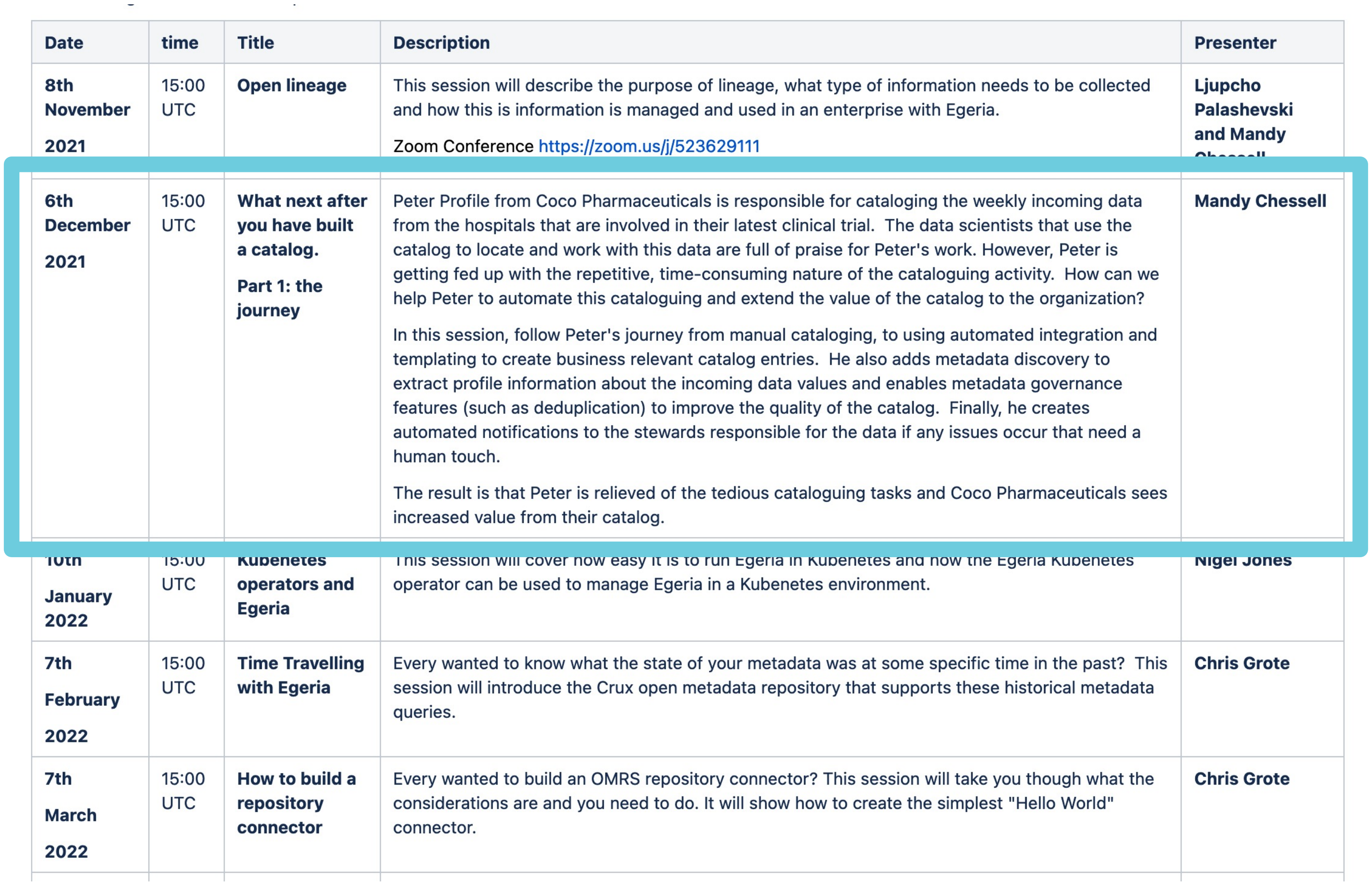
LF AI & DATA

The EGERIA Metadata Show

Egeria Webinar

WHAT NEXT AFTER YOU HAVE BUILT YOUR CATALOG?

Mandy Chessell CBE FREng
Egeria Open Source Project Lead

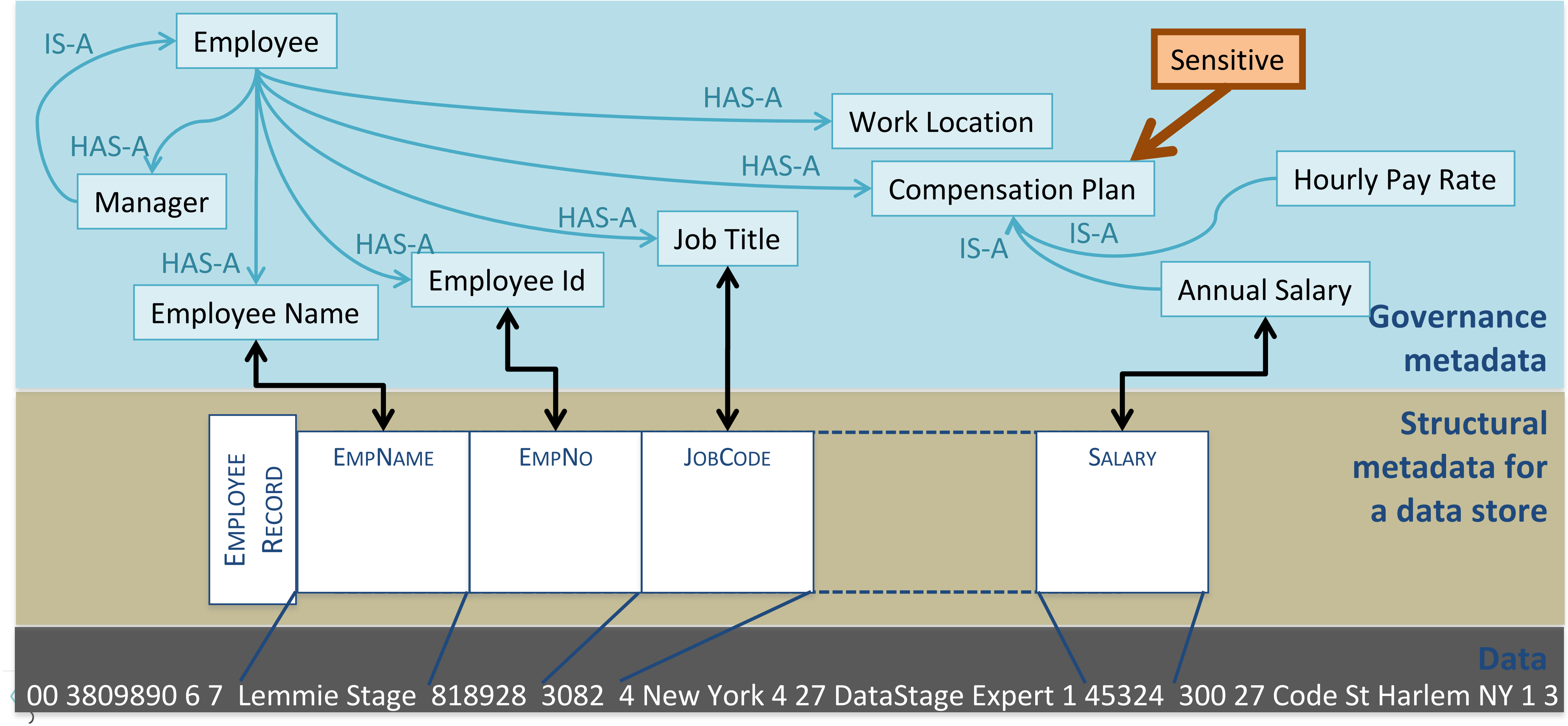


Date	time	Title	Description	Presenter
8th November 2021	15:00 UTC	Open lineage	This session will describe the purpose of lineage, what type of information needs to be collected and how this is information is managed and used in an enterprise with Egeria. Zoom Conference https://zoom.us/j/523629111	Ljupcho Palashevski and Mandy Chessell
6th December 2021	15:00 UTC	What next after you have built a catalog. Part 1: the journey	Peter Profile from Coco Pharmaceuticals is responsible for cataloging the weekly incoming data from the hospitals that are involved in their latest clinical trial. The data scientists that use the catalog to locate and work with this data are full of praise for Peter's work. However, Peter is getting fed up with the repetitive, time-consuming nature of the cataloguing activity. How can we help Peter to automate this cataloguing and extend the value of the catalog to the organization? In this session, follow Peter's journey from manual cataloging, to using automated integration and templating to create business relevant catalog entries. He also adds metadata discovery to extract profile information about the incoming data values and enables metadata governance features (such as deduplication) to improve the quality of the catalog. Finally, he creates automated notifications to the stewards responsible for the data if any issues occur that need a human touch. The result is that Peter is relieved of the tedious cataloguing tasks and Coco Pharmaceuticals sees increased value from their catalog.	Mandy Chessell
10th January 2022	15:00 UTC	Kubenetes operators and Egeria	This session will cover how easy it is to run Egeria in Kubenetes and how the Egeria Kubenetes operator can be used to manage Egeria in a Kubenetes environment.	Nigel Jones
7th February 2022	15:00 UTC	Time Travelling with Egeria	Every wanted to know what the state of your metadata was at some specific time in the past? This session will introduce the Crux open metadata repository that supports these historical metadata queries.	Chris Grote
7th March 2022	15:00 UTC	How to build a repository connector	Every wanted to build an OMRS repository connector? This session will take you though what the considerations are and you need to do. It will show how to create the simplest "Hello World" connector.	Chris Grote

What is Egeria?

- A set of open type definitions providing a common language for metadata
  - Guiding organizations in its use
  - Providing a basis for metadata exchange
- Basic metadata management (metadata repositories, metadata onboarding and management, user interface for search and display, automated metadata discovery and stewardship, …)
- Metadata integration with heterogeneous tools
  - Event-based, batch or scheduled
  - Developer toolkit and connector catalog
- Open metadata archives and content packs
- Metadata governance (metadata deduplication, validation, enrichment, …)

Different types of metadata



Coco Pharmaceuticals persona

Jules Keeper, CDO

Tessa Tube,
Chief Researcher

Faith Broker
Chief Privacy Officer

Nancy Noah
Cloud Specialist

Erin Overview,
Information Architect

Bob Nitter,
Integration Developer

Callie Quartile,
Data Scientist

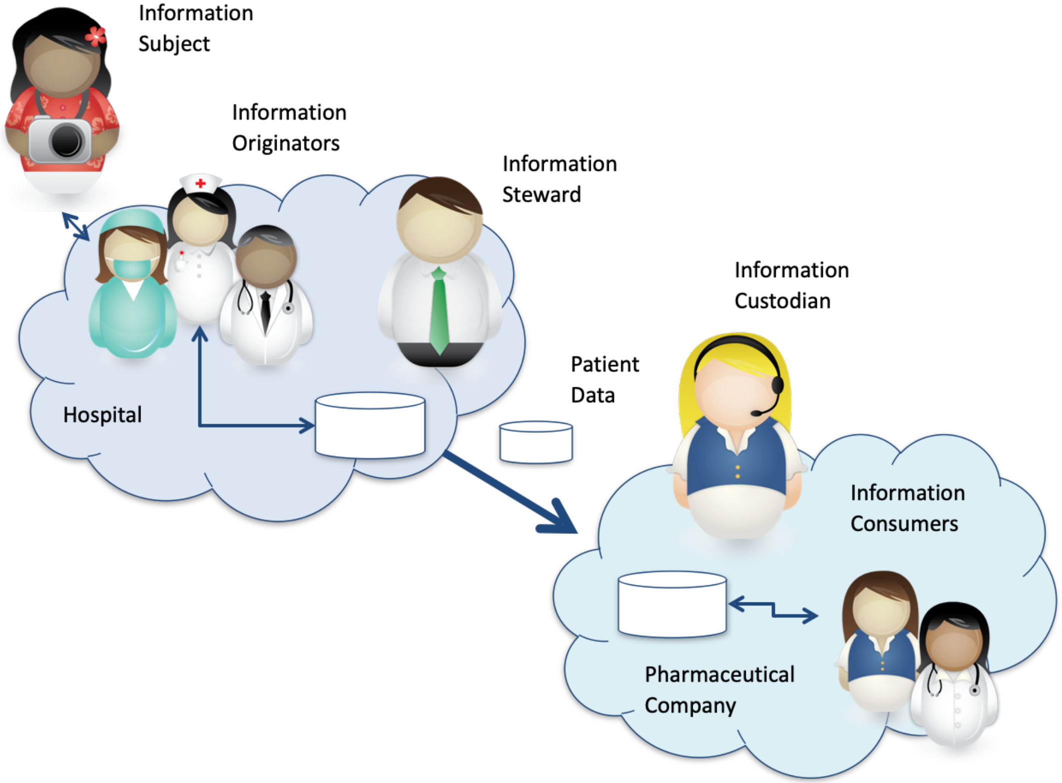
Gary Geeke
IT Infrastructure

EGERIA

https://odpi.github.io/data-governance/coco-pharmaceuticals/personas/

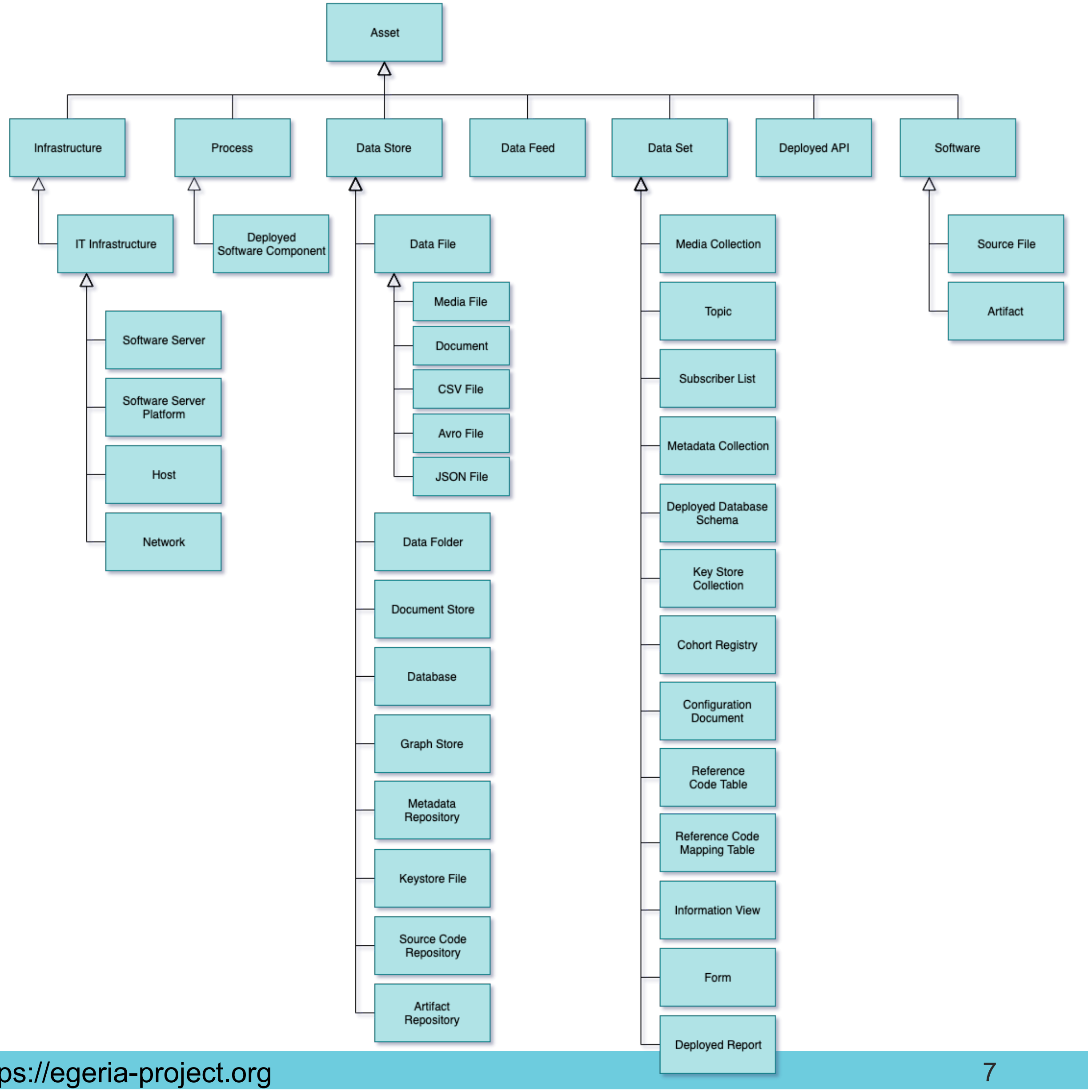
Data Transfer between Organizations

- Highly regulated process
- Trust is built from demonstrated governance on both sides.



