University of Piraeus
Department of Digital Systems

A.I. GROUP
Semantic Web & Multiagent Systems

Migrating Data to CRIS

Methods, Issues and Challenges

George Santipantakis

Antonis Koukourikos

George Vouros

DATA STORIES

Overview

- ➢Objective & Requirements
- ➢Migration Process
- ➢Data Sources Migrated
- ➢Major Issues
- ➢Challenges ahead

Objective & Requirements

Migration of multiple heterogeneous data sources to the CRIS repository maintained by NDC. Towards this goal, we have implemented a configurable tool and propose a semi-automated migration methodology.

This task requires:

- *The specification of mappings between heterogeneous data sources and CRIS*
- *duplicate entries detection and cleaning*
- *validity check in data by value (e.g. validation of electronic address)*
- *Normalization of data (e.g. contact information given in a single field in source, postal and electronic adresses in CRIS)*
- *Value conversions (e.g. unixtime to DateTime)*

Objective & Requirements

Also required:

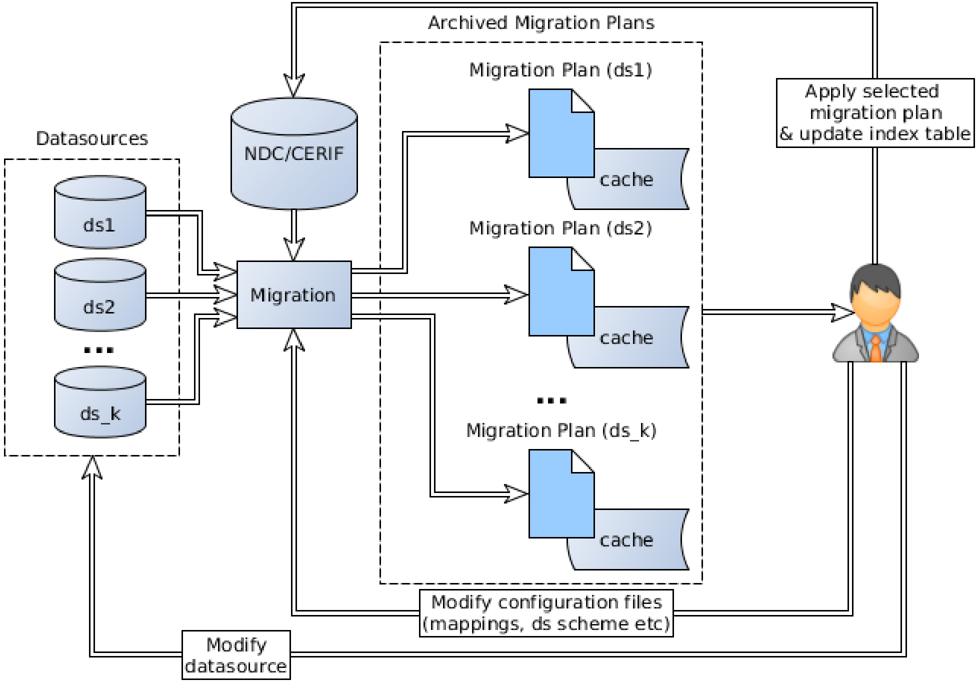
- *The construction of reports on validity and duplicate checks in data sources.*
- *A domain expert user for reviewing validity check reports and reconfigure the tool (or sources if possible).*
- *The construction of a migration plan for confirmation by a domain expert user.*
- *Humans should be in the loop. This is a challenge itself.*

