TÍTULO PROJETO: PLANO DE GESTÃO DE DADOS

1. EQUIPE TÉCNICA

Nome	Função	Instituição
Helena de Godoy Bergallo	Coordenação	UERJ
Paulo Eduardo Marques	Administrador de base de dados	UFES
Marcelo Segatto	Gestor da base de dados	UFES
William Ernest Magnusson	Concepção da base de dados	INPA
David Valentim Dias	Estruturador da base de dados	Profissional Liberal
Bolsista Técnico	Apoio técnico	

2. ESCOPO

O projeto “Programa de monitoramento da Biodiversidade Aquática da Área Ambiental I” irá gerar uma grande quantidade de dados que irá subsidiar a elaboração e a implementação de medidas para a recuperação e conservação da fauna aquática na Área Ambiental I. Porém, dados de inventários biológicos são muito variáveis dependendo do organismo e precisam de abordagens específicas. Assim, é muito difícil prever todos os possíveis tipos de dados e estrutura das amostragens a priori e, portanto, definir uma estrutura de banco de dados que funcione para tudo e todos. Embora não seja possível ter um banco de dados para tudo, é possível ter um repositório para tudo. Para que o repositório seja de boa qualidade é necessário que os dados venham com os metadados, que são as informações detalhadas sobre os dados.

Para que o repositório de dados funcione precisamos capacitar pessoas, que se dediquem a receber e checar os dados e metadados, que dialoguem com os geradores dos dados para corrigir os erros e finalmente disponibilizem os dados e metadados on line. Por outro lado, os pesquisadores precisam compreender o que são e como descrever os metadados, para que possam construir as tabelas de dados de forma que as informações originais não sejam sumarizadas e que as chaves primárias importantes estejam presentes. Quando possível, é importante que os tomadores de decisões estejam envolvidos nas decisões sobre a estrutura do repositório, especialmente em relação às chaves primárias.

3. OBJETIVO

O objetivo do presente projeto é de organizar, permitir a integração, disponibilizar e capacitar pessoas para gerenciar os dados e metadados, de forma que os dados sejam mantidos com segurança, e estejam disponíveis para as gerações de tomadores de decisão e pesquisadores ao longo do tempo. Os objetivos específicos são:

1- Conhecer as pesquisas que estão sendo desenvolvidas para verificar as unidades amostrais que estão sendo utilizadas e os dados coletados;

2- Definir chaves primárias para serem utilizadas nos monitoramentos, que permitam a integração dos dados entre os estudos;

3 – Definir o repositório para armazenamento dos dados e metadados;

4 - Capacitar e treinar técnicos, alunos e pesquisadores a gerenciar dados e metadados;

5 - Definir práticas de curadoria dos dados junto aos pesquisadores para preservar e manejar os dados;

6 - Criar e manter um portal de acesso aos dados e metadados;

7 - Definir a infraestrutura física e humana necessária para a implementação do Plano de Gestão.

8 – Criar um banco de dados dirigido às questões/temáticas cobertas pelas TRs, onde seja permitido visualizar os dados.

4. METAS E JUSTIFICATIVAS

Meta 1 – Definir as unidades amostrais e os tipos de dados coletados nos diferentes monitoramentos

É necessário realizar um workshop com a equipe técnica do Plano de Gestão de Dados e com os coordenadores da etapa 2 dos programas de monitoramentos, para conhecer as pesquisas que estão sendo realizadas.

Meta 2 – Repositório de dados e metadados definido

Baseado no conhecimento obtido sobre os monitoramentos durante o workshop, definir o tipo de repositório de dados que melhor se adeque aos dados dos monitoramentos.

Meta 3 – Capacitação no armazenamento de dados e metadados

Para se ter um bom repositório de dados é preciso capacitar alunos para trabalhar nos repositórios de dados, recebendo e checando os dados e metadados, dialogando com os geradores dos dados para corrigir os erros e disponibilizando os dados e metadados on line com qualidade.

