GESTÃO DE DADOS DE INVESTIGAÇÃO NO DOMÍNIO DA OCEANOGRAFIA BIOLÓGICA: CRIAÇÃO E AVALIAÇÃO DE UM PERFIL DE APLICAÇÃO BASEADO EM ONTOLOGIA

Resumo: No atual contexto científico os dados produzidos pelas atividades de investigação ganham crescente importância e visibilidade. Por este motivo, a questão da gestão de dados de investigação está no cerne da preocupação das comunidades científicas. Este artigo descreve a elaboração e a avaliação de uma ferramenta de curadoria digital desenvolvida para apoiar um pequeno grupo de investigadores da área da Oceanografia Biológica na gestão de seus conjuntos de dados. Esta ferramenta, chamada de perfil de aplicação, consiste em um padrão de metadados feito sob medida para descrição de dados do domínio citado. A mesma é formalizada em ontologia e incorporada no Dendro, uma plataforma colaborativa e multidisciplinar de gestão de dados concebida na Faculdade de Engenharia da Universidade do Porto. Investigadores da área podem armazenar e descrever seus dados na plataforma, e ainda tem a garantia de que os mesmos estão preparados para depósito em repositórios externos.

Palavras-chave: gestão de dados de investigação, curadoria digital, metadados, perfil de aplicação, ontologia.

Rúbia Tatiana Gattelli
Bibliotecária/Documentalista da Universidade Federal do Rio Grande, Brasil
rgattelli@gmail.com
;
Maria Cristina de Carvalho Alves Ribeiro
Departamento de Engenharia Informática da Universidade do Porto, Portugal
mcr@fe.up.pt

RESEARCH DATA MANAGEMENT IN BIOLOGICAL OCEANOGRAPHY DOMAIN: DEVELOPMENT AND EVALUATION OF A BASED ONTOLOGY APPLICATION PROFILE

Abstract: In the current scientific context the data produced by research activities gain incrisingly importance and visibility. For this reason, the research data management issue is at the center of scientific communities concerns. This article describes the development and evaluation of a digital curation tool designed to support a small group of researchers in the field of Biological Oceanography in managing their data sets. Such tool, called application profile, consists of a metadata standard tailored for describing data from the Biological Oceanography domain. It is formalized as an ontology and then it is incorporated into Dendro, a collaborative multidisciplinary data management platform designed at the Faculty of Engineering of the University of Porto. Researcher from this field can store and describe their data in the platform and still have the guarantee that they are prepared to deposit in external repositories.

Keywords: research data management, digital curation, metadata, application profile, ontology.

1 INTRODUÇÃO

Desde a última década tem havido um crescimento da importância de questões globais ligadas à ciência. Temas relacionados com gestão, acesso e reutilização de dados provenientes de atividades de investigação científica, sobretudo se estes forem resultantes de projetos de investigação financiados com recursos públicos, têm atraído atenção tanto de comunidades científicas quanto da esfera política internacional (RODRIGUES, *et al*., 2010).

Começa a se firmar em cenário global um novo paradigma científico fortemente ligado à tecnologia da informação e orientado por e para os dados de investigação, o qual se convencionou chamar de "e-Science". O termo, segundo Costa e Cunha (2014) foi cunhado por John Taylor, "Director General of Research Councils, Office of Science and Technology" do "National e-Science Centre" no Reino Unido. Taylor o definiu como uma ciência em grande escala executada através de colaborações globais possibilitadas pela Internet, as quais irão requisitar acesso a grandes coleções de dados, recursos computacionais em grande escala e visualização de alto desempenho para os cientistas.

Dentre as características da "e-Science", das quais podem ser citadas o trabalho colaborativo, a multidisciplinaridade e o uso de aparato tecnológico, os dados provenientes das atividades de investigação (referidos na literatura em língua portuguesa como "dados de investigação" – termo aqui adotado, ou "dados científicos", e na literatura em língua inglesa como "research data") desempenham um papel central. Os dados de investigação passam a ganhar status de importância e maior visibilidade neste novo paradigma científico, posto que cresce o entendimento de que são a base fundamental da ciência e da investigação científica. Os dados de investigação são o insumo a partir do qual a informação e o conhecimento são derivados, e têm o propósito de produzir e validar resultados de investigação.

De acordo com CORTI *et. al.* (2014) o período a partir dos anos 2000 tem visto uma série de empreendimentos em direção ao compartilhamento de dados, bem como ao desenvolvimento de capacidades humana e material para fazê-lo. São vistas várias iniciativas que visam assegurar a qualidade, sustentabilidade e acessibilidade aos dados de investigação. Da mesma forma, também tem crescido a demanda pela transparência e acesso aos dados, estas advindas não somente dos ambientes acadêmicos, onde em grande parte os dados são produzidos, mas de diferentes setores, como financiadores de investigação, editores de literatura científica, governos e a sociedade em geral.

Igualmente há que se referir a preocupação crescente das comunidades científicas em lidar com o aumento progressivo dos dados, impulsionado pelo incremento do aparato tecnológico utilizado para gerar ou recolher dados durante suas ações de investigação. Este fenômeno, constatado e reportado na literatura como "dilúvio de dados", (do inglês "data deluge") fez emergir a atenção em relação às políticas e práticas de gestão de dados de investigação.

A gestão de dados de investigação compreende todas as práticas, manipulações, melhoramentos e processos que possam assegurar a qualidade dos dados, bem como, garantir que estejam organizados, documentados, preservados, sustentáveis, acessíveis e reutilizáveis (CORTI *et al.*, 2014). Portanto, questões relacionadas ao armazenamento, utilização e preservação dos dados constituem problemas concretos para as comunidades de investigadores.

Na tentativa de oferecer soluções para tais problemas uma área relacionada com a curadoria digital de dados científicos tem se afirmado. A curadoria digital, de acordo com o Digital Curation Centre (DCC), centro de investigação em curadoria de informação digital no Reino Unido, envolve "a manutenção, preservação e adição de valor aos dados de investigação digitais ao longo do seu ciclo de vida". O DCC acrescenta ainda que a curadoria digital reduz a duplicação de esforços na criação de dados e aumenta o valor de longo prazo dos dados existentes, tornando-os disponíveis para futuras investigações. Além disso, a gestão ativa dos dados reduz ameaças ao valor de investigação de longo prazo e ainda diminui o risco de obsolescência digital.

A fragilidade inerente às mídias digitais e a obsolescência tecnológica que colocam os dados em risco, são problemas que podem ser sanados com as atividades de curadoria digital. Para Abbott (2008) todas as atividades envolvidas na gestão de dados, boas práticas de digitalização e documentação, e garantia de disponibilidade e adequação para descoberta e reuso dos dados no futuro, fazem parte da curadoria digital. Tais atividades podem envolver tecnologias para conversão dos dados em diferentes formatos digitais, descrição de conjuntos de dados (metadados), armazenamento e acesso aos dados de investigação (plataformas digitais).

Dentre as atividades a que se dedica a curadoria de dados de investigação, destaca-se aqui a criação de metadados normalizados, essenciais para descrição, acesso e reutilização dos dados. Eles são estruturas organizadas que orientam a forma como os dados serão descritos no

seu armazenamento em uma plataforma digital, e ainda fornecem pontos de acesso para que a informação seja posteriormente recuperada.

A boa documentação, isto é, boa descrição dos dados, é crítica para o entendimento dos dados em curto, médio e mais longo prazo, e é vital para o sucesso da preservação dos dados a longo prazo (CORTI *et al.*, 2014). Também é fundamental para as práticas de partilha e reutilização, pois é necessário que os dados sejam entendidos por outros investigadores, ou mesmo pelos próprios investigadores que os criaram depois de passado algum tempo.

Uma descrição clara requer um conhecimento profundo dos dados e do seu processo de criação, fatores que são diferentes em cada domínio de conhecimento. A criação de metadados normalizados é importante para garantir interoperabilidade, no entanto, estes podem não contemplar todas as características e especificidades das diferentes áreas de domínio. Logo, diferentes áreas de investigação demandam a criação de diferentes conjuntos de metadados, para que os dados possam ser descritos apropriadamente no que diz respeito às suas peculiaridades.

