Práticas e ferramentas de gestão de dados

Workflows para a “long-tail” da investigação

João Aguiar Castro, João Rocha da Silva, Ricardo Carvalho Amorim, João Correia Lopes, Gabriel David, Cristina Ribeiro

Faculdade de Engenharia da Universidade do Porto/ INESC TEC

Conteúdo

- Gestão de dados de investigação (RDM)
- RDM como parte do processo de investigação
- Limitações dos repositórios atuais
- O projeto TAIL
- Conclusões

Gestão de dados de investigação (RDM)

Stakeholders

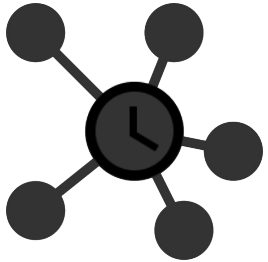
- Quem são os envolvidos no processo de RDM?
  - Instituições
  - Curadores
  - Investigadores
  - Entidades financiadoras

Instituições

- Cumprir com **obrigações legais**
- Contribuir para **políticas de acesso aberto**
- Implementar **planos de gestão de dados**

Investigadores

- **Partilhar** dados cedo com parceiros
- **Publicar** no final do projeto
- **Citar** dados em publicações e permitir citações
- Garantir a **sobrevivência dos dados** para além dos projetos (com suporte institucional)



Curadores

- **Receber** e **produzir** metadados em formatos **interoperáveis**
- Construir serviços de **disseminação** de dados
- Garantir as melhores condições para a **reutilização**
- Levar a cabo ações de **preservação**

Entidades financiadoras

- Especificar requisitos de gestão de dados para projetos de investigação
- Avaliar projetos e seus Data Management Plans (DMPs)
- Disseminar produção científica

Gestão de dados como parte do processo de investigação

Ecossistema diverso

- Grandes instituições ou comunidades
  - ✔ Infraestruturas de suporte
  - ✔ Bom financiamento
  - ✔ Práticas maduras
- Pequenos grupos
  - ✖ Práticas *ad-hoc*
  - ✖ Recursos insuficientes para curadoria
  - ✖ Necessidade de formação de recursos

Problema

Data Management Plans (DMPs) são requisito H2020.

- *Quais os standards para descrição?*
- *Como partilhar os dados e em que repositórios?*
- *Como garantir acesso aos dados após o projeto?*

Mas...

- Os dados são diferentes das publicações
  - Diversidade de domínios
  - Contextos de produção difíceis de capturar
  - Apenas os investigadores sabem como detalhar os processos de produção de dados

- Investigadores não são curadores
  - Não têm conhecimentos de gestão de dados
  - Mudam de equipa após final dos projetos de investigação
  - Podem não estar disponíveis aquando do depósito de dados

Outras questões

- Como prevenir a perda de dados durante a investigação?
- Como introduzir RDM no dia-a-dia dos investigadores?
  - Ferramentas devem proporcionar benefícios imediatos
- Como enriquecer as descrições dos dados?
  - Incorporar o investigador no processo de descrição

