ARCITECTA®
OPERATING SYSTEMS FOR META+DATA

RMACC

Metadata, the hidden enabler of advanced data management

Graham Beasley | 2022

Metadata is Awesome!

- What is metadata, history, analog examples, why now?
- Why talk about metadata at an HPC conference? Hint, it is a key ingredient for efficient Data Lifecycle Management
- Efficient HPC Scratch Storage use
- Metadata for compliancy and privacy protection
- Standards and Working Groups
- Enabler for other technologies such as AI

RMACC

What is metadata?

Metadata is Awesome!

- Why? It helps you find and organize things!
- Optimally!
- When you need them, even automatically!
- Other names “tags”, “search keys”, fields

Everyday examples of metadata?

- Metadata has been in use for a couple of centuries, we are just moving to the digital version.
- Card catalog since 1862 at Harvard
- Great example of a “database”
- Catalogs by Titles, authors, subjects
- Search Keys
- iPod, FM song + artist, etc.

Image courtesy of New York Society Library

Types of metadata

- "System Metadata"
  - File name, size, create, access, modify time, ownership permissions, etc.
  - In Unix/linux world this information comes from inodes and is often used by Backup or HSM software.
- Embedded File Metadata
  - Typically parsed out via MIME type.
- User Defined Metadata
  - This enables data life cycle management, notes, accounting.

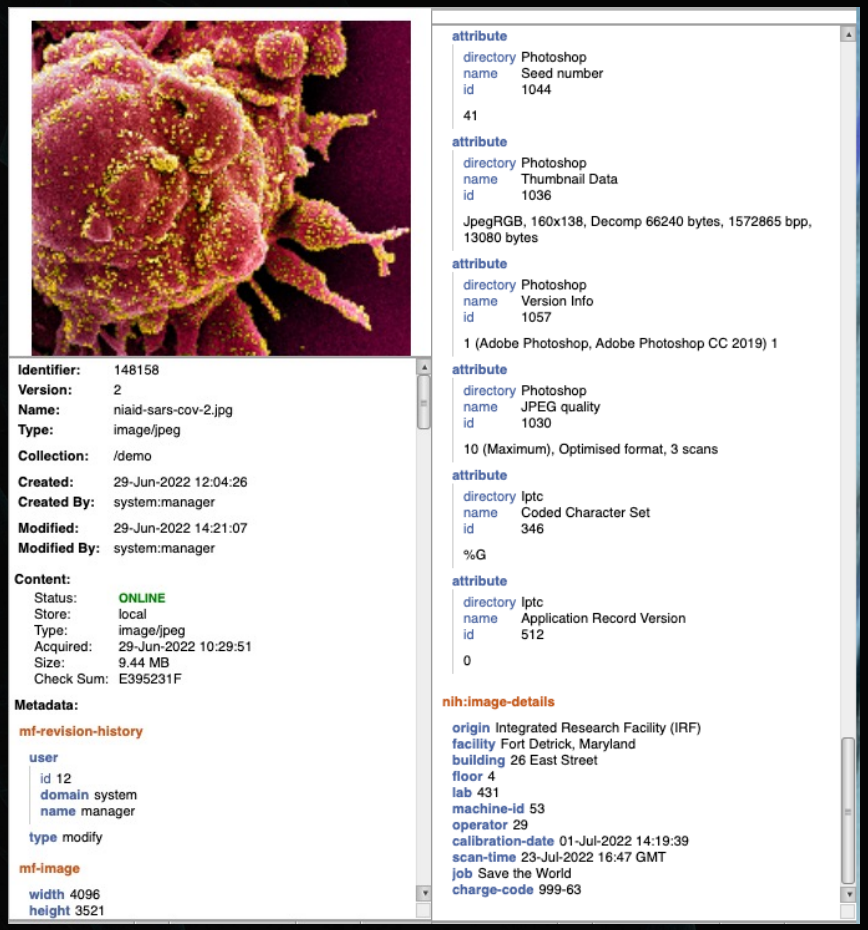


Image courtesy of NIH
National Institute of Allergy and Infectious Diseases (NIAID)

Metadata needs to Evolve

Nobody knows what the future will bring, The perfect metadata now might not be perfect in 2 years!

