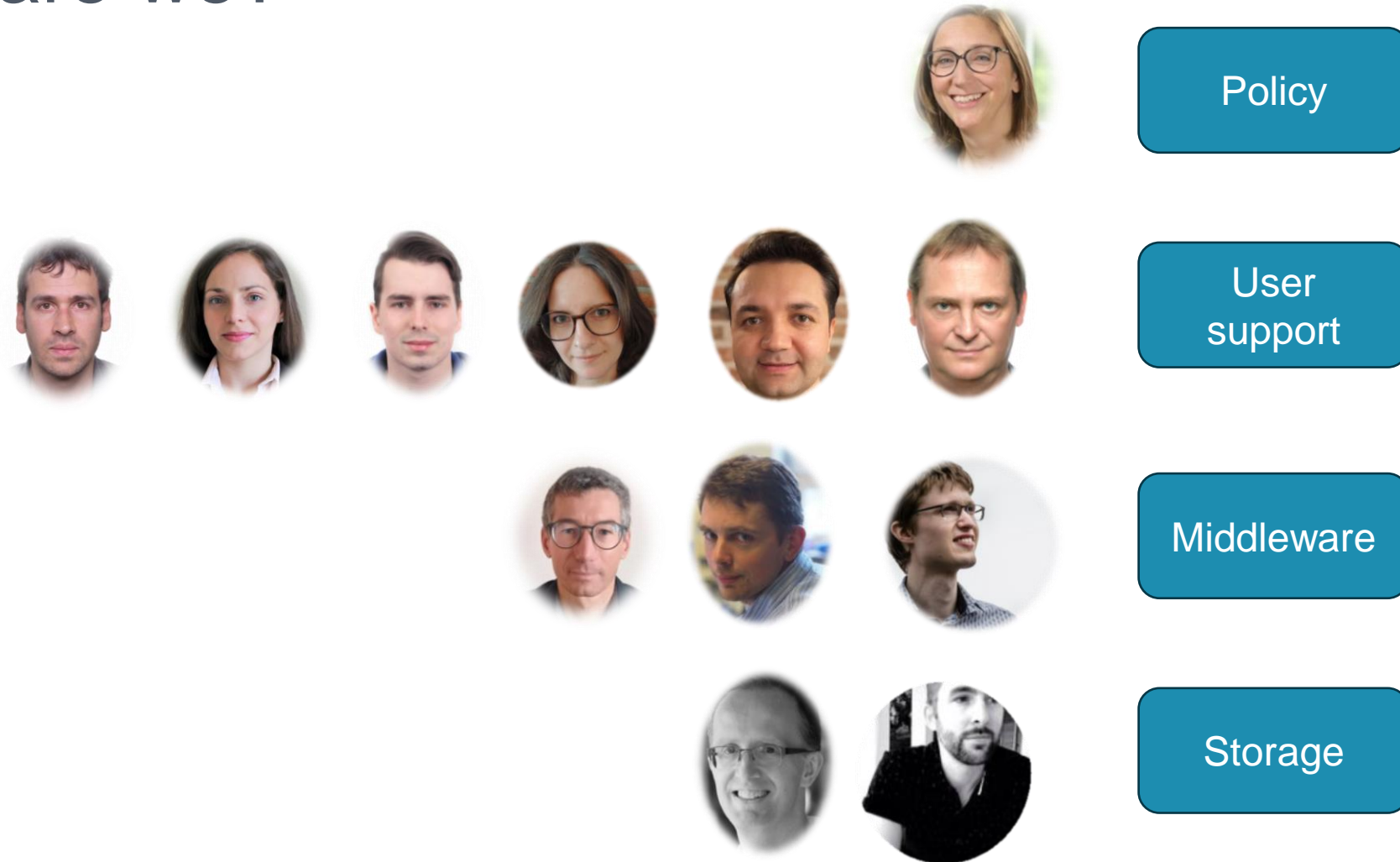


User Day 8 October 2024



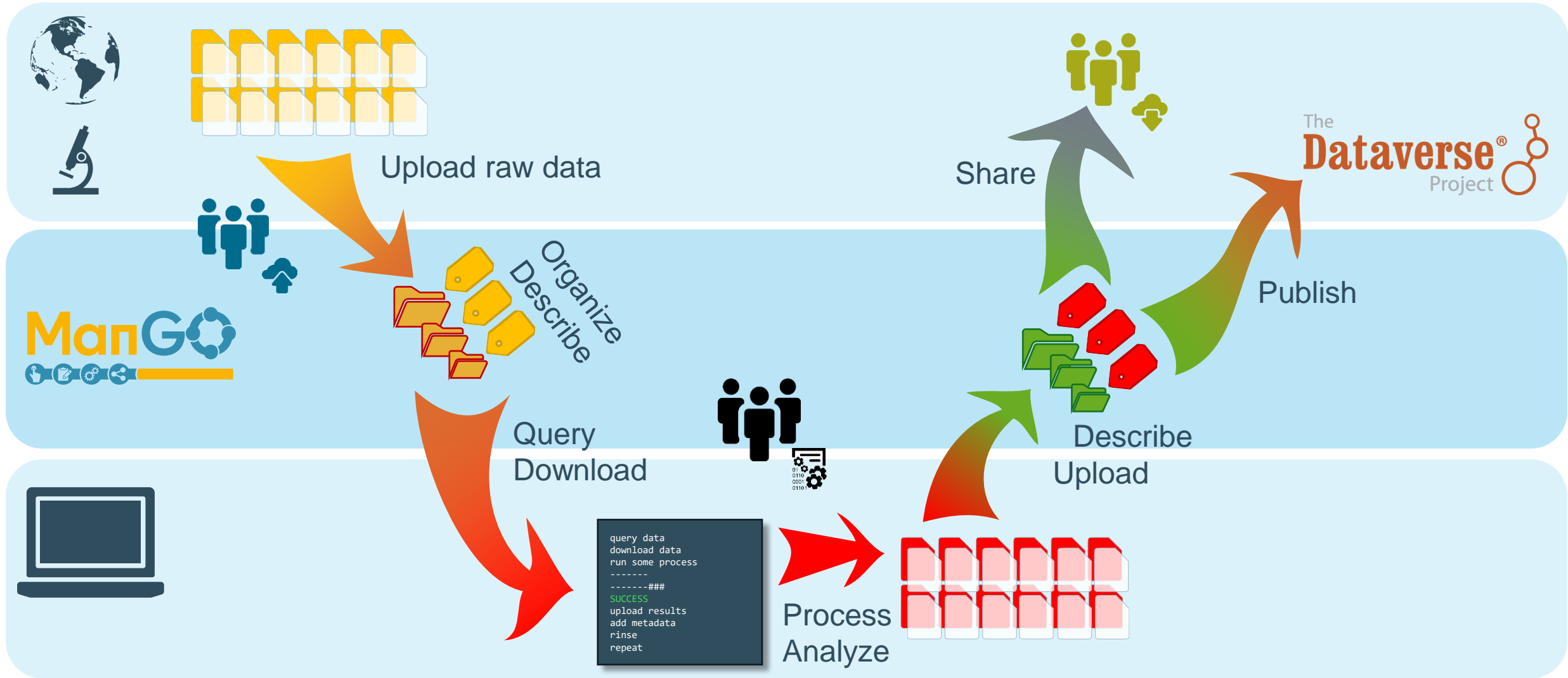
Deep dive into ManGO as a platform for research data management

Who are we?

