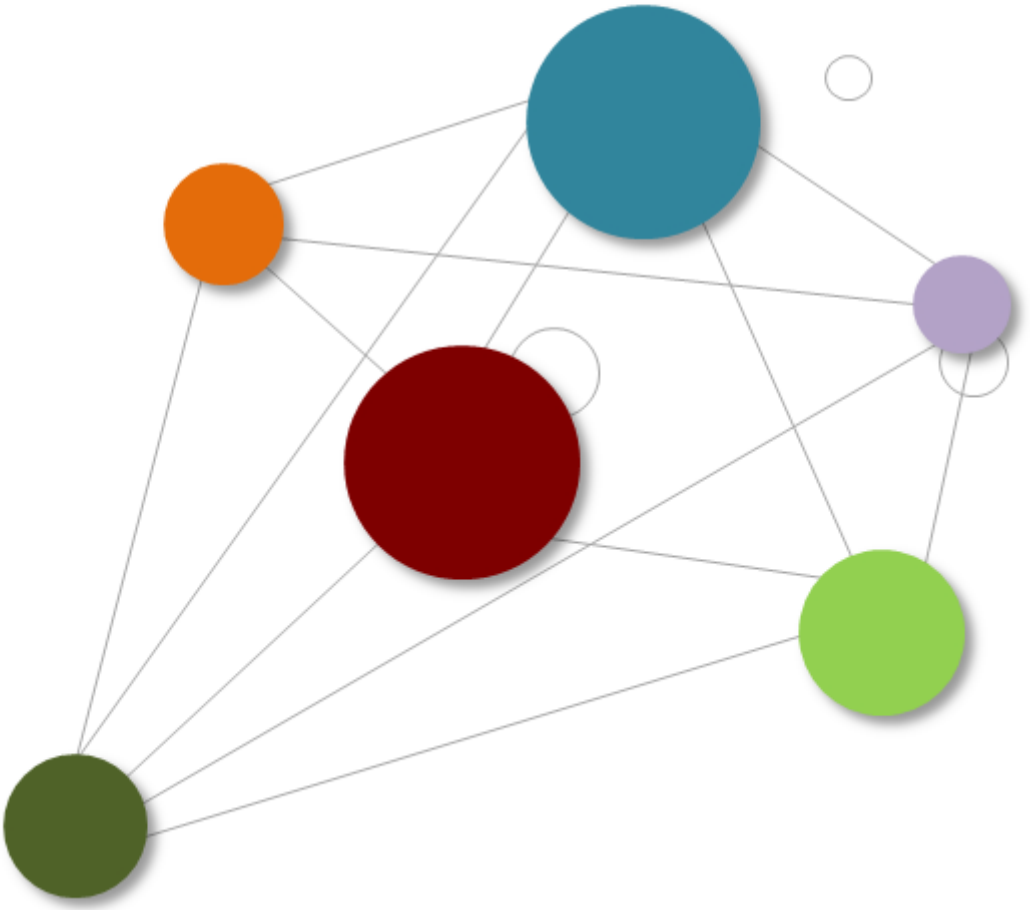
LOD Platform

Linked Data as a Service



@cult

Linked Data for the management of shared knowledge in libraries, archives, museums (LAM)

Table of content

For further information please send a message to info@atcult.it

1. Context

Libraries, archives and museums (LAM) hold a vast quantity of data and resources that, until now, have often remained hidden from sight in catalogues and archives. Unlocking the potential to exploit and diffuse this previously unknown information to a wider audience would bring opportunities to further enrich the World Wide Web, promote a culture of openness towards knowledge and create a number of advantages for each player in the information chain.
Thus, the world's cultural heritage may be carried forward for the benefit of future generations, research may be preserved in its original languages, and cultural variety and vitality may be kept alive.
Linked data is the instrument that can make this possible.

2. Semantic web and linked data

The term linked data (LD) refers to a set of rules for publishing and connecting structured data on the web and is based on the concept of a qualified relationship.
The central idea of LDs is to use hyperlinks not only to identify web documents (as happens in the traditional World Wide Web), but also arbitrary real world entities with the ultimate goal of creating a real network of unique and certified data that form the basis for navigation.
The problem of integrating catalogs of different objects is thus overcome, thanks to a system that interconnects them regardless of their type (books, works etc.) and their origin (libraries, museums, archives etc.), creating a very functional and complete information network for those who research, and at the same time making shared knowledge a collaborative tool for cultural institutions.
therefore, linked data lay the foundations for further improvements in the collaboration between institutions and for the reuse of data in different contexts, enabling more efficient Web-based identification and discovery possibilities.
This new approach is a precious opportunity for libraries, archives and museums to provide users with integrated research tools, increasingly sophisticated and in line with the evolution of new web technologies, while for data producers it is a decisive step for the creation and promotion of new forms of cooperation.
The potential benefits are significant; flexible and sustainable, this cooperative approach allows the creation of a well-structured system for organizing data, enhancing the specificity of the various cultural traditions and exploiting the opportunities offered by new technologies.
Key players in the sector will be able to contribute alongside smaller organizations, ready to share their experience and the uniqueness of their resources.
Libraries, archives, museums, publishers and suppliers will be key players in generating new data and discovering new resources, pushing the boundaries of their specific domains to create data enrichment opportunities that would previously have been unthinkable.

