TopQuadrant™

How Implementing Data Governance with Knowledge Graphs Enables Enterprise AI

TopBraid EDG™

Enterprise Data Governance

TOPQUADRANT COMPANY

FOUNDATION

- TopQuadrant was founded in 2001
- Strong commitment to standards-based approaches to data semantics

MISSION

- Empower people and drive results — by making enterprise information meaningful

FOCUS

- Provide comprehensive data governance solutions using knowledge graph technologies

CUSTOMERS

- Over 120 active customer organizations

TopQuadrant™

Today's Agenda


Irene Polikoff

- A brief history knowledge graphs in 90 seconds

  … and what are they anyway?
- How do knowledge graphs support data governance?
- A concise overview of AI and ML technologies
- Knowledge graphs provide a powerful platform for both integrated data governance and enterprise AI / ML
- Some examples of how TopBraid EDG uses AI (KR + ML) for enhanced information governance

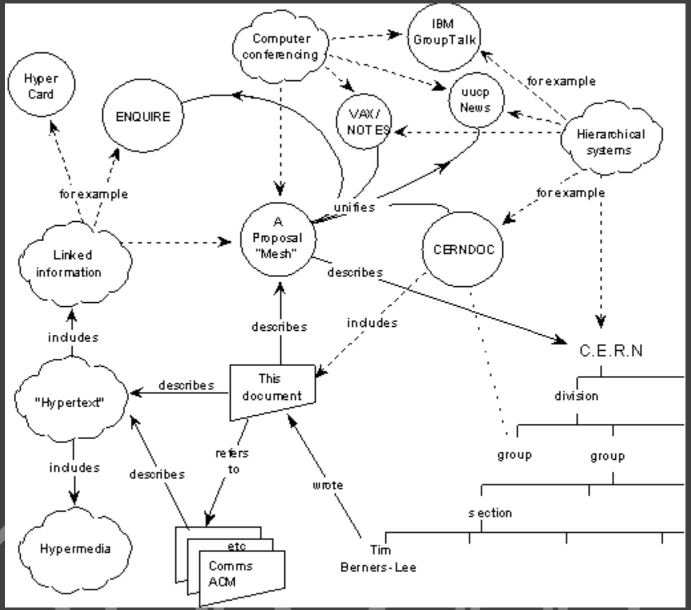
TopQuadrant™

Today's Agenda

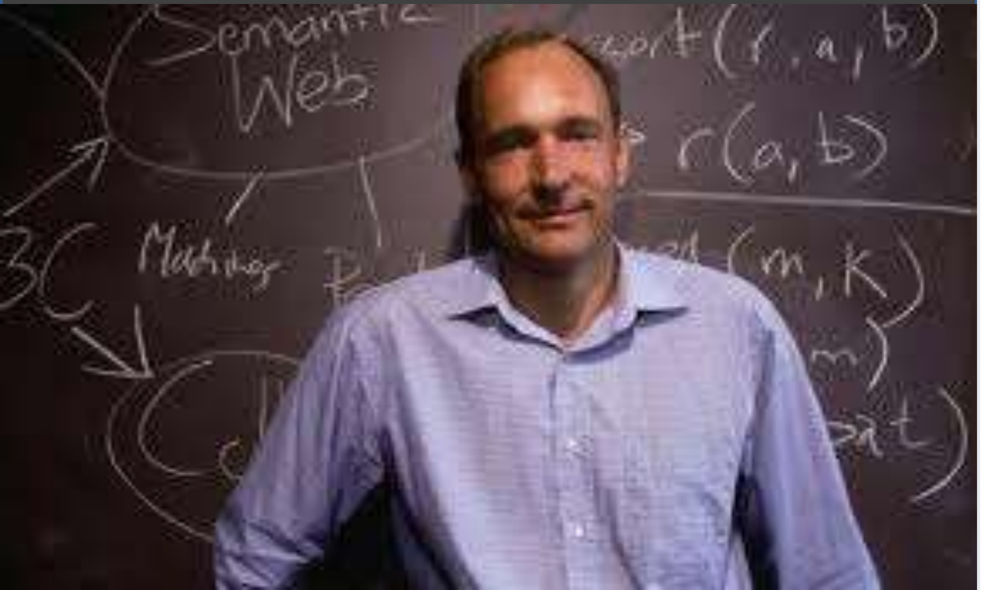
Irene Polikoff

- A brief history knowledge graphs in 90 seconds
  … and what are they anyway?
- How do knowledge graphs support data governance?
- A concise overview of AI and ML technologies
- Knowledge graphs provide a powerful platform for both integrated data governance and enterprise AI / ML
- Some examples of how TopBraid EDG uses AI (KR + ML) for enhanced information governance