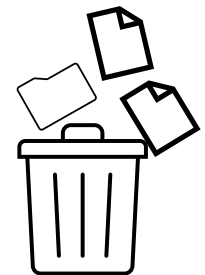
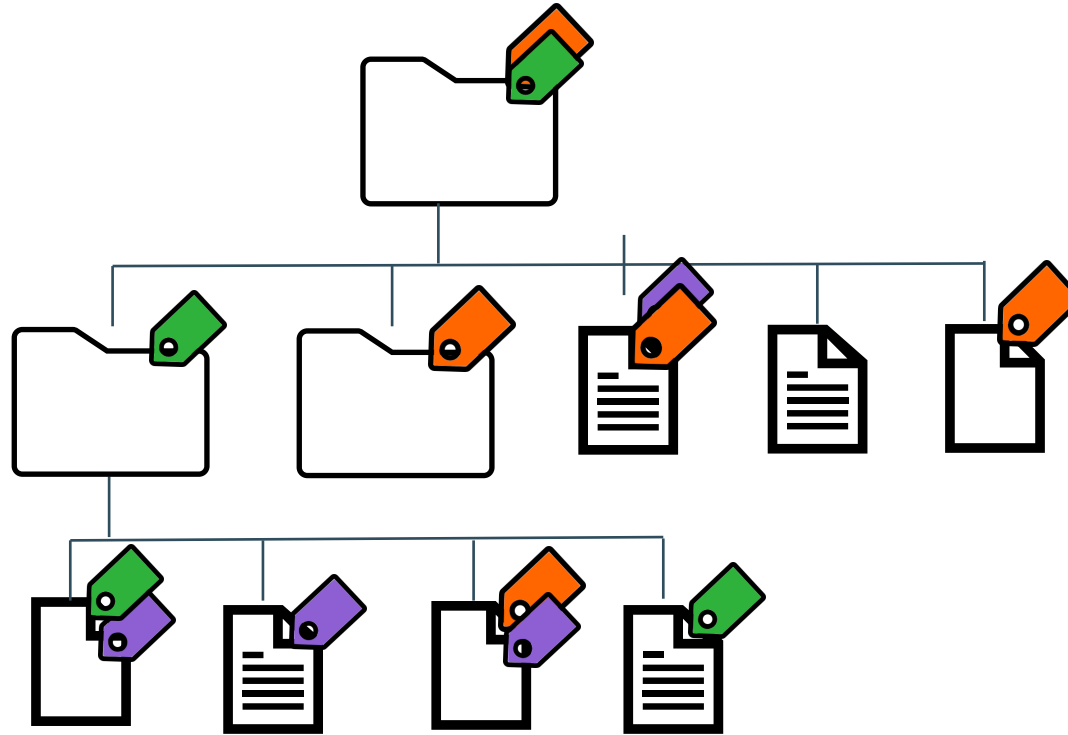




Realistic workflow example



iRODS data model



Metadata



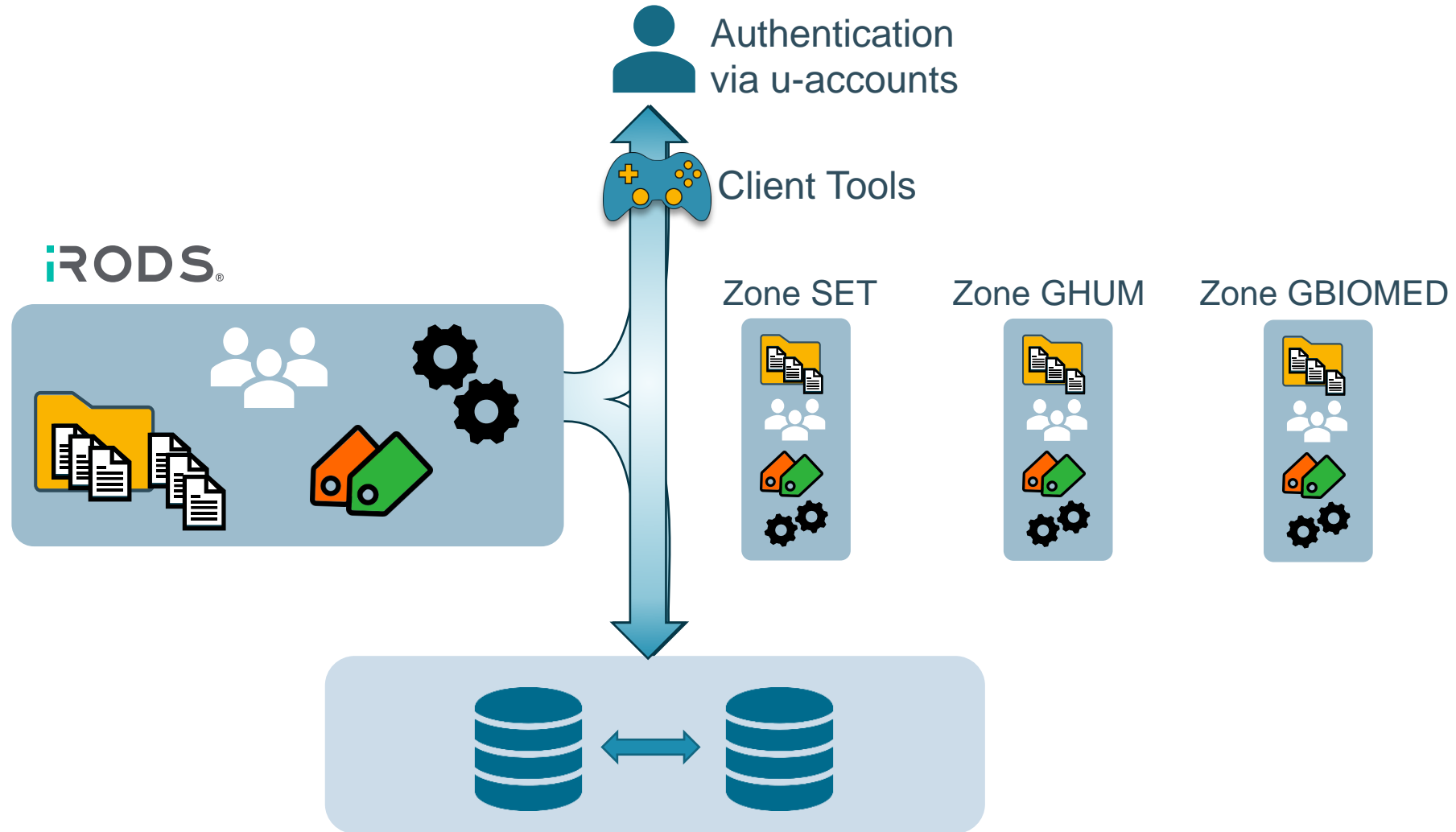
What for

- Discovery and search
- Description
- Workflows
 - status
 - validation flags
 - task driver

Sources

- Native from files
(Microscopic images)
- Via acquisition and lab
workflows (Labview)
- Ad hoc
- Structured

