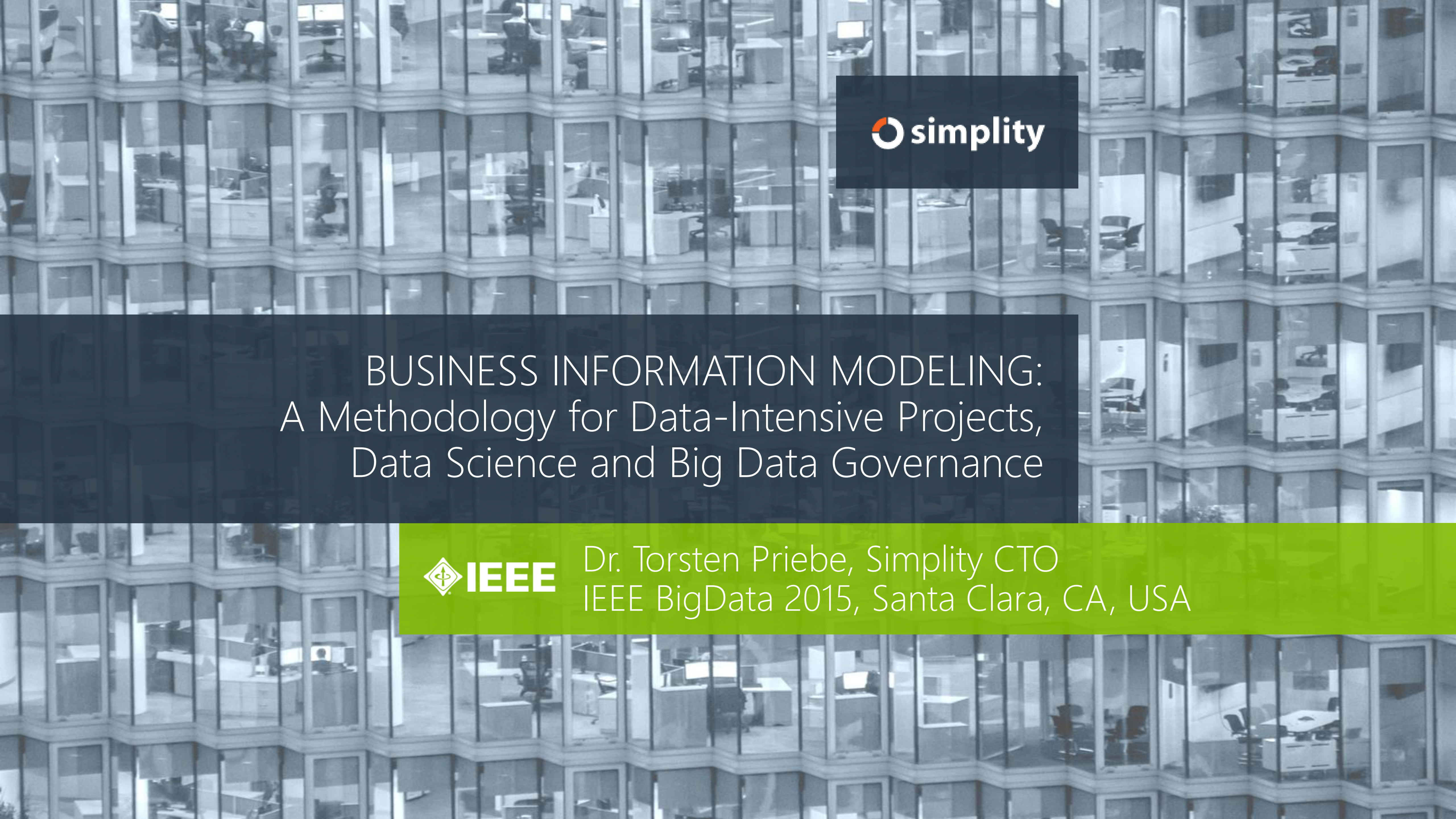


simplity

BUSINESS INFORMATION MODELING: A Methodology for Data-Intensive Projects, Data Science and Big Data Governance

IEEE

Dr. Torsten Priebe, Simplity CTO
IEEE BigData 2015, Santa Clara, CA, USA

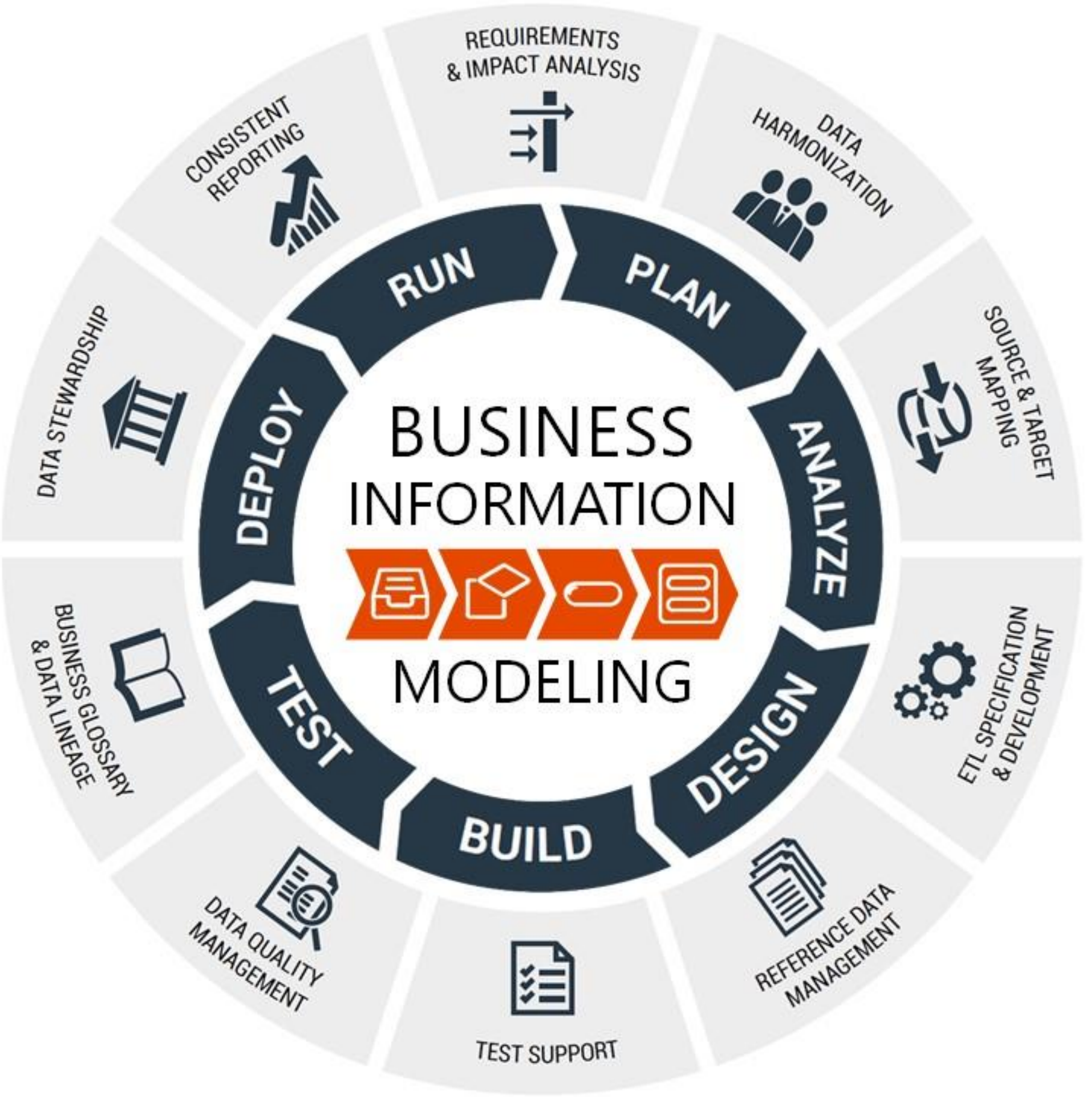
BUSINESS INFORMATION MODELING (BIM)

