cadernos BAD

7ª conferência luso-brasileira acessoaberto Instituto Politécnico de Viseu - Portugal

Construção de um repositório de dados oceanográficos

Ricardo Amorim
INESC TEC, Faculdade de Engenharia, Universidade do Porto
ricardo.amorim3@gmail.com

João Castro
INESC TEC, Faculdade de Engenharia, Universidade do Porto
joaoaguiarcastro@gmail.com

Inês Garganta
INESC TEC, Faculdade de Engenharia, Universidade do Porto
ei10162@fe.up.pt

Artur Rocha
INESC TEC, Faculdade de Engenharia, Universidade do Porto
artur.rocha@inesctec.pt

Gabriel David
INESC TEC, Faculdade de Engenharia, Universidade do Porto
gtd@fe.up.pt

Resumo

O artigo descreve a conceção e arquitetura de um repositório de dados de investigação de campanhas oceanográficas. O ponto de partida é o modelo de dados OGC Sensor Observation Service, complementado com uma extensão que visa responder às especificidades do caso de aplicação, as campanhas do projeto BIOMETORE, coordenado pelo Instituto Português do Mar e da Atmosfera. Escolheu-se uma implementação de software aberto do SOS, a qual é complementada com dados relativos às campanhas, equipas e documentos. Definiram-se os esquemas de metadados relevantes, por exemplo, para a classificação das espécies. Desenvolveu-se uma aplicação de carregamento dos dados baseada nos formulários usados pelos investigadores, com uma versão auxiliar móvel. E serão desenvolvidos serviços de

visualização dos dados e dos resultados processados, aptos a serem integrados em agregadores de dados marinhos.

Palavras-chave: Gestão de dados de investigação, Aquisição de metadados, Dados oceanográficos, Diretiva INSPIRE

Building an oceanographic data repository

Abstract

The paper describes the overall design and architecture of a research data repository for oceanographic campaigns. The starting point is the OGC Sensor Observation Service (SOS) data model, complemented by an extension intended to answer the specifics of the application case, the campaigns of the BIOMETORE project, coordinated by the Portuguese Sea and Atmosphere Institute. An open source SOS implementation has been chosen and it has been complemented with data on the campaigns, the teams, and the documents. The relevant metadata schemes have been defined, for instance, the species classification scheme. A data management application has been developed, based on the forms used by the researchers, with an auxiliary mobile version. Visualization services will be implemented, both for data and for processed results, able to be integrated in marine data aggregators.

Keywords: Research data management, Metadata acquisition, Oceanographic data, INSPIRE directive

Introdução

A investigação oceanográfica, tal como muitas outras observações da Natureza, produz dados em grande medida únicos, pois as campanhas de recolha são irrepetíveis, tal a quantidade de variáveis não controladas pelos investigadores. A isso acresce o custo habitualmente elevado das campanhas, pelo navio, equipamentos e equipas que mobilizam. O valor destes dados deriva não só do contexto de produção como também do seu uso potencial. Destinam-se a contribuir para monitorizar e suportar decisões relativas ao Bom Estado Ambiental das Águas Marinhas Europeias. É, portanto, imperioso que estes dados possam ser partilhados e replicados da forma mais alargada possível.

Acontece que tradicionalmente as técnicas de recolha, armazenamento e disseminação dos dados dos projetos de investigação oceanográfica não têm permitido uma partilha significativa. Os dados são muitas vezes registados em formulários em papel, cuidadosamente preparados, e posteriormente digitalizados ou copiados para folhas de cálculo, mas a ênfase tem estado mais na publicação dos artigos com os resultados da investigação do que na dos dados recolhidos. Existem também fluxos de dados provenientes de sensores e de satélites, recolhas de amostras de solo, de água, de plantas e de animais, diários de pesca, percursos de navios, gravações de fotografias, vídeo e som e resultados de análises laboratoriais.

Este problema reflete-se no projeto BIOMETORE (IPMA, 2015), um projeto de grande dimensão coordenado por investigadores do Instituto Português do Mar e da Atmosfera. A equipa de investigadores neste projeto é multidisciplinar e incluí geofísicos, biólogos e químicos que, em oito campanhas, produzem quantidades significativas de registos que devem ser tornados públicos, num suporte digital online, de forma a facilitar a sua reutilização e o seu processamento por diversos destinatários. É esse o objetivo do projeto complementar SeaBioData (INESCTEC, 2015).

O projeto SeaBioData visa organizar e armazenar todos os tipos de dados mencionados e os respetivos metadados de contexto, descritivos e técnicos. Pretende ainda disponibilizá-los, diretamente ou após processamento, em serviços de visualização e de interoperabilidade. Apresentam-se em seguida as principais decisões de arquitetura do sistema.