ManGO architecture





Client tools

ManGO Portal

iCommands

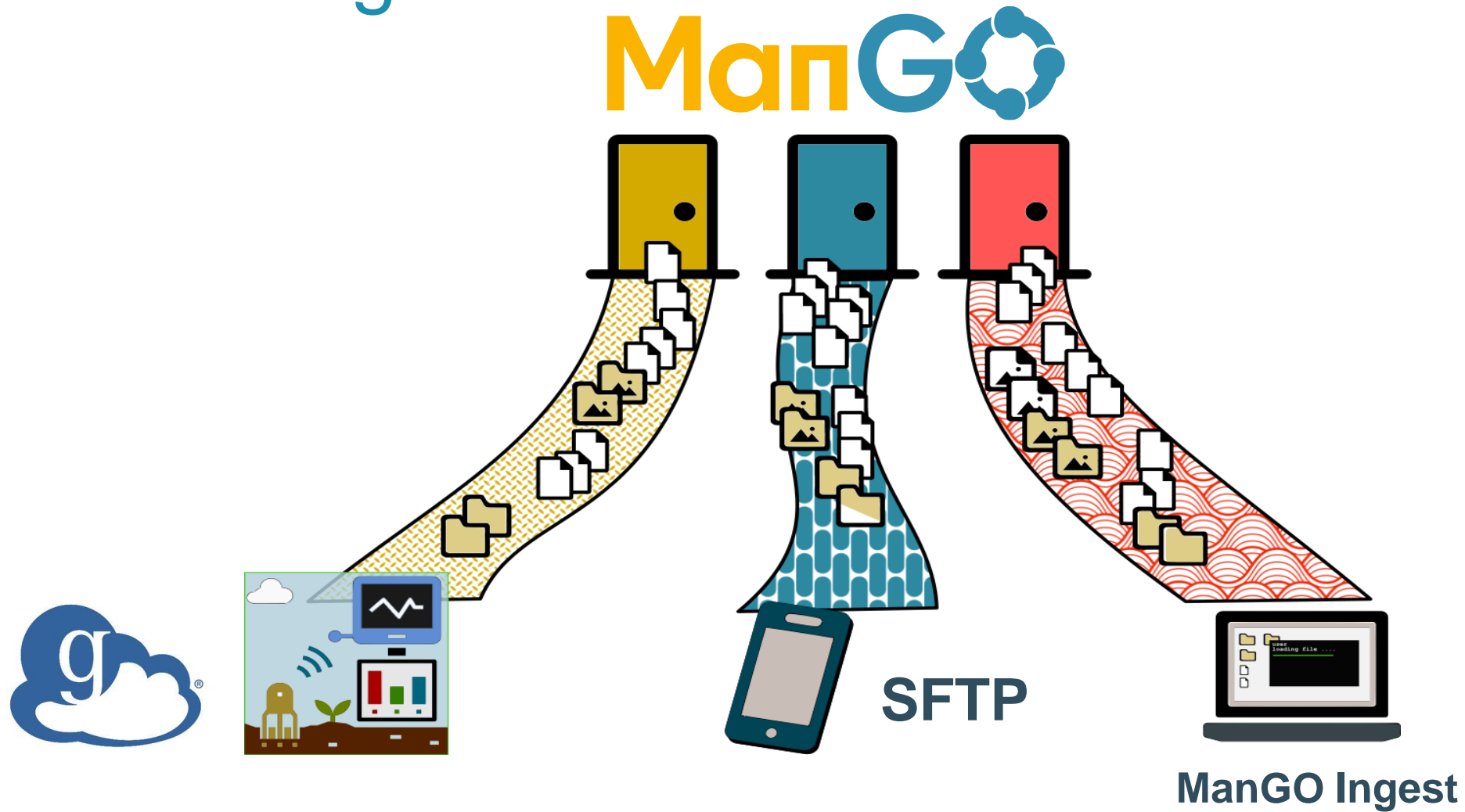


ManGO Ingest

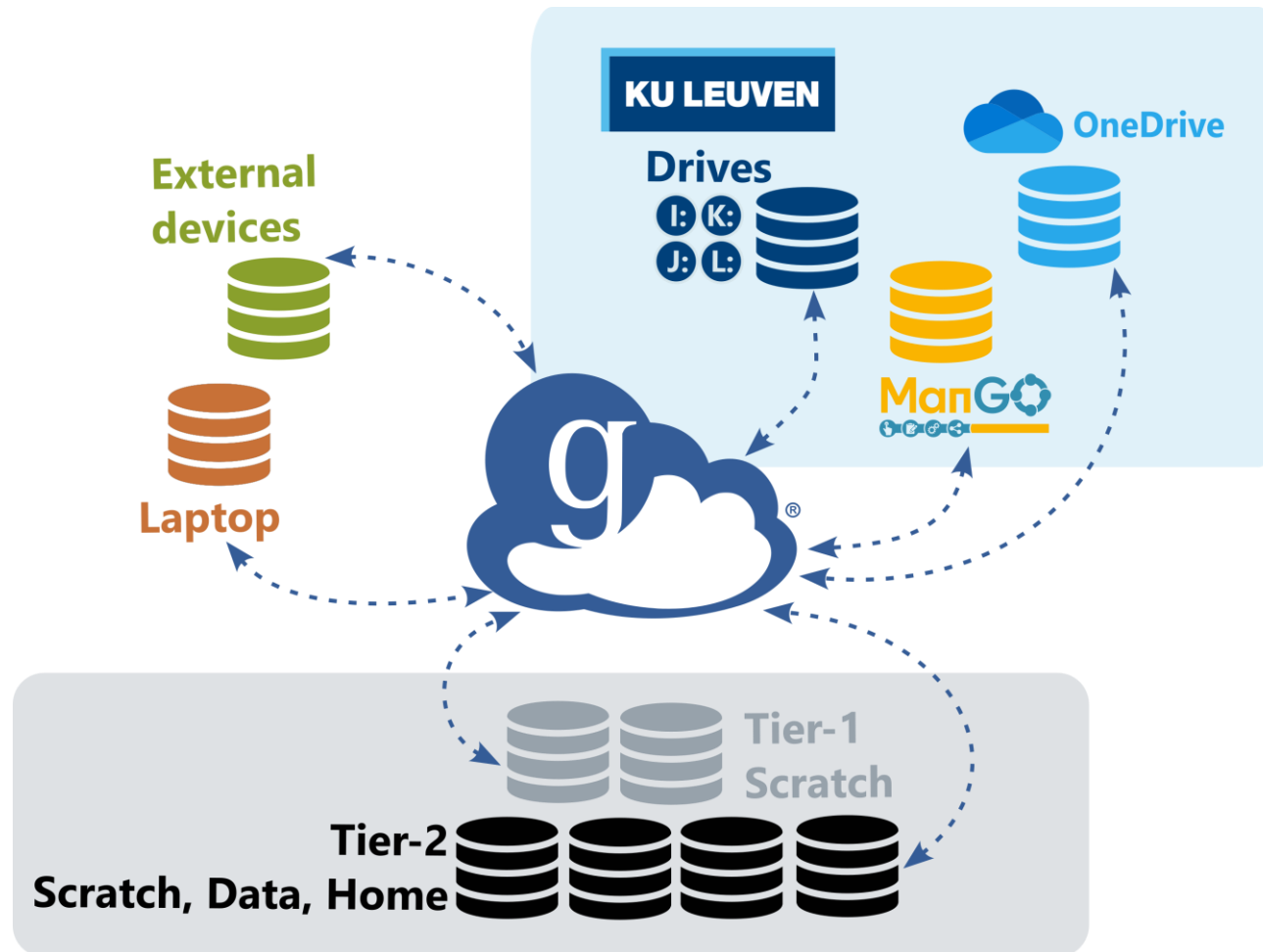
SFTP



Automatic ingest



Globus





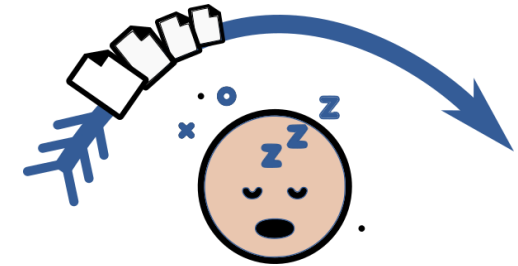
Globus



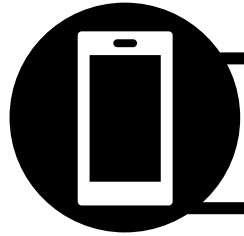
Clients (web, cli, Python, HTTP)
Secure, unattended syncing
Access by people outside of the project
Various file sizes



No metadata



SFTP Ingest tool



Semi-anonymous upload into predefined collection



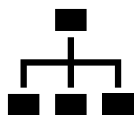
Access for untrusted devices



No metadata
Only upload



ManGO Ingest



results-gold-sublimation.csv
input-20241008.dat
results-cursed-nogood.csv
calculate.py



results-cursed*.csv



results-*.csv
input-*.dat



results-{{material}}-*.csv
input-{{exp_date}}.dat

results-gold-sublimation.csv



“material”: “gold”

input-20240101.dat




“exp_date” : 20241008

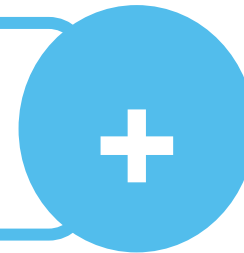
iRODS®

ManGO Ingest



Pure 
No cronjob needed
Can be used to sync too
Powerful path-based filtering
Automatic metadata extraction

Extensible with project specific filtering and
metadata extraction (eg [ExifTool](#) included)