Business Information Modeling (BIM)
is a holistic approach to structured business requirements definition, harmonization and model-driven implementation of data-intensive IT solutions

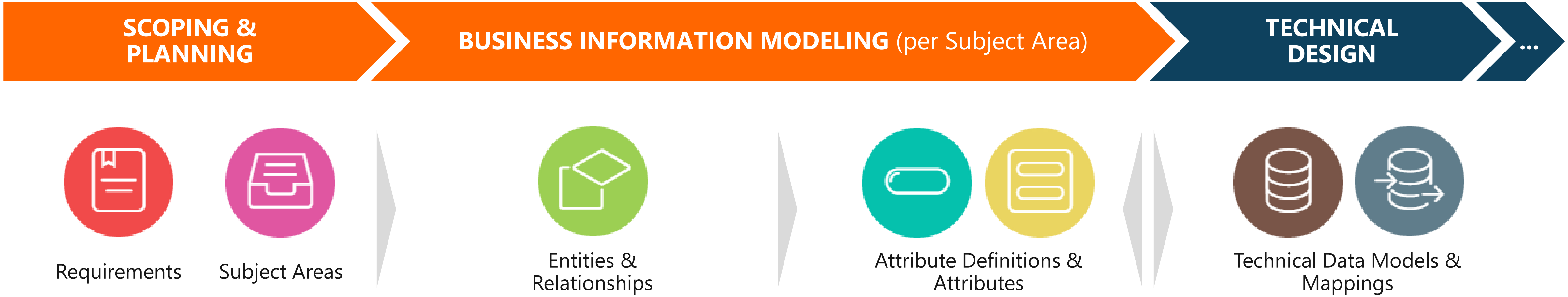
A business information model is defined in terms of:

- *Subject areas*
- *Entities*
- *Attribute definitions*
- *Attributes*

The model behind BIM is similar to ontology languages such as OWL. Our commercial implementation in Accurity Glossary is based on a relational database, but we continue to use OWL and Apache Jena for prototyping. W3C OWL



CORE ELEMENTS OF THE BIM METHODOLOGY



SCOPING & PLANNING

BUSINESS INFORMATION MODELING (per Subject Area)

TECHNICAL DESIGN

...

Requirements
Subject Areas
Entities & Relationships
Attribute Definitions & Attributes
Technical Data Models & Mappings

- Definition of **scope**, **business requirements** as input
- Structuring into **subject areas** (e.g. Customer, Loan, Collateral)
- **Project planning**, e.g. sequence of subject areas

- **Identification of entities** per subject area (e.g. Loan Account, but also subtypes like Mortgage)
- **Identification of relationships** between entities
- Definition and **harmonization of descriptions**

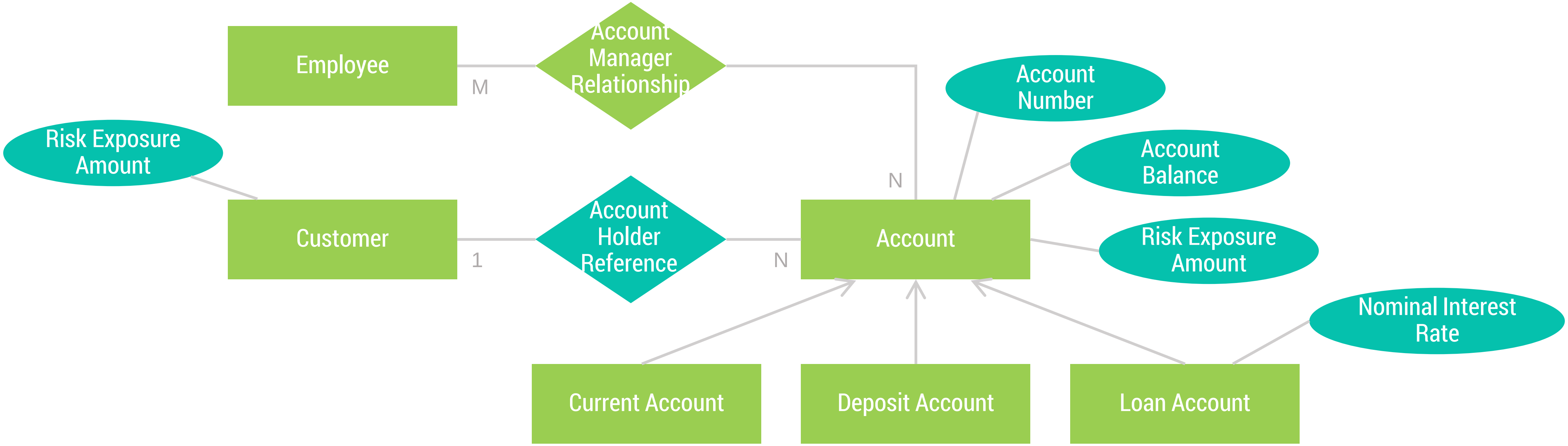
(cf. Classes in W3C OWL)

- **Precise definition of attributes** (e.g. Risk Exposure Amount)
- Assignment of "attribute definitions" to entities
- **Harmonization of attribute descriptions and calculation rules**

(cf. Properties in W3C OWL)