Arquitetura proposta

O ponto de partida, dada a complexidade dos dados, e também a existência da diretiva europeia INSPIRE para o desenvolvimento de uma Infraestrutura para a Informação Geoespacial (European Commission, 2007), aplicável ao registo de informação ambiental georreferenciável, foi a adoção do modelo de dados OGC Sensor Observation Service (Bröring et al., 2014). Este modelo assenta num conjunto de abstrações suficientemente genéricas para poderem ser úteis em múltiplos tipos de dados: a observação, as propriedades da observação, o processo utilizado, a referência geográfica, etc.

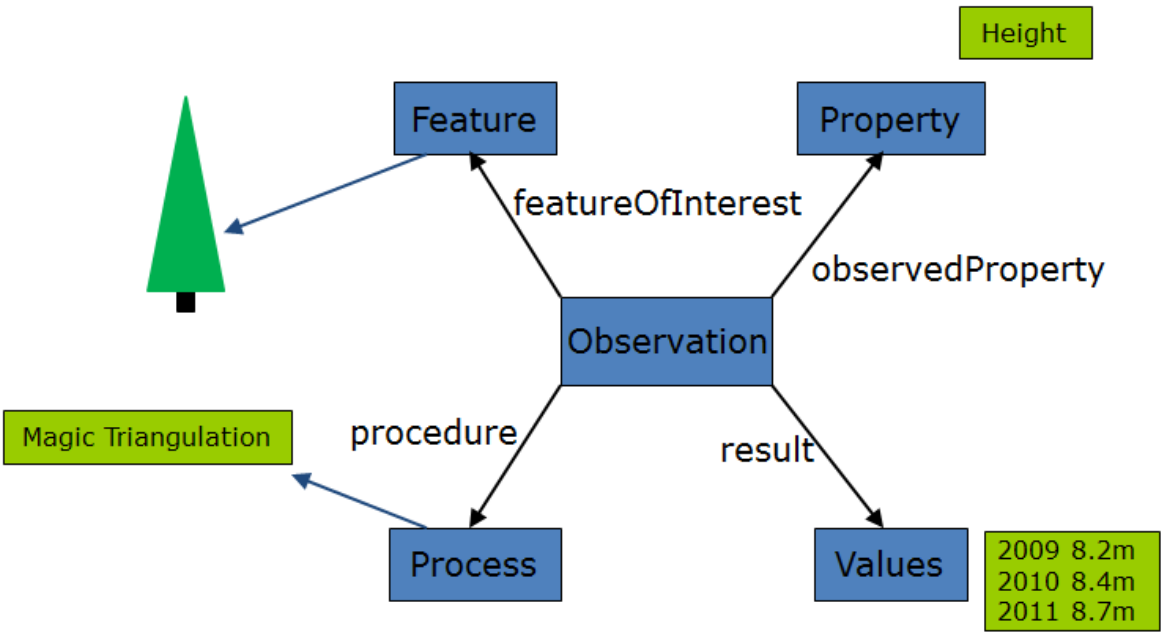

Figura 1: O conceito de observação (INSPIRE, 2011)

A figura 1 está centrada no conceito de observação. A observação é sobre um aspeto de interesse (Feature), habitualmente com uma expressão geográfica: um ponto, uma linha,

uma área, etc. Na figura trata-se de uma árvore, com determinadas coordenadas geográficas. A observação é realizada de acordo com um determinado procedimento (Process), que pode incluir referências ao método de obtenção e ao equipamento específico utilizado. Na figura, o método utilizado foi a triangulação mágica. A observação pode abranger diversas propriedades (Property), cada qual com o seu valor medido (Value). A propriedade observada foi a altura da árvore e poderia ter sido observada apenas num instante, produzindo um único valor. Neste exemplo, optou-se por mostrar uma série de três valores medidos, em três anos consecutivos.

Se, como é frequente no caso de aplicação da oceanografia, a uma observação corresponderem várias amostras, por exemplo, para procedimentos laboratoriais subsequentes, é possível considerar cada uma como observação subordinada à primeira, mudando o aspeto de interesse e considerando as respetivas propriedades. Por exemplo, se forem recolhidas folhas a várias alturas da árvore, cada uma passaria a ser o aspeto de interesse de uma observação subordinada, relacionada com a observação principal na árvore.

