MarkLogic®

MarkLogic Data Hub

INTELLIGENT APPROACH + SMART TECHNOLOGY

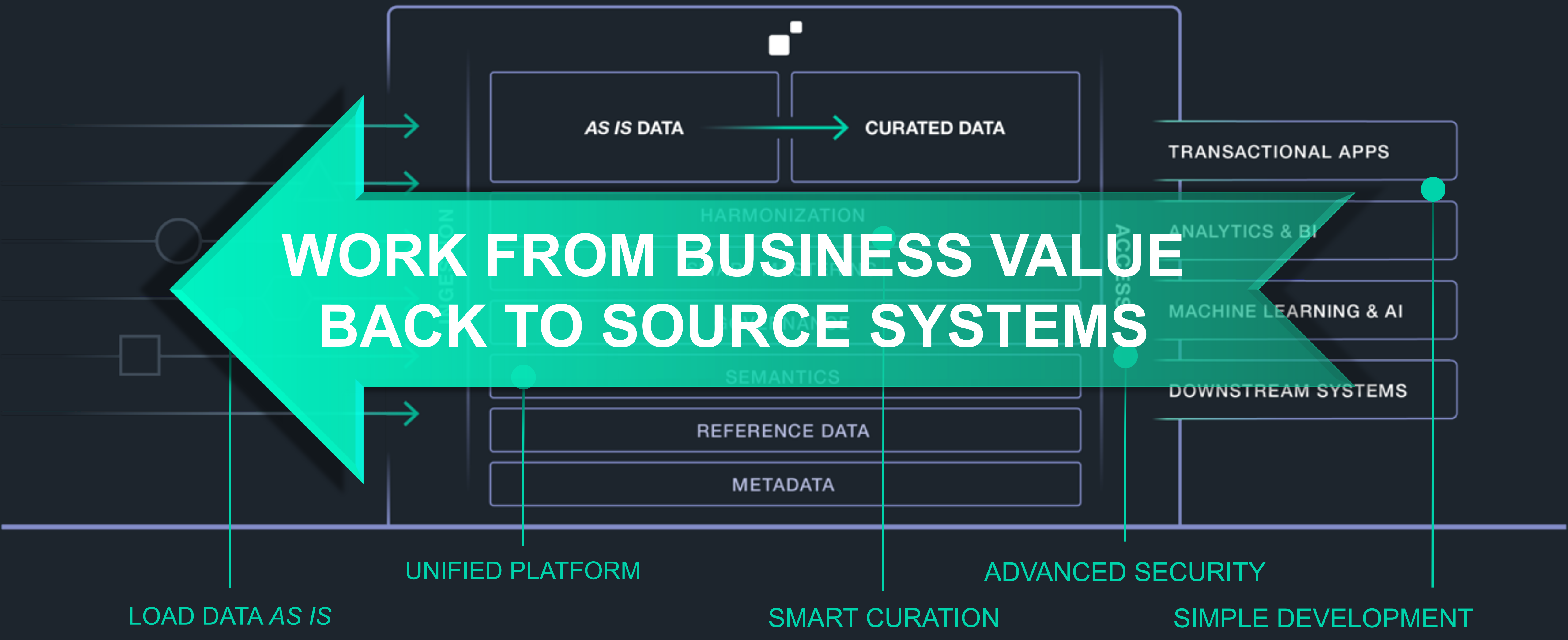
Michel de Ru, Senior Principal Solution Engineer

