Practical application of knowledge graphs using EDM Council’s CDMC information model

Thursday, July 13
11 AM EST

Knowledge Graph
Special Interest Group
Update
Knowledge Graph Special Interest Group Update
13 July 2023

A Graph industry update with:

Shaun Rolls
Senior Advisor
Head of Knowledge Graph SIG
EDM Council

Jim Halcomb
Head of Product Management
EDM Council

Terry Hedin
Director, Data Architect,
Product Manager
LSEG

Ben Clinch
Head of Information Architecture
BT Group
Today’s panel

**Moderator**

**Jim Halcomb**  
Head of Product Management  
EDM Council

**Shaun Rolls**  
Senior Advisor  
Head of Knowledge Graph SIG  
EDM Council

**Terry Hedin**  
Director, Data Architect, Product Manager  
LSEG

**Ben Clinch**  
Head of Information Architecture  
BT Group
Poll #1

What is your current role?

a. Analytics
b. Architecture
c. Technology
d. Ontology/Modeling
e. Data Governance
Data is Everywhere
KNOWLEDGE GRAPHS

- WHAT ARE THEY
- HOW ARE THEY USED
WHAT IS A KNOWLEDGE GRAPH

The “graph” in Knowledge Graph refers to a way of organizing and continually evolving your data that highlights relationships between data points.

Graph data looks like a network of interconnected points. This is in contrast to databases like Oracle or MySQL — relational systems — where data is stored in tables (rows, columns, cells).

While knowledge graphs are dynamic via evolving data nodes that are continually connected/expressed.

Distributed data sets that have shared references for interoperability.
KNOWLEDGE GRAPHS AND SEMANTIC DATA

- Knowledge graphs are structured via semantic data that focuses on the **nature of meaning**.

- The **connection** of words, concepts, symbols and language that are related to provide meaning.
WHO USES KNOWLEDGE GRAPHS TODAY?

- Medicine, Pharma & Health
- Technology
- Financial Services
Let’s Define Data Fabric

It’s an emerging data management design for attaining ... that utilizes ... in support of ... regardless of ... It is not one single tool or technology.

- Flexible, reusable and augmented data integration pipelines
- Knowledge graphs, semantics and ML/AI on active metadata
- Faster and, in some cases, automated data access and sharing
- Deployment options, use cases (operational or analytical) and/or architectural approaches
Gartner’s Evaluation of Data Fabric

Technology Pillars of the Data Fabric Design

Market Maturity: [High] [Medium] [Low]

As of June 2022:

<table>
<thead>
<tr>
<th>Benefit Rating</th>
<th>Transformational</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tech Maturity</td>
<td>Emerging</td>
</tr>
<tr>
<td>Market Penetration</td>
<td>1% to 5% of Target Audience</td>
</tr>
<tr>
<td>Time to Mainstream</td>
<td>5-10 Years</td>
</tr>
</tbody>
</table>

Source: Quick Answer: What Is Data Fabric Design? (G00767157)
Poll #2

What is your current knowledge graph experience level?

a. Explorer – curious about topic
b. Early Adopter – starting to learn and consider applications
c. Practitioner – applying KG in one or more projects
d. Expert – own and maintain knowledge graph models
Cloud Data Management Capabilities

CDMC Working Group

Cloud Challenges

- **Regulatory Mandate:** safeguard cloud computing services to protect customers’ sensitive information
- **Inefficiency:** data, technology, regulatory and planning challenges on nearly every cloud implementation
- **Complexity:** 89% of firms use 2 or more cloud providers*

CDMC Group Objectives

1. **Best Practices** for a hybrid-cloud world
2. **Cloud Data Controls** to meet regulatory obligations for protecting Sensitive Data
3. **Accelerate Trusted Cloud Adoption** modeled after the DCAM data management industry framework

* Source: Flexera, 2022 State of the Cloud
CDMC Industry Engagement
100+ leading firms – 300+ SME participants – 17 GSIB’s

CDMC Working Group
LSEG Morgan Stanley
J.P.Morgan
Standard Chartered
HSBC Societé Générale
TP ICAP
Deutsche Bank
BNP PARIBAS
Freddie Mac
CAPCO

Cloud & Technology Providers
Microsoft Azure
IBM Cloud
Google Cloud
AWS
collibra
snowflake
Informatica
BigID
data.world
PRIVITAR
Solidatus
securiti
KPMG

Regulatory Engagement
US: Federal Reserve, SEC, CFTC, FDIC
Canada: OSFI
UK: BoE, FCA, ICO
EU: ECB, ESMA (pending)
Germany: BaFin
Switzerland: FinMA
Australia: APRA
Singapore: MAS
Israel: Bank of Israel
India: RBI, SEBI (pending)
Africa/Middle East: 20+ Regulators
Others in process…

CDMC Adoption Support
• Training Courses
• Cloud Service Certification
• Open Source Tools
• CDMC Authorized Partner Program