Usar este modelo constitui o primeiro passo para garantir a interoperabilidade semântica com outros sistemas baseados no mesmo modelo. Optou-se, em seguida, por utilizar a implementação em software aberto 52° North do SOS, que suporta nas suas estruturas de dados a maior parte dos conceitos e possui API REST e de serviços Web.

No entanto, as especificidades do projeto Biometore, aconselharam a acrescentar uma extensão a este modelo, também com API REST similar à existente, para armazenar metadados relativos aos projetos, às campanhas, às equipas, aos documentos, etc. Com efeito, a produção dos dados de investigação é realizada num contexto específico, que é relevante conhecer, de modo a contribuir para a correta interpretação das condições da obtenção dos dados e para a autenticidade da informação.

Desta extensão local (ver figura 2), salienta-se a tabela de projetos, que regista os detalhes de todos os projetos no âmbito dos quais foram produzidos dados. Os projetos mais complexos organizam-se em work packages. Tanto os projetos como os work packages são realizados por equipas de investigadores (*user*). Os investigadores estão afiliados em departamentos (*department*) de organizações (*organization*) e podem ser registados contactos tanto para os investigadores como para as organizações (*ci_contact, ci_address, ci_telephone*). A tabela *ci_responsibleparty* faz a ligação entre os investigadores e o seu papel concreto num work package de um projeto. É esta tabela que se relaciona com a tabela de campanhas (*campaign*), as quais são vistas como estruturando as atividades de realização de observações. Com esta informação, será possível definir políticas de acesso aos dados, acompanhando as várias fases de desenvolvimento de um projeto.

A preparação de uma campanha inclui uma série de atividades que convém registar, com o objetivo de facilitar o trabalho de registo durante o decorrer da campanha e,

simultaneamente, melhorar a qualidade dos respetivos dados. Estas atividades incluem o planeamento das estações (*stations*) a visitar e a identificação dos procedimentos (*procedureslist*) a seguir em cada uma dessas estações.