“Begin with the end in mind.”
Stephen Covey

resembles

Data Services First approach

MarkLogic Data Hub Platform



WORK FROM BUSINESS VALUE BACK TO SOURCE SYSTEMS

INGESTION

AS IS DATA

CURATED DATA

HARMONIZATION

MASTERING

GOVERNANCE

SEMANTICS

REFERENCE DATA

METADATA

ACCESS

TRANSACTIONAL APPS

ANALYTICS & BI

MACHINE LEARNING & AI

DOWNSTREAM SYSTEMS

LOAD DATA *AS IS*

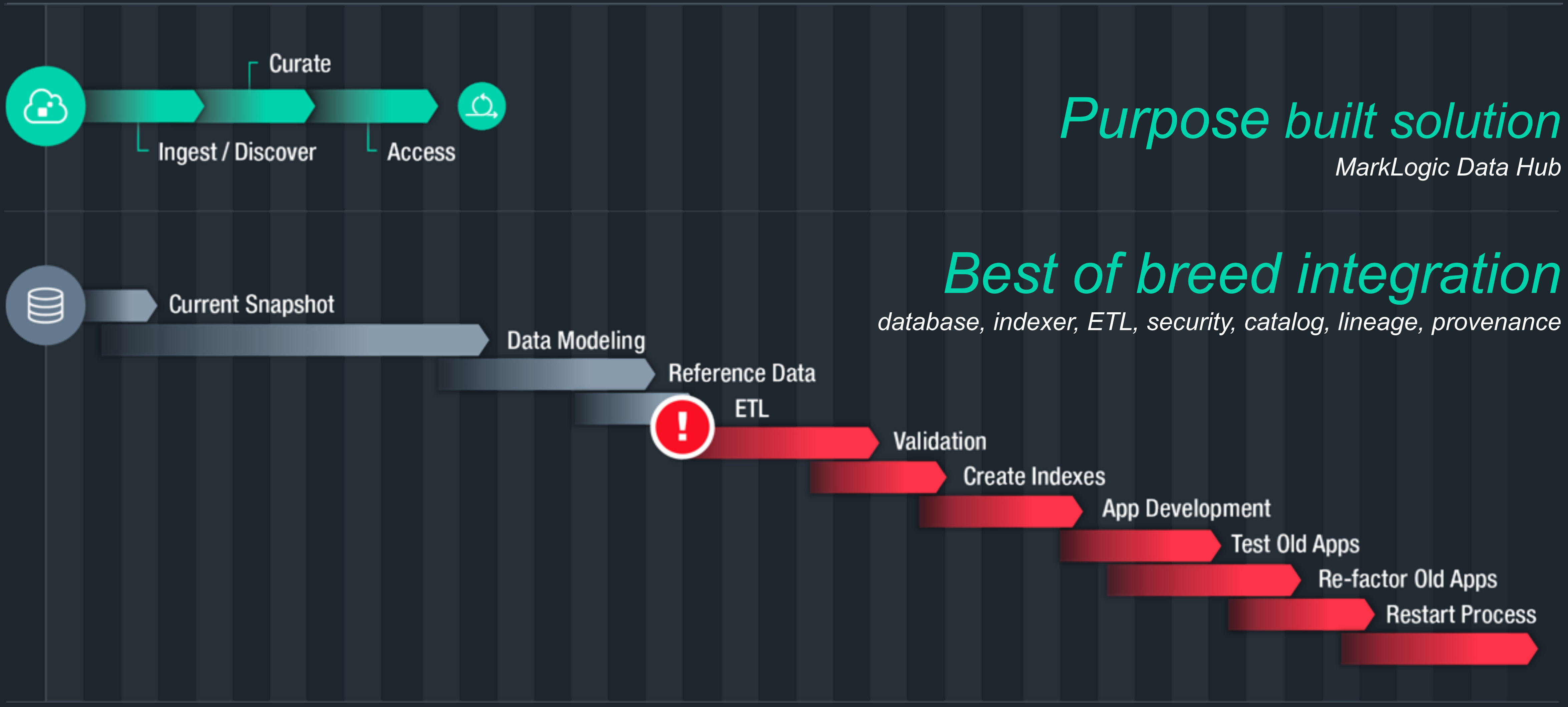
UNIFIED PLATFORM

SMART CURATION

ADVANCED SECURITY

SIMPLE DEVELOPMENT

Data Hub = 10-12x Faster



Purpose built solution

MarkLogic Data Hub

Best of breed integration

database, indexer, ETL, security, catalog, lineage, provenance

MarkLogic

Traceability allowing for Accountability

Michel de Ru, Senior Principal Solution Engineer

Why Traceability By Default?

REGULATORY CONTEXT

- Financial services companies operate in a heavily-regulated marketplace
- Governments, regulators and industry bodies are currently demanding greater levels of transparency, accountability and compliance from financial institutions