3. What is the LOD Platform

LOD - Linked Open Data Platform is a highly innovative technological framework, an integrated ecosystem for the management of bibliographic, archive and museum catalogues, and their conversion to linked data, extensible as needed for specific purposes.
The core of the LOD Platform was designed in the EU-funded project ALIADA, with the idea of creating scalable and configurable linked data management systems and discovery interfaces able to adapt to ontologies from different library, archive and museum (LAM) domains, capable of

automating the processes of creating and publishing linked open data, regardless of the data source format.
The aim of this framework is to open the possibilities offered by linked data to libraries, archives and museums by providing greater interoperability, visibility and availability for all types of resources.
The application of the LOD Platform obviously requires the careful analysis of the standards, formats and models used in the institution addressed; however, its coverage, based on BIBFRAME 2.0 as core ontology, can be enriched with a suite of additional ontologies, vocabularies and models according to specific needs.

By incorporating standards, models and technologies recognized as key elements for the creation of new processes of management and use of knowledge, the LOD Platform allows:

- the creation of a data structure based on Agent, Work, Instance, Item, Place entities, as defined by BIBFRAME, and extensible to reconcile other entities;
- data enrichment through the connection with external data sources;
- reconciliation and clusterization of entities created from the original data;
- the conversion of data according to the standard model indicated by the W3C for the LOD, RDF - Resource Description Framework;
- delivery of converted and enriched data to the target institution for reuse in their systems;
- the publication of the dataset in linked data on RDF storage (triplestore);
- the creation of a discovery portal with a web user interface based on BIBFRAME or other ontologies defined in specific projects.

4. Overall steps

In the implementation of a system that uses the LOD Platform, data from libraries, archives and museums are transformed into linked data through entity identification, reconciliation and enrichment processes.
Attributes are used to uniquely identify a person, work or other entity, with variant forms reconciled to form a cluster of data referring to the same entity. The data are subsequently reconciled and enriched with further external sources, so as to create a network of information and resources. The result is an open relationship database and Cluster Knowledge Base (CKB) in RDF.
The database uses the semantic web paradigms but allows the target institution to manage their data independently, and is able to provide:

- enrichment of data with URIs, both for the original library records and for the output linked data entities; for an overview of the sources used for the enrichment of the data see SVDE_URI_table_for_external_sources;
- conversion of data to RDF using the BIBFRAME vocabulary and other ontologies;
- creation of a virtual discovery platform with web user interface;
- creation of a database of relationships and clusters accessible in RDF through a triplestore;
- implementation of tools for direct interaction with the data, permitting the validation, update, long-term control and maintenance of the clusters and of the URIs identifying the entities (see below);
- batch/automated data updating procedures;
- batch/automated data dissemination to libraries.
- progressive implementation of additional workflows such as API for ILS, back-conversion for local acquisition and administration systems, reporting.

In the various stages of the project, the participating institutions are called upon to actively collaborate, so that decisions relating to the following stages are based on a solid foundation for large-scale implementation in the community of libraries, archives and museums.
The data used are not simply converted into RDF, but are also enriched with identifiers and interconnected with each other, so as to allow their use in the LD environment.
The goal is to ensure that a large amount of data, which often remains hidden or unexpressed in closed silos (“containers”), finally reveals its richness within existing collections.

5. Benefits

The LOD Platform, developed according to the principle of functionality, provides various environments and interfaces for the creation and enrichment of data and offers workflows capable of responding to the different needs of librarians / archivists / museum operators, professionals, scholars, researchers and participating students. There are several advantages:

- integration of the processes of a collaborative environment with local systems and tools;
- integration into the semantic web while maintaining ownership and control of the data, benefiting from the simplified administration of the environment and a large pool of data;
- continuous development and support for metadata functions according to semantic web standards;
- standards and infrastructures for "future-proof" data, ie ensuring that they are compatible with the structure of linked data and the semantic web;
- enrichment of data with further information and relationships not previously expressed in the established metadata formats in use (e.g. MARC), increasing the possibilities of discovery for all types of resources;
- create an environment that is useful for both end users and professionals (librarians, archivists, museum operators);
- allow librarians a wider and direct interaction with and editing of linked data entities through the Cluster Knowledge Base Editor (more details in the following sections);
- advanced search interfaces to improve the user experience and provide broader search results to users;
- optimised discoverability, revealing data that would otherwise have remained hidden in silos, allowing end users to access a large amount of information that can be both imported and exported by the library.

This approach fully harnesses the potential of linked data, connecting library information to the advantage of scholars, patrons and all library users in a dynamic research environment that unlocks new ways of accessing knowledge.

