KEEPSOLUTIONS
University of Minho SPIN-OFF

RETRIEVO

PORTAL AGREGEDOR DE CONTEÚDOS E PESQUISA FEDERADA

Características e requisitos técnicos

WHITE PAPER SERIES

Sobre o documento			
Identificador	WP11148.2		
Aprovado por	Luís Miguel Ferros	**Aprovado em**	2014-04-08
Classificação	Público		
Distribuição	N/A		

Revisões			
#	**Data**	**Autor**	**Alterações**
1	2011-09-23	Miguel Ferreira	Documento inicial
2	2014-04-08	Luís Miguel Ferros	Revisão geral das caraterísticas e vantagens

SUMÁRIO EXECUTIVO

Este documento tem como objetivo descrever as principais vantagens, características e funcionalidades do software RETRIEVO®, bem como os requisitos técnicos necessários à sua correta instalação.

DESCRIÇÃO DA SOLUÇÃO

O RETRIEVO® é um portal de pesquisa federada que representa um ponto de acesso único a todos os recursos de informação da sua instituição. O RETRIEVO® é capaz de pesquisar o seu catálogo da biblioteca, repositório institucional, sítio Web, bem como qualquer outra fonte de informação ou base de dados remota.

O RETRIEVO® disponibiliza funcionalidades de pesquisa avançada, filtragem de resultados, tags e comentários, RSS feeds, integração com redes sociais, etc. A consulta da informação é enriquecida pela apresentação de miniaturas dos documentos e pela possibilidade de consultar toda a metainformação associada ao documento.

O RETRIEVO® utiliza um índice local para aumentar a velocidade das pesquisas, armazenando localmente todos os resultados para um maior desempenho no acesso aos dados. Os resultados são consolidados e classificados por relevância, título, autor, assunto, descrição, editora, ano ou colaborador.

Este portal é compatível com os protocolos: OAI-PMH, Z39.50, SRU, serviços SOAP, conetores SQL, bem como qualquer outro gateway de acesso aos dados (e.g. EBSCO, ABI, ProQuest, etc.).

Para além de agregadores de conteúdos, estes portais actuam também como fornecedores de informação (OAI-PMH data provider e SRU) podendo ser integrados com agregadores internacionais como a Europeana, Driver, Repositório Europeu e APEnet ou serviços de pesquisa federada como a b-on.

VANTAGENS

INFORMAÇÃO DISPERSA NUNCA MAIS SERÁ UM PROBLEMA

O RETRIEVO® permite localizar informação que se encontra dispersa por vários sistemas de informação, mesmo que à partida estes sejam incompatíveis. A pesquisa é sempre realizada através de uma interface gráfica comum evitando a necessidade do utilizador aceder a cada um dos sistemas individualmente para localizar uma informação.

MILHÕES DE ARTIGOS CIENTÍFICOS À DISTÂNCIA DE UM CLIQUE

O RETRIEVO® é o produto perfeito para instituições de investigação. Este poderá vir pré-carregado com informação de mais de 180 repositórios em acesso aberto classificados pelo *World Ranking of Repositories* reunindo assim mais de 5 milhões de artigos científicos. O RETRIEVO® permite a localização de documentos tanto pela sua metainformação de elevada qualidade como pelo conteúdo dos mesmos.

SISTEMAS HETEROGÉNEOS PASSAM A FALAR ENTRE SI

O RETRIEVO® é compatível com vários protocolos de comunicação normalizados (e.g. OAI-PMH, Z39.50, SRU), bem como como serviços SOAP, conectores SQL e um conjunto alargado de gateways de acesso a bases de dados de artigos científicos (e.g. EBSCO, ABI, ProQuest, etc.). Independentemente dos protocolos implementados ou do sistema que suporta a informação que procura, o RETRIEVO® será capaz de a indexar.

