ReIReS
Research Infrastructure on Religious Studies

KU LEUVEN

Grant 730895, ReIReS
Public Report/ORDP

Data architecture

Name of the data providing institution:	KU Leuven - LIBIS
Contact:	Michiel.declerck@kuleuven.be; Roxannne.Wyns@kuleuven.be
Task:	6.3
Task leader:	KU Leuven - LIBIS
Type of data:	Summary of the data architecture of the ReIReS unified discovery environment
Purpose of data generation:	Technical description of the data architecture
Data's audience:	Public
License:	You are free to quote the content of the document by giving credits to the authors, ReIReS project and the EU Horizon 2020 Programme. You are not allowed to share the file, or any part of it, in the media without permission of the authors.
Open Research Data Pilot	No
Reusability duration:	Data remains reusable throughout the duration of the project.
Tags:	Data discovery – Integrated search – Data enrichment – Data aggregation – Federated search
Other, if applicable:	Date of the first training given by the training institution: [XX/XX/XXXX]
	Name of the first training given by the training institution: [..]

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730895.

Grant 730895, ReIReS
Public Report/ORDP

Data architecture

Document reference:	Data architecture		
Version number:	1.00		
Status:	Final		
Last revision date:	25/06/2019 09:07	by:	Michiel De Clerck
Verification date:	DD/MM/2019	by:	[Verifier Name]
Approval date:	DD/MM/2019	by:	DG RESEARCH
Subject:	Public Report/ORDP Data architecture		
Filename:	ReIReS_Deliverable_6.3_v1.00		

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730895.

Change History

Version Number	Date	Status	Summary of main or important changes
00.01	13/05/2019	WORKING	Working version for internal use
00.02	13/05/2019	WORKING	Internal review
00.03	17/06/2019	WORKING	Draft for review by Executive Board
01.00	28/06/2019	FINAL	Final Version

Distribution List

Name	Company	Role
Roxanne Wyns	LIBIS	WP6 leader
Tom Vanmechelen	LIBIS	Software architect
Peter O	LIBIS	Software architect
Michiel De Clerck	LIBIS	WP6 collaborator
Jan Driesen	Brepols	WP6 collaborator
Executive Board	RelReS	Validation
Allstaff@reires.eu	RelReS	Information on final version

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730895.

Table of Contents

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730895.

1. Introduction

This task translates the functional and non-functional requirements into the technical architecture of the ReIReS unified discovery environment. It starts with a research and design phase, selecting suitable hard- and software components to achieve the requirements specified in Deliverable 6.1 *Requirements Overview and Use Case Models*. The ReIReS unified discovery environment is intended for researchers in the field of religious studies to facilitate the discovery of larger sets of data. It aims to do this via both the aggregation of data from consortium members and a federated search in datasets that are accessible via API within the consortium.

These datasets will be published on a single location on a navigation and presentation interface designed for the searching, browsing and visualization of data. The data will be stored in a graph database, which uses graph structures that relate data items to each other to represent relationships between them. Graph databases hold the relationships between data as a priority.[1] This emphasis on expressing the relation between data is intended to further improve discoverability. Data will be stored in the JSON-LD format following a data model based on Schema.org.[2] JSON-LD and Schema.org were designed for encoding and structuring Linked Data respectively.[3] These elements are also used in search engine optimization because they increase visibility of search results.[4] This application profile created for ReIReS, based on Schema.org, is described in D6.2 *Integrated Metadata Model*.

This first version reports on the work done towards the development of the back-end of the infrastructure. An update of this report will follow as the work progresses with the build of the user interface and the refinement of the environment.

[1] https://en.wikipedia.org/wiki/Graph_database
[2] https://en.wikipedia.org/wiki/JSON-LD
https://en.wikipedia.org/wiki/Schema.org
[3] https://en.wikipedia.org/wiki/Linked_data
[4] https://en.wikipedia.org/wiki/Search_engine_optimization

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730895.