6. Added value

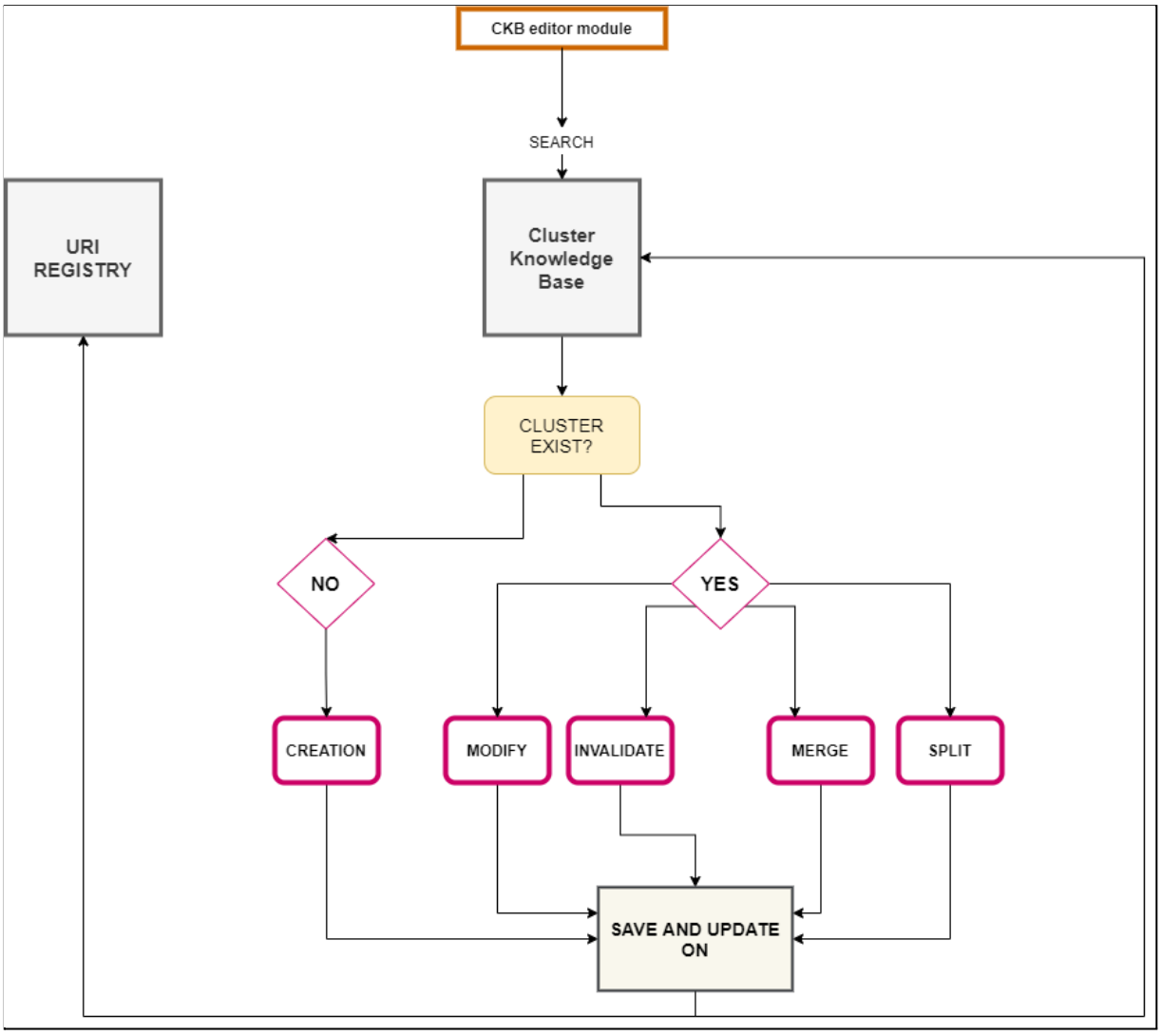
It’s particularly relevant to highlight that the LOD Platform is currently being enhanced with a module dedicated to edit and update entities in the Cluster Knowledge Base (CKB). This Cluster Knowledge Base editor has been named J.Cricket in the context of the Share-VDE system where it’s being designed, and is conceived as a collaborative environment with different levels of access and interaction with the data, enabling several manual and automatic actions on the clusters of

entities saved in the database, including creation, modification, merge of clusters of Works, of Agents etc.

