Discovering Analytical Concepts from User Profiles

Research Progress Report

Jovan Varga[1,2], Oscar Romero[1], Torben Bach Pedersen[2], and Christian Thomsen[2]

[1] Universitat Politècnica de Catalunya – BarcelonaTECH

[2] Aalborg Universitet

Outline

- Introduction
- Our Vision
- Analytical Metadata and Analytical Metadata Processing Types
- SM4AM: A Semantic Metamodel for Analytical Metadata
- Examples of Data Exploration Action and User Elements
- Running Example: World Bank Linked Data
- On the Quest for the Richer Schema Metadata
- From QB to QB4OLAP
- The Enrichment Automation Challenges
- Addressing Partial Semantics – Aggregation Functions and Dimension Hierarchies
- Publications and Future Work

Introduction

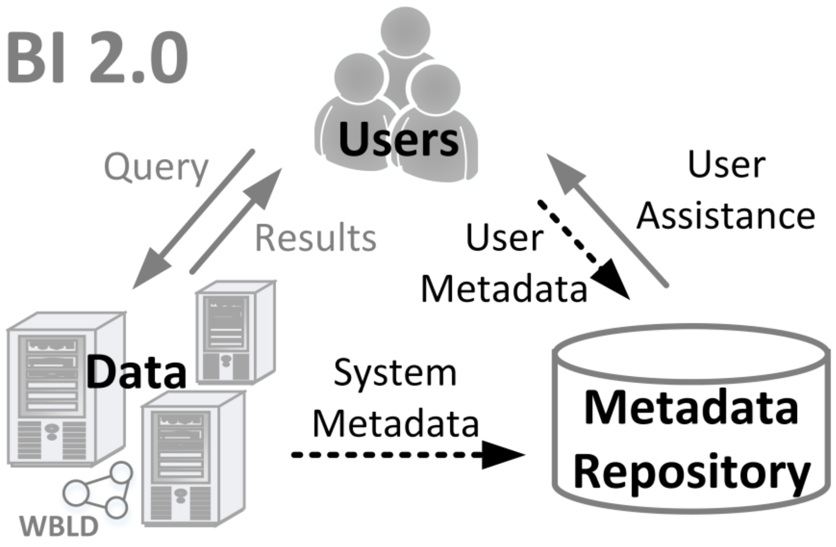
- Next generation Business Intelligence systems (i.e., BI 2.0 systems) need to be **user-centric**

- Non-expert **user**s need the **assistance** to navigate through the data landscape to perform their analysis

- **The metadata are the fuel for** different **user assistance** (e.g., query recommendation) algorithms and they directly determine the assistance possibilities

- However, the metadata **management** and **organization** are typically **overlooked**

