euroCRIS Strategic Membership Meeting 2024
Paris, 26 - 28 November, 2024

Supervised research information import into CRISs from global sources

Ivan Mršulja (University of Novi Sad, Serbia)
Miloš Popović (University of Novi Sad, Serbia)
Dragan Ivanović (University of Novi Sad, Serbia)

Agenda

- Motivation
- The TeslaRIS's implementation
  - Architecture
  - Implementation
- The further plans

Motivation

Motvation

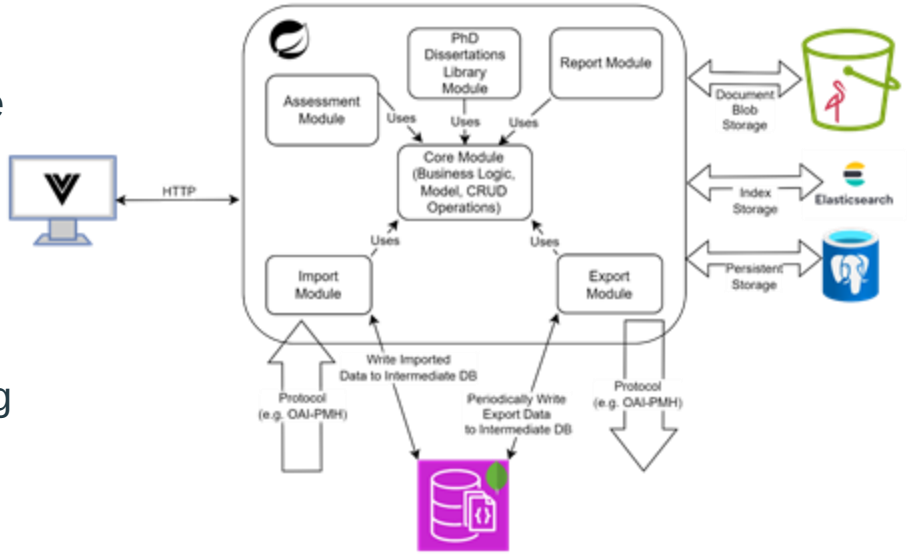
- Global platforms like Scopus provide partial data about researchers' publications, but these are insufficient for complete institutional bibliographies.
- Missing persistent identifiers in global sources lead to mislinked entities (authors, journals, conferences) during automated imports, risking data quality.
- Manual validation by researchers can ensure accurate linking of entities and metadata quality, even in the absence of unique identifiers.
- Existing tools often lack flexibility for use across multiple CRIS platforms or integration with diverse global sources like OpenAlex.

The TeslaRIS's implementation

Modern architecture

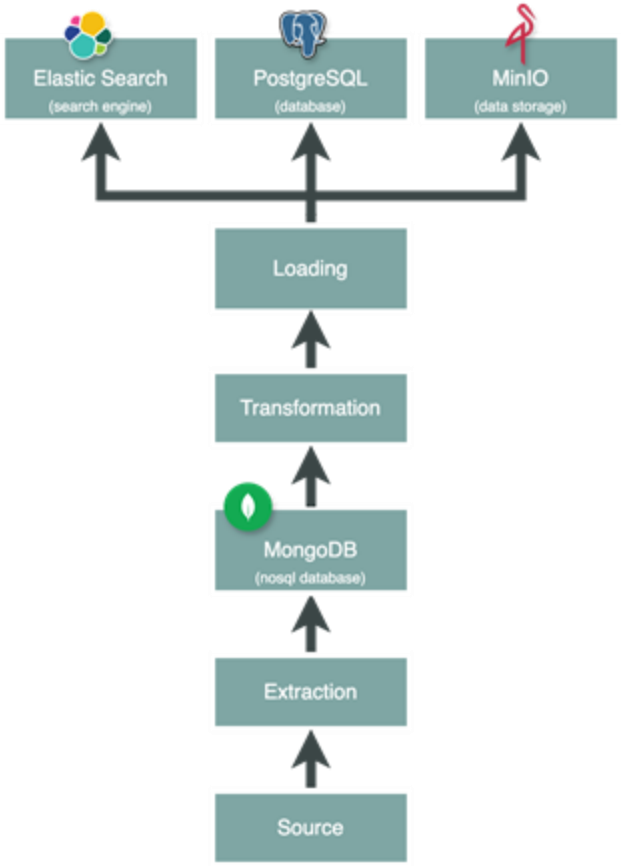
- TeslaRIS is featuring a modular, scalable architecture built on the Spring Boot framework.