Meta 4 – Portal de acesso aos dados e metadados

É preciso criar um portal de acesso aos dados e metadados onde pesquisadores, tomadores de decisão e outros usuários tenham acesso as informações geradas.

Meta 5 – Plano de Gestão dos Dados

Os monitoramentos que serão realizados irão gerar uma enorme quantidade de dados fundamentais para subsidiar a elaboração e a implementação de medidas para a recuperação e conservação da fauna aquática. Tais dados deverão ser disponibilizados e para tanto é fundamental promover um plano de gerenciamento das informações sobre os dados coletados, seus acessos, usos e disseminação.

Meta 6 – Banco de Dados

É preciso desenvolver ou definir um banco de dados dirigido às questões do TR 4 de forma a permitir a visualização de resultados do monitoramento.

5. CRONOGRAMA DE EXECUÇÃO

(*Preencher arquivo em Excel enviado*)

6. PRODUTOS

6.1. DADOS BRUTOS *(Listar os dados que serão entregues como produto, segundo o TR4, tanto os coletados in situ como os processados no laboratório)*	**RESPONSÁVEL** *(Pessoal Vinculado)*
Repositório de dados e metadados Banco de dados Plano de Gestão de Dados	Helena Bergallo, David Valentim, William Magnusson, Marcelo Segatto, Paulo Marques

6.2. ANÁLISE DE DADOS *(Relacionar as análises que serão feitas e entregues até o 15º mês de vigência do Projeto)*	**RESPONSÁVEL** *(Pessoal Vinculado)*
Os dados estarão disponíveis no repositório de dados e as sínteses poderão ser visualizadas no banco de dados. A disponibilização dos dados por parte dos pesquisadores permitirá análises mais robustas sobre o impacto causado pelo desastre de Mariana.	Helena Bergallo, David Valentim, William Magnusson, Marcelo Segatto, Paulo Marques

7. METODOLOGIA

Toda gestão de dados pode ser representada pelas etapas do ciclo de dados (Figura 1).