A necessidade de descrever conjuntos de dados de diferentes domínios tem levado à criação de perfis de aplicação (CASTRO, RIBEIRO e SILVA, 2013). Para Heery e Patel (2000) os perfis de aplicação (do inglês *application profile*) são um tipo de esquema de metadados que consistem de elementos de dados retirados de um ou mais esquemas existentes, combinados entre si e otimizados para uma aplicação local em particular. É o que as autoras chamam de abordagem “mix and match”.

Castro, Ribeiro e Silva (2013) afirmam que a quantidade de esforço que os investigadores precisam investir para descrever seus dados pode limitar a sua disposição em partilhá-los. Daí que, em vez de fazer os investigadores seguirem esquemas estritos de metadados, propõem que eles devem ter acesso a um perfil de aplicação feito sob medida para o seu próprio domínio.

Um perfil de aplicação, portanto, pode composto da seleção de descritores de um ou mais esquemas de metadados existentes, combinada com a criação de metadados específicos do domínio que se quer descrever. Este cumpre sua função em uma plataforma eletrônica onde se queira armazenar e descrever determinado recurso, e uma boa forma para representá-lo dentro de uma plataforma web é fazê-lo através do uso de uma ontologia. Uma ontologia é a ferramenta adequada para representar metadados criados para diferentes domínios, uma vez

que apresenta capacidade semântica para contemplar tais especificidades, pode evoluir facilmente e ainda ser integrada por sistemas de gestão de dados.

As ontologias podem ser de diversos tipos e são utilizadas em diversas áreas do conhecimento para organizar informação (ALMEIDA e BAX, 2003), sua adoção tem aumentado muito na medida em que cresce o volume de informação digital na web. Considerando o atual dilúvio de dados de investigação produzidos nas mais variadas áreas do conhecimento se pode inferir que as ontologias são ferramentas apropriadas para representar e organizar tal informação, absorvendo o vocabulário específico de cada área.

Levando em consideração o cenário da produção e gestão de dados por comunidades científicas, este artigo descreve a criação e posterior avaliação de um perfil de aplicação para descrição dos dados de investigação da área de Oceanografia Biológica. Os metadados do perfil (também chamados de descritores) são formalizados em ontologia e introduzidos em uma plataforma de gestão de dados de investigação chamada Dendro[1], desenvolvida na Faculdade de Engenharia da Universidade do Porto (FEUP).

O Dendro é uma plataforma baseada na web para depósito e troca de dados de investigação, projetada para ajudar os investigadores a armazenar os seus dados (à maneira da aplicação Dropbox), e para descrevê-los de forma colaborativa. Seu principal objetivo é ajudar os investigadores a descreverem seus dados na medida em eles são produzidos, tornando mais fácil a sua partilha com um repositório externo (institucional, por exemplo). Desta forma os dados podem ser citados por outros, bem como pelos seus próprios autores em publicações.

A plataforma Dendro foi desenhada com uma interface amigável para utilizadores sem conhecimento em gestão de dados. Utiliza conceitos de ontologias específicas de domínio, que se configuram em metadados criados sob medida para descrever dados destes domínios. Portanto, permite aos utilizadores construir uma base de conhecimento usando ontologias em segundo plano, que lhes dá prioridade em focar na escolha dos descritores semanticamente compatíveis com seus domínios sem se preocupar com questões de design e implementação que surgem do uso de ontologias. (Silva, *et al.* 2014).

O perfil de aplicação denominado “Biological Oceanography” passa a compor uma lista de conjuntos de metadados específicos de várias áreas do conhecimento existentes no Dendro. O objetivo é que investigadores de vários domínios possam fazer o depósito de seus

[1] Dendro - http://dendro.fe.up.pt/

dados na plataforma utilizando estes metadados para descrevê-los. Os dados ficam assim preparados para serem partilhados com outros repositórios ou sistemas de gestão de dados, caso seja do interesse do investigador.

A inserção do perfil na plataforma de gestão de dados permitiu que o mesmo pudesse ser utilizado e avaliado na prática, uma vez que pode fazer parte de uma campanha de experiências de interação dos investigadores da área com a plataforma Dendro. Na experiência estes procederam ao depósito de um conjunto de dados na plataforma e à sua descrição utilizando os vocabulários disponíveis. Concomitante com estas experiências os investigadores da Oceanografia Biológica foram convidados a realizar uma breve avaliação sobre o perfil de aplicação "Biological Oceanography, onde foram solicitados a responder um inquérito para avaliar a utilidade dos descritores do perfil de aplicação e sua compatibilidade com a investigação no domínio.

2 METODOLOGIA

Para empreender a criação do perfil de aplicação "Biological Oceanography" foi utilizada a abordagem metodológica que vem sendo executada no Information Systems Research Group (InfoLab) do Departamento de Engenharia Informática da FEUP. Também foram consultados os "Guidelines for Dublin Core Application Profiles[2]", que embasaram e complementaram a abordagem metodológica em questão, cujos passos estão abaixo descritos.

Para o desenho de um perfil de aplicação compatível com a descrição dos dados de investigação do domínio da Oceanografia Biológica foi necessário adquirir um algum conhecimento sobre a área. Obter informação sobre questões práticas, como procedimentos e instrumentos utilizados nas atividades de investigação, métodos de recolha, tratamento e armazenagem de dados, além de familiarização com a linguagem técnica do domínio, foram importantes pré-requisitos para se conseguir tal percepção.

Para o levantamento dos requisitos foi realizada uma entrevista semiestruturada com cada um dos investigadores que colaboraram com o estudo. As entrevistas foram conduzidas com base no "Guião de Entrevista Curadoria de Dados", adaptado de instrumento

[2] Dublin Core Metadata Initiative. **Guidelines for Dublin Core Application Profiles** - (http://dublincore.org/documents/profile-guidelines/#DCMI-MT),

disponibilizado pelo Data Curation Profile Toolkit[3], e gravadas para o áudio ser posteriormente analisado.

Para esta etapa de levantamento de requisitos sobre o domínio também se precedeu à análise de conteúdo de publicações científicas dos investigadores, onde se pode captar conceitos-chave para o entendimento da investigação na área. Para compreensão da linguagem técnica por vezes se recorreu a dicionários, wikis ou vocabulários controlados, e para entendimento de procedimentos específicos, eventualmente foi necessário recorrer ao próprio investigador.

As duas tarefas acima descritas forneceram embasamento para familiarização com termos específicos do domínio e entendimento da área e das experiências de investigação, e culminaram na elaboração de um mapa com os principais conceitos levantados, o qual foi validado pelos investigadores da área, após as correções necessárias. Para isso foi utilizado o software CMap Tool[4].

Deste mapa foram extraídos os conceitos mais importantes a serem adotados como potenciais descritores (metadados) no perfil de aplicação. No entanto, sentiu-se a necessidade de acrescentar conceitos que não constavam no mapa, mas que seriam importantes para a descrição de um conjunto de dados. Decidiu-se seguir a recomendação Dublin Core Application Profiles e estabelecer objetivos para o perfil de aplicação. A partir daí outros conceitos foram selecionados em conformidade com os objetivos.

Ainda, para a elaboração do perfil de aplicação foi necessário proceder ao estudo de esquemas de metadados e normas existentes para descrição de dados concernentes ao domínio da Oceanografia Biológica e áreas afins, tais como Ecologia e Biologia. Os padrões selecionados foram o “Ecological Metadata Language” (EML), “Darwin Core”, “OBIS Data Schema” e “Content Standard for Digital Geospatial Metadata, Part 1: Biological Data Profile”. Também foi estudado um padrão de metadados genérico, para o que foi escolhido o “Dublin Core”. Todos os esquemas mencionados foram selecionados a partir de sugestões do Digital Curation Centre.

Depois de elaborado o perfil de aplicação, foram pesquisadas ontologias com vocabulário que pudesse representar os descritores. O objetivo era importar estes vocabulários

[3] Data Curation Profiles Project. Data Curation Profile Toolkit - http://datacurationprofiles.org/

[4] Mais informações em: CMap Tool - http://cmap.ihmc.us/

e aproveitá-los para a formalização do perfil de aplicação dentro da plataforma de gestão de dados. As ontologias estudadas foram: “Dublin Core”, “Oboé”, “CERIF”, “Friend of a friend (FOAF)”, “TGWD Ontologies”, “The TaxonConcept Ontology”, “The Darwin-SW Ontology” e “Marine TLO”. Algumas destas ontologias foram recomendadas pelos colegas do InfoLab e outras recomendadas pelo World Wide Web Consortium (W3C)[5].