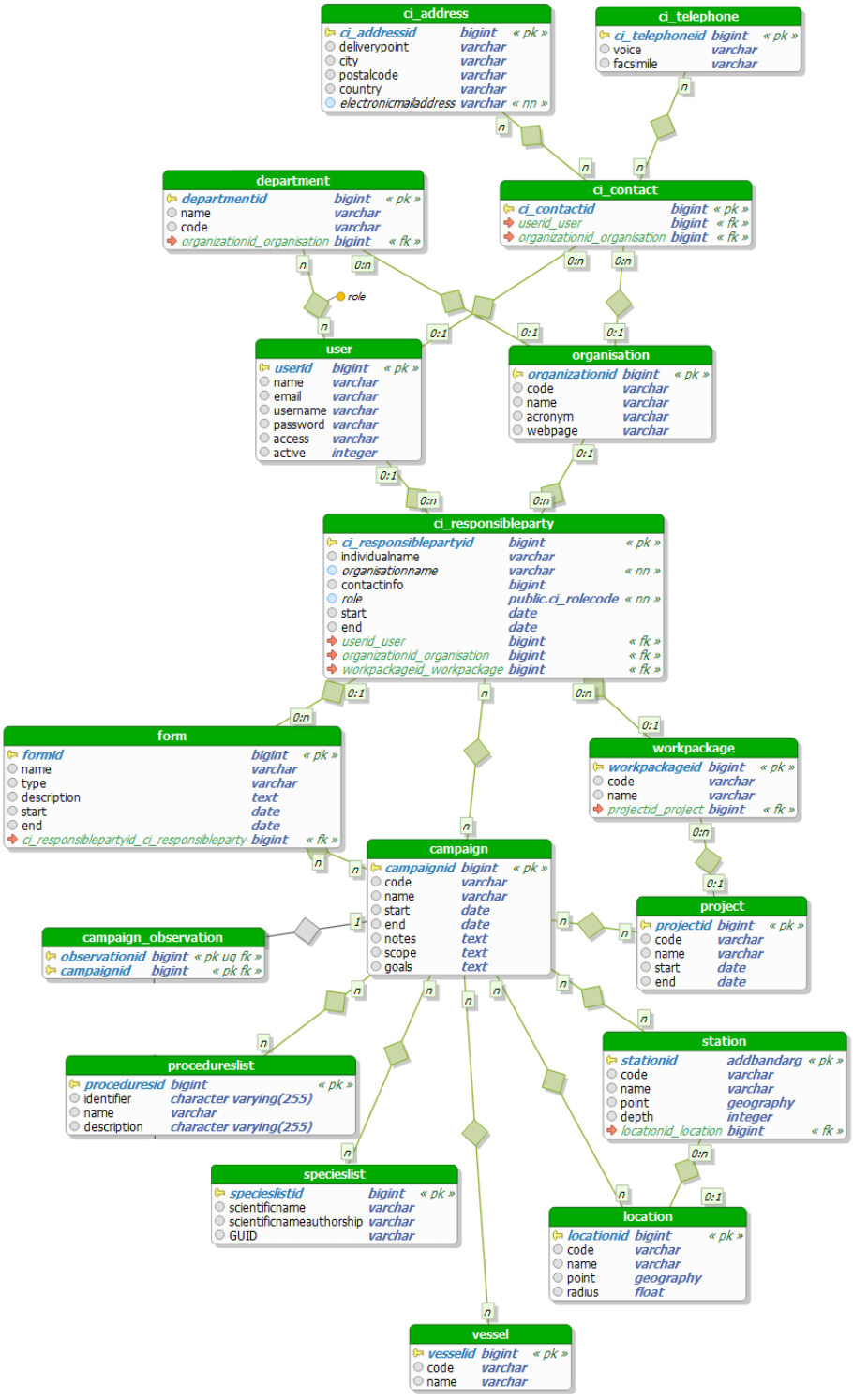

Figura 2: Modelo da base de dados local do SeaBioData (Garganta, 2016)

Um procedimento inclui um método e eventualmente equipamentos. Um exemplo de uma tabela de equipamentos é a tabela de navios (*vessel*), mas toda a lista de logística a

utilizar deve estar disponível, de forma a facilitar a seleção de um equipamento específico para uma instância de um procedimento. Os procedimentos são ainda associados a formulários (*form*), definidos a priori. Da experiência tida, obtiveram-se resultados importantes, a informação dos procedimentos e dos formulários deve ser flexível, admitindo evoluções que se adaptem às decisões tomadas durante a própria campanha. Isso pode significar a inclusão ou exclusão de propriedades para novas observações segundo o mesmo procedimento.

Devido às condições das operações de recolha ou às preferências dos investigadores, uma parte significativa dos registos começa por ser em formulários em papel, os quais são posteriormente transcritos para o repositório. Mas, e ainda na perspetiva de aumentar a autenticidade da informação, é requisito local que esses documentos primários sejam digitalizados e os respetivos ficheiros associados ao repositório. Neste caso, optou-se por armazenar os ficheiros numa hierarquia de pastas que mapeia a organização do repositório, por projeto, campanha e procedimento e, dentro disso, por ordem cronológica. Os metadados relevantes ficam na base de dados. Desta forma é possível eliminar o papel, sem perda de informação.

O armazenamento de ficheiros é também relevante para os vários tipos de dados que não sejam considerados adequados para carregamento como valores de propriedades de observações. Este é o caso das fotografias de aspetos de interesse, que podem ser consideradas como propriedades de uma dada observação, sendo o caminho até aos respetivos ficheiros registado enquanto tal. A fotografia pode por sua vez ser considerada como um aspeto de interesse ela própria, se for, por exemplo, sujeita a um processamento para se determinar a contagem de espécies. Os metadados técnicos da fotografia podem ficar na tabela de ficheiros. Algo de semelhante se passa com os vídeos, com os registos sonoros e até com os ficheiros CTD, caso se opte por não carregar as suas linhas individualmente.

Um outro elemento distintivo deste projeto é a preocupação com a definição prévia dos esquemas de metadados mais adequados a cada situação. A definição do modelo de metadados seguiu a estrutura da norma ISO 19115 para a descrição de dados geográficos, garantindo assim a descrição de dados em conformidade com a regulamentação INSPIRE para os metadados. Contudo, tendo em conta o perfil multidisciplinar do projeto BIOMETORE, foi identificada a necessidade de incluir descritores de forma a permitir a classificação de espécies e a descrição de processos metodológicos associados às amostras. Assim reutilizou-se um subconjunto de conceitos da Ecological Metadata Language e da Darwin Core. Todos os conceitos identificados para o perfil de metadados foram propostos em reuniões junto dos investigadores, para que estes tivessem a oportunidade de validar, excluir ou propor novos conceitos. Para além disso, para simplificar e reduzir o esforço dos investigadores na descrição de dados foram desenvolvidos vocabulários controlados tanto

previstos em normas existentes como definidos de acordo com as prioridades dos investigadores.

