

RETRIEVO 3

CONTENT AGGREGATION AND FEDERATED SEARCH

CHARACTERISTICS AND TECHNICAL
REQUIREMENTS

ABOUT THIS DOCUMENT

Identifier	WP201470		
Approved by	Luís Miguel Ferros	Approved on	2020-06-02
Classification	Public		
Distribution	N/A		

REVISIONS

#	Date	Authors	Modifications
1	2020-05-26	Márcia Fernandes	Initial document
2	2020-05-27	Miguel Ferreira	Final review
3	2020-06-02	Miguel Ferreira	Updated architecture

EXECUTIVE SUMMARY

Retrievio is a software capable of collecting and providing access to heterogeneous information assets from various different sources of information existing in an organisation. For example, this product can be used to search a library catalogue, a document management system and simultaneously the institution's website through a unique search interface.

Retrievio is not intended to replace existing information sources, but rather to provide a unique and privileged access point to information that is distributed across different silos within an organisation.

This document describes the main features and value propositions associated with this software. The document also outlines the technical requirements necessary to deploy the software in a production environment.

**SINGLE POINT
OF ACCESS
TO ALL
INFORMATION
ASSETS IN YOUR
ORGANISATION**

RETRIEVO

Retrievio is a software capable of collecting and providing access to heterogeneous information assets from various different sources of information existing in an organisation. For example, this product can be used simultaneously to search a library catalogue, a document management system and the institution's website through a unique and uniform search interface.

Retrievio is not intended to replace existing information systems that act as information sources, but rather to provide a unique and privileged access point to information that is distributed across different silos within an organisation.

After discovering information in Retrievio, the user will be redirected to the system that contains the original information in order to view it in context and in a complete way.

In addition to centralized search, this software also acts as a data provider to other systems. It can be integrated with international information aggregators such as Europeana, Archives Portal Europe, OpenAire or even external federated search services.

Single point of access to heterogeneous information sources

Retrievio allows the discovery of information that is dispersed throughout several unrelated information systems.

Searching is always performed via a standard graphical interface, avoiding the need to access every single system individually to find the information item one is looking for.

Increase your institution's visibility

In addition to consolidating information from multiple systems, Retrievio is able to validate it, clean it and enhance it to provide a better experience to the user.

Retrievio can also provide the consolidated information to external services that will add value via a new set of enhanced services, e.g. statistics, specialized portals, advanced analytics, etc.

Make heterogeneous systems more manageable

Retrievio is compatible with several standardised communication protocols (e.g. OAI-PMH, Z39.50, SRU), as well as SOAP and REST services, SQL connectors and a wide set of access gateways to scientific articles databases (e.g. EBSCO, ABI, ProQuest, etc.).

Regardless of the protocols implemented and of the system that supports the information, Retrievio will be able to collect, clean, index and present information in a convenient and straight forward way to the end-user.

All your information assets under control

Retrievio provides a wide set of statistics and reports that reflect the operating status of the system.

It allows each source of information to be monitored providing statistics such as number of records per system, most accessed records, monthly growth, aggregation reports, access statistics, among others.

The statistics system is configurable and can respond more easily to the needs of those who manage software.

Quality of information

To ensure the quality of the information, Retrievio has a compliance check engine that will validate the information collected according to established criteria.

The criteria are completely configurable by the system administrator and adaptable to any organizational requirement or context.

Additionally, the system can be configured to "clean" the collected information and even enhance it, producing results with higher quality that are more uniform across the board.

Fast as a **blink of an eye**

The large number of records managed by Retrievio that result from a constant aggregation of new information forced us to come up with innovative and scalable solutions.

Aggregations are now more efficient than ever as they allow the aggregation of information in parallel and provide faster and more accurate search results due to the new search engine adopted.

Your **information assets** have never been so valuable

Retrievio is the ideal product to locate information within your institution.

Using its vast number of connectors, it is possible to configure Retrievio to search shared folders, find records in your organization's ERP or find the latest news published on your institution's website.