A QUALIDADE NO CENTRO DAS ATENÇÕES

Para garantir a qualidade da informação recolhida, o RETRIEVO® dispõe de uma ferramenta de verificação de conformidade que lhe permitirá validar a informação recolhida de acordo com critérios bem estabelecidos. Os critérios de qualidade são totalmente configuráveis pelo gestor do sistema e adaptáveis a qualquer situação ou contexto organizacional.

GESTÃO EFICAZ DOS SEUS ATIVOS DE INFORMAÇÃO

O RETRIEVO® disponibiliza um vasto conjunto de estatísticas e relatórios de gestão que refletem o estado de operação do sistema. Entre estes encontram-se estatísticas de crescimento de cada fonte de informação, distribuição do nº de registos por base de dados recolhida, registos e fontes mais consultadas, crescimento mensal do acervo, relatórios de agregação e validação, estatísticas de acesso, entre outros.

RÁPIDO COMO UM PISCAR DE OLHOS

O aumento do número de registos na instância central do RETRIEVO®, resultado da constante incorporação de novas fontes de informação obrigou-nos a criar novas soluções de engenharia capazes de suportar eficazmente grandes volumes de informação. Com capacidade para pesquisar em dezenas de milhões de registos sem qualquer diminuição da performance, o RETRIEVO® é o produto certo para lidar com os seus problemas de grande dimensão.

ATRAENTE TAMBÉM POR FORA

A interface gráfica do RETRIEVO® é inteiramente parametrizável podendo ser configurada através do módulo de administração para ir ao encontro da identidade gráfica preferida pelo cliente. Destacamos o exemplo de dois agregadores baseados no RETRIEVO®: o Repositório Científico de Acesso Aberto de Portugal[1] e o Portal Português de Arquivos[2].

OS SEUS ATIVOS DE INFORMAÇÃO NUNCA FORAM TÃO VALIOSOS

O RETRIEVO® é produto ideal para localizar informação no interior da sua instituição. Recorrendo ao seu elevado número de conectores é possível configurar o RETRIEVO® para pesquisar pastas partilhadas, localizar registos no ERP da sua empresa, ou encontrar as últimas notícias publicadas no sítio Web da sua instituição. Tudo isto, na segurança da sua rede interna.

[1] http://www.rcaap.pt

[2] http://portal.arquivos.pt

CARACTERÍSTICAS

A Figura 1 apresenta os vários módulos funcionais do portal RETRIEVO®, nomeadamente, o módulo de verificação de conformidade, módulo de agregação, módulo de pesquisa e módulo de administração. A figura ilustra também alguns exemplos de sistemas que poderão agir como *data providers* na rede de arquivos aderentes e os principais atores que poderão interagir com o sistema.

