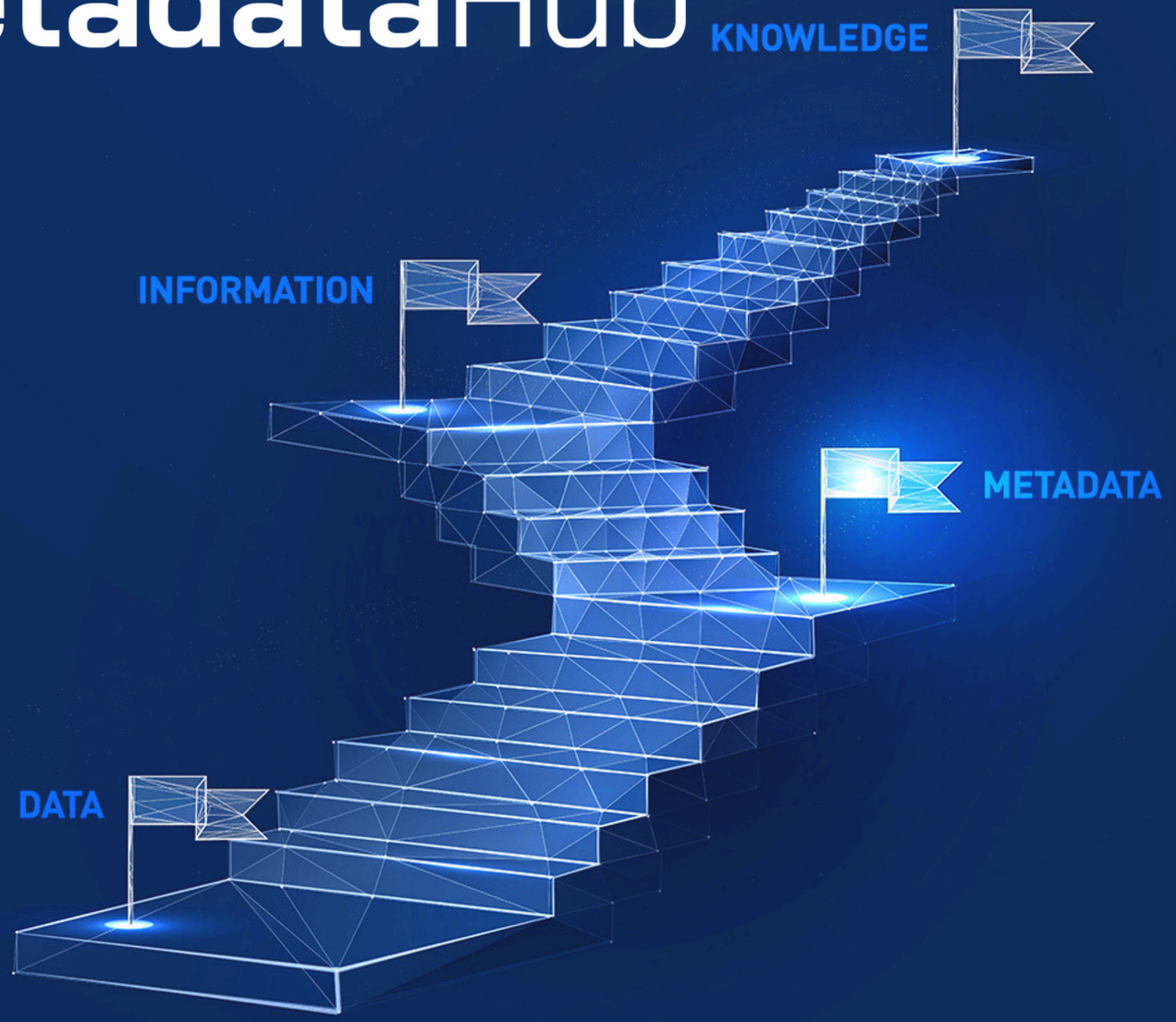
iRODS UGM Amsterdam 2024

MetadataHub

Unleash the Power of your Unstructured Data



David Cerf
Chief Data Evangelist
david@graudata.us

David Cerf

Chief Data Evangelist
GRAU DATA

DAVID@GRAUDATA.US

GRAU DATA

Your data \ Your control _

Session Goals

1. **Why is embedded metadata important**
2. **Metadata’s role in a data-driven world**
3. **How embedded metadata enhances iRODS**

What is Embedded Metadata?

POSIX Metadata

File name, file size, creation time, last access time, modification time, etc.

Embedded File Metadata

- Created by the application, defines the content and substance of the file.
- File can contain hundreds to tens of thousand metadata tags.

Machine-Generated File Content is Critical Embedded Metadata

From DNA, RNA, and Proteins to useful metadata

FASTA, FASTQ, FB Files

> 66.000 lines

Metadata

GC(%) = **50.79**

N50 = **4.641.652**

Q1 = **2.320.826**

Q2 = **4.641.652**

Q20(%) = **0**

Q3 = **2.320.826**

Q30(%) = **0**

U00096_3.Note = **Escherichia coli str. K-12 substr. MG1655**

avg_len = **4.641.652**

max_len = **4.641.652**

min_len = **4.641.652**

num_seqs = **1**

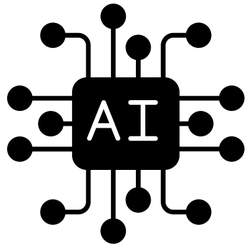
sum_gap = **0**

sum_len = **4.641.652**

type = **DNA**

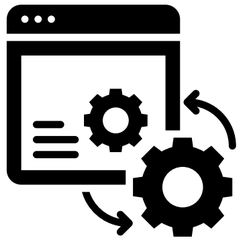
Embedded metadata is critical to data-intensive industries and for AI

Analytics



AI

Discovery



Workflow Automation

Data Orchestration

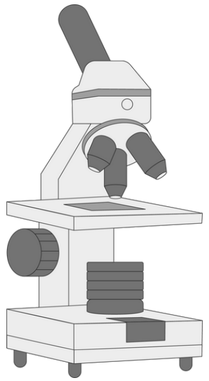
AI and ML have changed the data landscape

