Disponível em www.bad.pt/publicacoes

PAPER

12º CONGRESSO NACIONAL BAD

PT-CRIS: Um miradouro sobre o universo científico nacional

André Santos, José Carlos Ramalho, Miguel Ferreira, Hélder Silva, Luis Faria[a], João Mendes Moreira[b], Eloy Rodrigues, José Carvalho[c]

[a]KEEP SOLUTIONS, Portugal, {asantos,jcr,mferreira,hsilva,lfaria}@keep.pt
[b]Fundação para a Ciência e Tecnologia, Portugal, jmm@fccn.pt
[c]Universidade do Minho, Portugal, {eloy,jcarvalho}@sdum.uminho.pt

Resumo

Reconhecida a importância da ciência, tecnologia, inovação e do conhecimento gerado pela investigação científica, são inúmeros os sistemas de informação criados para dar resposta a necessidades de gestão e disseminação de informação em diferentes domínios.
Contudo, a dispersão de informação em vários sistemas, a não adoção de normas/práticas de referência e consequentemente a replicação de informação criam dificuldades às várias entidades que gerem ou consultam informação sobre ciência e respetivos indicadores na capacidade de gestão, execução, avaliação e tomada de decisão relativa a processos de investigação.
Surge assim a necessidade de criar um sistema que ofereça uma visão global do universo de ciência e tecnologia, agregando e relacionando informação de suporte à atividade científica desenvolvida em Portugal, i.e., informação sobre investigadores, organizações, programas de financiamento, projetos, resultados de investigação, instalações, equipamentos e serviços.
O sistema, ao relacionar e contextualizar a informação científica atualmente dispersa em vários sistemas, permitirá transformar informação em conhecimento, aumentar a visibilidade e difusão da ciência e simplificar processos na gestão da produção científica nacional.

Palavras-chave: Ecossistema, Informação, Ciência, VIVO, PT-CRIS

Introdução

A ciência, tecnologia, inovação e o conhecimento gerado pela investigação científica são de extrema importância uma vez que contribuem diretamente para o crescimento económico do país.

Em Portugal, são vários os sistemas de gestão de ciência e tecnologia que dão resposta às necessidades de gestão e disseminação de informação num ou vários domínios.

Contudo, a dispersão de informação através de vários sistemas, a não adoção de normas/práticas de referência e consequentemente a replicação de informação criam dificuldades às entidades que gerem ou consultam informação sobre o sistema científico nacional e respetivos indicadores no que diz respeito à capacidade de gestão, execução, avaliação e tomada de decisão relativa a processos de investigação.

Para dar resposta às necessidades e dificuldades identificadas, e no seguimento do PT-CRIS – projeto que tem como missão a criação e desenvolvimento de um ecossistema integrado de informação de suporte à atividade científica nacional (Moreira, 2015) – surge a necessidade de criar um sistema que ofereça uma visão global do universo de ciência e tecnologia, agregando e relacionando informação sobre investigadores, organizações, programas de financiamento, projetos, resultados de investigação,

instalações, equipamentos e serviços.

O sistema, ao relacionar e contextualizar a informação científica atualmente dispersa em vários sistemas, permite transformar informação em conhecimento, aumentar a visibilidade e difusão da ciência e simplificar processos na gestão da produção científica nacional.

Assim, planeia-se um sistema intuitivo, de fácil utilização e capaz de responder rapidamente e com precisão a perguntas como:

- Quantos artigos foram publicados por determinado autor em 2011 como primeiro autor?
- Quantas pessoas beneficiaram de emprego no decurso de projetos financiados pelo Sétimo Programa-Quadro da primeira chamada nos novos Estados-Membros?
- Quantos estudantes de doutoramento participaram em projetos de investigação nacionais? Em que países fizeram os seus doutoramentos?

Método

Neste projeto, é assegurado o planeamento e desenvolvimento de um sistema piloto, que permita avaliar a viabilidade do sistema final, na capacidade de agregar e relacionar grandes volumes de informação de várias fontes e domínios.

Em todas as fases do projeto procura-se seguir como referência iniciativas internacionais e respeitar as mais relevantes normas para identificação, interoperabilidade e representação de informação.