Alimentação do repositório

O melhor momento para registar metadados é o do registo dos próprios dados. Por isso, incluiu-se no projeto a preparação de um conjunto de tablets, com a aplicação LabTablet. As ferramentas para a descrição de dados incluem o caderno de laboratório eletrónico, Labtablet, e a definição de um modelo de metadados. No domínio da descrição de dados, existem várias ferramentas que dão apoio aos investigadores e se integram no ambiente laboratorial ou de campo. Neste contexto, os cadernos de laboratório eletrónicos são ferramentas essenciais para a captura de metadados que permitem, entre outros aspetos, a perceção do contexto de produção do conjunto de dados. O LabTablet é um caderno de laboratório eletrónico com ênfase no suporte para metadados compatíveis com normas existentes em cada domínio. Neste contexto, a aplicação também dá suporte ao registo de recolha de amostras, seguindo formatos para a recolha que estão já estabelecidos no domínio.

Está a ser desenvolvida uma aplicação para funcionar como interface de registo e pesquisa simples para os investigadores, a qual deve estar o mais próxima possível dos formulários que estes criaram e conhecem. Esta aplicação é encarada como interface local sobre dois componentes, a base de dados interoperável do SOS da 52° North e a extensão local para aspetos específicos do projeto BIOMETORE.

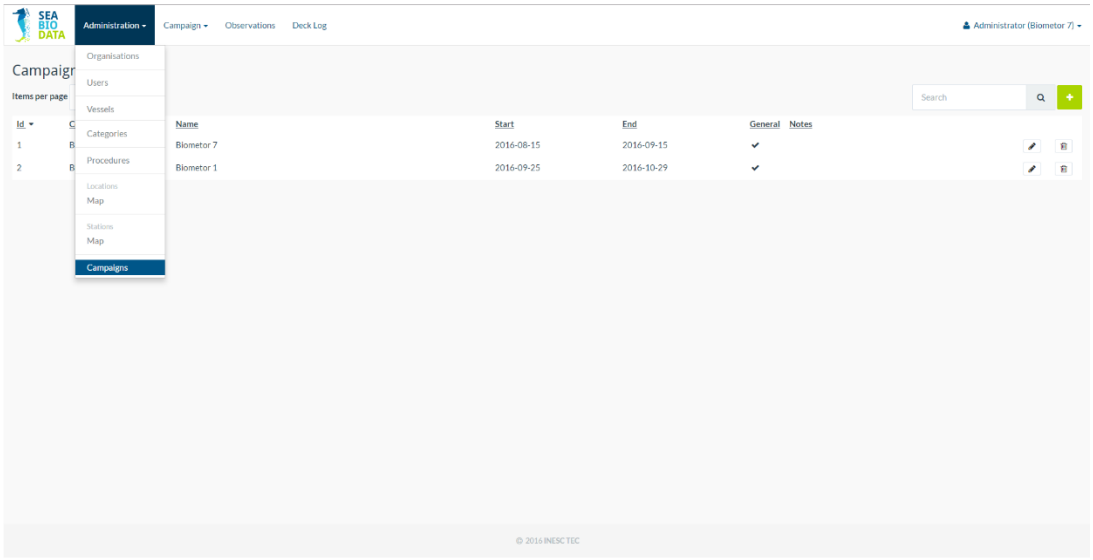

Figura 3: Interface de Administração

Aos formulários é necessário acrescentar visualizadores para os diversos tipos de dados e até para alguns dados processados. Serão também desenvolvidos serviços sobre os

dados aptos a serem importados por plataformas de agregação de dados marinhos. Nas figuras 3, 4, 5 e 6 está apresentada parte da interface desenvolvida.

A componente de administração (figura 3) possibilita a gestão das organizações, utilizadores, navios, procedimentos, estações, etc., associados ao projeto. Foi implementada ainda a funcionalidade de criação de procedimentos (figura 4), de modo a que os investigadores consigam gerir todas as componentes dos projetos/campanhas