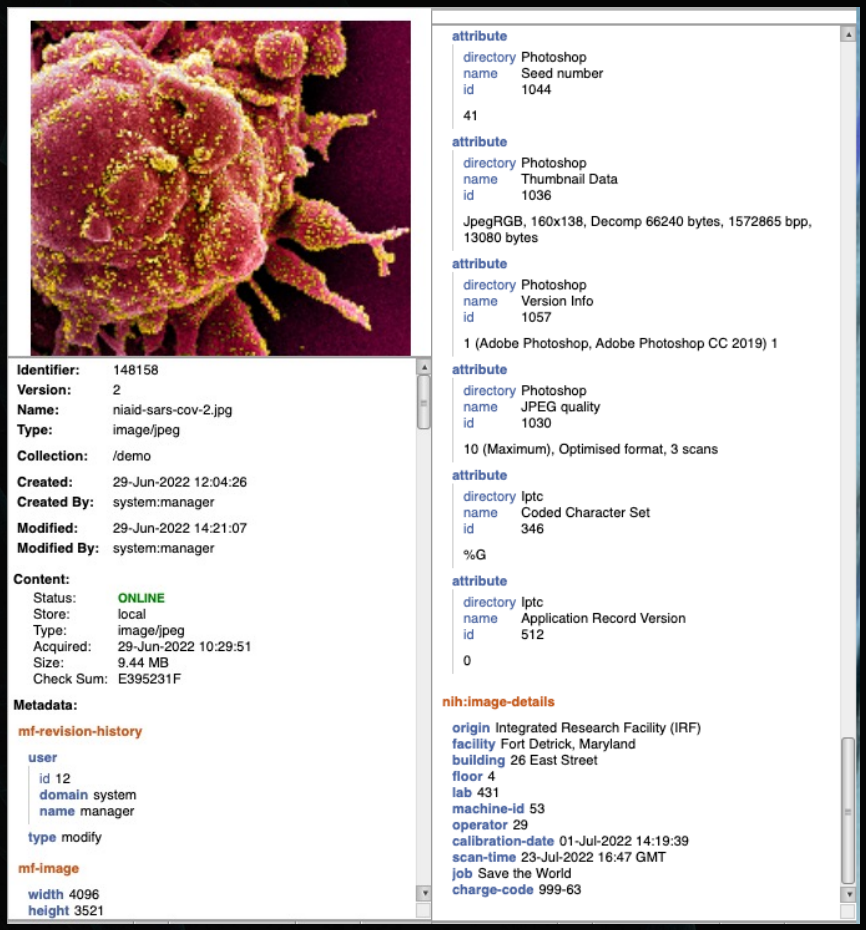
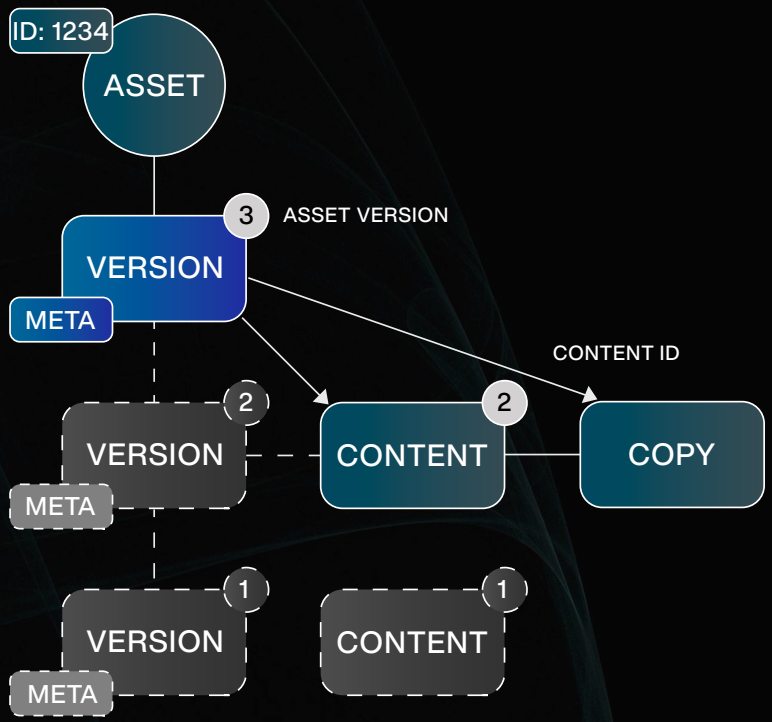


Image courtesy of NIH
National Institute of Allergy and Infectious Diseases (NIAID)

Digital Object model of an “ideal metadata db”