J.Cricket consists of two main layers:

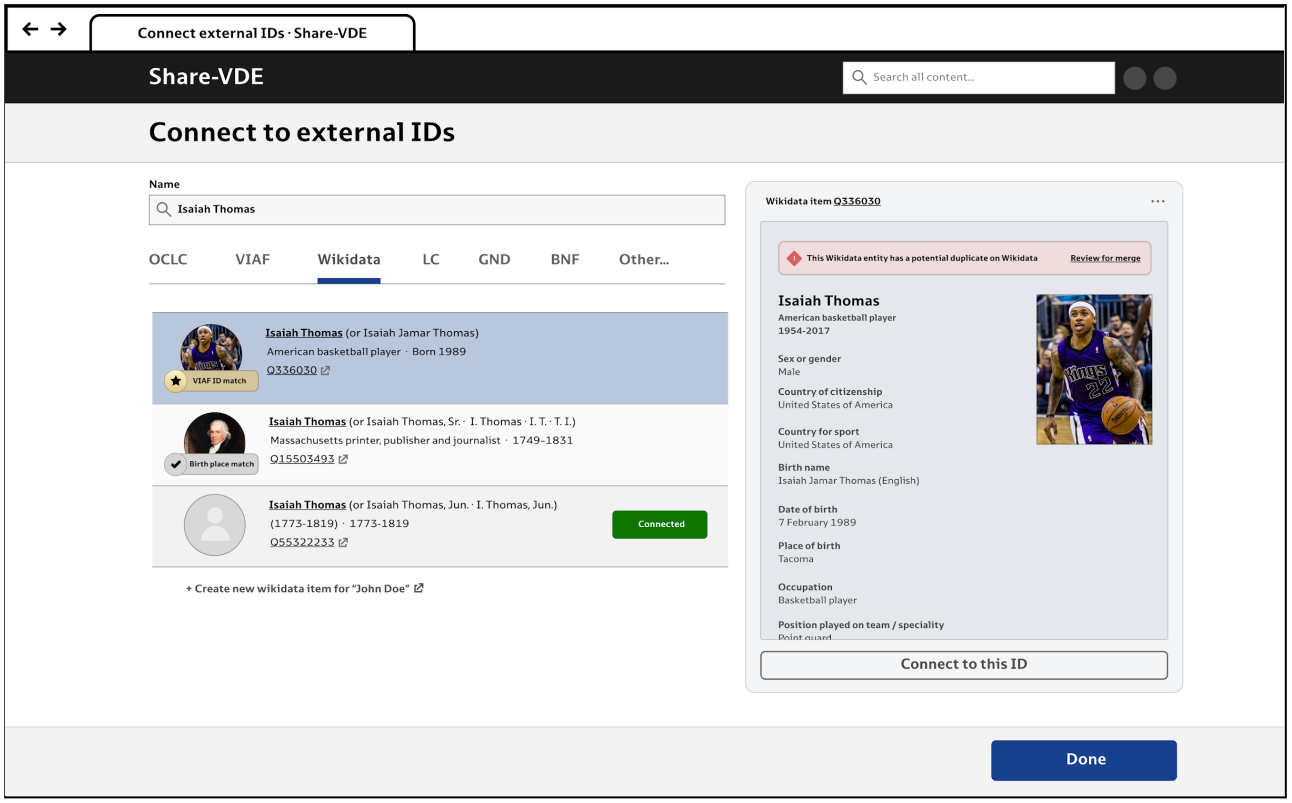
1. automatic checks and update of the data performed by the LOD system;
2. manual checks and edit of the data performed by the user through a web interface.

All changes to entities, both automatic and manual, are reported on the URI Registry, a source (also available in RDF) that tracks the updates of each entity, especially when this has an impact on the persistent entity URI.
The J.Cricket development plan is far-reaching and spans over the current and the next year. The system will be released in progressive steps and, roughly speaking, the authentication, advanced search and CKB entity merge functions that are currently under development are planned to be completed in the second half of 2021. Other functionalities will be addressed in a broader development plan over the year 2022.
Here follows the high level workflow of the Cluster Knowledge Base editor.



Here follows the link to a sketch of one of the manual functions of the CKB editor. This sketch represents the interface where the authorised user can edit the data of the work entity: edit of linked data entity.

A further added value is the ability of the LOD Platform to interact directly with external data sources such as ISNI and Wikidata. The interaction with Wikidata is currently under analysis and will be triggered from the CKB editor itself, allowing the search from the editor into Wikidata and the enrichment of the LOD Platform entities with information from Wikidata and vice versa. This way the editor will allow for the creation of new identifiers both in the external data sources (where possible or applicable) and in the Cluster Knowledge Base.
The following sketch illustrates the editor interface where the results from a query on Wikidata are displayed: the editor is ready to enrich the entity with Wikidata information that will be saved in the Cluster Knowledge Base.



7. How the LOD Platform works

The developed components and tools aim to create a useful environment for knowledge management, with advanced search interfaces to improve the user experience and provide wider results to libraries, archives, museums and their users:

- **Authify**, a RESTFul module that provides search and full-text services of external datasets (downloaded, stored and indexed in the system), mainly related to Authority files (VIAF, Library of Congress Name Authority files etc.) that can also be extended to other types of datasets. It consists of two main parts: a SOLR infrastructure for indexing the datasets and related search services, and a logical level that orchestrates these services to find a match within the clusters of the entities.