2. Architecture design phase

2.1. Preambles

Two major aspects of the work done in WP6 have an impact on the design choices made for the ReIReS discovery infrastructure: requirements and selection of a data model based on the principles of the semantic web for the interlinking of data. The main goal when choosing and configuring the architecture of the ReIReS unified discovery environment is to improve the findability, accessibility and visibility of datasets related to religious studies.

In order to tailor the ReIReS unified discovery environment to the needs and expectations of scholars in the field of religious studies, WP6 conducted a survey among members of the consortium and translated these into a set of requirements. The technical staff designs the data architecture based on these requirements according to the priorities assigned to them during the requirements analysis phase. The first requirements deal with back-end processes: the data storage and data retrieval layers. The requirements overview can be consulted in D6.1 *Requirements Overview and Use Case Models*. The main priorities that arose from the requirements analysis relate to the findability, accessibility and visibility of data.

Because the unified discovery environment has to store metadata from different sources and make it findable and accessible in a single location, it needs a shared data model. This is important in order to collect and present data from different data providers in a uniform way and to make it accessible in an interoperable and reusable format. The selected data model takes into account the requirements for standardized data access with a wide outreach potential. The data storage layer needs to be compatible with the data model so that data can be stored correctly. Therefore, the choice of data model also impacts the choice of technical solutions. The description of the data model and reasoning behind it can be found in D6.2 *Integrated Metadata Model*.

Schema.org was selected because:

- It enables semantic interlinking
- Is suitable for any type of data
- Is flexible and extendible
- Is human- and machine-readable

It is, however, also too extensive. Because of this, an application profile was created specifically for ReIReS with a selection of classes and properties. Data will be stored according to this data model in the JSON-LD format, which is mostly used for search engine optimization.[5]

2.2. Investigation of technology components

During the investigation phase both licensed and open source technologies where reviewed. This was done through both online research and test set-ups as well as consultations with commercial companies about their products. While some of the licensed products reviewed had potential, none were able to cover all aspects of the infrastructure and extra development would remain needed. At the same time, these

[5] https://en.wikipedia.org/wiki/JSON-LD

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730895.

products were expensive both in terms of one time purchase and in yearly maintenance fees. After careful consideration, it was decided to use mainly open source and freeware products for the development of the environment in order to save on those costs. Purchasing licensed technologies would also require following EU tendering procedures, which would slow down the architecture design process. Working with several open source technologies helps speed up the development work by using readily available components. This also frees up resources for staff to work on development. In addition, this helps guarantee the sustainability of the infrastructure after the lifetime of the project when available budget for licensing and maintenance costs is limited.