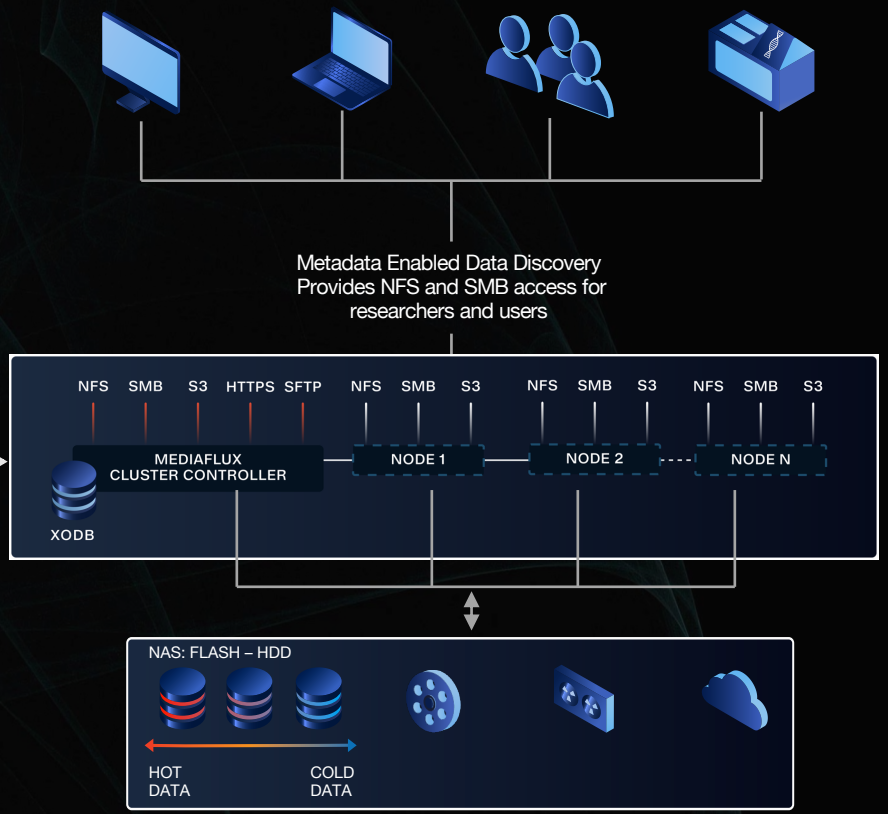
- A binary optimized database for the metadata is managed independently of the content.
- Data is not “held hostage”.
- User defined Metadata and System Metadata. Let’s take a closer look…



Data Lifecycle Management

Hybrid environment for HPC and research data management

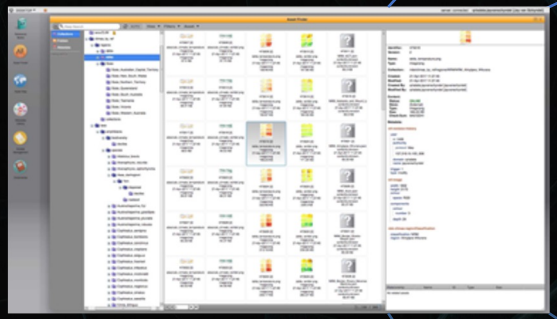
HPC CLUSTER AND
“SCRATCH” STORAGE



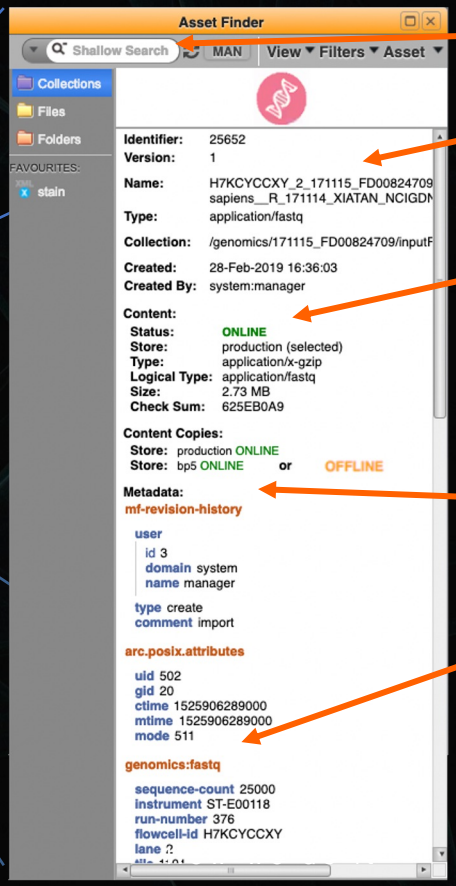
VIRTUALIZED STORAGE –
SINGLE GLOBAL NAMESPACE

Data Lifecycle Management

Access via a browser-based interface



Mediaflux Desktop



File Browser GUI Asset Finder searches metadata fields of interest to filter and create compound queries

File system level information (metadata) such as: Name, Location, Creation Date, etc.