- **Cluster Knowledge Base**, on PostgreSQL database, is the result of the data processing and enrichment procedures with external data sources for each entity; typically: clusters of names (authorized and variant forms of the names of Agents) and clusters of titles (authorized access points and variant forms for the titles of the Works).
- **Lodify**, RESTFul module that automates the entire process of converting and publishing data in RDF according to the BIBFRAME 2.0 ontology in a linear and scalable way. It is flexible and can be adapted to multiple situations: it allows, therefore, to manage the classes and properties not only of BIBFRAME but also of other ontologies as needed.
- **Triplestore**: the LOD Platform can currently be integrated with two triplestores, one open source (Blazegraph), more suitable for small or medium scale projects (up to about 2,000,000 bibliographic records), and a commercial one (Stardog) more suitable for larger data sets. In order for end users to access Stardog triplestore, the dedicated tool Stardog Studio is available (see https://stardog.studio and Stardog documentation). Moreover, Stardog provides a number of standard APIs to query the data available in the triplestore, see https://www.stardog.com/docs/#_api_overview.
- **Data presentation portal**, for retrieving and browsing data in a user-friendly discovery interface.

8. The data processing pipeline in a system using the LOD Platform

The diagrams in the next paragraph illustrate the high level workflow for the data processing cycle in the LOD Platform, from the delivery of original records to the publication on the web portal. The workflow diagrams have demonstrative purposes, but they express the overall steps of the process.

Starting from the left of **graph 1**, the data are imported from the library/the target institution (libraries, archives, museums etc.) in different formats (MARC, xml etc.). The type of data can be bibliographic and authority.

The data received are processed according to Text analysis and String matching processes (represented in the "Similarity's score" box), to identify the Entities included in the 'flat' texts (records), and prepare the creation of clusters.

This entity identification function is enhanced and extended through similar Text analysis and String matching processes launched on external sources (VIAF, ISNI, NAF, GND, Nuovo soggettario etc.), through the Authify framework: these processes generate the enrichment of the data with other variant forms coming from external sources and with the URIs through which the same entity is identified on these sources (reconciliation): the original cluster is enriched and will allow, in the process of conversion to linked data, to activate the function of interlinking, essential for sharing and reusing data on the web.

The result of these processes is threefold:

- Identification of entities;
- Data enrichment;
- Cluster creation through reconciliation processes.

The data thus obtained are ready to be processed again, through different channels:

- manual enrichment and quality check (in the event that the library requests a specific service from external agencies - such as Casalini libri - or internally manages the enriched data received);

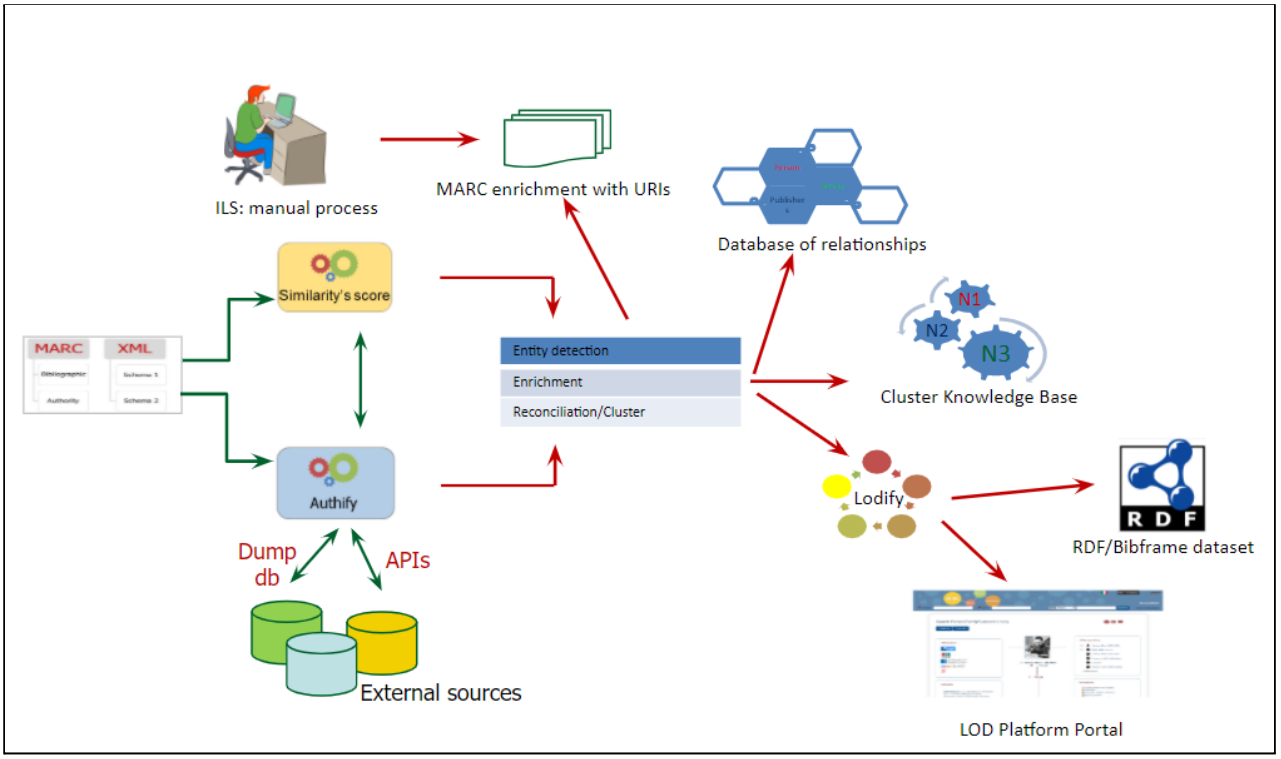
- extraction of “hidden” relations for the generation and feeding of a database of relations (which will be reused in possible subsequent steps to enrich the data and in the publication stages, to extend the links between data);
- creation of the Cluster Knowledge Base, available in RDF (therefore as Linked Open Data) and accessible via an end point for SPARQL and API queries;
- processing and conversion to RDF, following the BIBFRAME model and / or other ontologies and schemas proposed in the specific project.

