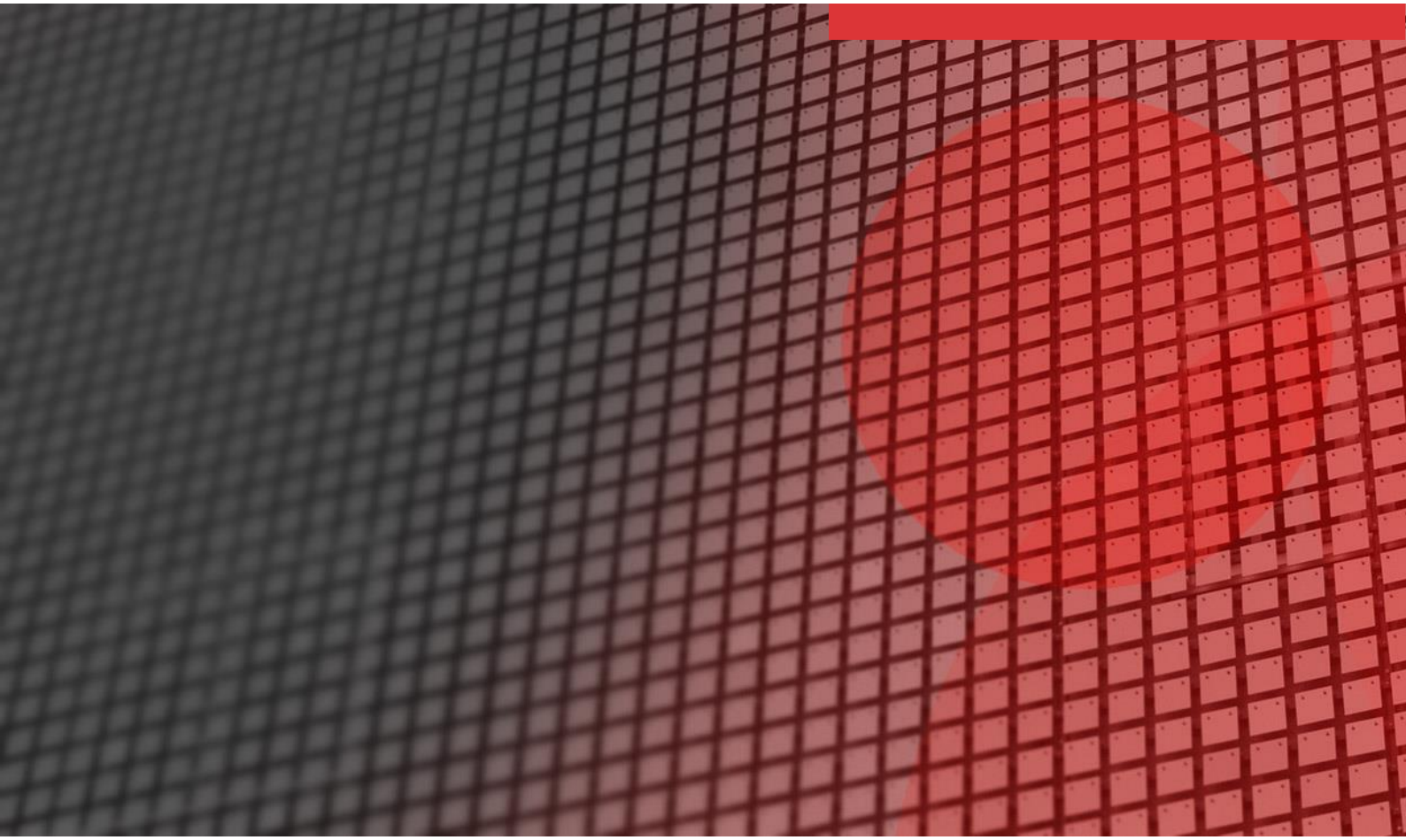
keep.
Preserving the future

RETRIEVO 2



CONTENT AGGREGATION AND FEDERATED SEARCH

CHARACTERISTICS AND TECHNICAL REQUIREMENTS

ABOUT THIS DOCUMENT

Identifier	WP181117		
Approved by	Luís Miguel Ferros	**Approved on**	2018-06-08
Classification	Public		
Distribution	N/A		

REVISIONS

#	Date	Authors	Modifications
1	2018-06-05	Miguel Ferreira	First version of the document

EXECUTIVE SUMMARY

Retrievo is a software capable of collecting and providing access to heterogeneous information assets from many different sources of information 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 uniform search interface.