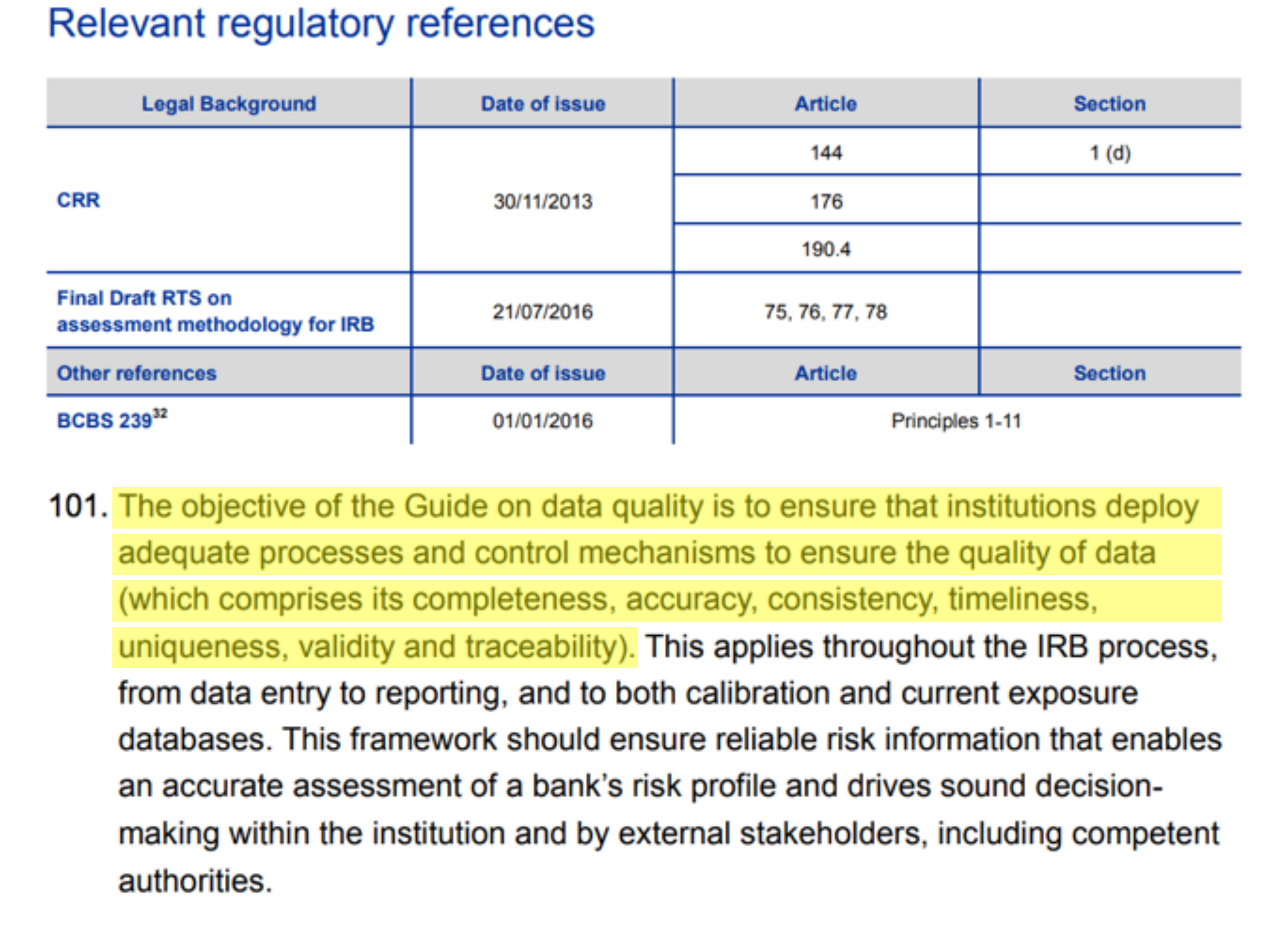
COMPLIANCE COST

- Falling short of your compliance requirements can lead to fines, penalties and legal actions. Even ***going*** through an audit costs significant time and money

FUNCTIONALITY COST

- Higher compliance standards can also translate into greater software functionality and ultimately, more work for your team

Example: TRIM regulation



Relevant regulatory references

Legal Background	Date of issue	Article	Section
CRR	30/11/2013	144	1 (d)
		176	
		190.4	
Final Draft RTS on assessment methodology for IRB	21/07/2016	75, 76, 77, 78	
Other references	**Date of issue**	**Article**	**Section**
BCBS 239[32]	01/01/2016	Principles 1-11	

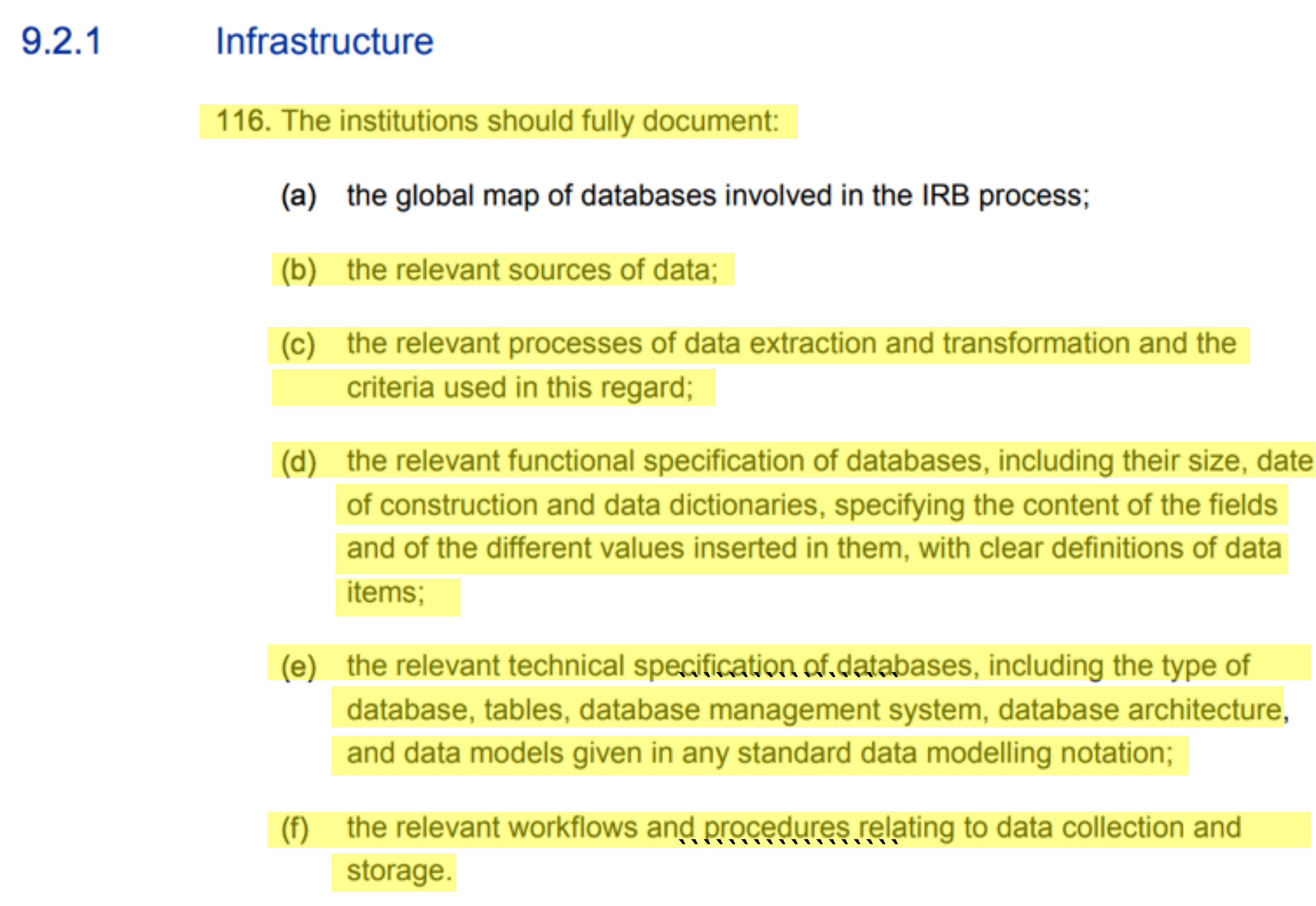
101. The objective of the Guide on data quality is to ensure that institutions deploy adequate processes and control mechanisms to ensure the quality of data (which comprises its completeness, accuracy, consistency, timeliness, uniqueness, validity and traceability). This applies throughout the IRB process, from data entry to reporting, and to both calibration and current exposure databases. This framework should ensure reliable risk information that enables an accurate assessment of a bank's risk profile and drives sound decision-making within the institution and by external stakeholders, including competent authorities.