A ontologia construída utilizando o software Protégé[6], constituiu-se apenas de algumas classes e propriedades, e por apresentar este formato simplificado é chamada de “ontologia ligeira” (“lightweight ontology”) (Castro, Silva e Ribeiro, 2014). Nesta ontologia, os descritores do perfil de aplicação são representados por propriedades. Após sua formalização a ontologia foi incorporada na plataforma Dendro, onde passou a cumprir sua função como modelo de metadados para descrição de dados de investigação.

Para sensibilizar os investigadores quanto à importância das boas práticas de gestão de dados, sobretudo no que se refere à questão de sua descrição, bem como para avaliar o desempenho da plataforma Dendro, uma série de experiências foi realizada com investigadores da Universidade do Porto. Investigadores das áreas em que foram construídas ontologias foram convidados a depositar e descrever seus conjuntos de dados e outros produtos gerados a partir destes, como artigos ou relatórios de investigação.

Para esta fase de utilização da plataforma foram convidados dois (2) investigadores da área de Ciências do Mar, pertencentes à Universidade do Porto. Após sua interação com a plataforma Dendro os mesmos foram convidados a fornecer um depoimento sobre sua experiência e a responder um breve inquérito, para avaliar os descritores do perfil de aplicação “Biological Oceanography”, composto por questões genéricas sobre validade e utilidade dos descritores do perfil.

3 DESENHO DE UM PERFIL DE APLICAÇÃO PARA O DOMÍNIO DA OCEANOGRAFIA BIOLÓGICA

O desenho do perfil de aplicação “Biological Oceanography” se enquadra na metodologia aplicada no InfoLab da FEUP. Investigadores deste laboratório têm trabalhado

[5] World Wide Web Consortium (W3C) - (http://www.w3.org/wiki/Good_Ontologies
[6] Mais informações em: Protégé - (http://protege.stanford.edu/

na elaboração de perfis de aplicação baseados em ontologia com o objetivo de estes serem incorporados à plataforma Dendro, a qual é destinada a suportar a gestão colaborativa de dados de investigação multi-domínio. Estes perfis são desenhados para situações e necessidades específicas com a colaboração de especialistas dos domínios. Retratam, portanto, atividades específicas de investigação, não tendo a pretensão de abranger o domínio em sua totalidade, antes pelo contrário, têm o objetivo de captar e desenhar os requisitos de gestão de dados em grupos de investigação específicos.

Portanto, justifica-se que o perfil de aplicação desenhado para o domínio da Oceanografia Biológica retrata especificamente as atividades de investigação de três grupos distintos oriundos da Universidade Federal do Rio Grande, no Sul do Brasil, sendo estes: Laboratório de Crustáceos Decápodes, Laboratório de Ictiologia, e Laboratório de Ecologia de Invertebrados Bentônicos. Um investigador de cada laboratório colaborou com o estudo, do que se pode depreender que o desenho do perfil de aplicação levou em consideração as atividades de investigação específicas destes investigadores. Embora tenha se tentado fazer uma generalização para o restante do grupo ao qual o investigador pertence, admite-se que alguns aspectos podem ter ficado de fora.

Nos tópicos seguintes estão descritos os passos realizados para a concretização de um conjunto de metadados para descrição de dados de investigação no domínio da Oceanografia Biológica.

3.1 LEVANTAMENTO DE REQUISITOS DO DOMÍNIO

O primeiro passo a ser dado em direção à elaboração de um perfil de aplicação compatível com as necessidades de descrição de um domínio deve ser o envolvimento dos investigadores da área. Para esta experiência contou-se com a colaboração de investigadores do Instituto de Oceanografia da Universidade Federal do Rio Grande, onde foi realizada uma apresentação que abordou o tema da gestão de dados de investigação com o objetivo de sensibilizar os investigadores em relação às boas práticas de gestão, e sobre o trabalho a ser empreendido para elaboração do perfil de aplicação.

A fase seguinte compreendeu o levantamento de requisitos necessários para elaboração do perfil de aplicação, na qual o contato direto com os investigadores/colaboradores do estudo foi importante. Para o levantamento dos requisitos,

que se constituem de elementos para a compreensão do domínio, foi programada uma entrevista com cada um dos colaboradores. Após a entrevista lhes foi solicitado o envio de pelo menos dois artigos seus já publicados e, se possível de um conjunto de dados. Caso o envio do conjunto de dados não fosse possível foi pedido que este fosse apenas mostrado, para se obter uma ideia mais concreta de como estes se apresentam.

A entrevista se baseou no guião usualmente utilizado pelo InfoLab/FEUP que foi traduzido e adaptado por um de seus membros de um instrumento disponibilizado pelo Data Curation Profile Toolkit (ressalta-se que a entrevista é um procedimento recomendado pelo Data Curation Profiles Project[7] adotado pelo InfoLab, outras atividades podem ser realizadas para levantamento de requisitos). O instrumento aborda questões sobre o fluxo dos dados produzidos durante o ciclo de investigação e questões sobre atividades de gestão, relativamente ao armazenamento e partilha de dados.

Após a entrevista já foi possível obter um entendimento dos conjuntos de dados gerados a partir das atividades de investigação da área. Ao lado do entendimento dos conjuntos de dados, a entrevista ainda possibilitou uma melhor compreensão da leitura dos artigos publicados pelos investigadores. Apesar disso, nesta fase ainda foi encontrada alguma dificuldade em compreender situações específicas, procedimentos envolvidos ou linguagem técnica. Para sanar estas dúvidas se recorreu a pesquisas na Internet, consulta a dicionários e vocabulários da área, e aos próprios investigadores, os quais foram a melhor fonte de informação. Por este motivo a disponibilidade deles é fundamental para o processo.

Durante a leitura de artigos se procedeu à análise de conteúdo dos mesmos, com identificação e seleção de conceitos considerados importantes. Análise de conteúdo é uma prática subjetiva, no entanto pode ser orientada pela Norma Portuguesa NP 3715, de 1989. A norma sugere quais as partes mais importantes de um documento escrito que devem ser analisadas, caso não haja disponibilidade para uma análise do texto integral. Ainda orienta na identificação de conceitos mais representativos do conteúdo do documento, que devem ser agrupados em grelhas segundo critérios.

Portanto, com base na norma, foram estabelecidos critérios gerais e dentro destes alocados os conceitos mais representativos do conteúdo do documento, que podem ser abaixo visualizados.

[7] Mais informações em: Data Curation Profiles Project - http://datacurationprofiles.org/about

- Identificação de título e autores
- Assuntos
- Variáveis de espaço
- Variáveis de tempo
- Metodologia e instrumentos utilizados
- Material coletado/triagem
- Parâmetros ambientais

3.2 A PRODUÇÃO DE DADOS NO DOMÍNIO DA OCEANOGRAFIA BIOLÓGICA

A produção de dados de investigação se dá no contexto de três subáreas do domínio da Oceanografia Biológica, mais concretamente nas áreas de ecologia de peixes, de crustáceos decápodes e de organismos invertebrados bentônicos. As três subáreas objetivam investigar as interações ecológicas entre organismos marinhos e estuarinos e seus parâmetros ambientais.

Os dados gerados a partir de suas atividades de investigação são sobretudo observacionais, ou seja, são captados em tempo real. As atividades dividem-se em duas etapas distintas, que foram designadas como atividades de campo e atividades de laboratório. As atividades de campo dizem respeito aos eventos de coleta de material biológico e ao registo dos parâmetros ambientais. As atividades realizadas em laboratório dizem respeito aos tipos de interações necessárias realizadas com o material biológico coletado em campo, o que pode envolver experimentos, dando origem também a dados experimentais.