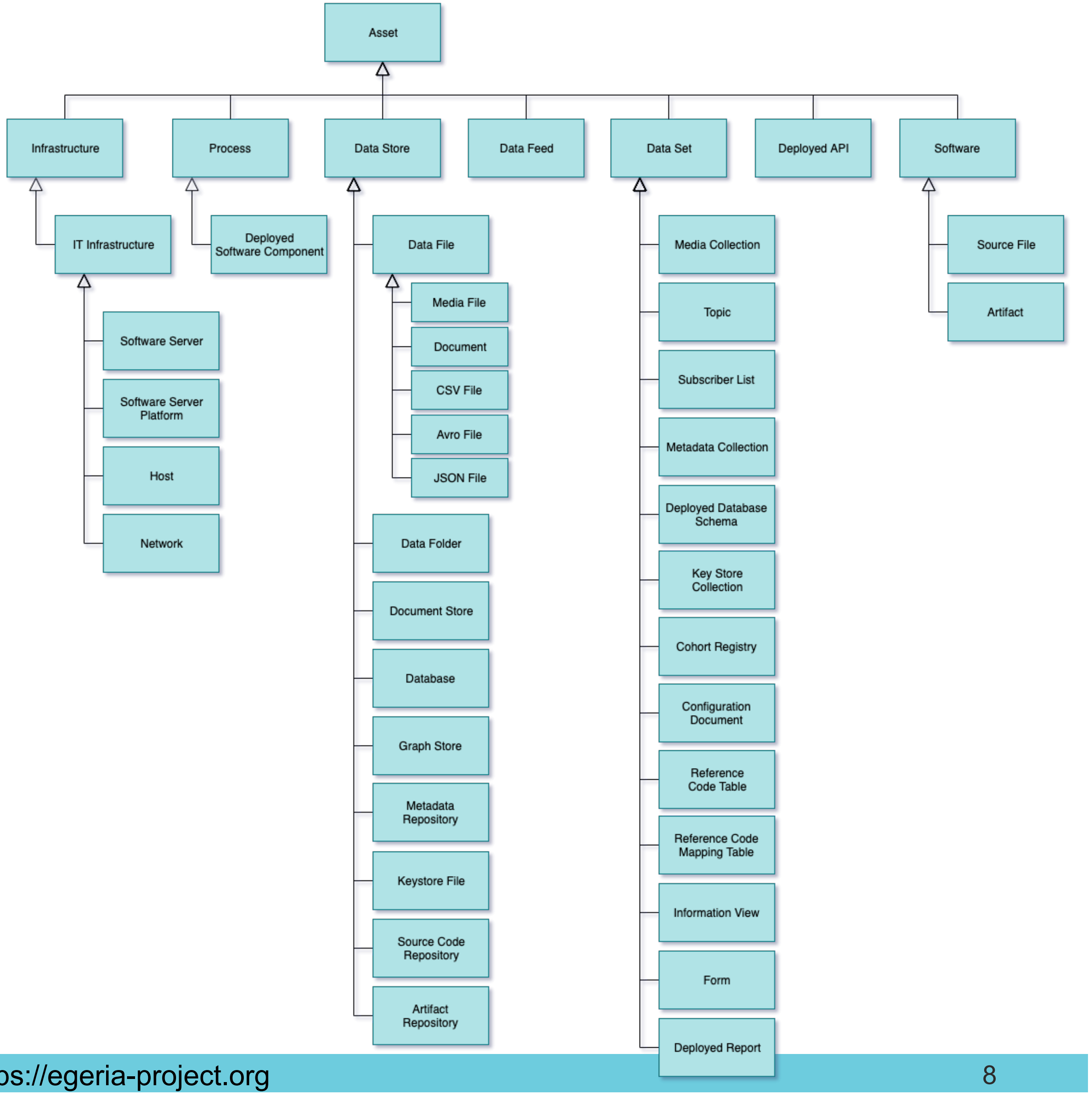
Types of Asset

- The open metadata asset type hierarchy enables metadata from multiple governance domains to be linked together in a single model.
    - IT Infrastructure
    - Data
    - Applications and Events
    - Software development and DevOps
    - Security



Types of Asset

- Infrastructure – hardware and software generic “platforms”.
- Process – well defined sequence of actions.
- DataStore – data at rest
- DataFeed - data in motion
- DataSet – collection of data
- DeployedAPI – interface
- Software – parts for IT system



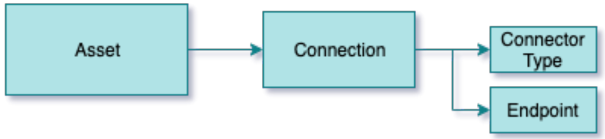
EGERIA

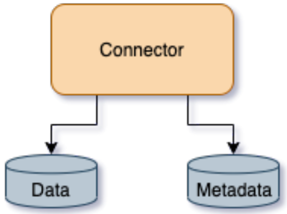
Asset Properties

Asset

- **Properties**:
  - **Open Metadata Unique identifier (GUID)** - this a globally unique id across all metadata instances. It is a string of letters and numbers and typically looks something like this 40d9520b-dbc0-4cc4-9bad-03ab72d027f3 and is assigned by Egeria.
  - **Qualified Name** - this is a globally unique name of the asset - it is unique across all assets It is assigned by the creator of the asset.
  - **Display Name** - typically qualified names are long in order to make them unique. The display name is a short name used in reports and other displays of asset information.
  - **Description** - description of the asset.
  - **Additional Properties** - names and values of additional properties that the organization wants to record about the asset.
- **Usage**
  - With the basic asset properties defined, the asset catalog provides a searchable list of the assets of the organization. The content provides in the names, descriptions and additional properties will determine how easy it is to retrieve specific assets.

Connections

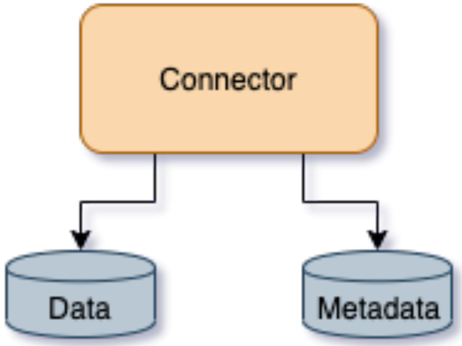


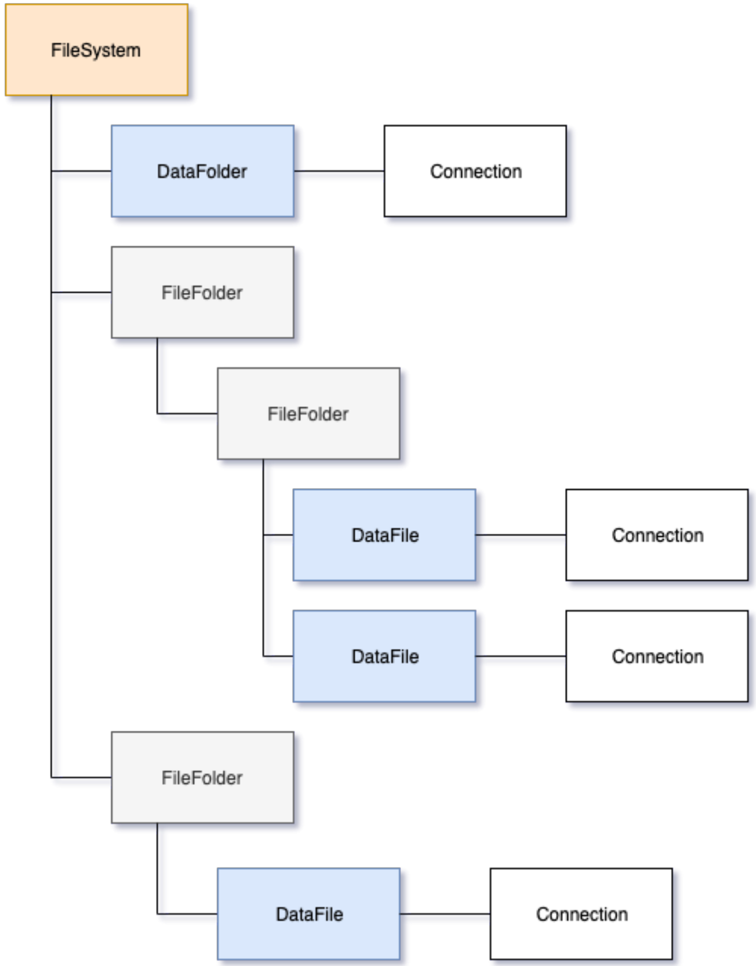


- A connection the information necessary to create a connector to the asset.
- The connector is a client to both the data and the information about the asset stored in open metadata. Some connectors use the metadata about the asset to control what data can be retrieved from the asset depending on the caller.
- **Usage**
  - The connector that is generated from the connection object enables both tools and applications to use the asset through a governed interface that provides metadata, data and, in some cases, metadata-driven access control.

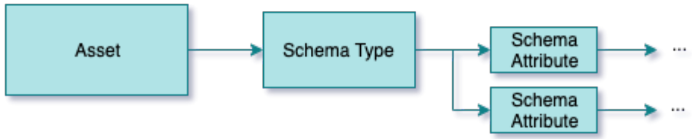
Not all assets need a connection

- A connection is only needed if applications/tools needs to access its contents.



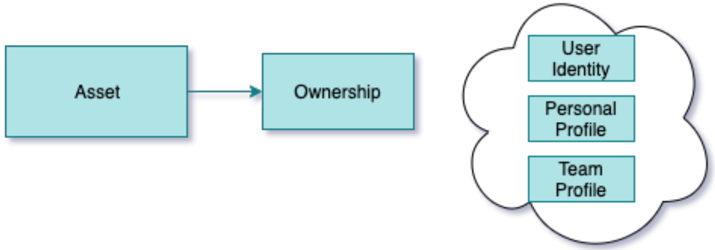


Schemas



- A schema describes the individual data fields and operations of the asset. It is organized to reflect the internal organization of the asset and so acts as a guide to the types of content in the asset and how to navigate around it.
- **Usage**
    - With the schema in place, it is possible to search for assets based on the type of data, or type of operations that the asset supports.

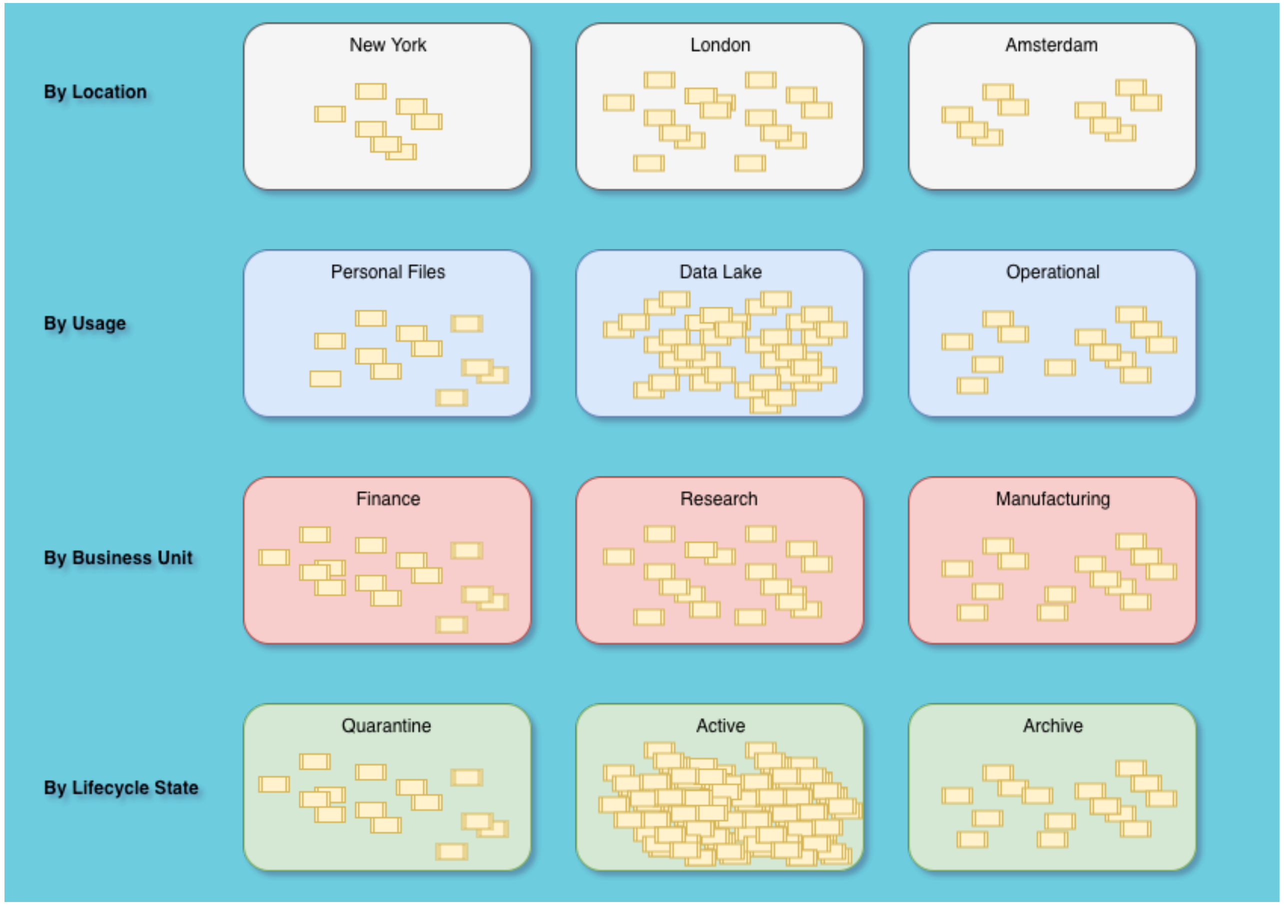
Asset Owners



- Asset ownership defines who is responsible for the asset. This covers ensuring the catalogue entry is correct, the contents of the asset are complete and correct and controlling access to the asset.
- The owner can be defined as a user identity, a personal profile or a team profile. These definitions are managed by the Community Profile OMAS.
- **Usage**
  - With an owner established, it records who is responsible for the protection and quality of the asset. It is possible to route requests from the consumers of the asset to the owner. An example of this is in managing queries about the content of the asset and requests for access to its contents.