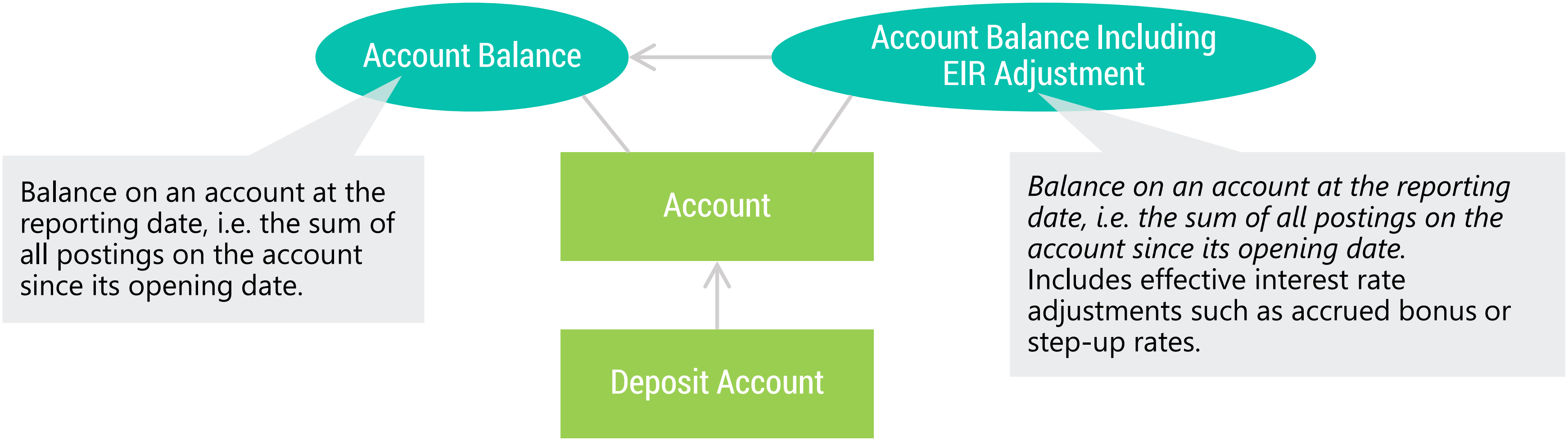
- Definition of (logical and physical) **technical data models**
- Definition of **source- and target mappings** as basis for a **model-driven ETL development**

BIM IS BASED ON AN EXTENDED ENTITY-RELATIONSHIP APPROACH



● = Entity (cf. Class in W3C OWL) ● = Attribute (Definition) (cf. Property in W3C OWL)

INHERITANCE ON ATTRIBUTE DEFINITION LEVEL



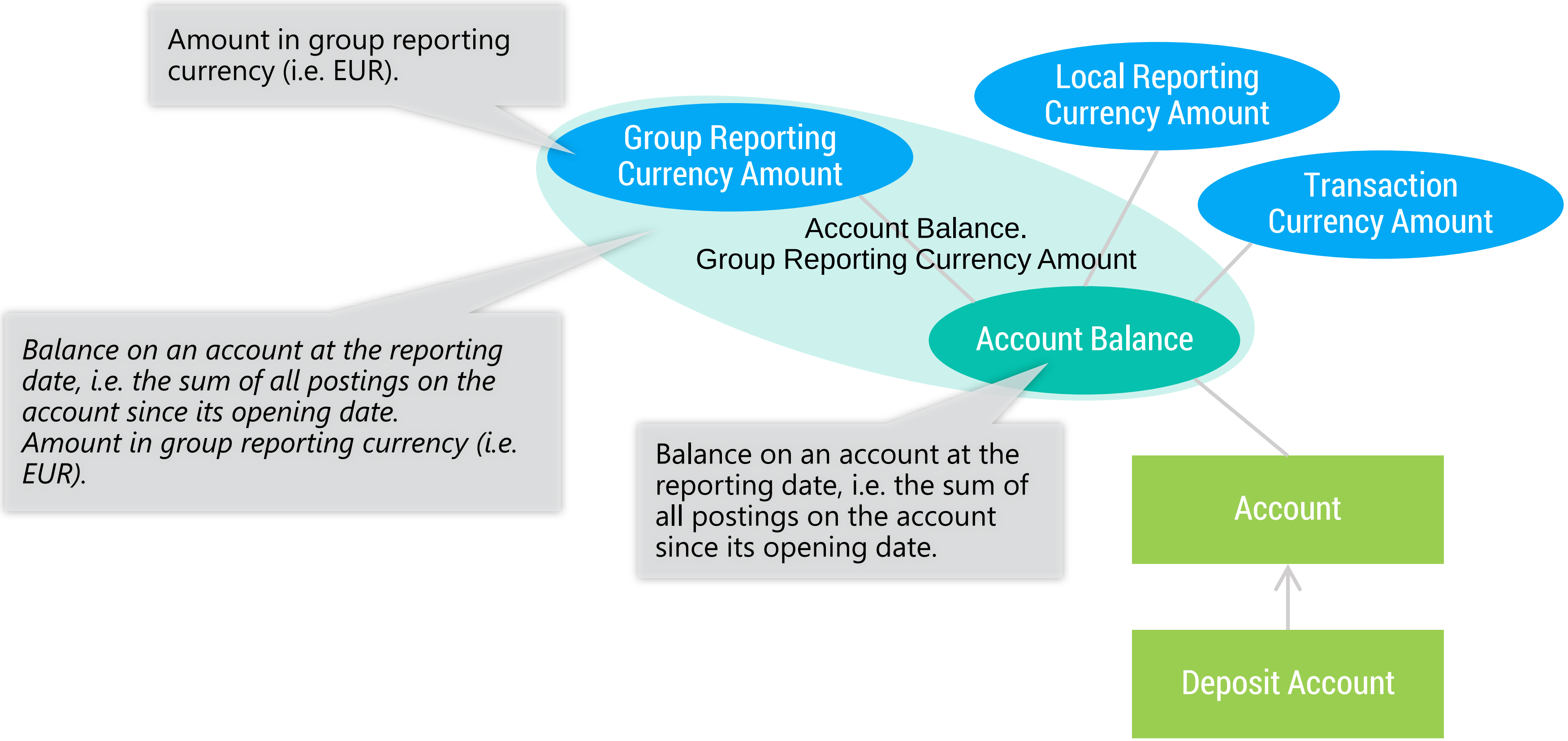
Balance on an account at the reporting date, i.e. the sum of all postings on the account since its opening date.

Balance on an account at the reporting date, i.e. the sum of all postings on the account since its opening date.
Includes effective interest rate adjustments such as accrued bonus or step-up rates.

● = Entity ● = Attribute (Definition)

cf. Parent Properties in W3C OWL

COMPOSITE ATTRIBUTES



Amount in group reporting currency (i.e. EUR).

