A Gestão de Entidades dos Repositórios aos Agregadores - o Caso LA Referencia e RCAAP

Autores

José Carvalho, University of Minho, jose.carvalho@usdb.uminho.pt;

Lautaro Matas, LA Referencia, lautaro.matas@redclara.net;

Washington Segundo, IBICT, washingtonsegundo@ibict.br ;

Paulo Graça, FCT|FCCN, paulo.graca@fccn.pt ;

Paulo Lopes, FCT|FCCN, plopes@fccn.pt

Resumo da Proposta

Esta proposta demonstra o trabalho que tem vindo a ser desenvolvido ao nível do agregador para incorporar o conceito de entidades e as suas relações existentes nos repositórios. Para o conseguir, foram implementadas diretrizes de interoperabilidade técnica que permitem a coexistência de repositórios com e sem entidades e diferentes tipos de agregação conforme os esquemas de metadados adotados localmente. Foram ainda definidos novos processos de agregação, processamento e indexação no sistema. O objetivo desses desenvolvimentos é oferecer suporte a uma representação completa do modelo de dados do repositório ao nível do agregador e fornecer serviços de valor agregado para todo o conteúdo incorporado.

Tipo de Proposta

- Comunicação

Tema da Conferência

Indique os temas abordados na sua proposta (remova os que não se aplicam):

- **Acesso Aberto e Dados de Investigação Abertos: sistemas, políticas e práticas**

o Repositórios digitais – institucionais, temáticos, de dados de investigação ou de património cultural

o Inovação na comunicação científica para a Ciência Aberta

o Modelos e padrões de metadados

- **Gestão de informação de Ciência e Tecnologia**

- CRIS – Sistemas de Gestão de informação de Ciência e Tecnologia
- Interoperabilidade entre sistemas de informação de apoio à atividade científica e académica
- Normas e diretrizes
- Identificadores persistentes

Palavras-chave

repositórios; agregador; entidades; interoperabilidade; diretrizes

Audiência

gestores de repositórios, bibliotecários, gestores de dados de investigação, programadores, decisores políticos, gestores de ciência, gestores de tecnologias de informação (programadores, administradores de sistemas e gestores de tecnologias de informação).

Proposta

Introdução

Este trabalho apresenta os resultados de um esforço colaborativo para estender a plataforma LA Referencia [1], no sentido de disponibilizar uma ferramenta para a criação de agregadores nacionais e regionais a partir de repositórios digitais, flexíveis e abrangentes com modelos de dados de entidades.

A evolução dos sistemas digitais e a necessidade de descrições dos registos mais pormenorizadas resultaram na criação de diretrizes de metadados como as diretrizes OpenAIRE para repositórios de literatura, versão 4 [11]. Estas consideram já alguns campos do registo como objetos mais ricos, por exemplo, uma descrição do autor pode ter mais atributos do que um simples "nome". Essa descrição pode conter ainda identificadores de autor, afiliação, entre outras informações relevantes que não podem ser facilmente expressas como textos simples ou como simples registos num campo.

Desse ponto de vista, a descrição de um autor é mais do que um campo simples, é designada de "entidade" que pode manter relações com outras entidades referenciadas, como Organizações ou Publicações. Nesse contexto, um modelo de dados pode ser visto como um grafo, onde nós e arestas representam, respectivamente, entidades e relações. Essa foi exatamente a abordagem adotada, por exemplo, pelo OpenAIRE, na construção do que foi designado de *OpenAIRE Research Graph* [4].

O software de repositórios DSpace 7 foi construído sobre um novo modelo de dados de entidades conforme aos requisitos das diretrizes OpenAIRE v4 [12] e permitirá uma interoperabilidade muito mais rica com outros sistemas. Essa característica é a base para a construção de sistemas agregadores e grafos de pesquisa a nível nacional, regional e internacional.