Simple, **attractive and intuitive**

The entire graphical interface of Retrievio looks simple and intuitive.

It can be completely configured to meet your institution's graphical identity.

Search options include basic search, full-text search, faceted search and advanced search.

Documented REST API using OpenAPI/Swagger

Retrievio offers a documented REST API that allows developers to use collected content in applications developed by third parties.

It is possible, for example, to develop a mobile application to access all of your organisation's valuable information assets by tapping into the REST API provided by Retrievio.

Compatible with any platform

Retrievio is 100% a web solution. To use the system, you only need a browser. The software is centralized on a server, and there is no need to install additional applications on the users' workstations.

Access to the system can be done from anywhere and any device, desktop or tablet. All of its graphical interfaces are responsive, meaning that they adjust automatically to the size of your device making it very user friendly independently of your screen size.

ARCHITECTURE AND APPLICATION MODULES

Retrievio is composed of 4 application modules. The following figure depicts the overall architecture of the product, as well as its technology stack.

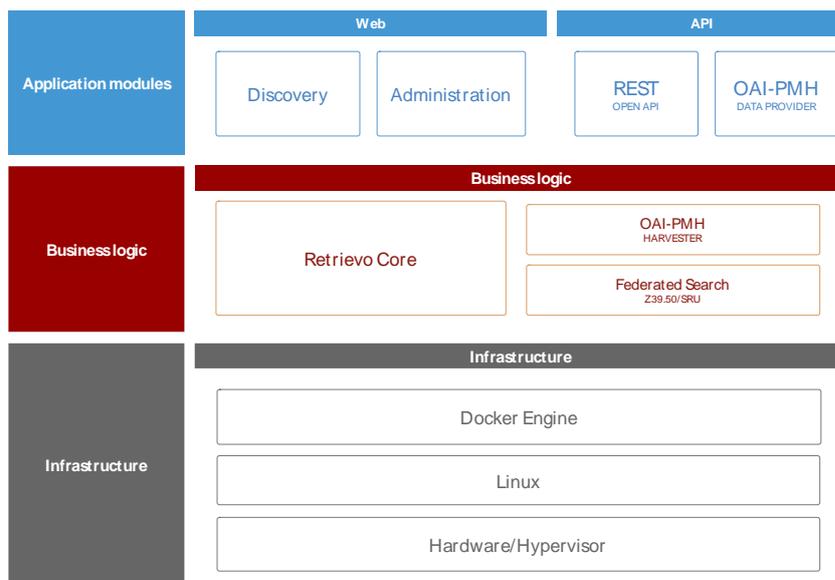


Figure 1 - Retrievio overall architecture.

DISCOVERY

The Discovery module provides a single access point for finding and retrieving information originating from different sources. The main objective is to facilitate access to information since it provides on a single interface data discovery and retrieval capabilities from heterogeneous data sources.

Users can perform cross-sectional searches on all collected records from multiple sources, carry out more specific filtering using existing information fields, define which information sources to search on, or by combining all of the above. She can, for example, search information on any data field from all data sources, or simply search in the Title of data sources A and B.

Search results are consolidated and indexed by title, author, subject, dates, or other attributes previously configured.

Typically, search results only show a subset of the metadata of the original record, which should be sufficient to identify the information being sought. To help users understand the information, a thumbnail of the associated document can also be displayed.

After locating the sought information, users can view the information in its original context, being forwarded to the data source that provided that record and be presented with a full set of metadata and associated files. This enables users to trigger additional services that only exist on the business specific application, e.g. borrow of documents using an integrated library management system.

The graphic design of this module can be adapted to meet the visual identity of the client.

ADMINISTRATION

Retrievio is accompanied by an Administration module that allows the systems administrator to configure and manage the entire system. Among the available operations, it is possible to list, modify and add new sources of information for aggregation, as well as configure their aggregation parameters.

It also allows users to view aggregation statistics, problem reports, growth statistics, as well as other indicators.

