

What is Open Metadata and why do we need it?

Our ongoing journey to a fully meta data driven information landscape

Ferd Scheepers, Chief Information Architect ING

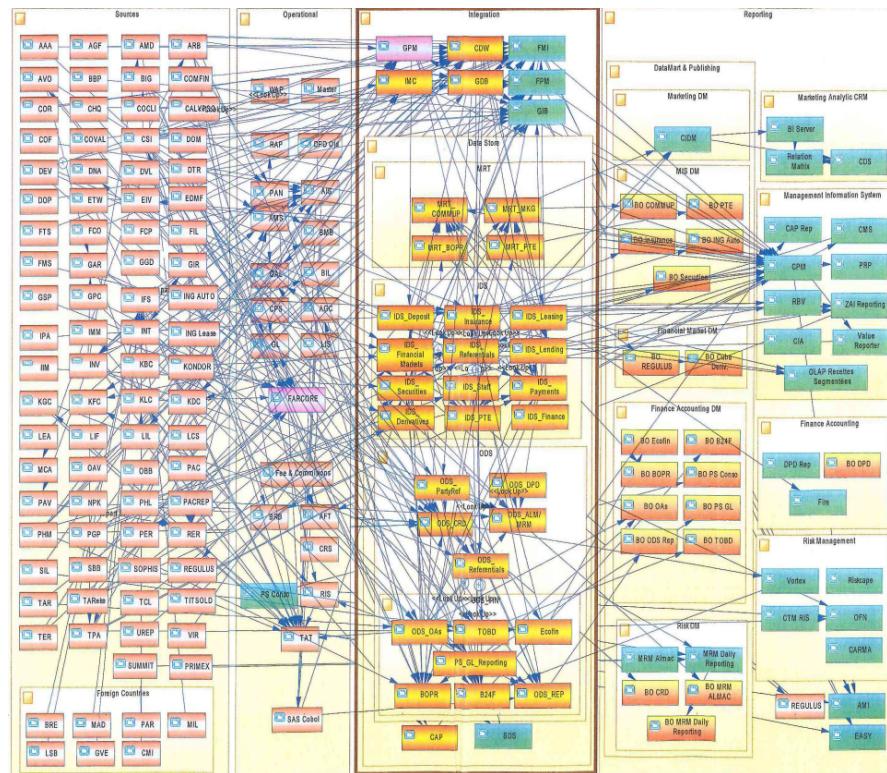
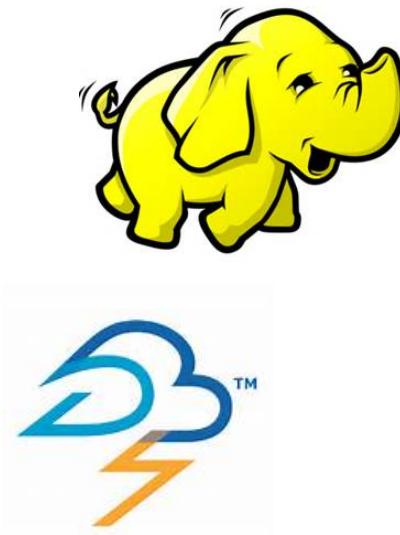
August 2018



Or maybe the title should be: Why would a bank get involved in an open source project around Open metadata?

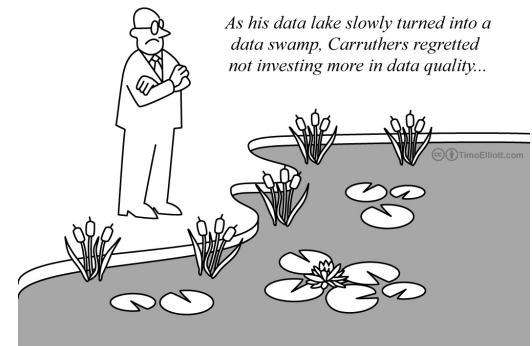


5 years ago we started on a transformational journey to get in control of our complex data landscape.