Command line tool only

ManGO Ingest: try it out now!

```
(venv) paul@CRD-L-11056:~/projects/mango-ingest/src$ mango_ingest -d /set/home/u0123318/test_ingest -v --glob="*.py"
[21:22:23] Reporting thread started mango_ingest.py:91
mango_ingest.py:91

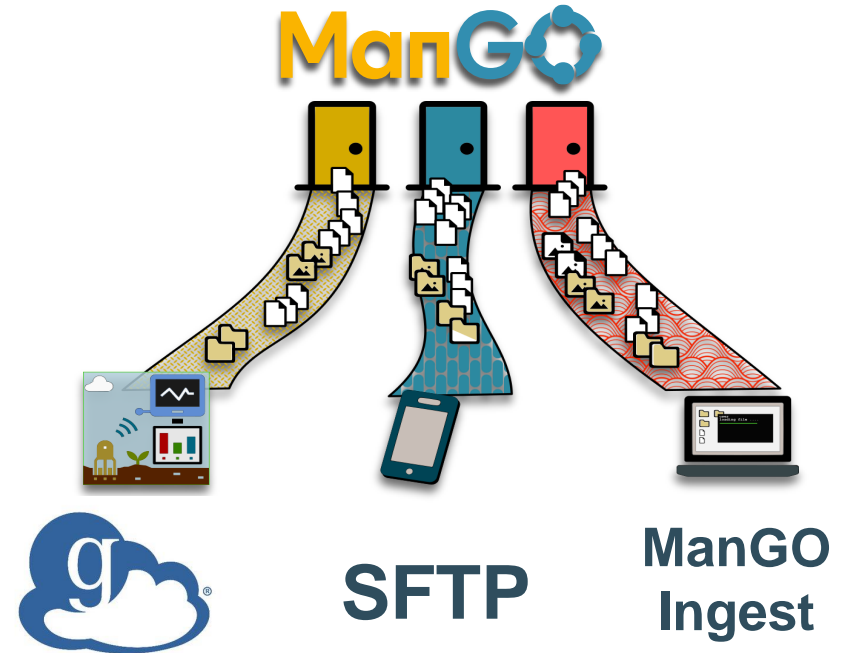
ManGO Ingest is now monitoring /home/paul/projects/mango-ingest/src
Recursive: False
Observer: <class 'watchdog.observers.polling.PollingObserver'>
Polling interval: 5 sec
Handler applied: ManGOIngestHandler
{'_case_sensitive': False,
 '_ignore_directories': True,
 '_ignore_regexes': [re.compile('(s:mango_ingest_results\\-.*\\.json)\\Z',
 re.IGNORECASE)],
 '_regexes': [re.compile('(s:.*\\.py)\\Z', re.IGNORECASE)],
 'delay_queue': {},
 'delay_queue_last_visit': 1728242543.345291,
 'delay_queue_lock': <unlocked_thread.lock object at 0x7fd86c564100>,
 'filter': None,
 'filter_kwargs': {},
 'irods_destination': '/set/home/u0123318/test_ingest',
 'metadata_handlers': [],
 'observer': 'polling',
 'path': PosixPath('/home/paul/projects/mango-ingest/src'),
 'path_list_to_treat': [],
 'verify_checksum': False}

Uploading ... 100% 0:00:00 0:00:00 50.3 kB ? 50.3 kB mango_ingest.py
```

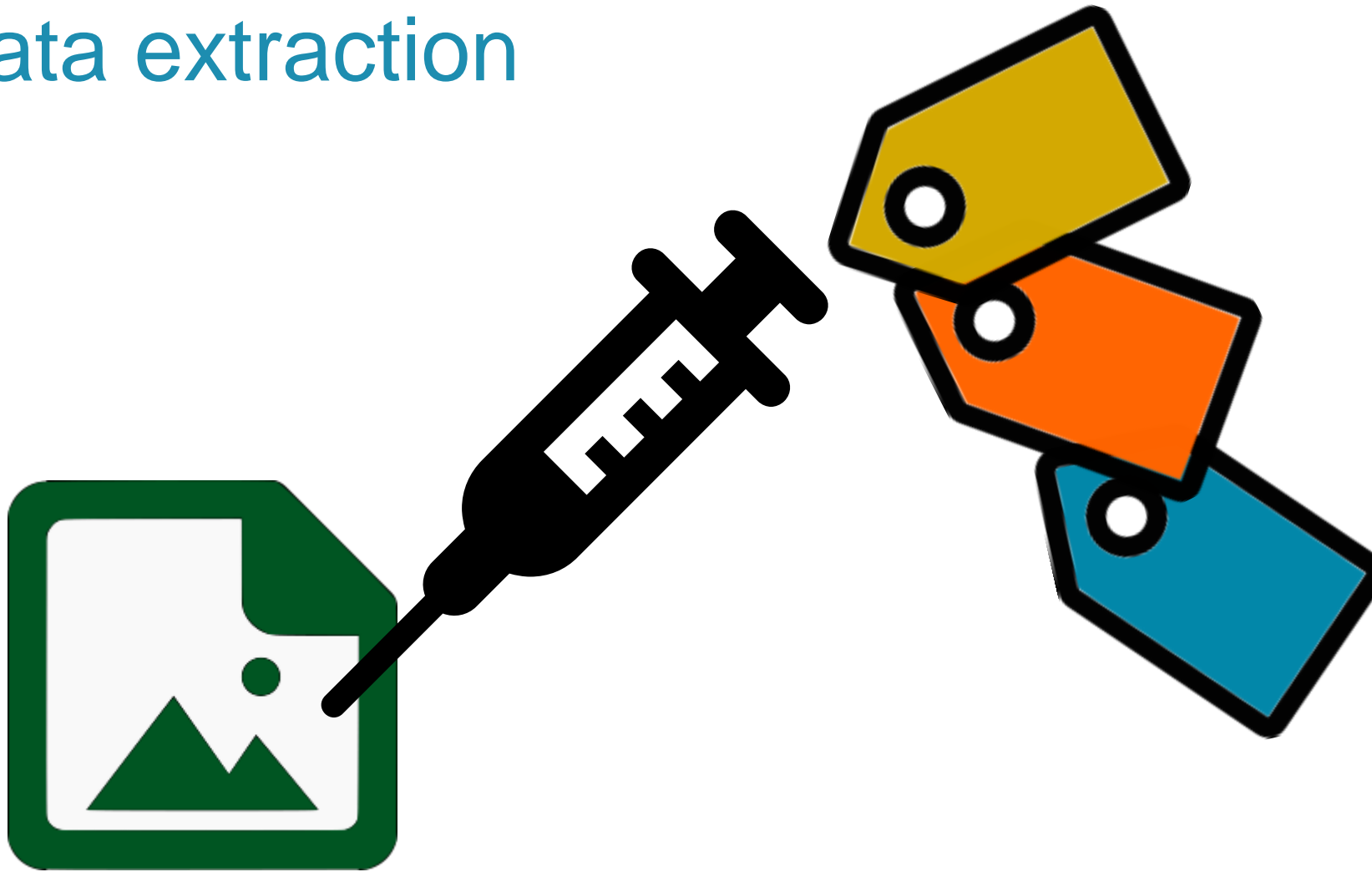
<https://github.com/kuleuven/mango-ingest>

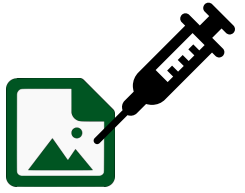
Automatic ingest – bottom line

- ☁ SFTP for unsafe devices, semi-anonymous uploads.
- ☁ Globus for connections between servers, synchronization, timed transfers.
- ☁ ManGO Ingest for lightweight scanning of local file system and customized treatment in uploads.



Metadata extraction





Metadata extraction from files

- Apache Tika
 - Discovery, integrated in ManGO Portal
- ExifTool
 - Similar as Apache Tika but even more file formats
- Specialized libraries
 - pylibCZlrw (Czi Microscopy images)