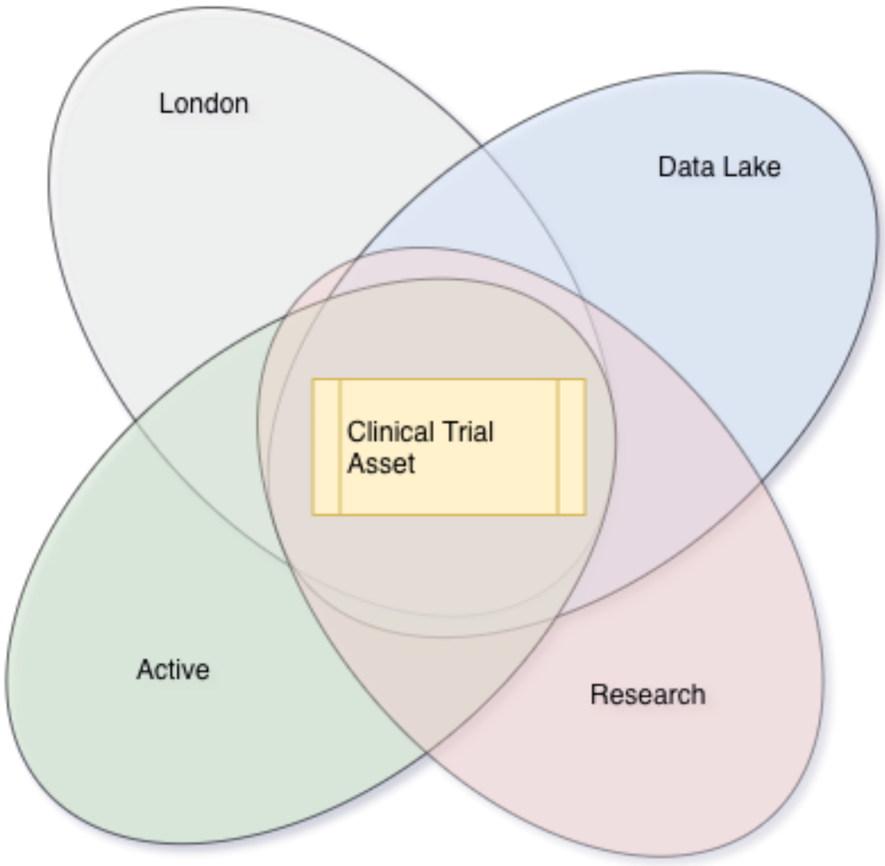
Governance Zones

- Metadata can become more valuable than data when it pulls together a complete view of the organization's assets and activities.
- A governance zone defines a set of assets that are related in some way



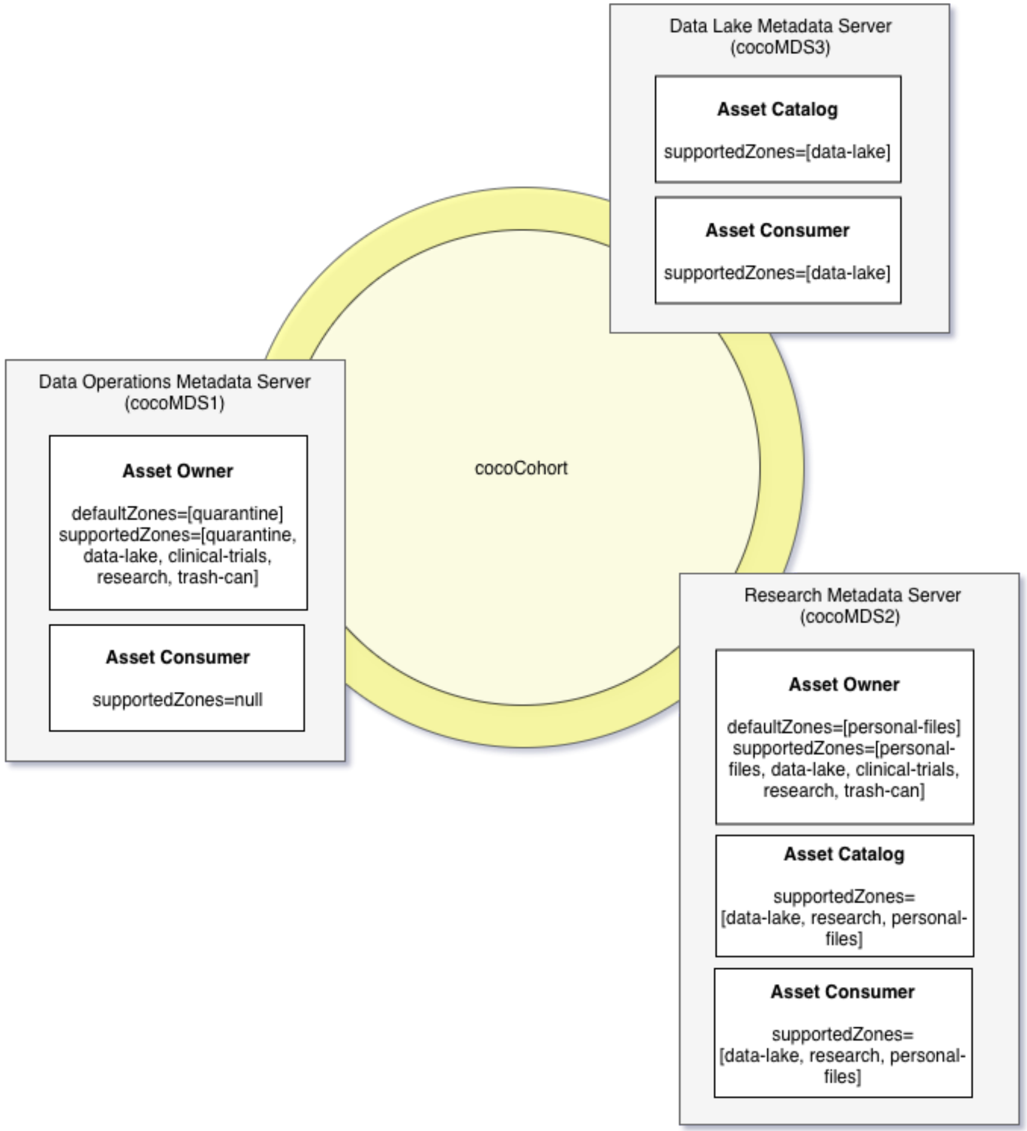
Zone Membership

- The zones that an asset belongs to determines its visibility to specific governance processes and API entry points.



Access services’ zone settings

- Default zones
  - Zone membership for new assets (unless explicitly set)
  - Null means not a member of any zone
- Supported zones
  - Scope of assets that are visible through this OMAS
  - Null means all assets
- Publish zones
  - Zones to set in an asset when ready for broader consumption



Data Lake Metadata Server (cocoMDS3)

Asset Catalog

supportedZones=[data-lake]

Asset Consumer

supportedZones=[data-lake]

cocoCohort

Data Operations Metadata Server (cocoMDS1)

Asset Owner

defaultZones=[quarantine]
supportedZones=[quarantine, data-lake, clinical-trials, research, trash-can]

Asset Consumer

supportedZones=null

Research Metadata Server (cocoMDS2)

Asset Owner

defaultZones=[personal-files]
supportedZones=[personal-files, data-lake, clinical-trials, research, trash-can]

Asset Catalog

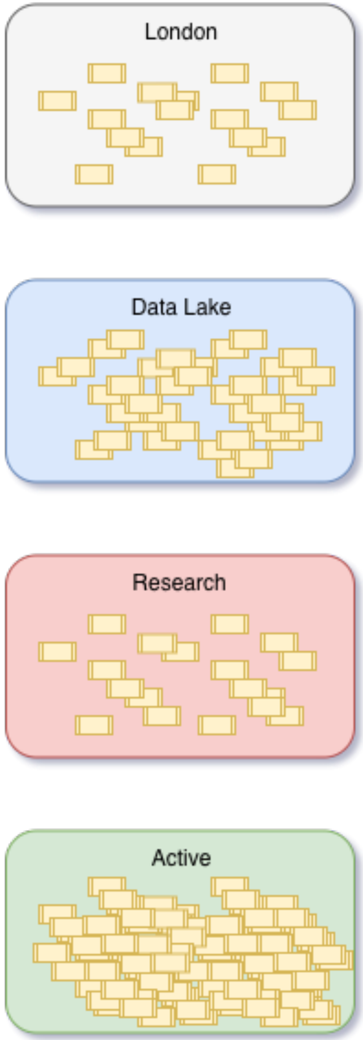
supportedZones=[data-lake, research, personal-files]

Asset Consumer

supportedZones=[data-lake, research, personal-files]

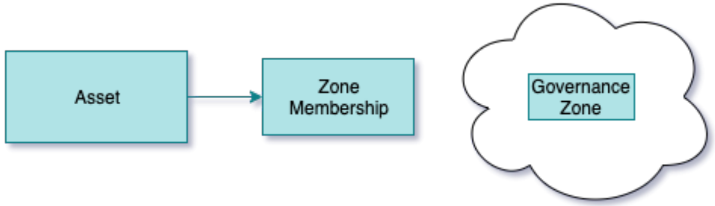
Use of governance zones

- Data sovereignty
- Adjustments for legal jurisdiction
- Asset visibility and access control
- Data access control
- Maintenance and backup processing
- Metering and billing
- Understanding dependencies



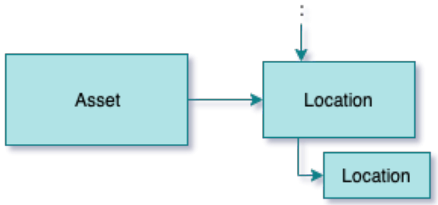
EGERIA

Governance Zones



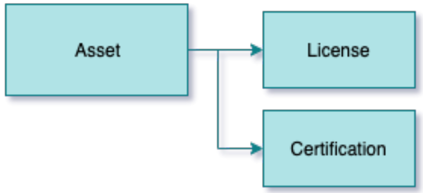
- Governance Zones allow assets to be grouped according to their usage. It is possible to assign supported zones to Egeria's Open Metadata Access Services (OMASs) to limit the scope of assets that are returned from searches.
- **Usage**
  - Using governance zones allows the organization to scope the assets that are returned to a community of users who are using the asset catalogue.
  - The governance zones can also be used to define the group of assets that an automated process should process.

Asset Location



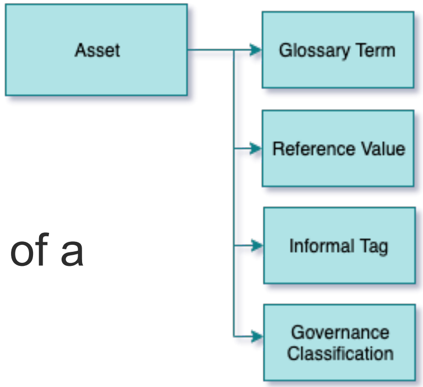
- Egeria supports the definition of a location model that divides both physical and digital space into hierarchies with cross links between the hierarchies. This means it is possible to link the assets to their location(s)
- **Usage**
  - Attaching assets to location definitions means it is possible to use details of the location as part of the search for assets.
  - Knowing the asset location, whether it is a physical or digital location can also help with demonstrating that data sovereignty is being respected and the level of risk that is allocated in a particular location.

License and Certifications



- An asset can have its license and/or certifications attached to it.
  - The license determines the terms and conditions of use for the asset. This becomes important particularly when assets come from an external organization.
  - Certifications typically relate to a regulation or standard or DevOps quality gate. When the certification is tied to the asset it means that the asset has passed the requirements.
- Usage
  - Attaching licences and certifications to assets raises awareness of the any restrictions on the use of the assets and to what standards they are managed to. If the licenses and certifications are machine readable, automated processes can used them to control the way that they manage the assets.

Classifiers



- Classifiers add labels and properties to the asset that identifies them as part of a specific group, or having particular characteristics.
- The types of classifiers are:
  - Glossary terms define the meaning of concepts and activities. When a glossary term is attached to a data field in the assets schema, it signifies that the data stored in that field has the meaning described in the glossary term.
  - Reference values identify sets of valid values of particular characteristics of the assets. For example, attaching a reference code for “personal data” to an asset indicates that it contains personal data.
  - Informal tags are labels that asset consumers create and attach to the asset and its data fields/operations. This is effectively a way of crowd sourcing knowledge about the asset.
  - Governance classifications provide formal classifiers for confidentiality, retention, confidence and criticality for the asset.
- **Usage**
  - Classifiers help to make assets more findable. They also identify which assets should be treated to certain types of processing. For example, data fields marked as sensitive could be masked when added to a sandbox.

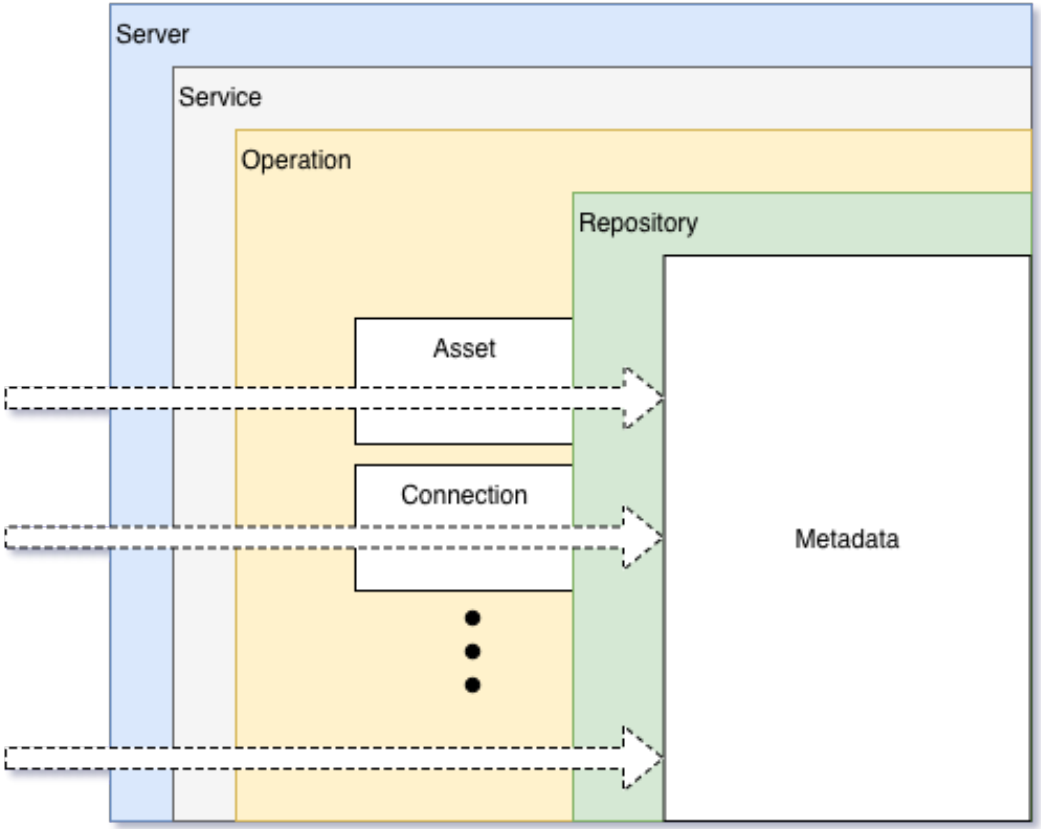
Using classifiers

- The classifiers can be added to any **Referenceable**.
  - A referenceable is a metadata element with a unique qualified name.
  - Assets are referenceables, so are elements of a schema, connections and many of the types of elements that can be attached to an asset.
- This means, for example, classifiers the whole asset or a specific field or operation in the schema.

Metadata Security

- Metadata ranges from public information to highly sensitive information.
- The consequence of integrating metadata is that metadata security needs to be more granular – instance based.
- Picture is an example of the different levels of security challenge in an Egeria server.
- Security is controlled through a connector
- Asset security can change asset settings (eg zone membership) based on calling user.