O material biológico capturado nos eventos de coleta consiste nos organismos estudados pelas áreas (peixes, crustáceos e bentos) e em sedimento (substrato do fundo de corpos d´água, neste caso o estuário da Lagoa dos Patos). Os parâmetros ambientais podem ser registados durante os eventos de coleta ou ser independente destes, de acordo com uma periodicidade previamente estipulada. Eles são chamados pelos investigadores de dados abióticos e os principais são temperatura da água, salinidade, transparência e profundidade da coluna d´água. Para estas duas atividades de campo são aplicados métodos e utilizados instrumentos e ferramentas específicos, que levam em consideração o tipo de material que se quer coletar ou o tipo de variável ambiental que se pretende registar.

Tanto para os eventos de coleta de material biológico quanto para os registros de parâmetros ambientais são definidos e registados alguns dados espaciais e alguns dados temporais, os quais, como os próprios nomes sugerem, são elementos importantes para contextualizar estas atividades territorial e temporalmente. Os dados espaciais referem-se a: nome do local onde será realizado o evento, nomes dos pontos específicos de coleta daquele local, suas coordenadas geográficas, sendo também definida a quantidade de pontos de coleta e a extensão do perfil. Os dados temporais referem-se a: data do evento, período de abrangência dos eventos, a periodicidade em que se realizam os eventos e a estação do ano em que se realizam.

Após os eventos de coleta, o material biológico capturado é levado para o laboratório onde é processado. Primeiramente é feita uma separação de todo o material coletado, a qual é chamada de triagem. Na triagem dos sedimentos estes são separados e calculados os elementos que os compõem, tais como argila, areia, silte e matéria orgânica. Na triagem dos organismos, os indivíduos são separados por espécies e então são feitas uma série de contagens (as quantidades de espécies, quantidade de indivíduos por espécie e quantidade total de indivíduos capturados) e tomadas uma série de medidas (biometrias dos indivíduos, que são diferentes para cada espécie, mas que consistem basicamente em medidas de largura, cumprimento e peso). Ainda podem ser verificadas outras variáveis, como o sexo, estágio de vida e o estádio de muda (para crustáceos) dos indivíduos. Todos os dados derivados destas interações com os organismos são designados pelos investigadores como dados bióticos. Da mesma forma que as atividades de campo, as atividades de laboratório também demandam métodos e instrumental específicos.

Em relação ao registro e organização dos dados normalmente estes são primeiramente anotados em planilha em papel, tanto os dados capturados em campo como os capturados em laboratório. Existe, então, uma fase posterior onde as planilhas em papel são transpostas para ficheiros Excel. Esta é considerada uma fase delicada onde se deve ter muita atenção para não se cometerem erros de digitação dos valores. As planilhas eletrónicas são muito simples, contêm nos cabeçalhos das colunas os nomes das medidas que são tomadas, e referências em relação a datas e locais onde são realizadas as coletas. Estas informações aparecem de maneira abreviada. Em alguns casos é feita uma legenda para as siglas, em outros não, por se entender que os investigadores que trabalham no mesmo laboratório já possuem

conhecimento para interpretar as planilhas (quando entram estudantes ou investigadores novos as planilhas são explicadas por um membro antigo).

Os ficheiros Excel são salvos em pastas com nomes que possam identificar seu conteúdo e são armazenados, normalmente nos computadores dos laboratórios. Às vezes são feitas cópias de segurança ou os dados são armazenados em drives na nuvem, mas isso depende da iniciativa individual de cada investigador.

Do ponto de vista da investigação os investigadores apontam quatro fases para o ciclo de vida dos dados: 1) coleta dos dados brutos (bióticos e abióticos); 2) transposição dos dados para planilhas Excel e armazenamento; 3) análise estatística dos dados (uso de softwares específicos); 4) elaboração de produtos finais (teses, artigos publicados, etc.).

O cuidado que os investigadores têm em relação ao armazenamento e preservação é em relação aos dados brutos, este foi o tipo de dados que todos referiram em entrevista ser o mais importante a preservar. Os investigadores têm noção de que estes dados preservados podem dar origem a novos estudos. Dados que já sofreram algum tipo de análise são dados a conhecer apenas nas dissertações, teses, relatórios ou artigos publicados.

Questões de gestão e partilha de dados não são institucionalizadas. Também não há exigência por parte das agências que financiam os projetos de investigação que os investigadores apresentem planos de gestão de dados de investigação ou pratiquem formalmente a sua partilha através de depósito em repositórios.

3.3 MAPA DE CONCEITOS E ESCOLHA DE DESCRITORES

O levantamento de requisitos (entrevista e análise de conteúdo dos artigos) levou à construção de um mapa de conceitos, onde foi possível modelar o conhecimento da área ao estabelecer relações entre os conceitos. Estes, fundamentalmente extraídos da análise de conteúdo das publicações, formaram um desenho das atividades práticas de investigação e variáveis intrínsecas para a produção de dados.

Por se tratar de três subáreas do mesmo domínio, que apesar de compartilharem características em comum ainda mantêm suas particularidades, foram construídos inicialmente três mapas. Depois de validados com cada um dos investigadores das áreas, estes

foram posteriormente unidos em apenas um mapa mais completo do domínio. Por ser um mapa bastante extenso optou-se por mostrar aqui uma versão reduzida a título de exemplo:

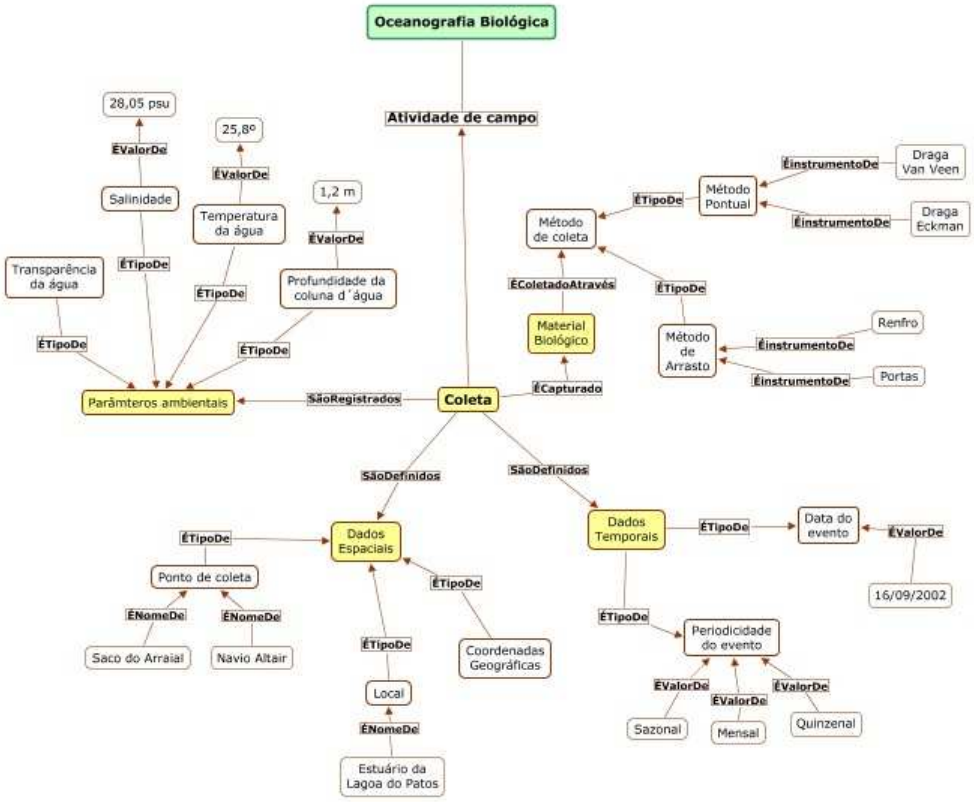

Figura 1 – Mapa de conceitos reduzido

A tarefa de modelar conceitos no mapa pode dispender algum tempo e ocasionar várias modificações do mapa original. Isto se deve ao fato de que, no decorrer desta ação, ocorre um aprimoramento sobre a visão de organização e de relações conceituais do conhecimento que se está a estruturar. Ademais, solicita-se sempre a intervenção dos reais conhecedores do domínio o que pode resultar em modificações adicionais. No entanto, tal exercício é relevante para o processo de aprendizagem, e culmina em um mapa conceitual consistente e, por consequência, na escolha mais adequada de potenciais descritores para conceber o perfil de aplicação.