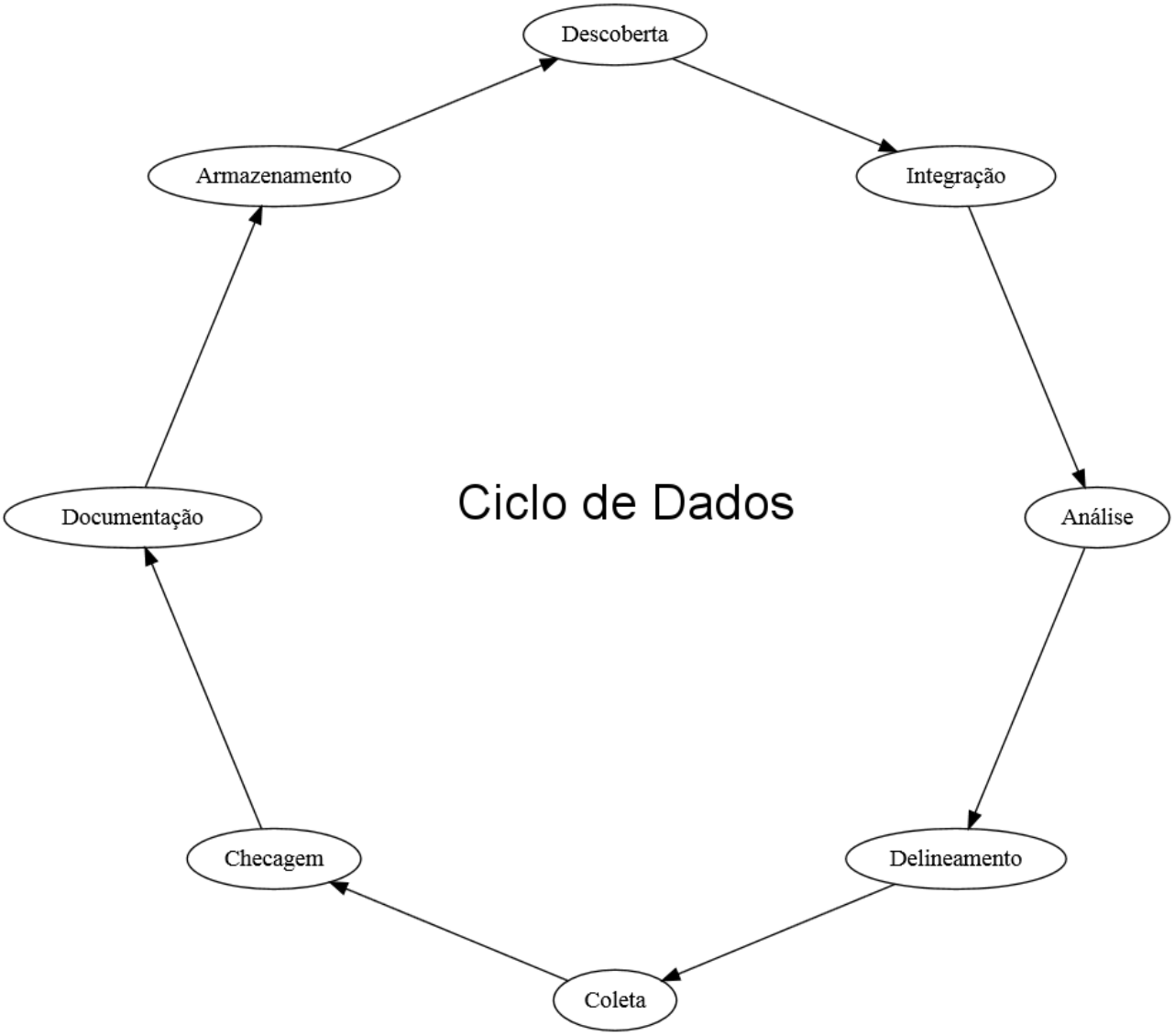

Figura 1 - Ciclo de dados adaptado do DataONE

Com o objetivo de ganhar ciência dos dados pretéritos existentes e da continuação do monitoramento é proposto um workshop para apresentação dos dados. Cada conjunto de dados deve ser apresentado e informações como unidade amostral, existência de ocorrências e variáveis ambientais deve ser registrada. Essas informações são vitais para a padronização dos códigos amostrais de coleta para fazer a relação de dados coletados por outros métodos e escolha da melhor forma de armazenamento.

As possibilidades de repositórios podem ser divididas pelo software, padrão de dados e metadados e local de hospedagem. A tabela abaixo lista as opções independente da natureza dos dados.

Software	Padrão de metadados	Padrão de dados	Hospedagem
IPT	EML	Darwin Core	SiBBr, on-premises
AttaPublica	EML modificado	Livre	SiBBr
Metacat	EML	Livre	on-premises
MAArE	Próprio	Livre	SKYMarket, on-premises

Como a maioria dos dados deve ser oriunda de monitoramento existe um grande favorecimento das opções com padrão de dados livre, mas é importante lembrar que será necessário compartilhar os dados com o ICMBio e o mesmo já possui sistemas legados com suporte a Darwin Core.

A documentação dos dados depende do padrão adotado, felizmente existe uma atual convergência para a adoção do EML (Ecological Metadata Language). É necessária uma capacitação dos coletores dos dados para que a documentação seja a mais fiel possível. A maioria das ferramentas já possui alguma forma de apostila básica, contudo para capacitação completa é recomendado curso presencial para minimizar desistências no meio do caminho e ajuda com suporte em tempo real.

Para descoberta dos dados e metadados os softwares já contam com alguma forma básica da mesma. Nenhuma das opções acima possui suporte a exibição de dados contendo variáveis ambientais. Para construção da mesma é preciso adotar uma solução de *Business Intelligence* (BI) conectada a um banco de dados ou algo mais manual como Shiny da RStudio (Figura 2). O ICMBio por exemplo emprega o uso BI através do painel dinâmico (Figura 3).