- Data flows smoothly between modular components, allowing for easy expansion and updates as needs evolve.



Import module

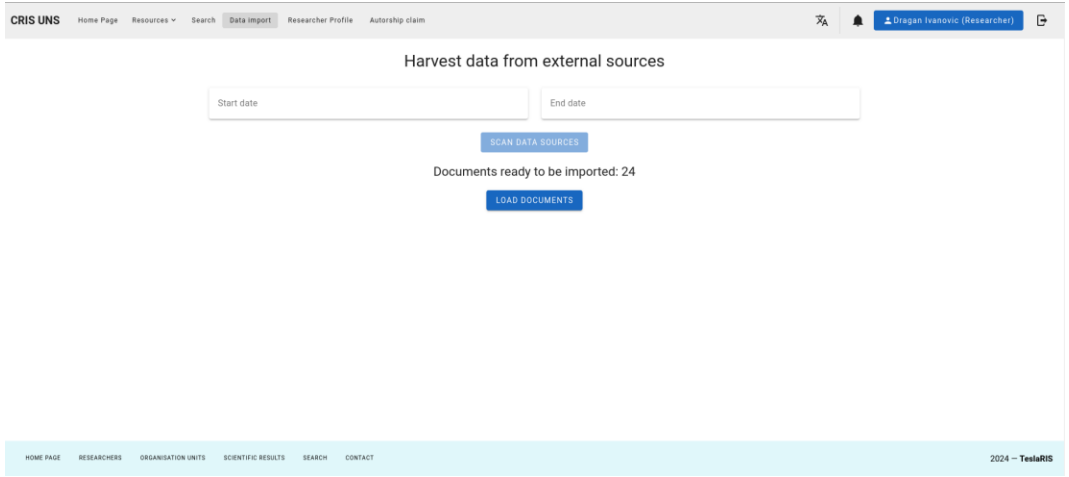
- Implements a robust ETL process, ensuring that data is accurately harvested and integrated. It supports various data formats and interfaces (REST APIs, OAI-PMH endpoints, etc.)
- An intermediate MongoDB database is used to temporarily store harvested data, allowing for necessary data cleaning and transformation before final integration.
- Platform enables re-scanning of sources. An AI model is used to detect whether new information is available on the already scanned records.



The import flow – scanning the sources

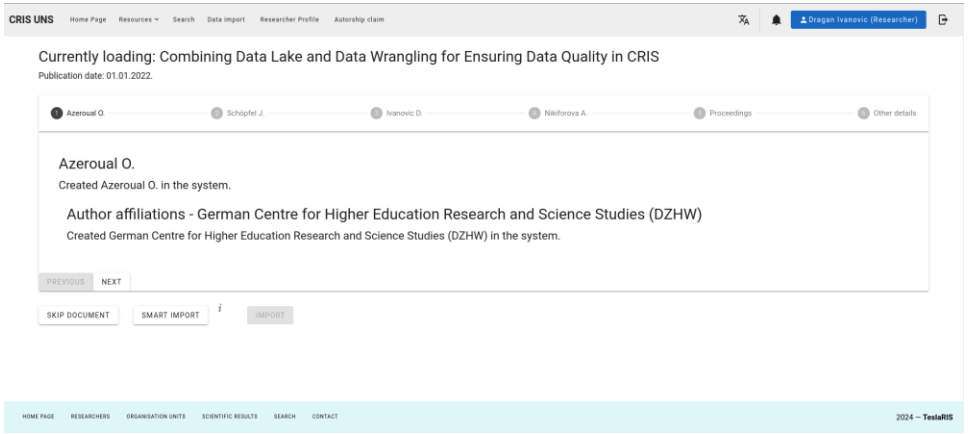
- Once user sets up necessary identifiers on his profile page, the data import becomes available