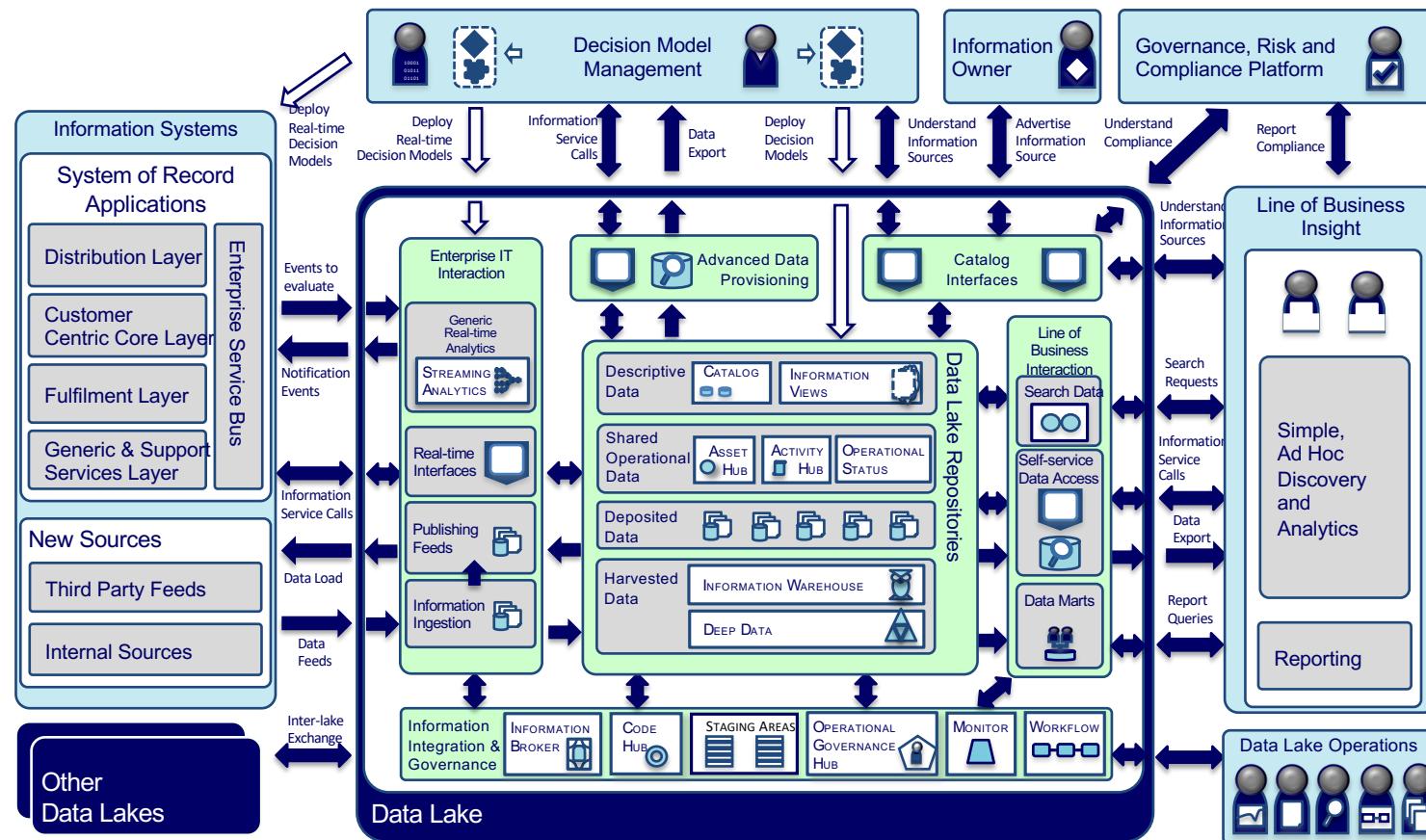


data
steward^{provider}
chief^{champion}
custodian
assistant^{user}
Officer
owner
czar

As his data lake slowly turned into a data swamp, Carruthers regretted not investing more in data quality...



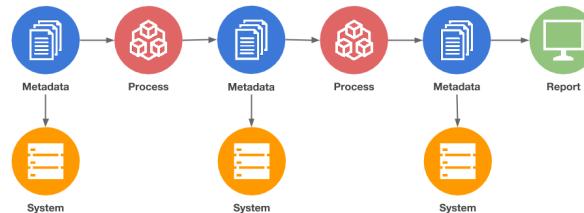
We created the Data Lake architecture together with IBM, we built the Data Lakes, and rolled them out globally.



Our Data lake is driven by metadata, which is much more than just a technical schema, a datatype or a technical asset name.



Location



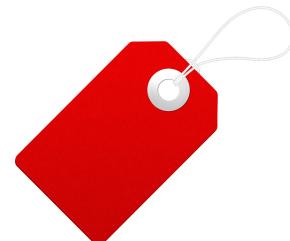
Lineage/Origin



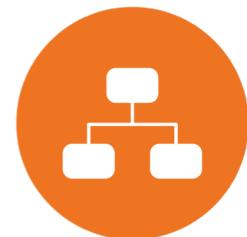
Owner



Quality



Classification
& security



Structure

IGC is our central Metadata repository where we manage all Metadata.



