

Research Project on Metadata Extraction, Exploration and Pooling

Ronald Steinhau
Entimo AG - Berlin/Germany

Content

- Motivation
- Project Setup
- Workflow
- Technical Base
- Project Status
- Use Cases

Industry Challenges | Motivation

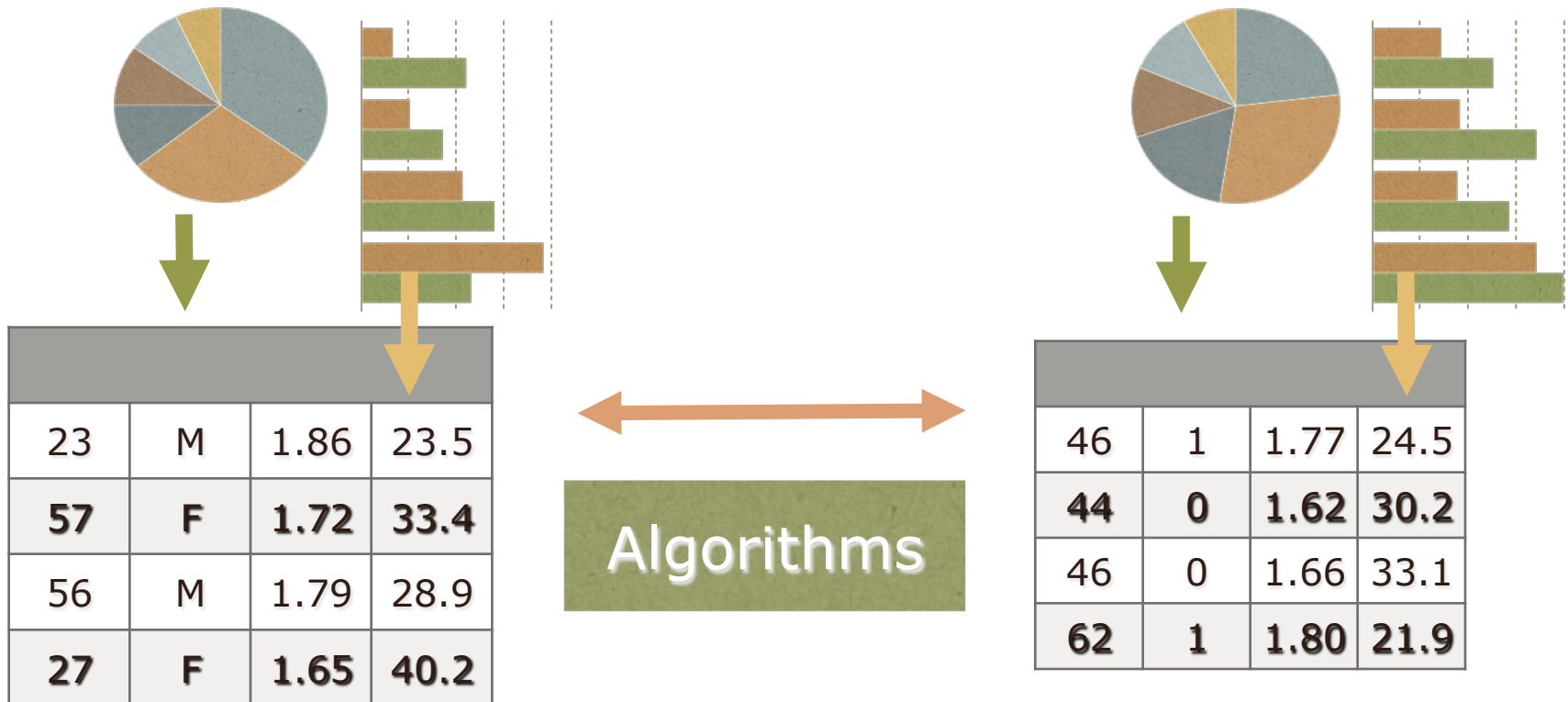
- Mapping of legacy studies to CDISC:
 - Insufficient and hardly retrievable specification
 - Knowledgeable experts might be not available anymore
 - High costs of mapping rule restoration and metadata creation
- Wave of Mergers and Acquisitions:
 - Study structure harmonization required
- CDISC Share Pilot Report, Conclusions:
 - Manual terminology harmonization possible for short lists
 - Technological support required for search, viewing and comparison of numerous, long lists