2001
Tim Berners-Lee

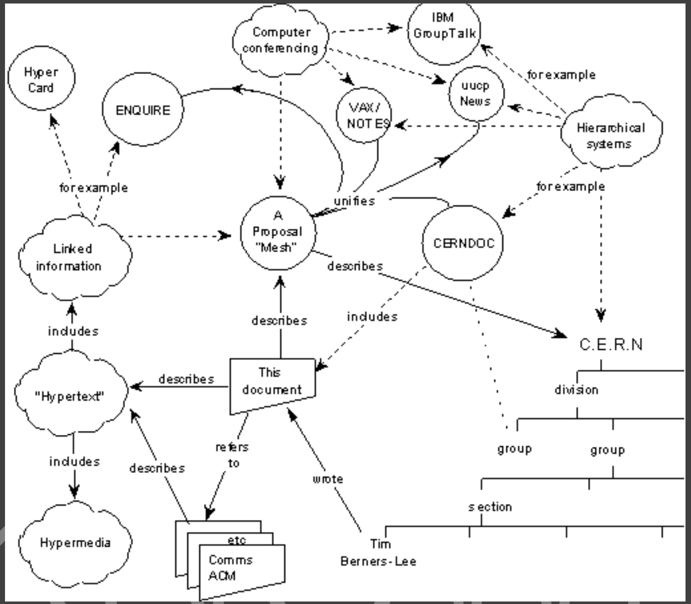


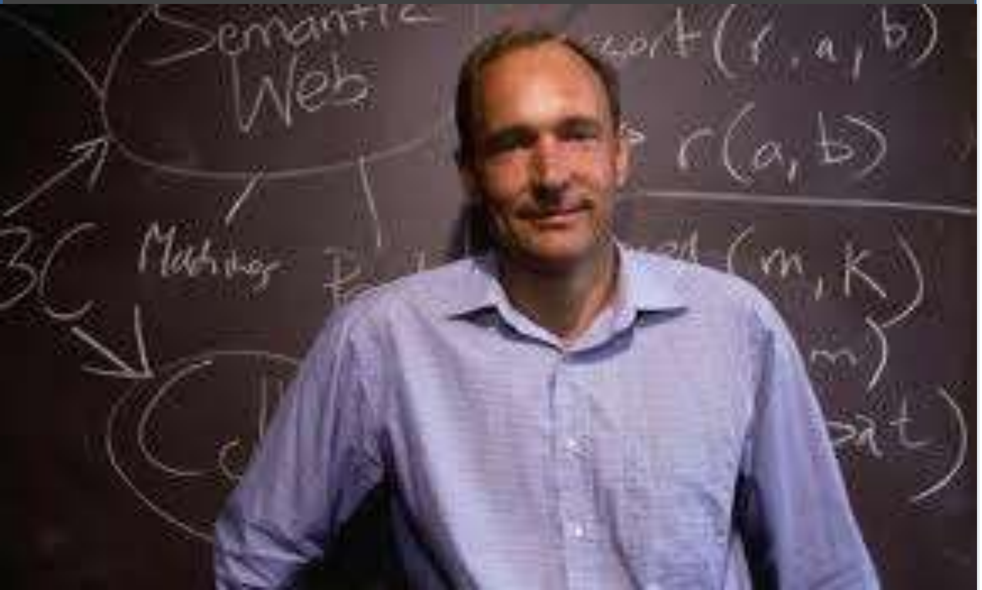
KNOWLEDGE GRAPHS



KNOWLEDGE GRAPHS

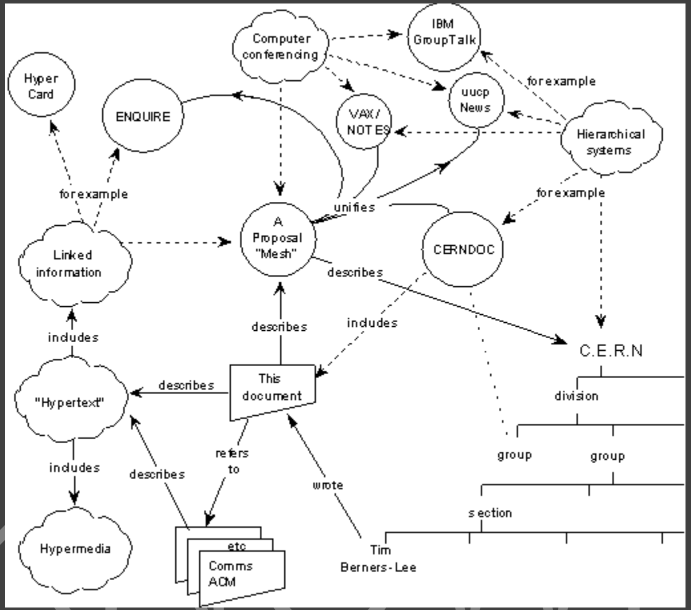
2001
Tim Berners-Lee

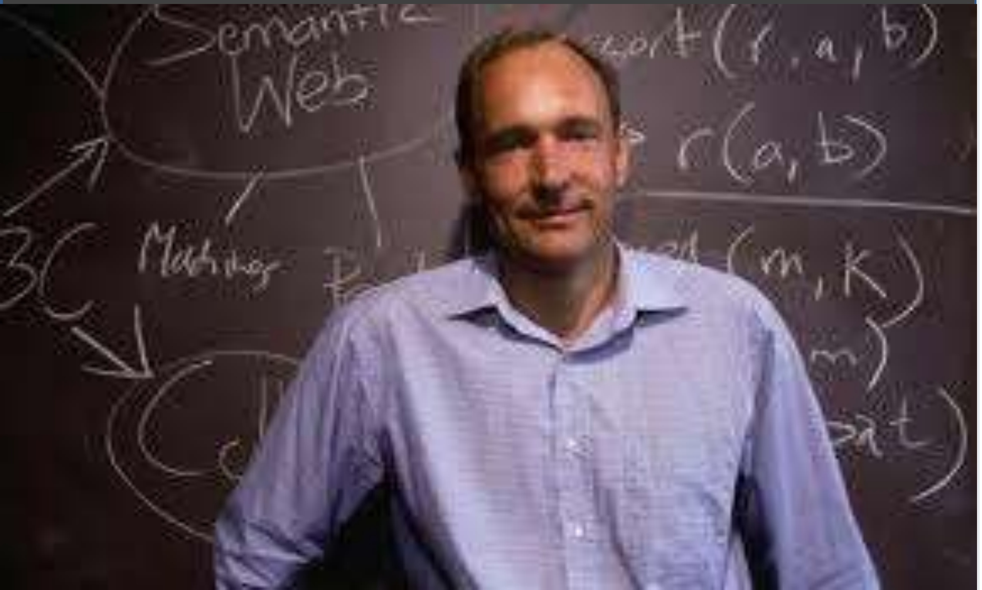




2012
Google

2001
Tim Berners-Lee





2012
Google

2019-2020

Knowledge Graphs in the news

> Google on AI and knowledge graphs in ZDNET

> Data governance 2.0 on Dataversity

> Conferences & workshops

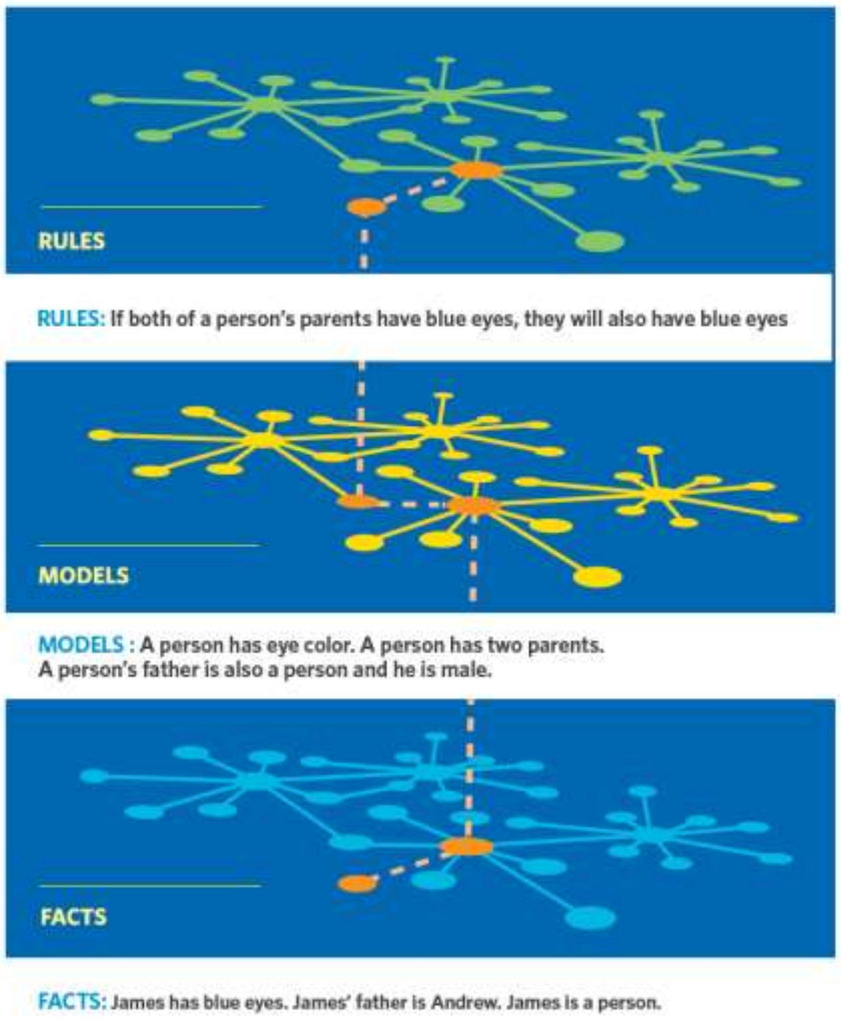
> Technology trends 2019

> Forbes article by Kurt Cagle

KNOWLEDGE GRAPHS

What are Knowledge Graphs?

- A Knowledge Graph represents a knowledge domain
- It represents knowledge as a graph
  - A network of nodes and links
  - Not tables of rows and columns
- It represents facts (data) and models (metadata) in the same way
  - Rich rules and inferencing
- It is based on open standards, from top to bottom
  - Readily connects to knowledge in private and public clouds

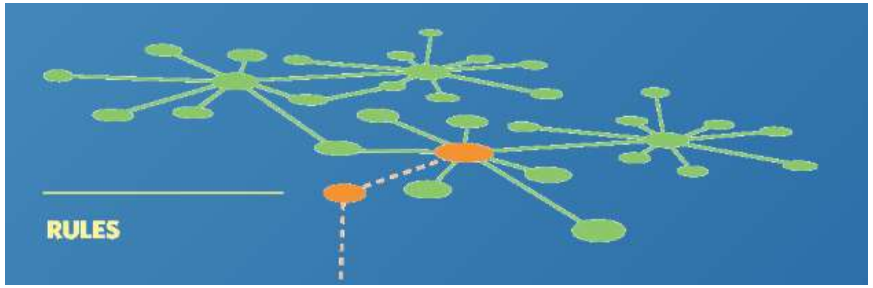


There can be different types and instances of Knowledge Graphs …

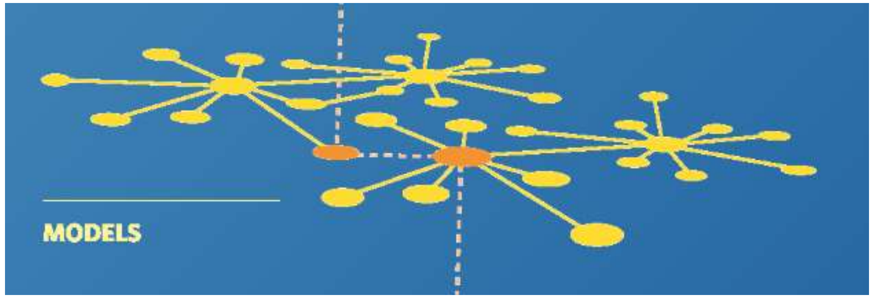
TopQuadrant™

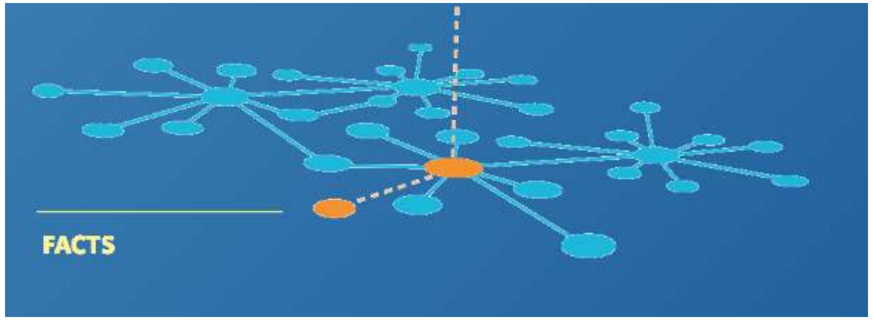
What are a Knowledge Graphs? (2)

- A knowledge graph contains information about entities in the world
  - organized as nodes connected by relationships
- The graphs capture data as well as semantics or meaning of data
- We can distinguish different types of graphs:
  - those containing basic data facts
  - and those containing higher level semantic information.



We now know that we can ask questions like “What is this person’s eye color?”





Graphs With:

Rules

If both of person’s parents have blue eyes, they will also have blue eyes

Models (data definition)

A person has eye color

A person has two parents

Person’s father is also a person and he is male

Facts (data)

James has blue eyes

James’ father is Andrew

James is a person

Today's Agenda

- A brief history knowledge graphs in 90 seconds

  … and what are they anyway?

- How do knowledge graphs support data governance?
- A concise overview of AI and ML technologies
- Knowledge graphs provide a powerful platform for both integrated data governance and enterprise AI / ML
- Some examples of how TopBraid EDG uses AI (KR + ML) for enhanced information governance

Irene Polikoff

How do Knowledge Graphs Support Data Governance?

TopQuadrant™

Data Governance Challenges

- Galaxies of data
- Diversity of perspectives
  - Business
  - Technical
  - Regulatory
- Diversity of representation

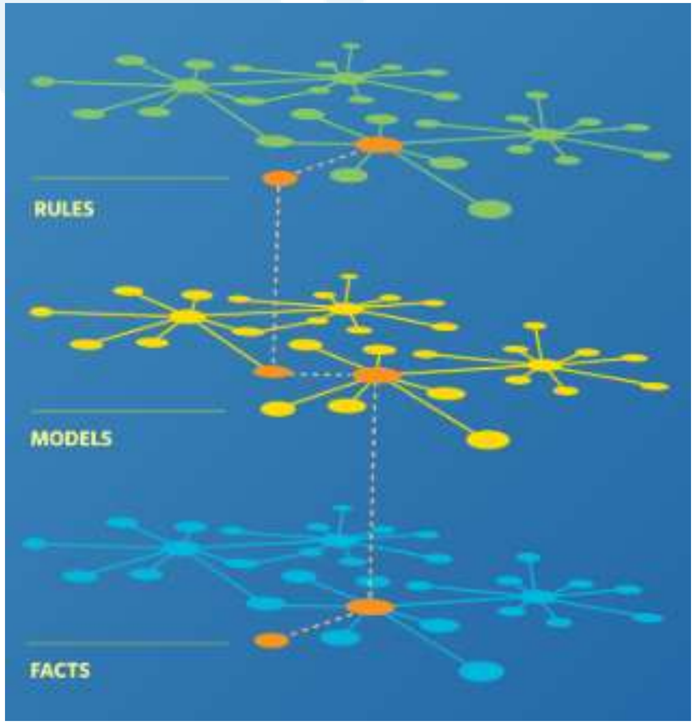
Solution

- Create a knowledge graph representing data sources
- Link to other relevant enterprise information e.g., systems, policies, infrastructure, activities
- Enrich, discover, connect
- Use to guide business decisions

Knowledge graphs overcome key challenges that data governance typically presents by:

- Providing common search for all types of stakeholders
- Connecting business terms to data elements
- Supporting regulatory compliance by tracing data lineage
- Representing regulations as knowledge graphs
- Inferring rules from data in knowledge graphs
- Connecting public and private knowledge graphs
- Creating enriched knowledge resources
- Providing knowledge graph APIs for applications
- Assessing the value of a data asset using graph-based network analysis

An Enterprise Knowledge Graph Infrastructure for Data Governance



Using Knowledge Graphs TopBraid EDG Delivers Data Governance 2.0 – Integrating Executive, Representative and Applied Governance Capabilities

TopBraid EDG supports integrated data governance across the ever growing numbers and types of data assets and governance needs

– because connections are important.