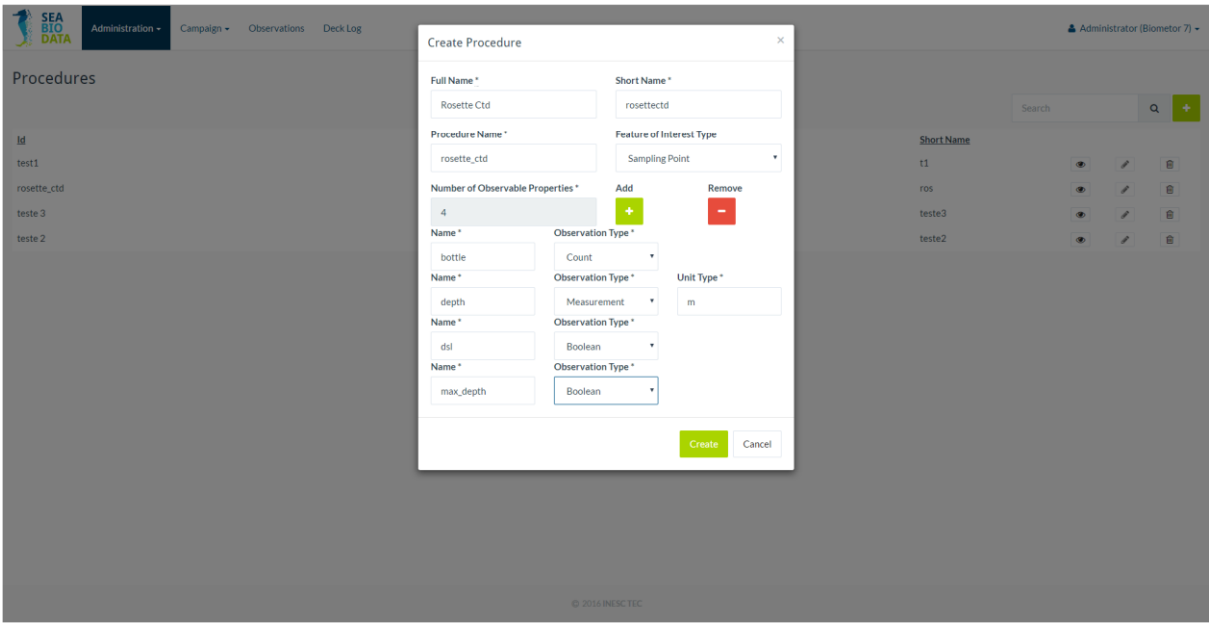

Figura 4: Interface de criação de procedimentos

A gestão da campanha (figura 5) permite associar utilizadores, navios, procedimentos, localizações e estações. Esta gestão deve ser feita antes do início de cada campanha, podendo ser atualizada conforme necessário durante a mesma.

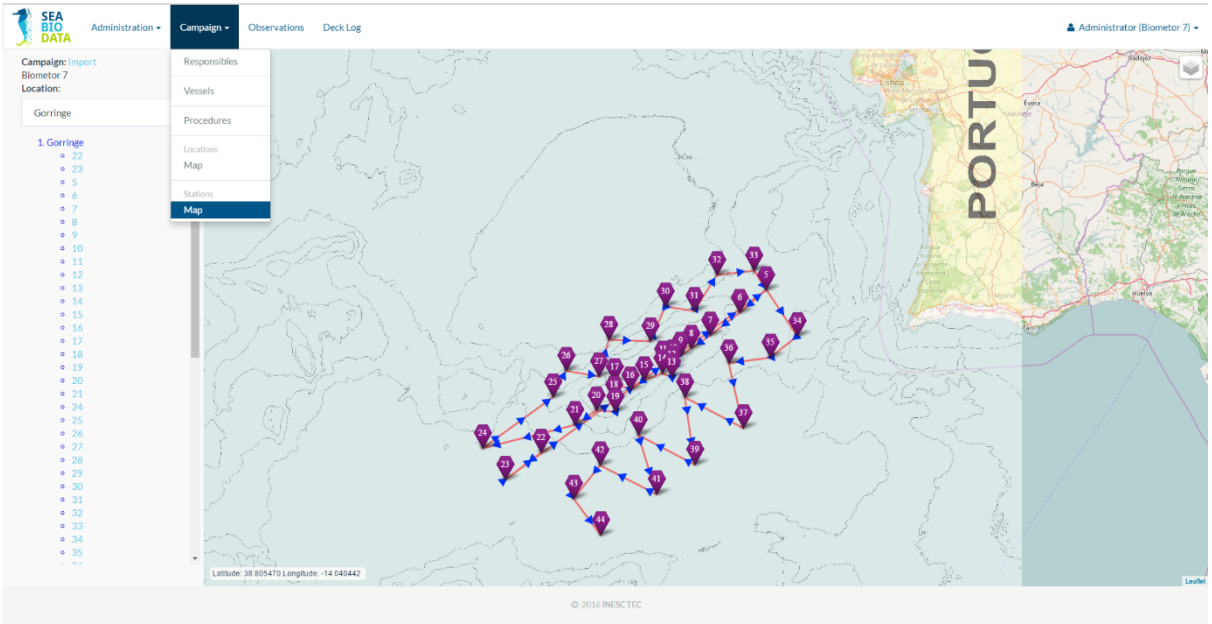

Figura 5: Interface de gestão da campanha ativa, com a opção mapa das estações selecionada

De destacar a funcionalidade de criação/remoção de estações, capaz de acrescentar pontos diretamente num mapa (API do Google Maps) e definir depois a ordem de visita. Na visualização das observações introduzidas (figura 6) foi implementada uma vista em tabela, de modo a assemelhar-se com a interface de uma folha de cálculo, facilitando assim o reconhecimento e a adaptação dos investigadores à nova plataforma. Acrescentou-se também aqui a funcionalidade de poder associar ficheiros, como fotografias, a uma determinada observação.