A etapa subsequente é da escolha dos conceitos do mapa que poderão se tornar os descritores do perfil de aplicação. Nesta fase foi importante dar aos investigadores noção do conceito de descritor (metadados) e qual sua função na gestão de dados de investigação. Comparações com exemplos concretos de anotação, que sejam conhecidos dos colaboradores,

são úteis. Foi necessário dotá-los de um nível de abstração um pouco maior no que se refere à organização e representação da informação, noção que a tarefa de descrição exige.

Para delimitar a escolha dos conceitos e tornar o processo simplificado foi necessário delinear objetivos para o perfil de aplicação. Estes foram estabelecidos de acordo com demonstrações de dúvidas dos investigadores durante o processo, tais como a quantidade de conceitos que se poderia escolher.

Da mesma maneira como surgiu dúvida em relação ao número de descritores, os investigadores também sentiram falta de descritores cujos conceitos não estavam expressos no mapa, descritores estes que seriam úteis numa situação real de descrição. Houve a necessidade de introduzir conceitos que descrevessem aspectos administrativos do projeto do qual derivam os dados, assim como aspectos que caracterizem o próprio conjunto de dados. Esta necessidade foi expressa em forma de objetivo e a partir de então foram selecionados alguns conceitos que os pudessem representar.

Para contornar estas situações os seguintes objetivos foram delimitados:

- Escolher descritores que cubram o domínio de forma abrangente, porém o mais completa possível. Caso o número de descritores fique um pouco extenso não é problema, pois no momento da descrição pode-se escolher usar apenas os que se quer. O importante é que estejam disponíveis em caso de necessidade.
- Definir aspetos de ordem administrativa, relativos ao projeto, tais como descrição ou resumo do projeto, responsável/chefe pelo laboratório, órgão de financiamento.
- Definir elementos que identifiquem o conjunto de dados, tais como título, autoria, assunto, natureza ou tipologia dos dados.
- Identificar material coletado.
- Quantificar espécies e indivíduos coletados.
- Localizar os eventos de coleta de dados no tempo e no espaço.
- Especificar métodos e instrumentos para amostragens.
- Especificar destinação de materiais coletados.

Doravante os conceitos do mapa e os outros que surgiram foram classificados e posicionados conforme os objetivos estabelecidos. Por exemplo, para o objetivo “Definir

aspectos de ordem administrativa relativos ao projeto" surgiram os conceitos "órgão de financiamento", "data de vigência do projeto", "descrição ou resumo do projeto".

3.4 TRANSPOSIÇÃO DE CONCEITOS PARA DESCRITORES CONFORME ESQUEMAS DE METADADOS

Durante a etapa de escolha dos conceitos teve início o processo de pesquisa e estudo de normas de descrição de dados e esquemas de metadados existentes no domínio da Oceanografia Biológica ou em domínios relacionados. O conhecimento destas normas também foi útil para orientar os investigadores na seleção de descritores, uma vez que estas poderiam abordar aspectos que não estavam contemplados até o momento, mas que, todavia, poderiam ser cruciais numa situação de descrição.

Depois da seleção definitiva de todos os conceitos iniciou-se a fase de transposição de conceitos para os descritores que formariam o perfil de aplicação. A maioria dos conceitos encontrou compatibilidade semântica nos descritores constantes nos esquemas estudados. Os que não puderam ser encontrados foram criados.

Os descritores atribuídos ao perfil de aplicação foram classificados em duas categorias: descritores gerais e específicos do domínio. Os descritores gerais possuem uma ampla abrangência e se aplicam para descrição de recursos dos mais variados suportes e de áreas de conhecimento diversificadas. São exemplos deste tipo os metadados que representam aspectos descritivos do projeto de investigação, do próprio conjunto de dados, ou de documentos gerados a partir dos dados (artigos, relatórios, etc.), tais como, "title", "creator", "date" ou "subject".

Os descritores específicos são os que buscam mapear o fluxo das atividades de investigação específicas do domínio. Também utilizam vocabulário próprio da área de investigação. Existem casos em que os descritores podem ser generalizados para outros domínios, como é o caso dos descritores "methods" e "instrumentation", porém a maioria só faz sentido dentro de uma mesma área de investigação ou área interdisciplinar, como é o caso dos descritores "scientific name" e "life stage", que se inserem no contexto das ciências biológicas.

Portanto, foram estudados esquemas e normas que contemplassem as duas categorias de descritores, de forma a elaborar um perfil de aplicação completo. Tais esquemas foram localizados no sítio eletrônico do Digital Curation Centre, onde há uma secção que lista padrões para descrição de dados, bem como casos de usos, em vários domínios. Esta secção organiza os padrões por disciplinas ou em lista geral.

De acordo com esta classificação por disciplinas foram escolhidos quatro padrões nas categorias "Biology" (Darwin Core, EML e OBIS) e "Earth Science" (FGDC/CSDGM) e um padrão da categoria "General Research Data" (Dublin Core).

Após análise dos padrões de metadados foi estabelecida a seguinte ordem de preferência de uso: 1º) Dublin Core, para representar os metadados de descrição genéricos e garantir interoperabilidade com outras plataformas e repositórios, uma vez que é amplamente utilizado; 2º) EML, norma bastante compatível para com a descrição na área da Oceanografia Biológica que estuda as interações ecológicas entre ambientes e organismos marinhos e estuarinos; 3º) Darwin Core, apresenta um vocabulário bastante compatível com o domínio, vindo a complementar a norma EML, além disso, também fornece uma ontologia para representação dos conceitos; 4º) OBIS, é uma extensão do Darwin Core e sua aplicabilidade se restringe ao repositório de dados Ocean Biogeographic Information System; 5º) CSDGM Part 1: Biological Data Profile, o vocabulário desta norma se apresentou mais restrito do que o das normas anteriormente citadas, motivo pelo qual foi preterido a estas; 6º) BIOLOGICAL OCEANOGRAPHY, nos casos em que nenhuma das normas estudadas conseguiu representar o conceito foi criado o vocabulário "Biological Oceanography", com prefixo "biocn".

Ao todo foram selecionados trinta e um (31) conceitos, entre gerais e específicos, para compor o perfil de aplicação. Para cada um dos conceitos foram analisadas as normas mencionadas acima a fim de se encontrar compatibilidade semântica entre conceito e descritor. Os descritores que correspondiam semanticamente aos conceitos selecionados eram escolhidos para compor o perfil conforme a ordem de preferência apresentada. Para os conceitos que não encontraram conformidade semântica foram criados descritores sob medida.

A fim de ilustrar como o processo de transposição dos conceitos para descritores aconteceu se pode ver o exemplo da Tabela 2. Nela é possível visualizar alguns dos conceitos selecionados a partir do objetivo "Definir aspectos de ordem administrativa relativos ao

projeto", ao lado destes se vê os descritores correspondentes e em sequencia os respectivos padrões de metadados de onde os descritores foram extraídos.

Tabela 2 – Transposição de conceitos para descritores

CONCEITO	DESCRITOR SUGERIDO	ESQUEMA
Nome do Laboratório /e/ou/ Nome da Instituição	Organization Name	EML
Órgão de financiamento	Funding	EML
Data de vigência do projeto	Date	Dublin Core
Descrição /ou/ resumo do projeto	Description	Dublin Core

Dos 31 descritores apenas 5 foram de fato criados, sendo os restante 26 descritores selecionados de padrões de metadados já existentes. Destes 26 descritores 5 pertencem ao padrão Darwin Core, 8 pertencem ao padrão Dublin Core, 12 ao EML e 1 ao OBIS.

Abaixo um exemplo do perfil de aplicação na forma prefixo:nome. O prefixo se refere ao padrão de metadados de onde o descritor foi retirado. Em seguida são apresentados uma breve descrição do que é o descritor e um exemplo de aplicação prático, ou seja, um exemplo de que informação colocar neste descritor no caso de uma situação real de descrição.