TopBraid EDG – Composition of Knowledge Graphs to Support Adaptive Data Governance

Schema *Data*

Enterprise Assets

Technical Assets

Big Data Assets

Data Assets

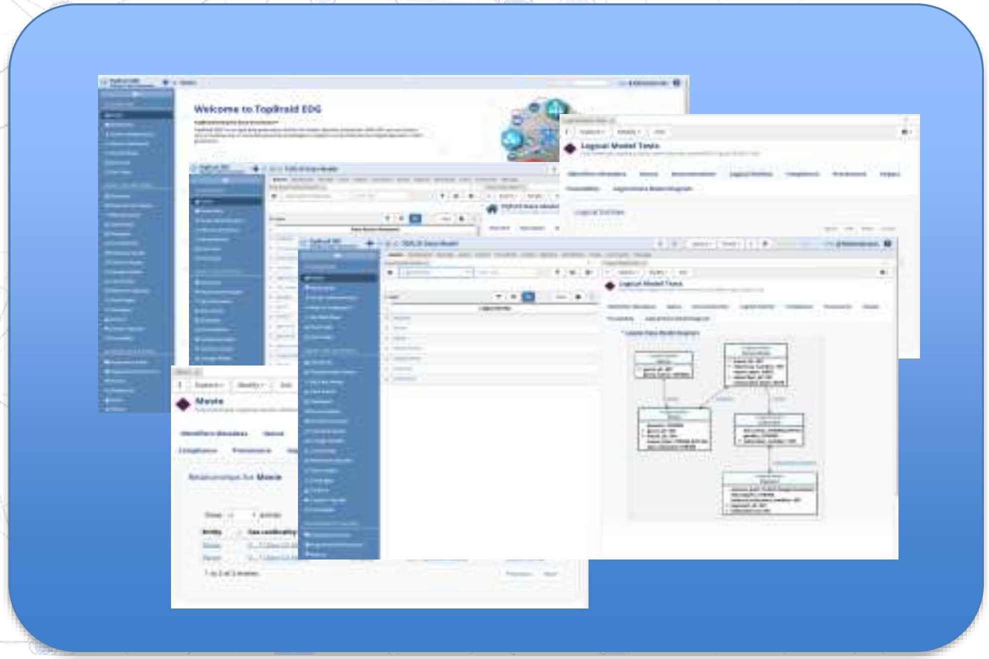
Glossaries

Terminologies

Governance Model

Compose Data and Schema

TopBraid EDG™
Enterprise Data Governance



Compose UI Behavior, Queries, Rules and Constraints

Business Logic

UX/UI

Access Control

+

Hundreds of pre-defined asset types

Today's Agenda

- A brief history knowledge graphs in 90 seconds
  … and what are they anyway?
- How do knowledge graphs support data governance?
- A concise overview of AI and ML technologies
- Knowledge graphs provide a powerful platform for both integrated data governance and enterprise AI / ML
- Some examples of how TopBraid EDG uses AI (KR + ML) for enhanced information governance

Irene Polikoff

A Concise Overview of AI and ML Technologies

- AI and ML are umbrella terms for a wide set of algorithms, technologies and approaches
- Applications using AI can be found in all industries and they span different functions
- Use of AI promises to deliver results:
  - that could not be achieved with more traditional technologies
  - or to achieve the results cheaper and faster
- Many organizations are just beginning their AI journey but are confronted with questions that are not yet well understood:

AI and ML Technologies: Concerns and Limitations?

- AI/ML have become dominant as a technology trend
- But there are strong concerns that the promise is over-hyped, e.g.:
  - heavy reliance on large data sets
  - susceptibility to machine bias
  - inability to handle abstract reasoning
  - work well only for problems where clear input can be mapped onto clear output
- Pure machine learning can be a black box approach producing results that people may not understand and can't examine

"A huge problem on the horizon is endowing AI programs with common sense. Even little kids have it, but no deep learning program does."

- Oren Etzioni, chief executive of the Allen Institute for Artificial Intelligence

→ **Enter Knowledge Graphs …**

"Why Artificial Intelligence Like AlphaZero Has Trouble With the Real World"

- Watson seemed to be endowed with the kind of clerical skills humans use on a host of real-world problems.
  - It could take a prompt in English, rummage through relevant documents at lightning speed, come up with the relevant snippets of information, and settle on a single best answer.
  - But seven years later, the real world continues to present stubborn challenges for AI.
  - A September report by the health publication *Stat* found that researching and designing personalized cancer treatments, as Watson's descendant Watson for Oncology aims to do, is proving difficult.
- **"The questions in *Jeopardy!* are easier in the sense that they don't need much common sense,"** wrote Bengio, who has collaborated with the Watson team, when asked to compare the two cases from the AI perspective. "Understanding a medical article is much harder. Again, much basic research is needed."
- As special as games are, there are still a few real-world problems they resemble.

https://www.quantamagazine.org/why-alphazeros-artificial-intelligence-has-trouble-with-the-real-world-20180221/

"Why we are in danger of overestimating AI"

- The more fundamental case against deep learning is that the technology cannot deal with many of the problems that humans will want computers to handle.
- It has no capacity for things the human mind can do easily, like abstraction or inference that make it possible for us to "understand" from very little information, or instantly apply an insight to another set of circumstances.
- "A huge problem on the horizon is endowing AI programs with common sense," says Mr Etzioni. "Even little kids have it, but no deep learning program does."

https://www.ft.com/content/4367e34e-db72-11e7-9504-59efdb70e12f

“Why Machine Learning Needs Semantics Not Just Statistics”

- The problem with likening machine learning to human learning is that when humans learn, they connect the patterns they identify to high order semantic abstractions of the underlying objects and activities. In turn, our background knowledge and experiences give us the necessary context to reason about those patterns and identify the ones most likely to represent robust actionable knowledge.
- In contrast, machines blindly search for the strongest signals in a pile of data. Lacking a cushion of background knowledge or life experiences to understand the meaning of those signals, deep learning algorithms cannot distinguish between meaningful and spurious indicators. Instead, they merely blindly encode the world according to statistics, rather than semantics.

https://www.forbes.com/sites/kalevleetaru/2019/01/15/why-machine-learning-needs-semantics-not-just-statistics/#74c2ee2f77b5

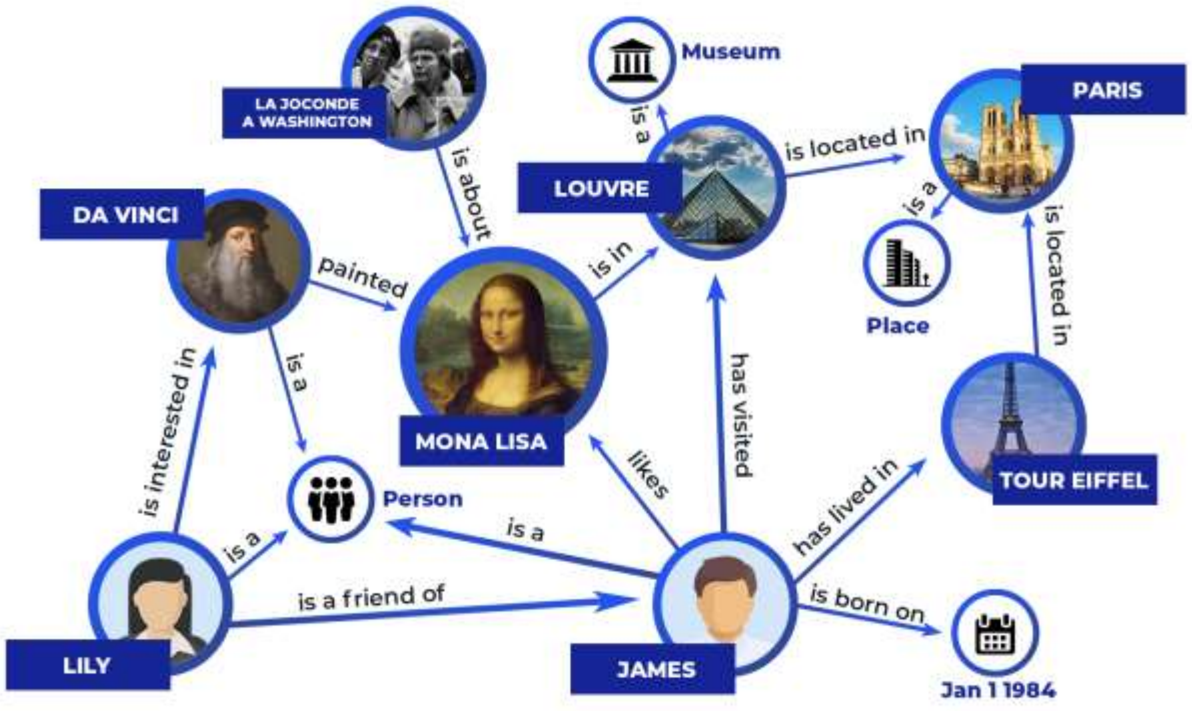
Knowledge Graphs Excel at Knowledge Representation

- Knowledge representation and reasoning (KR) is the field of AI dedicated to representing information about the world in a form that a computer system can utilize

- **Knowledge graphs** are part of the KR branch of AI
  - they can capture data *as well as semantics or the meaning of data*
  - they enable computers to reason based on the full available contextual and conceptual information

ML Technologies + Knowledge Graphs? Good Choice!

- In addition to supporting data governance, it is becoming widely accepted that **knowledge graphs** are also excellent at guiding and focusing ML



- Leading companies who are building **knowledge graphs** include Google, Apple, Amazon, Airbnb, Bloomberg, Facebook, LinkedIn, Thomson Reuters – and these are just a few.

Knowledge Graphs Form a Trusted Foundation for AI and Machine Learning by Providing Meaning to Information

Knowledge graphs which can provide:

- a supervisory capacity to direct productive application of ML
- an automated means for maintaining and improving data quality at any step in the data lifecycle
- well-understood, curated training data sets
- an unmatched way to make sense of the results
- Integration of structured and unstructured data sources as input

Today’s Agenda

- A brief history knowledge graphs in 90 seconds
  … and what are they anyway?
- How do knowledge graphs support data governance?
- A concise overview of AI and ML technologies
- Knowledge graphs provide a powerful platform for both integrated data governance and enterprise AI / ML
- Some examples of how TopBraid EDG uses AI (KR + ML) for enhanced information governance