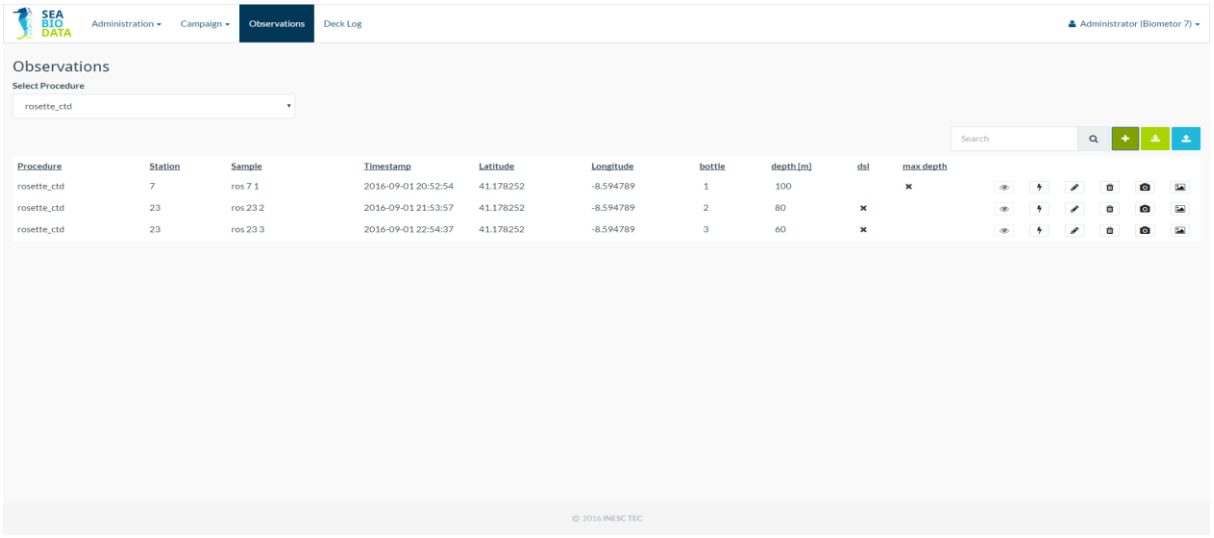

Figura 6: Interface de visualização das observações introduzidas para o procedimento selecionado

A interface de inserção de uma nova observação (figura 7) apresenta os campos que foram definidos como fazendo parte da observação durante a criação do respetivo procedimento, mais um cabeçalho com a estação da recolha, o código da amostra e ainda a posição (latitude e longitude) mais a data e hora da recolha.