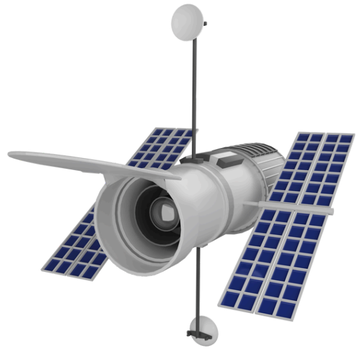
All data is needed all the time

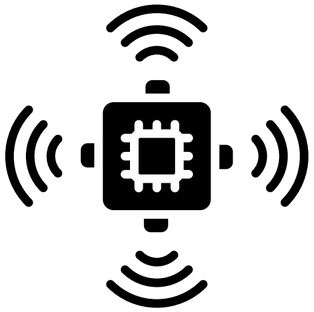
The Machine-Generated Data Problem

- 98% of all new data is machine-generated (by volume i.e., TB)
- Not typically read by humans but for computation

Data Generators



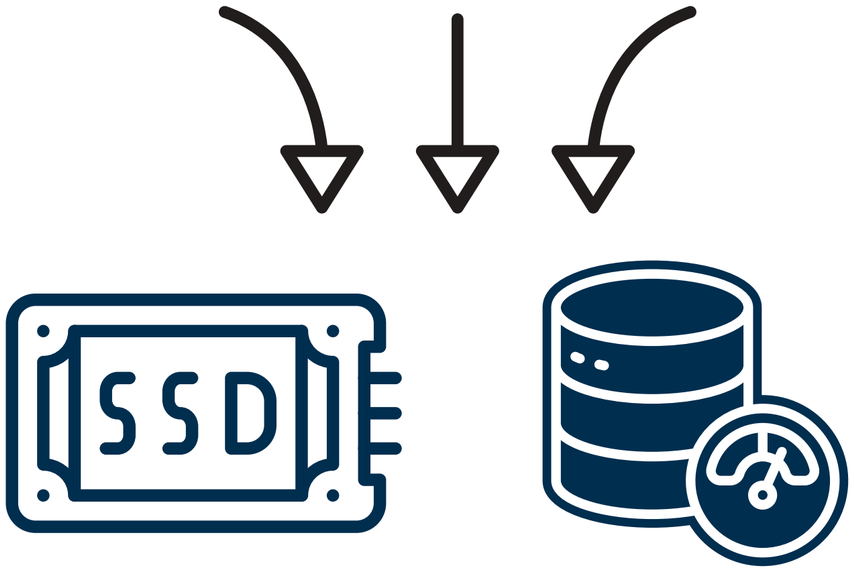




The Machine-Generated Data Problem

Machine-generated data resides on expensive storage

Data Generators



Primary Storage

The Machine-Generated Data Problem

Copies can go to archive

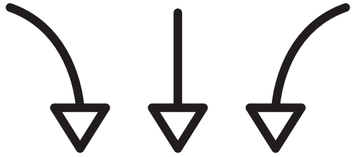
(cheap storage)