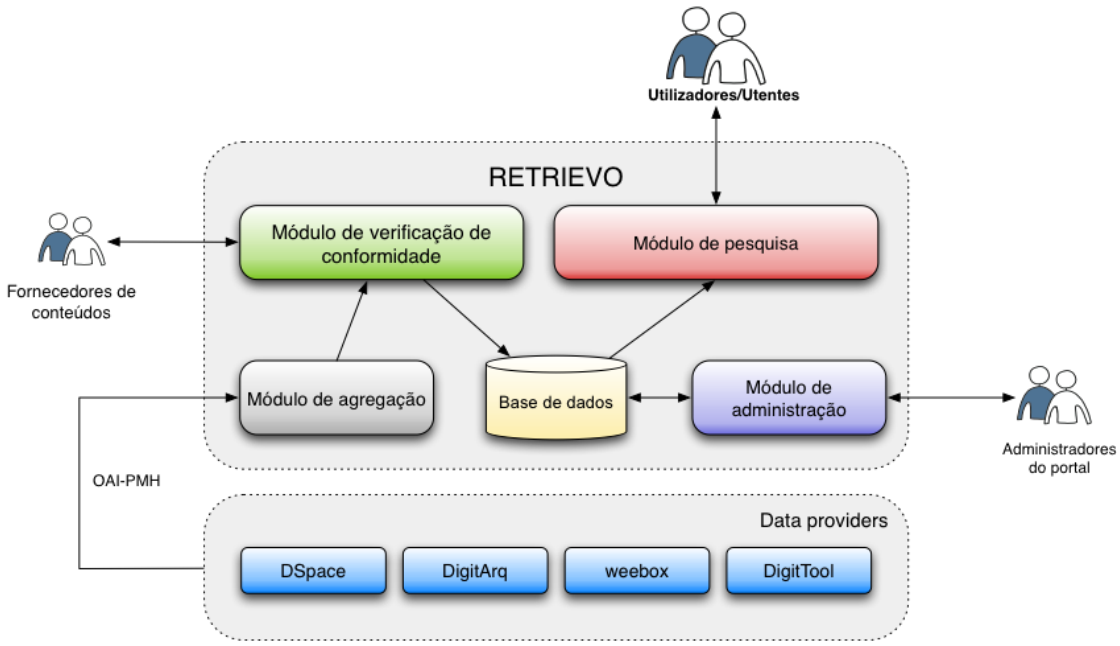

Figura 1 - Arquitetura detalhada do RETRIEVO®.

DATA PROVIDER

O RETRIEVO® é alimentado com informação proveniente de vários repositórios geograficamente distribuídos (i.e. os *data providers*). Os repositórios que fornecem conteúdos ao portal devem cumprir as diretrizes definidas pelo administrador do portal. O cumprimento dessas diretrizes será verificado recorrendo a uma ferramenta designada Módulo de verificação de conformidade.

Para que um repositório de conteúdos possa ser agregado pelo portal RETRIEVO®, este deve respeitar o formato de metainformação escolhido, bem como as diretrizes de agregação (regras de preenchimento obrigatório de metadados). Para esse efeito, um

repositório deverá dispor de *crosswalks* (i.e. mapeamentos) que assegurem a equivalência entre os seus formatos de metainformação internos e o formato exigido pelo RETRIEVO®. O portal está pré-configurado para aceitar metadados no formato OAI_DC[3].

MÓDULO DE VERIFICAÇÃO DE CONFORMIDADE

De modo a suportar/apoiar a adesão de novos fornecedores de conteúdos, o RETRIEVO® disponibiliza uma ferramenta de verificação de conformidade. Esta ferramenta pressupõe o registo de informação sobre o repositório a validar, nomeadamente, o nome do repositório, a instituição que o administra e quais as suas interfaces Web e OAI, bem como o *Set* a recolher para efeitos de verificação de conformidade.

Depois de efetuado o registo, um processo assíncrono é responsável por recolher todos os metadados do repositório e validá-los segundo o conjunto de diretrizes definidas pelo administrador do portal (Figura 2).

Figura 2 – Ferramenta de verificação de conformidade.

[3] http://www.openarchives.org/OAI/2.0/oai_dc.xsd

O conjunto de diretrizes a validar depende do domínio de aplicação. A sua implementação carece de análise no sentido de se elaborar um documento onde são definidas as regras de validação. Após a elaboração desse documento, o RETRIEVO® será configurado para validar todos os registos recolhidos de acordo com as regras de validação definidas.

Após a validação, é produzido um relatório detalhado contendo, para além de uma listagem de todas as anomalias encontradas, um conjunto de estatísticas que poderão ser úteis ao gestor do repositório de conteúdos. Após este processo, o relatório é enviado por e-mail para quem solicitou a validação e para os administradores do portal.

MÓDULO DE AGREGAÇÃO

O Módulo de agregação é responsável por recolher periodicamente os conteúdos fornecidos por cada repositório registado no portal. A agregação (i.e. *harvest*) é realizada de acordo com o protocolo OAI-PMH.

Faz também parte deste processo, a verificação de conformidade da metainformação recolhida e a sua adaptação de modo a alimentar adequadamente o Módulo de pesquisa do RETRIEVO®.