Levels of challenge for an inbound request



Types of information passed to the security connector

- Asset Properties
  - Identifiers
  - Type
  - Origin
  - License
  - Owner
  - Zone Membership
  - Classifications
  - Other Properties

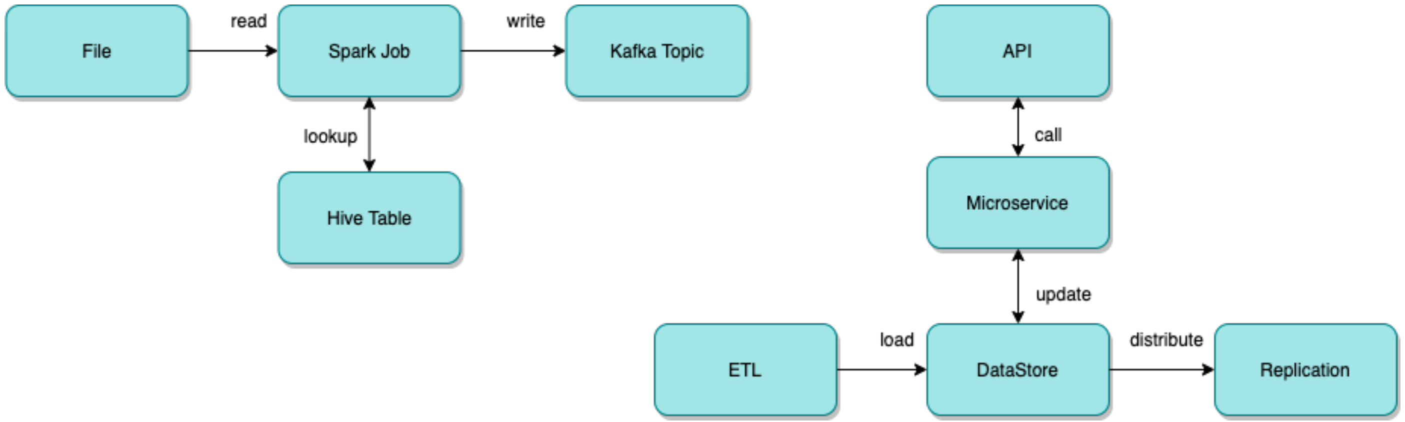
Values available through the Asset entity

- Asset Audit Header
  - Created By and Creation Time
  - Updated By and Update Time
  - Maintained By
  - Version number

Values maintained by the repository

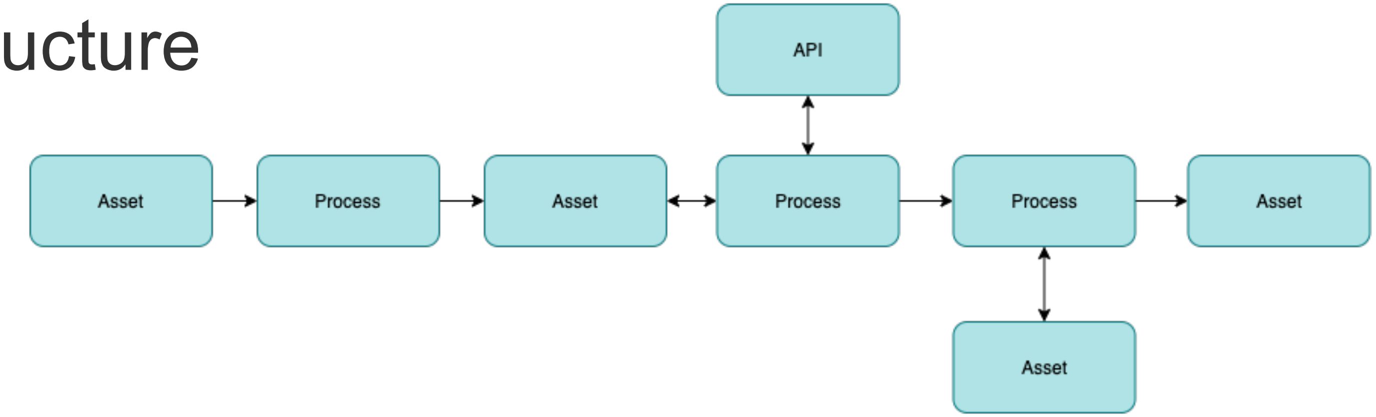
Lineage

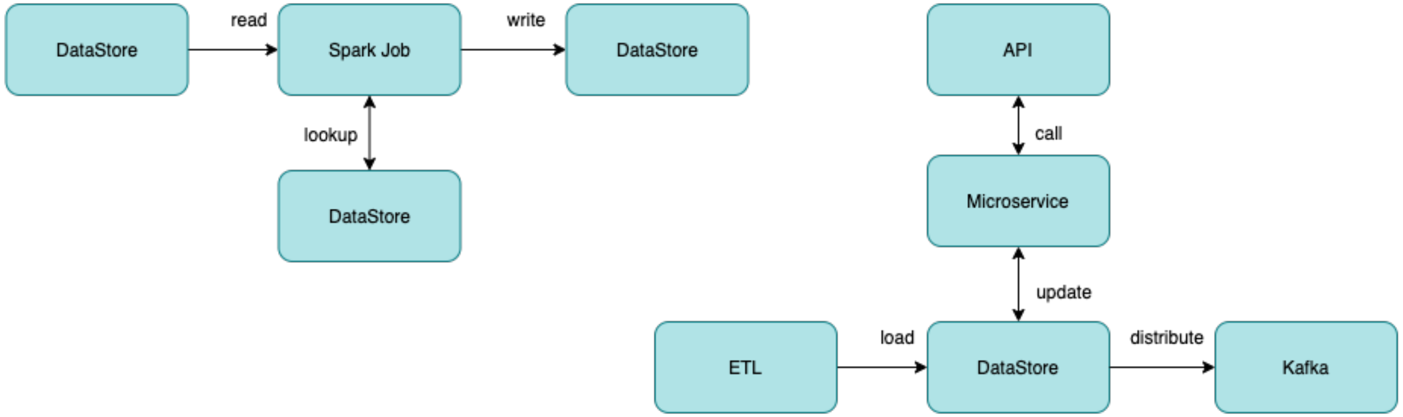
- Lineage shows how data flows from its origins to its various destinations. This includes details of the processing along the way. It is used both to understand:
  - whether the data used in reports and analytics models has come from the correct sources and has passed through the correct processing (traceability of data).
  - what would be the impact on downstream processing and consumers if something was changed (impact analysis).



https://egeria.odpi.org/open-metadata-publication/website/lineage/

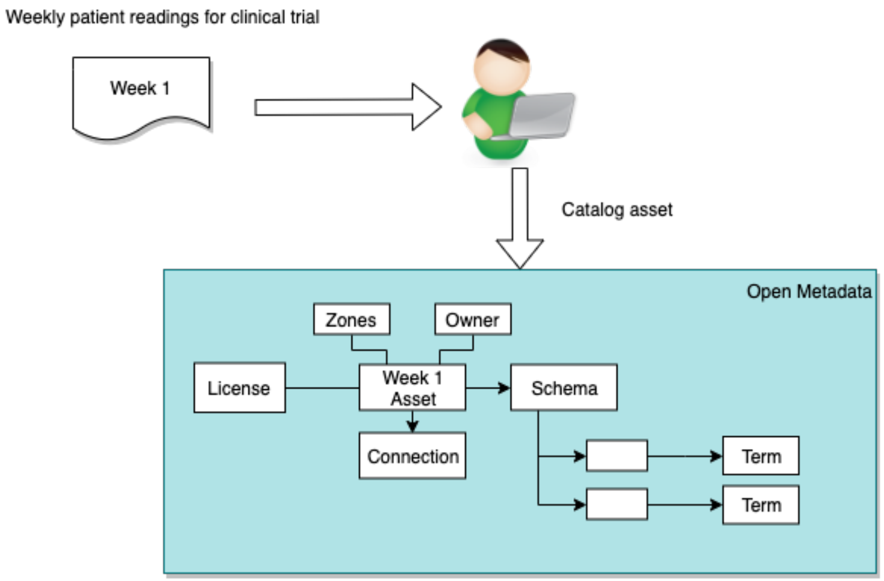
Basic lineage structure





Manual cataloging weekly patient readings

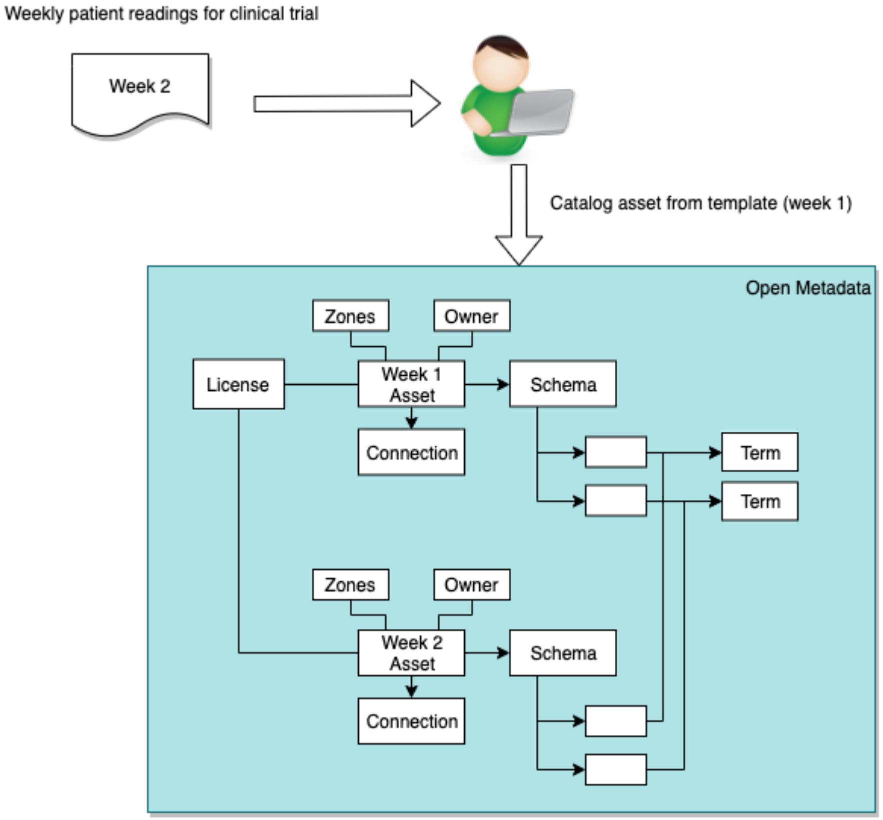
- Peter Profile is issuing calls to add the week 1 patient readings.
- Week 2 he gets to do it all over again
- And in week 3 …



Templated Cataloguing

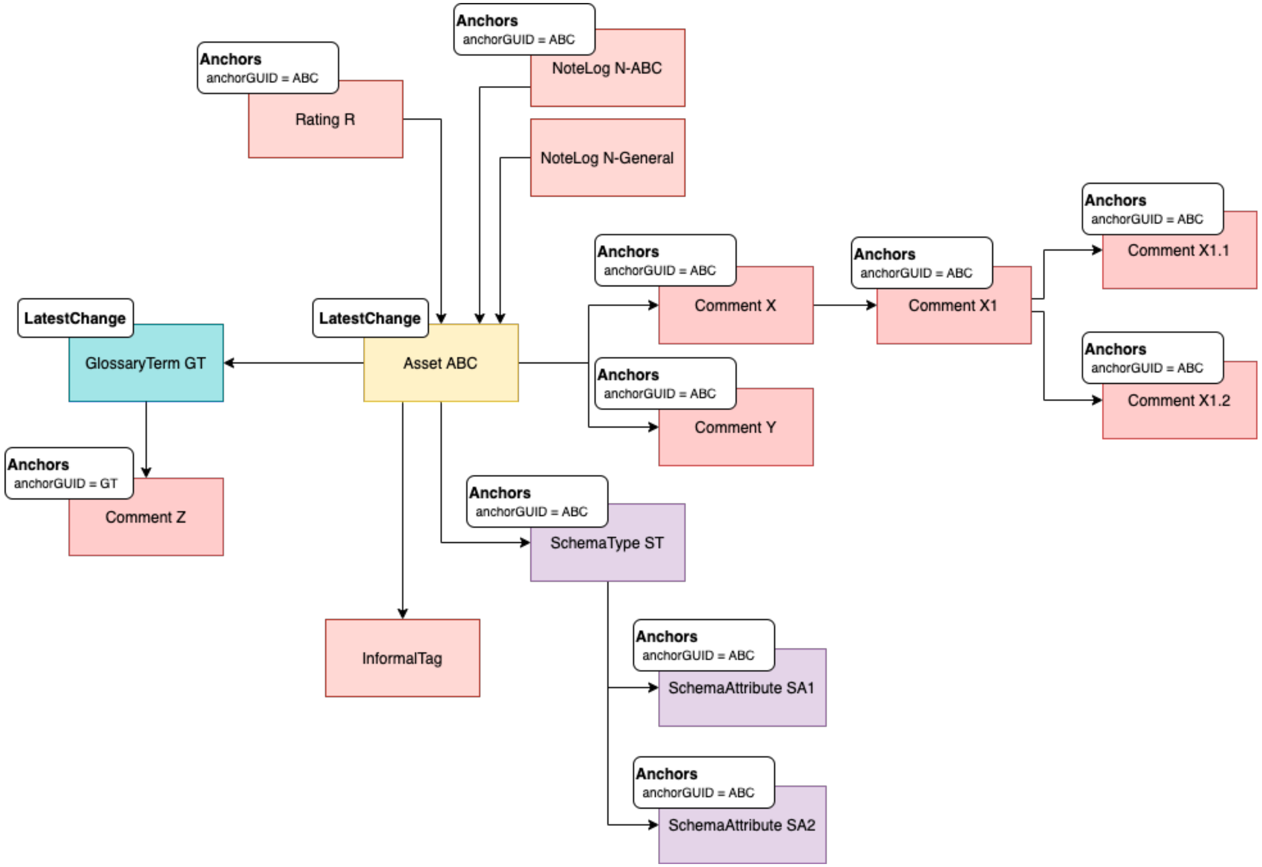
- For assets that are similar, it is possible to use a template that allows new assets to be cataloged with a single call.

- The template is set up in Egeria.