IBM INFOSPHERE INFORMATION GOVERNANCE CATALOG

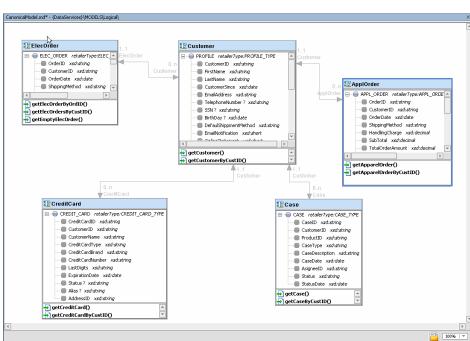
InfoSphere Information Governance Catalog

IBM InfoSphere Information Governance Catalog enables the enterprise to expose a central catalog of glossary assets and information assets to end users. The meaning of those assets can be broadened by adding technical descriptors and business context.

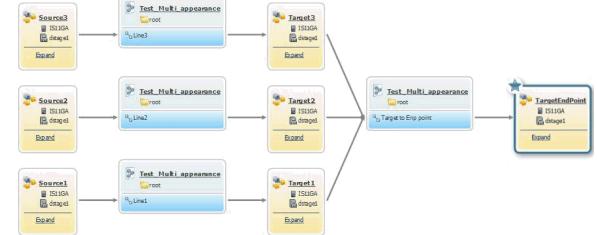
InfoSphere Information Governance Catalog provides search, browse, and query capabilities. In addition, you can establish asset collections and run lineage reports to examine data flow between assets.

Search Glossary Information Assets Labels Queries Collections Quick Term Finder

Categories & Terms Policies & Rules Information Assets Collections Queries

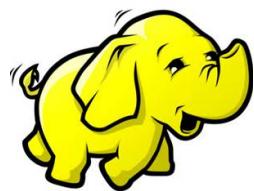


6



In a Data Lake we know where all the data is, what it means, and where it comes from. As long as we stick to IBM (IIS) tools.

But of course our landscape is more complex..