Content information, checksum, size and data accessibility

ONLINE – content is available on nearline storage

OFFLINE – content is on deep storage (on BlackPearl Tape)

ONLINE+OFFLINE – content is available on both

MISSING – content is missing

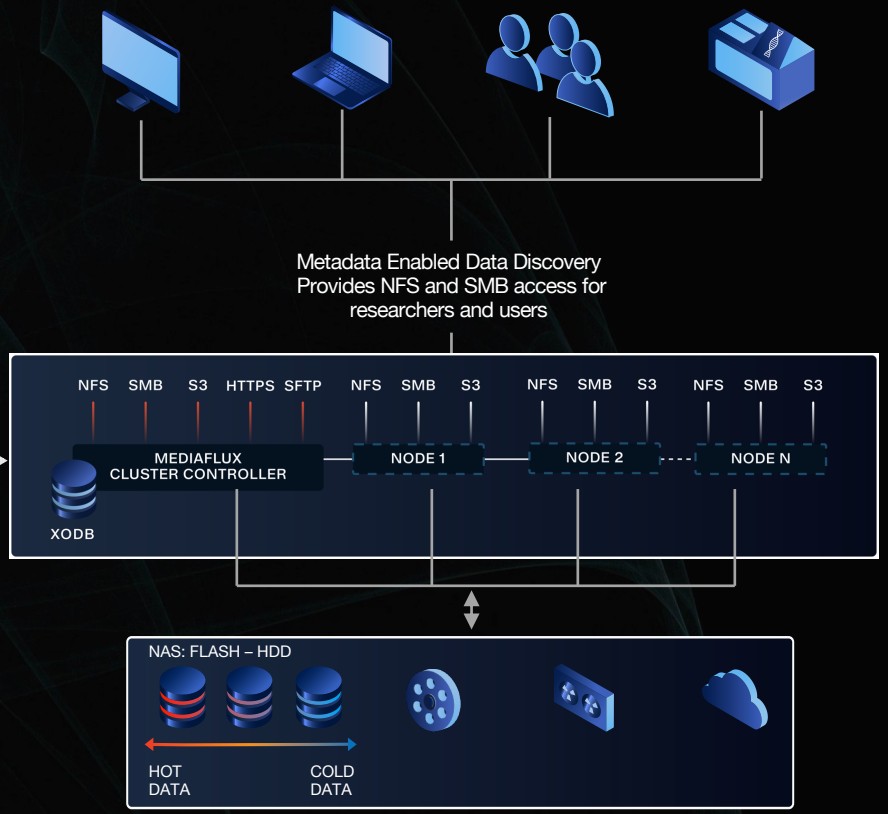
Revision history, audit trails, provenance and reproducibility

- Descriptive metadata
- Extracted during the ingest process
- Custom-built scraper that extracts specific metadata based on the data type. These plug-in content analyzers extract image type and resolution, scanner instrument information, project, study, population, PI, etc.
- User generated metadata such as: annotations, labels, tags, comments, and workflow-specific actions