REST API

In addition to supporting all the functionality of the previous 2 modules, the REST API broadens horizons by enabling the system to be integrated with other applications. This API enables third parties to control Retrievio and its information through programmatic interfaces.

Based on standards and best practices, Retrievio's REST API has been documented using OpenAPI (also known as Swagger), making its documentation available online in a format that is easy to read and test.

Retrievio's REST API is completely secure as only authenticated users can use the API using JWT authentication method.

OAI-PMH DATA PROVIDER

Retrievio collects and concentrates information from several sources of information. This information is validated, cleaned and properly indexed to support searches done via the Discovery module.

Once the previously dispersed information is collected, Retrievio can also provide information to other content aggregation portals (e.g. Europeana, Archives Portal Europe, OpenAire, etc.).

This module enables the creation of a single channel for providing information to other portals, avoiding the need to establish individual connections for each available data source.

This module makes information available through the OAI-PMH protocol or through files for sending via asynchronous mechanisms (e.g. FTP).

WEB CONTENT ACCESSIBILITY

The Web Content Accessibility Guidelines (WCAG) 2.0 are a set of recommendations issued by W3C that aim to make Web content more accessible. Compliance with these guidelines makes content published on the Web more accessible to people with disabilities, such as blindness and low vision, hearing loss and poor hearing, learning disabilities, cognitive limitations, movement limitations, speech impairment, photosensitivity, and others.

Following these guidelines also allows Web content to become more usable by users in general and by mobile devices such as smartphones, tablets, or wristwatches.

Given the importance of this issue, legislation was created to promote the adoption of these guidelines throughout public bodies of the European Union - *Directive (EU) 2016/2102 of the European Parliament and of the Council of 26 October 2016 on the accessibility of the websites and mobile applications of public sector bodies.*

KEEP SOLUTIONS supports this initiative and ensures that all of its products are in full compliance with the AA+ level of the Web Content Accessibility Guidelines (WCAG) 2.0.

Ideal for creating **networks of information sources**

Retrievio can support networks of heterogeneous information sources in a local, national or international context.

In a local context, for instance, it can concentrate cultural and heritage assets from a single municipality. In a national or international context, the portal can provide access to information collected from different organisations from different countries.

TECHNICAL REQUIREMENTS

Retrievio can be installed in any Linux server. All application modules are Web-based, so client workstations only require a browser to access the application user interfaces (e.g., Internet explorer, Chrome, Firefox, etc.).

The following table describes the technical characteristics of a server capable of hosting a Retrievio instance with 8 million records.

SERVER

RAM	8 GB 16 GB recommended
CPU	Intel Quad-Core or superior
HDD	100 GB Depends on the total number of records and their growth rate
Operating system	Ubuntu 20.04 LTS or compatible No licensing costs
Software	Docker engine No licensing costs
Network	100 Mbit/s or superior 1 Gbit/s recommended

WORKSTATION

RAM	4 GB
CPU	Intel Dual-Core or superior
Monitor	1280x768 pixels or superior
Operating system	Windows/Linux/MacOS
Software	Web browser
Network	100 Mbit/s or superior 1 Gbit/s recommended



www.keep.pt



+351 253 066 735



info@keep.pt



sales@keep.pt



KEEP SOLUTIONS, LDA.
Rua Rosalvo de Almeida, n° 5,
4710-429 Braga
Portugal

KEEP SOLUTIONS

KEEP SOLUTIONS is a company whose mission is to provide advanced solutions for information management and digital preservation.

Our approach consists in providing software and services to allow our customers to make a more efficient management of their information assets.

The company started its activity in 2008, having acquired the status of academic spin-off of the University of Minho, for being a business initiative with strong bonds with research centres and departments from this institution.

Our clients are mostly found in the public sector, more specifically in the areas related to archives, libraries and museums.

We invest in the continuous development of innovative solutions. To support that, we remain active in the production of scientific knowledge while engaging in large-scale R&D projects in cooperation with national and international institutions.