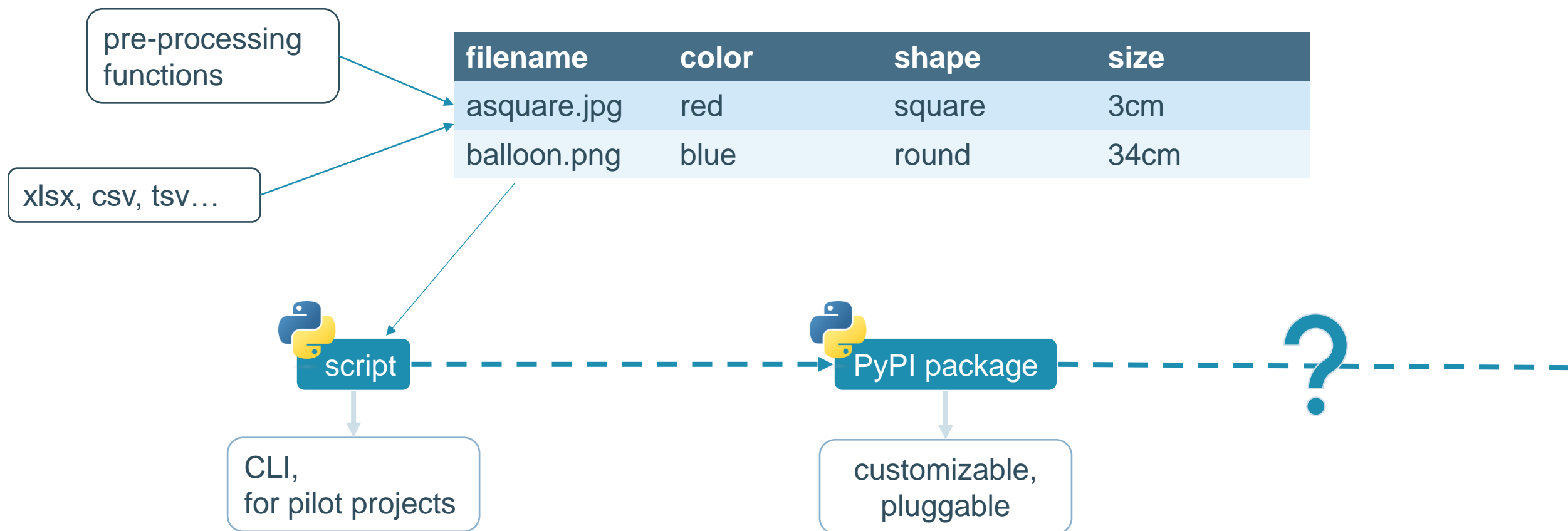


ExifTool





Metadata extraction from tabular files

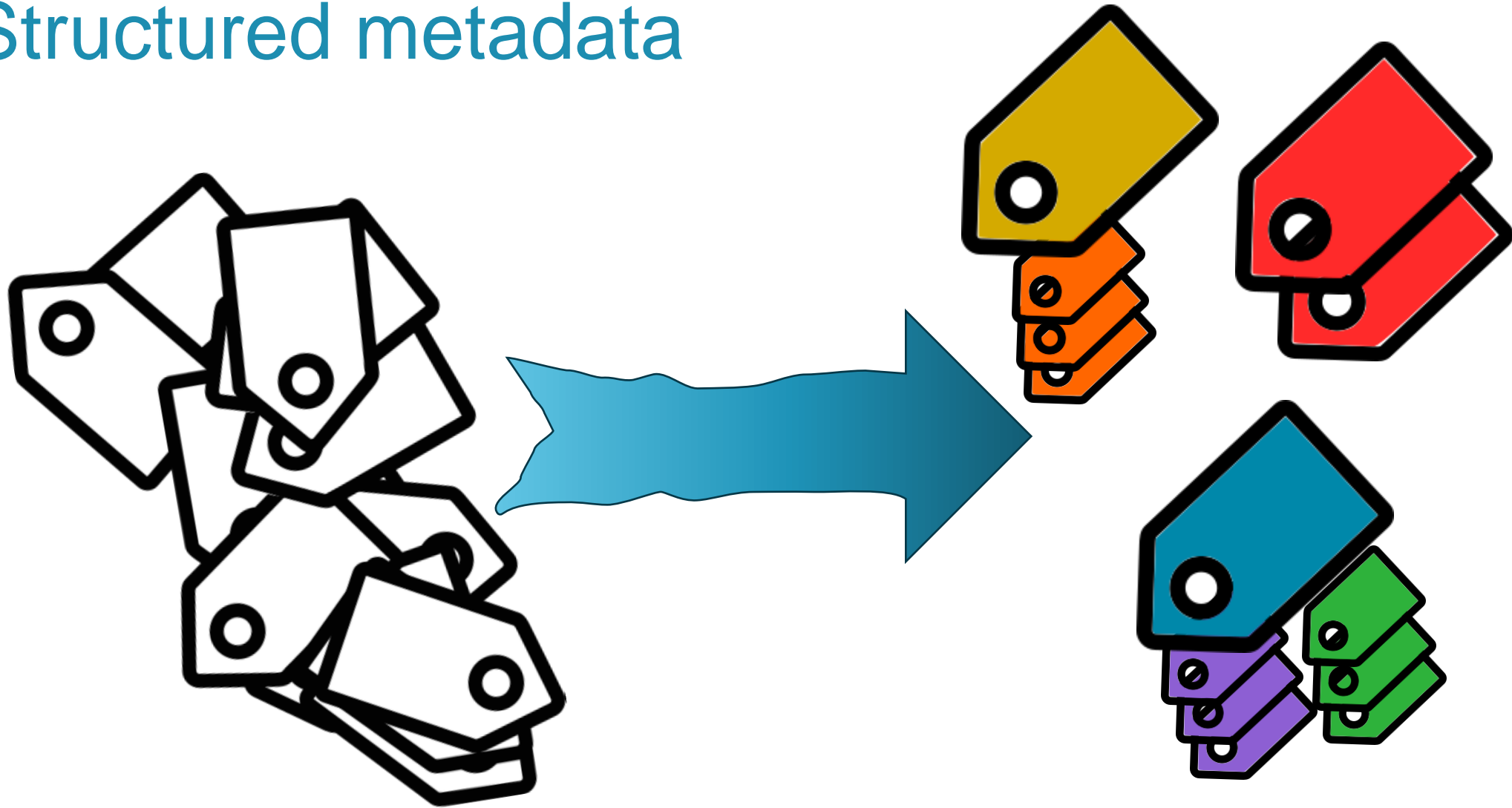


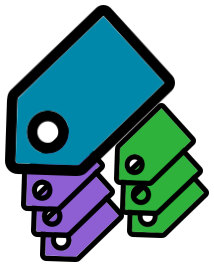
Metadata extraction – bottom line

- 🏷️ General or specialized packages to extract from inside the files
- 🏷️ Python package to extract from a tabular file
- 🏷️ Automated with ingest, triggered by rules, or ran ad hoc



Structured metadata





Structured Metadata

- Built on top of iRODS simple linear AVU model
- Hierarchies and validation
- GUI for editing and managing
- Schemas stored as JSON
- Python module for schema validation



The screenshot shows a web form for entering structured metadata. It includes the following fields:

- Material***: A dropdown menu with "wood" selected.
- Color**: A group of radio buttons with options: grey, red, blue, yellow, green, and other.
- Purchase date**: A date input field with the placeholder "mm/dd/yyyy" and a calendar icon. Below it, the text "Input type: date" is displayed.
- Width (in cm)***: A numeric input field with a unit icon. Below it, the text "Input type: float between 0.8 and 8.6" is displayed.

Hierarchies and namespaces

Author

Given name*


Mariana

Input type: text

Last name*


Montes

Input type: text

Email address 

mariana.montes@kuleuven.be

Input type: email

Email address 

montesmariana@gmail.com

Input type: email

Author

Given name*


Paul

Input type: text

Last name*

Borgermans

Input type: text

Email address 

paul.borgermans@kuleuven.be

Input type: email

Schema: “book”

Composite field: “author”

mgs.book.author.given_name: Mariana	1
mgs.book.author.last_name: Montes	1
mgs.book.author.email: mariana.montes@kuleuven.be	1
mgs.book.author.email: montesmariana@gmail.com	1
mgs.book.author.given_name: Paul	2
mgs.book.author.last_name: Borgermans	2
mgs.book.author.email: paul.borgermans@kuleuven.be	2

Metadata schema in action

Paint sample

version 1.0.0 published

View New (draft) version Copy to new schema Download JSON

Schema ID
paint_sample

Schema label
Paint sample

Add element

Colour of the sample*
other

Add element

Material composition of the pigment

☐ iron oxide
☐ limestone
☐ titanium dioxide

Add element

Save draft Publish

sample_001.png

Type: data_object

Realm: demo

Metadata schema: Paint sample 1.0

Colour of the sample*

other

Material composition of the pigment

- ☒ iron oxide
☒ limestone
☐ titanium dioxide

Save metadata

sample_001.png



Paint sample

Schema version: 1.0.0

Colour of the sample

grey

Material composition of the pigment

iron oxide, limestone

Edit

Delete metadata for schema
"Paint sample"

