

Research Project on Metadata Extraction, Exploration and Pooling: Challenges and Achievements

Ronald Steinhau

(Entimo AG - Berlin/Germany)

Content

- Project Goals
- Pre-Requisites
- Work Packages
- Advanced Workflows
- Conclusions and Outlook

Project Goals (1)

- Main Goals
 - Support different metadata systems
 - SDTM, ADaM, BRIDG, custom
 - Explore items dependent on contexts
 - Accelerate mapping process
 - Re-use information from comparable studies
 - Provide support in specification creation and issue resolution (full automation is illusionary)

Project Goals (2)

- Additional Goals
 - Immediate usage and classification of metadata
 - Advanced metadata management based on ISO 11179 for Metadata Repositories
 - Cross-linking between MD-Systems incl. terminology/codelists
 - Smart search and recommendation of attributes and mappings
 - Preserve history of user decisions after recommendations

Work Packages

1. Development Preparation
2. Specification / Modeling
3. Development
4. Test & Optimizations

Development Preparation

- Development Environment
 - Eclipse Helios / Scala IDE
- Advanced Libraries
 - Statistical analysis
 - Machine (“adaptive”) learning
- Infrastructure - Clinical Repository
 - Based on relational database
 - Fully generic tables (free schema)
 - Fast, minimal redundancy
 - Audit trail, versioning, SAS compliance

- Missing Values
- Codelists
- Formats



Specification / Modeling

- Metadata management & rules
- Data analysis
- Smart recommendations & history usage
- Finding and applying mapping specs
- Mapping / meta generator

Specification / Modeling (1)

Example Workflow: Import Clinical Data

- Analyze Data
 - Analyze data and retrieve statistical profiles
 - Extract all available metadata/data attributes:
 - Name (synonym support)
 - Label / Comment (Google like searches)
 - Profiles (statistics based searches)
 - Codelist analysis (context sensitive)...
 - Save all data in the clinical data repository
 - Save meta-information in the metadata repository
 - Keep links between data and metadata

Specification / Modeling (2)

Example Workflow: Import Clinical Data

- Provide recommendations:
 - Data types and their type length
 - Primary keys
 - Code lists
 - References to existing metadata (SDTM, BRIDG, custom)
- Find attributes used in mappings
 - SDTM/custom domain memberships
 - BRIDG references

Example: Schema Recommendation

Import Clinical Repository data set from SAS

Edit schema.
Edit schema information required for dataset conversion.

Source Path:

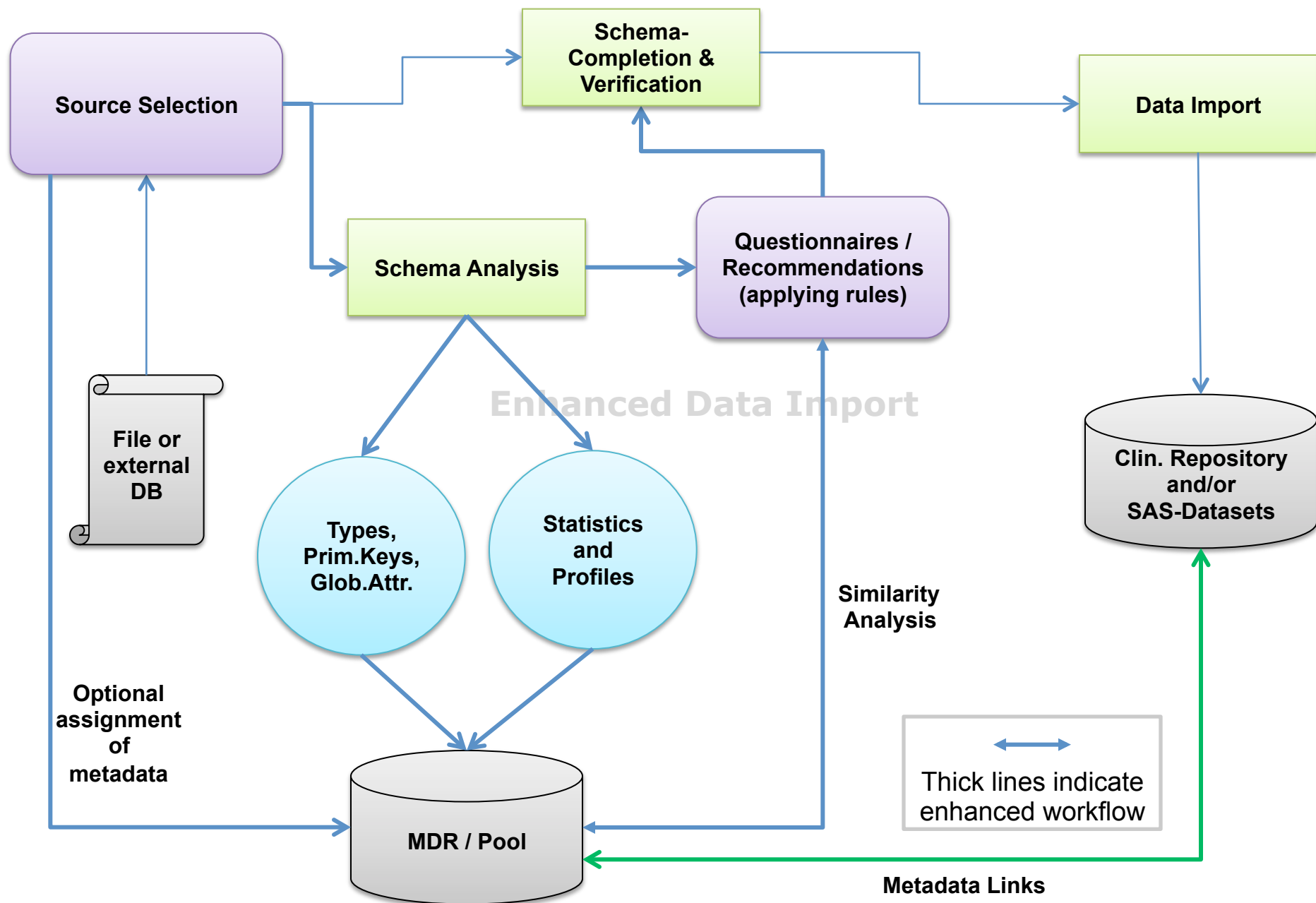
Meta Source Path:

Position	Name	Label	Type	Use Analyzed Type	Length	Default	Audit Required	Hidden	Not Null	Primary Key	Audit Info	Format	Codelist
1	study		Integer	☒	8		☐	☐	☐	☒	☐		
2	ptno		Integer	☒	8		☐	☐	☐	☒	☐		
3	recno		Integer	☒	8		☐	☐	☐	☒	☐		
4	ae_sd		Integer	☒	8		☐	☐	☐	☐	☐		
5	ae_sm		Integer	☒	8		☐	☐	☐	☐	☐		
6	ae_sy		Integer	☒	8		☐	☐	☐	☐	☐		
7	ae_ed		Double	☐	8		☐	☐	☐	☐	☐		
8	ae_em		Double	☐	8		☐	☐	☐	☐	☐		
9	ae_ey		Double	☐	8		☐	☐	☐	☐	☐		
10	aespec		Varchar	☐	10		☐	☐	☐	☐	☐		
11	llt_code		Integer	☒	8		☐	☐	☐	☐	☐		

Analyze data in order to retrieve more specific type information. Otherwise all numeric variables will be converted as data type 'double'.

Check availability of codelists before selecting them for conversion. ☒ update codelists ☐ replace codelists

< Back Next > Run Now Run Batch Run Scheduled Cancel



Mapping / Meta-Generator

- Finding mapping specifications
 - Find and recommend existing mappings
 - Support users with the completion (modification) of copied mappings
 - Tag mappings with metadata for smarter recognition
- Applying mappings
 - Generate mapping programs
 - Execute mapping programs with data

Conclusions

- Providing “smart” technical infrastructure is challenging, but necessary for complex systems
- Once in place, positive effects with growing usage and stored content
- Interconnected metadata systems and data provide better transparency and reusability
- Contextual knowledge (e.g. drug, study) leads to improved results

Outlook

- Define more metadata inter-connections
- Collect time saving statistics with larger studies
- Deeper Integration into entimICE

Embrace the new principle
“analyse → recommend → re-use”!

End

Thank you for your attention!

Questions?