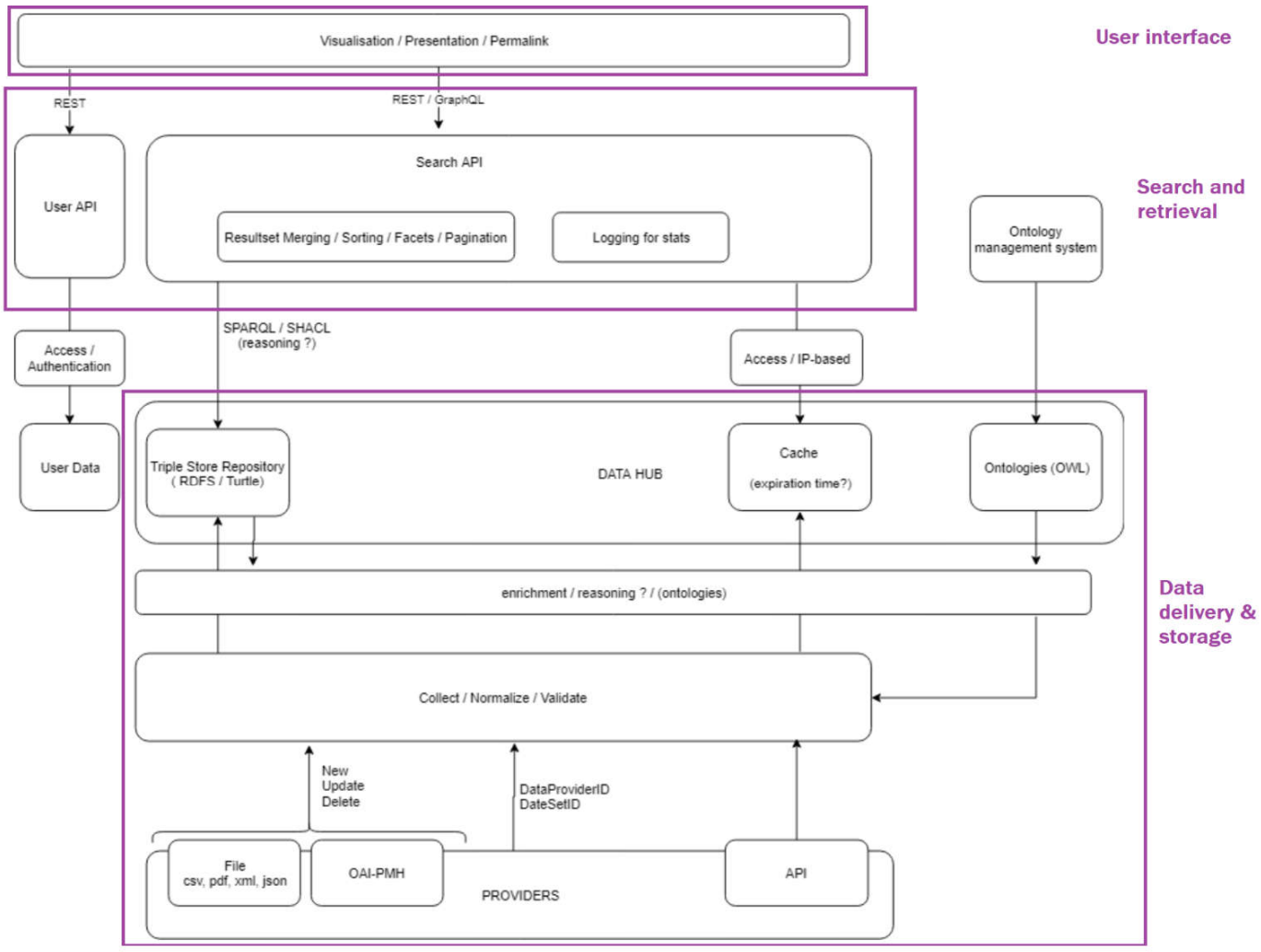
2.3. Architecture design

The design phase underwent several versions of architecture schematics and since development work is still in progress, it might still undergo additional changes and refinements when new parts of the architecture development are covered. The summaries in chapters 3 (Data delivery and storage), 4 (Search and retrieval), and 5 (User interface) give an overview of the components that were chosen and developed.

The design process is structured along three major builds. These development phases focus first on the back-end, which is the focus of this deliverable. Once a working Proof of Concept is available for testing and further configuration, work can begin on a preliminary version of the user interface. The final phase will mainly be concerned with the finalization of all the components and optimizing their integration.

The graphical outline below shows the architecture schematics as planned, from data delivery at the bottom to data representation in the user interface at the top. It visualizes all the steps along the way. The following chapters explain this more in depth.