iRODS[®] metadata model



Attribute

Name

Type: string



Hierarchies

Value

Value

Type: string



JSON

Type: via schema

Unit

Anything

Type: string

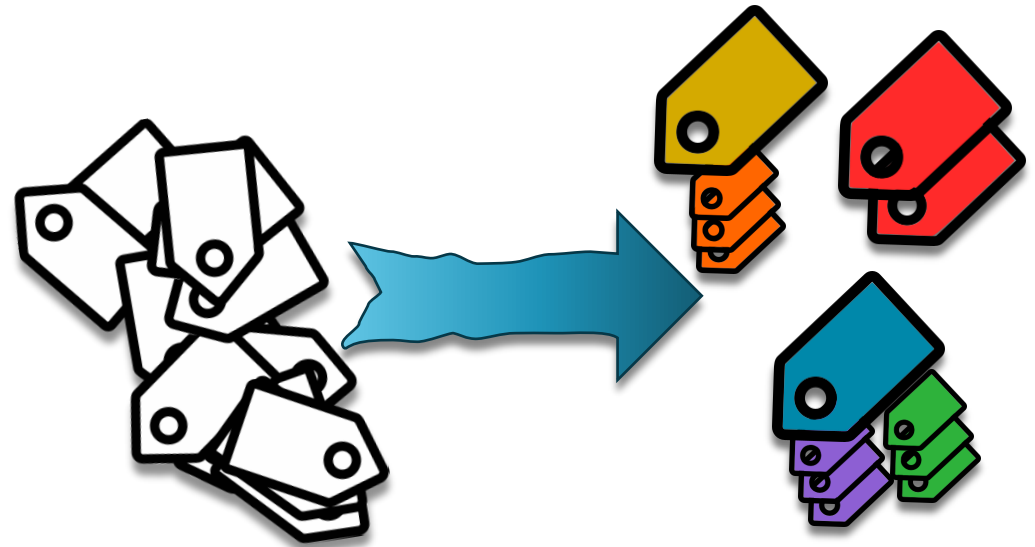


Status

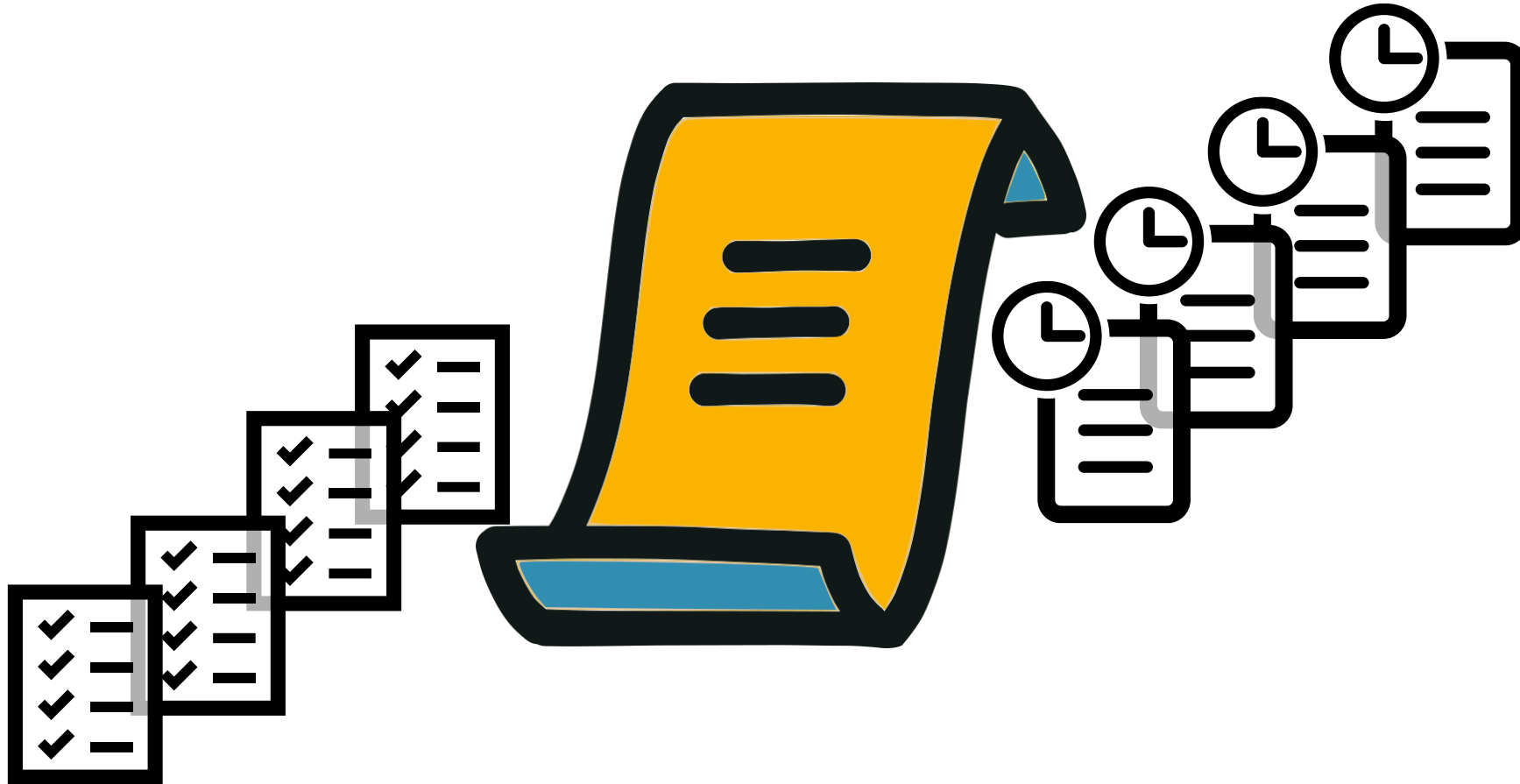
Relations

Structured metadata – bottom line

- 🏷️ Namespacing
- 🏷️ Hierarchies
- 🏷️ Grouping in ManGO portal
- 🏷️ Filtering, validation and versioning
- 🏷️ Documentation of the metadata



Auditing and reporting





Auditing capabilities

What

- Create
- Write, upload, sync
- Copy, move
- Delete
- Read, download
- Permission change
- Metadata change
- Checksum

Who

user

When

timestamp

Where

path

How

- ManGO portal
- iCommands
- Python client
- ...



History

5134230401747889603.jpg



System properties

Metadata

Permissions

Preview

Metadata inspection and extraction

History

Show 10 rows ▾ CSV

Time ▾	Realm	Action	Client	User	Path
2024-09-20 14:43:17	datateam_icts_icts_test	set_metadata_atomic	ManGO_portal	u0118974	/icts/home/datateam_icts_i
2024-09-20 13:32:14	datateam_icts_icts_test	set_metadata	mango_portal	u0118974	/icts/home/datateam_icts_i
2024-09-20 13:32:03	datateam_icts_icts_test	set_metadata	mango_portal	u0118974	/icts/home/datateam_icts_i
2024-09-20 13:27:22	datateam_icts_icts_test	set_metadata_atomic	ManGO_portal	u0118974	/icts/home/datateam_icts_i
2024-09-20 13:27:03	datateam_icts_icts_test	set_metadata_atomic	ManGO_portal	u0118974	/icts/home/datateam_icts_i
2024-09-20 13:26:46	datateam_icts_icts_test	read/download	ManGO_portal	u0118974	/icts/home/datateam_icts_i
2024-09-20 13:26:45	datateam_icts_icts_test	read/download	ManGO_portal	u0118974	/icts/home/datateam_icts_i
2024-09-20 13:26:33	datateam_icts_icts_test	read/download	mango_portal	u0118974	/icts/home/datateam_icts_i
2024-09-20 13:26:31	datateam_icts_icts_test	read/download	mango_portal	u0118974	/icts/home/datateam_icts_i
2024-09-20 13:26:31	datateam_icts_icts_test	set_metadata	mango_portal	u0118974	/icts/home/datateam_icts_i

Showing 1 to 10 of 12 entries



Report automated tasks

- ✓ Aggregate results of a set of tasks and report on them
- ✓ Tasks have been triggered automatically
 - ✓ Did they go well?
 - ✓ Did anything suspicious happen?
 - ✓ If something went wrong, what do we know?
 - ✓ When, where?!
- ✓ You may never read them, but if you ever need them, they will be there