EUROPEAN CENTRAL BANK

EUROSYSTEM

Guide for the Targeted Review of Internal Models (TRIM)

Example: TRIM regulation



9.2.1 Infrastructure

116. The institutions should fully document:

(a) the global map of databases involved in the IRB process;

(b) the relevant sources of data;

(c) the relevant processes of data extraction and transformation and the criteria used in this regard;

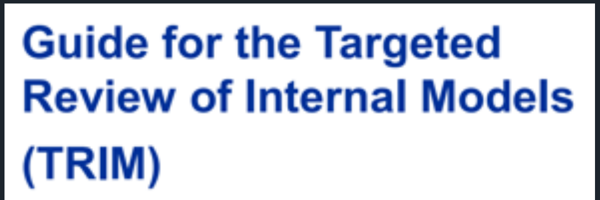
(d) the relevant functional specification of databases, including their size, date of construction and data dictionaries, specifying the content of the fields and of the different values inserted in them, with clear definitions of data items;

(e) the relevant technical specification of databases, including the type of database, tables, database management system, database architecture, and data models given in any standard data modelling notation;

(f) the relevant workflows and procedures relating to data collection and storage.

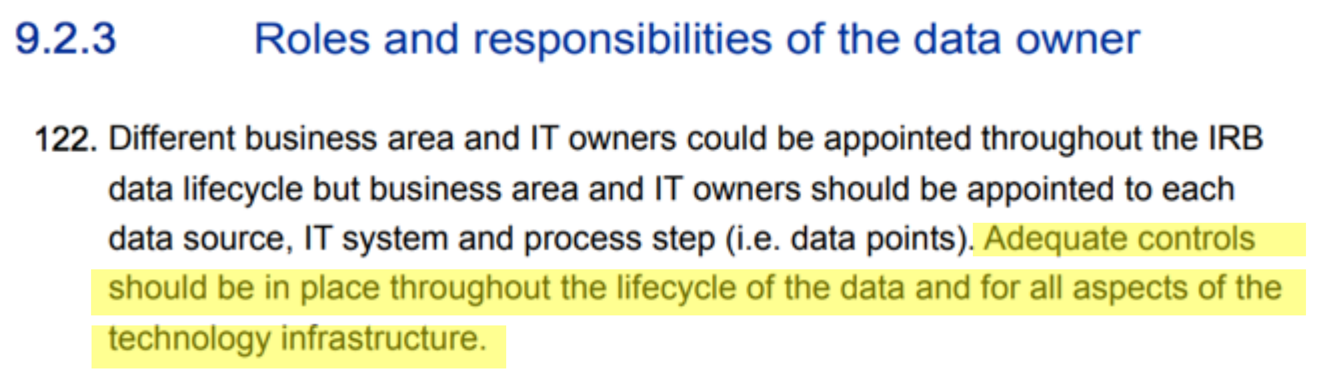
EUROPEAN CENTRAL BANK

EUROSYSTEM



Guide for the Targeted Review of Internal Models (TRIM)

Example: TRIM regulation

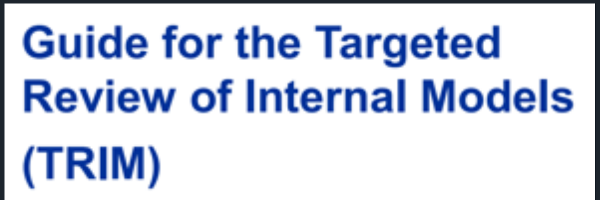


9.2.3 Roles and responsibilities of the data owner

122. Different business area and IT owners could be appointed throughout the IRB data lifecycle but business area and IT owners should be appointed to each data source, IT system and process step (i.e. data points). Adequate controls should be in place throughout the lifecycle of the data and for all aspects of the technology infrastructure.