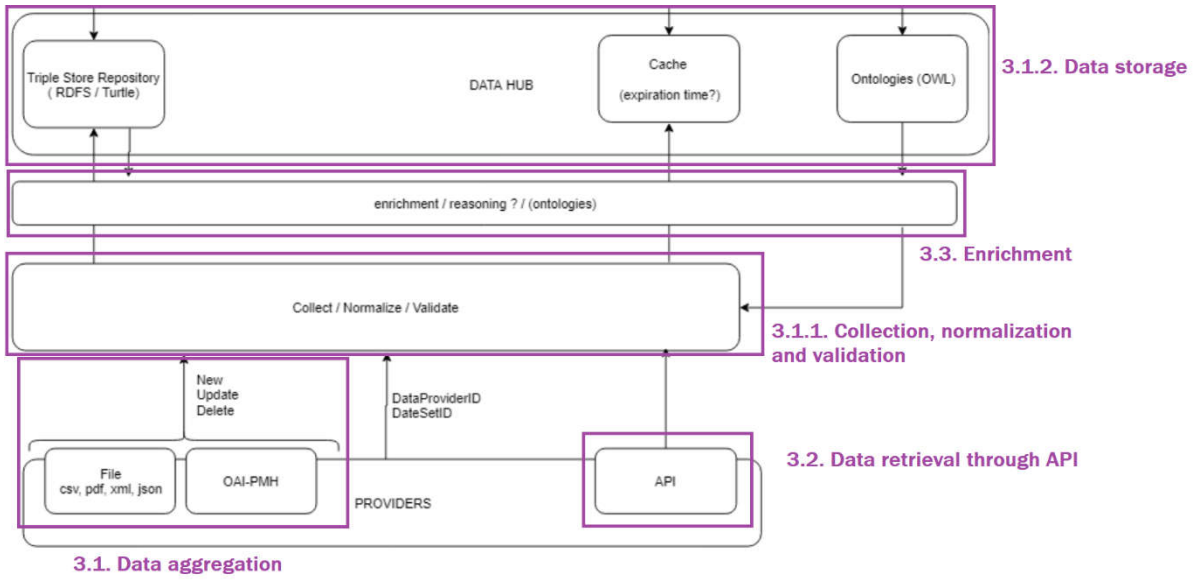
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730895.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730895.

3. Data delivery and storage

Most of the datasets are imported from different providers after mapping, normalization and enrichment into a central database. It provides the user with an integrated, consistent and governed view of the metadata. With regards to data storage, the unified discovery environment needs to import data from different databases (3.1. and 3.2.), store it according to the same data model (3.1.1.) and enrich and link that data together where relevant (3.3.)



3.1. Data aggregation

3.1.1. Collection, normalization and validation

For the uniform importing of datasets, WP6 and data providers have to follow a set of technical guidelines and metadata mapping specifications. The provided data is converted to the ReIReS Schema.org format. For this purpose a collector and a validator were developed.[6] The collector can process locally stored MARC-XML data or records supplied via an OAI-PMH client. This can be expanded to other formats in the future, depending on which other sources need to be added. The data aggregated from JGU Mainz is collected with an OAI-PMH-client in the MODS/METS format.[7] The data from KU Leuven is exported from ALMA, its library services platform, to the MARC 21 XML format.[8] The exact specifications of how data needs to be delivered will be documented in an ingest agreement that explains the steps data providers need to take to publish their data on the unified discovery environment. After normalizing the data to the ReIReS data model and validating it for mandatory elements, it is stored in an AgensGraph datahub.

[6] https://en.wikipedia.org/wiki/Data_collection_system
https://en.wikipedia.org/wiki/Data_validation
[7] https://en.wikipedia.org/wiki/Open_Archives_Initiative_Protocol_for_Metadata_Harvesting
http://www.loc.gov/standards/mods/
http://www.loc.gov/standards/mets/
[8] http://www.loc.gov/standards/marcxml/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730895.

3.1.2. Data storage

To store this data in a way that is compatible with the requirements, AgensGraph was installed on the servers intended for the ReIReS unified discovery environment.[9] AgensGraph is a graph database, meaning it uses graph structures rather than a relational model.[10] A graph database avoids the limitations of a relational database by explicitly stating the dependencies between data elements. In a relational model, data is organized into tables of columns and rows. Therefore, the relationship between data elements that have different structures or different data types cannot be stated explicitly. In a graph database, all data nodes can be connected to each other and their relationships can be stated. Considering that compatibility with Linked Data is a requirement, a graph database is an obvious choice. AgensGraph allows knowledge to be represented in a machine readable way and to express relationships between data from different data providers.

AgensGraph stores and manages various types of data including graph and relational data. This means that not all the data has to be stored according to graph structures when a relational database suffices. Storing administrative data related to the unified discovery environment in a graph structure, for example, would be redundant and take up more storage space. Properties of data elements and their relations are expressed in the JSON format, just as the ReIReS data model.[11] The development of the graph database for ReIReS required the integration of the Cypher Query Language to create an importer and exporter, because it allows for expressive and efficient querying of data.[12] AgensGraph is a free and open source software, which allows WP6 more flexibility in designing the data architecture while freeing up the budget required for more technical staff to design the unified discovery environment without relying on licensed software.