Tabela 3 – Perfil de aplicação Biological Oceanography

APPLICATION PROFILE BIOLOGICAL OCEANOGRAPHY		
DESCRIPTOR	**DESCRIPTION**	**EXAMPLE**
eml:beginDate	A single time stamp signifying the beginning of some time period.	2010-09-12
eml:commonName	Specification of applicable common names. These common names may be general descriptions of a group of organisms if appropriate.	Blue crab
dc:contributor	An entity responsible for making contributions to the resource.	Ictiology Lab.
dc:coverage	The spatial or temporal topic of the resource, the spatial applicability of the resource, or the jurisdiction under which the resource is relevant.	Lagoa dos Patos
dc:creator	An entity primarily responsible for making the resource.	Silva, João da
dc:date	A point or period of time associated with an event in the lifecycle of the resource.	2013-09-16 a 2015-07-30

3.5 FORMALIZAÇÃO DA ONTOLOGIA E SUA INGESTÃO NA PLATAFORMA DENDRO

A etapa posterior à elaboração do perfil de aplicação trata da sua formalização em ontologia, para que desta forma possa ser incorporada à plataforma de gestão de dados e nela os descritores possam atuar como parte do modelo de metadados para a descrição de dados de investigação e outros recursos.

A ontologia dentro da plataforma funciona como um vocabulário usado para descrever um domínio, onde as propriedades são a representação dos descritores do perfil de aplicação. A ontologia é a ferramenta apropriada para fazer tal representação uma vez que apresenta a flexibilidade semântica que a tarefa exige. É importante que a linguagem da ontologia seja a mesma do perfil de aplicação construído em colaboração com os investigadores, para não haver barreira no entendimento dos descritores.

Uma das ideias centrais de representar o perfil de aplicação através de uma ontologia é poder aproveitar o vocabulário de outras ontologias para um objetivo particular. Portanto, para a construção da ontologia "Biological Oceanography" se procedeu a uma pesquisa de ontologias existentes que pudessem ter uma representação dos descritores do perfil de aplicação. Tal processo foi semelhante ao que foi feito com o estudo dos padrões de metadados para compor o perfil de aplicação, o objetivo é encontrar compatibilidade semântica entre os descritores do perfil e os termos das ontologias. Para este caso, não foi preciso analisar as relações entre os conceitos das ontologias, uma vez que esta "lightweight ontology" só se ocupa de classes e propriedades.

Foram analisadas algumas ontologias recomendadas pelos colegas do InfoLab e outras recomendadas pelo World Wide Web Consortium (W3C)[8], sendo estas: Dublin Core, Oboé, CERIF, Friend of a friend (FOAF), TGWD Ontologies, The TaxonConcept Ontology. The Darwin-SW Ontology e Marine TLO.

Das ontologias acima, a Dublin Core e a Friend of a Friend já estão incorporadas na plataforma Dendro, portanto poderia haver compatibilidade de descritores. As ontologias OBOE, CERIF e Marine TLO, apesar de serem altamente recomendadas não apresentaram compatibilidade de vocabulário para com os descritores do perfil de aplicação, portanto não foram utilizadas para o vocabulário "Biological Oceanography". A ontologia "Darwin Core"

[8] World Wide Web Consortium (W3C) - (http://www.w3.org/wiki/Good_Ontologies

cuja norma já havia sido utilizada para compor o perfil de aplicação apresentou compatibilidade.

Após esta análise, outra foi procedida para verificar quais descritores poderiam ser aproveitados das ontologias já existentes na plataforma Dendro, quando verificou-se que muitos descritores do perfil de aplicação já estavam contemplados em outras ontologias da plataforma. Este mapeamento demonstrou que dos 31 descritores 14 já estavam formalizados em 5 ontologias no Dendro, sendo, portanto, 17 os descritores que precisaram ser formalizados na ontologia "Biological Oceanography".

Para a formalização dos descritores ainda não contemplados na plataforma Dendro foi utilizada como base a ontologia "Dendro Reasearch", onde foi criada uma classe chamada "Observation", por se tratar de um tipo de investigação de caráter, sobretudo observacional. Dentro desta foi criada uma subclasse com nome do domínio "BiologicalOceanography". A seguir foram criadas as *data properties* com os nomes dos descritores do perfil de aplicação.

Para cada uma das propriedades criadas foi inserida informação adicional para sua caracterização. Por exemplo, foram inseridas a *annotation property* "label", para estabelecer o formato como o descritor deve aparecer; "comment", onde constou uma descrição do conceito de cada descritor e um exemplo; e "isDefinedBy", que foi usada apenas nos casos de um descritor ter sido criado por alguma norma específica. Neste caso, foi colocada a URL da norma que definiu o descritor, por exemplo, foi colocada a URL da norma EML que definiu o descritor "Begin Date". Ainda foi feita a relação das propriedades com as respectivas classes, através de *Domains (intersection)*, a maioria das propriedades foi ligada à subclasse de "Observation" "BiologicalOceanography".

Depois de pronta a ontologia foi introduzida na plataforma de gestão de dados Dendro, a qual apresenta uma série de perfis de aplicação formalizados em ontologia. Alguns destes perfis são constituídos de descritores genéricos, que se aplicam a várias situações de descrição independentemente do domínio a ser descrito, como é o caso, por exemplo, dos vocabulários Dublin Core, Friend of a Friend e Dendro Research (este último com descritores genéricos aplicáveis às ciências). O restante dos perfis é constituído de descritores para anotação de dados de investigação e outros recursos de domínios específicos.

4 UTILIZAÇÃO E AVALIAÇÃO DO PERFIL DE APLICAÇÃO NA PLATAFORMA DENDRO

De forma a promover a plataforma Dendro como uma ferramenta adequada à gestão de dados de investigação, sobretudo no que se refere à questão da descrição de dados, bem como para avaliar o seu desempenho nesta tarefa, uma série de experiências foi conduzida com investigadores da Universidade do Porto. Participam da campanha de avaliação da plataforma investigadores das áreas em que foram construídas ontologias para descrição de dados. Estes foram convidados a depositar, pelo menos, um conjunto de dados e um artigo publicado e proceder à sua descrição.

Para as experiências no Dendro optou-se por convidar investigadores que não colaboraram com a elaboração dos conjuntos de descritores, para não enviesar a avaliação. Além do mais, sendo os vocabulários construídos com base em experiências específicas de investigação, a opinião de investigadores diferentes poderia fornecer subsídios para apreciação da qualidade e abrangência dos descritores.

A campanha de avaliação se realizou em duas sessões, sendo que na primeira um investigador de cada área faz o depósito e descrição de seus recursos numa primeira versão da plataforma, que não usa um sistema de recomendação de descritores. Na segunda sessão outro investigador da mesma área procede ao depósito e descrição de seus recursos com o sistema de recomendação de descritores ativado na plataforma. A ideia é que a plataforma aprenda com a interação de seus utilizadores e passe a recomendar descritores para facilitar a tarefa de anotação. Participaram da campanha investigadores área de Ciências do Mar e Ambiente do Centro Interdisciplinar de Investigação Marinha e Ambiental (CIIMAR)[9] da Universidade do Porto.

Foi elaborado um guião para as duas primeiras sessões para garantir que as mesmas sejam conduzidas de maneira uniforme. Como os investigadores convidados a participar da experiência não possuíam conhecimento da plataforma nem tampouco da atividade de descrição, as sessões se iniciaram com uma explicação do funcionamento do Dendro, suportado pelo Guião de Utilização do Dendro, e uma explicação da tarefa de descrição e do conceito de descritor.

[9] Centro Interdisciplinar de Investigação Marinha e Ambiental (CIIMAR) - http://www.ciimar.up.pt/

Apesar do objetivo geral da campanha de interações com o Dendro se centrar em testar a plataforma e seu sistema de recomendação de descritores, a experiência foi desenhada de forma a servir também os propósitos de validação das ontologias desenvolvidas. Neste caso, a experiência serviu para inferir sobre a atividade de anotação de dados por parte dos investigadores. Também se aproveitou a ocasião para solicitar aos investigadores que respondessem a um inquérito para avaliar os descritores do perfil de aplicação "Biological Oceanography", que foi encaminhado aos e-mails dos investigadores após sua interação com a plataforma Dendro, juntamente com o perfil de aplicação completo do domínio.