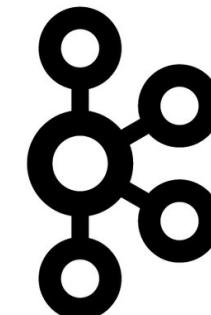


ORACLE



sas

IBM
Information
Governance
Catalog
YOUR METADATA



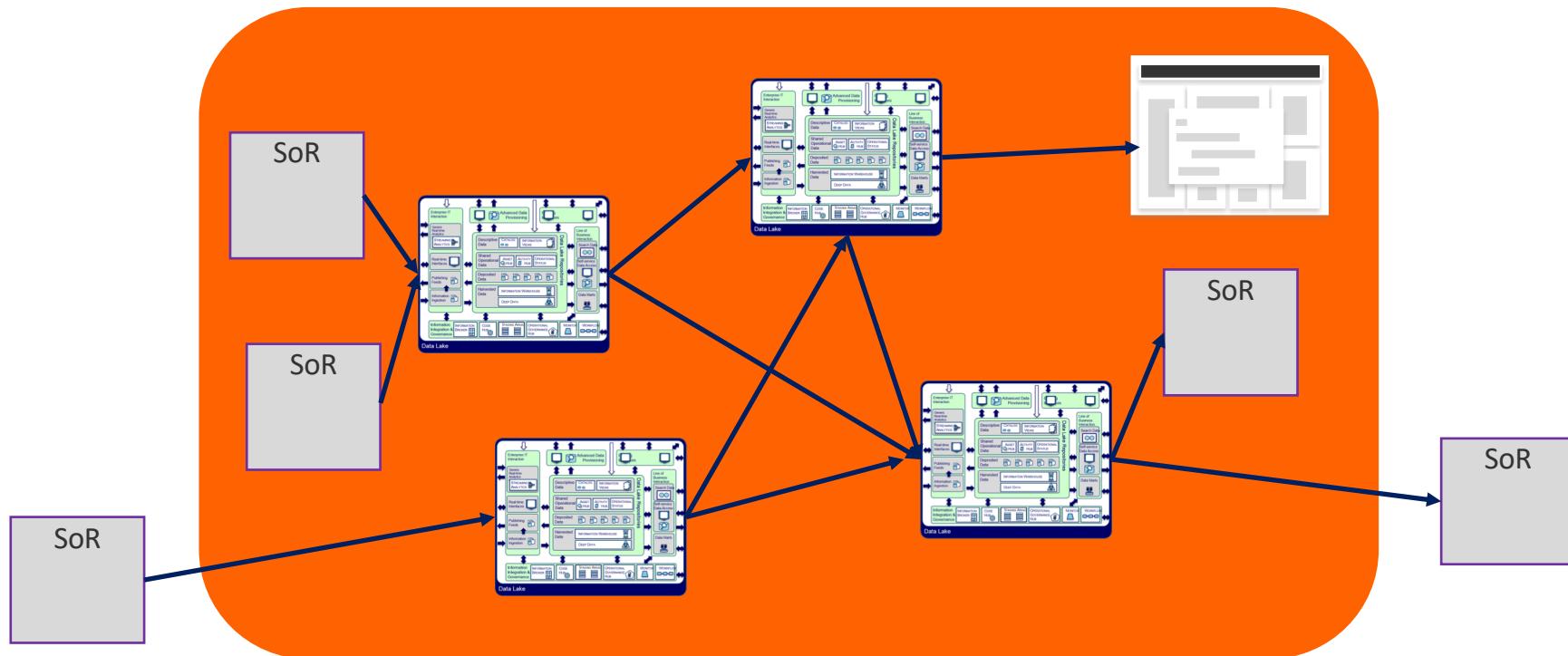
IBM
COGNOS

**Microsoft
Power BI**

collibra

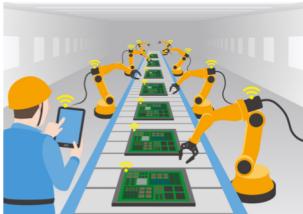
ING

And the overall data landscape is not limited to a single datalake. The data world is (becoming) very complex.



We need a way to manage our data in a consistent way across our heterogeneous landscape. Enter Open Metadata.