Understanding Object Boundaries

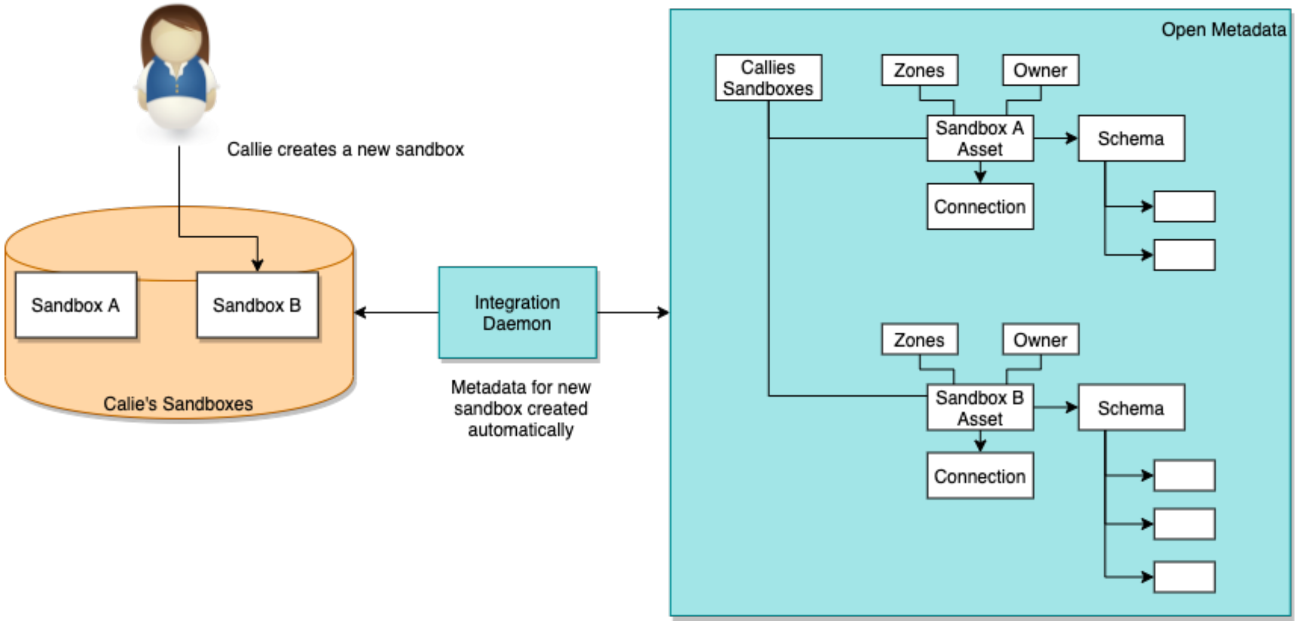
- Visibility
- Cascaded deletes
- Monitoring



EGERIA

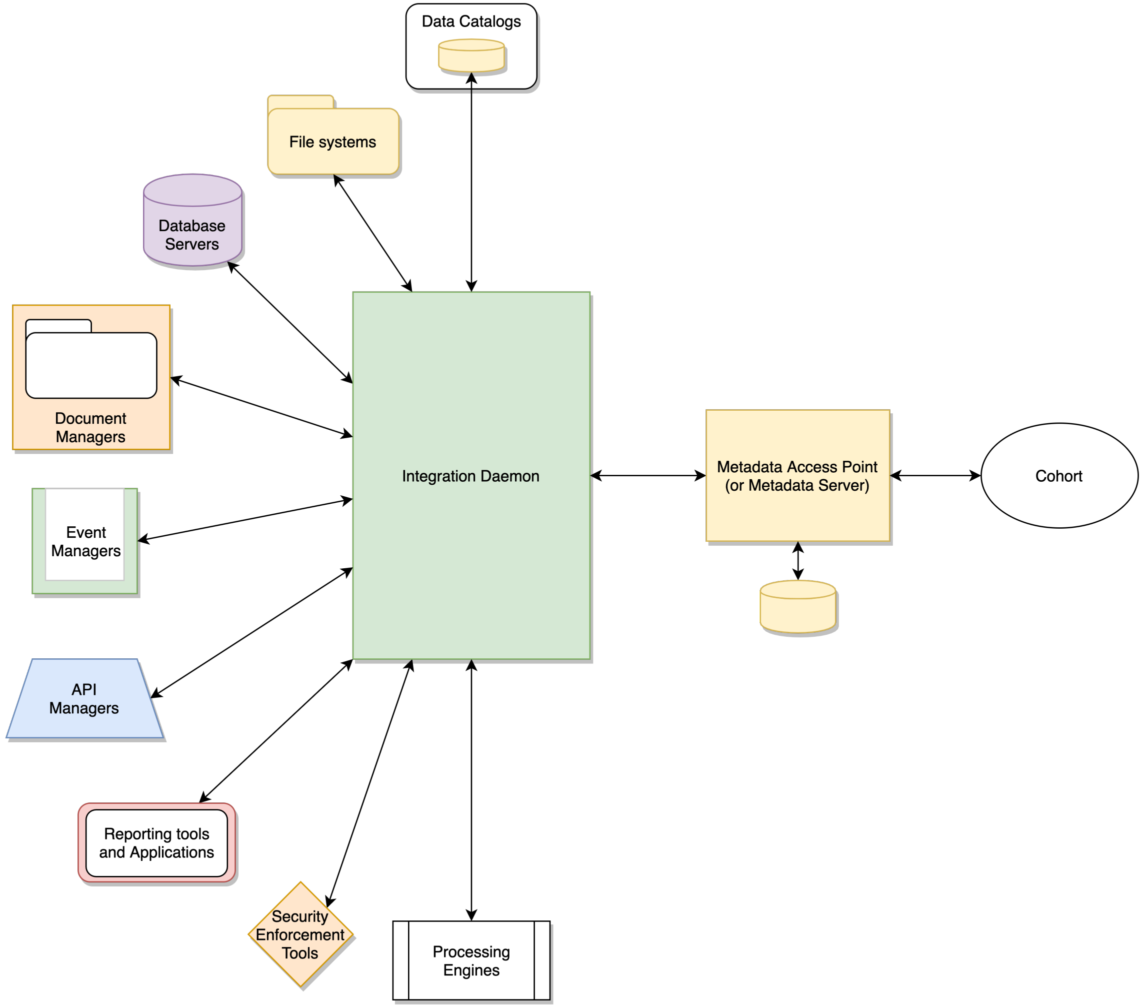
Integrated cataloguing

- Callie has a database server that she uses to analyze relational data.
- She creates a new sandbox for each type of analysis
- The data platform proxy automatically catalogs the sandbox.

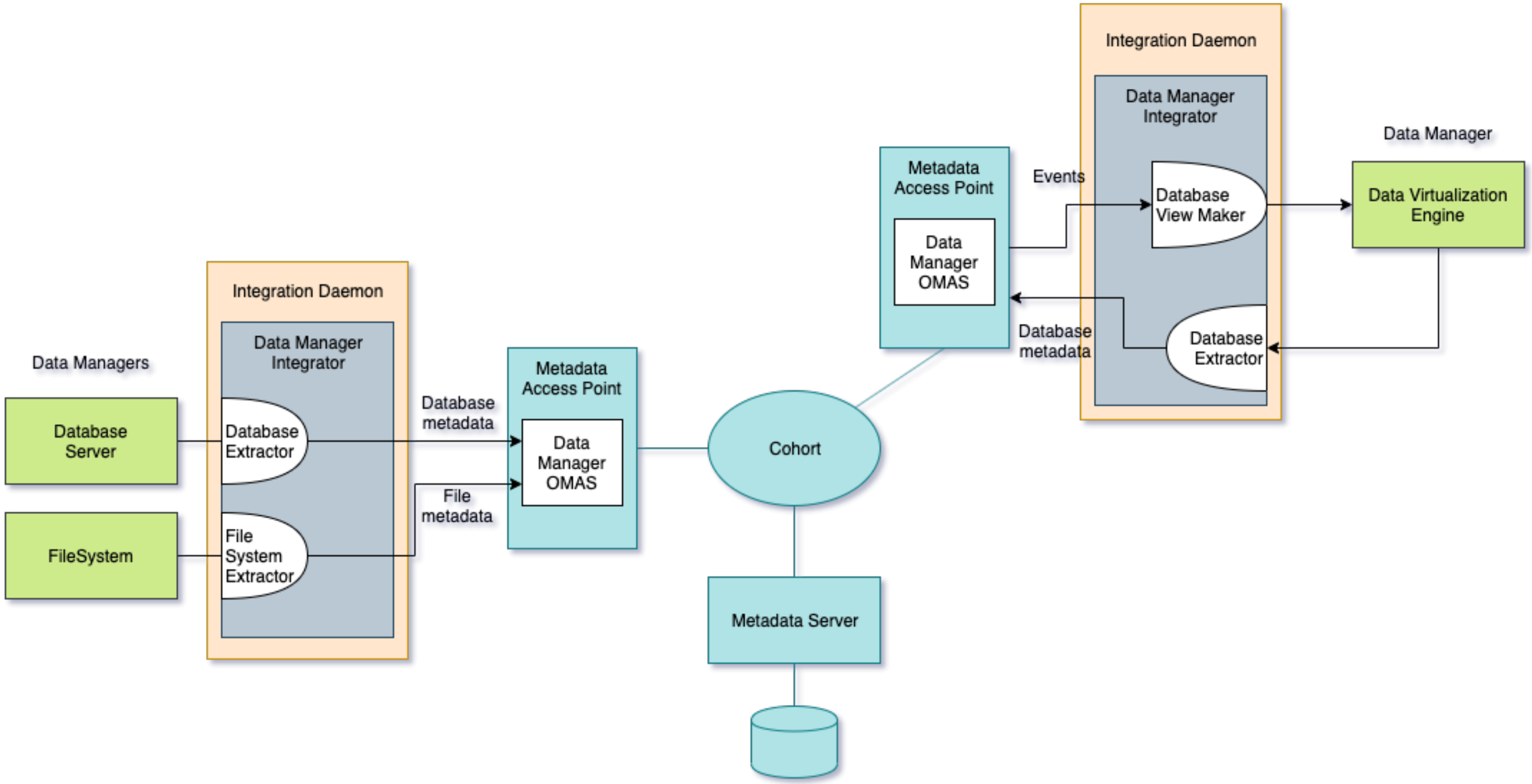


The Integration Daemon

- Metadata extraction, capture and delivery

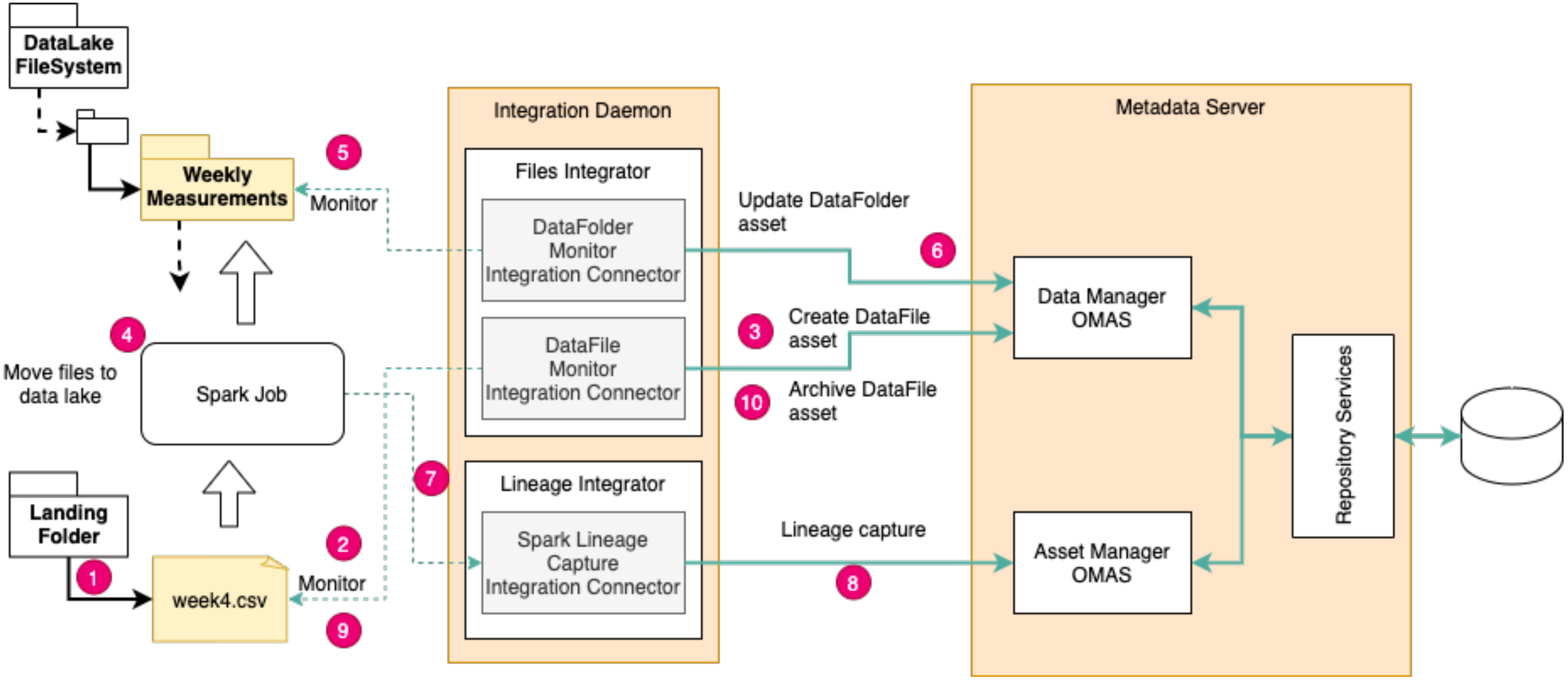


Example integrations for data managers (databases, file systems etc.)



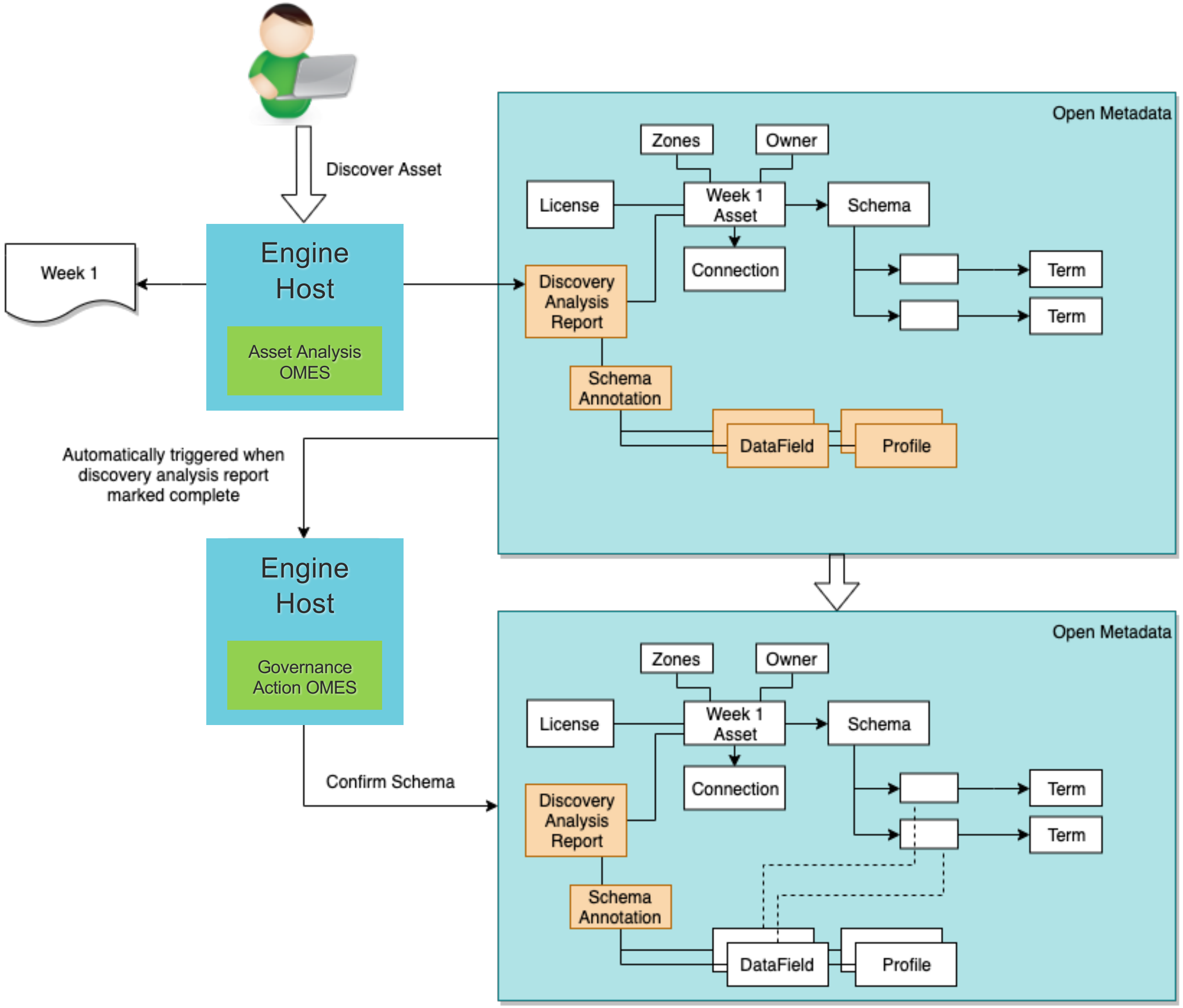
https://egeria.odpi.org/open-metadata-publication/website/solutions/data-manager-integration/