At the end of these processes, the data is ready to be indexed on the Portal and published on various sites, in RDF.
The system also provides procedures for periodic updating of the data (according to the frequency defined by the customer).

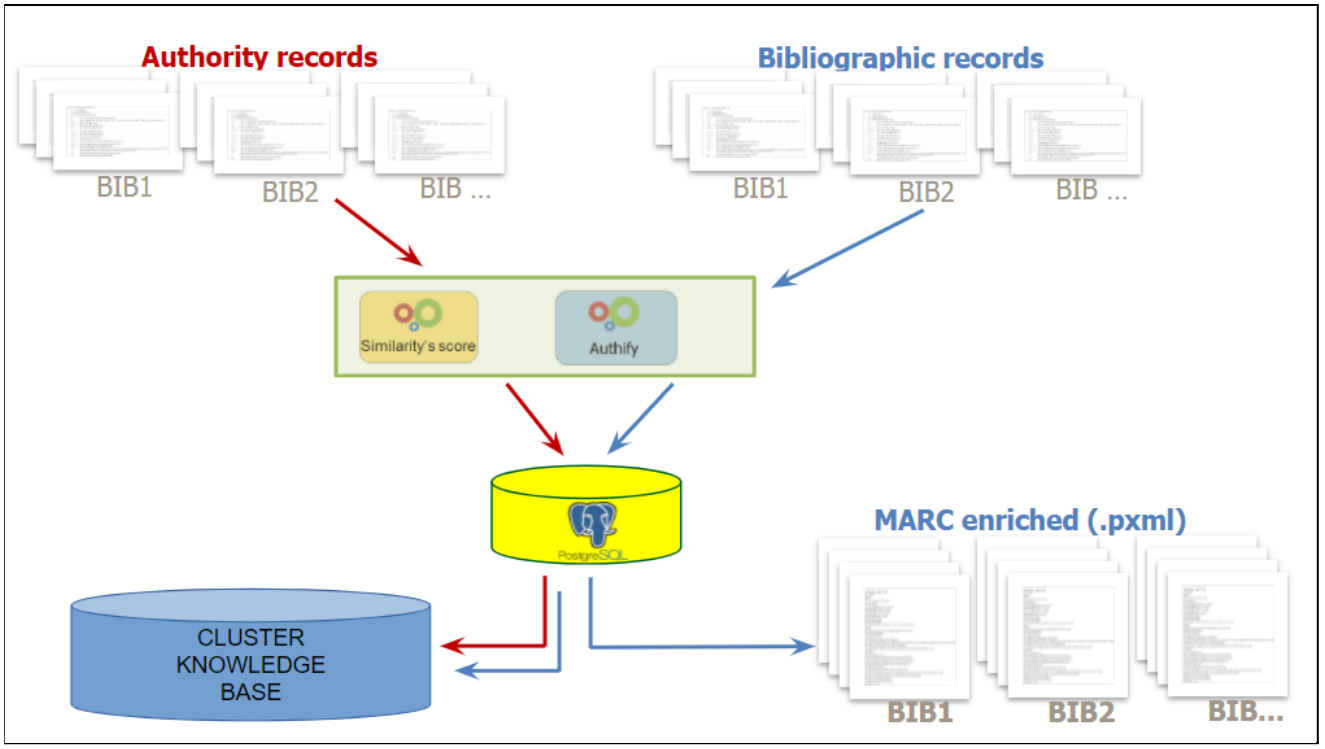
Graphs 2 and 3 show more in detail some steps of the overall workflow in graph 1.