Project Setup

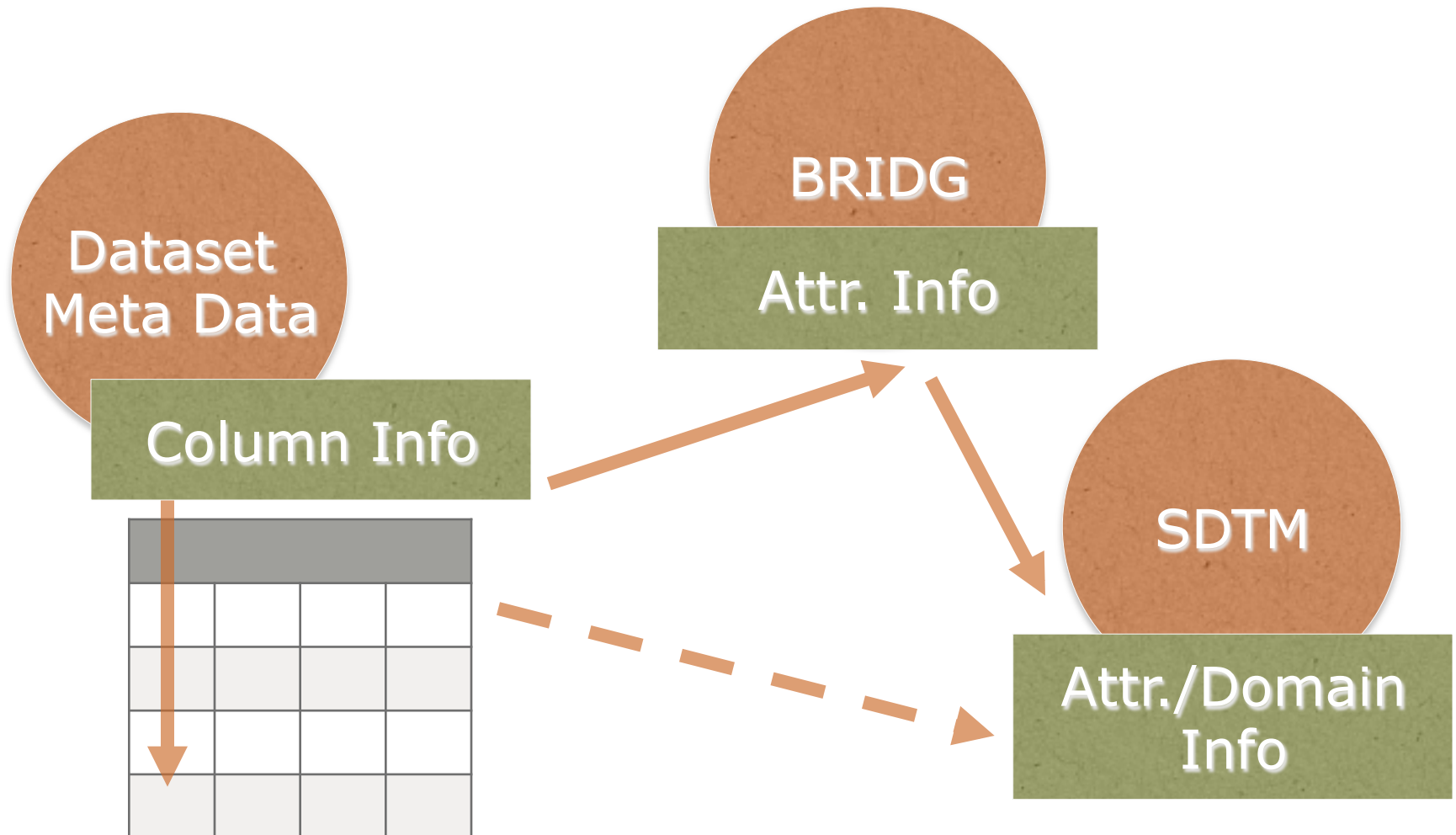
- Research funding from the German Federal Ministry of Economy and Technology
- High Level Goals:
 - Investigate mechanisms for automatic recognition, extraction of metadata
 - Allow context- and dependency-based metadata exploration
 - Include different types of metadata into analysis
 - Re-use clinical data and metadata from comparable studies managed in the central information pool
 - Provide support in specification creation and issue resolution > **Full automation is illusionary!!!**