EUROPEAN CENTRAL BANK

EUROSYSTEM

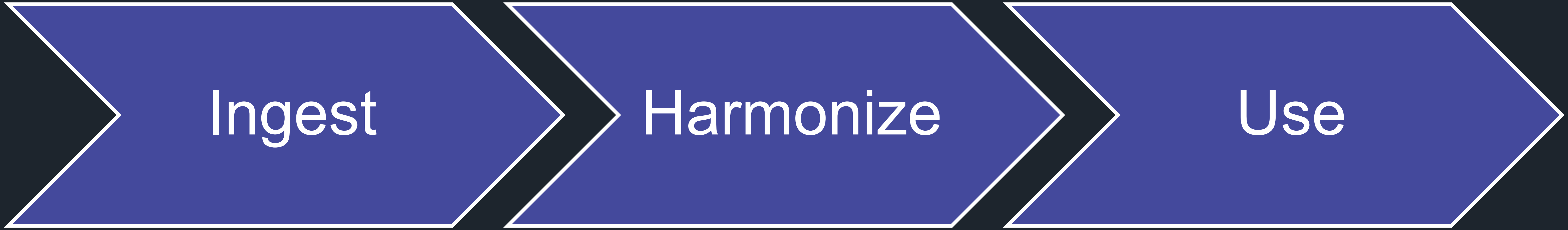


Guide for the Targeted Review of Internal Models (TRIM)

Typical traceability questions

- “How was this result conceived”
- “Which steps were involved in this calculation”
- “What do the data elements mean”
- “Which source data was used for this process”
- “What version of the source data do I have”
- “Who loaded it into the system and when”
- “When was the process called and how long did it take”
- “Who ran this process and why”
- “Which enrichments were done on this data”
- “What external data was used for this information”

Typical end-to-end process



Ingest

Harmonize

Use

“Which source data was used for this process”

“What version of the source data do I have”

“Who loaded it into the system and when”

“Which steps were involved in this calculation”

“Which enrichments were done on this data”

“What external data was used for this information”

“How was this result conceived”

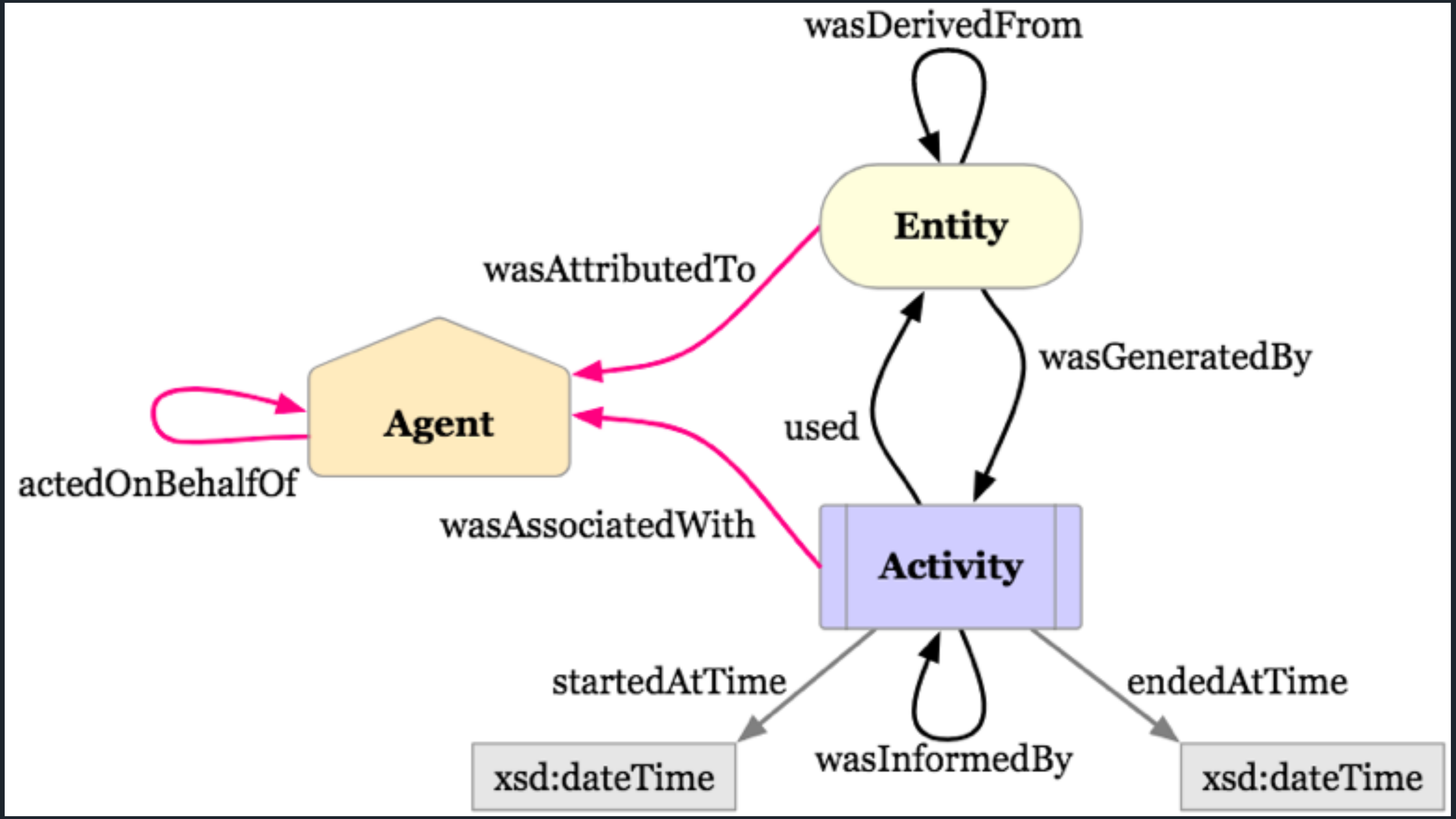
“What do the data elements mean”