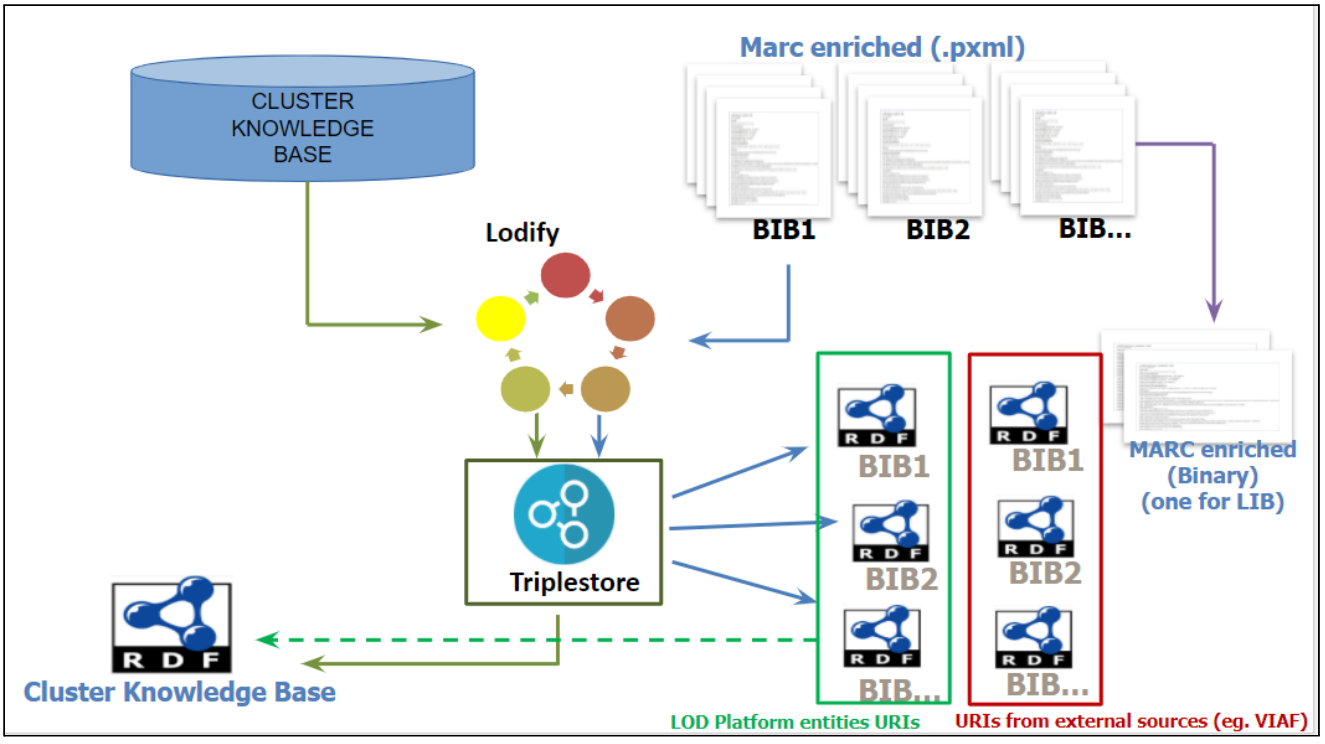
8.1. The LOD Platform workflow

Overall workflow - graph 1



Focus on processes - graphs 2 and 3





8.2. Data updates

The ability to handle data updates is covered by the LOD Platform.
The LOD Platform is able to manage entities created with data from internal and external sources, using different approaches that depend on the data source (update/change management, availability of OAI-PMH or other protocols, periodically update of the dump available for the web community etc.). These different approaches are to be analyzed in the course of the project and adapted to function with the target institution, so that the entities can be updated as much as possible in automated processes. In addition to this, an entity editor (J.Cricket) is being developed in order to allow managing entities manually, by authorized users, using a friendly user interface (see the dedicated section above).

Here follows an outline of the automated data update processes that have already been implemented for Share-VDE and that can be adapted to other projects.