MÓDULO DE PESQUISA

O RETRIEVO® incorpora um módulo de pesquisa que permite a localização e recuperação de conteúdos produzidos no âmbito de cada software aderente.

Este módulo permite ao utilizador recuperar registos de metainformação e ligar-se às representações digitais dos documentos descritos, desde que estes se encontrem em linha e estejam associados à metainformação recolhida. Por exemplo, caso haja imagens associadas a um registo descritivo de um documento, as miniaturas das mesmas serão recuperáveis através do portal de pesquisa. O acesso às imagens para consulta é efectuado, não de forma direta, mas através de uma ligação ao repositório que detém os dados e às suas interfaces de visualização de conteúdos.

O módulo de pesquisa permite ao utilizador realizar pesquisas inter-repositórios ou apenas em alguns repositórios. Ou seja, se um utilizador quiser apenas pesquisar no Arquivo Distrital do Porto (por exemplo) pode fazê-lo retornando apenas metainformação desse repositório. Se quiser pesquisar no Arquivo Distrital de Aveiro e na Câmara Municipal do Corvo, deve poder fazê-lo. Se quiser pesquisar em todos os repositórios simultaneamente, também poderá fazê-lo.

A Figura 3 apresenta o mapa de navegação no portal RETRIEVO®.

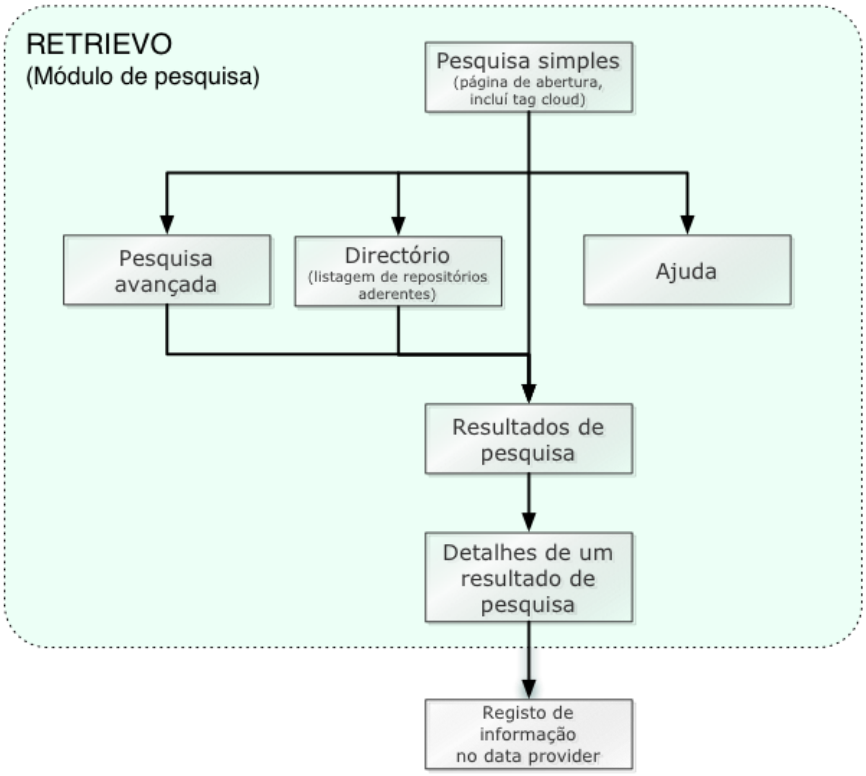

Figura 3 – Mapa de navegação do portal RETRIEVO®.

Este módulo calcula e apresenta também uma *tag cloud* com os termos de pesquisa mais utilizados pelos seus utilizadores.

Figura 4 - Página de abertura do RETRIEVO®.