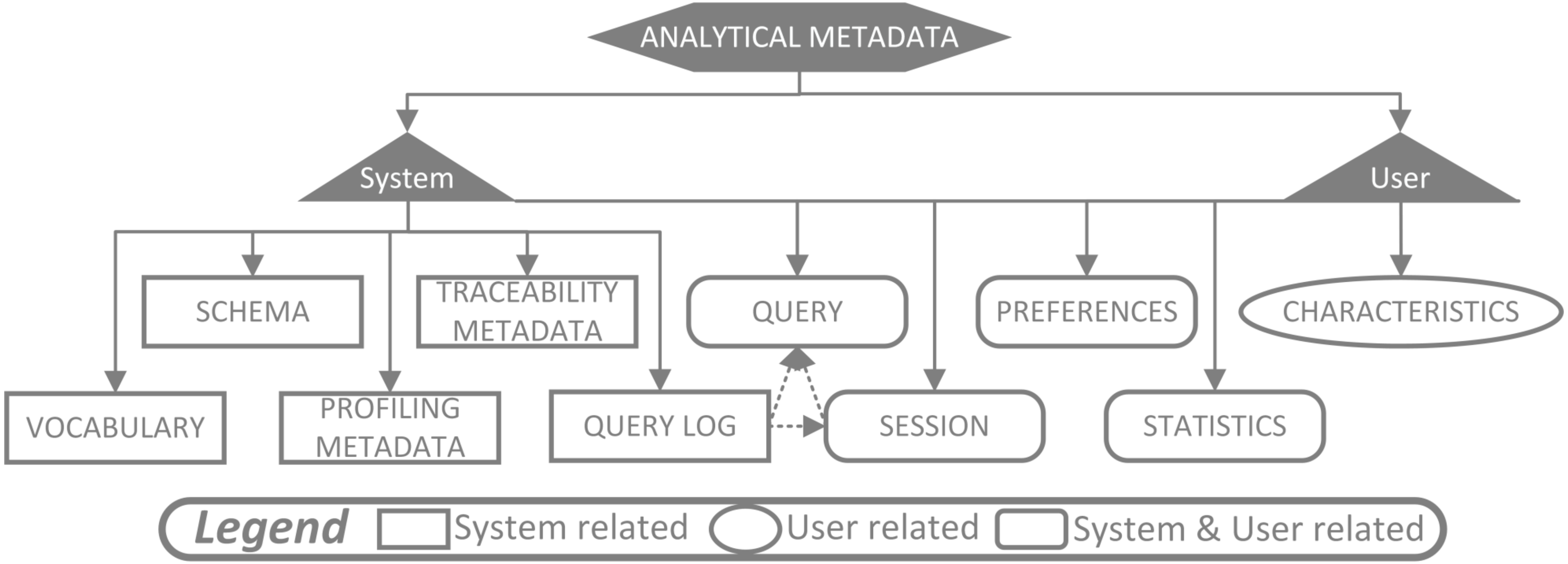
Our Vision



- The novel settings bring data coming from **external and non-controlled data sources**

- Therefore, the metadata need to be:
  - Considered and handled **as a first-class citizen**
  - Designed in a **flexible and reusable** manner
  - The base to support **automation**