Retrievo 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 in a given 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.

RETRIEVO

SINGLE POINT OF ACCESS TO ALL INFORMATION ASSETS IN YOUR ORGANISATION

Retrievo is a software capable of collecting and providing access to heterogeneous information assets from many different sources of information 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 uniform search interface.

Retrievo 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 in a given organisation.

After discovering information in Retrievo, 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

Retrievo 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.

Increase your institution's visibility

By using Retrievo, a particular institution can make their public information available on the Internet.

In addition to consolidating information from multiple systems in a single place, it provides users mechanisms that facilitate information search and retrieval in a uniform way.

Make heterogeneous systems more manageable

Retrievo 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.).

All your information assets under control

Retrievo provides a wide set of management 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.

Quality of the information ensured

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

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

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

ARCHITECTURE AND APPLICATION MODULES

Retrievo is composed of 3 application modules. The following figure depicts the overall architecture of the product, as well as its technology stack.

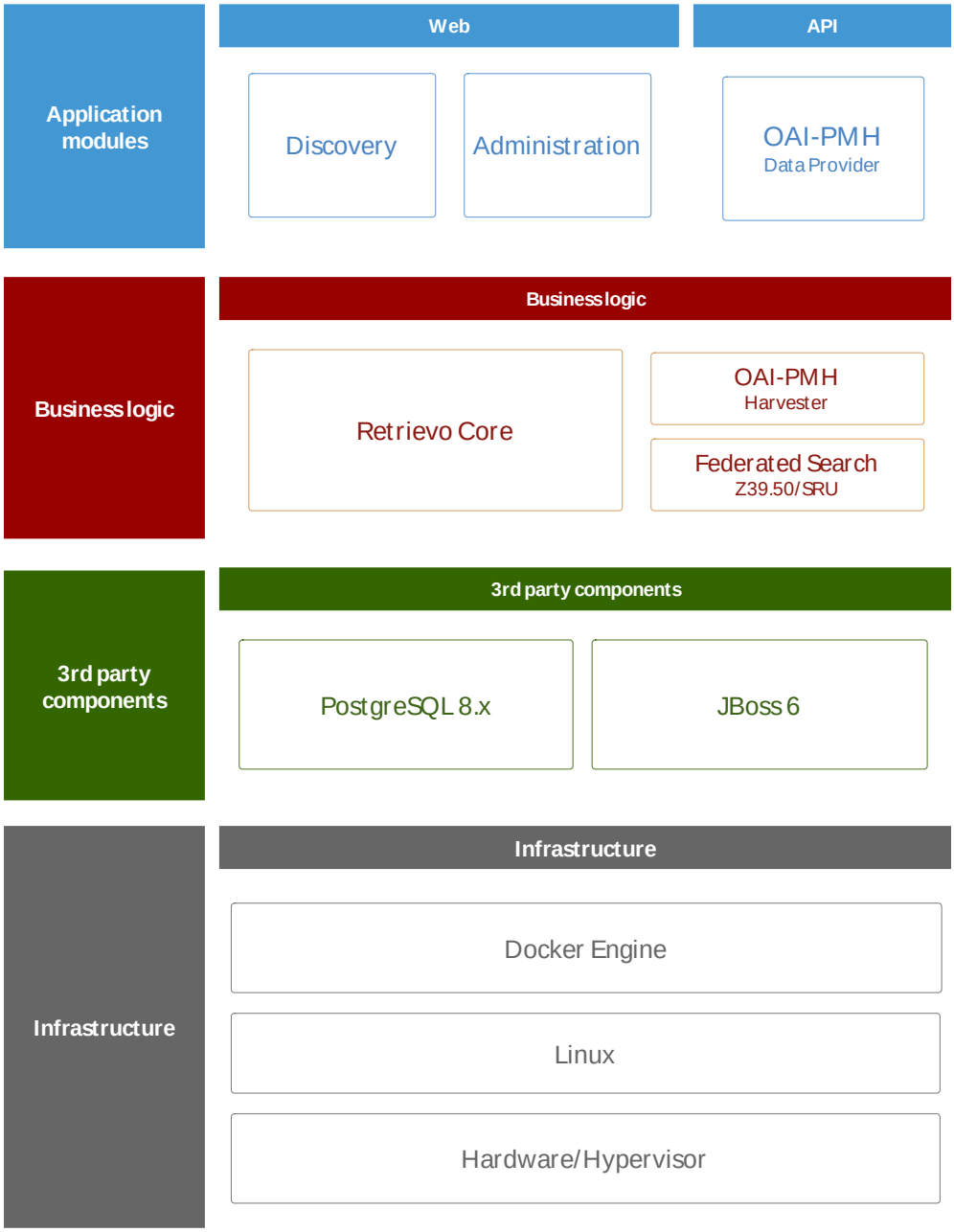


Figure 1 - Retrievo overall architecture.

DISCOVERY

The Discovery module provides the user with a single access point for finding and retrieving information originating from different sources. The main objective is to facilitate the 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, carry out more specific filtering using existing information fields, define which information sources to search on, or by combining all of the above. Search results are consolidated and presented in uniform way.

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 identify this information, a thumbnail of the associated document can also be displayed.

After locating the sought information, users can view records on their original system, with full metadata being shown, in the right context and with access to additional services, if available.

Answers at the **blink of an eye**

Retrievo is the right product to deal with the retrieval of information in environments that gather large volumes of information records.

Retrievo is capable of searching tens of millions of records without any decrease in performance due to its advanced text search engine.

ADMINISTRATION

Retrievo 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.

OAI-PMH DATA PROVIDER

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