Irene Polikoff

Auto-suggesting Mapping of Data Elements to Business Terms

Problem

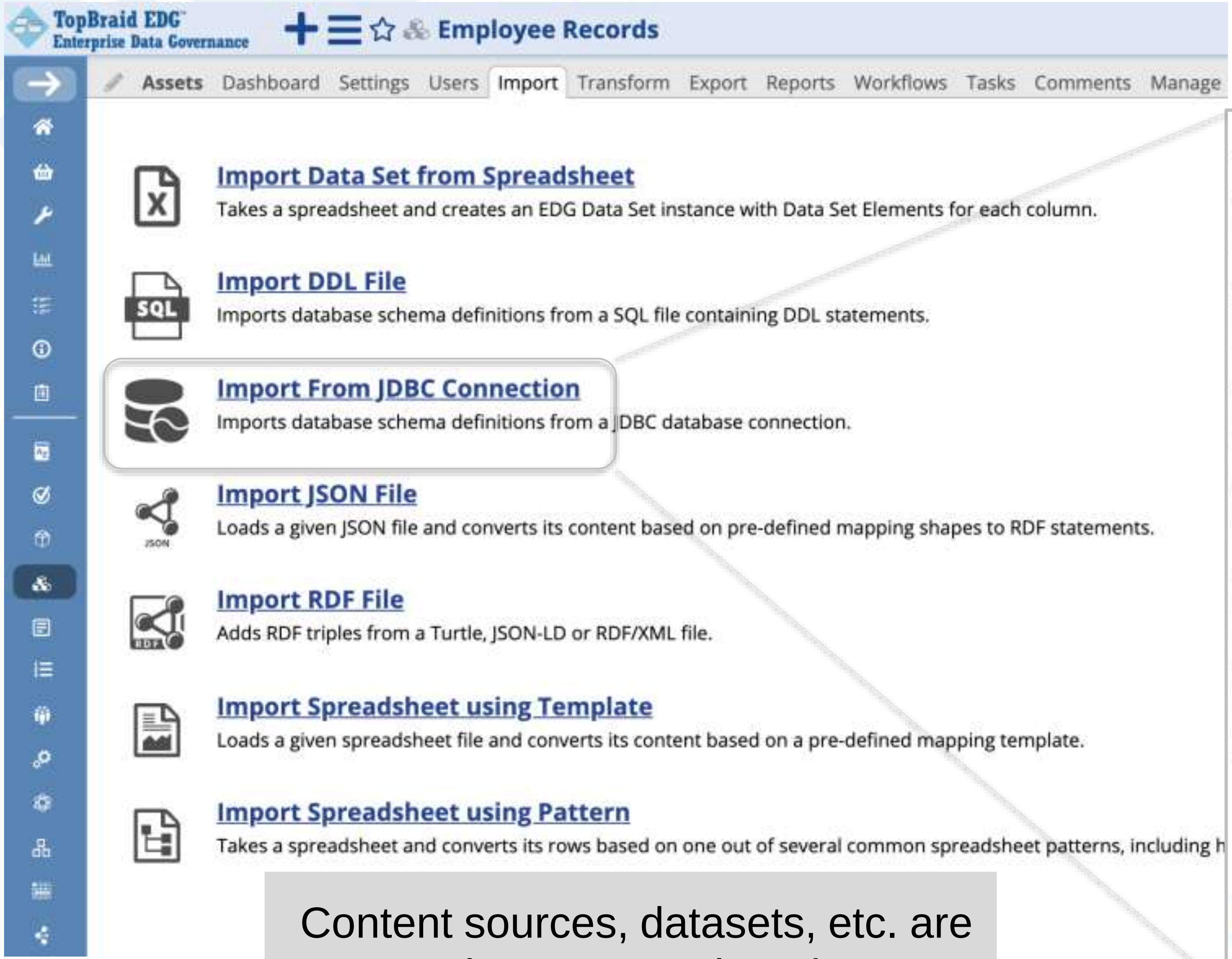
- Business stewards and subject matter experts work on agreed business glossaries and data dictionaries, establishing business metadata
- Data sources can be examined to capture technical metadata
- Manually connecting technical and business metadata requires a lot of effort and time; enterprises can easily have millions of data elements across all key sources

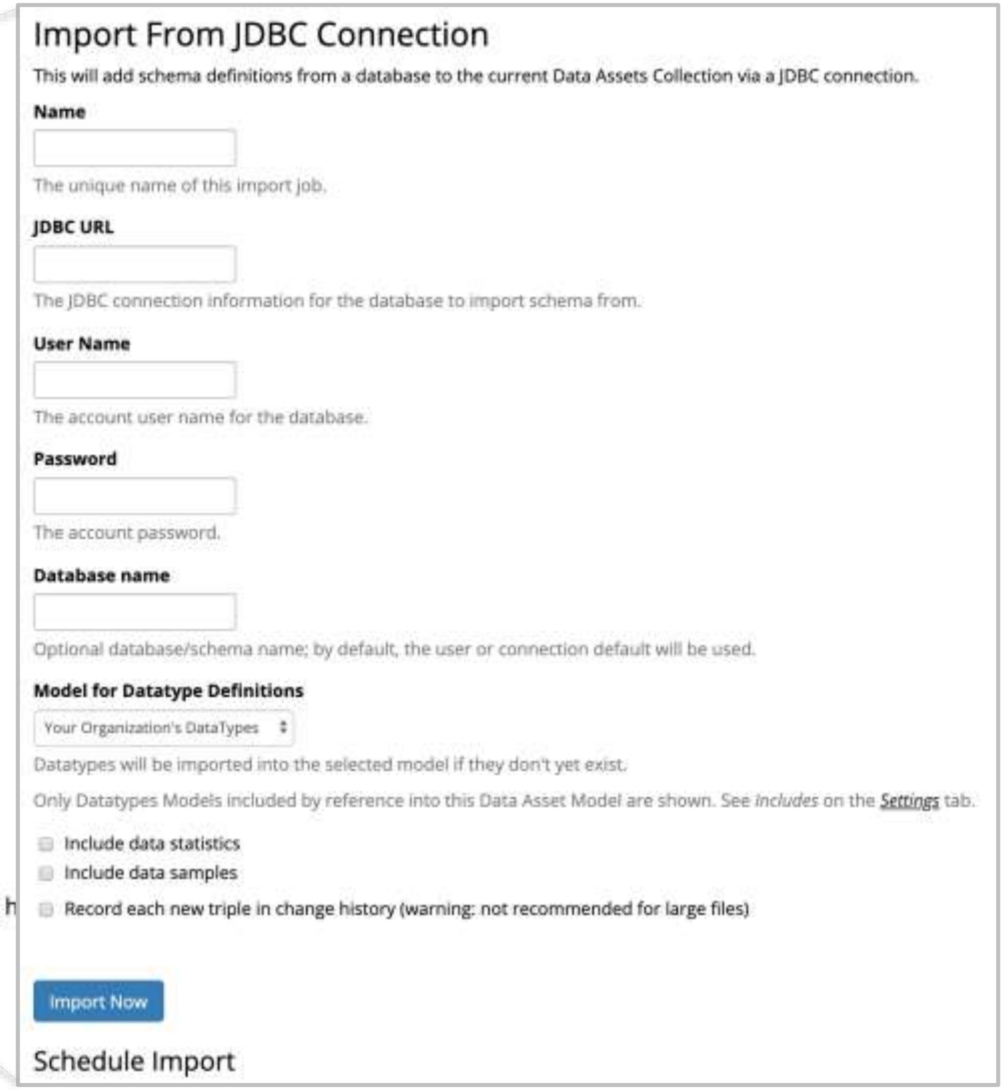
Solution

- TopBraid EDG, as an enterprise knowledge graph for data governance describes the shape of data and rules associated with business terms
- It then uses the rules to infer connections between technical representation of data sources and their business meaning

TopQuadrant™

Automated Data Cataloging





Content sources, datasets, etc. are also auto-cataloged

TopQuadrant™

Define Employee ID as a Term

Employee ID ×

⋮ Explore ▾ Modify ▾ Edit

Employee ID

URI http://example.org/glossaries/new#First_Name

Identifiers Metadata

type: PII Term

label: Employee ID

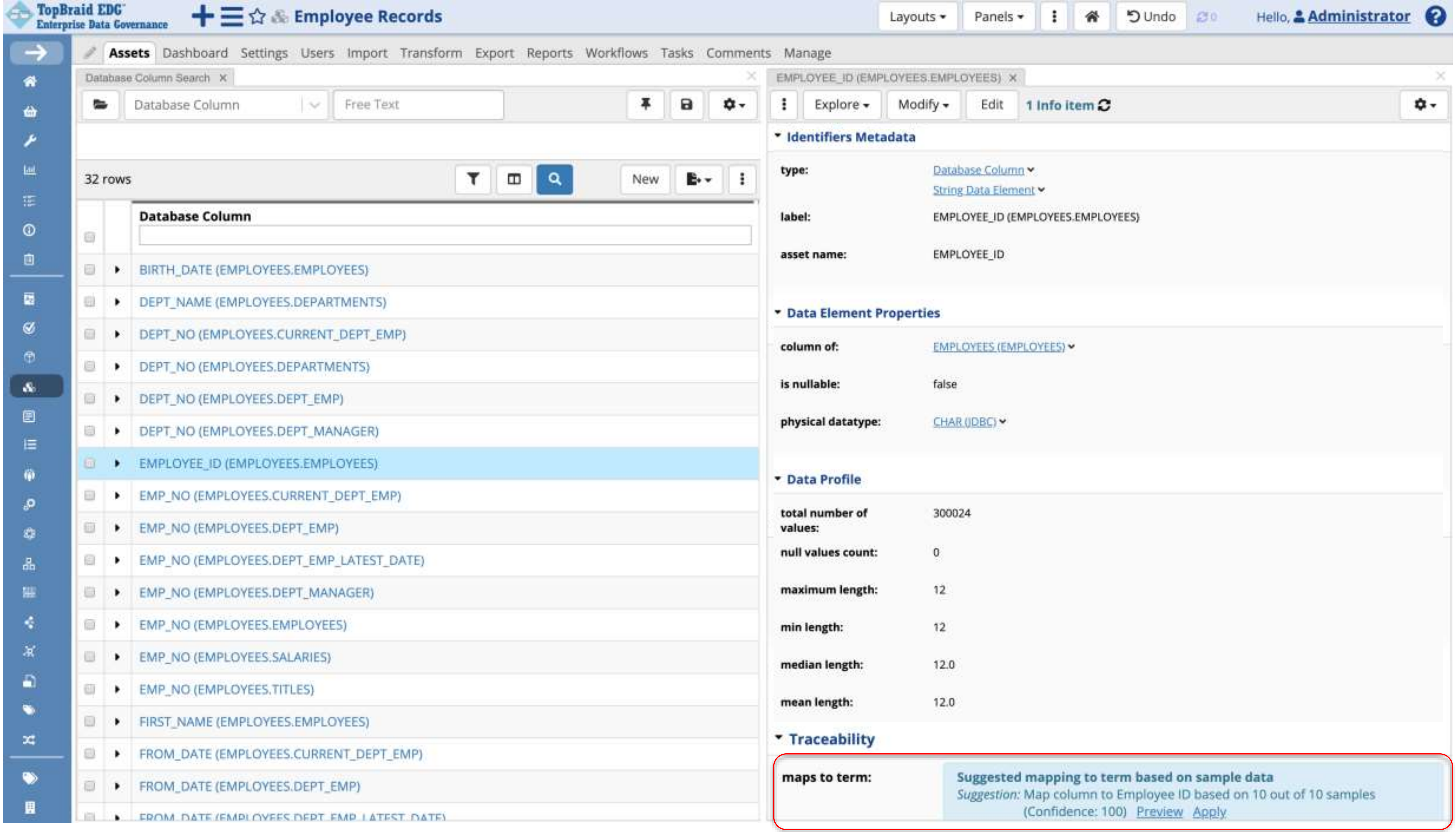
definition: Employee ID is defined as the single unique code of reference used by the enterprise to identify an individual person having permanent full time employ by the enterprise