Após as explicações sobre a plataforma deu-se início a primeira sessão de interações onde o "investigador 1" foi convidado a criar uma pasta e depositar seus ficheiros e a descrever ambos utilizando um sistema manual de escolha de metadados.

O investigador, a princípio, teve dificuldades em distinguir a descrição das pastas da descrição dos ficheiros, e demonstrou dificuldades em selecionar os descritores apropriados para a descrição de ambos. Percebeu-se que, no fundo, a tarefa de anotação se apresentou complexa, o que foi agravado pela falta de conhecimento dos vocabulários (ontologias) e respectivos descritores.

O investigador não demonstrou interesse em percorrer descritores das ontologias específicas de outras áreas, fixando-se nas ontologias genéricas e na ontologia de sua área. Ao final da experiência mencionou a dificuldade de navegação nas ontologias, achou que não era muito prática a forma de escolha dos descritores e sugeriu outro tipo de agrupamento dos descritores. Achou a plataforma em si relativamente fácil de usar e intuitiva, depois da demonstração. Também considerou o Dendro uma ferramenta importante para gestão de dados.

Relativamente ao inquérito que avalia o perfil de aplicação, as respostas foram bastante positivas, demonstrando que os descritores do perfil são úteis e compatíveis com o domínio da Oceanografia Biológica. Quando questionado se usou os descritores deste vocabulário nas descrições nas interações com o Dendro ou se os usaria, o investigador respondeu de forma positiva para as duas opções. Também afirmou que considera os descritores úteis e que os usaria para descrever seus dados se os fosse depositar em alguma plataforma.

Relativamente à compatibilidade semântica dos descritores com a linguagem utilizada no domínio, o investigador declarou que considerou o vocabulário criado semanticamente

adequado à investigação no domínio. Também considerou que os descritores eram compatíveis com o domínio no que concerne à sua capacidade de descrever suas atividades de investigação, como por exemplo, os eventos de coleta.

Quanto à abrangência e completude do perfil de aplicação considerou o perfil bastante completo e abrangente em relação aos aspectos de investigação do domínio da Oceanografia Biológica, não modificaria ou suprimiria qualquer um dos descritores. Porém, como sugestão, acrescentaria algo relacionado ao tamanho dos indivíduos coletados, por exemplo, “Mean size of the individual collected”.

Na segunda sessão de experiência de interação com a plataforma de gestão de dados o “investigador 2” utilizou uma versão do Dendro habilitada com um sistema de recomendação de descritores. Ou seja, a plataforma indicou para o investigador metadados compatíveis com sua área de atuação, os quais poderim ser úteis para a atividade de anotação de seus dados.

Para dar início à experiência foi feita uma explanação sobre a plataforma Dendro, seus principais objetivos e funcionalidades com relação à descrição, preservação e partilha de dados de investigação, e ainda foi dada ao investigador uma noção sobre descrição de dados. Então, foi solicitado a este que criasse uma pasta no mesmo projeto anteriormente utilizado pelo investigador que participou da “sessão 1” da campanha de experiências com o Dendro. No caso, os dois investigadores trabalham colaborativamente no mesmo projeto. Depois de criada a pasta o investigador adicionou a ela um conjunto de dados.

O investigador não apresentou dificuldades no uso da plataforma e na escolha dos metadados para descrever os recursos. Percorreu a lista de descritores recomendados e selecionou somente aqueles que considerou necessários para a anotação de seus dados. Percebeu-se que, dos descritores recomendados, muitos dos que foram selecionados pelo investigador pertencem ao vocabulário específico de sua área, “Biological Oceanography”.

Como os descritores recomendados já não satisfaziam sua necessidade de descrição foi aconselhado a percorrer todos os vocabulários no modo de seleção manual, para o caso de haver algum conjunto de descritores mais específico para o seu domínio.

Não foram detectados quaisquer problemas de sistema. Ao final averiguou-se que as descrições ficaram todas guardadas. A experiência toda levou cerca de 30 minutos, metade do tempo que levou a experiência da sessão 1, o que se atribui ao sistema de recomendação de descritores, que escusa o investigador de ter que percorrer muitas listas de ontologias para encontrar os descritores pretendidos.

O investigador 2 foi também convidado a responder ao inquérito que avalia o perfil de aplicação “Biological Oceanography”. Assim como as respostas anteriores estas demonstraram a utilidade e compatibilidade dos descritores do perfil para com o domínio da Oceanografia Biológica. O investigador respondeu que usou os descritores deste vocabulário nas suas interações com o Dendro e os usaria em uma futura atividade de descrição. Também afirmou que considera os descritores úteis para a atividade de descrição de dados de investigação e produtos gerados a partir destes.

Quanto à conformidade semântica dos descritores com a linguagem utilizada no domínio e sua capacidade de descrever as atividades de investigação o investigador declarou que considera os descritores do vocabulário adequado em ambos os casos.

Considerou o perfil de aplicação completo e abrangente, portanto não modificaria ou suprimiria qualquer um dos descritores. Apenas sugeriu a inclusão de opções pré-definidas em um descritor chamado “subárea”, por exemplo, Ecologia, Genética, Imunologia, Aquicultura, etc. A escolha da subárea agilizaria a tarefa de descrição.

A qualidade das descrições feitas pelos investigadores não foi avaliada nas duas etapas de experimentação do Dendro, no entanto, notou-se um empenho de sua parte para descrever os dados da maneira mais correta possível. As perguntas dos investigadores acerca de padrões para descrição, como a língua ou unidades de medida a serem usadas, deixaram este fato evidente. Em nenhum dos casos houve uma descrição fictícia.

Em ambas as sessões foram usados descritores que compõem o perfil de aplicação “Biological Oceanography”, os quais estão distribuídos nas diversas ontologias da plataforma Dendro, e não apenas os descritores da ontologia “Biological Oceanography”.

5 CONSIDERAÇÕES FINAIS

Os dados de investigação já têm sua importância consolidada no atual contexto científico, porém, percebe-se que a questão da gestão de dados de investigação ainda é bastante incipiente. Embora existam organizações com iniciativas proeminentes no desenvolvimento e fornecimento de ferramentas de curadoria digital, políticas de gestão, programas de treinamento e educação, a prática da gestão de dados como uma tarefa planeada que segue determinados padrões ainda não é comum entre grupos de investigação.

Este cenário é compreensível se pensarmos na quantidade de trabalho que os investigadores já possuem para gerar e analisar dados, produzir, validar e comunicar resultados. Para além disso, ainda necessitam se envolver com questões burocráticas dos projetos e das agências de financiamento. Para não citar quando o investigador tem sob sua responsabilidade a tarefa de gestão de pessoas ou bens materiais. Isto, por um lado, justifica a falta de iniciativa dos investigadores para as boas práticas de gestão, e por outro, prova que as iniciativas devem advir de gestores e curadores de dados.

Neste sentido este estudo pretendeu fornecer apoio à gestão de dados de investigação a pequenos grupos de investigadores, buscando abranger várias fases do ciclo de vida dos dados, logo depois de sua criação. Particularmente, se preocupou com questões de documentação, armazenamento e possibilidade de partilha dos mesmos.

Para resolver a questão da documentação de dados, que é essencial para que estes sejam inteligíveis, partilhados entre colaboradores e preservados ao longo do tempo, foi desenhado um perfil de aplicação para atividades de investigação no domínio da Oceanografia Biológica. Para resolver a questão do armazenamento dos dados foi pensada e desenvolvida a plataforma Dendro, onde estes poderiam ser depositados, descritos, e se for do interesse dos investigadores, partilhados com parceiros ou copiados para repositórios externos.

Um perfil de aplicação pode ser desenhado por curadores com a colaboração de investigadores do domínio, ou sem esta colaboração, desde que o curador tenha algum conhecimento sobre ele. No entanto, a experiência de elaboração do perfil com a colaboração dos investigadores mostrou-se valiosa, pelo que se recomenda. É importante que eles sejam vistos como atores do processo, e não apenas como utilizadores da ferramenta.