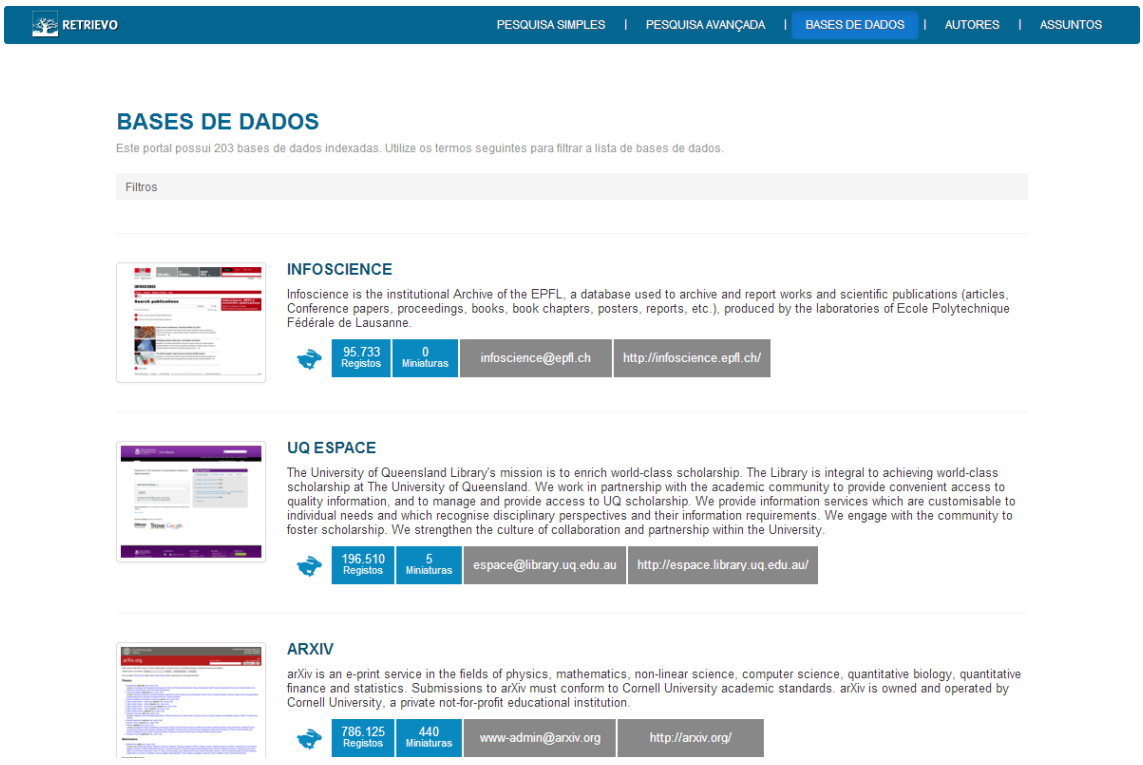

Figura 5 – Lista das fontes de informação do RETRIEVO®.

MÓDULO DE ADMINISTRAÇÃO

O portal RETRIEVO® é acompanhado de um Módulo de administração que permite ao gestor do portal realizar as seguintes operações:

1. Consultar estatísticas de agregação;
2. Consultar relatórios de problemas detetados durante a agregação;
3. Listar, adicionar, modificar novos repositórios para agregação e disponibilização nas interfaces de pesquisa, bem como configurar os seus parâmetros de agregação;
4. Consultar estatísticas de crescimento do portal e de cada repositório individual;
5. Consultar os 10 registos mais visualizados;
6. Consultar os 10 fornecedores de conteúdos mais visualizadas;
7. Consultar um resumo dos erros de agregação verificados:
    a. Data de alteração do registo inválida
    b. Sem nível de descrição
    c. Código de referência ou identificador inválido
    d. Sem título
    e. Datas extremas inválidas
    f. Idioma inválido ou inexistente
    g. Sem datas extremas
    h. Nível de descrição desconhecido
8. Consultar indicadores variados, como:
    a. N.º de entidades detentoras
    b. N.º de entidades detentoras ativas
    c. N.º de registos
    d. N.º de registos visíveis
    e. N.º de registos com problemas ligeiros
    f. % de registos com problemas ligeiros
    g. N.º de registos recolhidos
    h. N.º de registos aceites
    i. N.º de registos rejeitados
    j. % de registos rejeitados