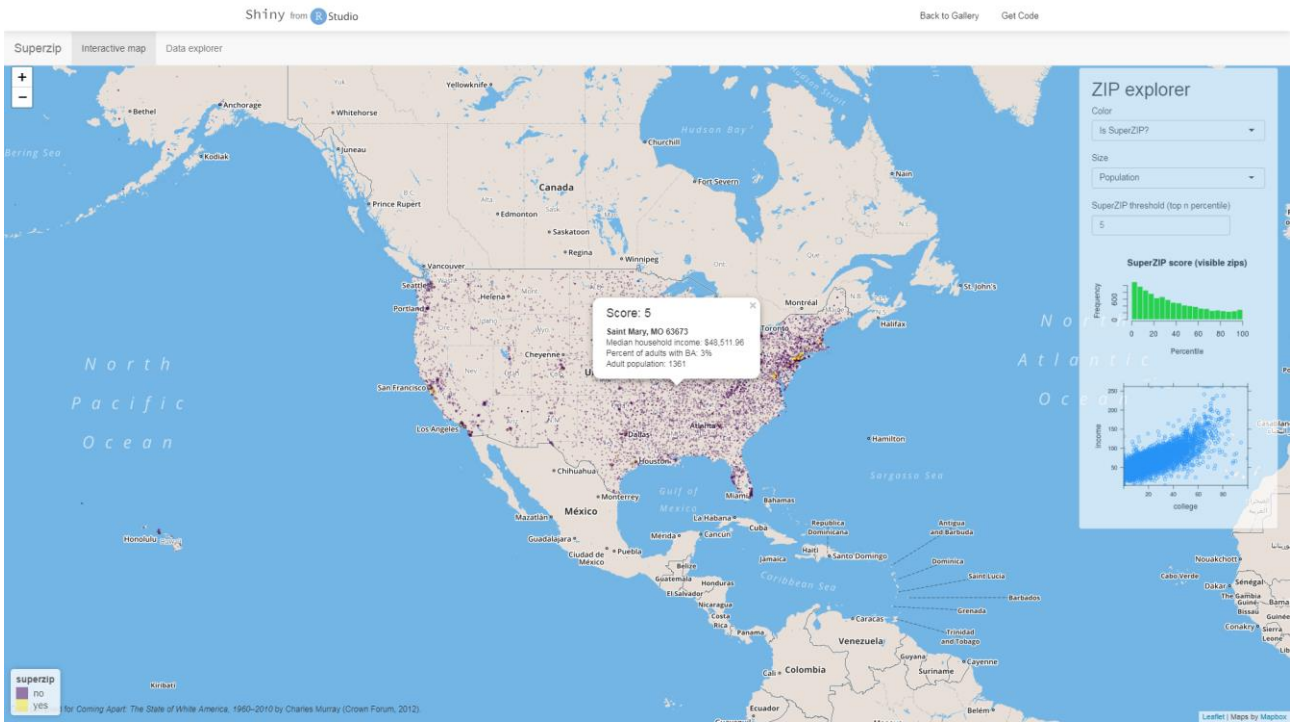

Figura 2 - Exibição de dados usando Shiny da RStudio

Figura 3 - Painel Dinâmico do ICMBio usando ferramenta de BI

A simples exibição dos pontos de coletas em um mapa é um desafio sem a existência de um banco de dados. Projetos nacionais como SiBBr e CRIA criam um banco de dados contendo uma porção reduzida dos dados originais apenas para permitir pesquisa e visualização dos dados. Cada conjunto de dados depositado no repositório devera passar por um processo de ETL (Extract, Transform, Load) para o banco de dados. Esta etapa visa a criação de uma base normalizada e outra desnormalizada na forma de cubo de dados. Cada dimensão do cubo representa o eixo de uma pergunta que se deseja fazer aos dados. Os produtos do ETL são uma base que permite visualização e análise de dados de forma padronizada pela maioria das ferramentas disponíveis, tanto de BI como de análise estatística como o R.