Lineage capture example with files



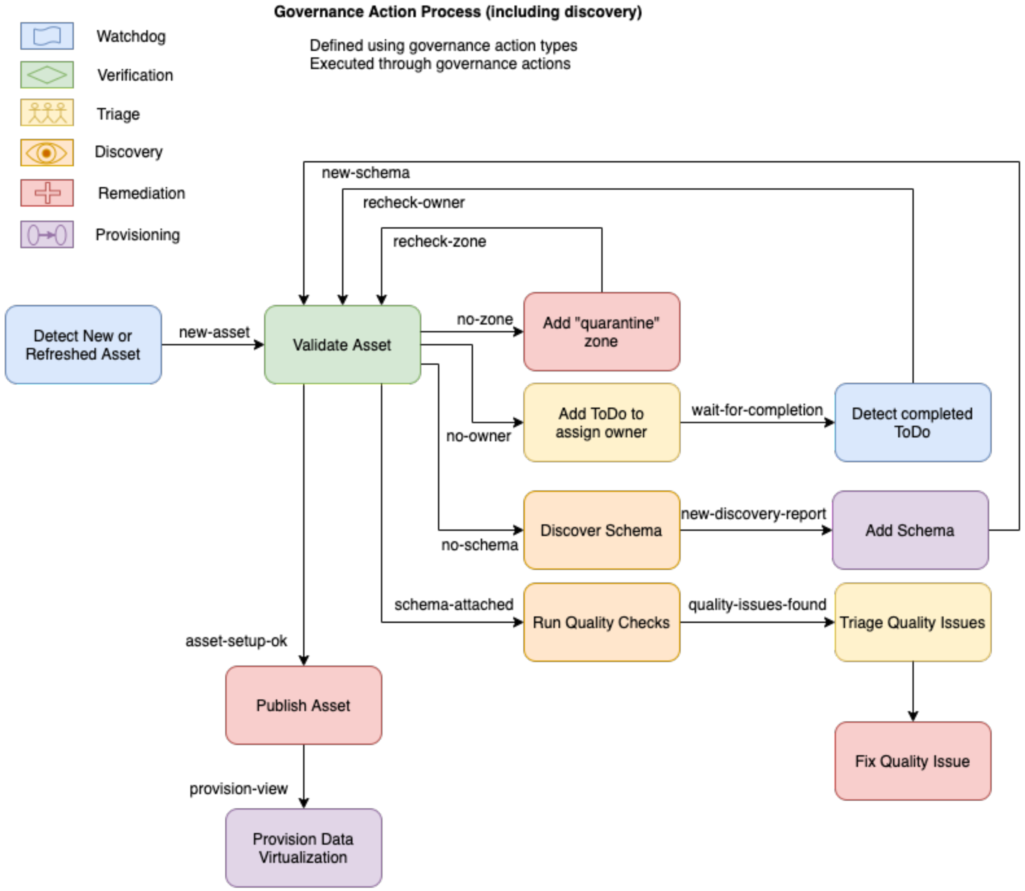
Asset Analysis and Governance Action

- Accessing resource to create new insights about an asset and storing them as annotations linked to the asset.
- Manages the transition of annotations to standard metadata elements.
  - With steward interaction
  - Completely automatically



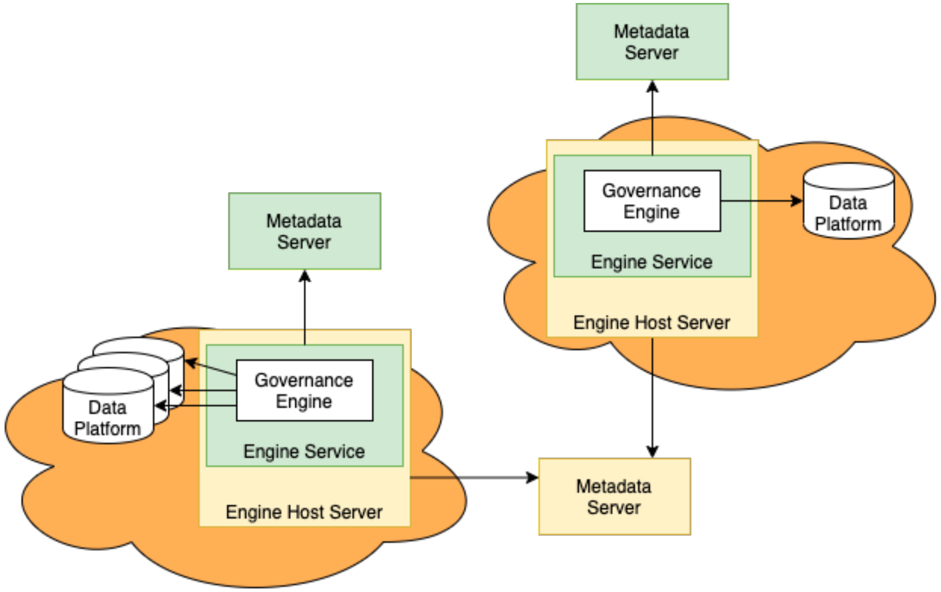
Governance Action Processes

- Pre-defined process templates used to choreograph governance actions



Maintaining coherence

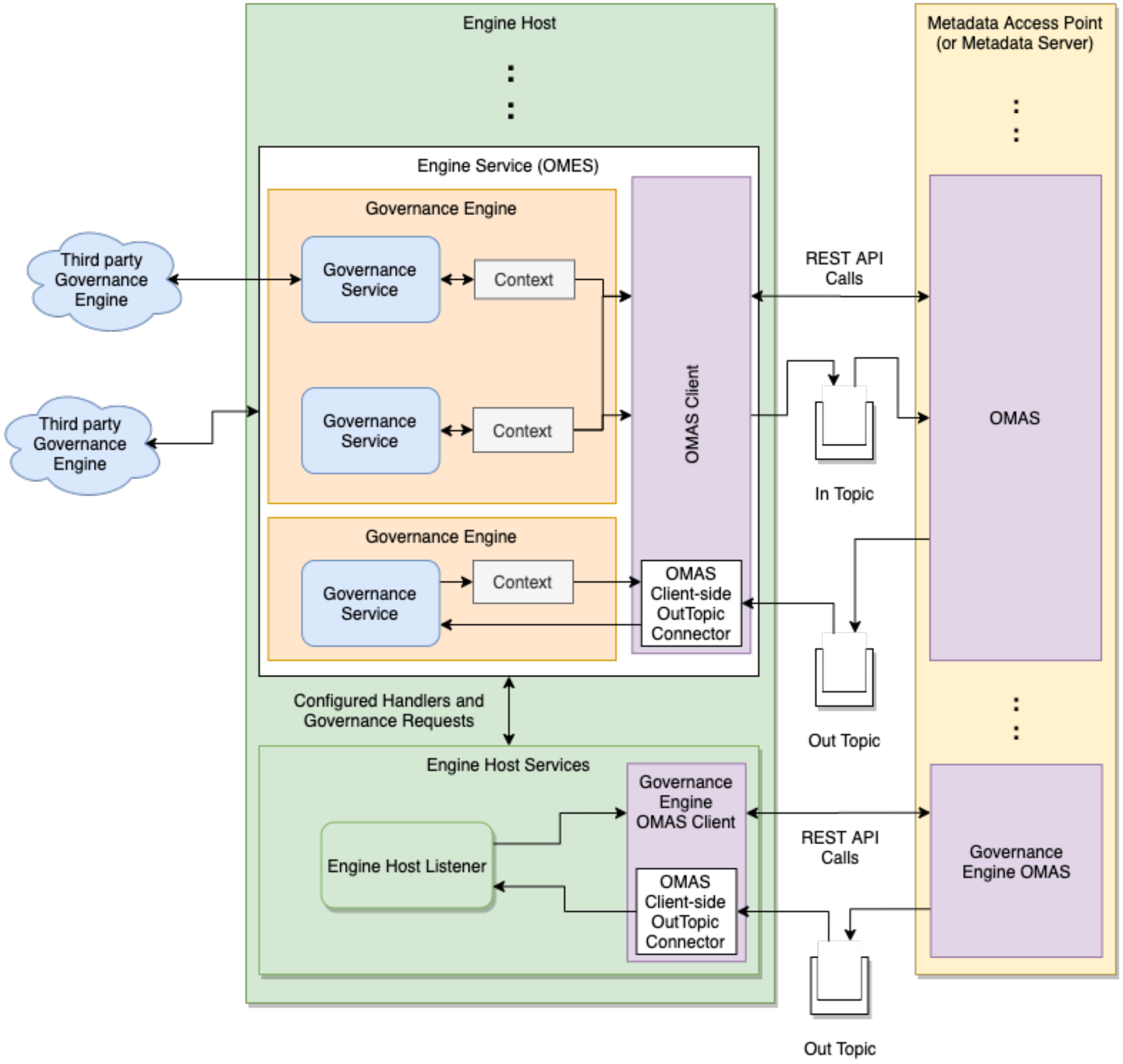
- Metadata from heterogenous sources is:
  - Inconsistent
  - Incomplete
  - Duplicated



- Governance services (running in governance engines) attempt to:
  - Discover new metadata by analyzing resources
  - Improve the quality of metadata across the ecosystem
  - Coordinate governance across multiple governance tools
  - Provide an audit trail of all action taken