Numa fase inicial, são identificados os diferentes domínios, fontes e interfaces de informação e definido o modelo de dados do sistema.

De seguida, são avaliados os sistemas de software existentes que permitem a concretização dos objetivos propostos, tendo em consideração diversos fatores como o preço, aceitação, personalização, métodos de recolha e agregação de informação, mecanismos de normalização, formatos de exportação e outras funcionalidades relevantes.

Selecionado o sistema mais adequado aos objetivos dos projeto, procede-se à instalação e configuração do mesmo.

Durante a fase de inserção de informação no sistema, adota-se um método iterativo, representado na figura 1, em que em cada iteração são selecionadas as fontes e interfaces de informação para o domínio selecionado, são agregadas amostras de informação e mapeadas para o modelo de dados do sistema, identificadas relações com a informação presente no sistema e, por fim, validada a informação agregada.

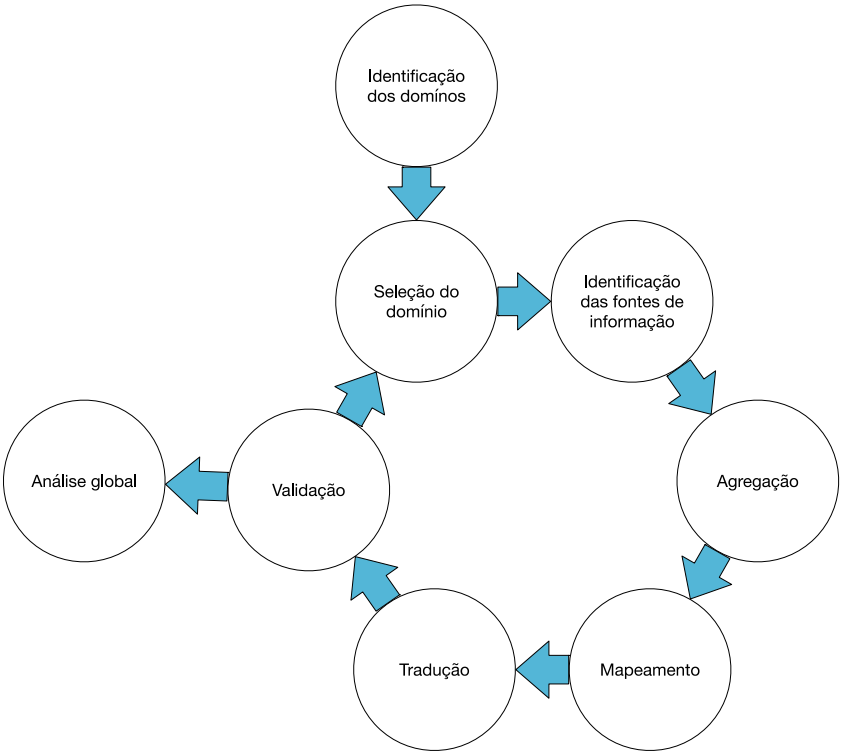

Figura 1 - Ciclo iterativo de agregação e validação de informação

Para assegurar a utilidade e relevância do sistema, é sempre preferível obter de menos informação mas completa, fidedigna e relacionada em detrimento de mais informação com menor validade.

Depois de agregados os domínios de informação pretendidos, procede-se à validação global da informação obtida e à validação da capacidade do sistema para responder às perguntas identificadas anteriormente.

Com base na validação, são revistas todas as decisões efetuadas ao nível dos domínios, fontes e interfaces de informação, modelo de dados, arquitetura e, caso se justifique, é iniciada uma nova iteração com diferentes decisões para que se consiga encontrar a melhor concretização para os objetivos do projeto.

Resultados

No decorrer do projeto, e no âmbito do projeto PT-CRIS, foram recolhidas várias informações e tomadas decisões relevantes para o desenvolvimento do projeto que são expostas de seguida.

Domínios de informação

Numa fase inicial foi necessário identificar os domínios de informação que se pretendem representar no projeto piloto, tendo em atenção a sua relevância e a disponibilidade atual de informação e respetivas fontes.