Auditing and reporting – bottom line

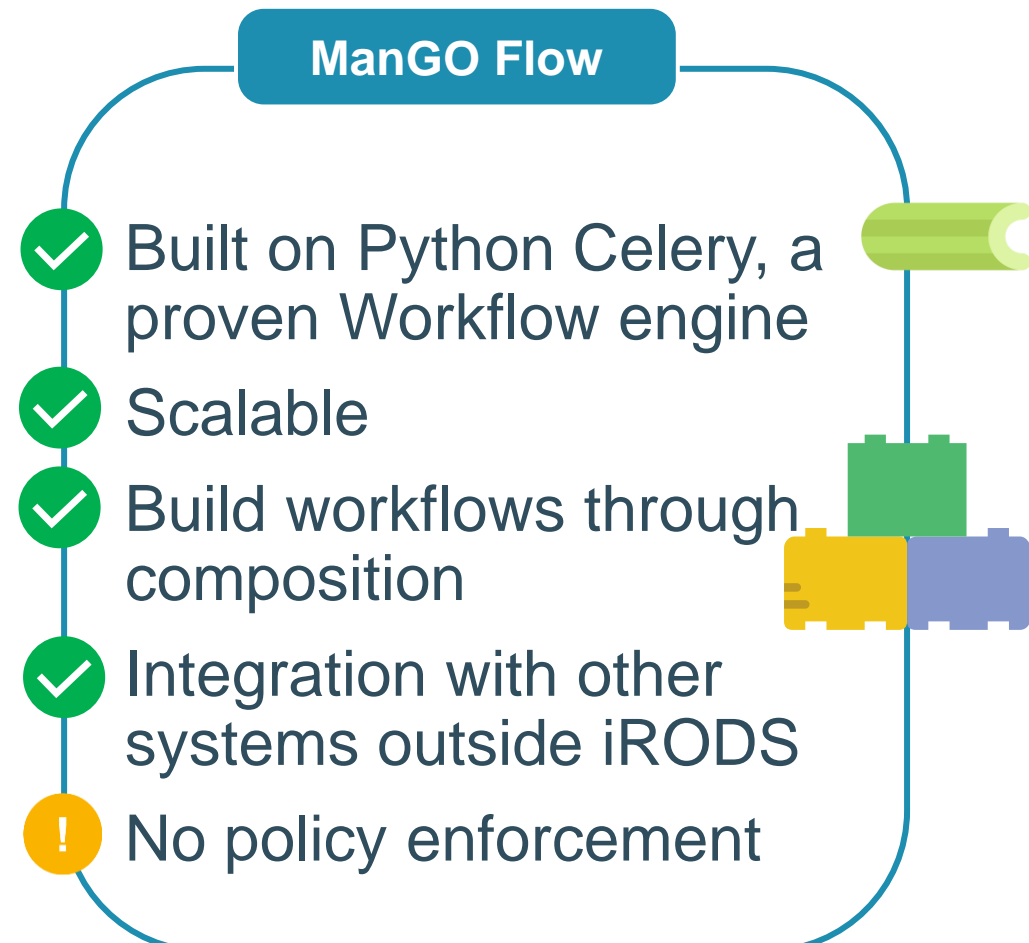
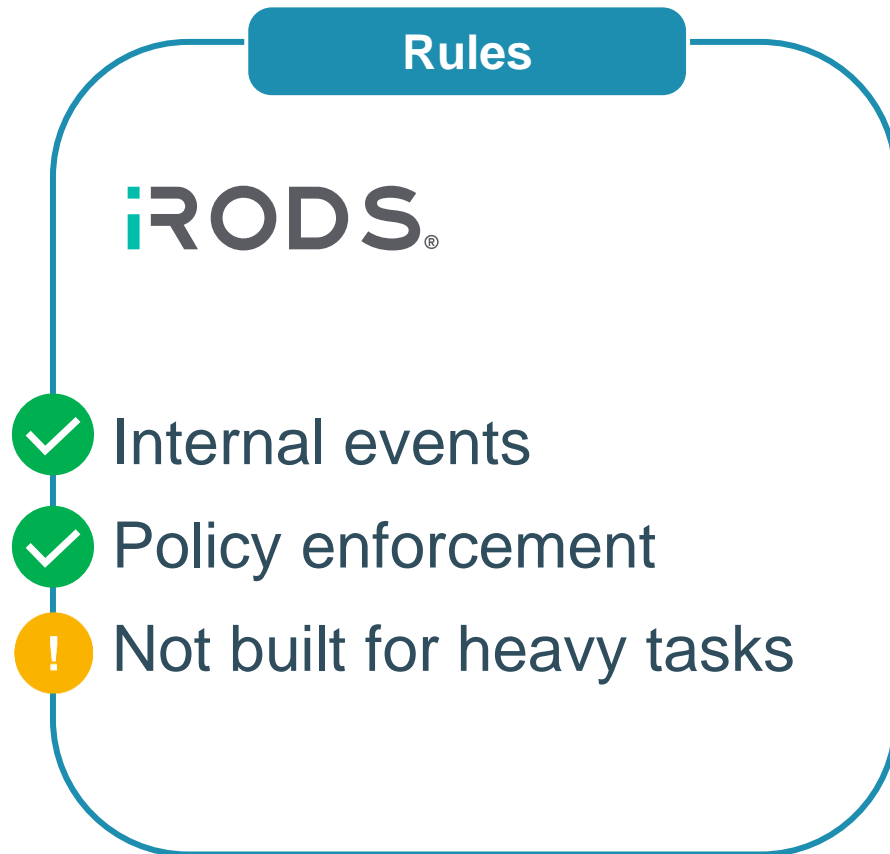
- 📜 History of objects recorded and reported
- 📜 Record of activity by request
- 📜 Register of automatic actions on logs



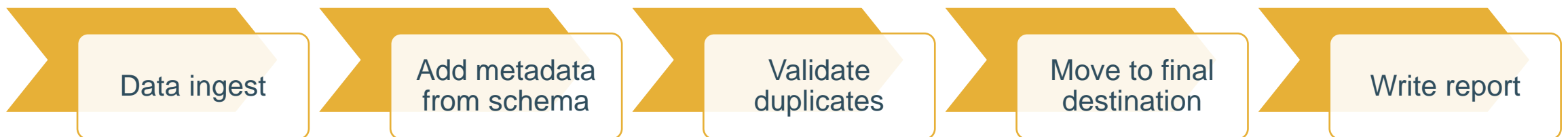
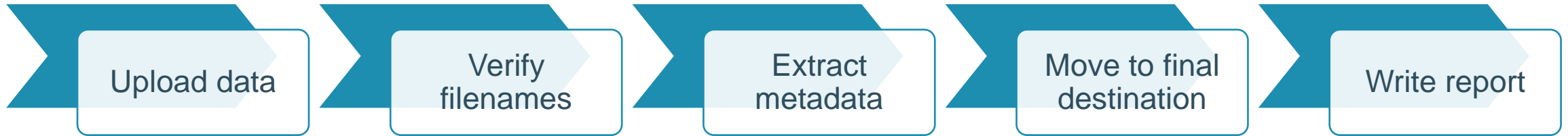
ManGO Flow