A partir desta nova geração de repositórios digitais, será possível coletar dados complexos e alimentar agregadores nacionais e regionais como o oasisbr [5], o Portal RCAAP [6] e o LA Referencia [7]. No entanto, esses sistemas também devem ser estendidos para explorar a expressividade dos dados agregados no modelo de dados da entidade. Além disso, esses sistemas agregadores podem servir para a implementação de sistemas CRIS nacionais em Portugal, Brasil e Peru (respectivamente designados PTCRIS, BRCRIS e PerúCRIS) e a interoperabilidade adjacente entre agregadores e plataformas curriculares, como o CiênciaVITAE [8], o CTI Vitae [9] e a Plataforma de Currículos Lattes [10].

Entidades no Repositório

O DSpace e a sua comunidade técnica, com o apoio do OpenAIRE, têm trabalhado no lançamento do DSpace 7 que fornecerá novas funcionalidades importantes que possibilitam a conformidade com estes novos requisitos. O Grupo de Trabalho de Entidades do DSpace trabalhou na introdução do conceito de entidades para facilitar a sua conformidade, enquanto o novo Grupo de Trabalho do DSpace-OpenAIRE liderado pela FCT|FCCN que se concentrou em tornar realidade o cumprimento de requisitos das diretrizes do OpenAIRE 4 no DSpace 7. Este trabalho criou as bases para melhorar a integração dos repositórios no ecossistema de gestão de informação científica. Está prevista a disponibilização do novo DSpace 7 para junho de 2021 e está já previsto um teste alargado pela comunidade à versão beta 5 entre abril e maio de 2021.

Entidades no Agregador

O software LA Referencia (LRHarvester 3.5) é uma plataforma para coleta, validação e transformação de metadados (enriquecimento / curadoria), atualmente instalado em dez nós nacionais da América Latina e que periodicamente agrega e processa mais de 1,9 milhões de registros de metadados de repositórios regionais. Além disso, os principais componentes deste software estão a ser usados como parte do Portal de Pesquisa RCAAP em Portugal para gestão do processo de agregação.

O esquema de metadados "mais rico" introduzido pelas diretrizes do OpenAIRE v4 e o modelo de entidades do DSpace 7 motivaram a evolução do LA Referencia (LRHarvester) para uma nova arquitetura. Um esforço colaborativo entre as equipas do LA Referencia, RCAAP e IBICT durante 2019 e 2020 derivou um conjunto de resultados importantes: uma nova arquitetura da plataforma LRHarvester v4, um roteiro de desenvolvimento colaborativo e uma implementação ao nível beta de um modelo de relação de entidades no agregador (Figura 1).