Data Lifecycle Management

Feed the Beast with Data Wrangling tools

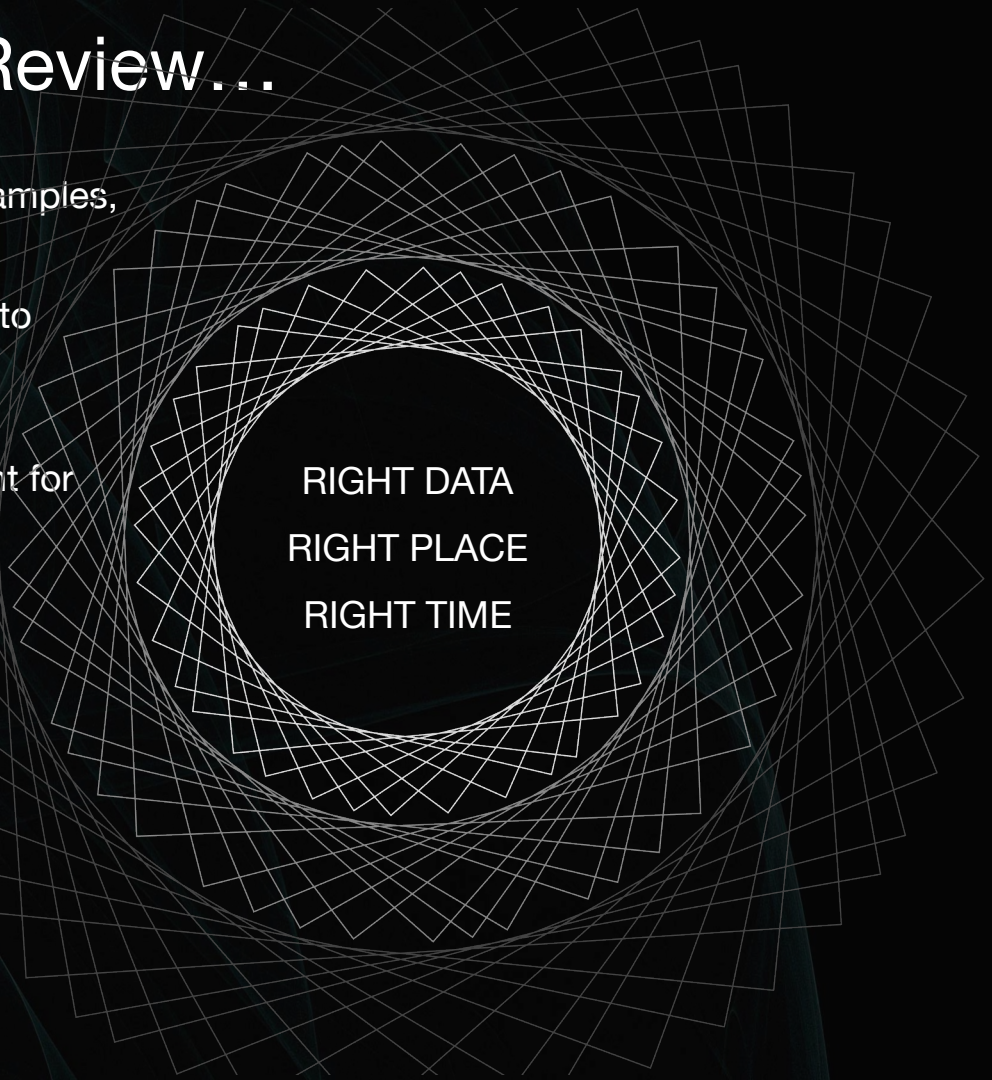
HPC CLUSTER AND
“SCRATCH” STORAGE



VIRTUALIZED STORAGE –
SINGLE GLOBAL NAMESPACE

Time Check! Quick Review…

- ✓ What is metadata, history, analog examples, why now?
- ✓ Different types of metadata and how to efficiently manage it.
- ✓ Why talk about metadata at an HPC conference? Hint, it is a key ingredient for efficient Data Lifecycle Management
- ✓ Efficient HPC Scratch Storage use
- Metadata for compliancy and privacy protection
- Data Explosion and cost of storage
- Standards and Working Groups