but machine-generated data

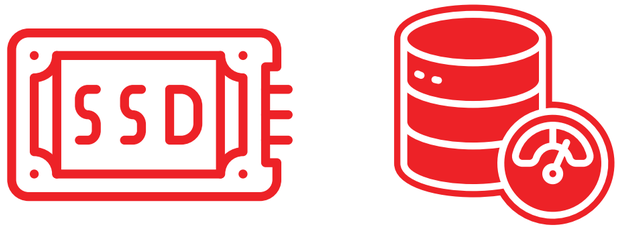
is still on primary storage

- Critical for AI / ML, analytics
- Requires fast access

Data Generators



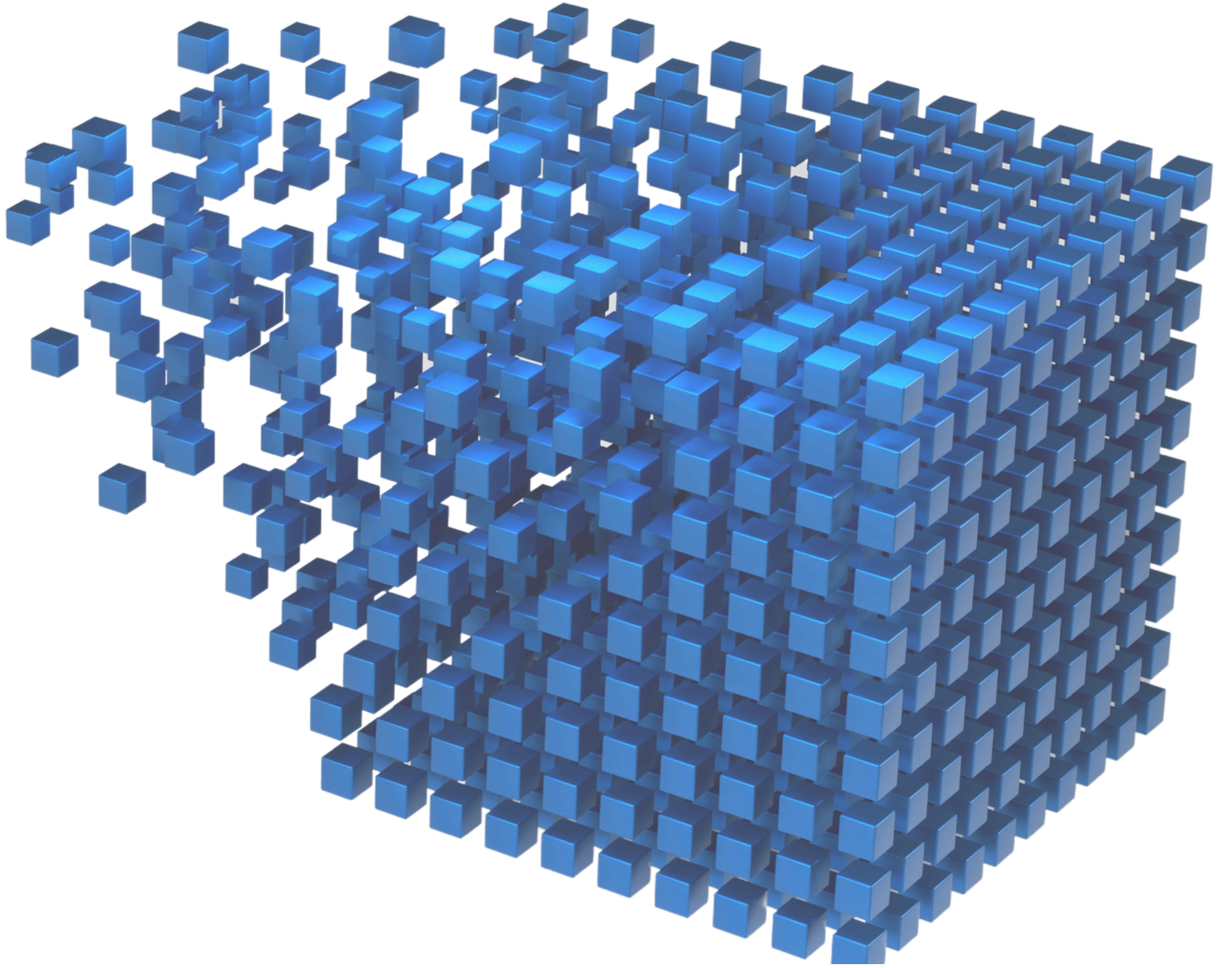
Primary Storage



Archive / Copy

Unleash the Embedded Metadata

MetadataHub



Transforms unstructured data into structured, easily searchable data sets by harvesting the embedded metadata

The Challenge of Diverse Unstructured Data Formats

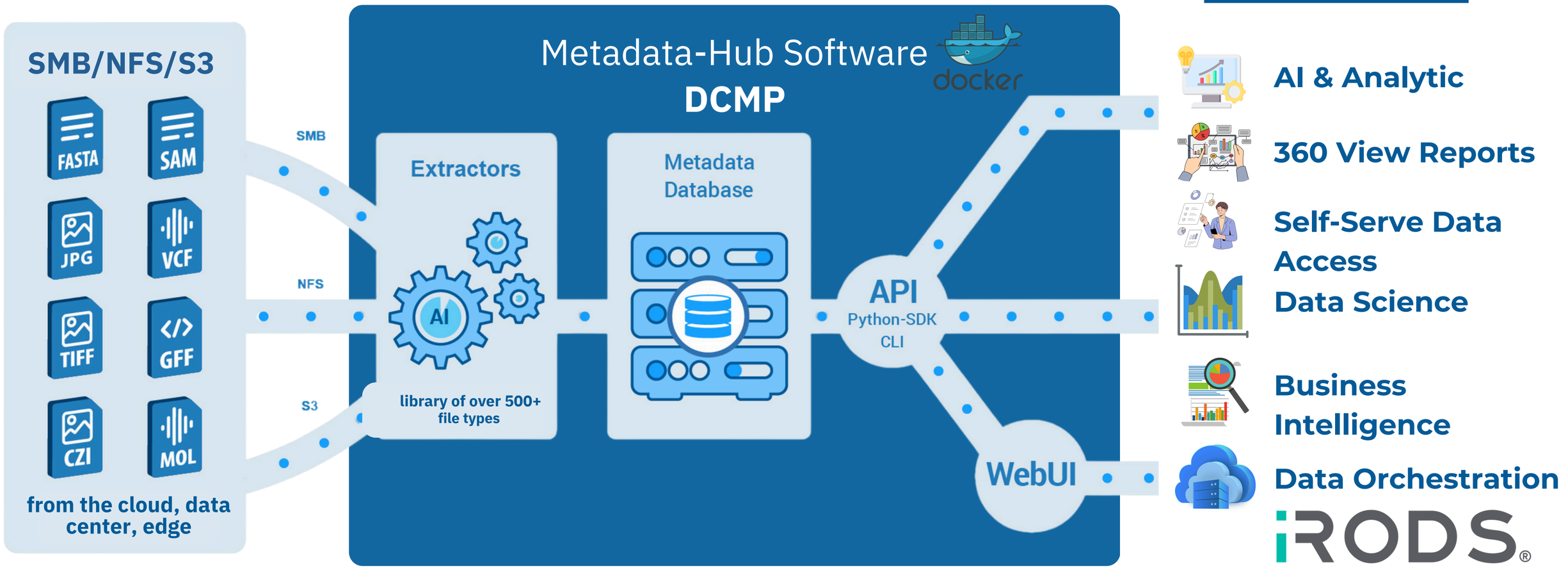
Metadata-Hub support 100's of unique file formats

360 **3FR** 3G2 **3GP** 3GP2 **3GPP** 1SCA **AA** AAE **AAX** ACFF **ACFM** ACR **AFM** AI **AIF** AIFC **AIFF** AIT **AIM** AL3D **ALI** AMFM **APE** APNG **ARF** ARQ **ARW** ASF **AVI** AVIF **AZW** AZW3 **BMP** BPG **BTF** CDF **CH5** CHM **CIF** CIFF **COS** CR2 **CR3** CRM **CRW** CS1 **CSV** CXD **CZI** DC3 **DCM** DCP **DCR** DFONT **DIB** DIC **DICM** DIVX **DJV** DJVU **DLL** DM2 **DM3** DM4 **DNG** DOC **DOCM** DOCX **DOT** DOTM **DOTX** DPX **DR4** DS2 **DSS** DV **DVB** DVR-MS **DYLIB** EIP **EMI** EPS **EPS2** EPS3 **EPSF** EPUB **ERF** ETS **EXE** EXIF **EXR** EXV **F4A** F4B **F4P** F4V **FFF** FIT **FITS** FLA **FLAC** FLEX **FLI** FLIF **FLIR** FLV **FRM** FPF **FPX** FTS **GEL** GIF **GPR** GZ **GZIP** H5 **HDP** HDR **HEIC** HEIF **HIF** HIS **HTM** HTML **I2I** ICAL **ICC** ICM **ICS** IDML **IIQ** IMG **IMS** IND **INDD** INDT **INR** INSP **INSV** INX **ISO** ITC **J2C** J2K **JNG** JP2 **JPC** JPE **JPEG** JPF **JPG** JPM **JPS** JPX **JSON** JXL **JXR** K25 **KDC** KEY **KTH** LA **LFP** LFR **LIF** LNK **LRV** M2T **M2TS** M2V **M4A** M4B **M4P** M4V **MACOS** MAX **MDF** MEF MIE **MIF** MIFF **MKA** MKS **MKV** MNG **MOBI** MODD **MOI** MOS **MOV** MP3 **MP4** MPC **MPEG** MPG **MPO** MQV **MRC** MRW **MSR** MTS **MXF** NAF **ND2** NEF **NEWER** NII **NKSC** NMBTEMPLATE **NRW** NUMBERS **O** ODB **ODC** ODF **ODG** ODI **ODP** ODS **ODT** OFR **OGG** OGV **ONPOPUS** ORF **ORI** OTF **PAC** PAGES **PBM** PCD **PCT** PCX **PDB** PDF **PEF** PFA **PFB** PFM **PGF** PGM **PICT** PLIST **PMP** PNG **PNL** POT **POTM** POTX **PPAM** PPAX **PPM** PPS **PPSM** PPSX **PPT** PPTM **PPTX** PRC **PS** PS2 **PS3** PSB **PSD** PSDT **PSP** PSPFRAME **PSPIMAGE** PSPSHAPE **PSPTUBE** QIF **QT** QTI **QTIF** R3D **RA** RAF **RAM** RAR **RAW** REC **RIF** RIFF **RM** RMVB **RPM** RSRC **RTF** RV **RW2** RWL **RWZ** SCN **SEQ** SER **SIF** SKETCH **SLD** SO **SR2** SRF **SRW** SVG **SVS** SWF **THM** THMX **TIF** TIFF **TORRENT** TS **TTC** TTF **TUB** TXT **VCARD** VCF **VOB** VRD **VSD** VSI **WAV** WDP **WEBM** WEBP **WMA** WMV **WOFF** WOFF2 **WTV** WV **X3F** XCF **XDCE** XHTML **XLA** XLAM **XLEF** XLS **XLSB** XLSM **XLSX** XLT **XLTM** XLTX **XMP** XQD **XQF** XV **ZIP** **ANY MANY MORE …**