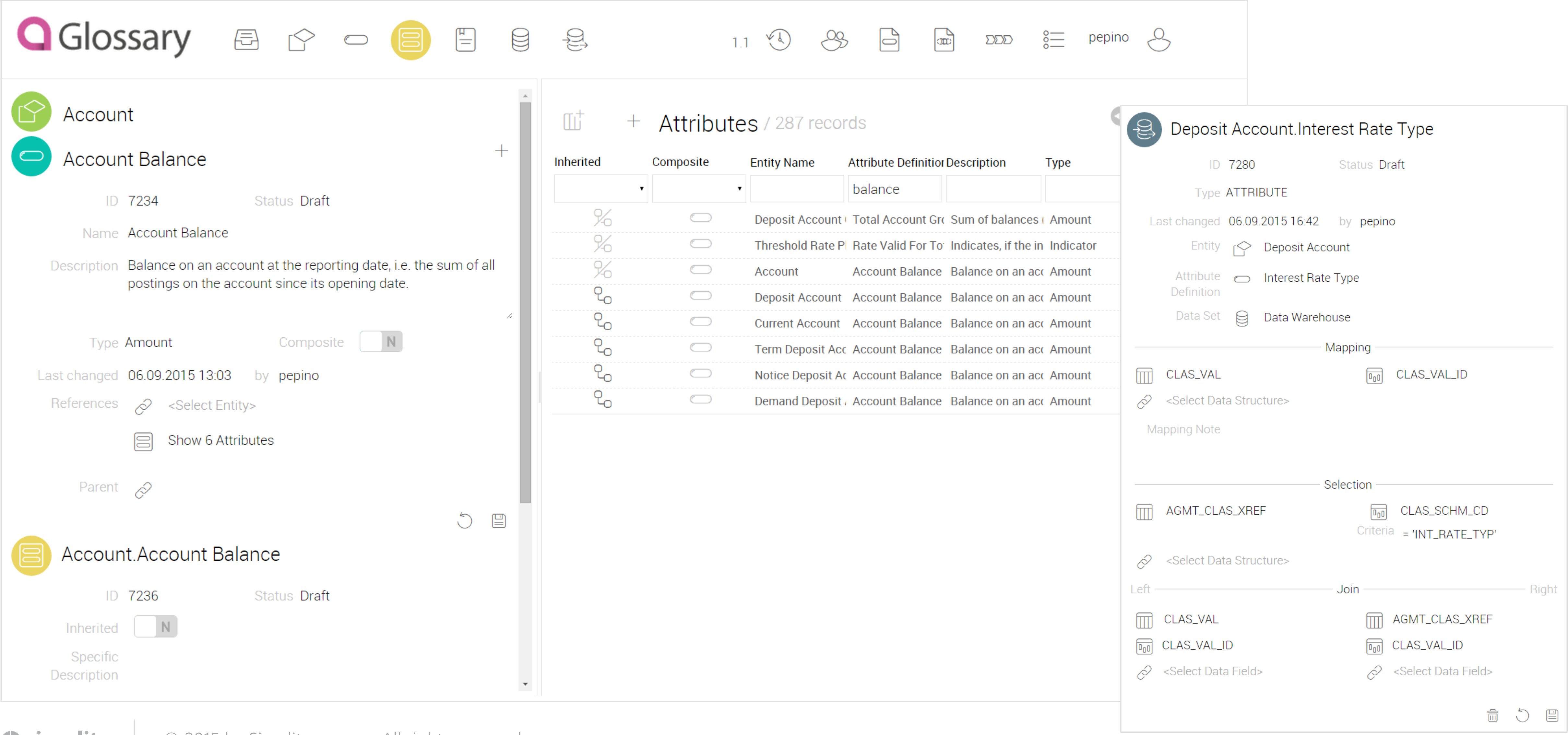
Balance on an account at the reporting date, i.e. the sum of all postings on the account since its opening date.
Amount in group reporting currency (i.e. EUR).

Balance on an account at the reporting date, i.e. the sum of all postings on the account since its opening date.

Group Reporting Currency Amount
Local Reporting Currency Amount
Transaction Currency Amount
Account Balance.
Group Reporting Currency Amount
Account Balance
Account
Deposit Account

🟩 = Entity ● = Attribute (Definition) ● = Attribute Component

BUSINESS INFORMATION MODEL IN ACCURITY GLOSSARY



Account

Account Balance

ID 7234 Status Draft

Name Account Balance

Description Balance on an account at the reporting date, i.e. the sum of all postings on the account since its opening date.

Type Amount Composite N

Last changed 06.09.2015 13:03 by pepino

References <Select Entity>

Show 6 Attributes

Parent

Account.Account Balance

ID 7236 Status Draft

Inherited N

Specific Description

Attributes / 287 records

Inherited	Composite	Entity Name	Attribute Definitior	Description	Type
			balance		
		Deposit Account (	Total Account Gr	Sum of balances (	Amount
		Threshold Rate P	Rate Valid For To	Indicates, if the in	Indicator
		Account	Account Balance	Balance on an acc	Amount
		Deposit Account	Account Balance	Balance on an acc	Amount
		Current Account	Account Balance	Balance on an acc	Amount
		Term Deposit Acc	Account Balance	Balance on an acc	Amount
		Notice Deposit Ac	Account Balance	Balance on an acc	Amount
		Demand Deposit	Account Balance	Balance on an acc	Amount

Deposit Account.Interest Rate Type

ID 7280 Status Draft

Type ATTRIBUTE

Last changed 06.09.2015 16:42 by pepino

Entity Deposit Account

Attribute Definition Interest Rate Type

Data Set Data Warehouse

Mapping

CLAS_VAL CLAS_VAL_ID

<Select Data Structure>

Mapping Note

Selection

AGMT_CLAS_XREF CLAS_SCHM_CD

Criteria = 'INT_RATE_TYP'