Procurou-se perceber as entidades de informação com maior importância para as funções da Fundação para a Ciência e a Tecnologia (FCT) e que domínios de informação são representados pela norma

internacional *Common European Research Information Format* (CERIF), que permite a representação de informação científica.

A FCT, apesar de se centrar no financiamento do sistema científico, esta interage com todo o universo de investigação científica, representado na figura 2, para efeitos de analise análise bibliográfica e transferência de tecnologia para a indústria (Moreira, 2015).

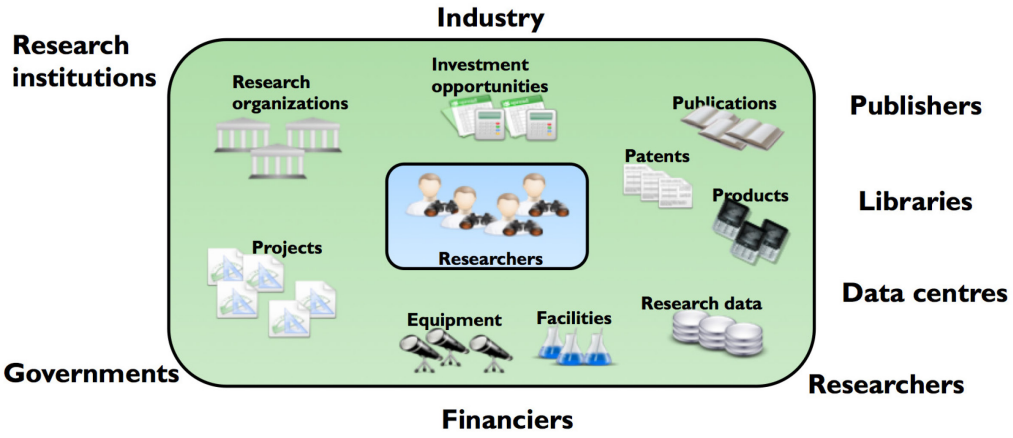

Figura 2 - Universo de ciência e tecnologia

Na figura são ilustrados ao centro os investigadores como produtores de ciência que pertencem a organizações, executam projetos, usam equipamento para as suas experiências e produzem resultados de investigação. À volta do ecossistema encontram-se outros intervenientes que representam os consumidores de informação científica.

O modelo CERIF, representado numa visão global na figura 3 (Simons & Danica Zendulková, 2013), permite representar e relacionar informação relativa a pessoas, organizações, projetos, resultados de investigação (publicações, patentes e produtos), serviços, programas de financiamento, equipamentos e informações curriculares.

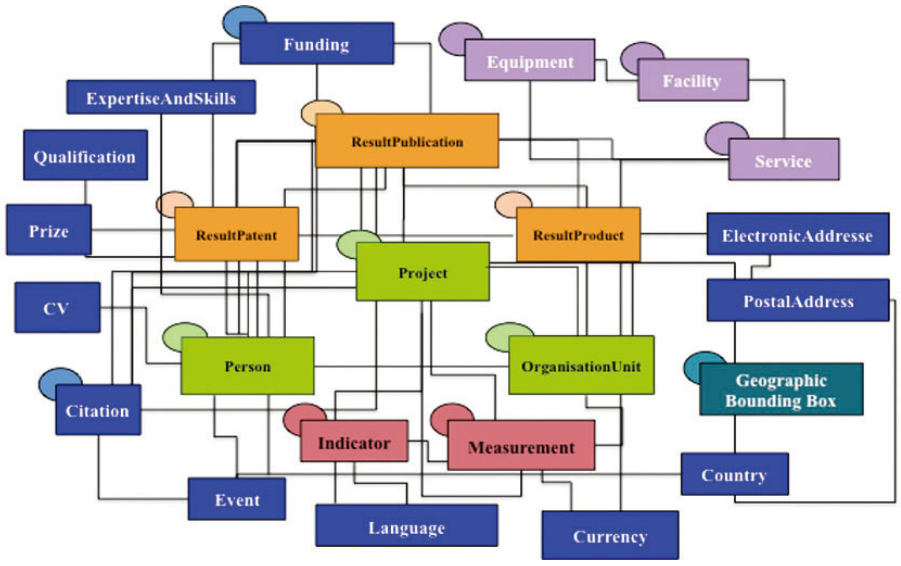

Figura 3 - Visão global das entidades e relações do modelo CERIF

Pela direta relação entre as entidades representadas pela FCT e as entidades do modelo CERIF selecionaram-se os seguintes domínios de informação:

- Investigadores
- Resultados de investigação (patentes, publicações e produtos)
- Organizações
- Financiamentos
- Infraestruturas
- Equipamento
- Projetos

Dos domínios selecionados identificam-se como domínios prioritários para o piloto a implementar os investigadores, resultados de investigação e os projetos.

Fontes de informação

Identificados os domínios de informação, prosseguiu-se para identificação das fontes de informação e respetivas interfaces para a agregação de informação.

Investigadores

No modelo que rege o ecossistema PT-CRIS, o ORCID - um esforço orientado pela comunidade científica, aberto e sem fins lucrativos para manter um registo único e universal de investigadores e uma ligação transparente às suas atividades de pesquisa - é apresentado como eixo central de produções científicas e sistema de identificação única de autores (Haak, Fenner, Paglione, Pentz, & Ratner, 2012; Moreira, 2015).

Considerado uma fonte de inequívoca de informação sobre investigadores, este sistema disponibiliza uma API REST para consulta, criação e atualização de informação.

Na reunião de trabalho do PT-CRIS, realizada em fevereiro de 2015, nas jornadas anuais da FCCN, foi apresentada a nova versão da plataforma de curricula DeGóis e que pode representar uma futura fonte de informação, uma vez que contemplará componentes, como os serviços, que não estão presentes no ORCID. O lançamento da plataforma está previsto para o inicio de setembro de 2015 (Moreira, 2015).

Publicações

São diversas as possíveis fontes de informação para obter resultados de investigação, sejam estes publicações, patentes ou produtos.

Uma possível abordagem poderia passar por agregar informações diretamente nos repositórios de acesso aberto das instituições ou em *Current Research Information Systems* (CRISs) locais. Contudo, essa abordagem representaria mais complexidade na implementação e manutenção de interfaces e agregadores para os diferentes sistemas.

Outra possível abordagem podia passar por agregar em plataformas com informação de investigadores e das suas atividades de investigação, como o ORCID ou a nova plataforma de curricula DeGóis.

Contudo, a política sobre acesso aberto a publicações científicas resultantes de projetos de I&D financiados pela FCT[1] indica que "todas as publicações sujeitas a arbitragem por pares ou a outros processos de revisão ou validação científica que incluam resultados de I&D financiados total ou parcialmente pela FCT devem ser obrigatoriamente depositadas pelos autores, em versão, pelo menos num repositório integrante da rede RCAAP – Repositório Científico de Acesso Aberto de Portugal" o que faz com que o portal RCAAP represente assim uma fonte completa de informação de publicações financiadas pela FCT.

Na sua versão atual o portal RCAAP apresenta uma interface OAI-PMH para agregação de publicações. Contudo, está em desenvolvimento uma API REST e uma API CERIF-XML que constituirá uma melhor interface de agregação uma vez que a informação será recebida num formato conhecido e poderão ser utilizados agregadores comuns a sistemas que implementem a mesma norma.

Projetos

Atualmente, é mantida pela FCT, uma folha de cálculo com informação relativa a projetos de investigação financiados completamente ou parcialmente pela mesma. Assim, é expectável que o sistema central seja capaz de processar informação presente em folhas de cálculo ou em ficheiros CSV.

Para a agregação de informação sobre projetos internacionais, podem ser utilizadas as interfaces REST ou OAI-PMH disponibilizadas pela *Open Access Infrastructure for Research in Europe* (OpenAIRE), que disponibilizam informação relativa a projetos financiados pela Comissão Europeia e pela fundação *Wellcome Trust*.