3.2. Data retrieval through API

Because some datasets are accessible through API, their metadata can be queried without having to import their records into the database. To connect their data sources to the ReIReS unified discovery environment through API, data providers will have to follow similar technical guidelines and metadata mapping specifications as mentioned previously. During the pilot stage of the project, the *Index Religiosus* database maintained by Brepols will be integrated with the aggregated data in AgensGraph to test the viability of combining aggregated and federated search. As this data was not yet accessible via data transfer protocols, Brepols designed an API through which the unified discovery environment could access *Index Religiosus*. Search results from both forms of data delivery are then combined in one result list through the Blender API.

[9] https://bitnine.net/agensgraph/
[10] https://en.wikipedia.org/wiki/Graph_(abstract_data_type)
https://en.wikipedia.org/wiki/Relational_model
[11] https://bitnine.net/documentations/manual/agens_graph_quick_guide.html#data-model
[12] https://en.wikipedia.org/wiki/Cypher_Query_Language

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730895.

3.3. Enrichment

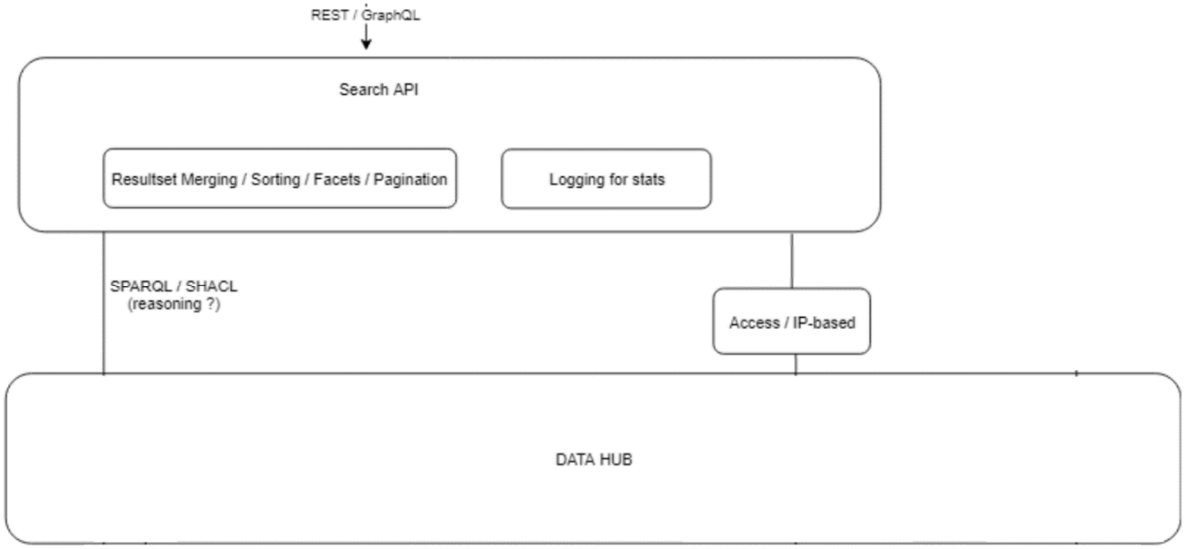
In the long run, the unified discovery environment will also need to integrate tools needed for the enrichment of metadata and the integration of multilingual thesauri and authority resources. In the future, GraphDB will be used for enrichment and interpretation of the data.[13] GraphDB includes tools to explore datasets and the links between them, to discover and experiment with data and to quickly and easily integrate hundreds of datasets to enrich research. This includes knowledge graphs, metadata management and data integration tools.

[13] https://www.ontotext.com/products/graphdb/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730895.