NOTE: ALL SUB-FORMATS OF THE LISTED FILES ARE SUPPORTED TOO ➔ FOR EXAMPLE: JPG ALONE HAS 41 SUB-FORMATS

Metadata-Hub: Creating a Metadata Fabric

Storage



Data Users

AI & Analytic

360 View Reports

Self-Serve Data Access
Data Science

Business Intelligence

Data Orchestration

iRODS®

iRODS Metadata Powered Workflow

Machine-Generated Data

Performance Storage

Metadata Harvested

Metadata sent to iRODS
Instructions to move files to archive

iRODS®

Files are moved from performance storage to the archive upon creation

Archive

Object Storage and Tape

Cloud Storage

GRAU DATA
Your data \ Your control _

MetadataHub Helps Free Disk Capacity

Machine-Generated Data

25% — 75%

Improvement in Storage

Performance Storage

iRODS®

Metadata Harvested

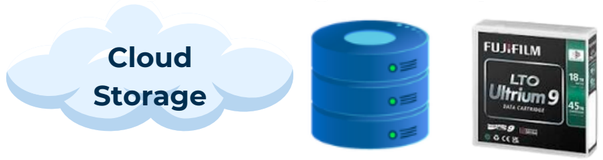
Metadata sent to iRODS
Instructions to move files to archive

100%

All data is allways in the Archive

Archive

Object Storage and Tape



GRAU DATA
Your data \ Your control _

Feeding Data to Data Users

Machine-Generated Data

Performance Storage

Metadata Harvested

MetadataHub becomes the proxy for:

Query, Search, Provisioning

Data Users

- AI & Analytic
- 360 View
- Self-Serve Analytics
- Data Science
- Business Intelligence
- Data Orchestration

metadata

iRODS

Archive

Object Storage and Tape

Cloud Storage

Files can be re-called if ever required

GRAU DATA
Your data \ Your control _

S3 Glacier Tape Archive

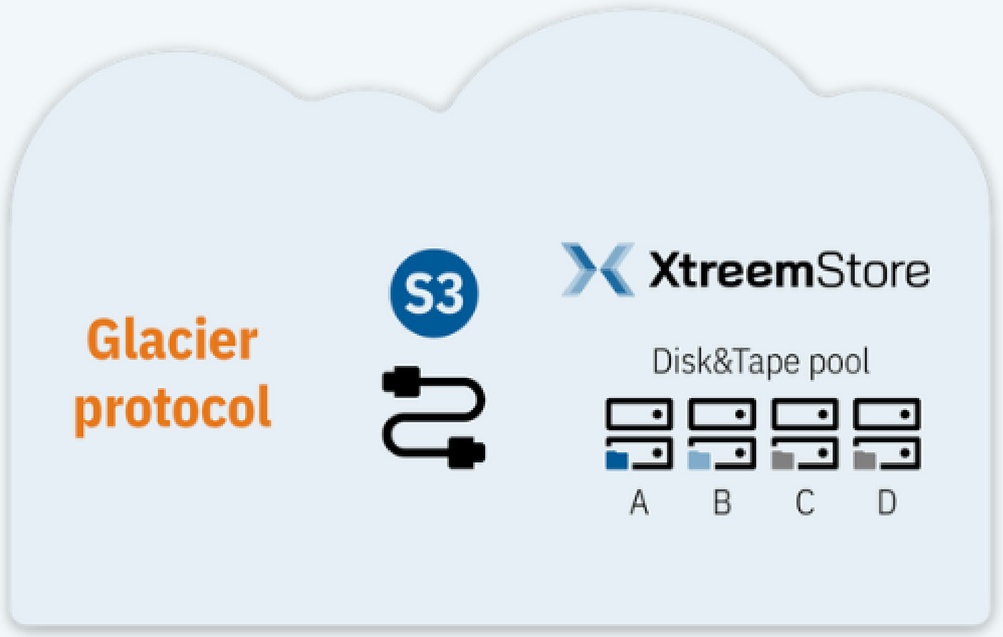
S3 / Glacier

XtreemStore

iRODS

TURN ANY TAPE LIBRARY INTO A GLACIER CLOUD ARCHIVE

- Simplicity - S3 & Glacier compatible
- Open standards and vendor-neutral hardware
- No hidden costs



GRAU DATA

Applications, Data Lakes, Reporting, etc

SMB, NFS, S3, API

API, SDK, CLI

Data Management & Orchestration

iRODS

Metadata Provisioning

powered by metadata

Global Metadata Repository

Server 1

Metadata-Hub
Discover the value of your data

Hub 1 | Hub 2 | Hub 3

Server 2

Metadata-Hub
Discover the value of your data

Hub 4 | Hub 5

Metadata Harvesting *

WAN

XtreemStore

SMB/NFS | SMB/NFS | SMB/NFS | S3 | SQL

Share 1 | Share 2 | Share 3 | Share 4 | Share 5 | Share 6 | Share 7 | Bucket | Database
Media managers