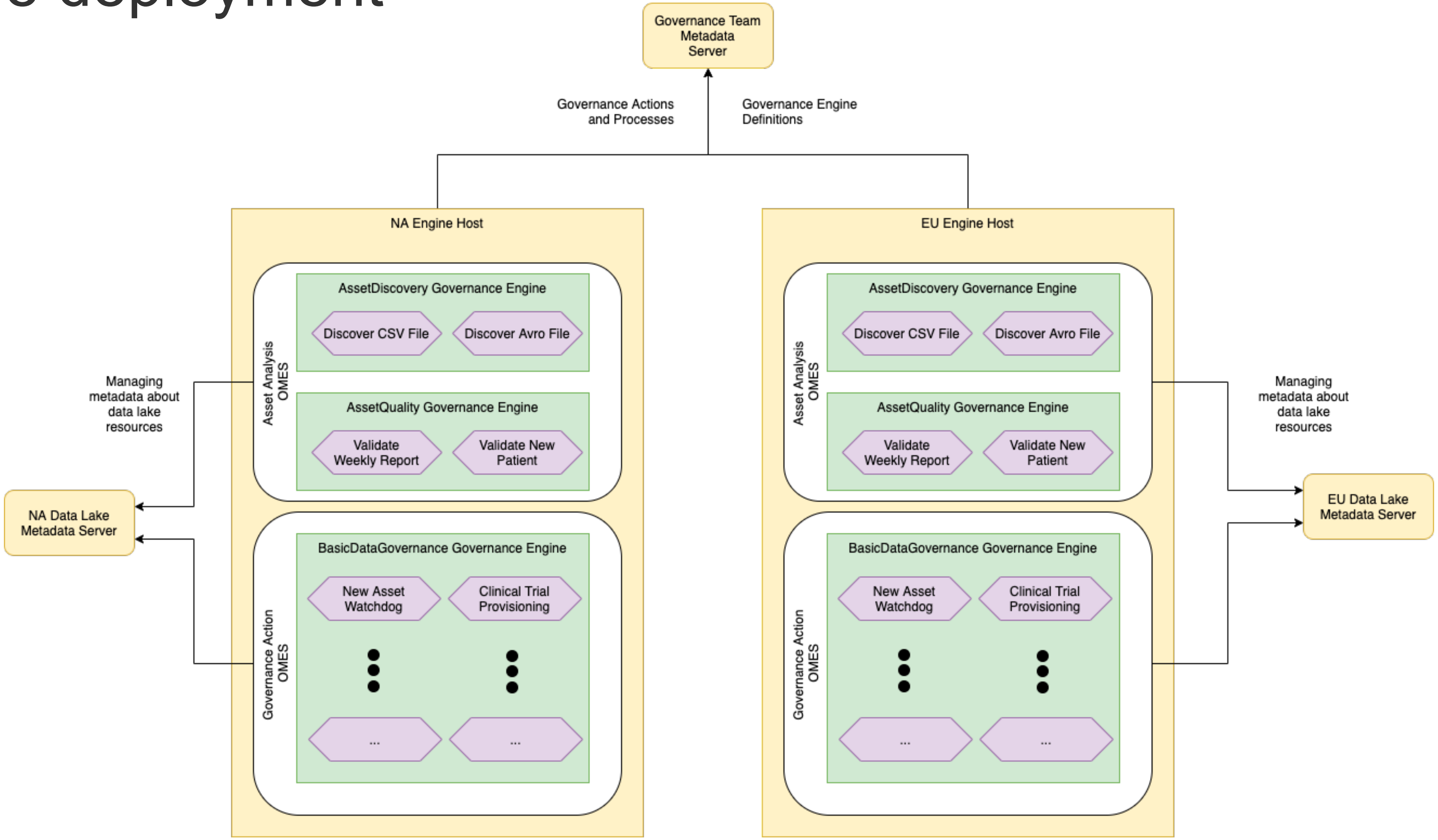
Inside the engine host

- Asset Analysis OMES
- Governance Action OMES

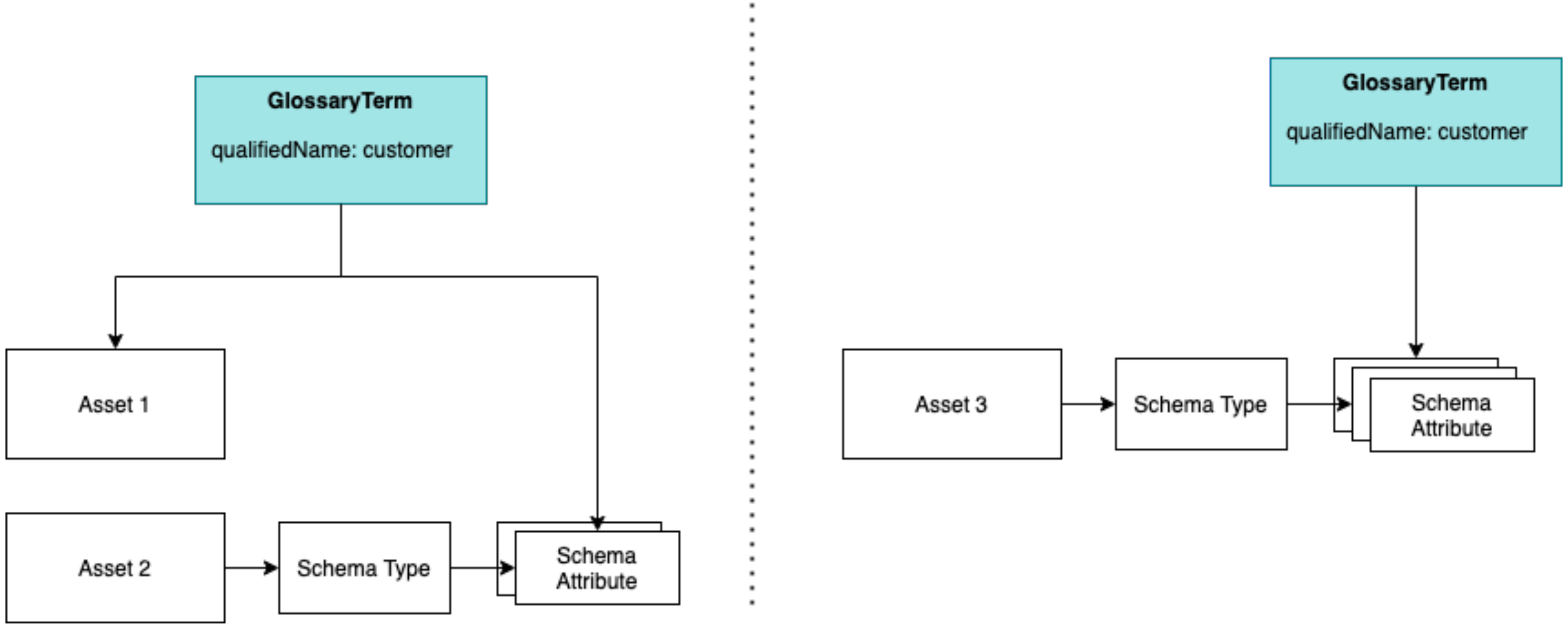


EGERIA

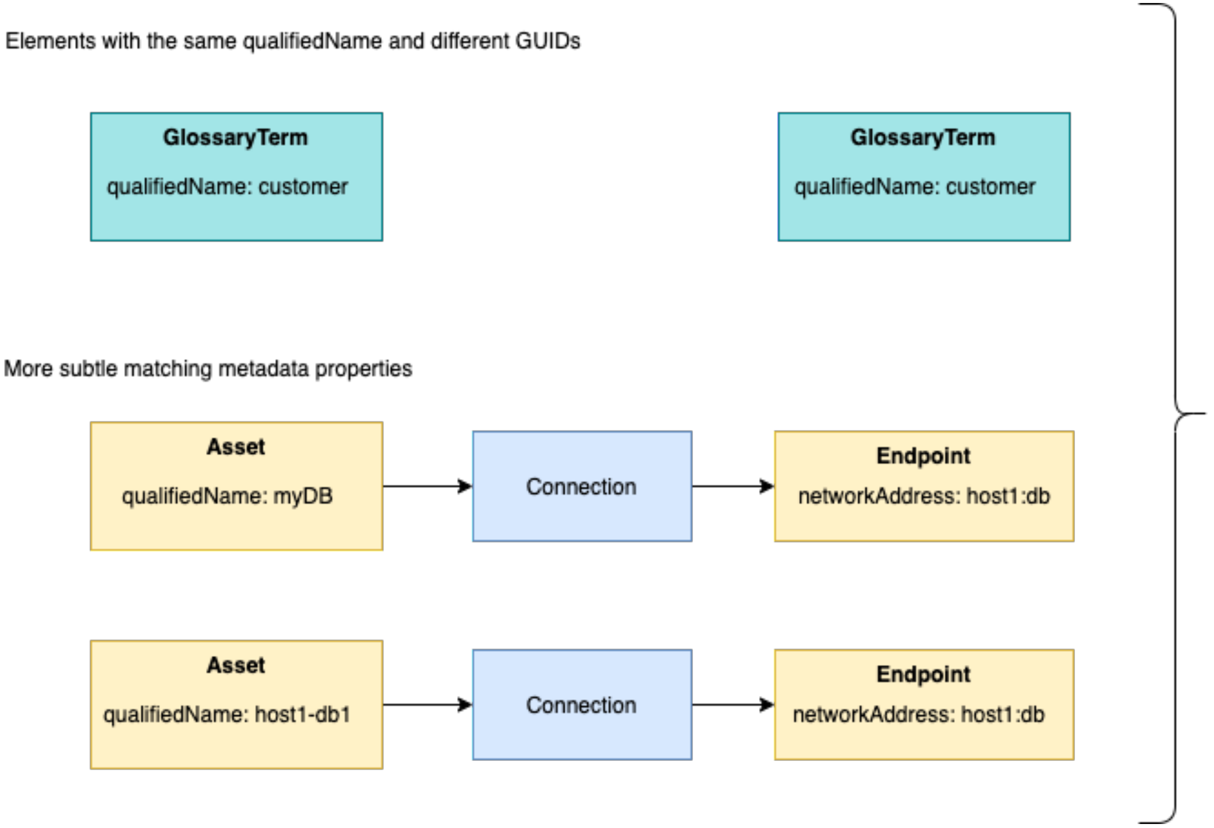
Example deployment


Governance Team Metadata Server
Governance Actions and Processes
Governance Engine Definitions
NA Engine Host
EU Engine Host
Asset Analysis OMES
AssetDiscovery Governance Engine
Discover CSV File
Discover Avro File
AssetQuality Governance Engine
Validate Weekly Report
Validate New Patient
Governance Action OMES
BasicDataGovernance Governance Engine
New Asset Watchdog
Clinical Trial Provisioning
...
...
Managing metadata about data lake resources
NA Data Lake Metadata Server
Managing metadata about data lake resources
EU Data Lake Metadata Server

Challenges with Duplicates



Types of Duplicates

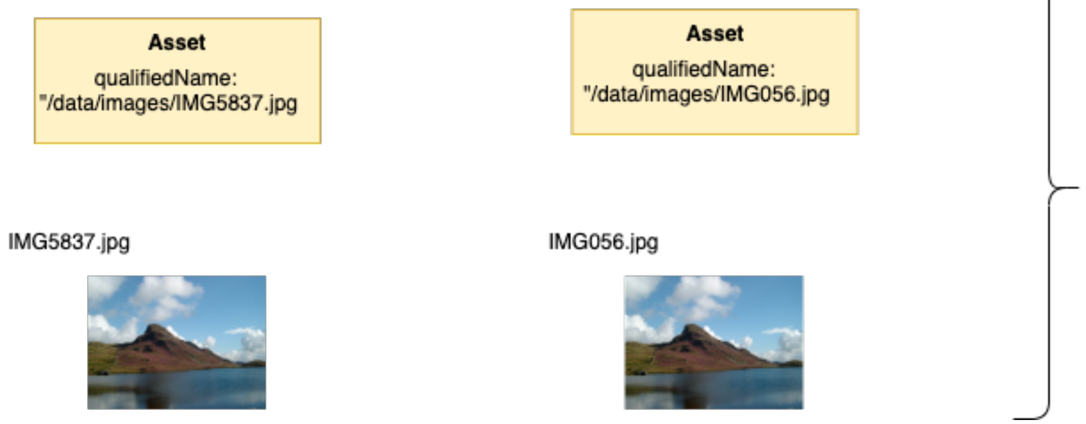


Avoided in OMASs by checking for previously defined elements before create/update/restore.

Detected through monitoring and scanning metadata from the connected repositories

Types of Duplicates

No correspondence in metadata but the contents of the real-world counterpart are the same.



Asset

qualifiedName: "/data/images/IMG5837.jpg

Asset

qualifiedName: "/data/images/IMG056.jpg

IMG5837.jpg

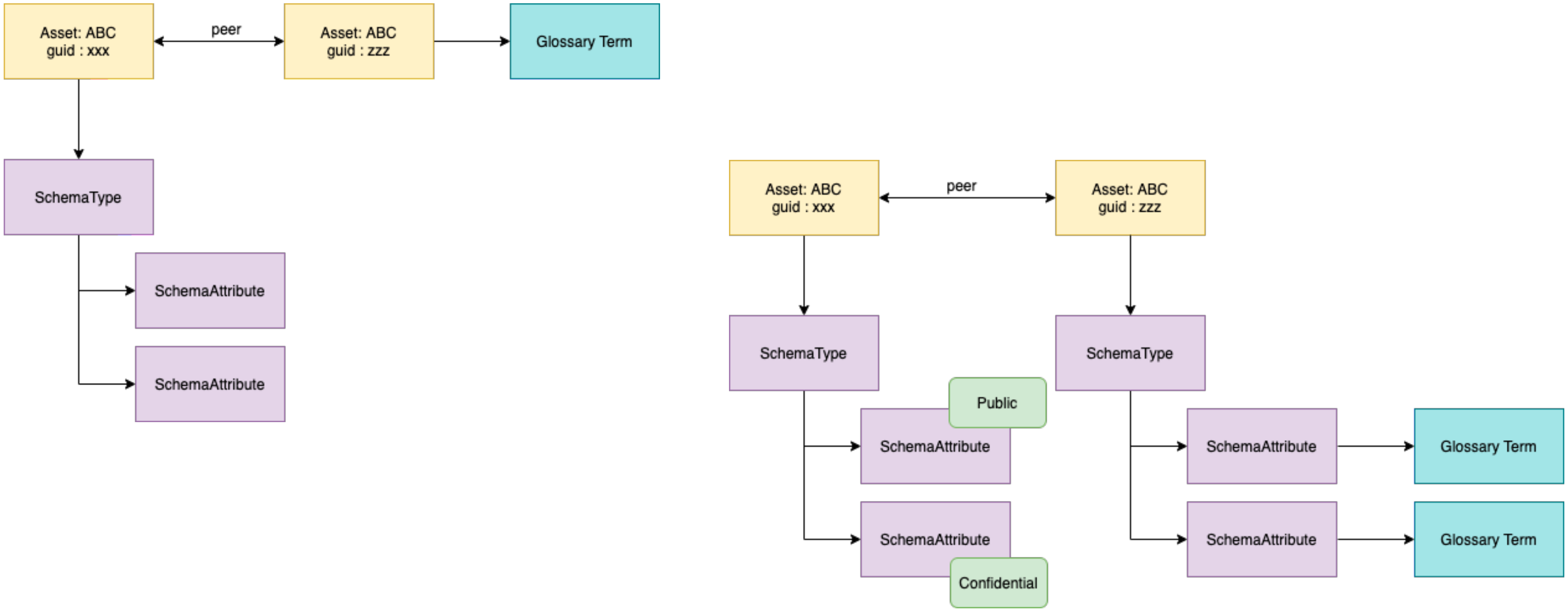
IMG056.jpg

Avoided through sound data management practices

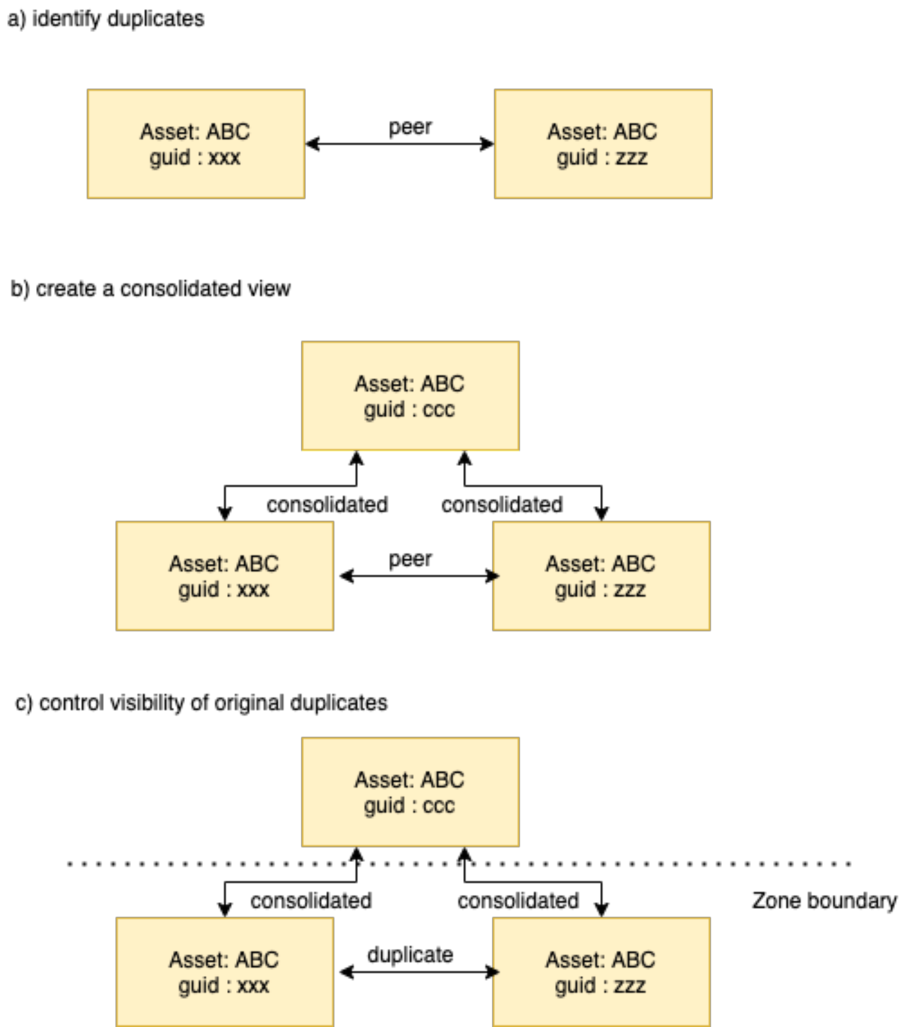
Metadata created by Automated Metadata Discovery; recorded in open metadata types as a FingerprintAnnotation

Annotations monitored using the metadata monitoring and scanning processes described above.