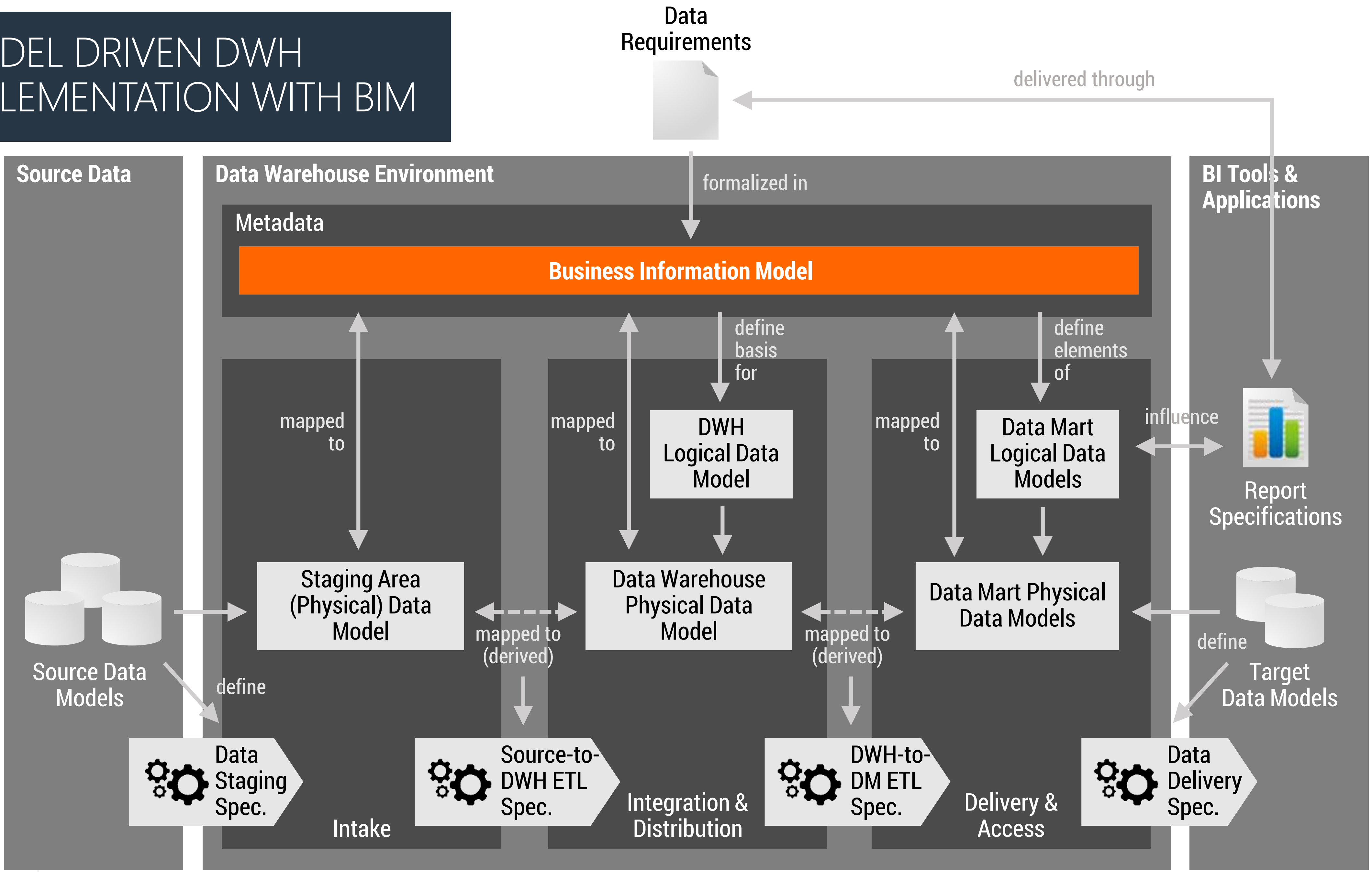
<Select Data Structure>

Left Join Right

Left	Right
CLAS_VAL	AGMT_CLAS_XREF
CLAS_VAL_ID	CLAS_VAL_ID
<Select Data Field>	<Select Data Field>