Edge

Data center

Cloud

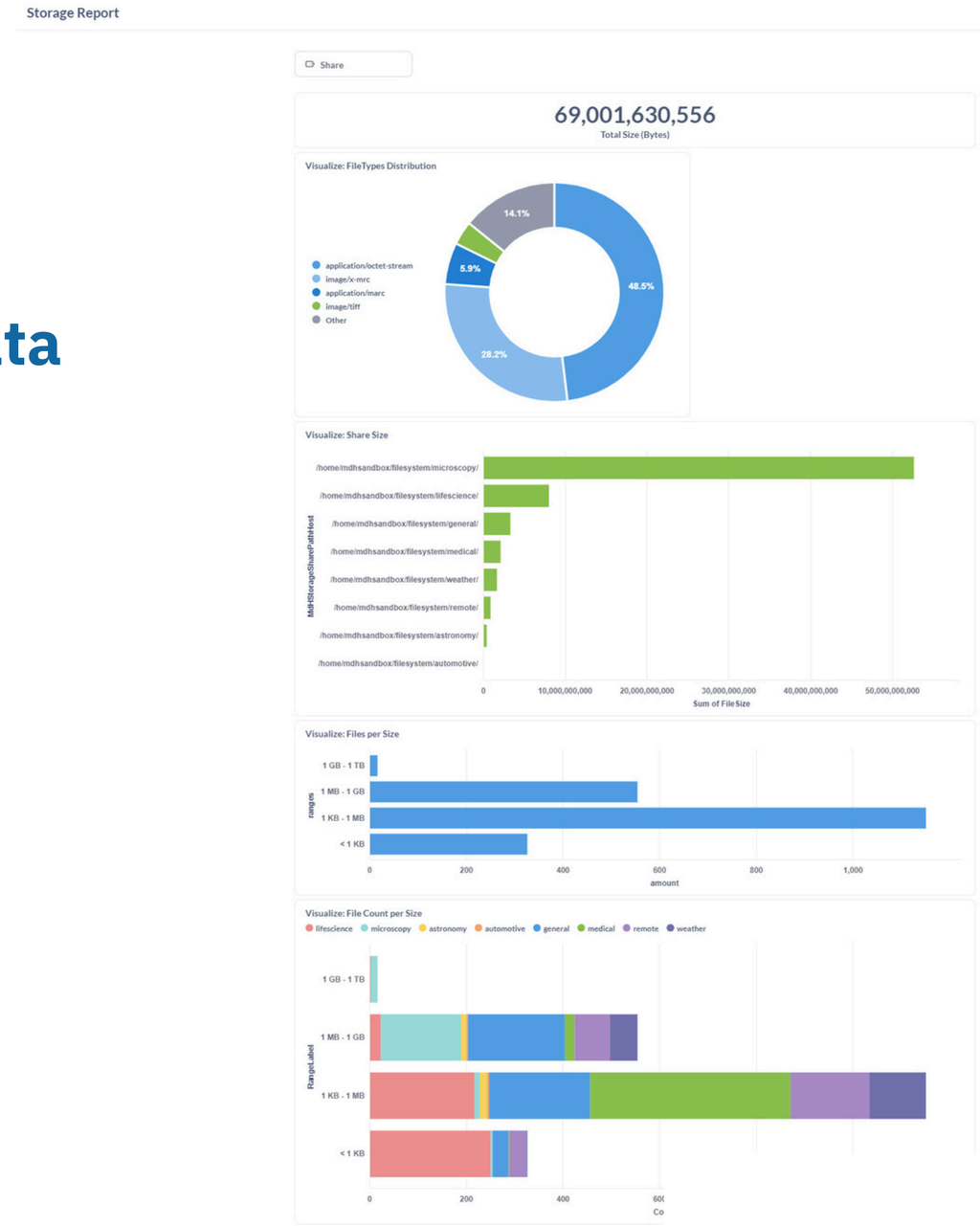
GRAU DATA
Your data \ Your control _

DATA LANDSCAPE REPORTS

Gain Unparalleled Visibility into Your Data Assets

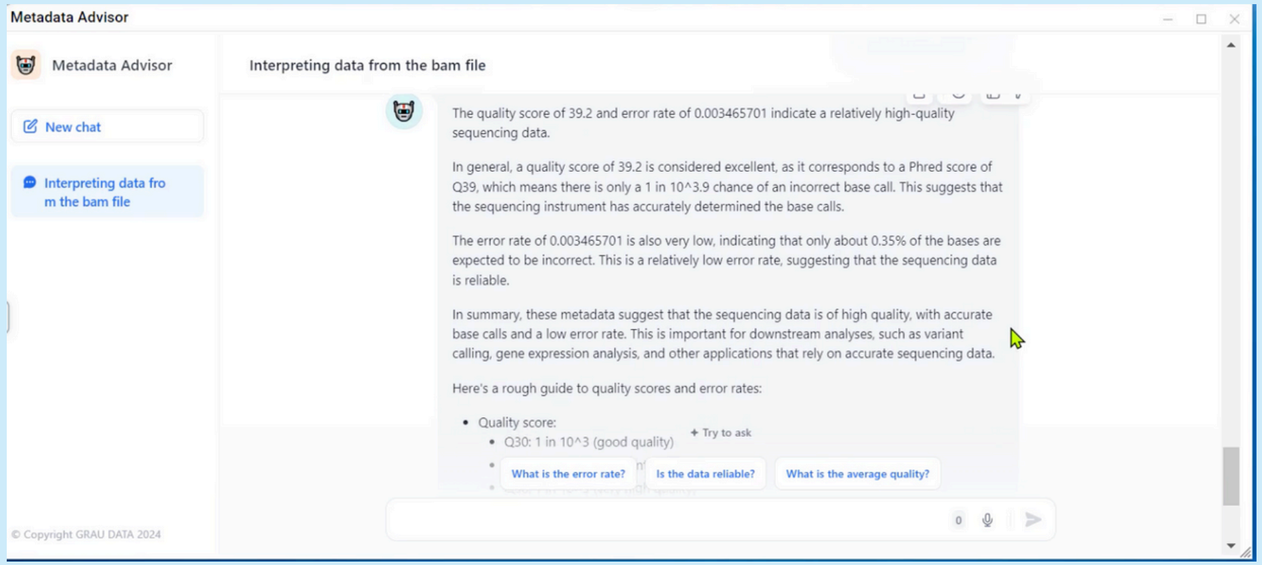
Unified view of data across cloud, data centers, and edge devices

Insights into data distribution, storage utilization, and growth trends



GRAU DATA
Your data \ Your control _

METADATA ADVISOR



AI-powered features to provide users with advanced insights into their unstructured data and metadata.