O envolvimento dos investigadores é fundamental para um entendimento consistente do domínio, o que evita erros que podem afetar a linguagem do perfil de aplicação. Por este motivo a disponibilidade de contato com investigadores e a sua compreensão sobre a importância do processo é importante. Entender e utilizar a mesma linguagem do investigador no perfil de aplicação é outro fator a ser destacado, uma vez que esta não pode ser uma barreira para a descrição de dados, antes pelo contrário, deve ser facilitadora do processo de descrição.

Em relação à prática da documentação dos dados esta tem que ser o mais simples possível para o investigador, uma vez que suas atividades de investigação já lhe demandam um esforço considerável. A tarefa adicional de descrever dados pode ser um peso extra que o

investigador tem que carregar. Somado a isso, o desconhecimento e inexperiência em relação à atividade de anotar pode ser um impedimento para que a realizem. Portanto, todos os recursos para facilitar a tarefa de descrição devem estar disponíveis ao utilizador, desde a criação de metadados compreensíveis até o emprego de tecnologia inteligente em repositórios ou plataformas de gestão de dados.

Num cenário ideal existiria um gestor de dados junto a cada grupo de investigação para fazer o tratamento desta informação tão logo ela fosse gerada. Esta tarefa poderia ficar a cargo de profissionais da informação, que têm conhecimento e experiência em representação e organização do conhecimento, bem como em gestão da informação. No entanto, este cenário está muito longe de se tornar realidade. Logo, o que poderia tornar-se uma alternativa viável, é que estes profissionais tomassem para si a responsabilidade de promover as boas práticas de gestão de dados junto aos investigadores e fornecessem a eles as ferramentas necessárias para praticá-las.

Por outro lado, agências de financiamento também deveriam fornecer suporte para que investigadores preservem e partilhem seus dados, através de iniciativas tomadas em direção à implantação de práticas uniformizadas de gestão de dados, ao desenvolvimento de políticas ou fornecimento programas de treinamento. Além de fornecer o apoio financeiro necessário à produção dos dados, é preciso também assegurar que estes se mantenham relevantes de modo que possam ser reutilizados ao longo do tempo. Assim, o ciclo de investimentos nas atividades de investigação é otimizado, evitando repetições nos processos de recolha de dados.

De fato, é imperativo sensibilizar os investigadores para a elaboração de planos de gestão de dados, para a importância de documentar, preservar e partilhar dados de investigação. Este é um compromisso que deve ser partilhado por várias instâncias, porém deve haver um esforço dos próprios investigadores para que seja instalada a cultura do depósito, descrição e partilha de dados.

Neste momento, em muitos grupos de investigação, ainda prevalece uma visão de que os dados, mesmo que gerados com recursos públicos, pertencem a eles, portanto, é sua a decisão do que fazer com os dados. São iniciativas, como a que está a decorrer na FEUP, que contribuem para uma mudança de paradigma. A disponibilização aos investigadores de uma plataforma onde possam, de forma segura e fácil, depositar e documentar seus dados, pode ser o primeiro passo para que os mesmos sejam partilhados com repositórios externos. A

campanha de interações dos investigadores com a plataforma Dendro a decorrer está a ser positiva neste sentido, mostrando abertura e disponibilidade dos investigadores que dela participaram.

A experiência de elaboração de um perfil de aplicação para integrar a lista de conjuntos de descritores da plataforma Dendro foi muito satisfatória, uma vez que sua utilização para descrição de dados foi efetiva. Embora tenha sido elaborado para atividades de investigação específicas do domínio da Oceanografia Biológica no Brasil, na campanha de interação com o Dendro os descritores do perfil foram selecionados por investigadores de área correlata na Universidade do Porto, o que demonstra sua utilidade em situações práticas de descrição. Estes investigadores ainda avaliaram o perfil como completo e abrangente. Durante a experimentação do Dendro notou-se ainda que alguns descritores foram usados por investigadores de outras áreas, o que reforça a contribuição do perfil de aplicação para com uma plataforma que funciona de forma colaborativa e multidisciplinar.

A abordagem metodológica utilizada para elaborar o perfil de aplicação “Biological Oceanography” é flexível e pode ser repetida para construção de perfis em outras áreas do conhecimento. As experiências do InfoLab demonstraram que isso é factível, e que as descrições podem ser feitas em qualquer domínio, que trabalhe com qualquer tipo de dados. Os mapas de conceito se mostraram muito úteis para a modelagem do conhecimento nos domínios, tanto para o curador quanto para os investigadores. E a ontologia é uma ferramenta eficaz para formalização dos descritores e seu funcionamento dentro da plataforma de gestão de dados.

AGRADECIMENTOS

Este trabalho contou com a estimada colaboração de investigadores do Instituto de Oceanografia da Universidade Federal do Rio Grande (IO/FURG) para elaboração do perfil de aplicação, com a colaboração de investigadores do Centro Interdisciplinar de Investigação Marinha e Ambiental da Universidade do Porto (CIIMAR/UP), para testar e avaliar o perfil criado, e com a colaboração da equipe do projeto Dendro (InfoLab/FEUP) na integração do perfil na plataforma de gestão de dados. A todos estes profissionais nosso agradecimento.

REFERÊNCIAS

ABBOTT, Daisy. **What is digital curation?** DCC Briefing Papers: Introduction to Curation. Digital Curation Centre: Edinburgh, 2008. Disponível em: <http://www.dcc.ac.uk/resources/briefing-papers/introduction-curation/what-digital-curation>Acesso em: 29 Dez. 2014.

ALMEIDA, Maurício B.; BAX, Marcello P. Visão geral sobre ontologias: pesquisa sobre definições, tipos, aplicações, métodos de avaliação e de construção. **Ciência da Informação**, Brasília, v. 32, n. 3, p. 7-20, set./dez. 2003.

CASTRO, João Aguiar; RIBEIRO, Cristina e SILVA, João Rocha da. **Designing an Application Profile using qualified Dublin Core**: a case study with fracture mechanics datasets. Proc. Int'l Conf. on Dublin Core and Metadata Applications 2013, Lisboa, 2 a 6 de Setembro de 2013. Disponível em: http://dcpapers.dublincore.org/pubs/article/view/3685 Acesso em: 30 Set. 2014.

CASTRO, João Aguiar; SILVA, João Rocha da e RIBEIRO, Cristina. **Creating lightweight ontologies for dataset description: practical applications in a cross-domain research data management workflow**. Joint Conference on Digital Libraries, Londres, 8 a 12 de Setembro de 2014.

CORTI, Louise; VAN DEN EYNDEN, Veerle; BISHOP, Libby; WOOLLARD, Matthew. **Managing and sharing research data**: a guide to good practice. Los Angeles: SAGE, 2014. ISBN: 978-1-4462-6726-4.

COSTA, Maíra Murrieta e CUNHA, Murilo Bastos da. O bibliotecário no tratamento de dados oriundos da e-science: considerações iniciais. **Perspectivas em Ciência da Informação**,Belo Horizonte, vol. 19, n. 3, Jul./Set. 2014. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362014000300010&lng=pt&nrm=iso&tlng=en> Acesso em: 26 Mar. 2015.

DIGITAL CURATION CENTRE. What is digital curation? Disponível em: <http://www.dcc.ac.uk/digital-curation/what-digital-curation> Acesso em: 29 Dez. 2014.

HEERY, Rachel e PATEL, Manjula. Application profiles: mixing and matching metadata schemas. **Ariadne**, v. 25, Set. 2000. Disponível em: http://www.ariadne.ac.uk/issue25/app-profiles Acesso em: 22 Out. 2014.

RODRIGUES, Eloy *et al.* **Os Repositórios de Dados Científicos**: estado da arte. Projecto RCAAP D24 – Relatório, 2010. Disponível em: <http://hdl.handle.net/1822/10830> Acesso em: 24 Out. 2014.

SILVA, João Rocha da *et. al.* **Dendro**: Collaborative Research Data Management Built on Linked Open Data. In: The Semantic Web: ESWC 2014 Satellite Events. Série Lecture Notes in Computer Science. Creta: Springer, 2014. Disponível em:

http://link.springer.com/chapter/10.1007/978-3-319-11955-7_71 Acesso em: 01 de Outubro de 2014.

TAYLOR, John. Definig eScience. National eScience Centre. Disponível em: <http://www.nesc.ac.uk/nesc/define.html> Acesso em: 09 Dez. 2014.