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

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

BUSINESS INFORMATION MODELING (per Subject Area)

TECHNICAL DESIGN

- 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 [WSDG 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 [WSDG 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

- Employee
- Customer
- Account Manager Relationship
- Account Holder Reference
- Account
- Current Account
- Deposit Account
- Loan Account
- Account Number
- Account Balance
- Risk Exposure Amount
- Nominal Interest Rate
- Risk Exposure Amount

= Entity (cf. Class in [W3C OWL])
= Attribute (Definition) (cf. Property in [W3C OWL])
INHERITANCE ON ATTRIBUTE DEFINITION LEVEL

Account Balance

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

Account Balance Including EIR Adjustment

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.

Account

Deposit Account

Deposit Account = Entity

Account = Attribute (Definition)

cf. Parent Properties in [W3C OWL]
Account Balance

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

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

Local Reporting Currency Amount

Transaction Currency Amount

Deposit Account

Account

Balance on an account at the reporting date, i.e. the sum of all postings on the account since its opening date.
BUSINESS INFORMATION MODEL IN ACCURITY GLOSSARY
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


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

- 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

Source: Mike Ferguson, Intelligent Business Strategies, Juni 2015
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**<sup>a</sup> | **Document Database**<sup>c</sup> | **Graph Database**<sup>f</sup>
Subject Area | Data Set | Database, Schema | structured | -- | --
Entity | Data Structure | Table, View | -- | Document | Node, Relationship
Attribute | Data Element | Column | Field<sup>c</sup> | Field<sup>e</sup> | Property

---

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

How to deal with the manual tagging effort, is manual tagging feasible?
Thank you for your attention!