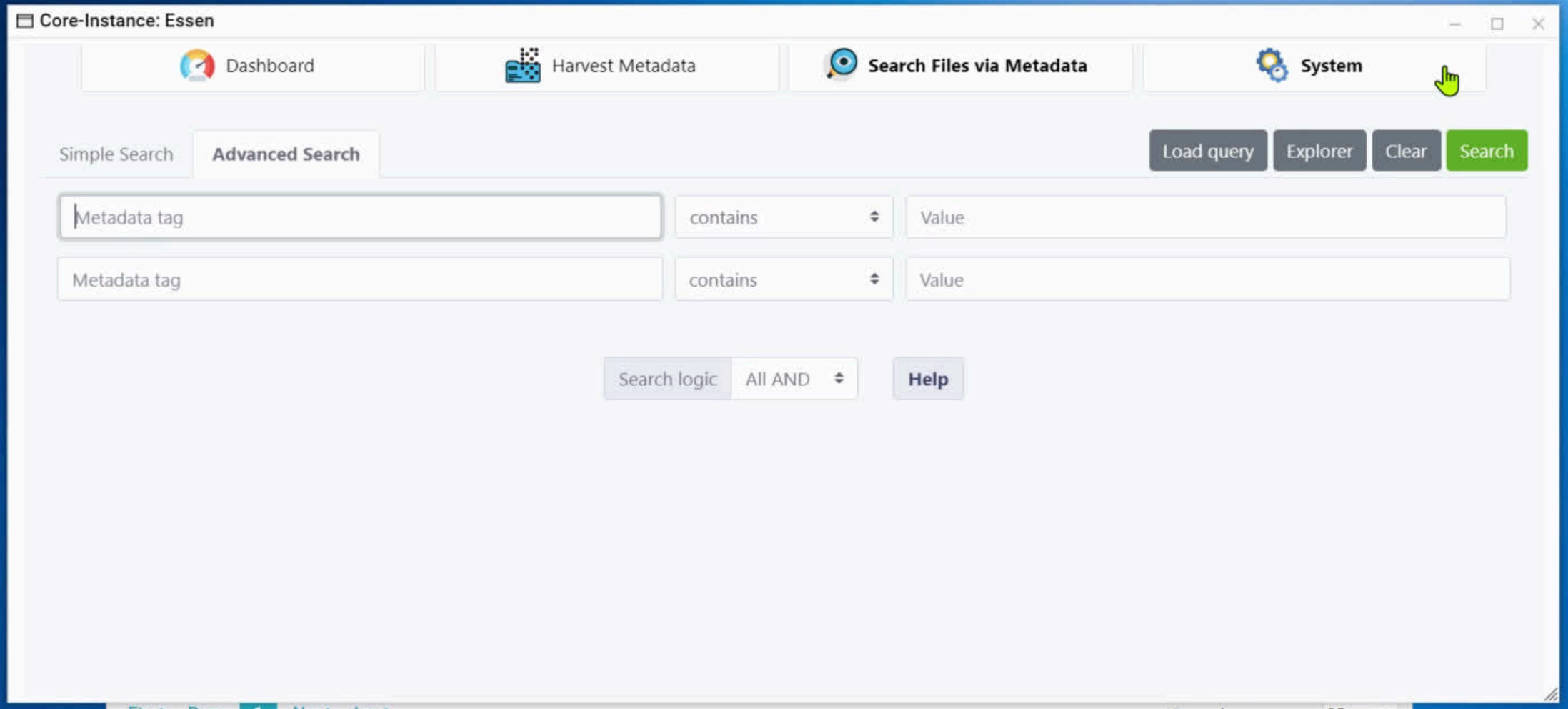
Provides expert guidance to:

- **Interpretation** metadata extracted by Metadata-Hub.

.

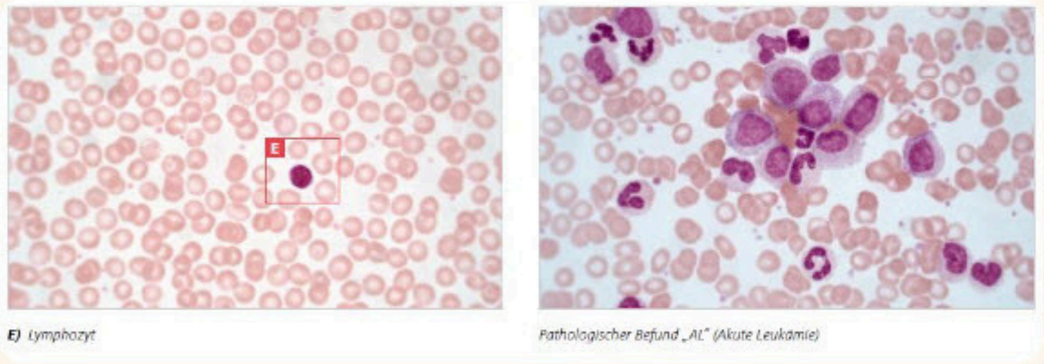
- **Confidence in Data Usage**: Make informed use of your data by ensuring a clear grasp of its context and significance.

GRAU DATA



MENU

STUDIO: DEMO-SERVER



MetaSens:

Creates text descriptions and keywords from files content.

Prompt:
Describe the images and compare them.

The image shows two microscopic views of blood smears, which are used for various medical tests. The left side is labeled "E) Lymphozyt" (German for "lymphocyte"), indicating that the image is focused on lymphocytes, a type of white blood cell. The right side is labeled "Pathologischer Befund, AL (Aktue Leukamie)" (German for "Pathological finding, AL (Acute Leukemia)"), suggesting that the image is showing a pathological finding related to acute leukemia.

In the left image, you can see a cluster of lymphocytes, which are typically round or oval-shaped cells with a distinct nucleus. In the right image, there is a more disorganized pattern of cells, with some cells appearing abnormal or atypical, which could be indicative of a leukemia diagnosis. The cells in the right image are more irregular in shape and size, which is characteristic of leukemia cells.