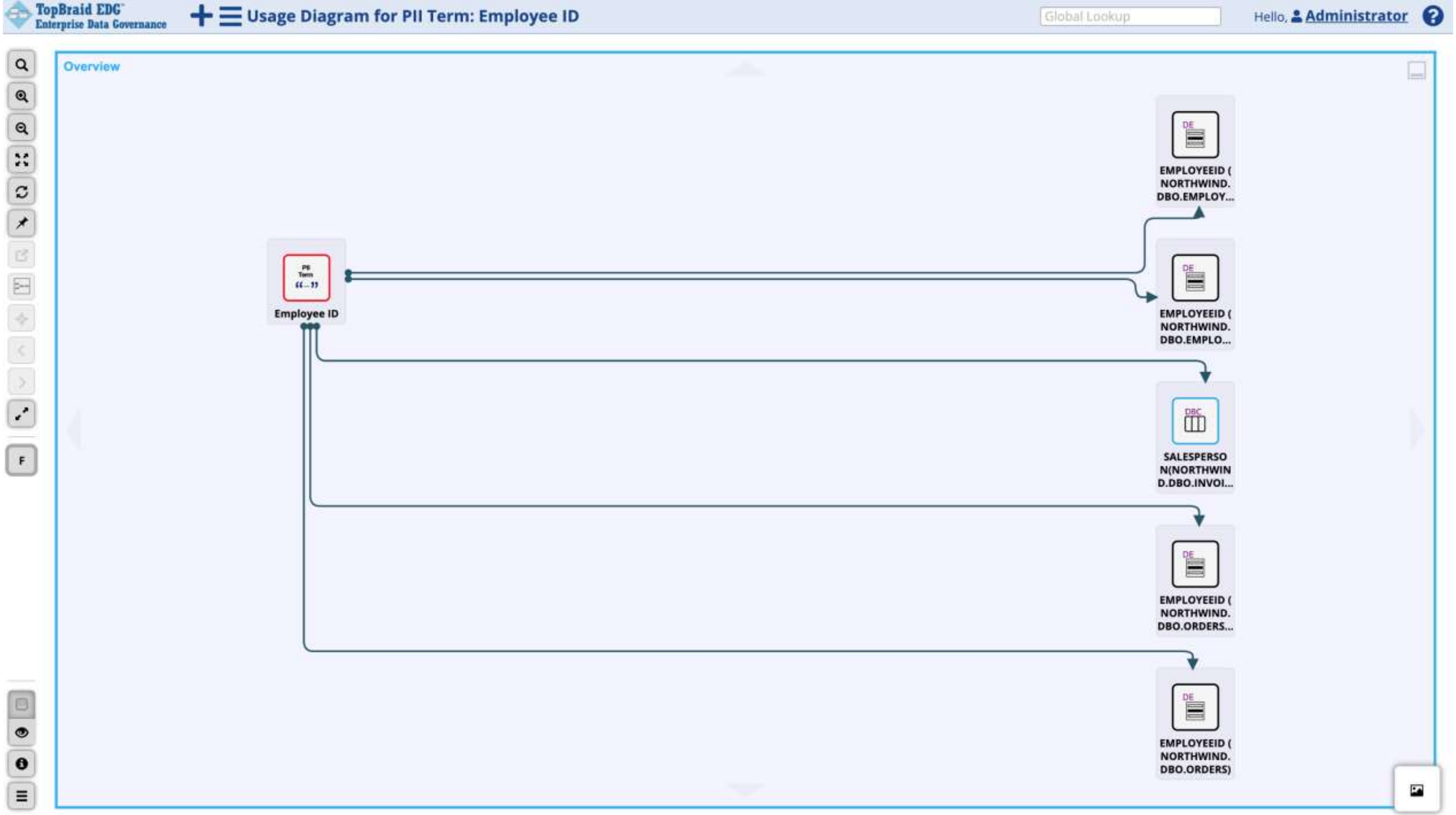
Employee ID is NOT used as identification for part-time, contract or seasonal employees. It is also not used to identify employees of partner, client or vendor companies.

Glossary Term Metadata

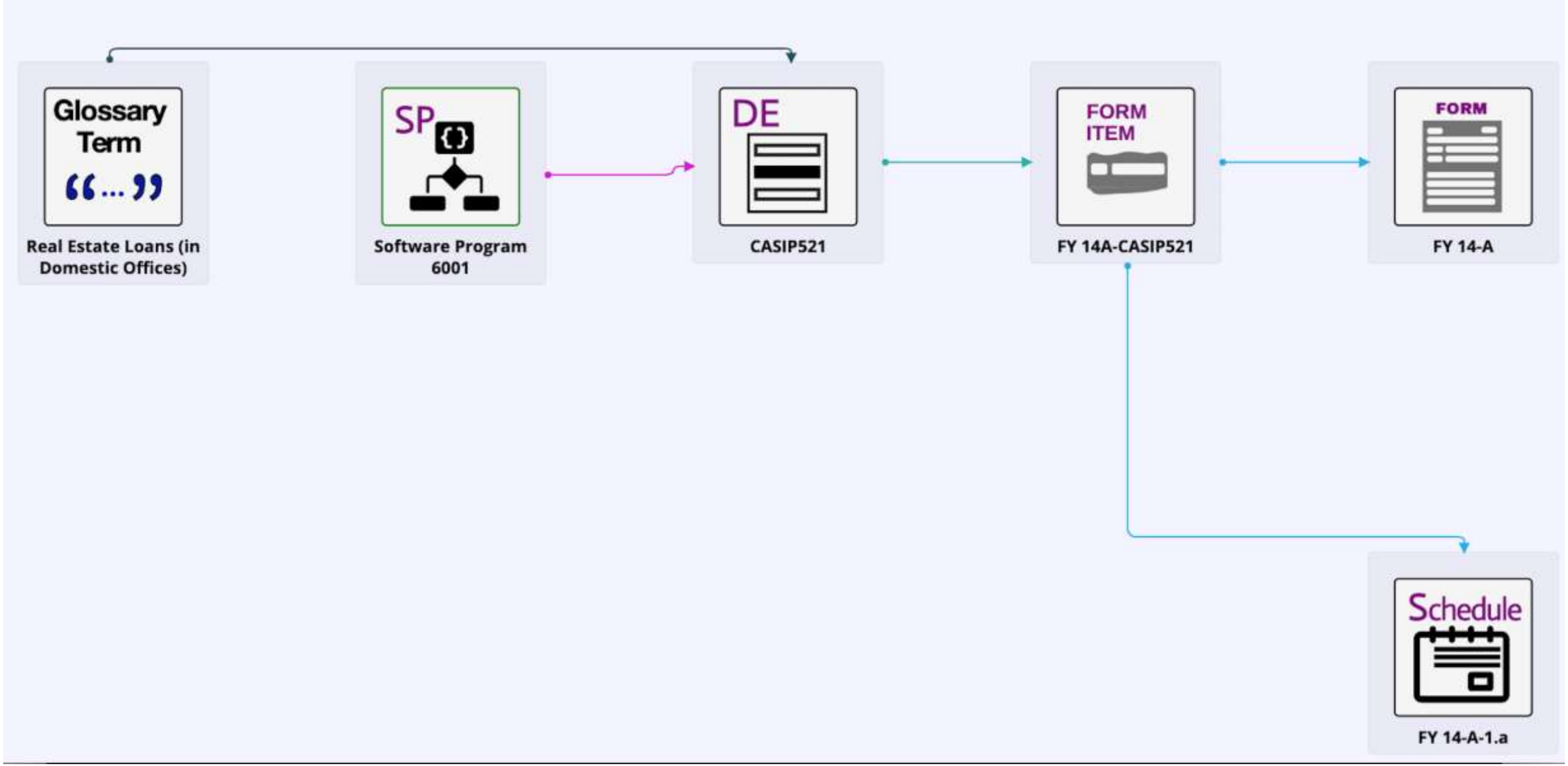
data value rule: Employee ID rule

label:	Employee ID rule
datatypes:	string
min length:	12
max length:	12
regex pattern:	^[ACGT][678][34]-?\d{4}-?\d{3}$


TopBraid EDG
Enterprise Data Governance
Employee Records
Layouts
Panels
Undo
Hello, Administrator
Assets
Dashboard
Settings
Users
Import
Transform
Export
Reports
Workflows
Tasks
Comments
Manage
Database Column Search
Database Column
Free Text
32 rows
New
Database Column
BIRTH_DATE (EMPLOYEES.EMPLOYEES)
DEPT_NAME (EMPLOYEES.DEPARTMENTS)
DEPT_NO (EMPLOYEES.CURRENT_DEPT_EMP)
DEPT_NO (EMPLOYEES.DEPARTMENTS)
DEPT_NO (EMPLOYEES.DEPT_EMP)
DEPT_NO (EMPLOYEES.DEPT_MANAGER)
EMPLOYEE_ID (EMPLOYEES.EMPLOYEES)
EMP_NO (EMPLOYEES.CURRENT_DEPT_EMP)
EMP_NO (EMPLOYEES.DEPT_EMP)
EMP_NO (EMPLOYEES.DEPT_EMP_LATEST_DATE)
EMP_NO (EMPLOYEES.DEPT_MANAGER)
EMP_NO (EMPLOYEES.EMPLOYEES)
EMP_NO (EMPLOYEES.SALARIES)
EMP_NO (EMPLOYEES.TITLES)
FIRST_NAME (EMPLOYEES.EMPLOYEES)
FROM_DATE (EMPLOYEES.CURRENT_DEPT_EMP)
FROM_DATE (EMPLOYEES.DEPT_EMP)
EMPLOYEE_ID (EMPLOYEES.EMPLOYEES)
Explore
Modify
Edit
1 Info item
Identifiers Metadata
type:
Database Column
String Data Element
label:
EMPLOYEE_ID (EMPLOYEES.EMPLOYEES)
asset name:
EMPLOYEE_ID
Data Element Properties
column of:
EMPLOYEES (EMPLOYEES)
is nullable:
false
physical datatype:
CHAR (JDBC)
Data Profile
total number of values:
300024
null values count:
0
maximum length:
12
min length:
12
median length:
12.0
mean length:
12.0
Traceability
maps to term:
Suggested mapping to term based on sample data
Suggestion: Map column to Employee ID based on 10 out of 10 samples
(Confidence: 100)
Preview
Apply