CDMC Working Group Delivered release v1.1
28 September 2021

CDMC Framework Validated
Cross Industry
• Life Sciences
• Telecommunications
• Manufacturing
• Retail / Services / CPG
• Consumer Tech
• Government / Defense
CDMC Framework

160+ Pages
6 Components
14 Capabilities
37 Sub-capabilities
14 Key Controls

Best practice framework for the migration, management, and enablement of sensitive data in cloud

Free license – Download at: EDMCouncil.org > Frameworks > CDMC
Cost Metrics

14 Key Controls for Managing Data Risk

Data Sovereignty and Cross-Border Movement of sensitive data

Authoritative Data Sources and Provisioning Points

Data Ownership

Data Control Compliance

Data Lineage

Data Quality Measurement

Data Retention, Archiving and Purging

Data Privacy Impact Assessments

4.0 Protection & Privacy

3.0 Accessibility & Usage

2.0 Cataloguing & Classification

1.0 Governance & Accountability

Data Cataloguing

Entitlements and Access for Sensitive Data

Data Classification

Data Consumption Purpose

Security Controls

Data Retention, Archiving and Purging

Sensitive Data includes classifications such as:

- Personal Information (PI) / Sensitive Personal Data
- Personally Identifiable Information (PII)
- Personal Health Information (PHI)
- Company or Client Identifiable Information
- Material Non-Public Information (MNPI)
- Specific Information Sensitivity Classifications (such as ‘Highly Restricted’ and ‘Confidential’)
- Critical Data Elements used for important business processes
- Licensed data

CDMC 14 Controls
Free Download License: EDMCouncil.org > Frameworks
Poll #3

Why are you interested in Knowledge Graph?

a. Analytics
b. Fraud Detection
c. Network Management
d. Data Lineage
e. Data Fabric/Catalog
f. Data Science
g. Other
What is it?

Conceptual graph metadata model of entities and relationships required to deliver CDMC.

Simple machine-readable (RDF) vocabulary aligned to data management capabilities.

Developed by millions of geniuses
Objectives

- Provide a **reference information metadata model** for designing information systems in line with the CDMC Framework
- Provide a **formal vocabulary** and point of connection between the capabilities of the CDMC framework
- Create **interoperability** between existing open standards to harmonize data sharing
- Allow practitioners to **evidence data management controls** using a consistent model across technologies
- Demonstrate how the information model **supports regulatory review**
CDMC provides a top-level vocabulary based on existing and open standards.

Influence from 16 open vocabularies
How the Information Model Support the Key Controls

1. Data Control Compliance must be monitored for all data assets containing sensitive data via metrics and automated notifications. The metrics must be calculated from the extent of implementation of the CDMC Key Controls specified in subsequent sections.

2. Confirm data asset controls are continuously operational and tracked for sensitive assets for all other CDMC controls.

3. Generate report/dashboard that presents the metrics for sensitive data controls.

4. Confirm the ability to set an escalation threshold for each metric.

5. Test the automated reporting against threshold limits:
   4.1 Impact the effectiveness of a Key Control in a manner that causes a metric threshold to be outside threshold limits.
   4.2 Confirm the automated reporting of the move outside threshold limits.

6. The Data Sovereignty and Cross-Border Movement of sensitive data must be recorded, auditable and controlled according to defined policy.

7. Classification must be automated for all data at the point of creation or ingestion and must be always on.

8. Data Consumption Purpose must be provided for all Data Sharing Agreements involving sensitive data.

9. Appropriate Security Controls must be enabled for sensitive data and evidence must be recorded.

10. Data Protection Impact Assessments must be automatically triggered for all personal data according to its jurisdiction.

11. Data Retention, Archiving and Purging must be managed according to a defined retention schedule.

12. Data Quality Measurement must be enabled for sensitive data with metrics distributed when available.

13. Cost Metrics directly associated with data use, storage and movement must be available in the catalog.

14. Data Lineage Information must be available for all sensitive data.

Term Mapping

<table>
<thead>
<tr>
<th>Control Test Term</th>
<th>Information Model Term</th>
<th>Notes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Asset</td>
<td>Data Asset</td>
<td></td>
</tr>
<tr>
<td>Metric</td>
<td>Control Metric / Control Metric Measurement</td>
<td></td>
</tr>
<tr>
<td>Sensitive Data</td>
<td>Data Asset / Classification Occurrence / Information Sensitivity</td>
<td></td>
</tr>
<tr>
<td>Escalation Threshold</td>
<td>Control Metric Threshold</td>
<td></td>
</tr>
<tr>
<td>Key Control</td>
<td>Control</td>
<td></td>
</tr>
</tbody>
</table>

"Sensitive Data" as used in this document should be implemented as defined in Information Sensitivity, classifications that, if applied to a Data Asset, indicate that the Data Asset should be treated as sensitive.
Information Model : Live Demo
Snowflake, AWS & Microsoft Announce CDMC Automated Data Risk Control Framework