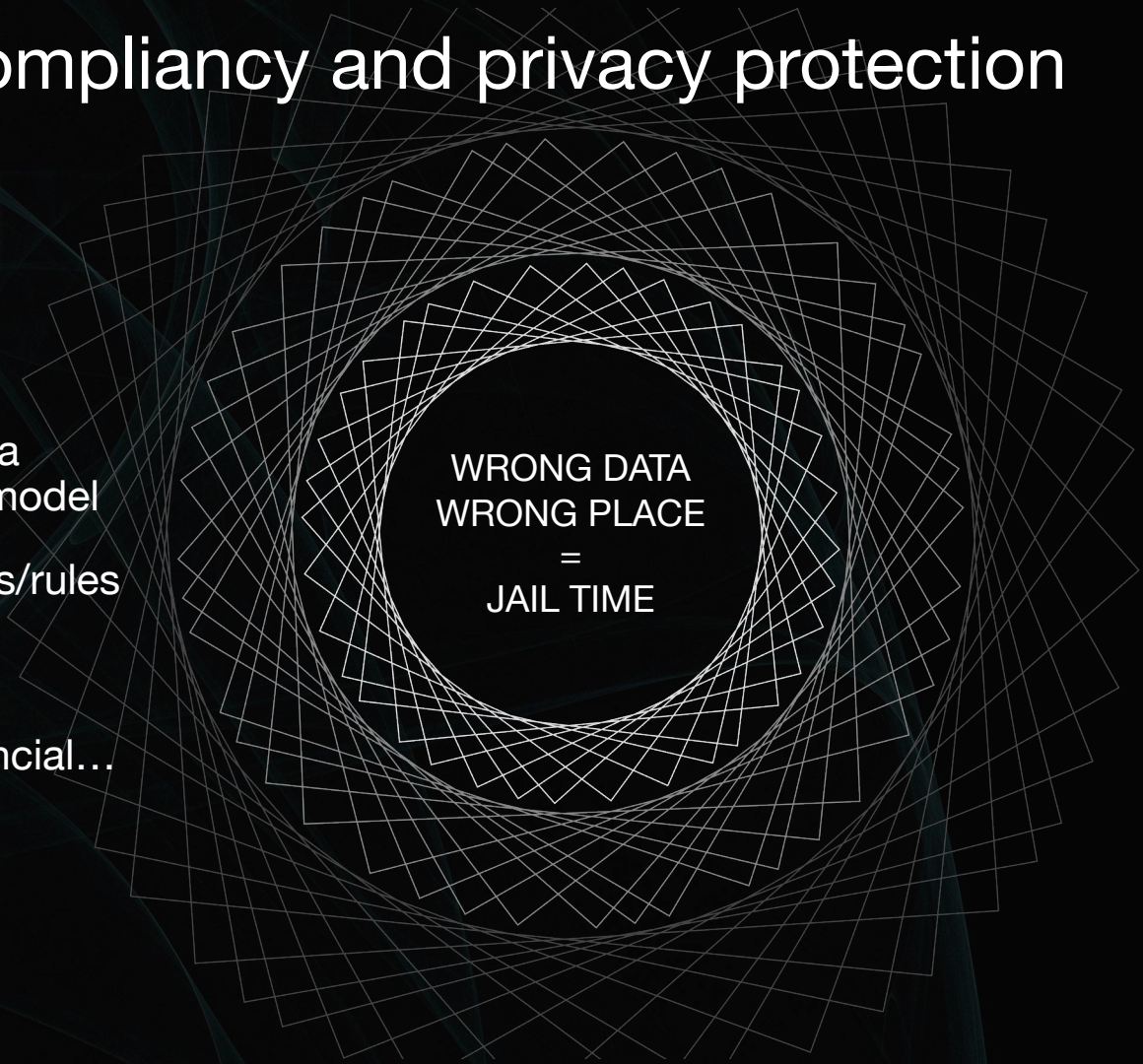


Metadata example

```
Paris_example — -bash — 80×10
grahams-imac:Paris_example gb_on_imac$
grahams-imac:Paris_example gb_on_imac$
grahams-imac:Paris_example gb_on_imac$ pwd
/Users/gb_on_imac/Dropbox/AUSA/content/RMACC/Paris_example
grahams-imac:Paris_example gb_on_imac$ ls -i
13171260943 eiffel-tower.HEIC
grahams-imac:Paris_example gb_on_imac$ ls -l
total 2624
-rw-r--r--@ 1 gb_on_imac  staff  1340531 May  7 10:30 eiffel-tower.HEIC
grahams-imac:Paris_example gb_on_imac$
```

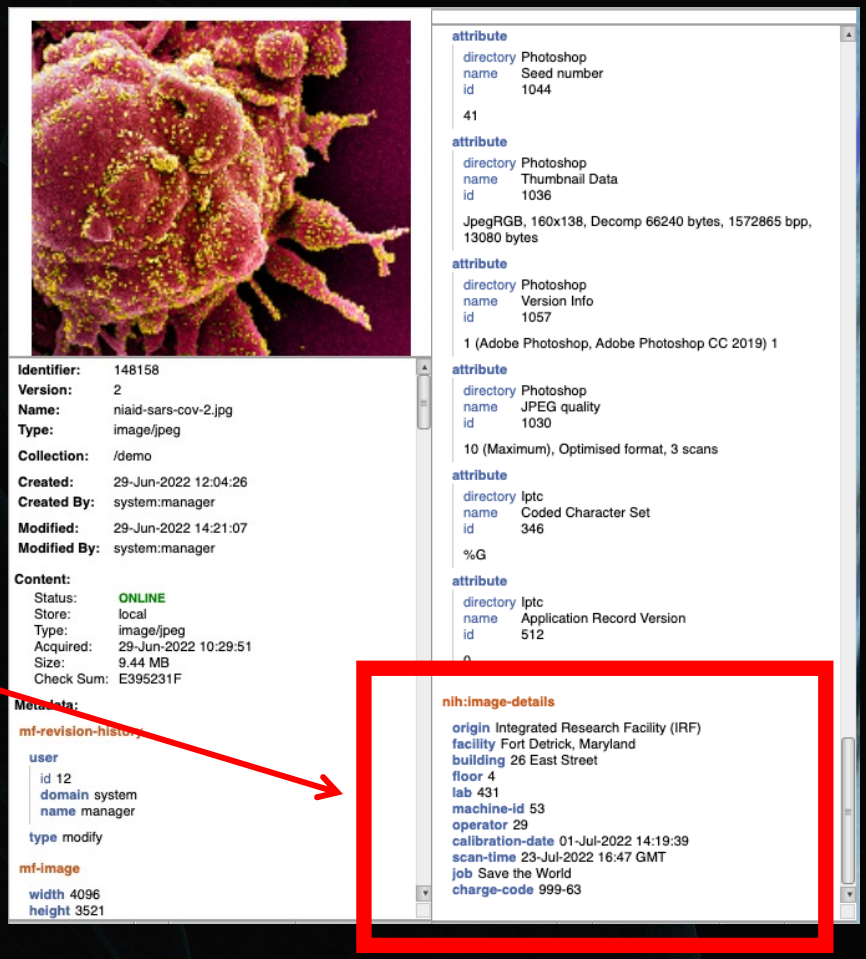
Metadata for compliancy and privacy protection

- Data Provenance
- Access Controls
- GDPR, HIPAA, CPRA
- What can you do? Build a system with Actor/Role model
- Update Metadata as laws/rules change
- Who needs this… Life Sciences, Defense, Financial… just about everyone!