4. Search and retrieval

The creation of a ReIReS application profile also facilitates an integrated search and retrieval of the datasets. It includes elements for indexing and display of metadata, and specifies the mandatory, recommended and optional metadata elements the data providers have to include when contributing data. The search engine indexing takes into account the specifications of this application profile, as well as the needs of researchers as indicated by the requirements analysis.[14] For this purpose, the unified discovery environment needs to efficiently and reliably retrieve data (4.1.) from the data storage hub, while also accessing databases included via API (4.2.) and presenting those results from different sources in a uniform result list (4.3.).



4.1. Elastic search

To make the aggregated data ingested into the AgensGraph database searchable, the technical staff of WP6 installed Elasticsearch on the ReIReS servers.[15] Elasticsearch is a distributed, RESTful search and analytics engine. As a distributed engine, its components are located on different parts of a computer network. [16] To provide more reliability, it was therefore installed across three nodes on the ReIReS servers. REST refers to a software architectural style that provides interoperability between computer systems and the internet and aims for fast performance, reliability and the ability to expand without affecting the system.[17] Elasticsearch is an open-source tool with JSON support that supports faceted search.[18] The

[14] https://en.wikipedia.org/wiki/Search_engine_indexing
[15] https://www.elastic.co/products/elasticsearch
https://en.wikipedia.org/wiki/Elasticsearch
[16] https://en.wikipedia.org/wiki/Distributed_computing
[17] https://en.wikipedia.org/wiki/Representational_state_transfer
[18] https://en.wikipedia.org/wiki/Faceted_search

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730895.

technical staff of WP6 developed load-scripts to pre-process the ReIReS application profile format. This helps improve indexing and search options in Elasticsearch and gets the data in line with the Elasticsearch mappings of the application profile. A query made in the user interface is sent through the Elasticsearch engine and the results of this search are then combined with those of the federated search.

4.2. Federated search

In order to process requests made by the ReIReS unified discovery environment for the search and retrieval of data in the *Index Religiosus*, Brepols is designing an application programming interface.[19] The integration of API's such as this allows the unified discovery environment to access data through a federated search by distributing a single query request to different participating databases.[20] A Proof of Concept for the Brepols API is already available for WP6, allowing the technical staff to develop a solution for the combination of aggregated and federated search results in a single result set. This enables the input of search queries to *Index Religiosus* via the ReIReS unified discovery environment. The API can then deliver search results mapped to the ReIReS application profile in a JSON-LD format. This testing API delivers a result that the technical staff can work with while it is being designed alongside the rest of the unified discovery environment.

4.3. Blender

In order to merge the search results from the aggregated data from the Elasticsearch engine and from the federated data retrieved via the Brepols API, the WP6 technical staff implemented a blender tool. The Blender is an API that receives a search query in a Lucene based query language and returns a combined result set from the different data providers.[21] The Blender can handle request from the ReIReS Elasticsearch engine, with data from JGU Mainz and KU Leuven, and the Brepols API. Without this tool, search results from different databases would not be able to be merged in a single sortable list. This Proof of Concept shows the feasibility of a combined aggregated and federated search.

[19] https://en.wikipedia.org/wiki/Application_programming_interface
[20] https://en.wikipedia.org/wiki/Federated_search
[21] https://en.wikipedia.org/wiki/Apache_Lucene