Examples of asset attachments that are combined



Example 3 step process



a) identify duplicates

Asset: ABC
guid : xxx

peer

Asset: ABC
guid : zzz

b) create a consolidated view

Asset: ABC
guid : ccc

consolidated

consolidated

Asset: ABC
guid : xxx

peer

Asset: ABC
guid : zzz

c) control visibility of original duplicates

Asset: ABC
guid : ccc

consolidated

consolidated

Zone boundary

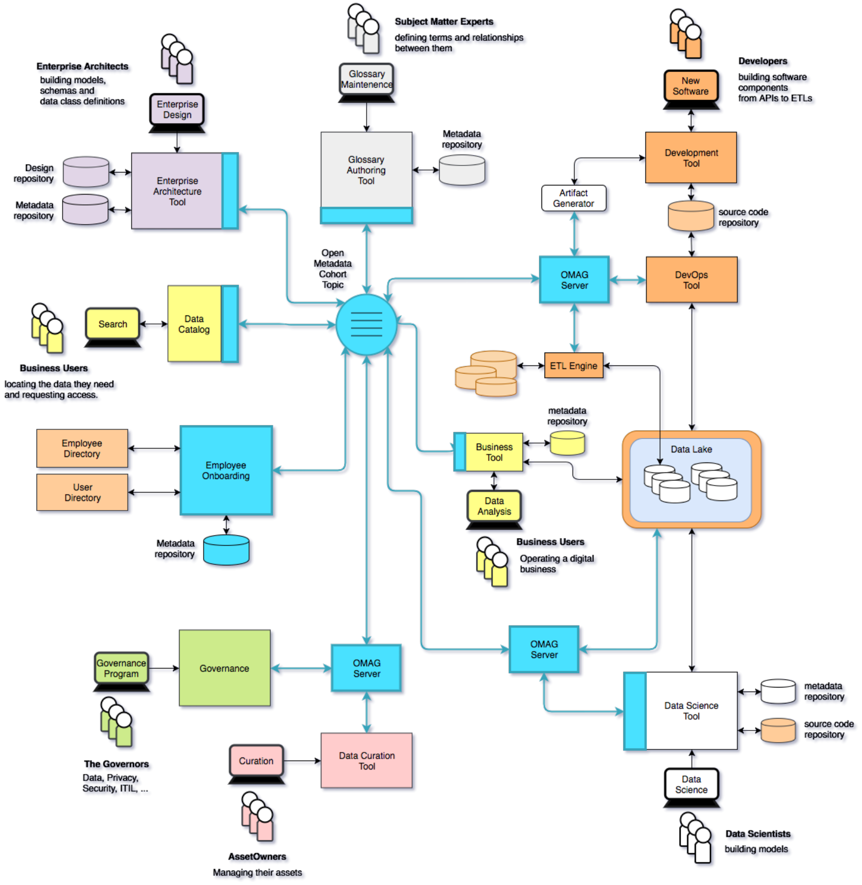
Asset: ABC
guid : xxx

duplicate

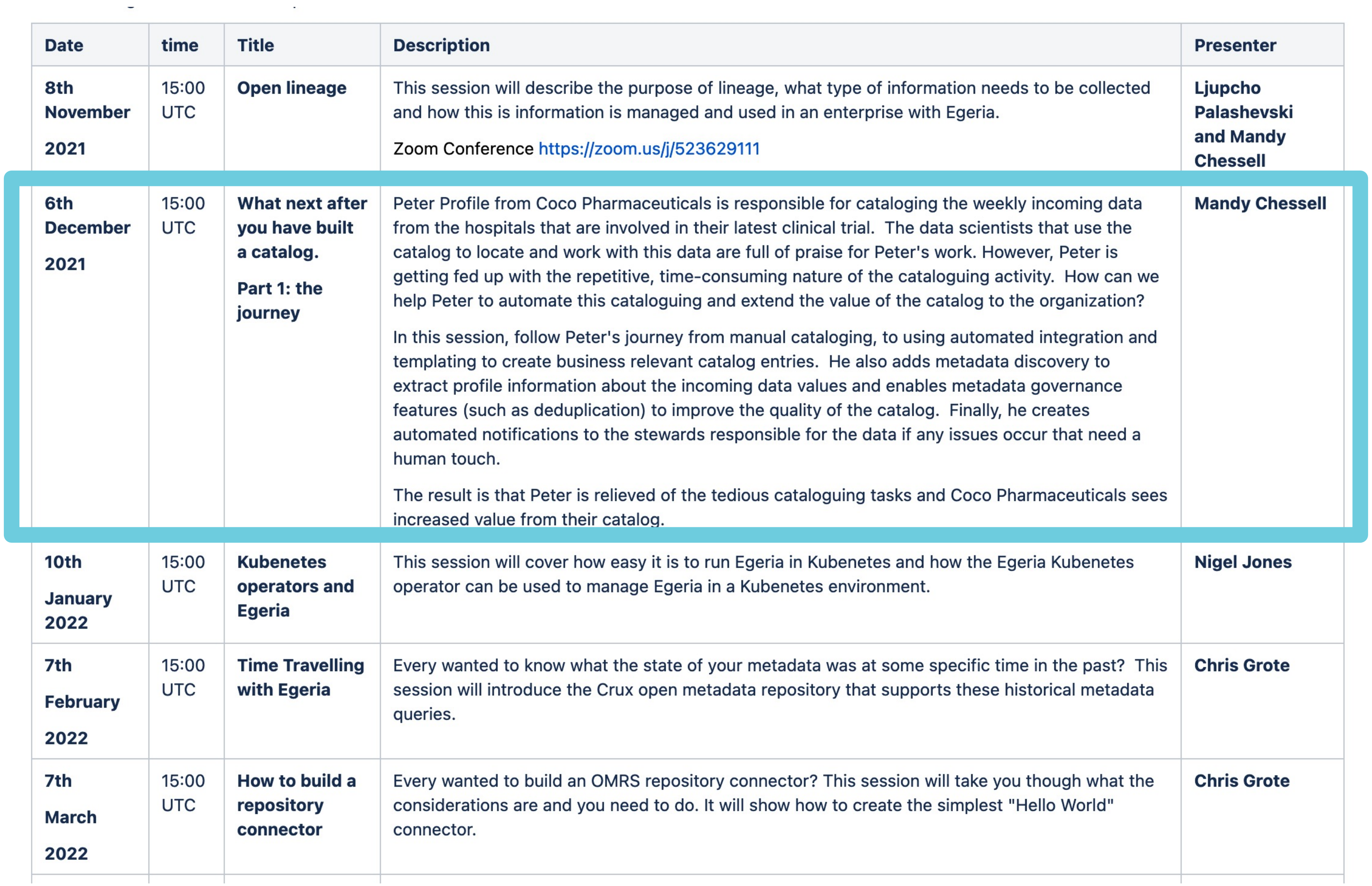
Asset: ABC
guid : zzz

Using Egeria …

- Eases the cost of metadata integration through
  - Comprehensive standards and libraries.
  - Active vendor recruitment program.
- Provides direct support to many governance roles, filling the gaps between function offered through commercial tools.
- Provides best practices and content packs to accelerate an organization’s journey to becoming data driven.



EGERIA



Date	time	Title	Description	Presenter
8th November 2021	15:00 UTC	Open lineage	This session will describe the purpose of lineage, what type of information needs to be collected and how this is information is managed and used in an enterprise with Egeria. Zoom Conference https://zoom.us/j/523629111	Ljupcho Palashevski and Mandy Chessell
6th December 2021	15:00 UTC	What next after you have built a catalog. Part 1: the journey	Peter Profile from Coco Pharmaceuticals is responsible for cataloging the weekly incoming data from the hospitals that are involved in their latest clinical trial. The data scientists that use the catalog to locate and work with this data are full of praise for Peter's work. However, Peter is getting fed up with the repetitive, time-consuming nature of the cataloguing activity. How can we help Peter to automate this cataloguing and extend the value of the catalog to the organization? In this session, follow Peter's journey from manual cataloging, to using automated integration and templating to create business relevant catalog entries. He also adds metadata discovery to extract profile information about the incoming data values and enables metadata governance features (such as deduplication) to improve the quality of the catalog. Finally, he creates automated notifications to the stewards responsible for the data if any issues occur that need a human touch. The result is that Peter is relieved of the tedious cataloguing tasks and Coco Pharmaceuticals sees increased value from their catalog.	Mandy Chessell
10th January 2022	15:00 UTC	Kubenetes operators and Egeria	This session will cover how easy it is to run Egeria in Kubenetes and how the Egeria Kubenetes operator can be used to manage Egeria in a Kubenetes environment.	Nigel Jones
7th February 2022	15:00 UTC	Time Travelling with Egeria	Every wanted to know what the state of your metadata was at some specific time in the past? This session will introduce the Crux open metadata repository that supports these historical metadata queries.	Chris Grote
7th March 2022	15:00 UTC	How to build a repository connector	Every wanted to build an OMRS repository connector? This session will take you though what the considerations are and you need to do. It will show how to create the simplest "Hello World" connector.	Chris Grote

Open forum

Questions?

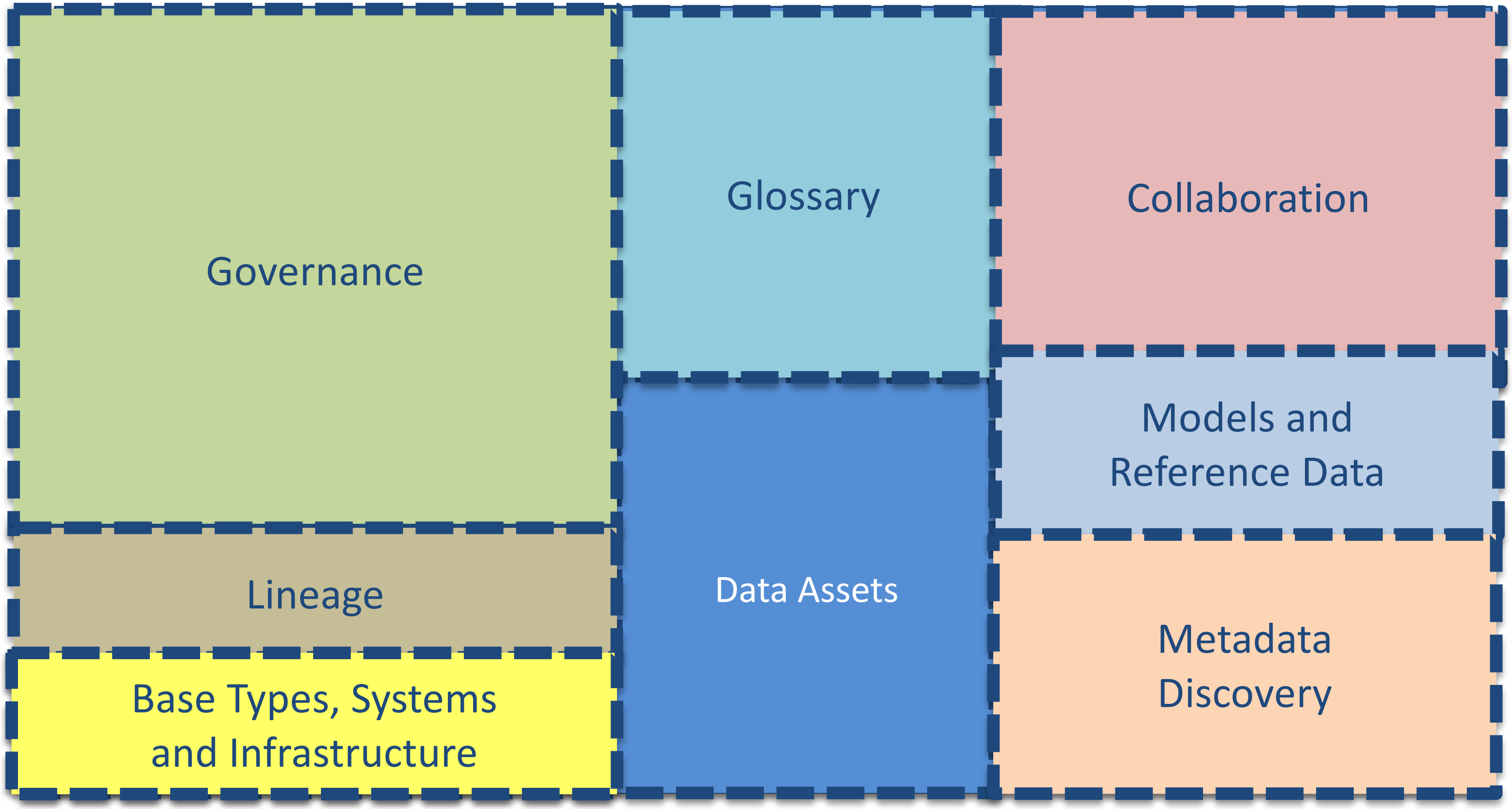
THANK YOU!

EGERIA

Achievements

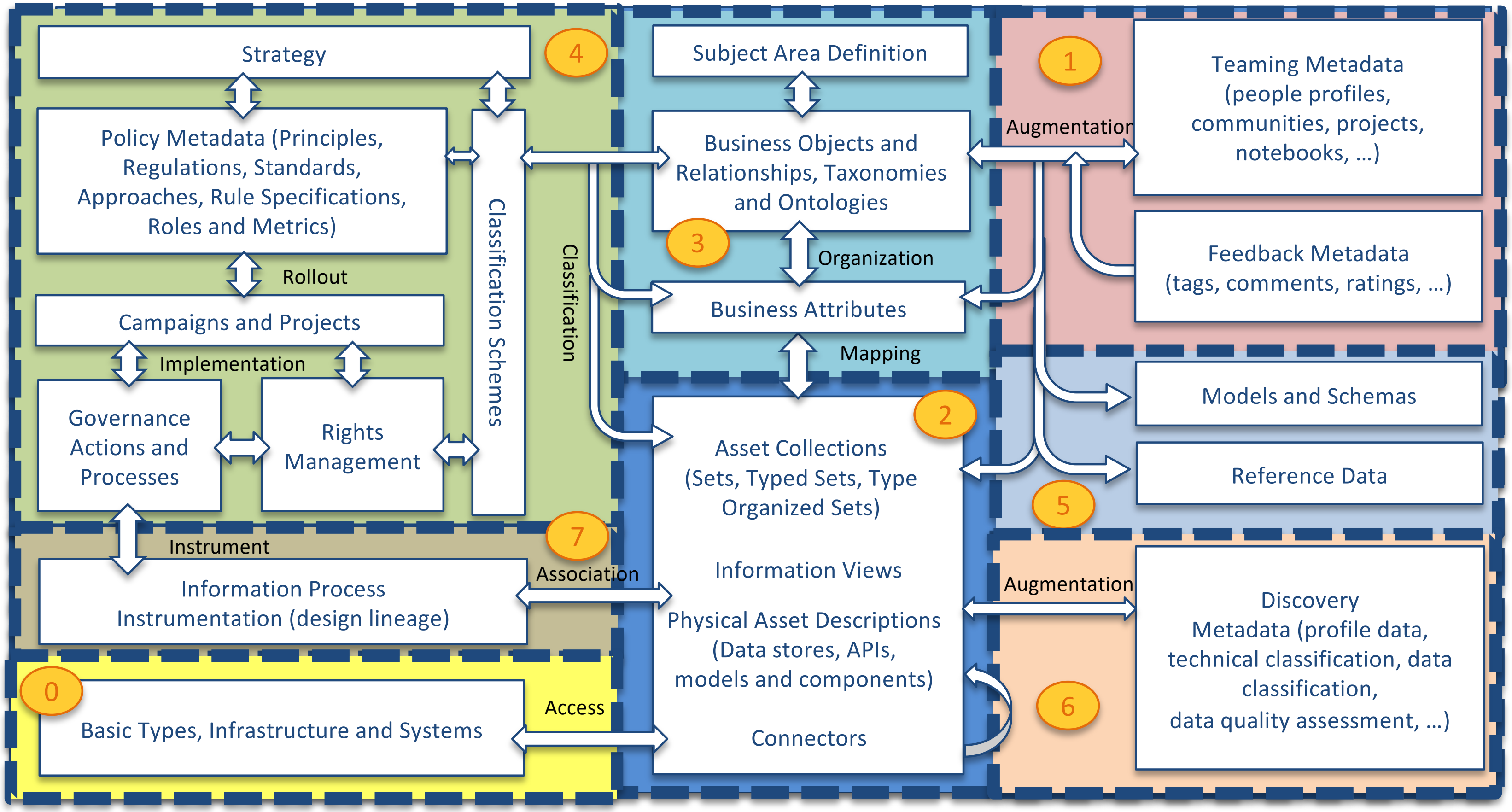
- 700 linked open metadata types demonstrating how the knowledge from many tools can be linked together.
- Open metadata repository interface proven for table, graph and hierarchical DB stores.
- Enterprise queries and replication across heterogeneous technologies
- Conformance test suite and mark
- Automated configuration of data virtualization technology and security as new data sets are added to a data lake
- Suite of persona-based labs and tutorial using Jupyter Notebooks.
- Virtual graph of metadata maintained across distributed heterogenous metadata repositories.
- Frameworks, APIs and connectors for minimizing integration cost for different types of technologies
- Virtual repository explorer UI
- Instance based security
- Controlling visibility of assets through zones
- Scalable, secure platform configurable and customizable through connectors
- Purpose-based data access
- Metadata versioning and provenance
- Multi-tenant UI based on carbon
- W3C semantic standards pattern for data model exchange
- Automation of metadata acquisition through templates, daemons, discovery services and stewardship.
- Classification of assets
- Reference data management
- Multi-technology collaboration and feedback
- Multi-domain governance model
- Digital service lifecycle, from business design, development, devOps and use.
- Comprehensive open lineage services.
- Metadata deduplication

Scope of metadata covered



https://egeria.odpi.org/open-metadata-publication/website/open-metadata-types/

Scope of metadata covered


Strategy
4
Policy Metadata (Principles, Regulations, Standards, Approaches, Rule Specifications, Roles and Metrics)
Rollout
Campaigns and Projects
Implementation
Governance Actions and Processes
Rights Management
Classification Schemes
Classification
Subject Area Definition
Business Objects and Relationships, Taxonomies and Ontologies
3
Organization
Business Attributes
Mapping
1
Augmentation
Teaming Metadata (people profiles, communities, projects, notebooks, …)
Feedback Metadata (tags, comments, ratings, …)
Models and Schemas
Reference Data
5
2
Asset Collections (Sets, Typed Sets, Type Organized Sets)
Information Views
Physical Asset Descriptions (Data stores, APIs, models and components)
Connectors
7
Instrument
Association
Information Process Instrumentation (design lineage)
0
Basic Types, Infrastructure and Systems
Access
Augmentation
6
Discovery Metadata (profile data, technical classification, data classification, data quality assessment, …)

EGERIA

https://egeria.odpi.org/open-metadata-publication/website/open-metadata-types/