Types of metadata

- "System Metadata", File name, size, create, access, modify time, ownership permissions, etc.
  - In Unix/linux world this information comes from inodes and is often used by Backup or HSM software.
- Embedded File Metadata
  - Typically parsed out via MIME type.
- User Defined Metadata
  - This enables data life cycle management, notes, accounting.
- Privacy information goes here and the fields can evolve
  - Information that influences Actor/Role access models can also go here



Data is everywhere

Our technology is at the core of some of the most demanding big data environments.

LIFE SCIENCES

RESEARCH

CLINICAL

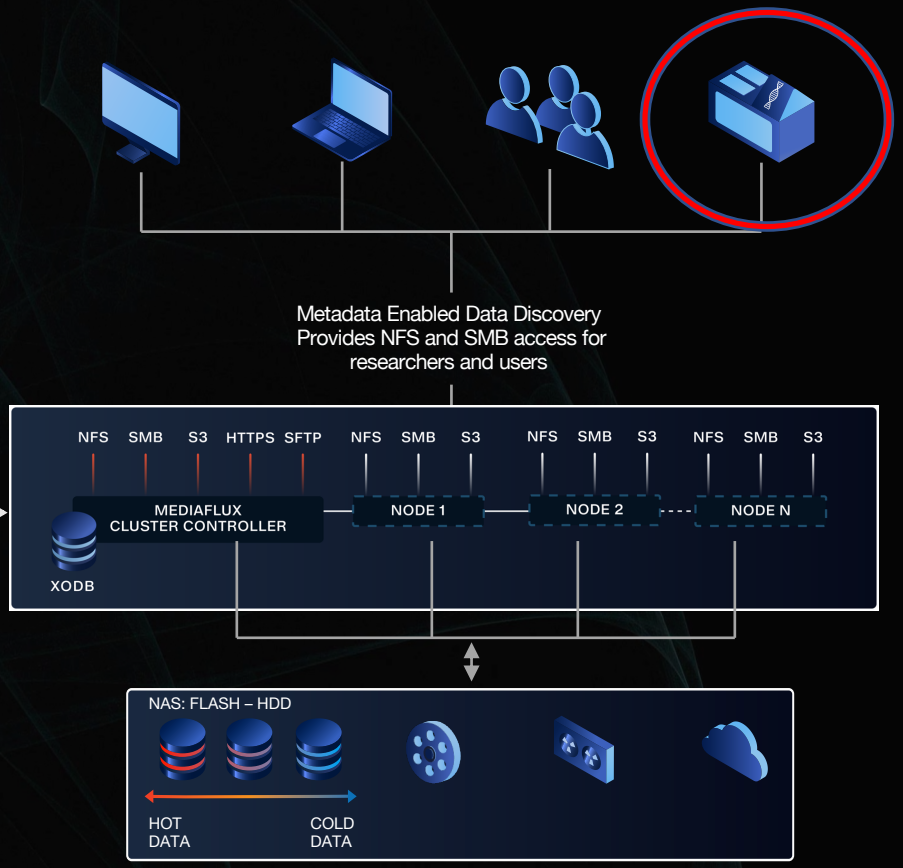
GEOSPATIAL

MEDIA & ENTERTAINMENT

DEFENSE & INTELLIGENCE

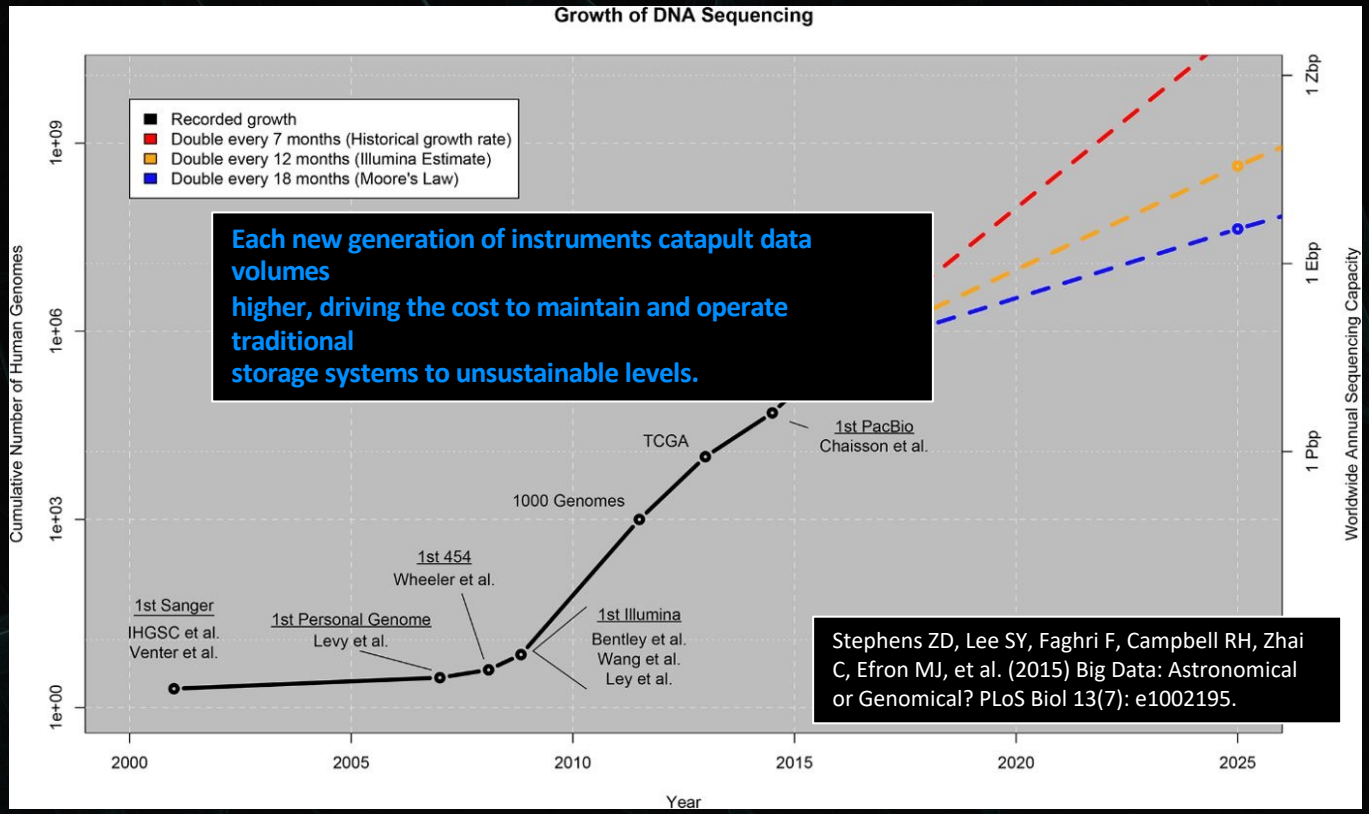
Nobody wants less resolution!

HPC CLUSTER AND
“SCRATCH” STORAGE



VIRTUALIZED STORAGE –
SINGLE GLOBAL NAMESPACE

A Known Problem: Exp. Data Growth in Genomics



Standards and Working Groups

“FAIR” Principles (metadata is key aspect)

From: go-fair.org

- Findable: “The first step in (re)using data is to find them. Metadata and data should be easy to find for both humans and computers. Machine-readable metadata are essential for automatic discovery of datasets and services, so this is an essential component of the FAIRification process.”
- Accesible: “Once the user finds the required data, she/he needs to know how can they be accessed, possibly including authentication and authorisation.”