“When was the process called and how long did it take”

“Who ran this process and why”

A model for end-to-end traceability

- The PROV Ontology
- W3C recommendation 2013



Actors for traceability

Entity

This is about the data itself

Activity

A process that was called using or producing the data

Agent

The responsible actor using the data (Entity) or process (Activity)

Relationships between actors

Was Generated By

Was Derived From

Was Attributed To

Started At Time

Ended At Time

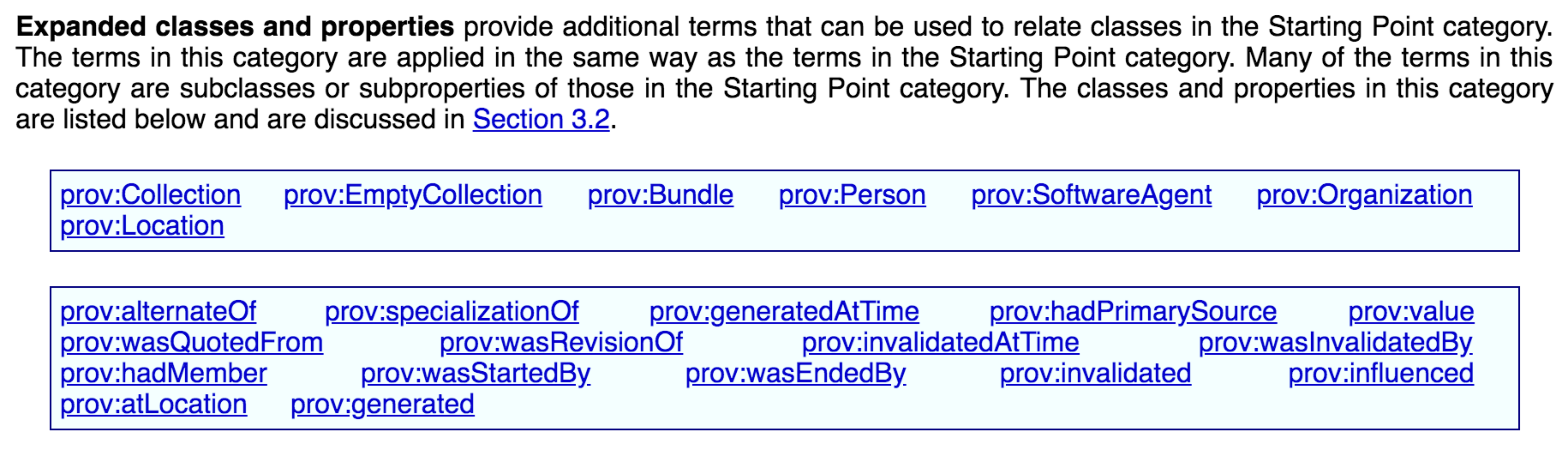
Used

Was Informed By

Was Associated With

Acted On Behalf Of

And many more

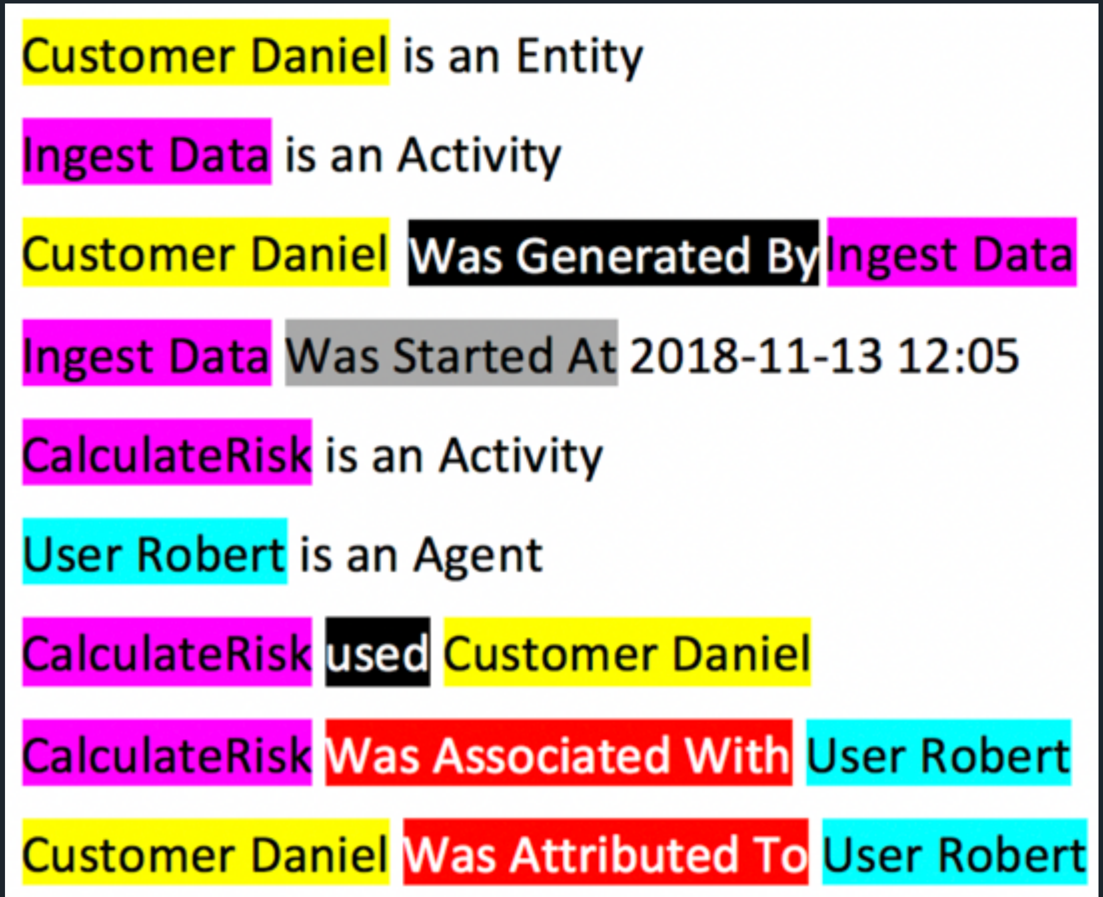


Expanded classes and properties provide additional terms that can be used to relate classes in the Starting Point category. The terms in this category are applied in the same way as the terms in the Starting Point category. Many of the terms in this category are subclasses or subproperties of those in the Starting Point category. The classes and properties in this category are listed below and are discussed in Section 3.2.