- User can re-scan his publications at any time for potential updates

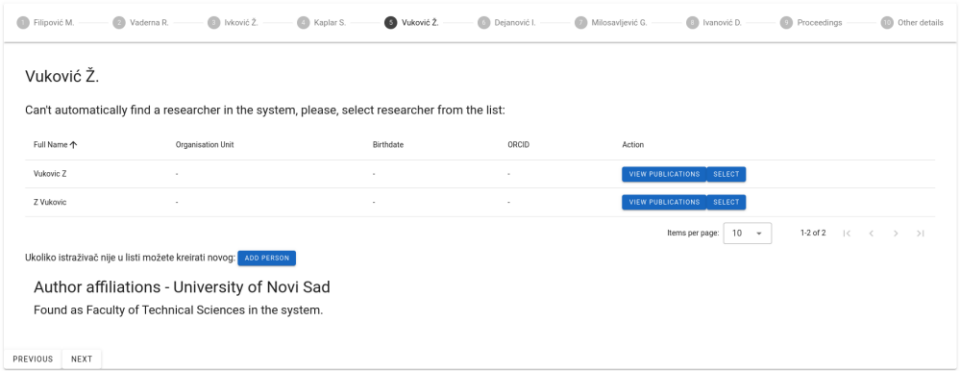


The import flow – author information

- The system tries to perform this step automatically using the import platform's unique identifier (e.g. Scopus Author ID, ORCID ID etc.).



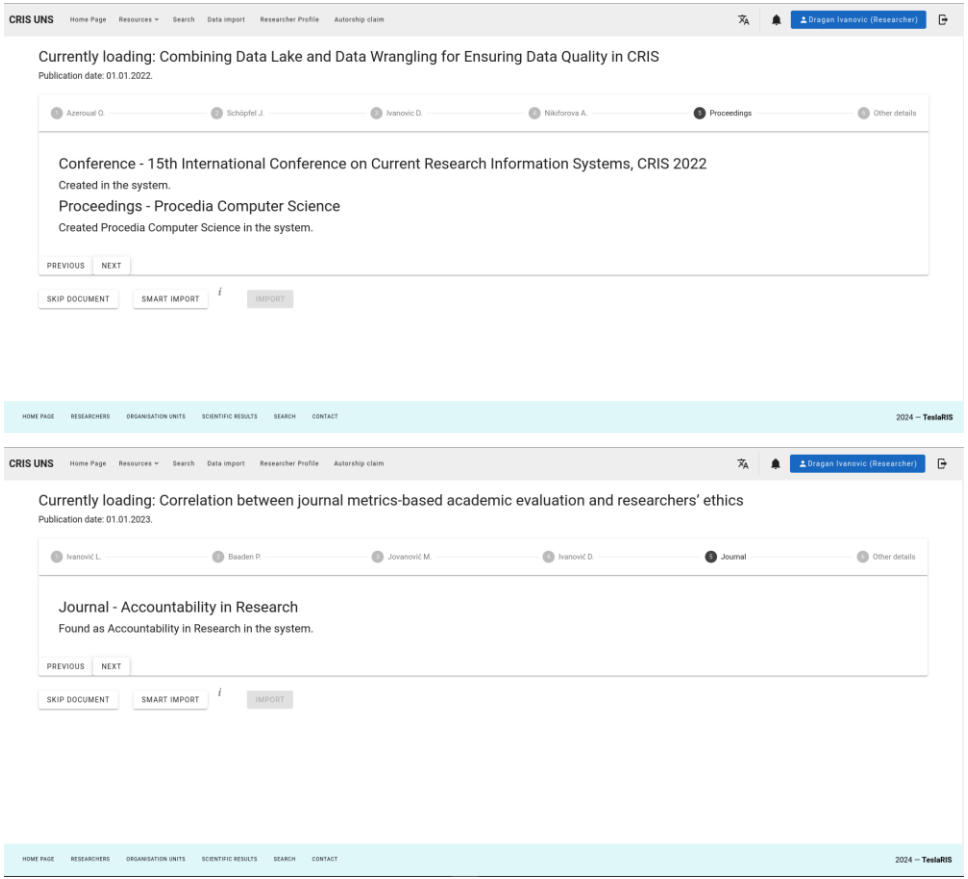
- If it fails to do so, our database is queried by author name and the user is given a list of potential candidates to manually choose from.