É importante referir que na nova versão do portal RCAAP, atualmente em desenvolvimento, está prevista a inclusão de informação relativa a projetos e a sua disponibilização numa API REST, com capacidade de exportação no formato CERIF-XML. Espera-se que o sistema num seja capaz de integrar com a referida interface, para que a informação sobre projetos financiados pela FCT seja atualizada num único local.

Organizações, financiamentos, infraestruturas e equipamento

Identificados como domínios de informação importantes mas menos prioritários, é relevante identificar possíveis fontes e interfaces de informação para organizações, financiamentos, infraestruturas e equipamento científico.

Para os domínios referidos, e no seguimento do projeto PT-CRIS, está em desenvolvimento um sistema de informação de infraestruturas científicas, um sistema de gestão de subvenções e uma base de dados CERIF "ponto único de verdade" que disponibilizará uma API que permitirá obter respostas em CERIF-XML para os domínios de informação do modelo.

Espera-se que a base de dados e os sistemas referidos sejam uma fonte inequívoca de informação que até agora era de difícil obtenção uma vez que os dados encontravam-se segmentados pelos vários sistemas legados existentes.

[1] *Política sobre Acesso Aberto a Publicações Científicas resultantes de Projetos de I&D Financiados pela FCT, https://www.fct.pt/documentos/PoliticaAcessoAberto_Publicacoes.pdf*

Sistema de software

Para a implementação do sistema piloto é necessário selecionar que sistema de software será adotado para concretização dos objetivos propostos, tento em consideração diversos fatores como o preço, aceitação, personalização, métodos de agregação, mecanismos de normalização de informação, formatos de exportação e serviços de valor acrescentado disponibilizados.

O *CTSA Research Networking Affinity Group* (RNAG), que apresenta um longo historial na identificação de necessidades para sistemas de gestão e disseminação de ciência e orientação para sistemas disponíveis, realizaram uma extensa análise aos sistemas atualmente disponíveis, e disponibilizaram a mesma numa página da Wikipédia[2], o que facilita a tomada de decisão relativa à seleção do sistema a adotar.

Após uma extensa análise às diversas tabelas comparativas, resultantes da análise efetuada, distingue-se o sistema VIVO como sistema mais indicado para a concretização dos objetivos propostos.

O VIVO é uma ferramenta semântica, de código aberto, que quando instalada e devidamente povoada de informação, permite a descoberta de pessoas e da investigação que estas produzem. Esta ferramenta suporta a edição, pesquisa, navegação e visualização de atividades de investigação de diferentes áreas.

Segundo as tabelas comparativas referidas, esta ferramenta distingue-se por ser uma ferramenta de código aberto, amplamente adotada, que apresenta mecanismos automáticos de ingestão de informação, implementa um grande número de tecnologias de web semântica, permite a agregação de dados de diversas fontes de informação, permite a interoperabilidade com diversos sistemas e apresenta mecanismos de desambiguação de autores.

Modelo de dados

Como referido, o VIVO é uma aplicação que implementa várias tecnologias e princípios de web semântica e permite a pesquisa sobre diferentes domínios de informação.

Uma aplicação VIVO utiliza como modelo de dados uma ontologia VIVO-ISF que permite representar investigadores contextualizados com a sua experiência, os seus resultados de investigação, interesses, feitos e organizações associadas.

Esta ontologia utiliza elementos de mais de 12 ontologias diferentes, o que facilita a interoperabilidade entre sistemas. Esta pode ainda ser facilmente personalizada e estendida, apesar das alterações poderem prejudicar a interoperabilidade com outros sistemas. É importante referir que a aplicação VIVO é facilmente adaptável para refletir novas classes e propriedades adicionadas à ontologia.

Na figura 4, como exemplo, é representado um diagrama do módulo académico da ontologia. Podem assim ser identificadas classes que representam pessoas, organizações, funções, relações, processos, informações, documentos e entidades complementares às anteriores.