The maintenance of metadata must be automated



Metadata management must become ubiquitous



Metadata access must become open
and remotely accessible



Metadata should be used to drive the governance of data

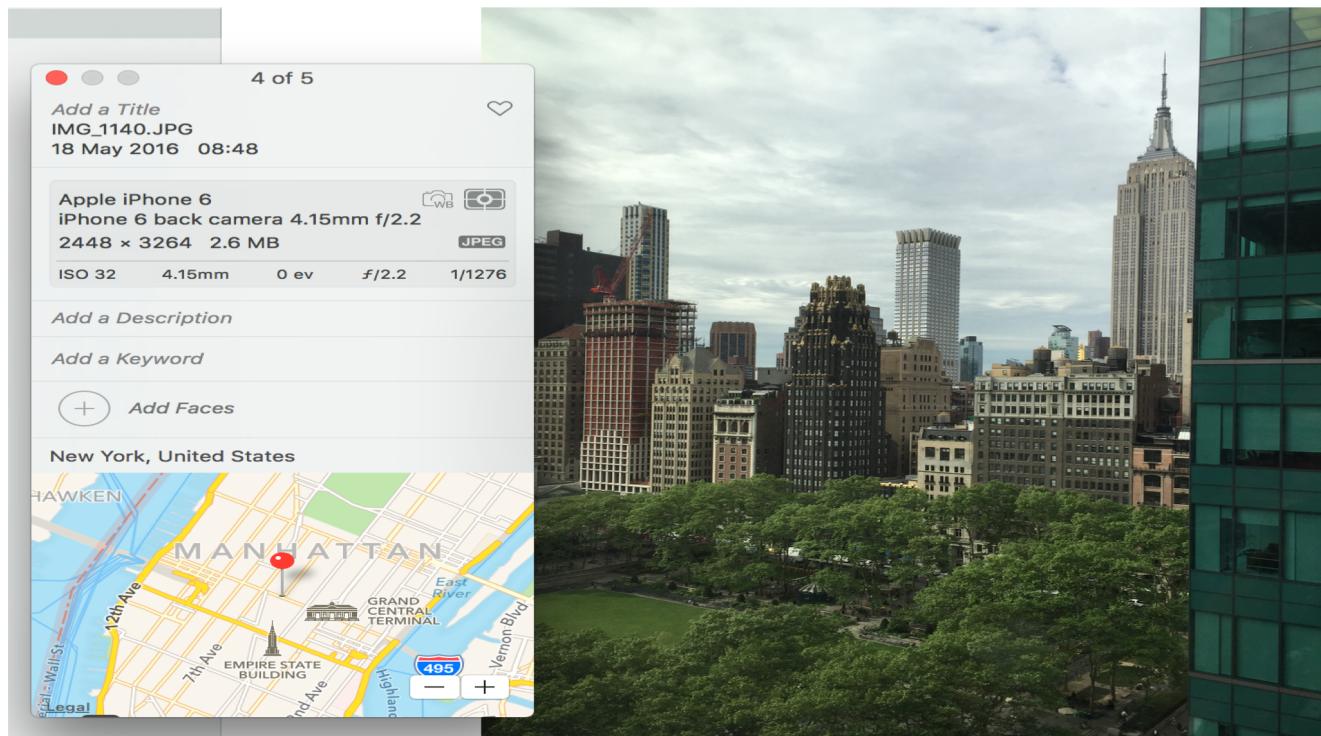


Wherever possible, discovery and maintenance of metadata has to be an integral part of all tools that access, change and move information.

Everything is metadata

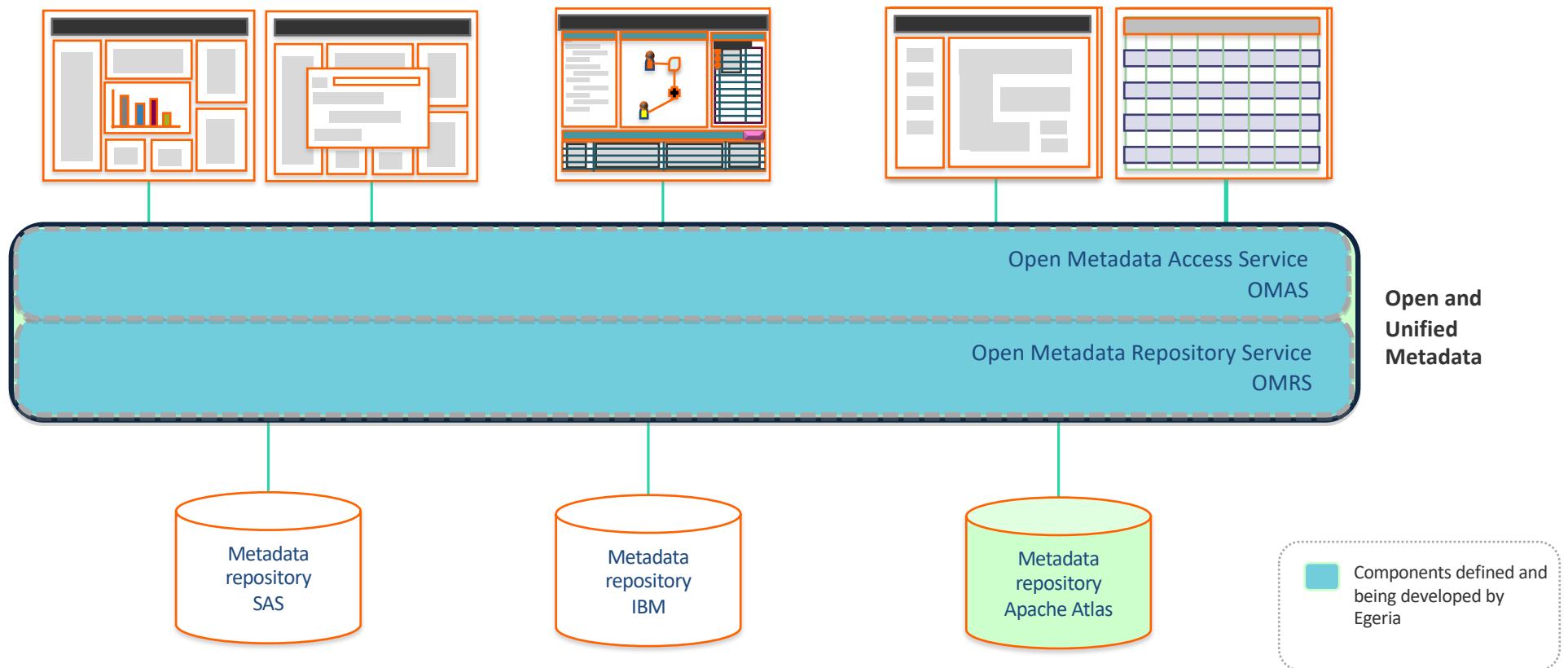
David Weinberger

There are really good examples of Metadata standardisation, think about JPEG or MP3.



metadata is captured when the data is created, moved with the data and can be augmented and processed by any of the vendor tools.

Open and Unified Metadata makes metadata exchange and out-of-the-box end user tools integration possible.



Open Metadata is comprised of open types, APIs, exchange protocols and frameworks.