The import flow – event / journal

- The same process is repeated for authors' affiliations and journal/proceedings that the publication is published in.

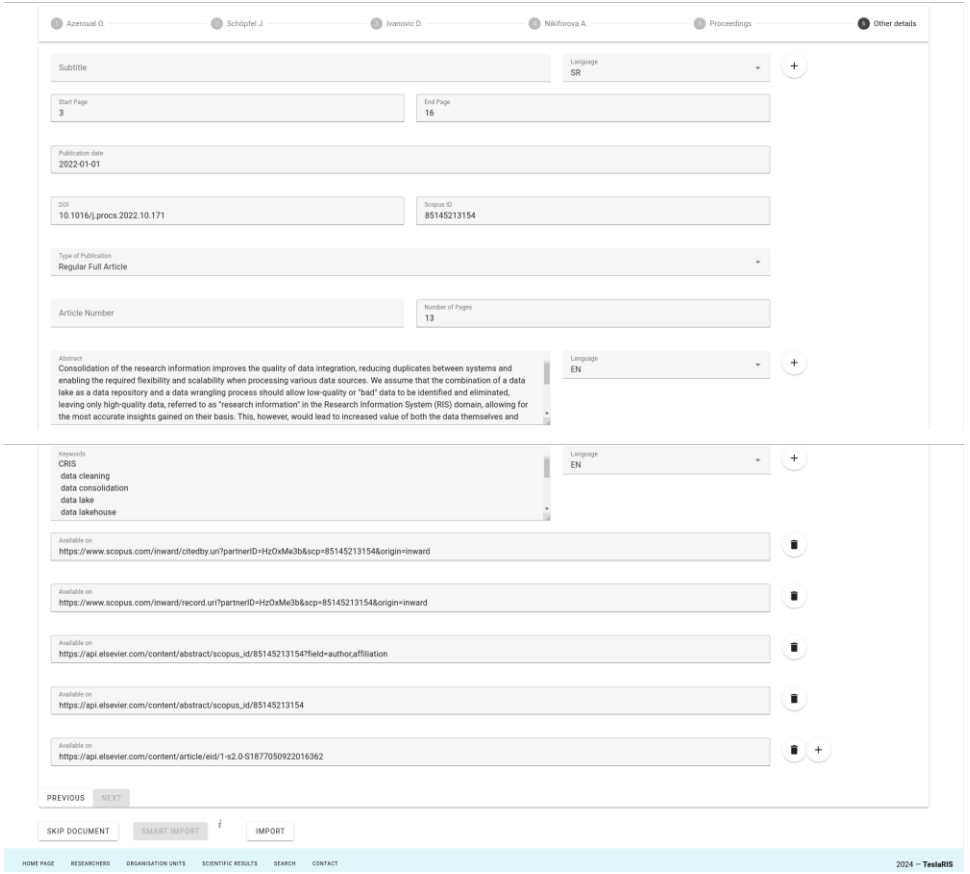
- The platform's unique ID will be automatically bound to chosen/created entity, ensuring that next time this process is completed automatically by the system.