[2] Comparison of research networking tools and research profiling systems. http://en.wikipedia.org/w/index.php?title=Comparison_of_research_networking_tools_and_research_profiling_systems&oldid=645697650

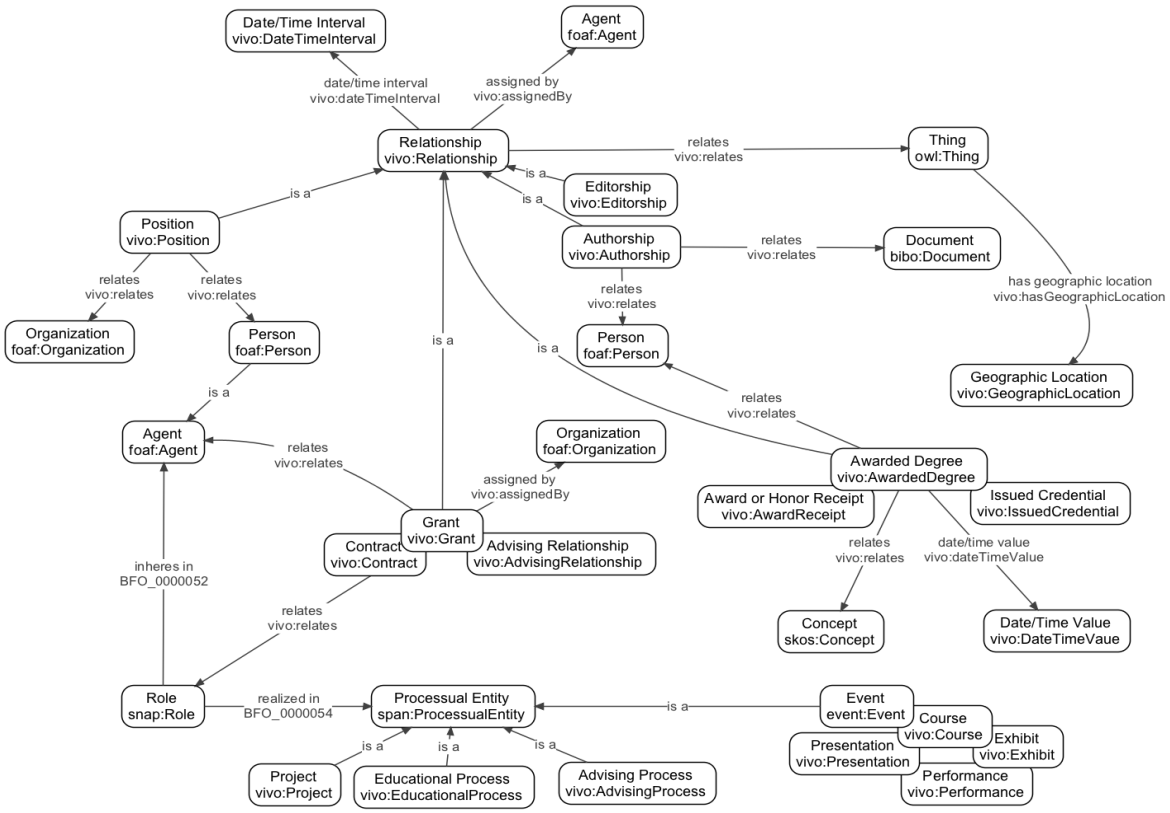

Figura 4 - Módulo académico da ontologia VIVO-ISF

Para o piloto proposto são efetuadas o mínimo número de alterações possíveis de forma a potenciar a interoperabilidade do sistema desenvolvido como outros sistemas e assim minimizar/eliminar a perda de informação.

Arquitetura do sistema

Identificados os domínios, fontes e interfaces de informação e definido o modelo de dados, procede-se à definição da primeira versão da arquitetura do sistema.

Como identificado anteriormente, o projeto piloto deve ser capaz de agregar e relacionar informação relativa a projetos, pessoas e resultados de investigação e deve estar preparado para futuras integrações com sistemas que seguem normas de interoperabilidade conhecidas, e.g. CERIF-XML.

Para a agregação de informação de projetos de investigação, numa primeira fase e numa base regular, é agregado um ficheiro CSV com informação atualizada sobre os mesmos e criado um transformador para mapear a informação agregada para a ontologia definida.