Microsoft achieves first native Cloud Data Management Capabilities certification

Mike Flasko  Vice President and General Manager, Data Governance
Google Cloud becomes the next CDMC Cloud Certified Solution
BT becomes the first telecommunications organization to complete CDMC assessment
Poll #4

What are your current blockers to knowledge graph adoption in your organization?

a. Lack of expertise
b. No clear path to adoption
c. Unclear business benefits / ROI
d. Insufficient funding
e. Lack of stakeholder buy-in
f. Confusion regarding tooling and integration
The Data Fabric and Data Mesh movement
What it means for collaboration

“Data Fabric” and “Data Mesh” as terms, point towards Information Architecture thinking bringing together disciplines across the business.
Data Fabric: A practical definition for BT

Data Fabric is **Smart**, **Automated** **Data Management at Scale**

Use of knowledge graphs, semantic understanding of data and its context (active metadata)

The practice of managing the risks and the value of the data within an enterprise (establishing ownership, policies, standards, controls, improving data quality and data value)

Consistent and auditable, automation of common data governance activities such as metadata, data handling policies and controls using machine learning, AI and repeatable patterns

An order of magnitude usually considered within or above the petabyte range

**Point to Note:** Data Fabric doesn’t replace the need for data management; it automates it
Wrapping the Fabric around the Mesh

The data fabric wraps sustainable, automated data governance around the data mesh.

The fabric provides policies, standards and tools to make federated data governance consistent and work for all of BT.
Information Architecture Conceptual View: A simplified view

Data Marketpace, Data Usage

- Data Handling & Controls
- Metadata Harvesting, Lineage & Data Landscaping
- Information Modelling, Metadata & Data Ownership
- Trusted Sources & Data Quality
- Data Retention & Archiving

Integration

- Controlling access to data, recording purpose of use, semantically describing the data for improved search, structuring and grouping data for maximum re-use and interoperability
- Connecting sources, measuring data quality
- Recognising record types and sending to archives, communicating and applying maximum retention actions
- Support for constructing the information model / knowledge graph and connecting it to the enterprise.
- Recording, disseminating and applying machine readable data handling policies
- Automatically recognising, storing and propagating metadata, including provenance and purpose of processing
- Highly performant integration and transport paths

- Recording, disseminating and applying machine readable data handling policies
“The Information Model is for our data what the organisation chart is for our people.

In a digital company data is one of our most valuable assets.

The information model needs to be as pervasive in our everyday thinking and activities as the org chart.”

The Information Model (a knowledge graph) provides Data Fabric with a structure and a rich set of semantic meaning that can be used to inform entity resolution and automate controls while simplifying the way we manage data.
## Information Model Usage: A Worked Example

<table>
<thead>
<tr>
<th>Instance Data Terms</th>
<th>Information Model Term</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification Occurrence</td>
<td>instance of applying a Data Classification to a Resource</td>
<td></td>
</tr>
<tr>
<td>Control</td>
<td>defined way to ensure control of data assets</td>
<td></td>
</tr>
<tr>
<td>Control Instance</td>
<td>implementation of a Control for a Data Asset</td>
<td></td>
</tr>
<tr>
<td>Control Metric</td>
<td>definition of a measurement that determines the effectiveness of a Control</td>
<td></td>
</tr>
<tr>
<td>Control Metric Measurement</td>
<td>value of a Control Metric at a point in time</td>
<td></td>
</tr>
<tr>
<td>Control Metric Threshold</td>
<td>value against which a Control Metric Measurement is compared to assess the effectiveness of a Control</td>
<td></td>
</tr>
<tr>
<td>Data Asset</td>
<td>collection of data owned by an organization that is considered to have intrinsic value</td>
<td></td>
</tr>
<tr>
<td>Information Sensitivity</td>
<td>classification of the information within a data asset that indicates the level of control and protection that must be applied to the asset due to the nature of the data and its sensitivity or importance to the organization or to the subject of the data</td>
<td></td>
</tr>
<tr>
<td>Threshold Constraint</td>
<td>relationship between a Control Metric Measurement and the corresponding Control Metric Threshold that determines the effectiveness of a Control</td>
<td></td>
</tr>
</tbody>
</table>
Summary

☑️ How a distributed architecture benefits from common structures and standards (a.k.a. the information model)

☑️ Four modes in which you can use the information model and the comparative benefits:

   Reference – Authoritative interpretation of the metadata requirements for CDMC compliance

   Mapping – Scalable implementation to produce a catalog of catalogs (Key Control 1)

   Digital twin – Smart query across data sets that is consistent with the business concepts

   Interoperability – Integration with vendor solutions and harmonize data sharing across supply chain
Cloud Data Management Capabilities

Thank You!

Opportunities for Industry Collaboration

• New Knowledge Graph Course – Q4 2023
• Volunteer to help w/enhancement of the Information Model extensions to DCAM v3
• Volunteer to test the CDMC Information Model for existing use cases to improve or enhance the model

• CDMC Information Model Graph Viewer

• CDMC Information Model Test Mapping