Analytical Metadata



ANALYTICAL METADATA

System

User

SCHEMA

TRACEABILITY METADATA

QUERY

PREFERENCES

CHARACTERISTICS

VOCABULARY

PROFILING METADATA

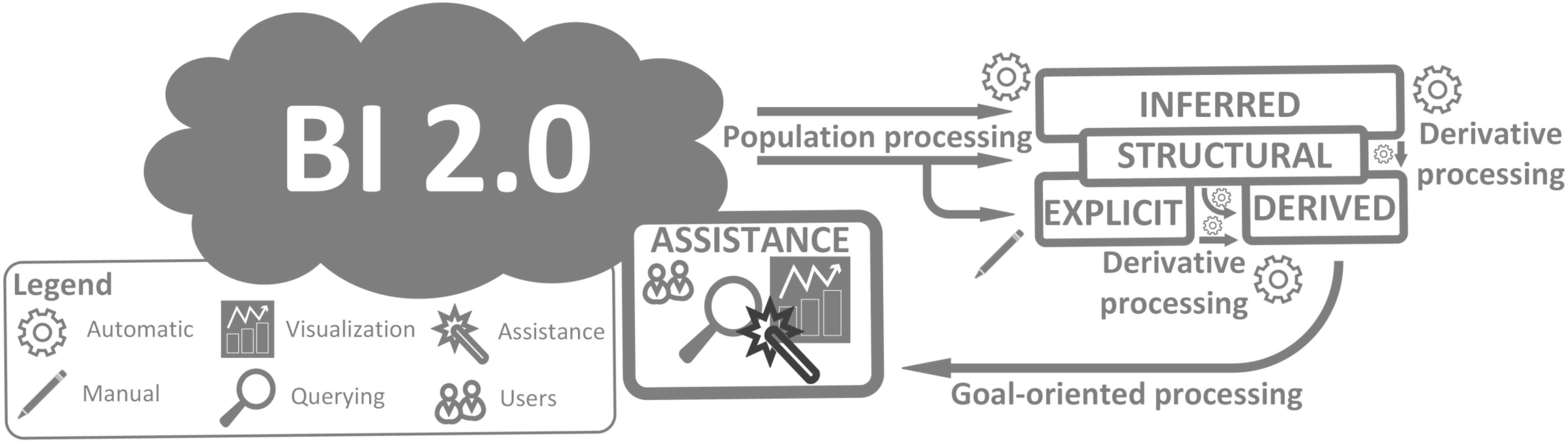
QUERY LOG

SESSION

STATISTICS

Legend □ System related ○ User related ▢ System & User related

Analytical Metadata Processing Types



SM4AM: A Semantic Metamodel for Analytical Metadata

- An **RDF-based** formalization of AM

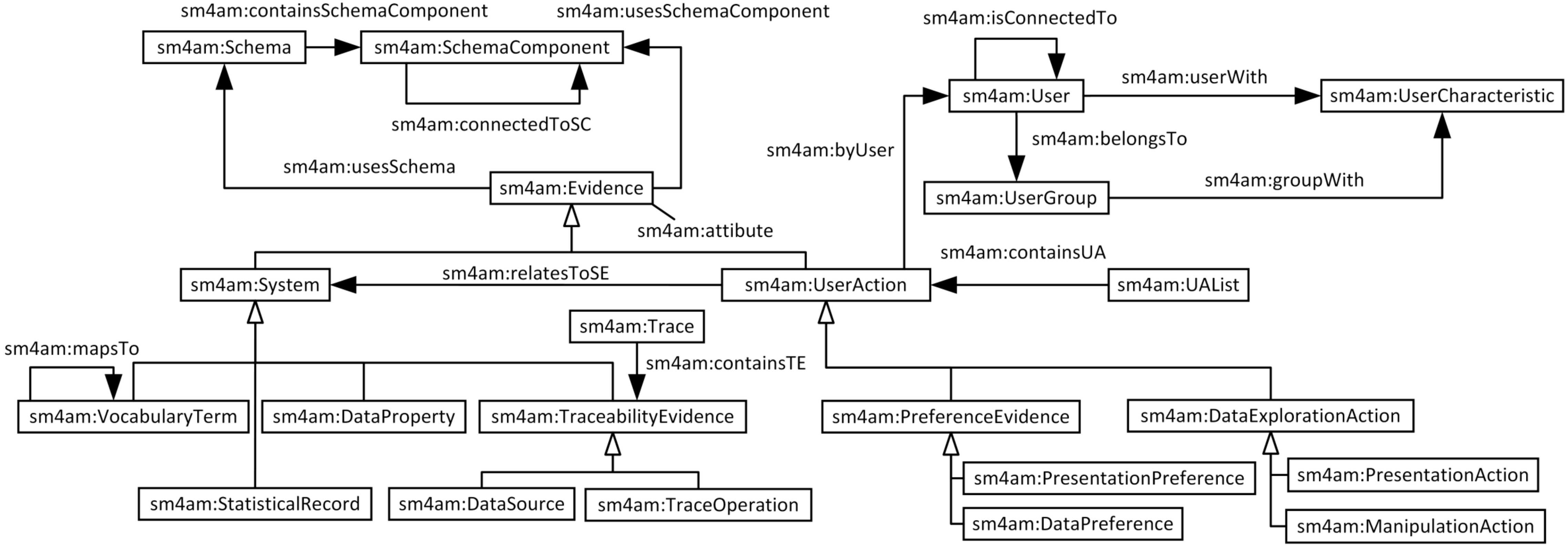
- The metamodel **to overcome the heterogeneity** between different system-specific models