TopBraid EDG
Enterprise Data Governance
Usage Diagram for PII Term: Employee ID
Global Lookup
Hello, Administrator
Overview
Employee ID
EMPLOYEEID (NORTHWIND. DBO.EMPLOY...
EMPLOYEEID (NORTHWIND. DBO.EMPLO...
SALESPERSO N(NORTHWIN D.DBO.INVOI...
EMPLOYEEID (NORTHWIND. DBO.ORDERS...
EMPLOYEEID (NORTHWIND. DBO.ORDERS)

Connected View of Information



Extracting Implicit Semantics in Data to Explicit Semantics in a Knowledge Graph

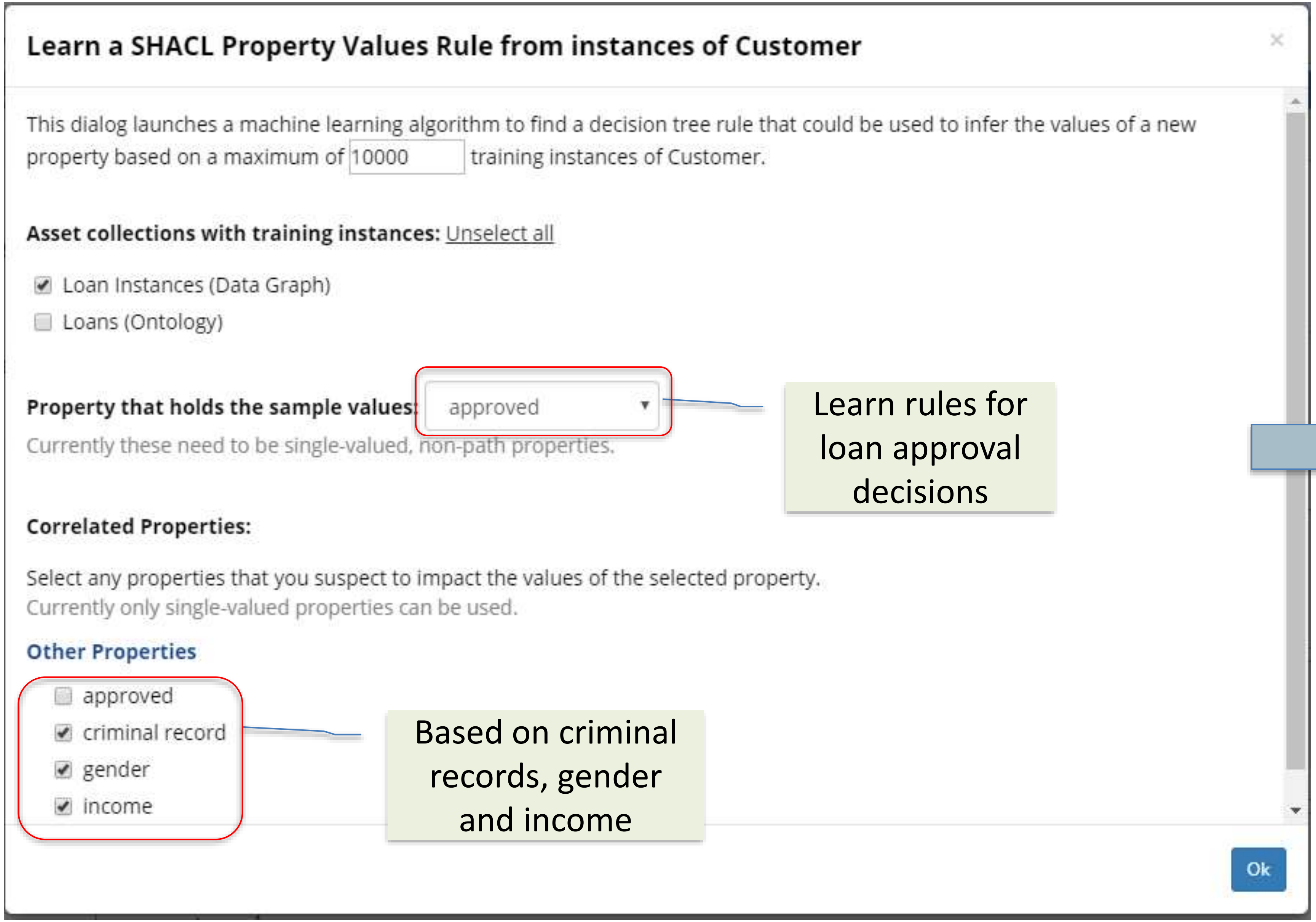
Problem

- Business rules are encoded in different places within an enterprise ecosystem
- Error prone and hard to change
- Difficult to check by business experts

Solution

- TopBraid EDG can catalog your data sources, pulling schema information from a source into a knowledge graph
- Then use machine learning to extract rich semantics about relationships between data values (business rules) directly from sample datasets
- Business users can participate in and control this knowledge acquisition process

Extracting Implicit Semantics in Data to Explicit Semantics in a Knowledge Graph



Rules

- Reject if criminal record
- Reject if income less than $65K
- Approve if no criminal record and income greater than $65K

Auto-mapping Disconnected Vocabularies

Problem

- Individual datasets are specified in isolation using different local terminologies, depending on the commissioners of the data collection and the questions they require answering at that point in time.
- As a result, linkage of data is time consuming, require providers to specifically collate, check and submit against a wide range of specification and collection methods.
- Inconsistency of conclusions from data
- Long lead time to data

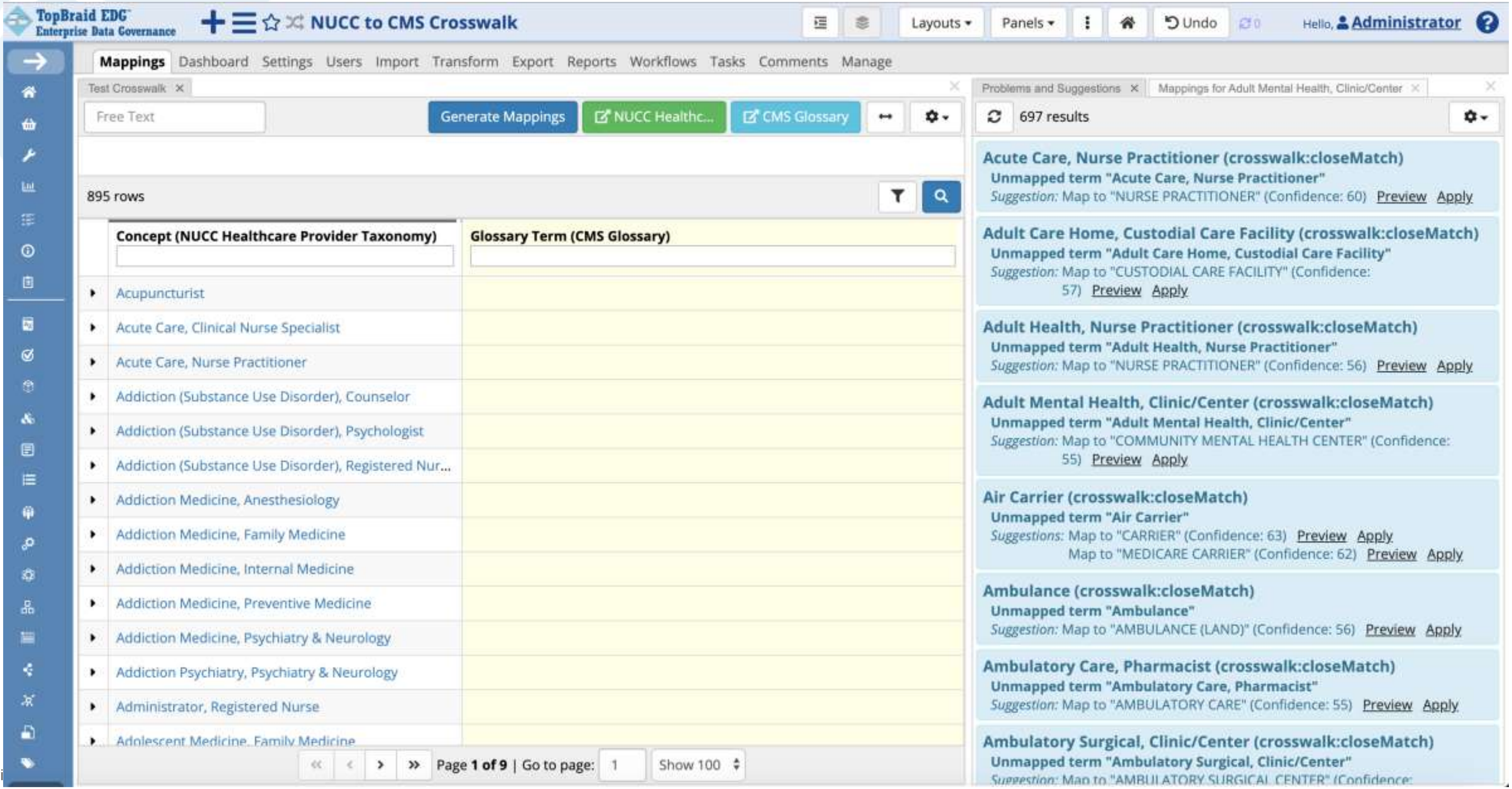
Solution

- TopBraid EDG, can manage different terminologies and automatically infer connections between them, enabling:
  - Safe, appropriate and consented linkage of de-identified data from multiple sources
  - Facilitation of data reuse
  - Efficient mechanisms to expand the coverage of data collected as required by all consumers

In the UK NHS extra annual cost to data providers of producing and submitting just Admitted Patient Care and Outpatient Care data collections and audits is **in excess of £654 million**.