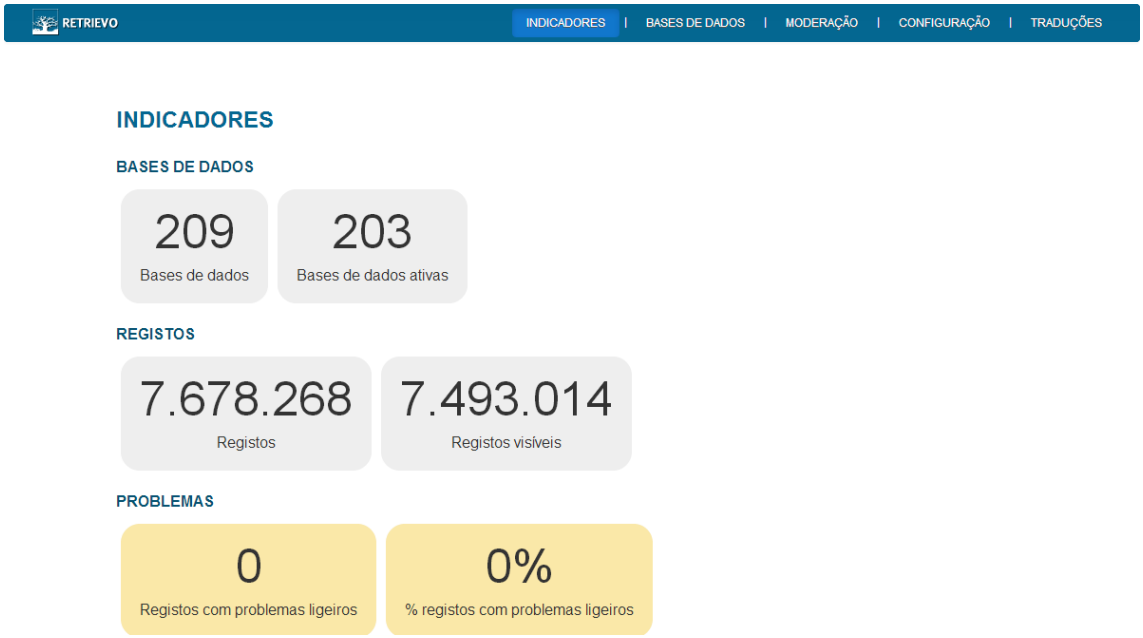

Figura 6 - Painel de administração, secção de indicadores.