prov:Collection prov:EmptyCollection prov:Bundle prov:Person prov:SoftwareAgent prov:Organization prov:Location

prov:alternateOf prov:specializationOf prov:generatedAtTime prov:hadPrimarySource prov:value prov:wasQuotedFrom prov:wasRevisionOf prov:invalidatedAtTime prov:wasInvalidatedBy prov:hadMember prov:wasStartedBy prov:wasEndedBy prov:invalidated prov:influenced prov:atLocation prov:generated

Example



Customer Daniel is an Entity

Ingest Data is an Activity

Customer Daniel Was Generated By Ingest Data

Ingest Data Was Started At 2018-11-13 12:05

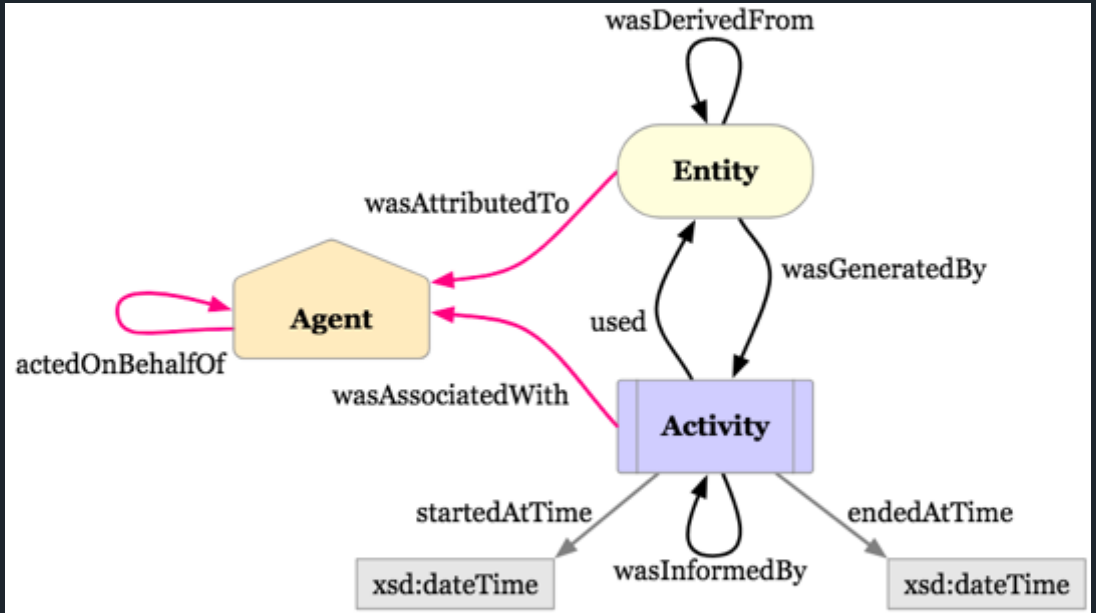
CalculateRisk is an Activity

User Robert is an Agent

CalculateRisk used Customer Daniel

CalculateRisk Was Associated With User Robert

Customer Daniel Was Attributed To User Robert



Conclusion

- PROV-O is a world wide standard, kept by W3C
- PROV-O allows for full end-to-end lineage and provenance
- Full lineage and provenance allows for traceability
- **Traceability is a pre-requisite for accountability**
- **Accountability provides for a good night sleep in a regulated vertical!**