TopQuadrant™

Auto-mapping of Vocabularies In EDG



Extracting Meaning from Unstructured Data

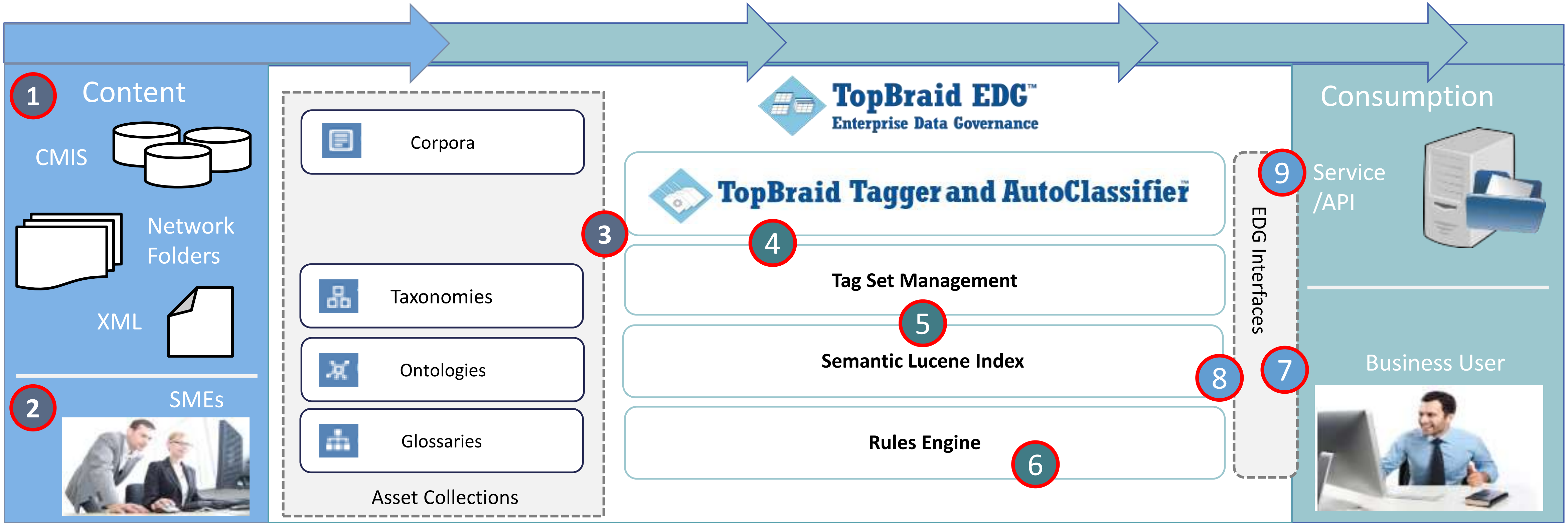
Problem

- Much enterprise information is in documents
- Understanding enterprise information requires closing the gap between structured and unstructured information … But typically, these sources are governed separately
- Further, most machine learning algorithms work well either with text or with structured data

Solution

- Knowledge graphs can bridge this gap since they can contain relationships between all information irrespective of the format including relational data, XML, JSON, CSV, and text
- In EDG, machine learning guided by controlled vocabularies can associate documents with key topics; rules can further process this information
- As a result, enterprises can now have a unifying view across all metadata

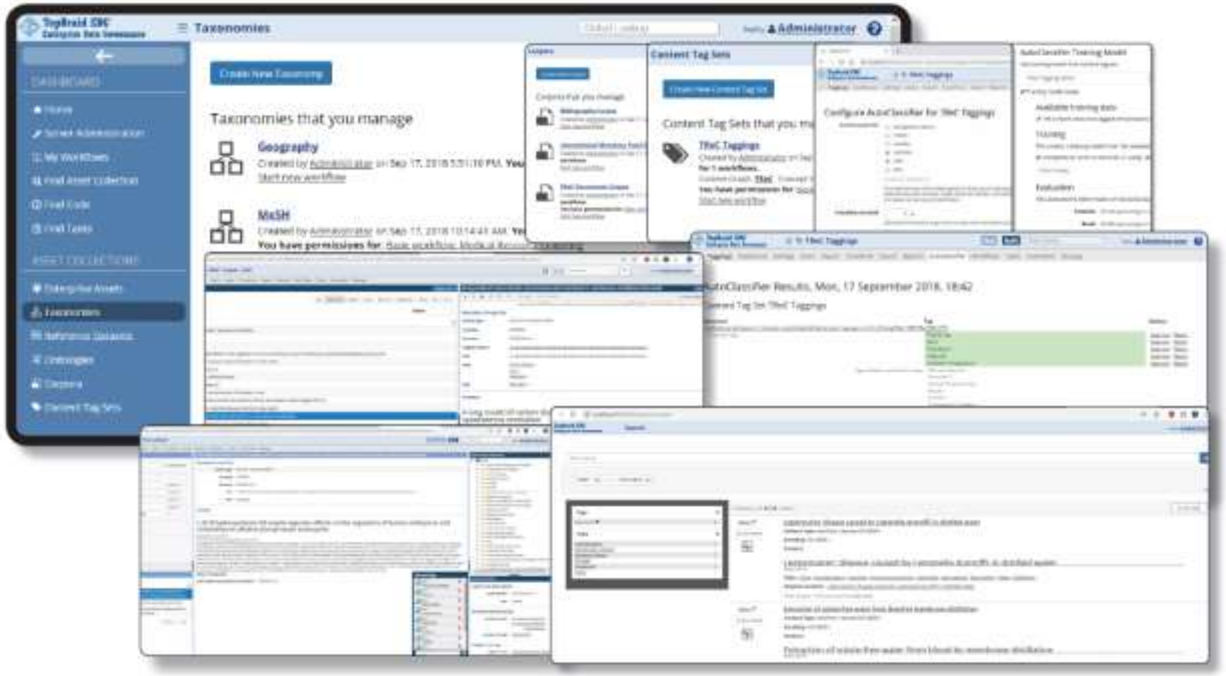
TopBraid EDG: Search Enrichment

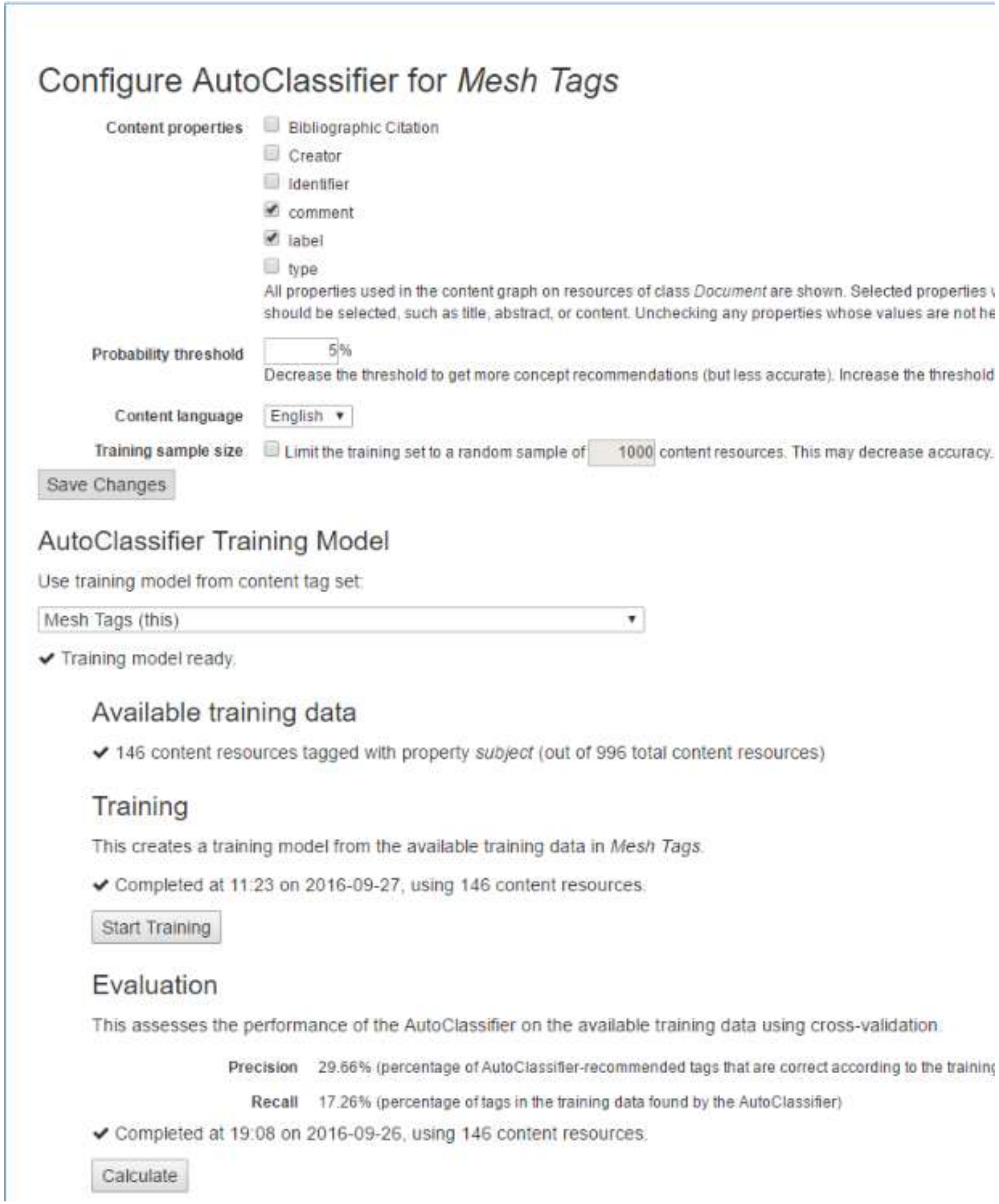


1. Content is retrieved/received over many connectors.
2. SMEs of all sorts develop/integrate vocabularies. As needed, they also review and curate auto-classification results providing input that improves results accuracy over time.
3. Vocabularies and curation feedback are leveraged by AutoClassifier to extract information from unstructured content inside Asset Collections.
4. AutoClassifier uses machine learning to populate maintainable Tag Sets (relationships between 'Content' and 'Concepts').
5. Text, Tags, and other relevant characteristics of Assets are indexed for optimized search and discovery.
6. Semantic models and business rules can be used to further post process and extend results
7. Search user submits a request and the interface sends the request on to the query engine.
8. The query is interpreted and semantic search is executed. Results include relevant metadata for customized user experience.
9. Similar to the search user, applications submit requests to an EDG interface. Determined by the interface (Web Service, Saved Search, SPARQL endpoint), results are prepared and sent in appropriate format.

Extracting Meaning from Unstructured Data

TopBraid Tagger's auto-classification capability delivers tagging results that are as consistent and accurate as those created by content indexing experts using typical domain taxonomies.





