

Supervised research information import into CRISs from global sources

Authors

Ivan Mršulja¹ (0000-0002-5376-4106), Miloš Popović¹ (0009-0007-2938-3673), Dragan Ivanović¹ (0000-0002-9942-5521)

¹Faculty of Technical Sciences, University of Novi Sad

Keywords

TeslaRIS, MongoDB, Scopus, import, embeddings, LLMs

Information about publications are partially cataloged in global sources. Those information might be used as a starting point for collecting complete bibliographies of researchers in institutional research information systems. The information might be checked by publication authors in this process in order to be well ingested into the graph of research information. Persistent identifiers might be useful in this process when identifiers exist in global sources. Unfortunately, in some cases those identifiers are missing which might lead to mislinking entities due to unavailable entity recognition (journal, conference, researcher, etc.) in the case automatic import is applied. Therefore, we are suggesting a semi-automated process supervised by publication authors who will help a CRIS platform in proper linking of entities in the case of missing persistent identifiers. We implemented a module which is supporting this process. The module is implemented in the TeslaRIS platform in the case of importing data from Scopus global source. However, this module can be reused for other CRIS platforms and easily extended to support other global sources (OpenAlex for instance).

TeslaRIS is centered around a modular and scalable architecture implemented using the Spring Boot framework and follows the principles of a modular monolith (a.k.a. “modulith”). The **Import Module** is responsible for efficiently collecting data from external sources such as other CRIS systems, global publication platforms, files, etc. It implements the ETL (extraction, transformation, loading) process. The result of this submodule is temporarily stored in a MongoDB intermediate database in common format so data from multiple sources can be loaded by the same loading submodule. Upon harvesting, the record embedding is calculated using LLM (all-MiniLM-L6-v2). As embeddings are calculated by context, we can easily perform a vector search and see whether the record is already loaded from some other source (deduplication) or whether we possess the older version of the same record which has to be updated (merging) even in multilingual context. If it is absolutely necessary, system administrators can consolidate and clean up data, before its transformation and loading into the TeslaRIS database. After harvesting is complete, every author from the publication that has a user profile on our platform will receive a notification stating that he/she can perform the loading process. The loading process guides the researcher through a wizard that performs supervised loading of data. First, the user has to bind every author from the imported publication to an author in our system. The system tries to perform this step

automatically using the import platform's unique identified (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. Should the user not find an adequate match or should our database query return no results, the new researcher is automatically created and the process is continued to the next author until we reach the end of the list. The important detail here is that once this step is completed with a certain author, the platform's unique ID will be automatically bound to said researcher, ensuring that next time this process is completed automatically by the system, without additional user's interaction. After that, the same process is repeated for authors' affiliations and journal/proceedings that the publication is published in. 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. This asynchronous harvesting process ensures seamless integration with external systems and enhances data interoperability.

The image displays two screenshots of the TeslaRIS web application interface, illustrating the supervised import module. The top screenshot shows the initial step of the process, where the user is currently loading 'Publications from Serbia in the Science Citation Index Expanded: a bibliometric analysis' for the year 01.10.2015. A progress bar at the top indicates the sequence of steps: Ivanović D., Fu H., Ho Y., Journal, and Other details. The 'Fu H.' step is active, showing author information: 'Found as Fu H. in the system.', 'Author affiliations - Peking University' (Found as Peking University in the system.), and 'Author affiliations - Asia University' (Found as Asia University in the system.). Navigation buttons include 'PREVIOUS', 'NEXT', 'SKIP DOCUMENT', 'SMART IMPORT', and 'IMPORT'. The bottom screenshot shows the same process at a later stage, where the 'Journal' step is active. This screen displays various publication metadata fields: 'Subtitle', 'Language' (set to 'SR'), 'Volume', 'Issue', 'Start Page' (145), 'End Page' (160), 'Year of Publication' (2015-10-01), 'DOI' (10.1007/s11192-015-1664-9), 'Scopus ID' (84942990602), and 'Type of Publication' (Research Article). The user's profile 'Janko Jankovic (Researcher)' is visible in the top right corner of both screenshots.

Figure 1 - UI for TeslaRIS's supervised import module