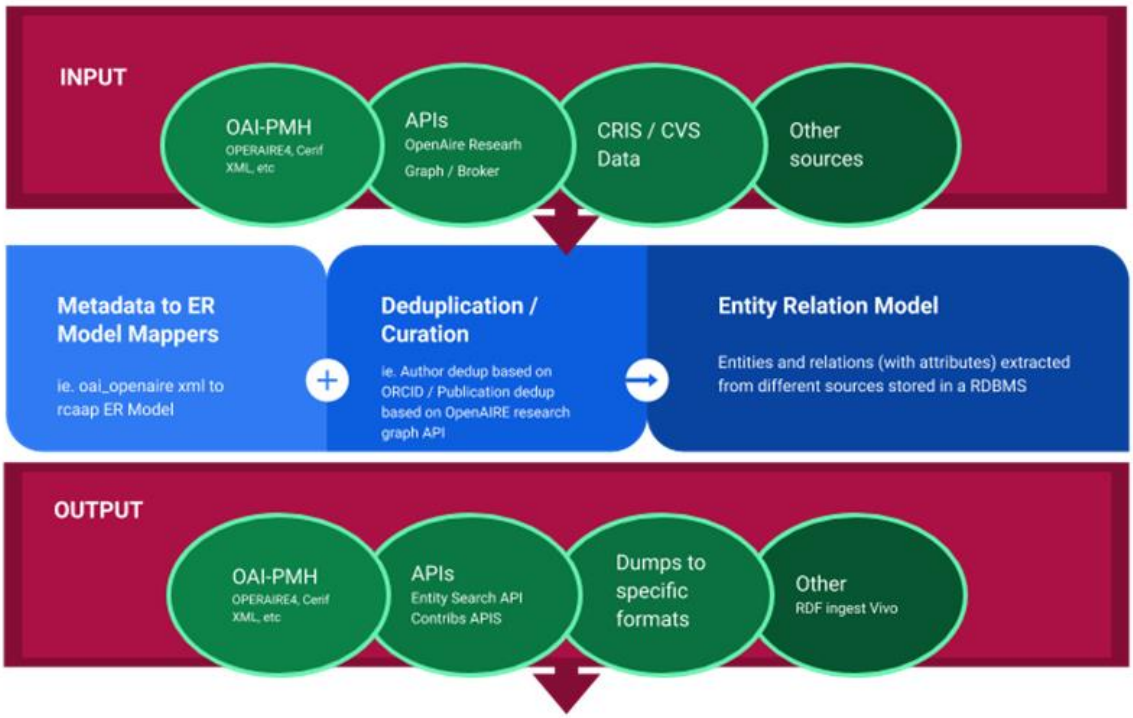

Figura 1: Organização do Software La Referencia (LRHarvester)

A plataforma LA Referencia LRHarvester 4.0 visa fornecer serviços de valor acrescentado ao nível do agregador nacional e regional, como:

- Construir, enriquecer e armazenar um modelo de relação de entidade configurável capaz de representar um ecossistema de investigação nacional / regional;
- Capturar entidades e relações de diferentes fontes de metadados (repositórios de literatura, agregadores, APIs de grafos de pesquisa, serviços de CVs nacionais, metadados de financiadores);
- Devolver informação sobre o enriquecimento de metadados resultante para as fontes originais (ou seja: repositórios);
- Criar APIs do serviço com base nos dados de relação de entidades para interoperar com outros atores do ecossistema de ciência e tecnologia (CV s, CRIS).

Como prova de conceito, foi focado o desenvolvimento no caso da entidade autor pois no caso nacional é que tem uma representatividade maior ao nível dos identificadores de autor, fruto de uma funcionalidade implementada no DSpace 5 para inclusão do ORCID e/ou Ciência ID nos repositórios.

Em 2020 e 2021, o trabalho desenvolvido focou-se principalmente na reformulação dos diferentes processos internos da aplicação para garantir um funcionamento com recursos que adotem ou não esquemas de metadados mais ricos e dessa forma garantir uma coexistência no sistema.

Figura 2: Workflow Interno do Software La Referencia (LRHarvester)

A abordagem adotada ao nível do agregador permite a integração de repositórios institucionais com esquema de metadados oai_openaire, de acordo com as guidelines OpenAIRE 4.0, mas também a integração de revistas científicas com o plugin OpenAIRE[1] que expõe os metadados sob a forma de entidades.

Serviços para Repositórios (ao nível do agregador)

Com base nesse novo contexto de informação, onde o conceito de entidades está amplamente disponível nos diferentes serviços, está aberto o caminho para novos serviços integrados para os diferentes atores.

Existem trabalhos realizados com base em cinco casos de uso de integração relacionados com tarefas de *claim* (integração de identificadores de autores), depósito em repositórios através de sistemas externos (CRIS), sincronização de metadados entre sistemas, controlo de autoridades para entidades de CRIS (autores, organizações e financiamento) e tarefas de curadoria para novas entidades.

Considerando o exemplo do sistema CiênciaVITAE (Sistema de Currículos Científicos em Portugal) com o Portal de Pesquisa RCAAP, os utilizadores podem reivindicar as suas publicações disponíveis no agregador e importá-las para o seu currículo. Além disso, podem depositar trabalhos diretamente do seu sistema de currículo para um repositório institucional.

Quando os identificadores de autor estão disponíveis, o autor pode ter sugestões sobre qualquer publicação existente no agregador associada a qualquer um de seus identificadores. Isso pode levar-nos a um contexto sincronizado de informações científicas, como já é feito pelo CiênciaVITAE e o ORCID. Além disso, esse ecossistema partilhado permite a curadoria de dados por diferentes utilizadores, pelos próprios autores e pelos gestores de repositório ou gestores de ciência ao nível da instituição. A representação das entidades por meio dos diferentes sistemas reforça o uso de serviços autoritativos com base em autores, financiamento, publicações, afiliações, revistas e conecta-os com base nas necessidades de cada serviço e de cada parte interessada.