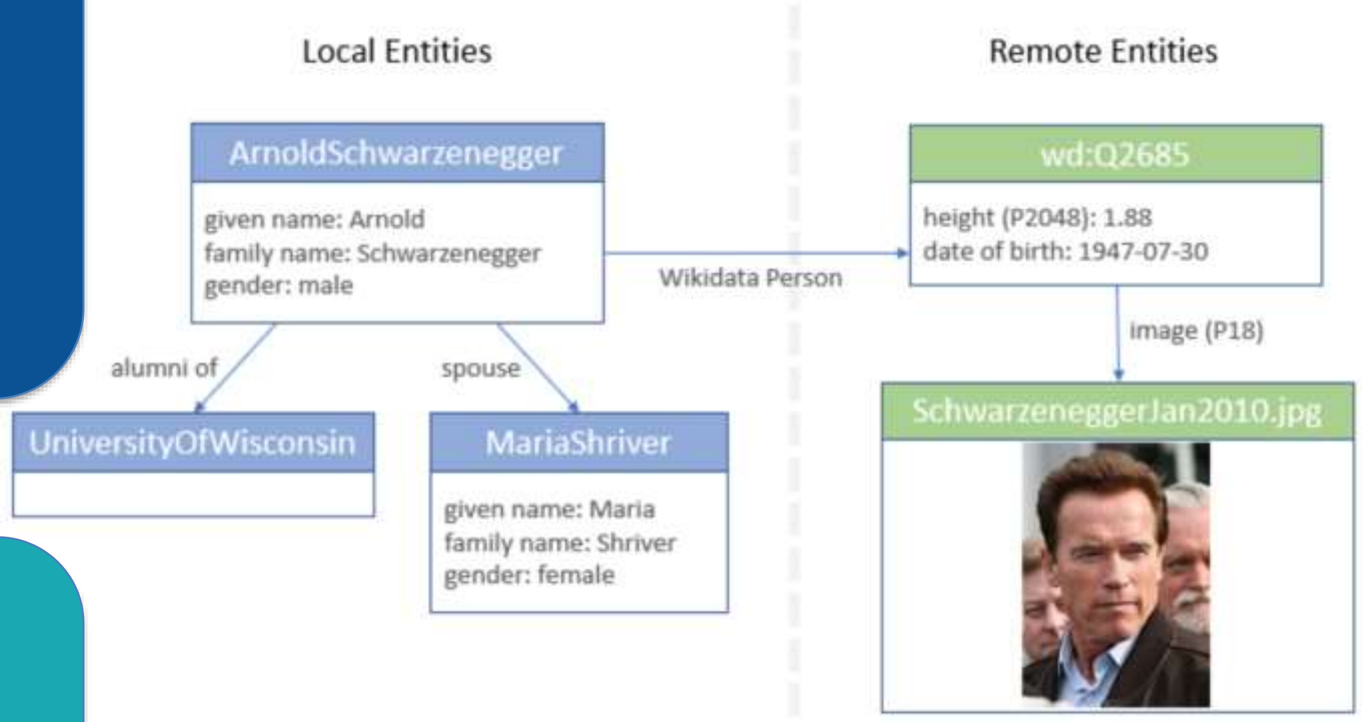
Connecting Public and Private Knowledge Graphs

Opportunity

- Knowledge sharing is an important driving force behind progress
- Increasingly, organizations are publishing knowledge graphs covering scientific, financial and general information. Examples include:
  - Wikidata - https://www.wikidata.org/
  - Blue Brain Nexus - https://tinyurl.com/y9ncet4c
  - Thomson Reuters/REFINITIV - https://tinyurl.com/y2e39pn5

Solution

- Leverage external knowledge graphs by reaching out to them as the source of reference data and other important information
- You no longer have to maintain all information yourself
- Instead, you can directly take advantage of authoritative, curated and maintained information assembled by people you trust – without having to write and run specialized importers that transform data.



TopQuadrant™

Today's Agenda

- A brief history knowledge graphs in 90 seconds
  … and what are they anyway?
- How do knowledge graphs support data governance?
- A concise overview of AI and ML technologies
- Knowledge graphs provide a powerful platform for both integrated data governance and enterprise AI / ML
- Some examples of how TopBraid EDG uses AI (KR + ML) for enhanced information governance

* *And one more thing – Food for Thought:*
What about the governance of the use of AI / ML itself?


Irene Polikoff

TopQuadrant™

What about the governance of the use of AI / ML itself?

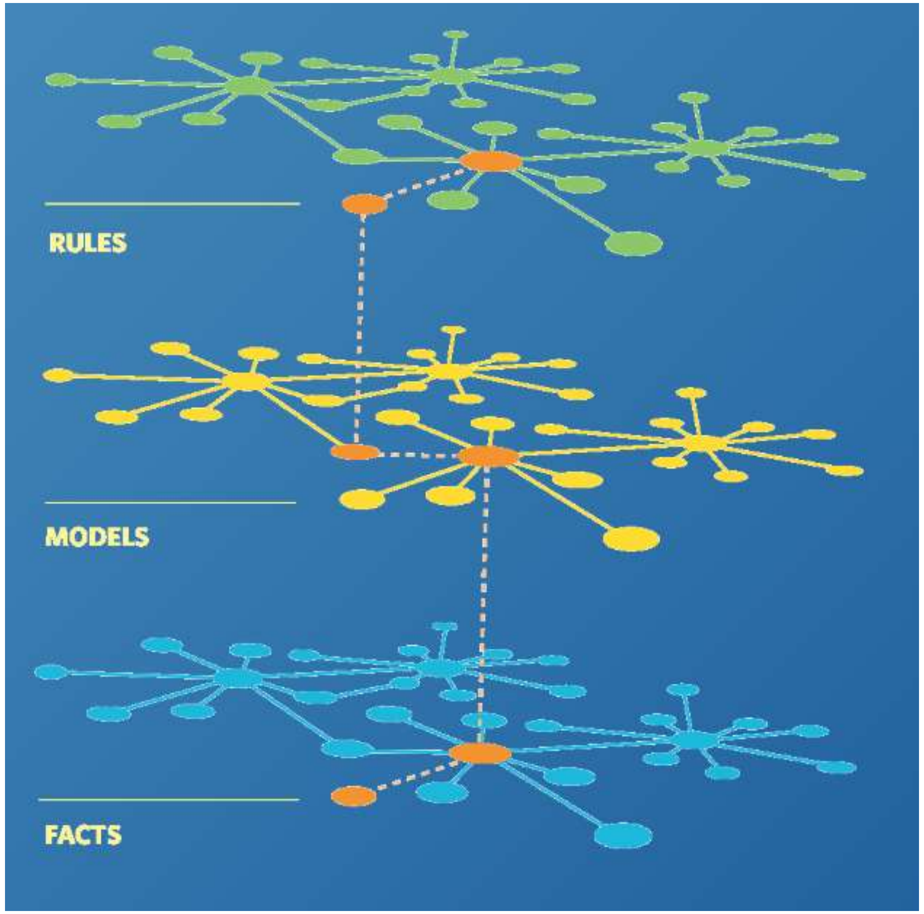
- As companies integrate machine learning into their products and systems, there are important foundational technologies that come into play
- Not surprisingly – current machine learning and AI technologies require large amounts of data—specifically, labeled data for training models
  - A common theme in conversations with data engineers, data scientists, and AI researchers is the need for solutions that can help track data lineage and provenance, **i.e., data governance solutions**
  - Having reliable metadata for datasets, such as what job created the dataset, where data came from, etc., is crucially important for anyone using a dataset.
  - This includes not only people responsible for audit, compliance, reliability, debugging and other activities, but also AI algorithms
  - As an enterprise starts to employ AI/ML on an ongoing basis, **the governance of the use of AI itself will need to be addressed.**

What Governance is Needed to Support the Use of AI in the enterprise?

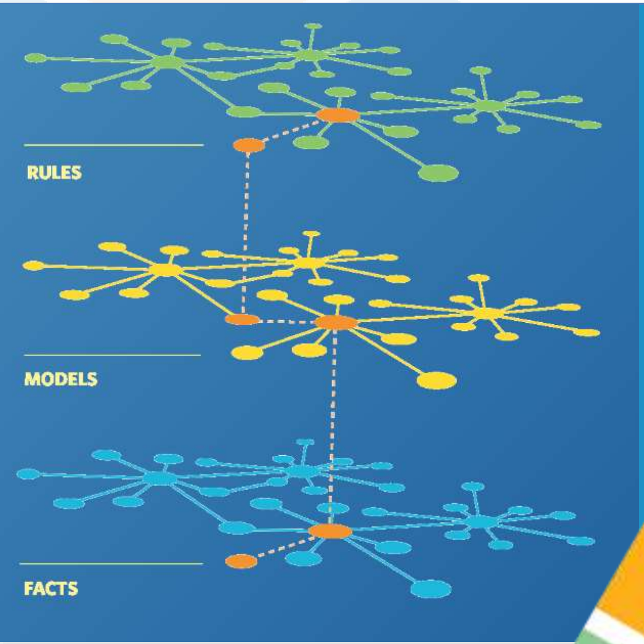
- Management of training datasets
- Help in reliably and cost effectively combining data across heterogeneous data sources
  - to provide data objects as training data sets
  - including those composed of information gathered from structured data as well as text
- Capturing what AI algorithms are being used and for what purposes
- Metrics for understanding and evaluating the usefulness of results delivered by different AI algorithms
- …?

Conclusion: Benefits of a Knowledge Graph based Platform for Data Governance 2.0

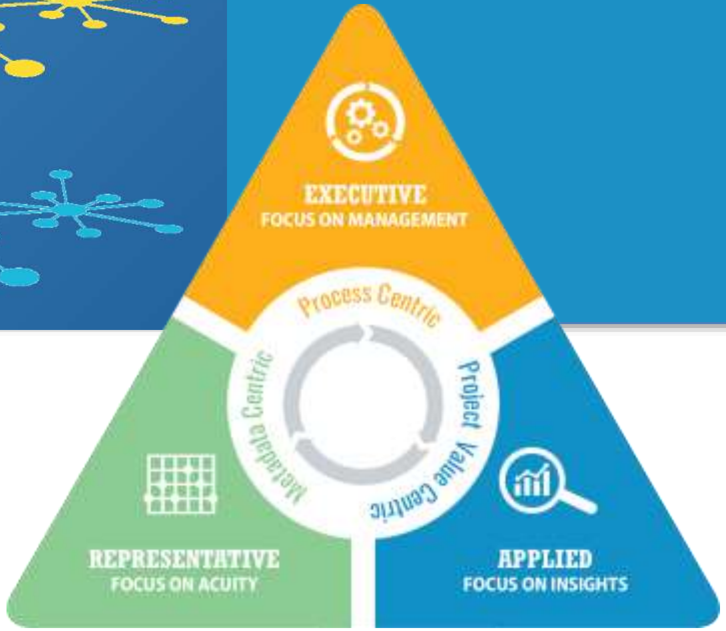
- Flexibility and extensibility based on standards
- Integration of reasoning and machine learning
- Enabling people (UI) and software (APIs/web services) to view, follow and query
- Bridging of data and metadata “silos” to provide seamless data governance
- Delivery of Knowledge-driven data governance



As an enterprise knowledge graph infrastructure, TopBraid EDG supports Data Governance 2.0 and applications of AI / ML



Thank You !



… Questions?

TopQuadrant™

To Learn More about TopBraid EDG and Knowledge Graphs:

EDG Product Info:

- TopBraid Enterprise Data Governance (TopBraid EDG)

(https://www.topquadrant.com/products/topbraid-enterprise-data-governance/)

Contact us: at info@topquadrant.com to:

- Discuss vocabulary management solutions (glossaries, taxonomies, ontologies)
- Request a more targeted demo of TopBraid EDG
- Ask for a free EDG evaluation account

TopQuadrant™

More Resources ...

More Webinar Recordings, Slides, Q&A:

- https://www.topquadrant.com/knowledge-assets/topquadrant-webinars/

Short Videos:

- TopBraid EDG "Quick Grok" Videos
  https://www.topquadrant.com/knowledge-assets/videos/

- TopBraid EDG Animated Video
  https://www.topquadrant.com/project/edg_agile_modular/

Blog:

- https://www.topquadrant.com/the-semantic-ecosystems-journal/

White Papers

- https://www.topquadrant.com/knowledge-assets/whitepapers/