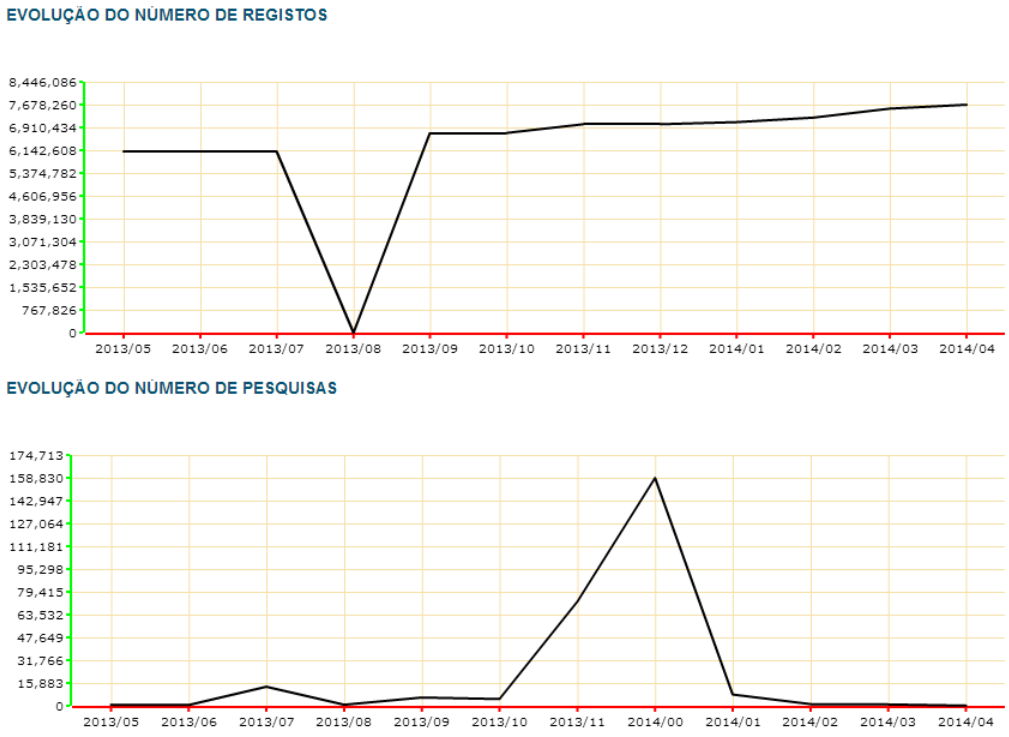

Figura 7 - Painel de administração, evolução do número de registos e pesquisas.

REQUISITOS TÉCNICOS

O portal RETRIEVO® pode ser instalado em qualquer servidor Linux com características de *hardware* convencionais. A Tabela 1 ilustra as características técnicas de um servidor capaz de suportar o portal RETRIEVO® num ambiente em que são agregados vários milhões de registos de informação.

Atributo	Valor
Hardware	CPU 2 x CoreQuad 2.0 GHz 4 GB RAM Disco 70GB (em RAID 1)
Sistema operativo	Ubuntu Server 9.10 x64
Versão do Kernel	Linux 2.6.31-14-server x86_64
Base de dados	PostgreSQL 8.x
Servidor aplicacional	JBoss 5.x
Servidor MTA	Postfix 2.6.x

Tabela 1 – Características técnicas do servidor.

KEEP SOLUTIONS

A KEEP SOLUTIONS é uma empresa de base tecnológica nascida no seio da Universidade do Minho que oferece um conjunto de produtos e serviços na área da gestão e preservação de informação digital.

A KEEP SOLUTIONS tem-se especializado na prestação de serviços de consultoria em preservação digital, recuperação de suportes, migração de dados, digitalização, análise e concepção de sistemas de informação, manutenção, alojamento e suporte de repositórios digitais e no desenvolvimento de soluções para publicação electrónica.

A estreita ligação que a KEEP SOLUTIONS mantém com a Universidade do Minho garante-lhe acesso privilegiado às mais recentes linhas de investigação desenvolvidas a nível nacional e internacional. Tratando-se de uma spin-off académica, faz parte da sua missão, transformar conhecimento científico em produtos de valor acrescentado adaptados às necessidades do mercado, contribuindo, assim, para o desenvolvimento e competitividade dos seus clientes.

Tendo nascido de uma plataforma de I&D, a KEEP SOLUTIONS permanece ativa na produção de conhecimento científico. Prova disso são as inúmeras publicações e participações em eventos científicos onde os seus colaboradores têm marcado presença.

Endereço Web	http://www.keep.pt
Telefone	+351 253 066 735
Fax	+351 253 067 248
Correio-electrónico	info@keep.pt
Orçamentos	sales@keep.pt
Morada	KEEP SOLUTIONS, LDA. Rua Rosalvo de Almeida, nº 5 4710-429 Braga, Portugal
NIPC	508 496 870