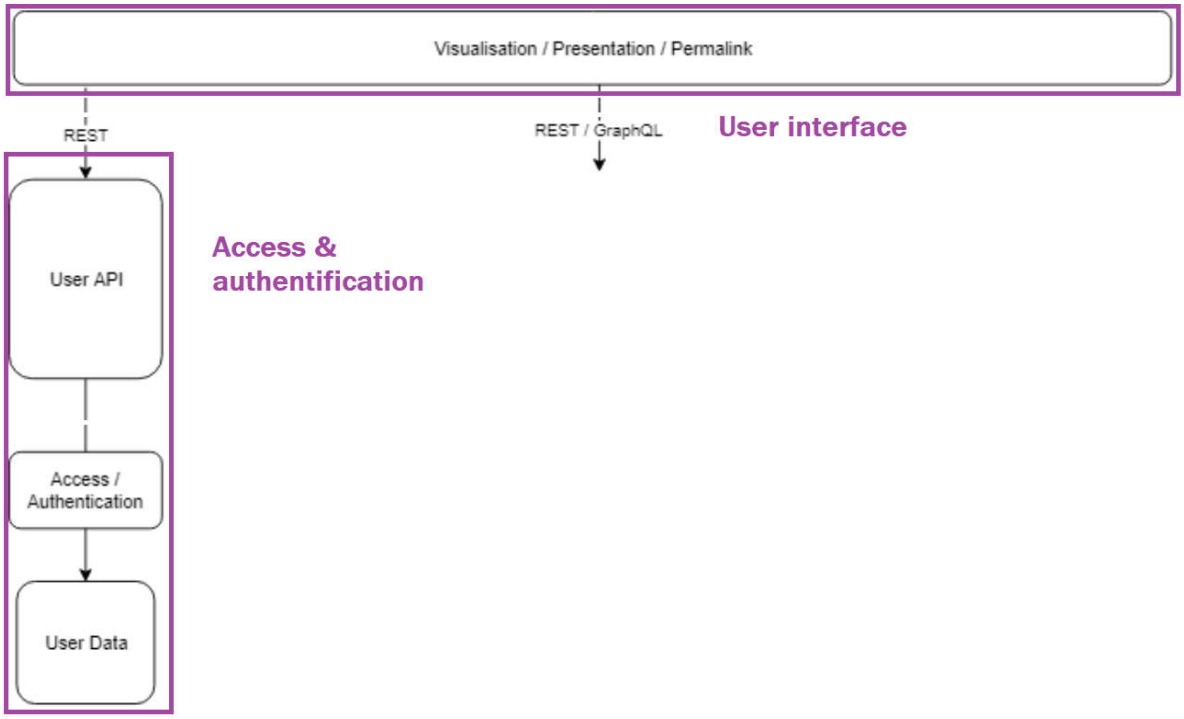
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730895.

5. User interface

With a Proof of Concept for data storage and search and retrieval available, the WP6 technical staff has started work on a first basic version of a user interface for the ReIReS unified discovery environment. Based on the requirements and the ReIReS application profile, basic functionalities for a user interface have been defined and visualized in a mock-up using the wireframing tool Balsamiq. This version of the interface will be suitable for presentation purposes and subsequent feedback.

5.1. Access

Because Brepols is a commercial data provider, the user interface will take into account a different visualization of their data depending on whether or not users have access to Brepols data via IP range based access rights.



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730895.

6. Conclusion

Presented in this document is the first version of what is to become the ReIReS unified discovery environment. During this phase of the development plan, WP6 focuses on testing and validation activities of the Proof of Concept. Testing and validation then proceeds in parallel with the development of new technical and functional features. This leads to updates to the data repository and discovery system in order to achieve a first generation release of the environment by the end of the ReIReS project. It is expected that the unified discovery environment will also encompass changing requirements and new evolutions later in the development process. In this stage, the viability of aggregating datasets in a graph database according to the ReIReS data model and rendering that data searchable together with a federated search of other databases via API was explored. The necessary technical solutions for this were designed as a prototype.

The WP6 technical staff, up to this point, has set up the ReIReS servers and installed an AgensGraph database on them. Data from JGU Mainz and KU Leuven has been mapped to the ReIReS Schema.org application profile and imported to that database. With the installation of Elasticsearch on the servers, that database can be queried and search results retrieved from it. The design of an API by Brepols, makes the *Index Religiosus* searchable via the prototype of the unified discovery environment. The query results of both the aggregated and federated searches can be presented in a uniform result list with the implementation of a blender tool.

The next steps in the architecture design phase will focus on updates of the prototype for the Proof of Concept and further integration work.

- A first release of the user interface is planned for early 2020
- New features will be added to the user interface after testing and evaluation with users
- Further configuration and refinement of the implemented components
- More datasets will be added in the coming months and throughout the project

The finished version of the ReIReS unified discovery environment will be described in the final version of this deliverable in month 34 of the project.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730895.

(End of Document)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 730895.