Para a agregação de resultados de investigação, é utilizada uma versão em desenvolvimento da API CERIF-XML do Portal RCAAP. A informação agregada é mapeada utilizando um transformador CERIF2VIVO desenvolvido no âmbito do projeto AgriVIVO (Nogales, Sicilia, & Jörg, 2014). Ao agregar informação relativa a projetos de investigação, é necessário agregar informação complementar relacionada com pessoas, organizações e financiamento que podem ser obtidas a partir da interface REST do ORCID ou na base de dados CERIF em desenvolvimento. Depois de agregar e transformar a

informação, segue-se uma fase de tradução onde se procurará relacionar a nova informação com a informação já presente no sistema.

Importante será referir que, desenvolvidos e melhorados os transformadores para CERIF-XML e CSV, a futura integração do sistema com interfaces semelhantes fica simplificada.

Depois de povoado o sistema, este deve permitir a pesquisa e consulta de informação a partir de páginas anotadas com RDF, disponibilizar uma interface SPARQL, um formulário de pesquisa sobre diferentes domínios de informação, assim como serviços de valor acrescentado como mapas de ciência e mapas de autores.

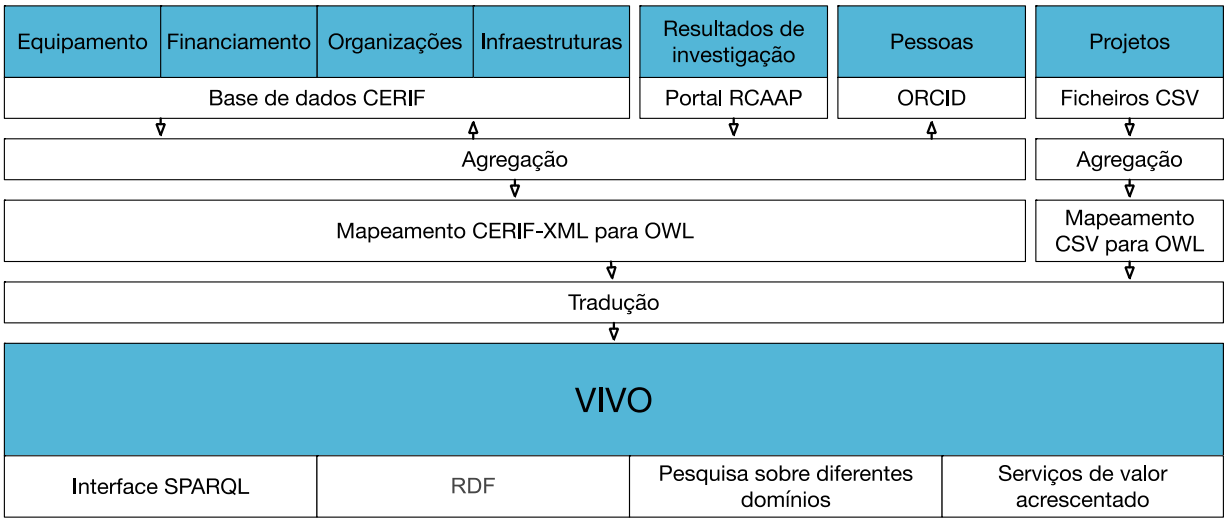

Figura 5 - Primeira versão da arquitetura do projeto piloto

A arquitetura definida para o projeto piloto, representada na figura 5, prevê a rápida expansão de fontes, formatos e domínios de informação uma vez que é esperado que o sistema final agregue domínios de informação adicionais e sistemas ainda em desenvolvimento no âmbito do projeto PT-CRIS.

Discussão

Ao decidir implementar um sistema VIVO de dimensão nacional, é relevante procurar exemplos de iniciativas semelhantes para suportar as decisões mais relevantes.

O VIVO surgiu como uma ferramenta para facilitar a colaboração científica entre diferentes domínios de informação na Universidade Cornell. Este começou a ser amplamente usado por diversas instituições dos Estados Unidos e mais tarde começaram a surgir implementações do sistema a nível nacional, muitas vezes utilizando como fontes de informação outros sistemas VIVO institucionais. A ferramenta continuou a evoluir e agora é já usada em contextos internacionais.