The import flow – publication details

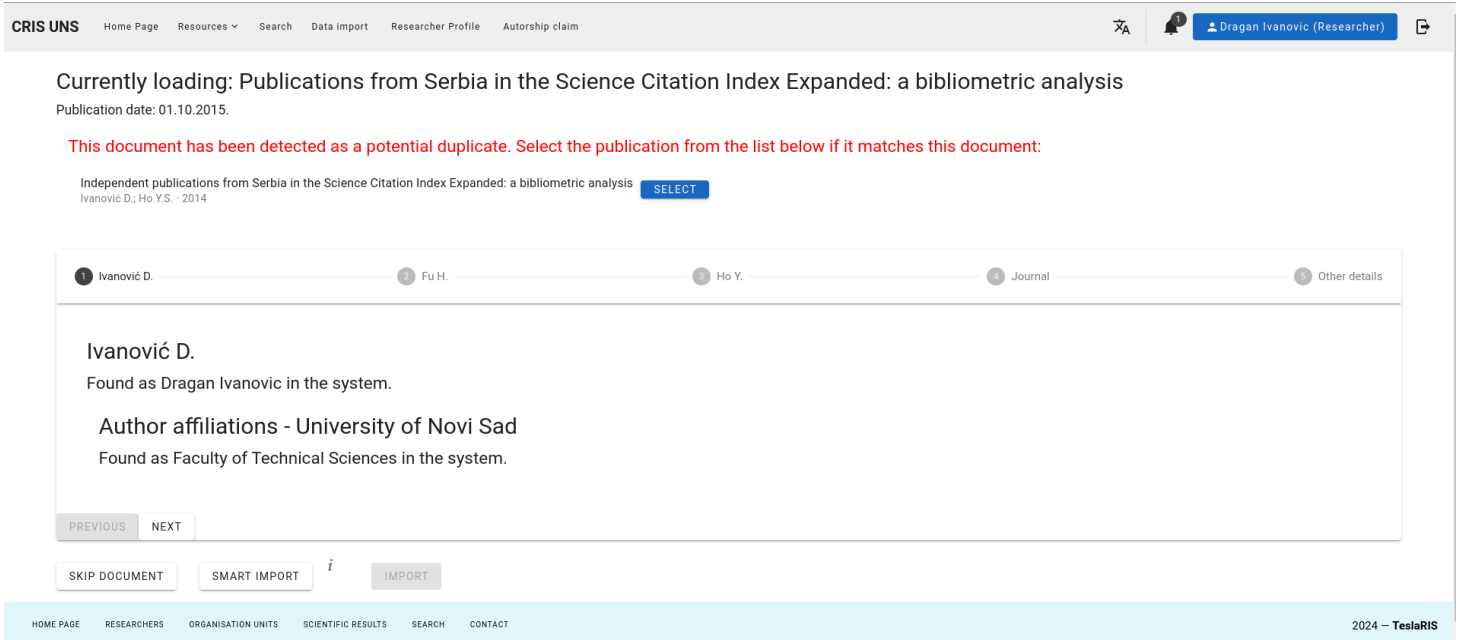
- In the end, the user is presented with other publication metadata, such as description, keywords etc., so he/she can edit all other details about the publication.

- Submitting this last form means that publication will be successfully loaded into our system and the user can continue to the next one



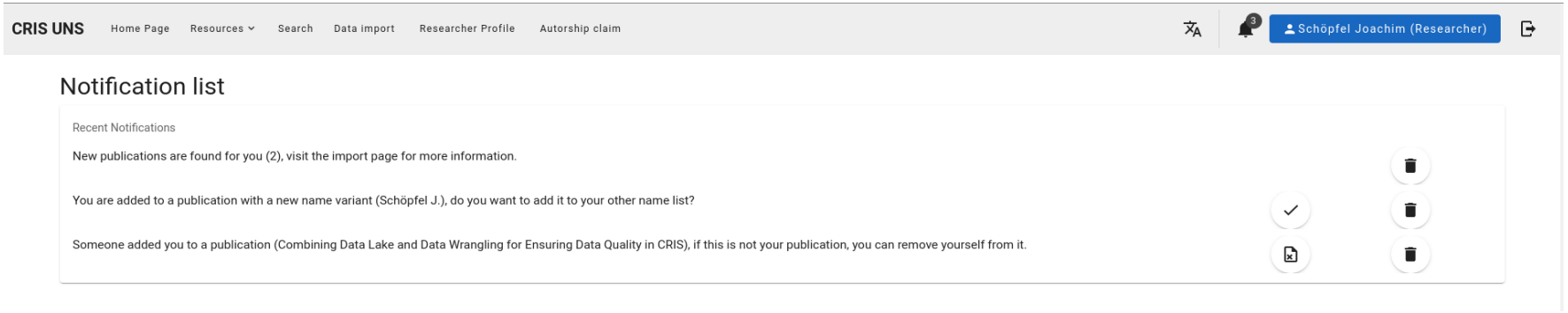
The import flow – built-in deduplication

- Leveraging Elasticsearch's search capabilities, deduplication is performed during the import.



The import flow – side effects

- Every author can import publications, regardless of who performed the initial harvest
- Every author receives notifications when they are added on new publication and if they are added with different name



Further plans

Development plan

- Integrate new data sources (e.g. OpenAlex, WOS...)
- Perform testing with bigger user group to identify potential flaws and further room for improvement