Conformidade

Prevê-se o desenvolvimento de uma metodologia para a preparação dos repositórios DSpace para o processo de atualização para a versão 7 para garantir uma reutilização mais alargada da informação existente.

1 https://github.com/ojsde/openAIRE/blob/master/readme.md

Contudo, será necessário outros processos de curadoria, através de ferramentas do próprio repositório ou validadores externos como o Validador RCAAP ou o validador OpenAIRE. Finalmente, prevê-se ainda um trabalho adicional de ligação de entidades para os dados pré-existentes no sentido de enriquecer as ligações e a informação que é transmitida para outros serviços. Toda a nova informação inserida nos repositórios já terá como base o conceito de entidades e será por isso já devidamente organizada sem necessidade de curadoria adicional. A conformidade dos dados é relevante para a prestação de serviços de qualidade e processos de sincronização automáticos ou semiautomáticos.

Conclusão

Os repositórios e os agregadores têm funções diferentes para diferentes atores, mas podem partilhar e reutilizar as mesmas informações. O aspecto principal, para criar relações confiáveis, é a validação das relações pelos curadores, neste caso particular, pelos autores e gestores de ciência. Estes desenvolvimentos misturam a curadoria realizada em diferentes níveis e demonstram o valor acrescentado para o sistema de informação.

Ainda existem alguns desafios para alcançar a interoperabilidade total entre os serviços. A interoperabilidade semântica, fornecida pelas diretrizes do OpenAIRE 4, é um bom começo para ter uma pequena descrição de um ecossistema CERIF para permitir que mais serviços possam interoperar. Podemos, nesse contexto, fazer uma comparação com o Dublin Core como um modelo básico básico de metadados que se adapta às necessidades básicas de interoperabilidade. Nesse novo contexto, as diretrizes do OpenAIRE 4 descrevem um modelo simplificado de metadados do modelo CERIF para promover a interoperabilidade semântica entre sistemas.

Os desafios futuros serão a adaptação das informações para se adequarem aos contextos disciplinares e a obtenção de informações suficientes para permitir a geração de relatórios e a avaliação de investigadores, organizações e revistas.

Adicionalmente, estes desenvolvimentos permitirão interligar de uma forma mais automática as ligações com outros serviços existentes, como estatísticas de uso, grafos de informação científica para promover o enriquecimento dos registos agregados. Os agregadores têm também objetivos ocultos na maneira como criam comunidades locais, nacionais ou internacionais em torno dos recursos que recolhem. Este trabalho realizado com o software LAHarvester é uma consequência natural da necessidade de obter informações mais descritivas baseadas em identificadores para permitir uma melhor reutilização de informações em diferentes ecossistemas.

Referências

[1] *LA Referencia platform* < https://github.com/lareferencia >. Accessed in jan/2020.

[2] *BASE* < https://www.base-search.net/ >. Accessed in jan/2020.

[3] *CORE* < https://core.ac.uk/ >. Accessed in jan/2020.

[4] *OpenAIRE Research Graph* < https://www.openaire.eu/openaire-research-graph-open-for-comments >. Accessed in jan/2020.

[5] *oasisbr* < http://oasisbr.ibict.br/ >. Accessed in jan/2020.

[6] *RCAAP* < https://www.rcaap.pt/ >. Accessed in jan/2020.

[7] *LA Referencia search portal* < http://www.lareferencia.info/ >. Accessed in jan/2020.

[8] *CiênciaVitae* < https://cienciavitae.pt/ >. Accessed in jan/2020.

[9] CTI Vitae < https://ctivitae.concytec.gob.pe/appDirectorioCTI >. Accessed in jan/2020.

[10] *Lattes Curriculum Platform* < http://lattes.cnpq.br/ >. Accessed in jan/2020.

[11] OpenAIRE Guidelines for Literature Repository Managers v4 <https://openaire-guidelines-for-literature-repository-managers.readthedocs.io/en/v4.0.0/> Accessed in jan/2020

[12] OpenAIRE open letter to Duraspace - Request for adoption of technical recommendations <https://drive.google.com/file/d/17vPJQcOk3WBB4wkeO38K4wtZRDVsic_1/view>. Accessed in jan/2020.