- The use of RDF **provides flexibility** and **reuse** potential

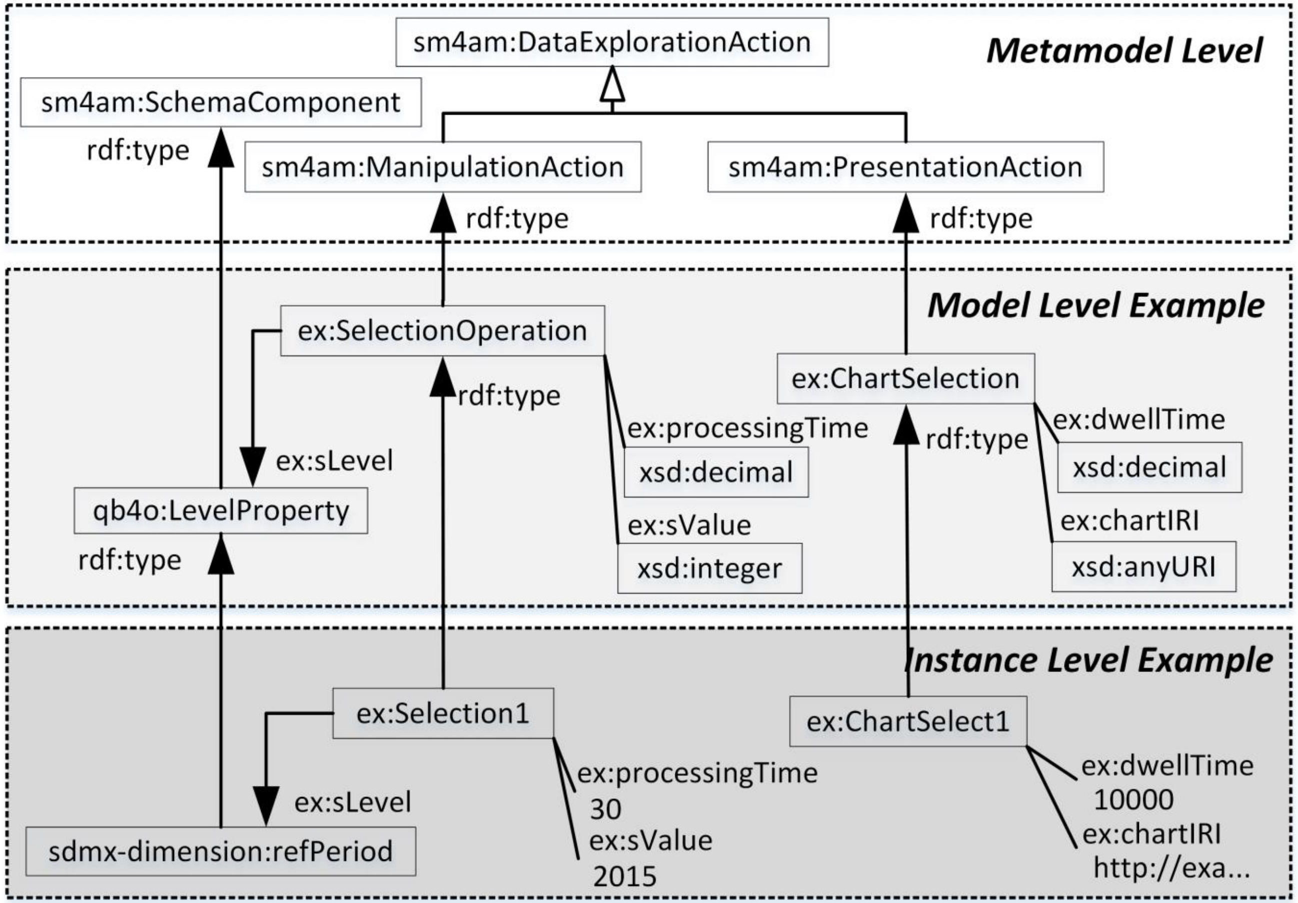
- RDF led us to the **Linked Data** initiative that brings new data available from various source

- The **Linked Data** sources are typically hard for exploration and they can also significantly **benefit** from the use of AM **for the user assistance**