Implemented Migration Workflow

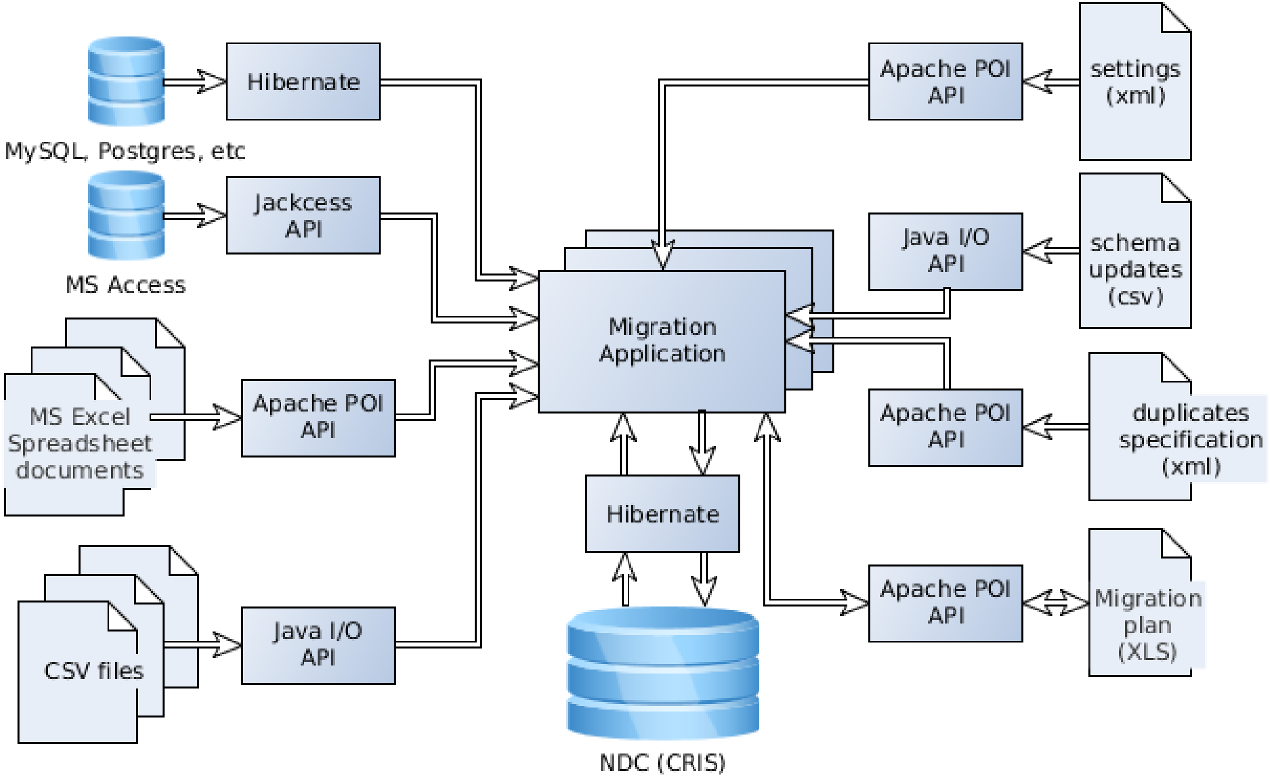
1. Construction of a *migration plan* for each source data collection

2. Review and approval of the actions in a migration plan by a domain expert

3. Execution of the migration plan over the target CRIS database



The migration tool



Mapping Specifications for Migration Tool

Mappings can be provided in a text file specifying:

a. what entity should be created
b. the source file/table to be processed
c. the field in the source table and the corresponding table and field in CRIS database
d. the value in the source field can be also used as argument in a function
e. The specification allows joined tables, If more than one tables in the source are involved for an entity.

Mapping Specifications for Migration Tool

Example (Join tables):

- OrgUnit “TableA".“V.A.T.2"=“TableB"."AFM"

Example (use of functions):