- Common base



an open set
of APIs



an open set of
metadata
types



an open set of
exchange
protocols

- Frameworks



Governance



Discovery

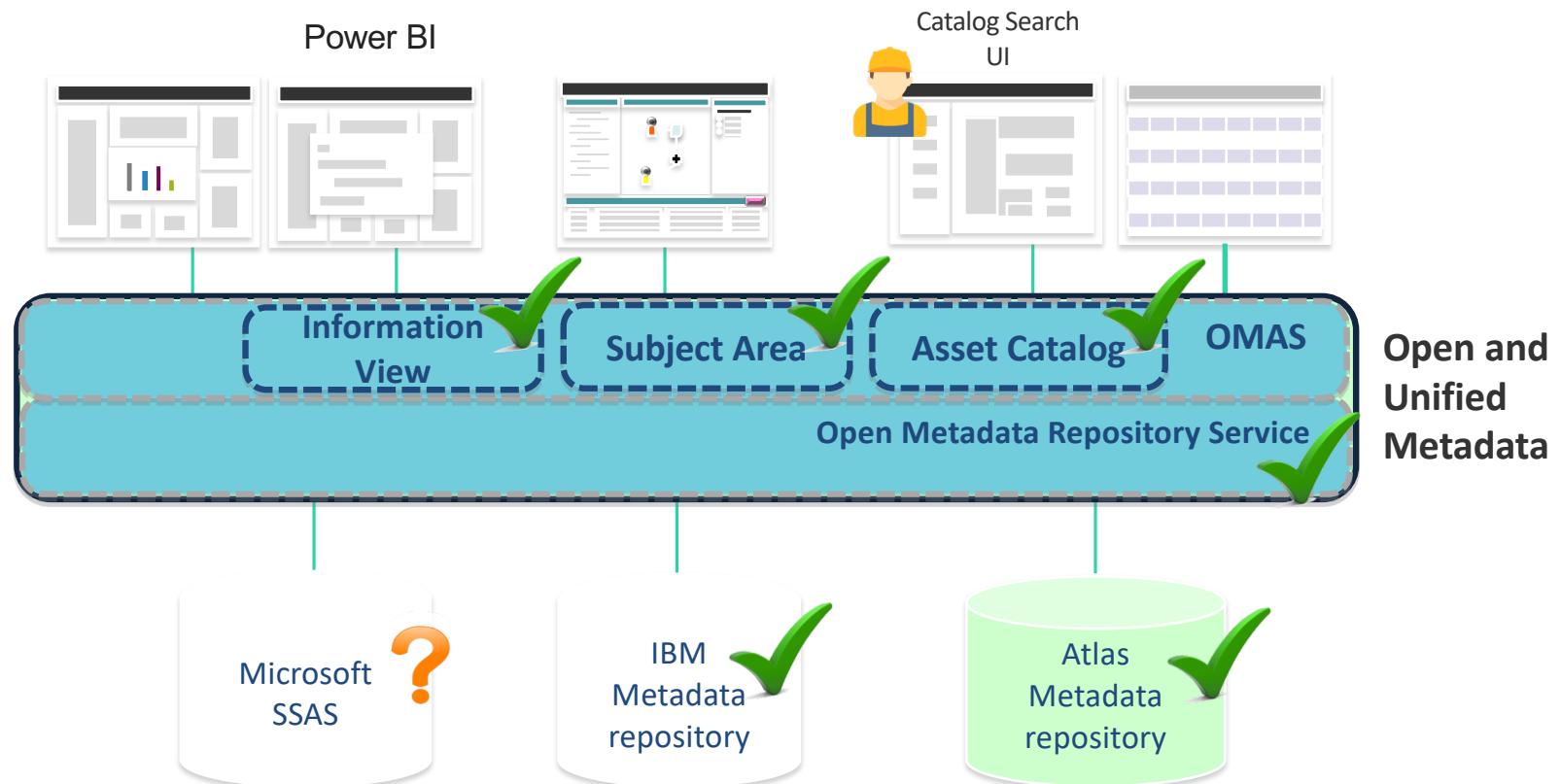


Access
Frameworks

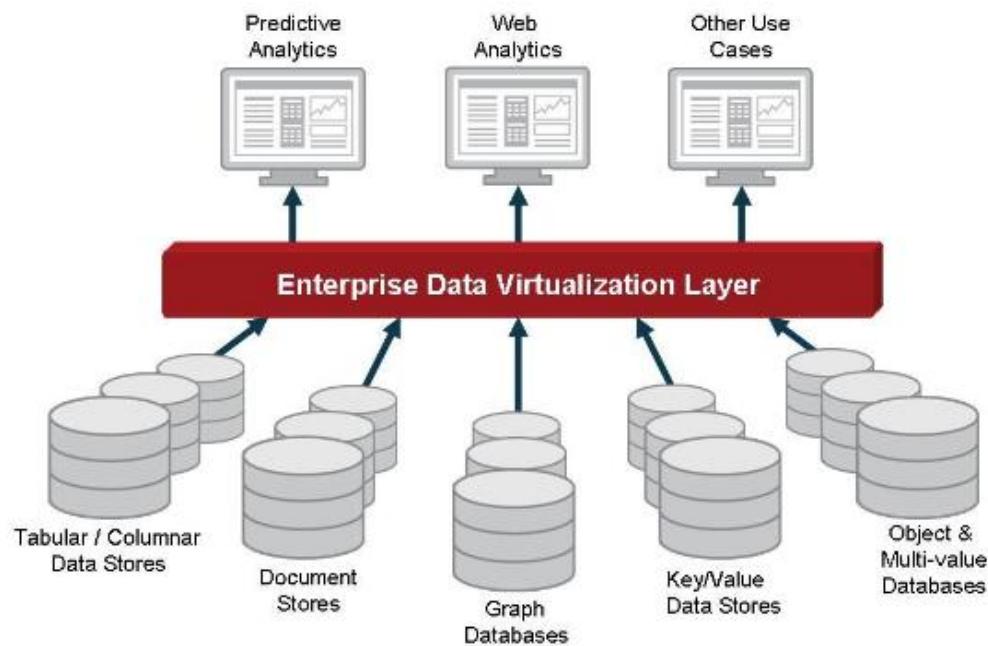
This is not the first attempt to standardise metadata. Why will we succeed where others haven't?