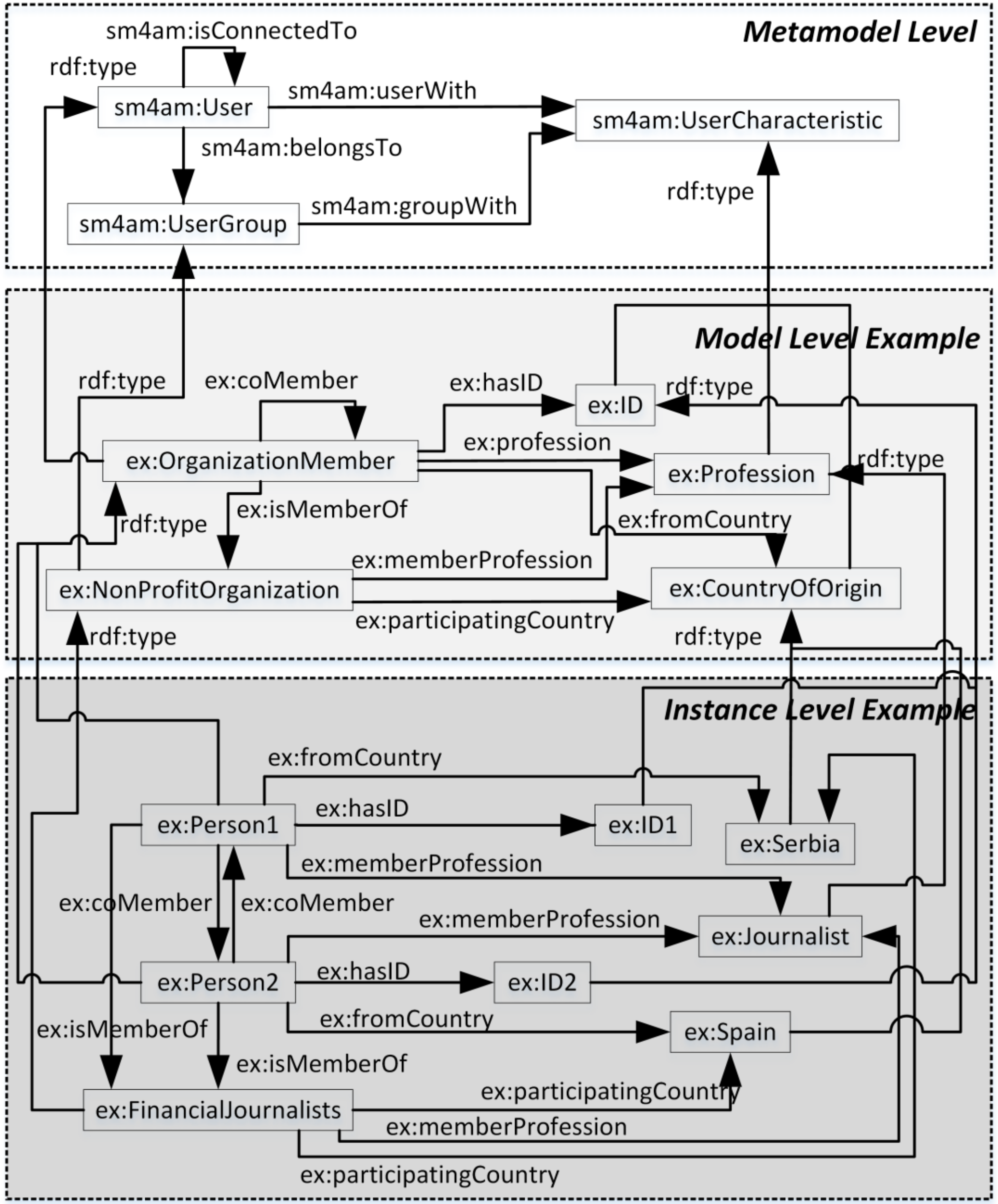
Analytical Metadata Modeling for Next Generation BI Systems