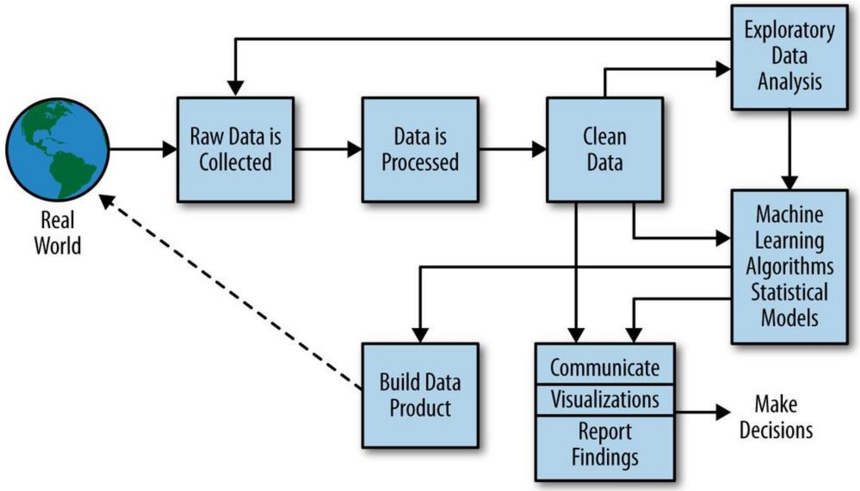
simplity

MODEL DRIVEN DWH IMPLEMENTATION WITH BIM

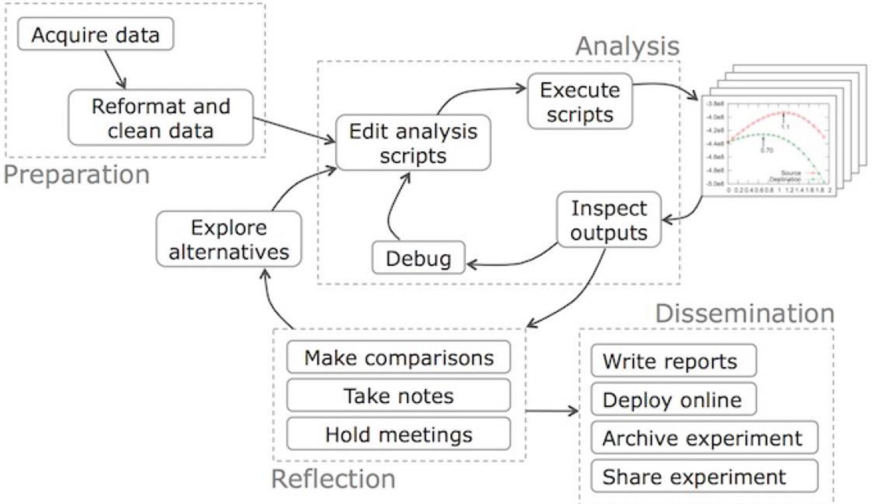


simplity

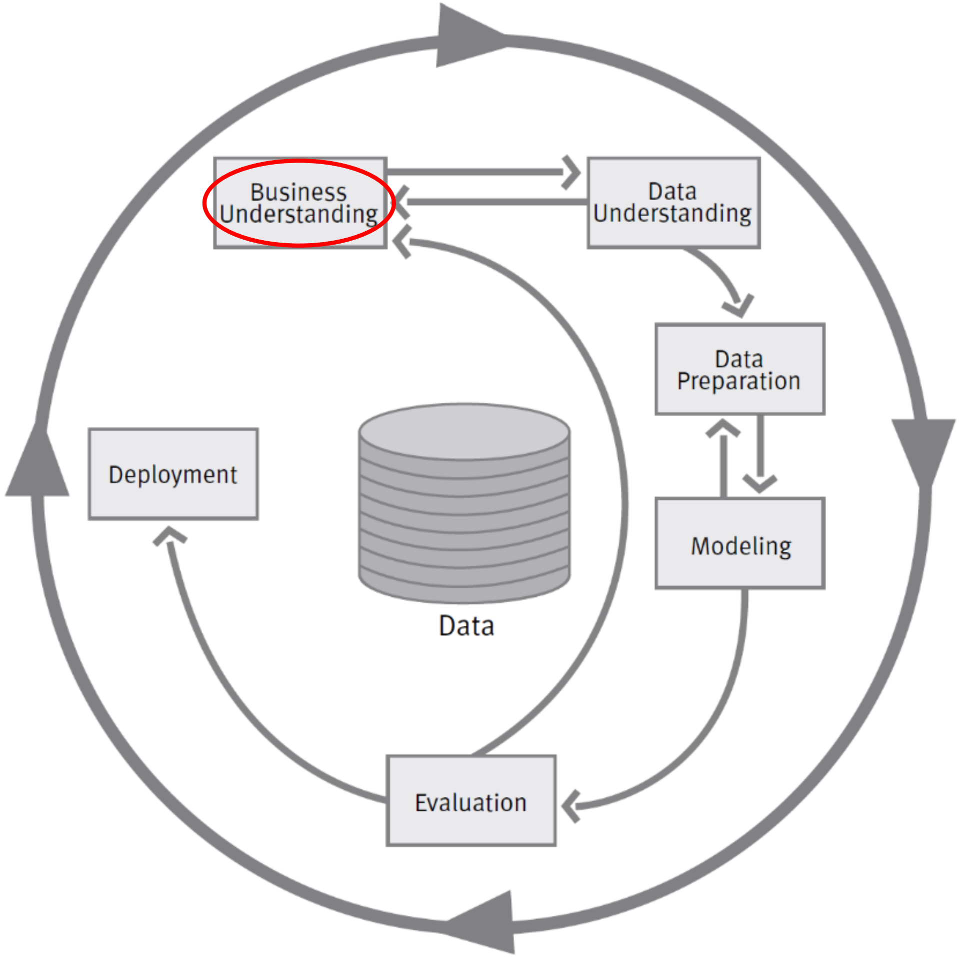
MOST DATA SCIENCE WORKFLOW PROPOSALS NEGLECT BUSINESS REQUIREMENTS



Source: C. O'Neil, R. Schutt, Doing Data Science: Straight Talk from the Frontline, O'Reilly, 2013



Source: P. Guo, “Data Science Workflow: Overview and Challenges”, blog@CACM, Communications of the ACM, 2013



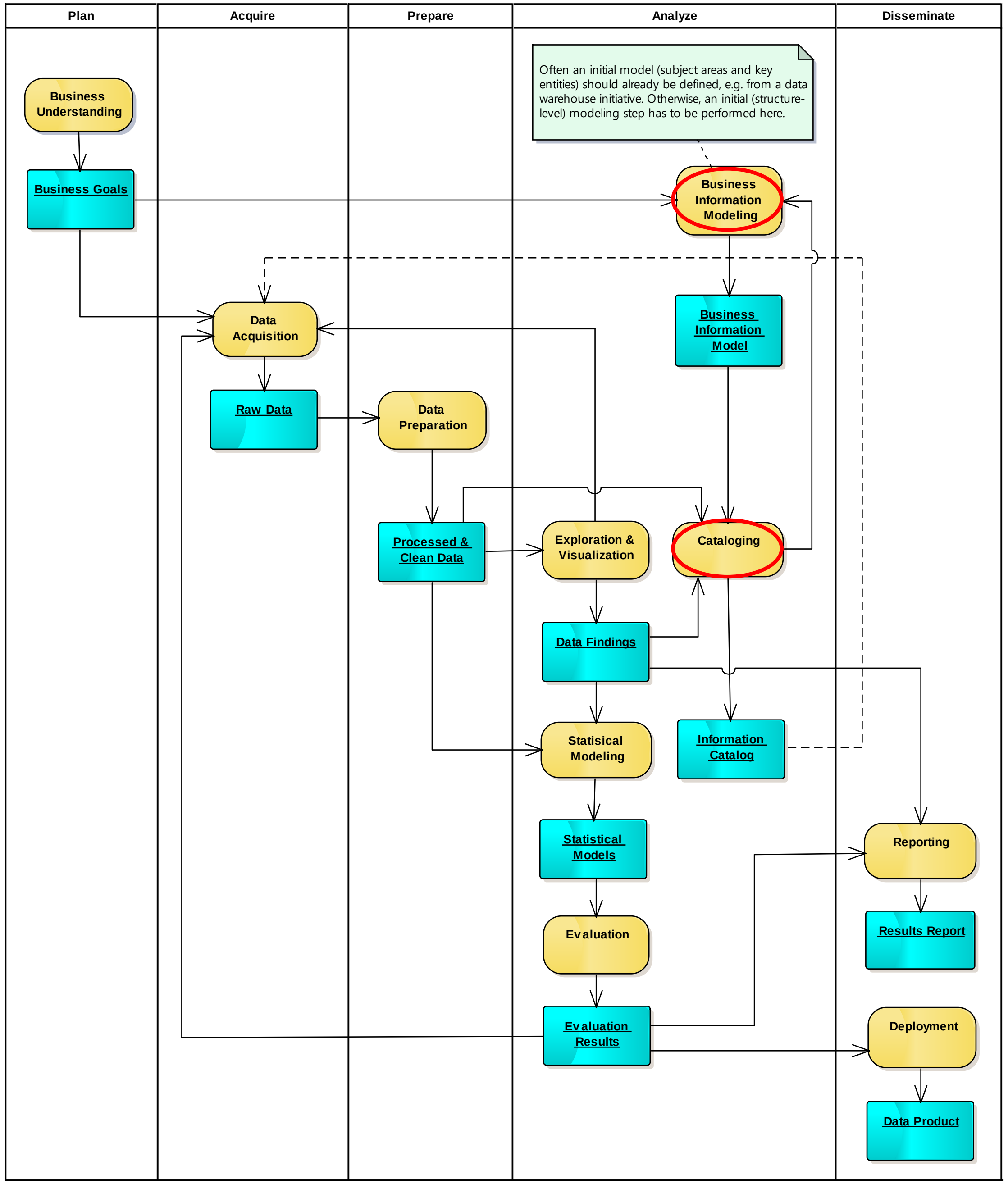
Source: C. Shearer, “The CRISP-DM Model: The New Blueprint for Data Mining”, Journal of Data Warehousing, 5(4), 2000

BIM & THE DATA SCIENCE PROCESS

- Based on data mining process models like CRISP-DM, but also other “data science workflow” proposals, we defined a **consolidated process model**

- On the one hand, the business information model serves as a good basis to **capture at least high-level business requirements** or goals (expected output), e.g.
  - *Explore the available information on Customers*
  - *Project the Probability of Default of a Customer*