- Interoperable: “The data usually need to be integrated with other data. In addition, the data need to interoperate with applications or workflows for analysis, storage, and processing.”
- Reusable: “The ultimate goal of FAIR is to optimise the reuse of data. To achieve this, metadata and data should be well-described so that they can be replicated and/or combined in different settings.”

The perfect metadata now might not be the perfect metadata in 2 years, 10 years or 20 years.

More information

- Data Management: tools, whitepapers, products, etc. www.arcitecta.com
- From 2022 Gartner report ”Strategic Roadmap for Storage” key take aways.. Not so much about the storage but more about “Data Management” with suggested action plans like: *Appoint “data champions” to develop policy-aware data classification and retention strategies for business-critical applications and compliance and privacy issues.*
- June 2022 white paper “Scientific Data” by Devan Ray Donaldson who is PhD Professor at Luddy School of Informatics, Computing, and Engineering at Indiana University https://www.nature.com/articles/s41597-022-01428-w
- Citable ID’s etc. https://libguides.library.nd.edu/research-data-services/make-citable
- Data Management plans, etc.
- https://researchdatamanagement.harvard.edu/data-management-plans
- https://researchdata.princeton.edu/research-lifecycle-guide/data-management-plans
- https://dataservices.research.unimelb.edu.au/services/40/