- All vendors have gone though, or are going through a similar internal standardisation
- We build this with a reference application, all in Open Source. No paper standards
- We drive it from the consumers and vendors together. Vendors alone will not come to a standard

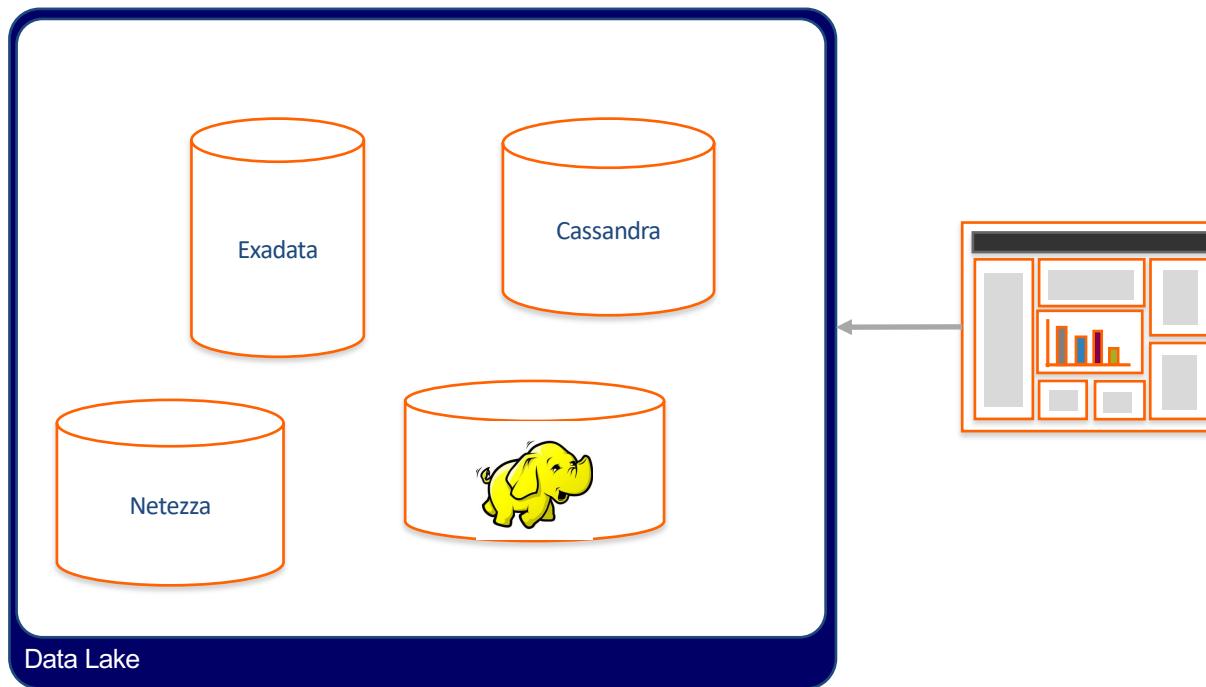
Core components of Open Metadata have been designed and have an initial implementation.