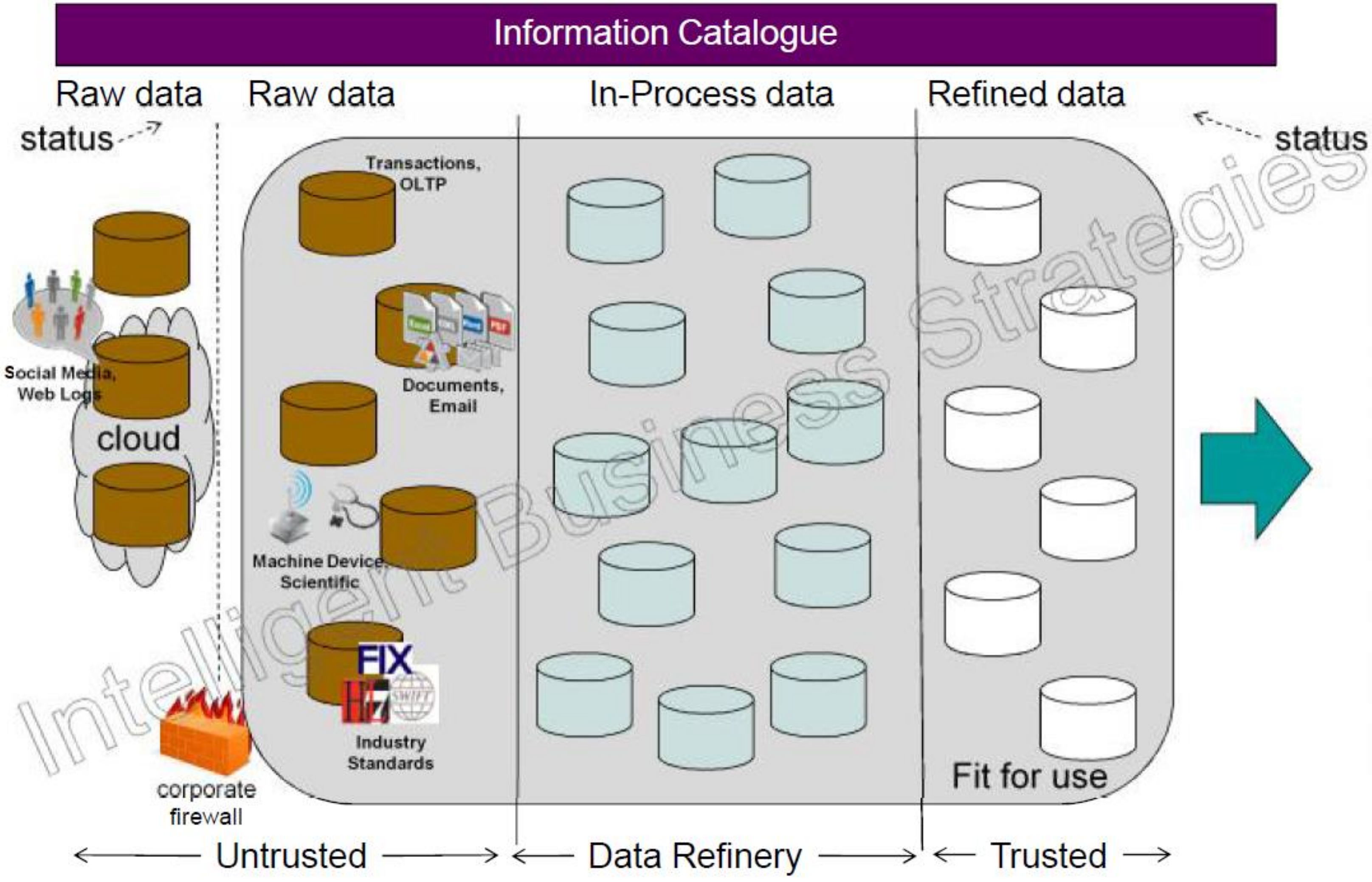
- On the other hand, the business information model is suitable for **cataloging the data sources** (input data)



Often an initial model (subject areas and key entities) should already be defined, e.g. from a data warehouse initiative. Otherwise, an initial (structure-level) modeling step has to be performed here.

BACKUP

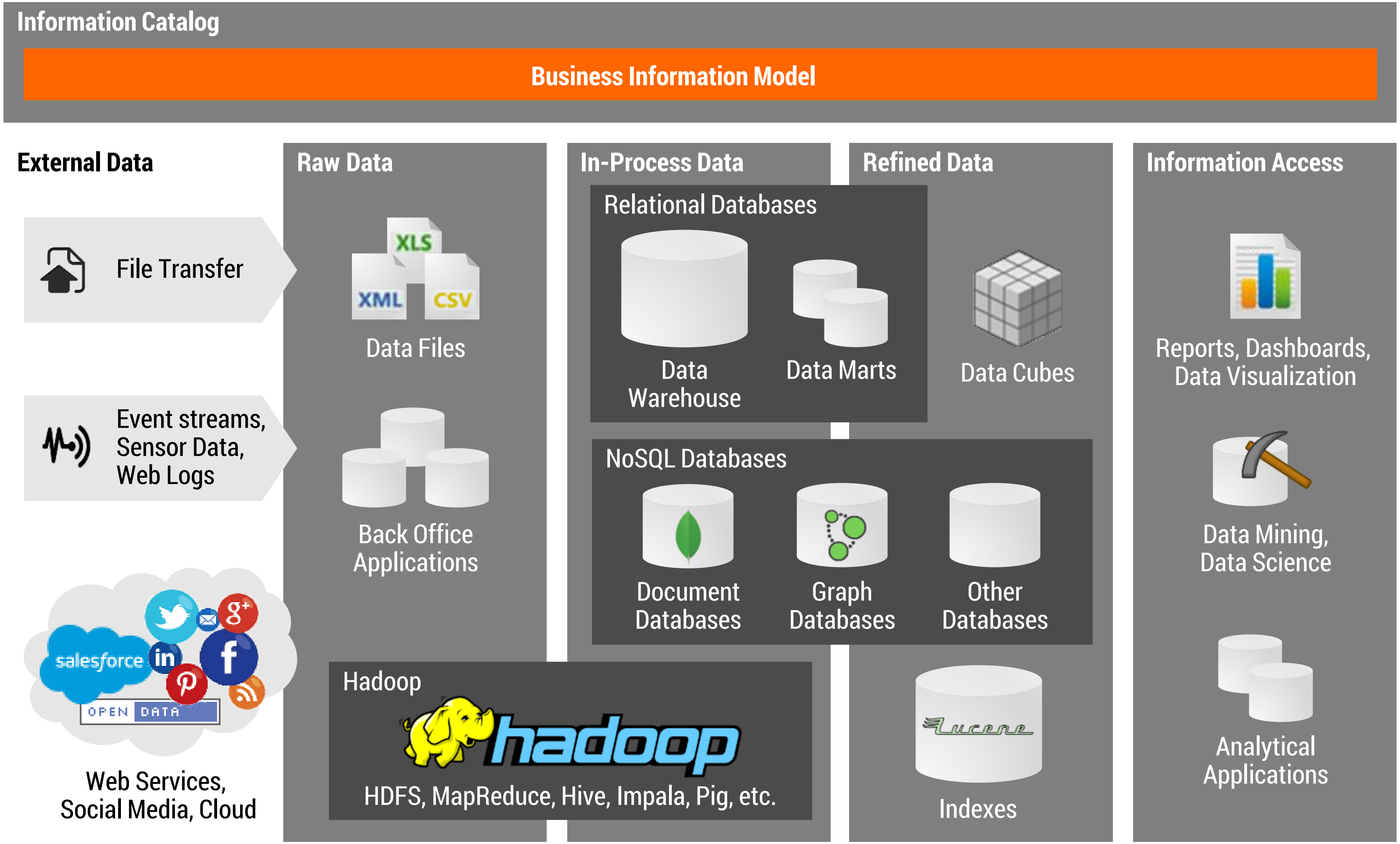
MOTIVATION: THE ROLE OF AN INFORMATION CATALOG IN A DATA RESERVOIR



Source: Mike Ferguson, Intelligent Business Strategies, Juni 2015

- Document where data is so others can find out what information is available
- View metadata lineage about the data
  - See where it came from
- Name and describe data
  - Define shared business vocabulary terms
- Classify data, e.g.
  - Personal data
  - Sensitive data – to indicate protection needed from unauthorized access
  - Governance rules can be applied to different data classifications
- Define data governance policies
- Shop for data (Data As A Service – DaaS)
- Create subscriptions