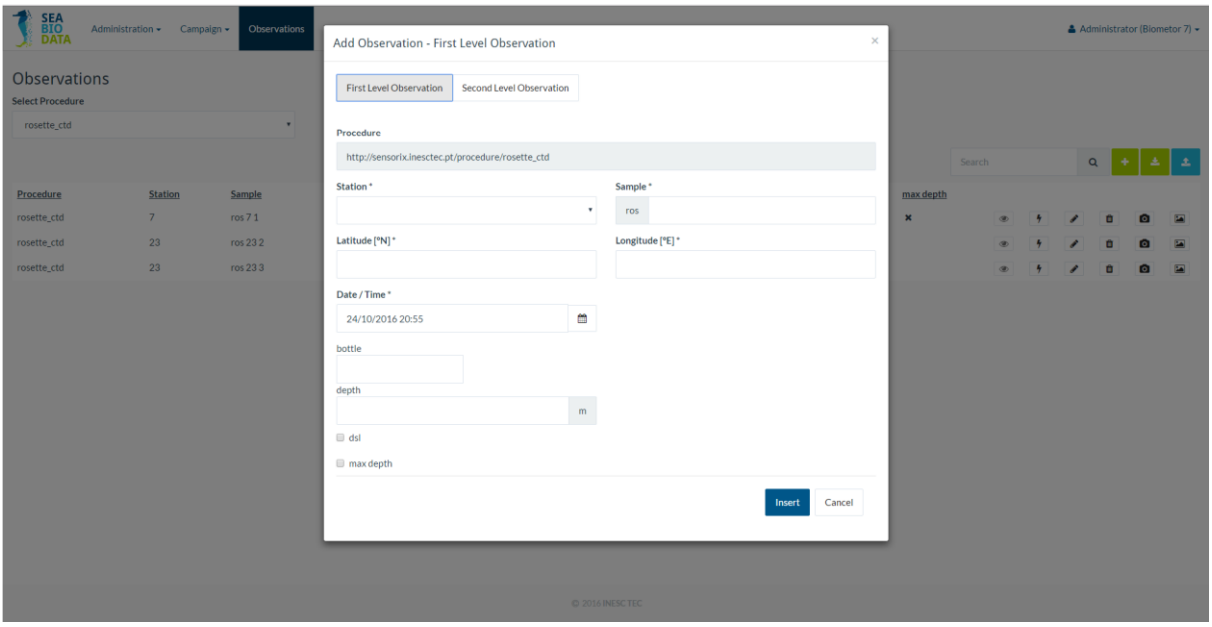

Figura 7: Interface de introdução de uma observação

A interface apresentada foi desenvolvida com técnicas de adaptabilidade que a tornam adequada tanto para computadores como para tablets.

Conclusão

Uma das tarefas prevista para o projeto era a realização de um primeiro teste em ambiente de campanha, que permitisse avaliar os pressupostos do ciclo de vida da produção de dados, especialmente nas primeiras fases de planeamento de campanhas e de utilização a bordo, e a usabilidade da interface de recolha de dados. Esse teste já teve lugar, no mês de setembro de 2016.

As reações obtidas já permitiram corrigir aspetos de usabilidade e reconhecer as vantagens do preenchimento por omissão de alguns valores, como as coordenadas GPS e o instante correntes, ou o procedimento, a estação e o operador do registo anterior, sem prejuízo de qualquer destes valores poder ser corrigido manualmente. Os tablets são ainda úteis para cronometrar e registar durações e tirar fotografias. Um segundo aspeto que ficou claro foi a necessidade de flexibilidade no planeamento, permitindo correções no percurso, com a inclusão de novas estações e a eliminação de estações ainda não utilizadas. Foi necessário até efetuar modificações nos procedimentos, de molde a permitir incluir propriedades não previstas. Para que tal seja possível, é necessário que nos resultados seja possível registar um valor nulo ou "não aplicável", para preencher a nova propriedade nas observações anteriores à modificação.

O teste realizado não permite ainda estimar a percentagem de registos primários que continuarão a ser efetuados em formulários em papel, dado o caráter experimental do teste. É, no entanto, de crer, que essa percentagem continuará a ser não desprezável, o que coloca a questão da digitalização desses documentos e da sua associação às respetivas transcrições no repositório.

Os próximos passos do projeto são a realização de um teste em ambiente laboratorial, o carregamento sistemático dos dados das campanhas anteriores e a elaboração de serviços de visualização.

Nota: Este trabalho foi suportado pela EEA Grant PT02_Aviso5_0002 SeaBioData – Portuguese Seamounts Biodiversity Data Management.

Referências bibliográficas

BRÖRING, A.; STASCH, C.; ECHTERHOFF, J. (2014) – *OGC Sensor Observation Service Interface Standard*, Tech. Rep. OGC and ISO 19156:2011(E), Open Geospatial Consortium (2014).

EUROPEAN COMMISSION (2007). *Council Directive 2007/2/EC of the European Parliament and of the Council of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE)*. OJ L108/1.

GARGANTA, Inês (2016) – *Exploring the Sea: Heterogeneous Geo-Referenced Data Repository*, dissertação de mestrado em Engenharia Informática e Computação, Faculdade de Engenharia da Universidade do Porto.

INSPIRE Working Group on Observations & Measurements (2011). *D2.9 Guidelines for the use of Observations & Measurements and Sensor Web Enablement – related standards in INSPIRE Annex II and III data specification development*, p16 [Em linha: http://inspire.ec.europa.eu/documents/Data_Specifications/D2.9_O&M_Guidelines_v2.0rc3.pdf] [Consult. 2016-10-23].

IPMA (2014) – *Projeto EEA Grants PT02-0018 BIOMETORE – Biodiversity in seamounts: the Madeira-Tore and Great Meteor* [Em linha: http://www.biometore.pt] [Consult. 2016-10-23].

INESCTEC; IPMA (2015) – *Projeto EEA Grants, PT02_Aviso5_0002 SeaBioData – Portuguese Seamounts Biodiversity Data Management* [Em linha: http://proj.inesctec.pt/seabiodata] [Consult. 2016-10-23].