Example: Dataset Profiling

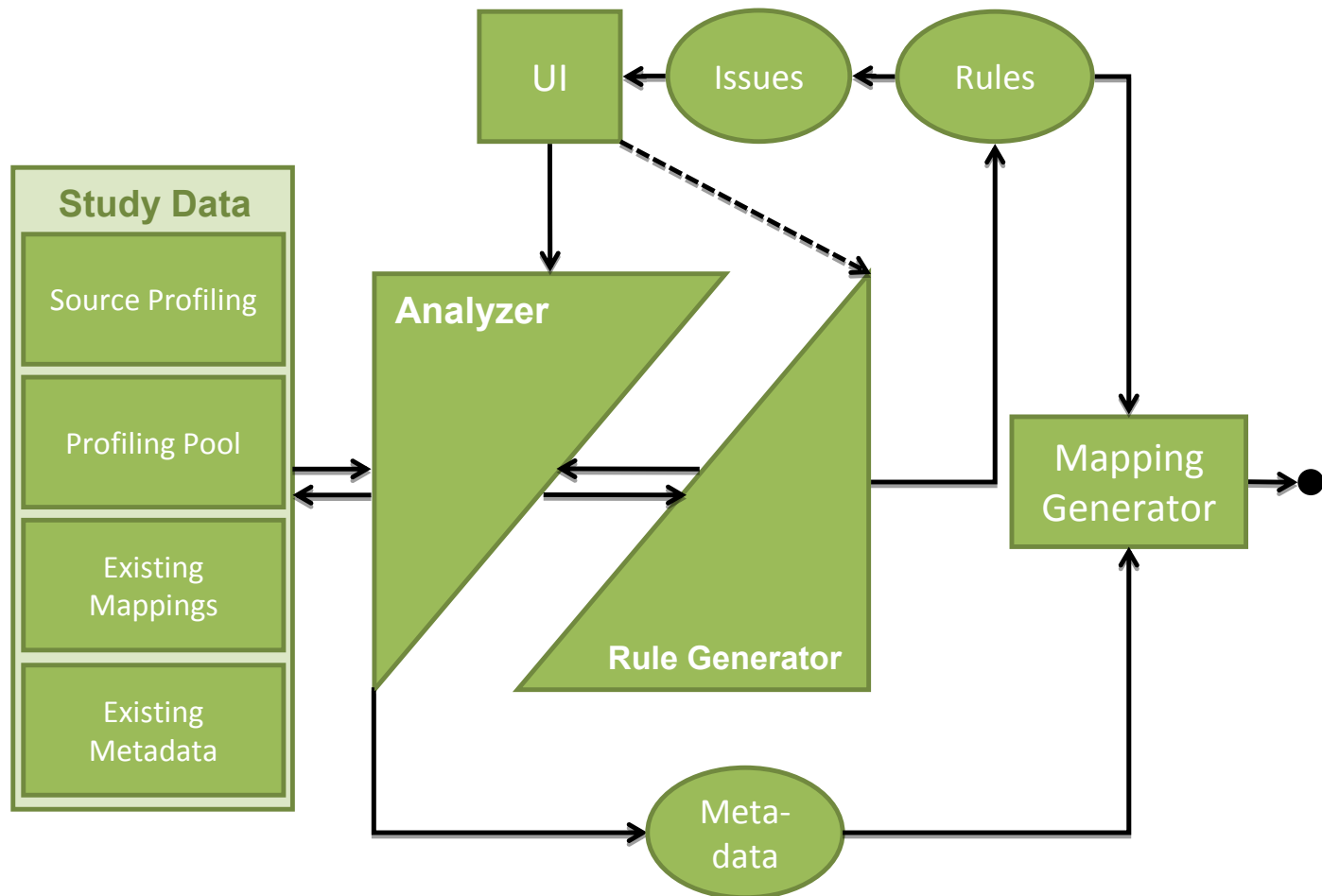
- Search for existing mapping algorithms based on the metadata and profiling results



Example: Linked Metadata



Workflow



Metadata Extraction

- Information Collection / Source Profiling
 - Name patterns
(e.g. variable names, labels, comments)
 - Statistical analysis of values
 - Analysis of formats and associated values
 - Cross-reference analysis, relationship analysis between entities

Analysis: Used Principles and Entities

- Existing profiling pools and metadata library (existing studies)
- Existing mappings (and associated data)
- Similarity detection based on probabilities (e.g. Bayesian), fuzzy logic, categorization, clustering...
- Generation of a “minimum set of questions to a study”
- Profiling, its results, generation of issues is steered by the rules managed in the rule engine
- User input increases the detection grade and facilitates adaptive learning – **Maintenance!**

Rule Generator

- Rule catalogs created per target domain or on a cross-domain basis
- Domain-specific language (DSL) for the formulation of rules and profiling issues