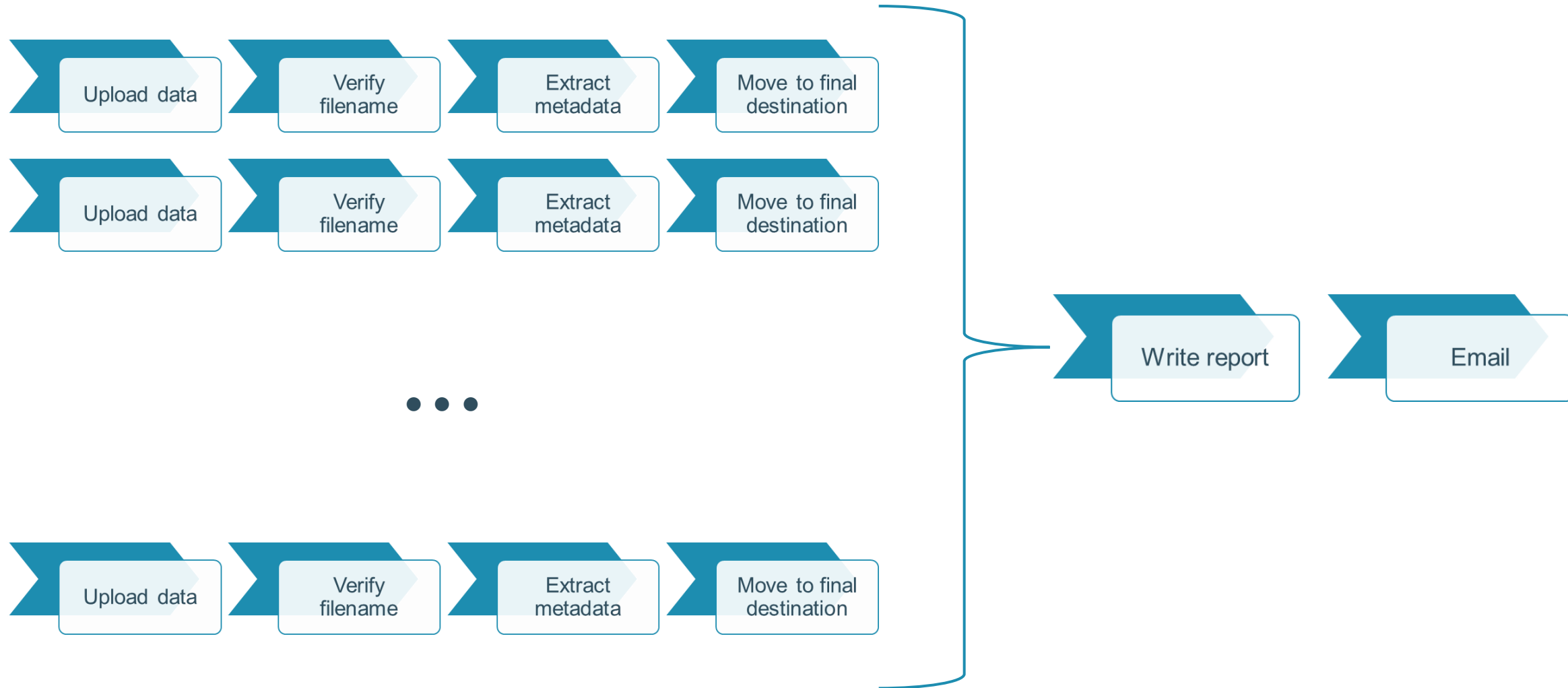
Automation: rules and ManGO Flow



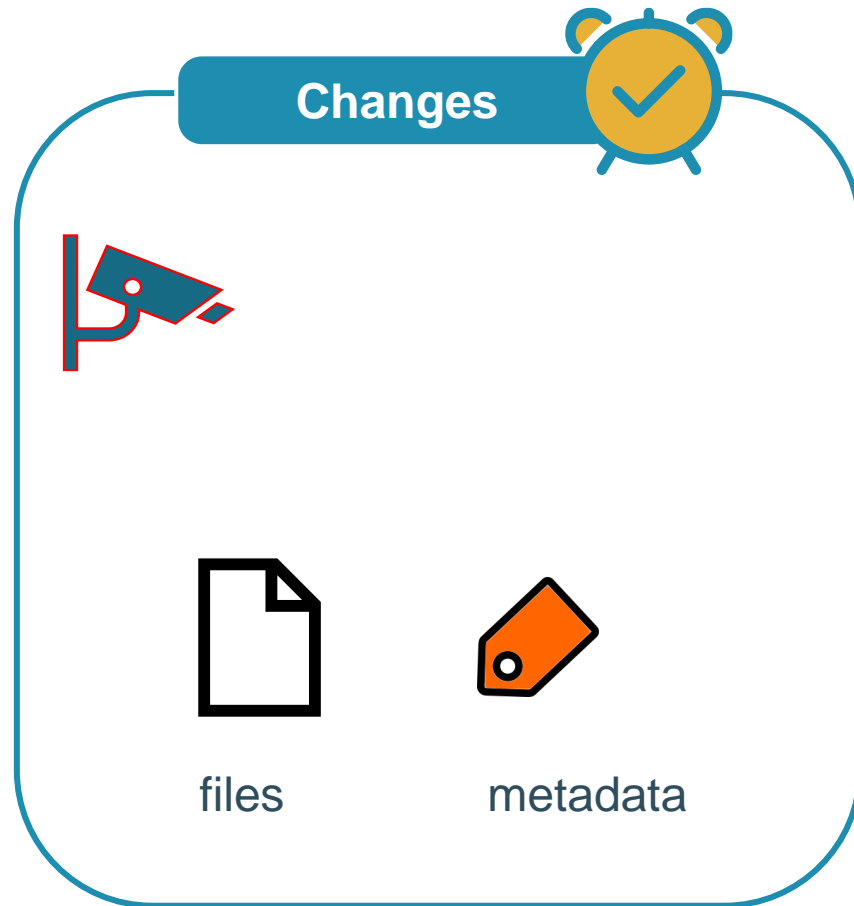
Workflows: linear composition



Workflows: aggregate pattern



ManGO *Flow* triggers



ManGO *Flow* base actions



Files

Copy
Move
Validate

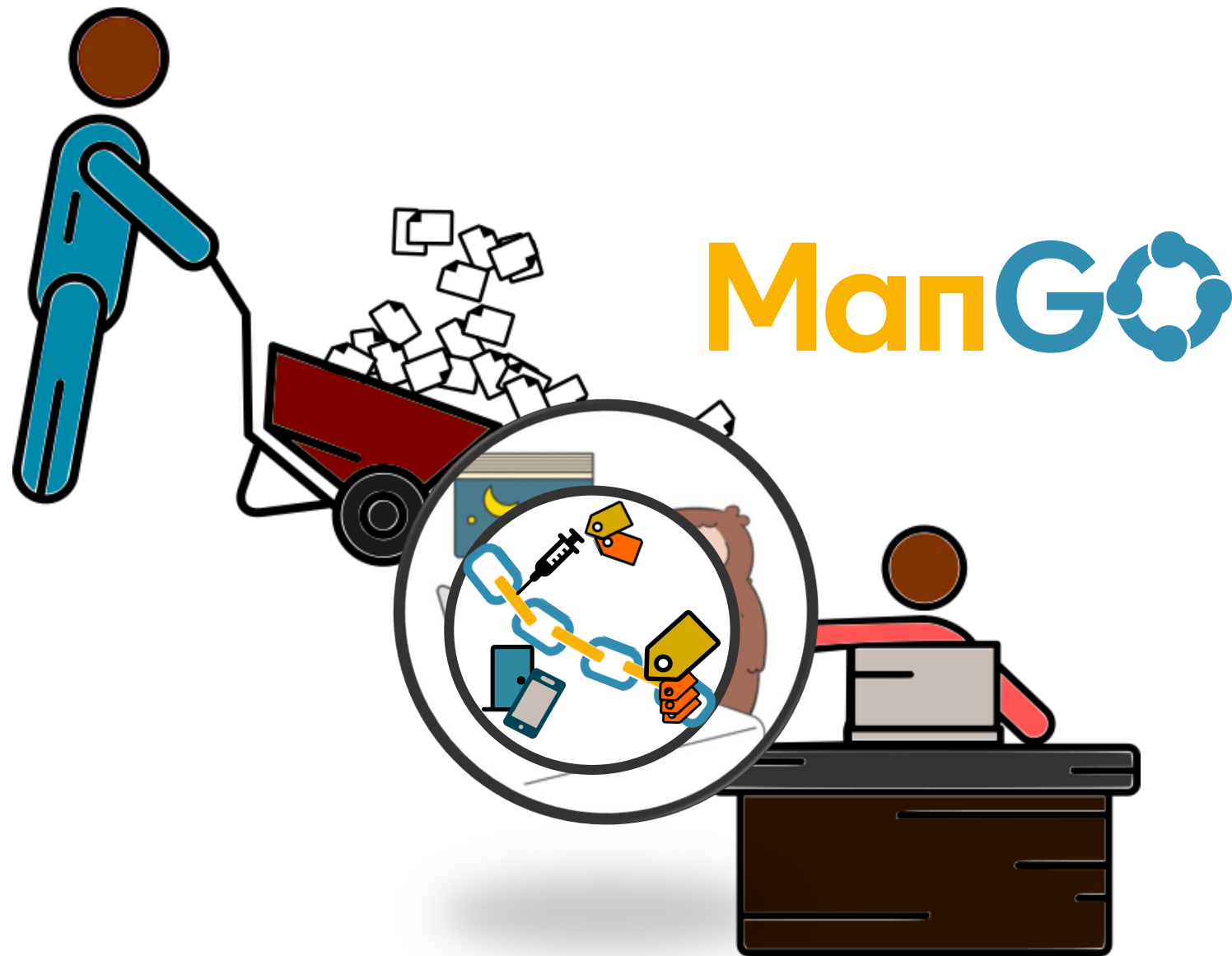
Metadata



Path
Headers
ExifTool
Apache Tika

Reporting

Metadata
Email
Report
Task Logs



What else is on the horizon

- Cold Storage
- Integration of workflows with HPC

The End

Questions?