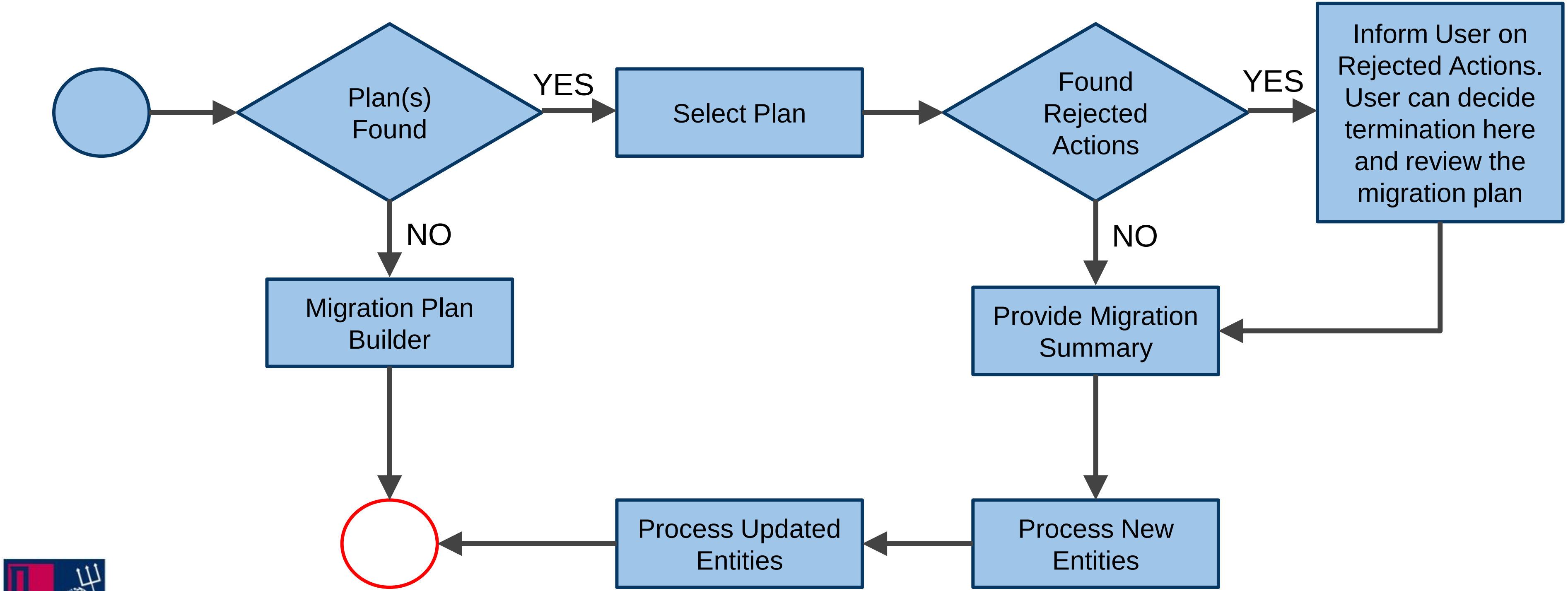
Person	Researcher	fldName	Cfpername	cfname	#getname($v)
Person	Researcher	fldName	Cfpername	cffamilyname	#getFamilyName($v)

User Review & Migration Entry Validation

The migration plan contains:

- the identical entities found (no changes to be made in CRIS)
- the new entities to be created in CRIS (requires user confirmation)
- the updates on existing entities (requires user confirmation)

Migration Process



Data Sources Migrated

We have migrated data from the following sources (table is indicative, entries for other entities such as fund, equipment, etc have been also created):

- ➢ *National Strategic Reference Framework (NSRF)*
- ➢ *Greek Research and Technology Network (GRNET)*
- ➢ *Seventh Framework Programme (FP7)*
- ➢ *Horizon 2020 (H2020)*

Data Source	Projects	Organisation Units	Persons
H2020 Projects	1410	4873	-
H2020 Proposals	25903	30199	-
FP7 Projects	25238	29352	81569
FP7 Proposals	158562	55612	41407
GRNET	5007	1456	8523
NSRF	232556	34153	11068

Major Issues Addressed

- **Heterogeneity of sources**: a variety of data connectors is supported and configurations had to be implemented (CSV, XLS, MDB and JDBC)
- **Entity type indentification**: Heuristics have been applied for some sources, even for distinguishing a person from an organization (these were not distinguished in the schema of the source)
- **Duplicate Detection**: Some times the same entity has two or more representations in the source. The tool can identify duplicates, w.r.t. user defined detection rules for each entity.
- **Versioning and Integrity**: The implemented solution records changes in the database per user to enable tracking of changes

Challenges Ahead

Size of the datasets is a part of the problem itself:

- Thousands or Millions of entries in the sources
- Difficult to process using conventional data processing applications
- State of the art scalable solutions for partitioning data must be used to make the whole process much more efficient

Further more:

- automation in the process for deduplication, and computation of similar entities
- putting humans in the loop so as to ease the validation of results and impose further requirements to the process (e.g. refine deduplication, or similarity checks)
- making the tool as generic as possible to be configurable for new data sources.
- open the data and linking them with open data sources.

Thank you