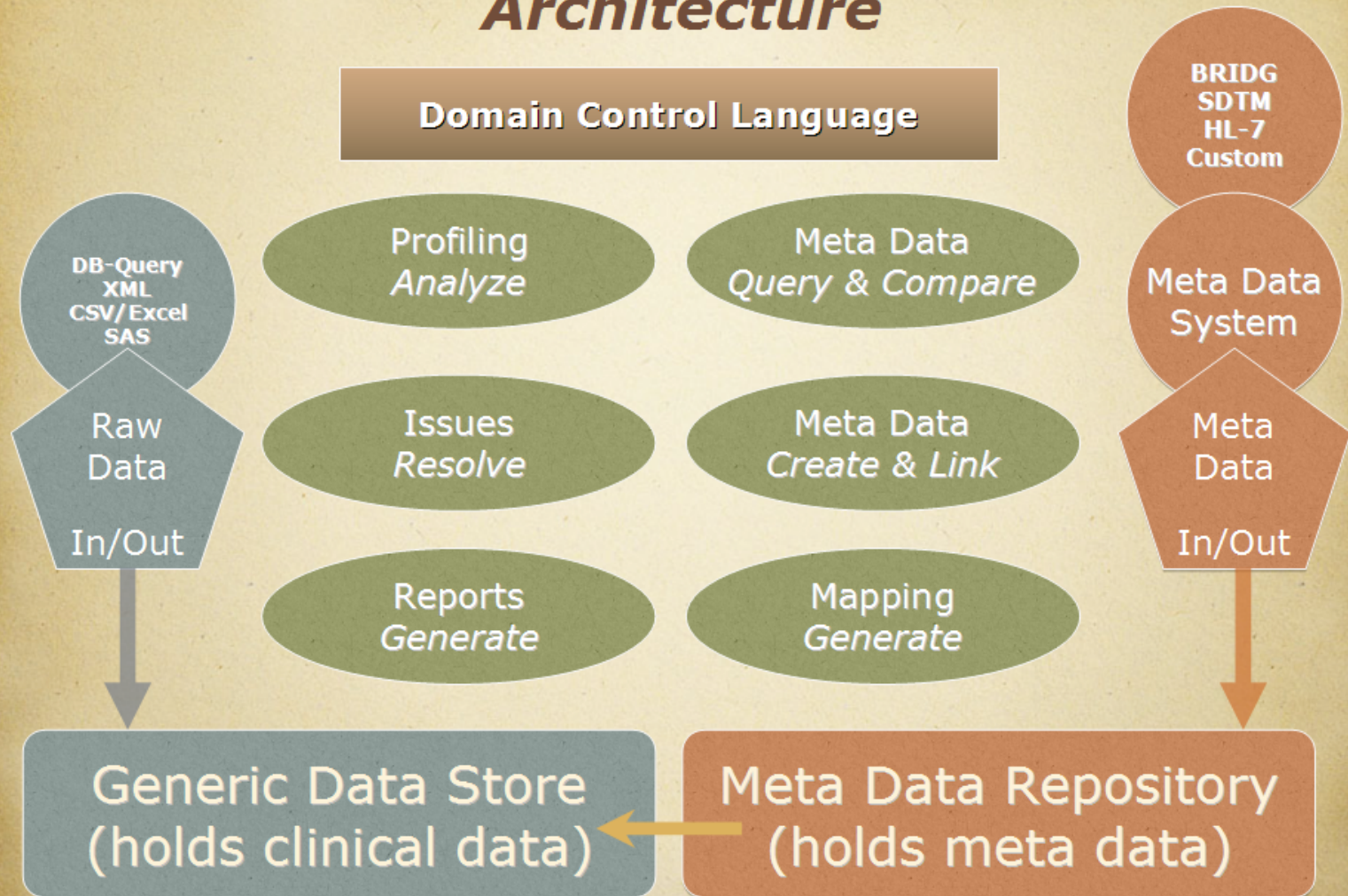
Metadata Visualization and Exploration

- **Cross-model capability with model tagging!**
- **Assignment of contexts and relating metadata from different contexts**
- **Reporting capabilities**
 - Metadata of a selected entity (e.g. dataset or data column) represented in one or more metadata contexts
 - Relations between metadata in different contexts
 - Relation of different metadata types (code lists, domain definitions, mappings...)
 - Context switch
 - Available mapping specifications in different layouts
 - Specification export in formats optimized for human readability and automatic post-processing

Technical Base

- Model-driven approach - Eclipse Modelling Framework (EMF)
- XText for DSL creation and grammar definition
- Rule systems:
 - JTP (Java Theorem Prover) - object oriented "Reasoning System"
 - JFuzzyLogic library for the implementation of fuzzy logic aspects and Bayesian classification
 - Other forward chaining and backward reasoning rule systems
- Open connectivity principles to exchange data, metadata, specification with external systems

Architecture



Use Cases

- **Extract metadata by comparison/rules**
- **Find similar mapping strategies**
- **Harmonize codelists and terminology**
- **Generate mappings sequences**
- **Benefits:**
 - Consistency: Non-redundant and non-ambiguous metadata definition
 - Massive time savings

Project Status

- Infrastructure in place: DSL, Modules, Repositories
- Feasibility studies and pilots of modules successful
 - Multiple metadata and raw data importers and extractors implemented
 - Evaluation and selection of principles ongoing
 - Cross-model metadata navigator/explorer growing

End

Many thanks for your
attention!
Questions?