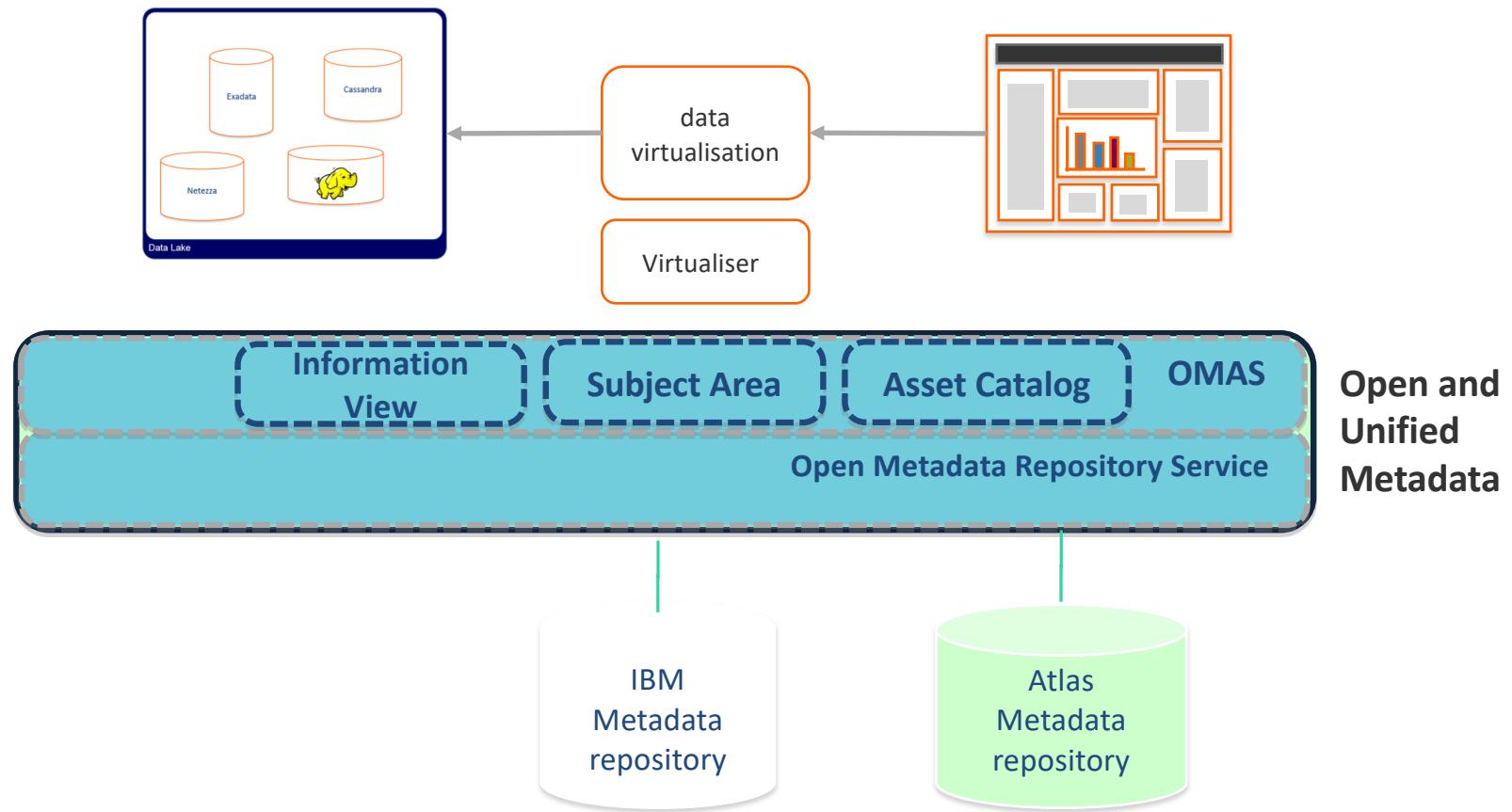
Having a vision is great, but we need a concrete deliverable to prove that we can make open meta data work



Inside the Data Lake, multiple repositories hold data, security and names of physical assets are not standardised



Data virtualisation will create a semantic layer around the Data Lake that also enforces security based on business terms



Questions, Thoughts, Suggestions.



Egeria Open Metadata and Governance Project
www.odpi.org