These images are likely used for diagnostic purposes in a clinical setting to identify and classify different types of blood cells and to detect abnormalities that could be indicative of a disease like leukemia.

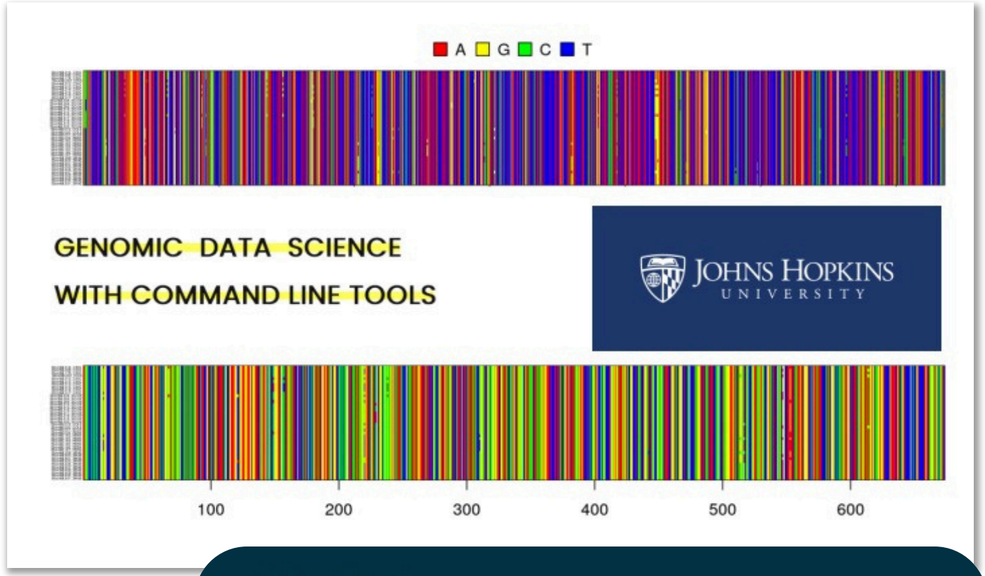
Use Cases


MdH Studio
GLOBAL SEARCH
METADATA CORE
METADATA CORE
METADATA CORE
METADATA CORE
METADATA CORE
METADATA CORE
METADATA CORE

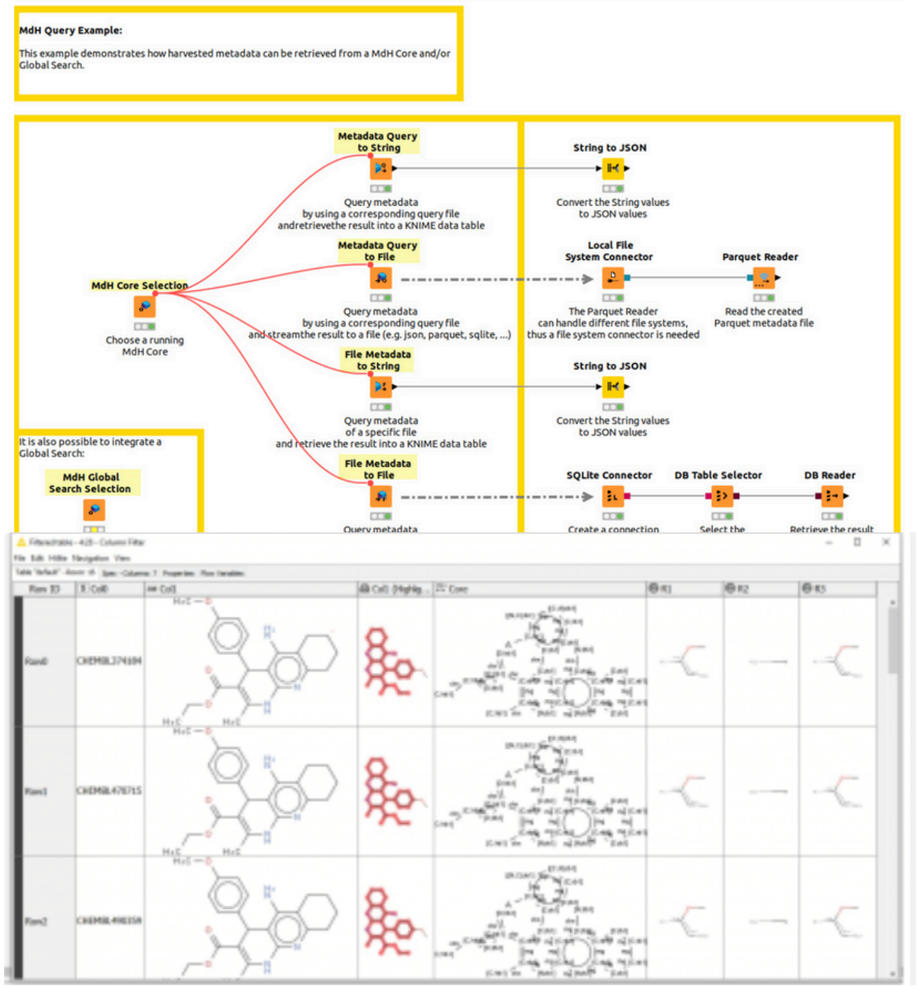
Use Case: ZUSE Institute Clinical Center

Challenges

1. **Improve data accessibility, search and analysis**
2. **Support Diverse Format Extraction**:
   - **Medical Imaging**: NIFTI (brain scans), DICOM (CT & MRI images)
   - **Genomic Formats** (FASTA, FASTQ, SAM, VCF): enable precise genetic data analysis
   - **Proteomics Format** (PDB): in-depth protein structure studies
3. **Enable collaboration in accordance with FAIR practices**
4. **Reduce Data Preparation**
5. **Security & Governance**



ZUSE Institute - Management of Research Data / Life Science Data



Results

Enhancing Workflow Automation:

- **Streamlined data preparation** saving countless hours
- **Global data accessibility** to all metadata enforced by access controls
- **Accelerated insight**