Um exemplo da utilização do sistema num nível global, e também uma referencia para o sistema central que se pretende desenvolver, é o sistema AgriVIVO, um portal que permite a colaboração entre

diferentes intervenientes da área agrícola, agregando mais de 10 fontes de informação dispersas por diferentes países (Global Forum on Agricultural Research, 2011).

O AgriVIVO assemelha-se ao projeto proposto porque, para além de também agregar fontes dispares de informação, foi também responsável por grande parte dos transformadores e exemplos para a interoperabilidade entre sistemas CERIF e o VIVO (Nogales et al., 2014).

A preocupação existente para que os sistemas nacionais disponibilizem interfaces baseadas em normas internacionais, como o formato CERIF-XML, permite facilitar a interoperabilidade entre sistemas e reduz significativamente os custos de desenvolvimento e manutenção de interfaces de interoperabilidade. Esta preocupação é também uma preocupação de outros interlocutores, e um excelente exemplo é o anuncio de cooperação entre o VIVO e a *European Organization for International Research Information* (EuroCRIS) para convergência semântica e interoperabilidade entre os dois sistemas.

Estes são exemplos dos crescentes esforços para a cooperação mundial, onde se acredita que o PT-CRIS e o portal central aqui proposto sejam uma referência.

Conclusões

O planeamento e desenvolvimento de um piloto de um sistema desta dimensão e importância é muito relevante uma vez que permite antecipar problemas, encontrar soluções e avaliar a viabilidade e utilidade do sistema final que permitirá agregar e relacionar informações de várias fontes e domínios.

Durante o planeamento e primeiros desenvolvimentos foram tomadas várias decisões, seguindo exemplos de iniciativas internacionais, e procurou-se respeitar as principais normas/práticas internacionais alinhando assim o projeto com esforços para a colaboração na gestão da ciência a nível mundial.

Uma das principais dificuldades encontrada durante o planeamento passou pela elevada dependência do sistema para com outros desenvolvimentos no âmbito projeto PT-CRIS. Assim, considera-se essencial manter a comunicação e coordenação com os diferentes intervenientes do projeto para que se consiga reduzir dependências e esforços.

De seguida dar-se-á continuidade ao desenvolvimento do projeto, onde esperam-se dificuldades relacionadas com a qualidade da informação e com a escalabilidade do sistema, por isso, o desenvolvimento deve ter inicio com a realidade de que poderão ser necessárias várias iterações até que se encontrem as melhores soluções para os objetivos pretendidos.

A data de conclusão do projeto piloto apresentado, está prevista para antes do final do ano de 2015.

Acredita-se que o projeto será uma relevante contribuição para a promoção da ciência nacional, dando-lhe uma maior visibilidade para todos os intervenientes nacionais e internacionais e facilitando os processos de gestão e produção de ciência e inovação, através do acesso facilitado a informação autoritária, completa e fidedigna.

Referências

(GFAR) Global Forum on Agricultural Research. (2011, September 30). AgriVIVO for enabling global networking for agriculture. Concept Note | EGFAR. Global Forum on Agricultural Research (GFAR). Retrieved from http://www.egfar.org/documents/agrivivo-enabling-global-networking-agriculture-concept-note

Haak, L. L., Fenner, M., Paglione, L., Pentz, E., & Ratner, H. (2012). ORCID: a system to uniquely identify researchers. *Learned Publishing*, *25*(4), 259–264. doi:10.1087/20120404

Moreira, J. M. (2015). PTCRIS - Parte I. In *Jornadas FCCN 2015*. Retrieved from http://pt.slideshare.net/JooMoreira3/jornadas-2015-ptcris-parte-i

Nogales, A., Sicilia, M.-A., & Jörg, B. (2014). Combining VIVO and Google Scholar Data as Sources for CERIF Linked Data: A Case in the Agricultural Domain. *Procedia Computer Science*, *33*, 266–271. doi:10.1016/j.procs.2014.06.042

Simons, E., & Danica Zendulková. (2013). CRIS – Repository Connection. Possibilities and Values. In *OpenAIRE Interoperability Workshop*. Retrieved from http://pt.slideshare.net/OpenAIRE_eu/simons