Data Exploration Action Example

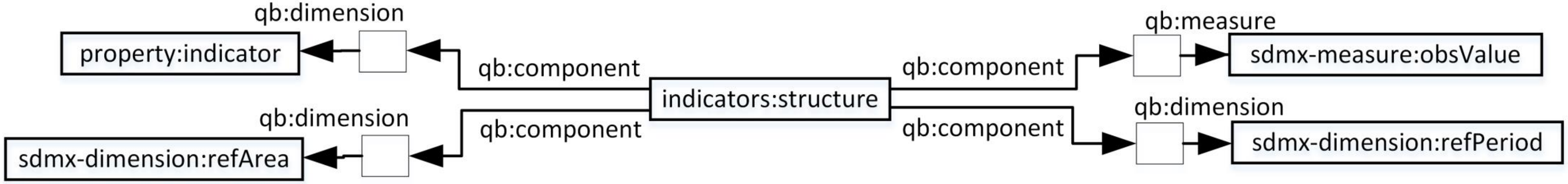


User Example



Running Example:
World Bank Linked Data

- World Bank provides financial and technical support for developing countries around the world and it publishes data about its projects, indicators, about the countries in development, and related information as World Bank Open Data on its website
- These data were used for the creation of the **World Bank Linked Data** (WBLD) data set (http://worldbank.270a.info/)
- We focus on *Market capitalization of listed companies (current US$)* (http://worldbank.270a.info/dataset/CM.MKT.LCAP.CD.html)



On the Quest for the Richer Schema Metadata

- WBLD are represented with the RDF Data Cube (**QB**) **Vocabulary** (http://www.w3.org/TR/vocab-data-cube/)

- The **QB4OLAP** vocabulary (https://code.google.com/p/publishing-multidimensional-data/wiki/VocabularyOutline) extends the QB vocabulary with the additional schema semantics (i.e., metadata) necessary for the OLAP analysis

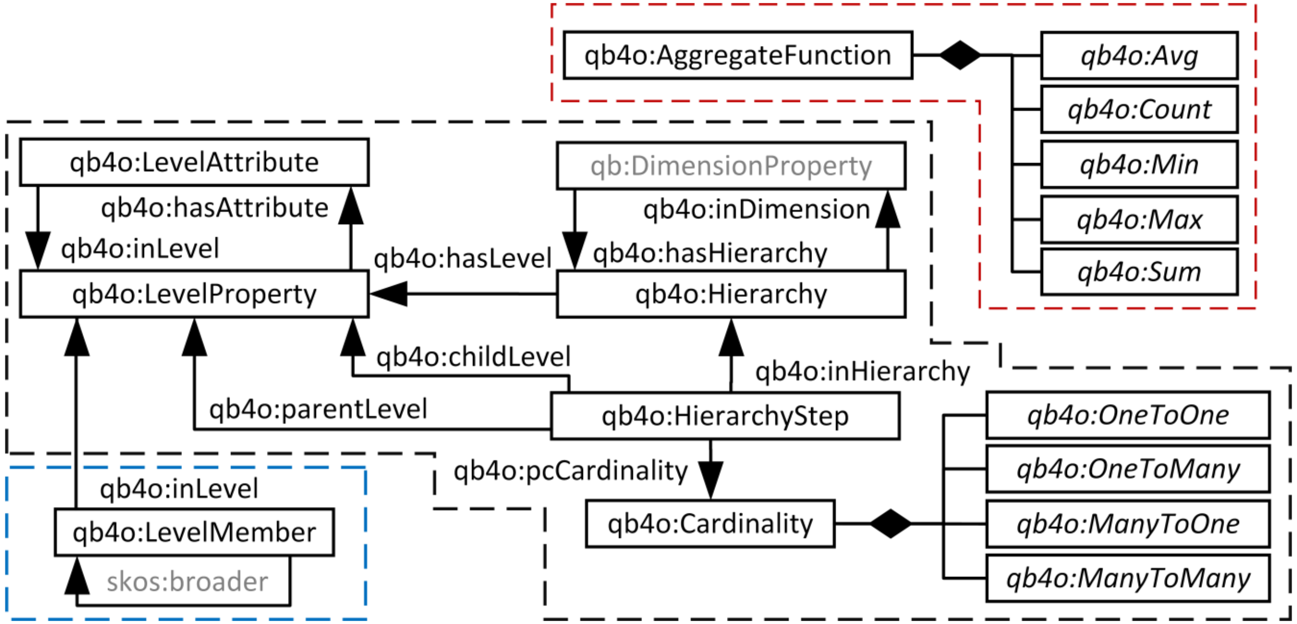
- The challenge we face here is how to enrich the WBLD (or another QB) data set with QB4OLAP

- As a solution, we jointly work with Lorena Etcheverry and Alejandro Vaisman on defining the **methodology** and **tool** for addressing this challenge in a (semi-)automatic manner