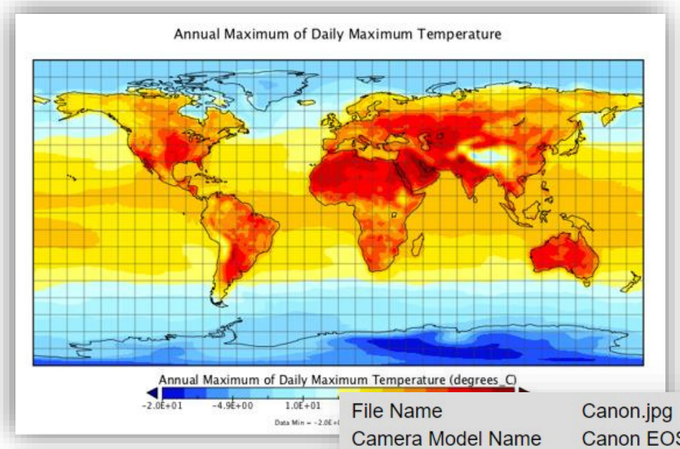
Optimizing Research Data Management:

- Holistic Data Landscape View
- Reduced Data Transfers
- Improved storage untilization

Use Case: Weather & Climate

- Climate research –NetCDF , GRIB, BUFR, ZARR, GRID
  - File format which saves multidimensional scientific data such as temperature, humidity, pressure, wind speed and wind direction

- Drone shots –RAW Image Format
  - Master metainformation from around 7000 different camera models of 105 camera manufacturers
  - Used in various industries such as forensics, geo-information services, logistic, agronomics usw.



File Name	Canon.jpg
Camera Model Name	Canon EOS DIGITAL REBEL
Date/Time Original	2003:12:04 06:46:52
Shooting Mode	Bulb
Shutter Speed	4
Aperture	14.0
Metering Mode	Center-weighted average
Exposure Compensation	0
ISO	100
Lens	18.0 - 55.0 mm
ocal Length	34.0 mm
age Size	8x8
uality	RAW
ash	No Flash
hite Balance	Auto
ocus Mode	Manual Focus (3)
ontrast	+1
harpness	+1
aturation	+1
olor Tone	Normal
olor Space	sRGB
e Size	2.6 kB
e Number	118-1861
rive Mode	Continuous Shooting
Owner Name	Phil
Serial Number	0560018150

Use Case: Max Planck Institute for Radio Astronomy

Challenges

Initial 45 PB of storage (telescope images, blueprints, etc.).

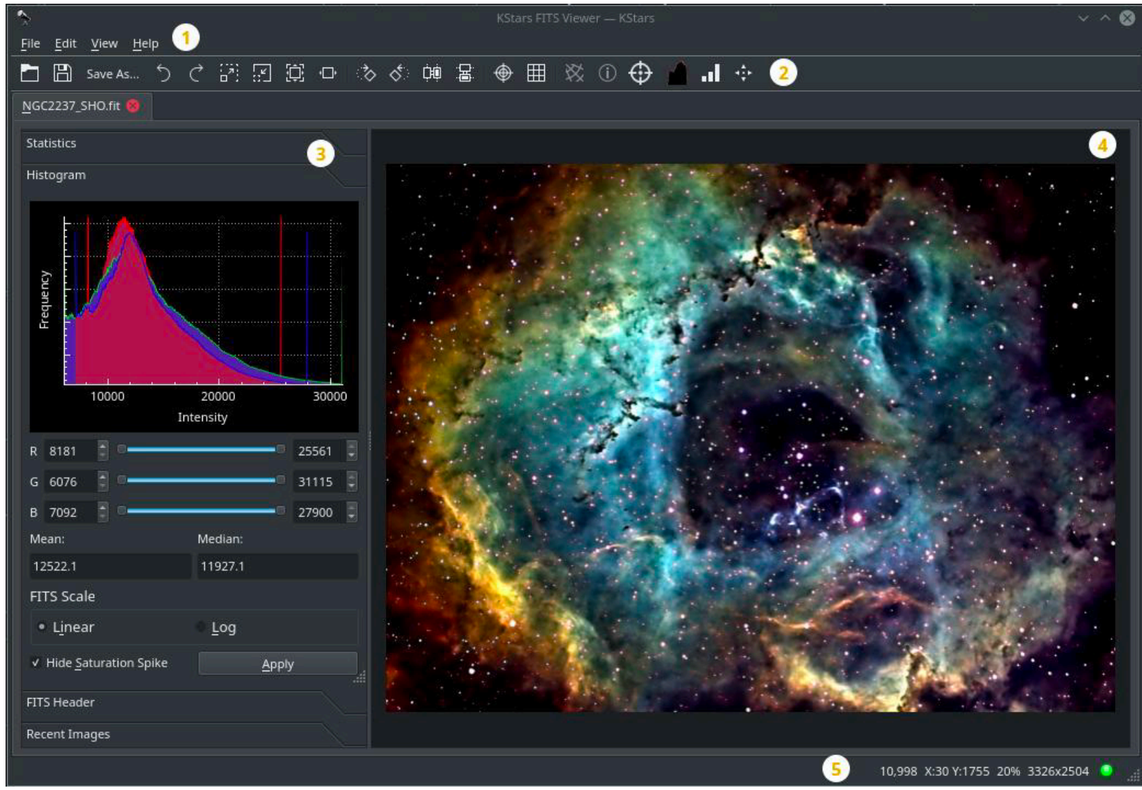
“No idea what data we have or where it is!”

No transparency to billions of files → "a black hole"

No possibility of filtering out relevant data from complex, heterogeneous datasets

Requirement:

1. Organize, catalog, and share research data effectively.
2. Improve collaboration, reproducibility, and data integrity.



FITS
(Flexible Image Transport System) was designed specifically for astronomical data, and includes provisions such as describing photometric and spatial calibration information, together with image origin metadata.

More info: fits.gsfc.nasa.gov

GRAU DATA
Your data \ Your control _

Use Case: Max Planck Institute for Radio Astronomy

MAX PLANCK GESELLSCHAFT

Results

Metadata-Hub created a centralized metadata warehouse.

- Enforcing data governance policies, and tracking data lineage.
- Streamlining data migration, integration, and compliance processes.
- Created custom extractors.

Enabled Metadata Mart:

- Extract embedded metadata from FITS file format (NASA.org).
- Reduced data prep in evaluation and search.
- Automated data provisioning to Jupyter Notebooks.
- Introduced “no-code/low-code” self-service data access.

CONTACT

David Cerf

Chief Data Evangelist

david@graudata.us

Phone: +1 972 896 3164

Nina Mangold

Channel Manager

nina.mangold@graudata.com

Phone: +49 7171 187-125

Learn more at:

www.Moremetadata.com

www.graudata.com

GRAU DATA

Your data \ Your control _