Movie Palace Demo

Michel de Ru, Senior Principal Solution Engineer

Introducing Movie Palace

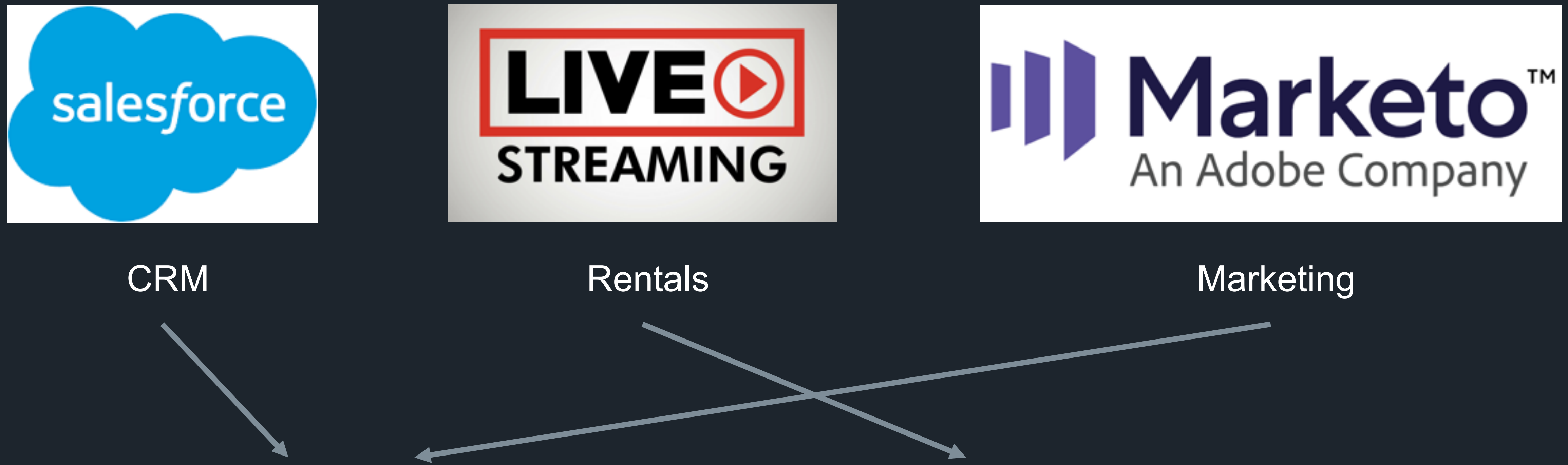
Disconnected data increases churn

CRM

Rentals

Marketing

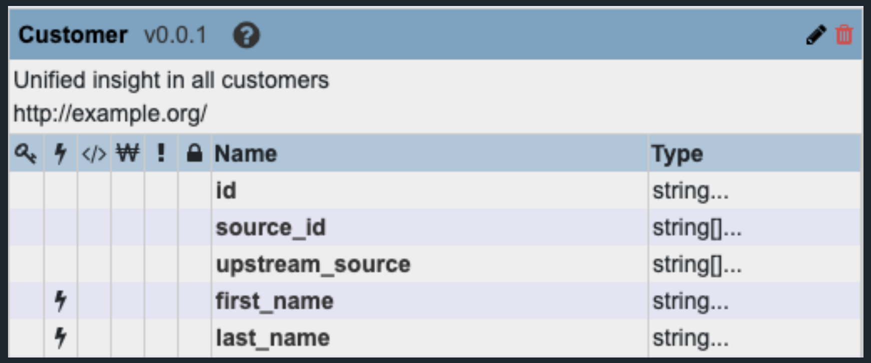
Business entities



CRM

Rentals

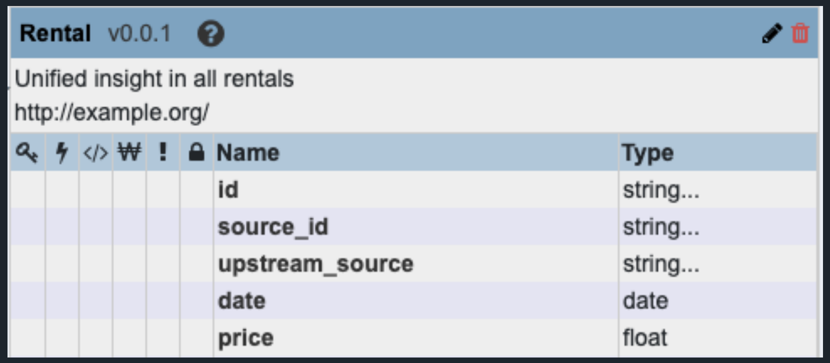
Marketing



Customer v0.0.1

Unified insight in all customers
http://example.org/

🔑	⚡	</>	₩	!	🔒	Name	Type
						id	string...
						source_id	string[]...
						upstream_source	string[]...
	⚡					first_name	string...
	⚡					last_name	string...



Rental v0.0.1

Unified insight in all rentals
http://example.org/

🔑	⚡	</>	₩	!	🔒	Name	Type
						id	string...
						source_id	string...
						upstream_source	string...
						date	date
						price	float