Once the previously dispersed information is collected, Retrievo can provide information to other content aggregation portals using the OAI-PMH Data Provider module (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.

Take your information **even further**

Information collected by Retrievo can be disseminated through the Internet using standardised protocols, greatly increasing the visibility of the institution and its most valuable information assets.

Retrievo may provide information to other content aggregation portals via its OAI-PMH Data Provider module. Retrievo facilitates the process of adhering to international content aggregation platforms such as Europeana or Archives Portal Europe.

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**

Retrievo 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.

COMPATIBILITY MATRIX

Retrievo is able to collect and provide access to information from various data sources using two different methods:

1) Aggregation
2) Federated search

The first method is characterised by the fact that the records are periodically collected and indexed in a centralized database. Subsequent searches are extremely fast because records are already collected and indexed.

The second method performs real-time searches on the remote information systems that contain the original data. This makes the response time depend on how long these systems take to provide results. In any case, results are always cached so that subsequent similar searches become instantaneous.

The following tables provide information about the compatibility between Retrievo and several other sources of information. If you have a system not listed herein, please contact us.

ARCHIVAL MANAGEMENT SYSTEMS

DATA SOURCE	METHOD	PROTOCOL	FORMAT
Atom	Aggregation	OAI-PMH	OAI_DC
X-Arq	Aggregation	OAI-PMH	OAI_DC
Archeevo	Aggregation	OAI-PMH	OAI_DC or EAD
DigitArq	Aggregation	OAI-PMH	OAI_DC or EAD

INTEGRATED LIBRARY SYSTEMS

DATA SOURCE	METHOD	PROTOCOL	FORMAT
Koha ILS	Aggregation	OAI-PMH	OAI_DC or MARCXML
	Federated search	Z39.50	OAI_DC or MARCXML
SirsiDynix Horizon	Aggregation	OAI-PMH	OAI_DC

SCIENTIFIC REPOSITORIES

DATA SOURCE	METHOD	PROTOCOL	FORMAT
DSpace	Aggregation	OAI-PMH	OAI_DC
EBSCO (special licensing applies)	Federated search	Z39.50	-
World Bank Publications	Federated search	HTTP	-
ProQuest	Federated search	Z39.50	-
Scopus	Federated search	HTTP	-
ICPSR	Federated search	HTTP	-

OTHER TYPES OF DATA SOURCES

DATA SOURCE	METHOD	PROTOCOL	FORMAT
Excel	Aggregation	File import	CSV
Database	Aggregation	File import	CSV
	Federated search	DB Driver	-
SOAP Web services	Federated search	HTTP	-
REST Web services	Federated search	HTTP	-

TECHNICAL REQUIREMENTS

Retrievo 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 Retrievo 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 16.04 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

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.

www.keep.pt
+351 253 066 735
info@keep.pt
sales@keep.pt
KEEP SOLUTIONS, LDA.
Rua Rosalvo de Almeida, nº 5,
4710-429 Braga
Portugal