From QB to QB4OLAP

- The methodology steps for (semi-) automatic schema enrichment of existing QB data sets with the additional QB4OLAP semantics:

1. Redefinition of the cube schema
2. Specifying the aggregation functions
3. Definition of the dimension hierarchies
4. Annotation of the cube instances



The Enrichment Automation Challenges

- Semantics related:
  - **Partial semantics** – for instance, a data set might lack the information about the aggregation functions that can be applied over measures
  - **Imperfect semantics** - for instance:
    - *Heterogeneity* of semantics where data are defined with different but related semantic concepts
    - *Errors* in semantics
    - *Outliers* in semantics
- Instance related:
  - **Partial data** – for instance, the aggregation value at the month level needs the data about all related days
  - **Imperfect data** – for instance:
    - *Heterogeneity* of data as settings where not all data are formatted well or they do not follow explicit or implicit constraints
    - Data *errors*
    - *Outliers* in data

Addressing Partial Semantics – Aggregation Functions

- The aggregation function definitions:
  - Appropriate **aggregation function definition** depends on the **measure type** and it is a well known challenge in the literature as the concept of **summarizability** in OLAP and statistical databases (H. Lenz, A. Shoshani, *Summarizability in OLAP and statistical data bases*, SSDBM 1997). In that context, measure types are:
    - Stock (e.g. inventory of a product)
    - Flow (e.g., monthly sales value)
    - Value-per-unit (e.g., product item price)
  - The other two conditions are **disjointness** and **completeness**

- **Our idea** is to use the metadata to guide this process:
  - General knowledge metadata (measure type, measure name, and data type)
  - Traced knowledge metadata (statistics)

Addressing Partial Semantics – Dimension Hierarchies

- Typically, the potential multidimensional space is **identified by the functional dependencies** between fact and dimensions and between the levels inside the dimension hierarchy

- Detection by the analysis of RDF data sets where we **benefit from the QB semantics** that is the starting point for discovering possible new dimension levels, hierarchies, and level attributes

- Our goal is to show **the feasibility of this task**

Publications

- Published papers:
  - Jovan Varga, Oscar Romero, Torben Bach Pedersen, Christian Thomsen, *Towards Next Generation BI Systems: The Analytical Metadata Challenge*, DAWAK 2014
  - Jovan Varga, Oscar Romero, Torben Bach Pedersen, Christian Thomsen, *SM4AM: A Semantic Metamodel for Analytical Metadata*, DOLAP 2014
- *Submitted papers:*
  - Jovan Varga, Oscar Romero, Torben Bach Pedersen, Christian Thomsen, *Analytical Metadata Modeling for Next Generation BI Systems*, Information Systems (submitted in May 2015)
- In progress papers:
  - Jovan Varga, Lorena Etcheverry, Oscar Romero, Alejandro Vaisman, Torben Bach Pedersen, Christian Thomsen, *QB2OLAP: An Approach for Enriching QB Data sets with Additional QB4OLAP Semantics*, Journal of Web Semantics (submission planed for the end of July 2015)

Future Work

- Planned publications:
  - Jovan Varga, Lorena Etcheverry, Oscar Romero, Alejandro Vaisman, Torben Bach Pedersen, Christian Thomsen, *A Tool for Enriching QB Data sets with OLAP Semantics*, EDBT conference Demonstrations 2016 (expected submission by 28th September 2015)
  - Jovan Varga, Oscar Romero, Torben Bach Pedersen, Christian Thomsen, *Query Recommendation Based on User-Generated Analytical Metadata Evidences*, SIGMOD 2016 conference (expected submission by 19th November 2015)
  - Jovan Varga, Oscar Romero, Torben Bach Pedersen, Christian Thomsen, *Query Recommendation Meets Next Generation BI Systems*, VLDB conference Demonstrations 2016 (expected submission by 31st March 2016)
  - Jovan Varga, Oscar Romero, Torben Bach Pedersen, Christian Thomsen, *Context-Aware Recommendations for User-Centric BI Systems*, ICDE 2017 conference (expected submission by 5th August 2016)

Thank you!

Questions?

JVARGA@ESSI.UPC.EDU