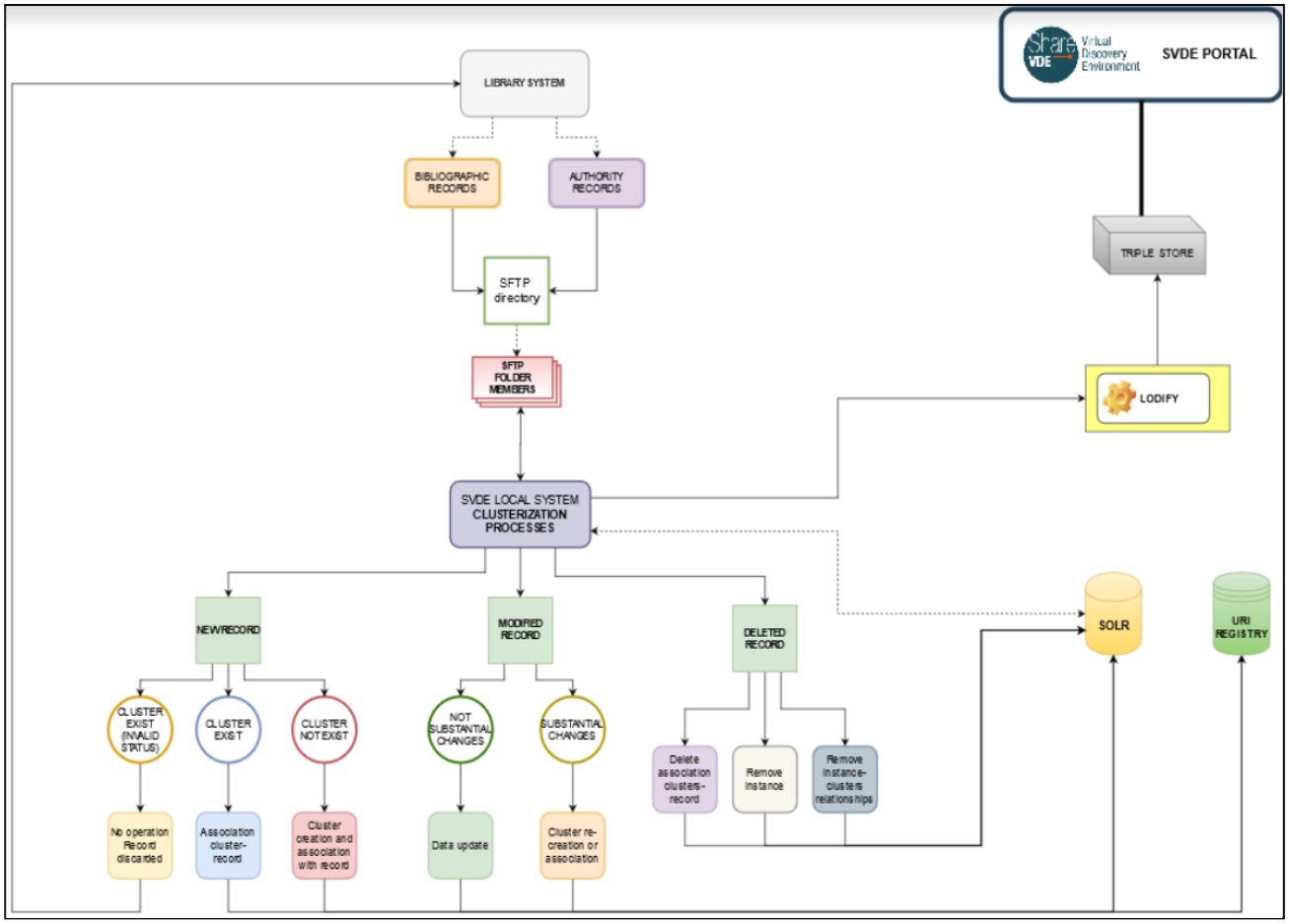
Share-VDE delta updates

By “delta” update we mean the changes that occur to the library records that are periodically pushed to Share-VDE. The automation of the ingestion in Share-VDE of updated library records has the purpose of regularly updating the data available through the discovery interface and the other endpoints of the workflow where the data are available. This means updating the data of the clustered entities and the related resources searchable on the discovery interface and in the triplestore, according to the frequency requested by the library.
Steps of the process:

1. the library delivers bibliographic and authority records to the Share-VDE SFTP directory, in the sub-directory dedicated to their institution. Share-VDE expects to receive from each library only the delta of their records, i.e. only those records that have been changed or added or deleted, compared to the previous dispatch;
2. the script running on Share-VDE side processes the records in sequential order, by file name, and accepts in input .mrc files (for new and modified records) and .txt files (for deleted records);
3. download of library MARC records in Share-VDE local system: after MARC records are uploaded from the library to the SFTP server, a script on SVDE end running regularly connects Share-VDE internal system to the individual SFTP folders of libraries, checks if a new file has been uploaded to the SFTP and downloads the MARC records in Share-VDE local system. The files submitted by Share-VDE libraries are automatically transferred from the SFTP sub-directory of the institution that has uploaded the files to the corresponding sub-directory of Share-VDE internal repository;
4. MARC records processing: the delta updates MARC records are processed according to SVDE procedures and enriched (i.e. addition of Share-VDE original URIs and of URIs from external sources such as ISNI, VIAF, enrichment of other tags etc.). The data are saved in Postgres database;
5. upload to Solr: the records processed are uploaded to Solr platform for indexing, before populating Share-VDE portal. Among the processes involved, data from library records are processed and indexed in a way that the autocomplete function in the search fields on https://share-vde.org/ displays the indexed data (e.g. author, title) as suggested result to the user performing a search for a resource;
6. updated data online: after the indexing phase, the information processed by Share-VDE is ready to go live on https://share-vde.org/.

The delta updates process triggers: the update of clustered entities on SVDE portal; the update of the data available on Stardog triplestore, the delivery of enriched MARC records to libraries.

Here follows a diagram showing the data flow for the elaboration of the delta update records in the SVDE system.



9. User interface

Here follow some details of the user interface of the discovery portal where the linked data entities converted in the LOD Platform can be browsed (in the examples below, Agent and Work entities).