VISION: BIG DATA LANDSCAPE WITH BIM AS INFORMATION CATALOG



Information Catalog

Business Information Model

External Data

File Transfer

Event streams, Sensor Data, Web Logs

Web Services, Social Media, Cloud

Raw Data

Data Files

Back Office Applications

In-Process Data

Relational Databases

Data Warehouse

Data Marts

NoSQL Databases

Document Databases

Graph Databases

Other Databases

Refined Data

Data Cubes

Indexes

Information Access

Reports, Dashboards, Data Visualization

Data Mining, Data Science

Analytical Applications

Hadoop

HDFS, MapReduce, Hive, Impala, Pig, etc.

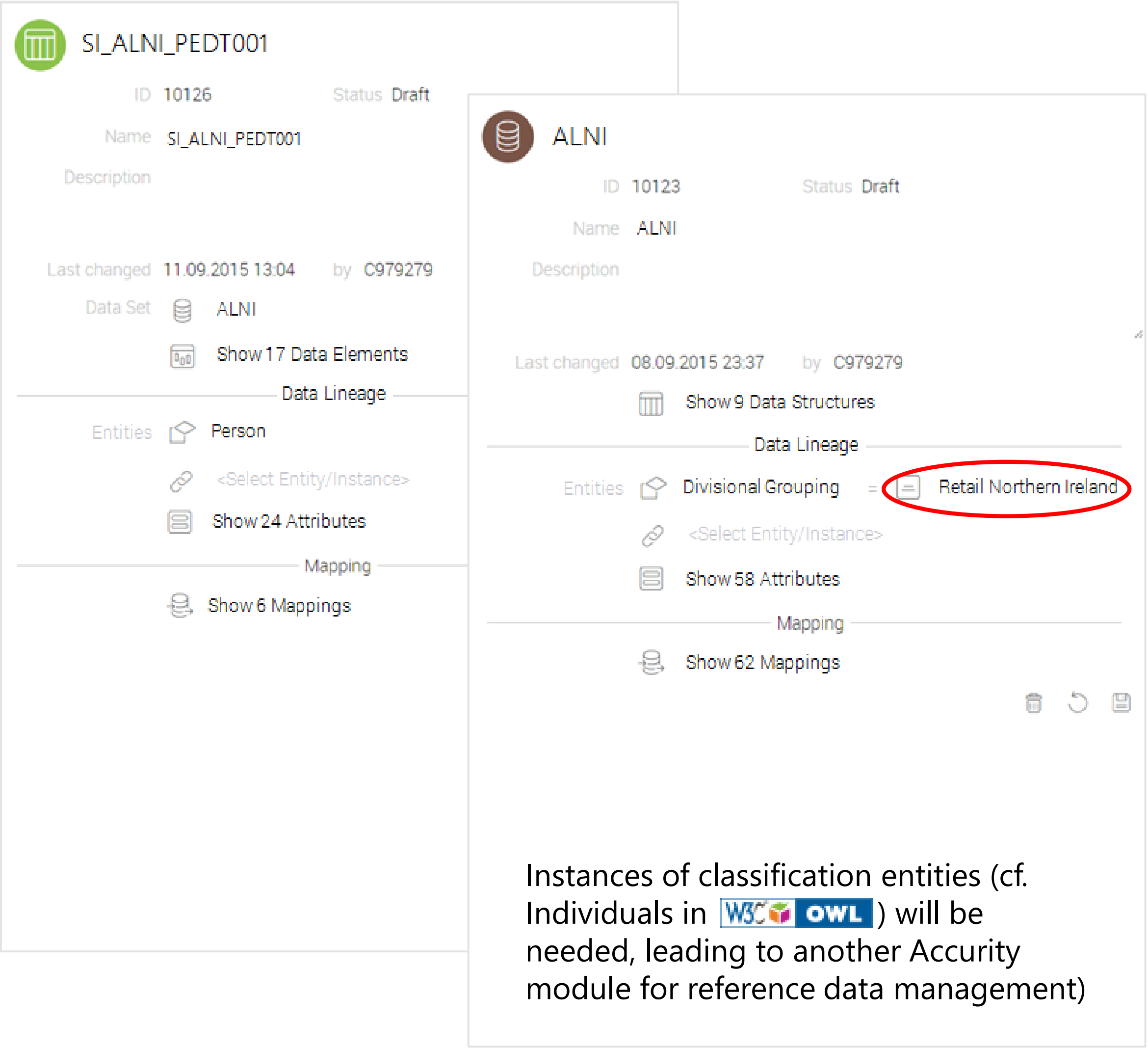
TAGGING DATA SETS, STRUCTURES AND ELEMENTS IN ACCURITY GLOSSARY

- Tagging data sets and data structures provides a quick way of **cataloging big data assets** based on the business model
- Also unstructured documents can be tagged accordingly
- See also: **Semi-automatic annotation of text documents** with semantic metadata using machine-learning algorithms (Priebe et al. 2005)
- The common model enables an **integrated view on both structured and unstructured data**

BIM Terminology		**Data Store Terminology**				
Business Model	***Technical Model***	***Relational Database***	***File System[a]***		***Document Database[d]***	***Graph Database[f]***
			structured	*unstructured*		
Subject Area	Data Set	Database, Schema	Directory		Database	Database
Entity	Data Structure	Table, View	File[b]	Document	Collection	Node, Relationship
Attribute	Data Element	Column	Field[c]	–	Field[e]	Property

Do we need to extend the BIM model to suport (big) data cataloging?

How to deal with the manual tagging effort, is manual tagging feasible?



SI_ALNI_PEDT001

ID 10126 Status Draft

Name SI_ALNI_PEDT001

Description

Last changed 11.09.2015 13:04 by C979279

Data Set ALNI

Show 17 Data Elements

Data Lineage

Entities Person

<Select Entity/Instance>

Show 24 Attributes

Mapping

Show 6 Mappings

ALNI

ID 10123 Status Draft

Name ALNI

Description

Last changed 08.09.2015 23:37 by C979279

Show 9 Data Structures

Data Lineage

Entities Divisional Grouping = Retail Northern Ireland

<Select Entity/Instance>

Show 58 Attributes

Mapping

Show 62 Mappings

Instances of classification entities (cf. Individuals in W3C OWL) will be needed, leading to another Accurity module for reference data management)

simplity

Dr. Torsten Priebe
Chief Technology Officer
Phone: +43 699 14166193
Email: torsten.priebe@simplity.eu

IEEE BigData 2015, Santa Clara, CA, USA

Thank you for your attention!

IEEE