Análise de soluções de repositórios atuais

Convergência em 3 fatores fundamentais

1 Envolver investigadores desde cedo

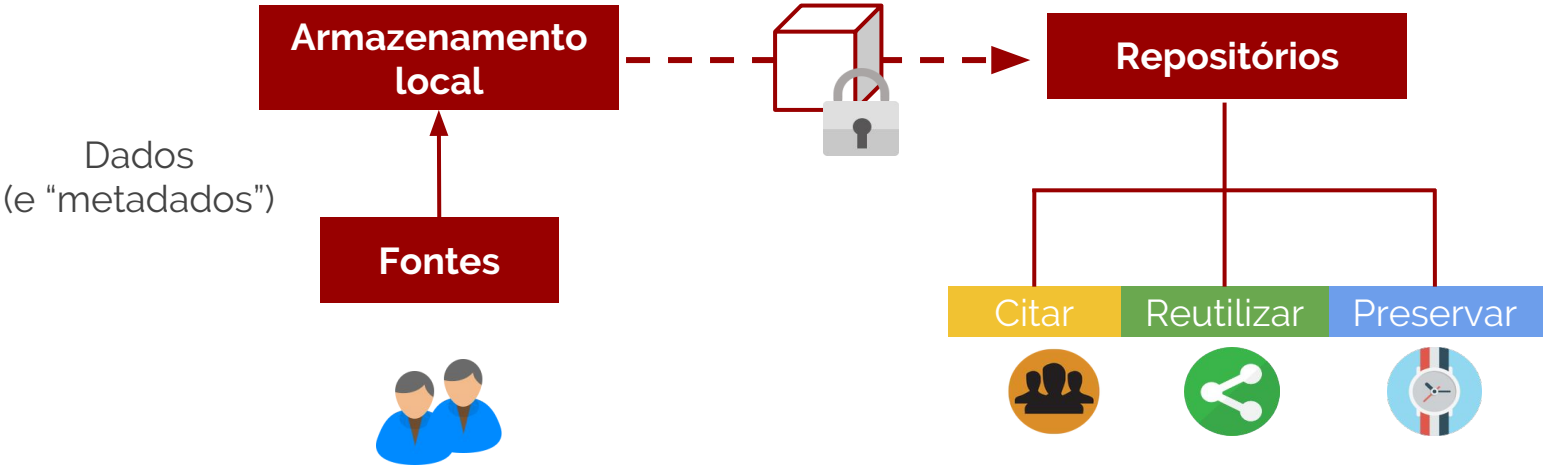
2 Produzir metadados do domínio

3 Promover a visibilidade na comunidade

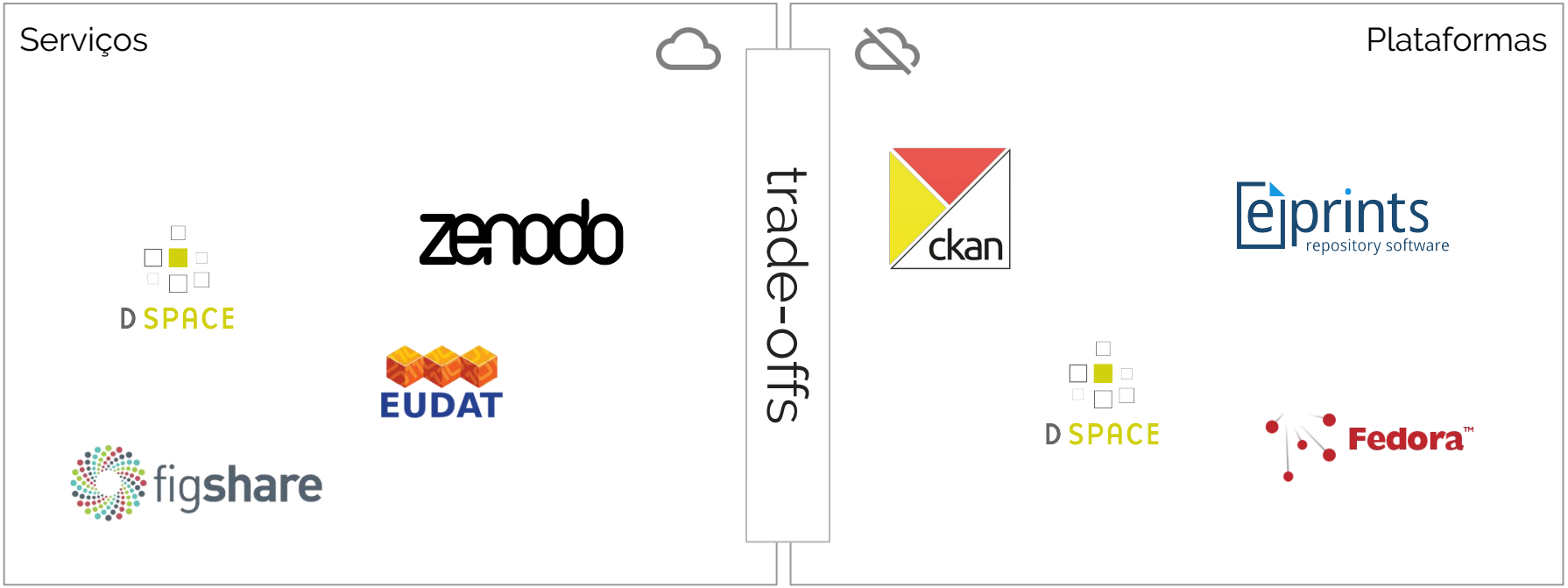
Princípios

- Dados têm de estar associados a metadados de domínio
- É impraticável atribuir a produção de todos os metadados a curadores
- Investigadores já os produzem no seu dia a dia
  - Mas em suportes frágeis e dispersos
- Registos devem facilitar a interoperabilidade

“O” workflow

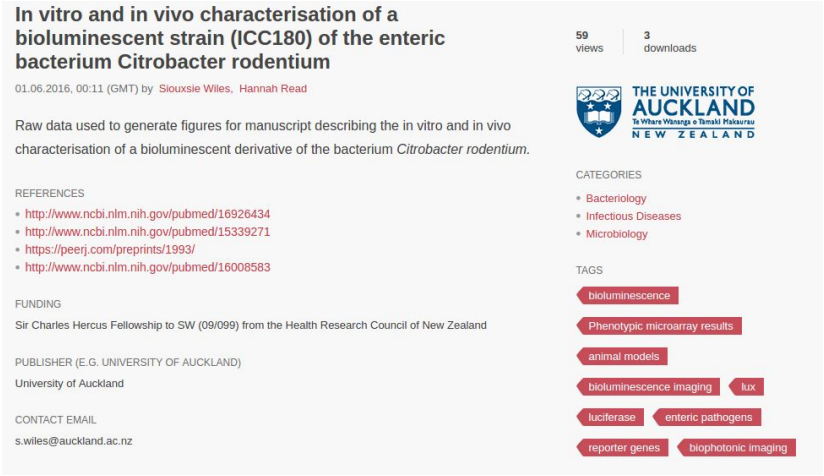

Armazenamento local
Repositórios
Dados
(e “metadados”)
Fontes
Citar
Reutilizar
Preservar

Repositórios de dados de investigação



figshare

- Sistema de estatísticas de citações de dados
- Exportação para sistemas de referências bibliográficas
- Sem custos de manutenção



In vitro and in vivo characterisation of a bioluminescent strain (ICC180) of the enteric bacterium Citrobacter rodentium

59 views | 3 downloads

01.06.2016, 00:11 (GMT) by Siouxsie Wiles, Hannah Read

Raw data used to generate figures for manuscript describing the in vitro and in vivo characterisation of a bioluminescent derivative of the bacterium *Citrobacter rodentium*.

REFERENCES

- http://www.ncbi.nlm.nih.gov/pubmed/16926434
- http://www.ncbi.nlm.nih.gov/pubmed/15339271
- https://peerj.com/preprints/1993/
- http://www.ncbi.nlm.nih.gov/pubmed/16008583

FUNDING

Sir Charles Hercus Fellowship to SW (09/099) from the Health Research Council of New Zealand

PUBLISHER (E.G. UNIVERSITY OF AUCKLAND)

University of Auckland

CONTACT EMAIL

s.wiles@auckland.ac.nz

THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

CATEGORIES

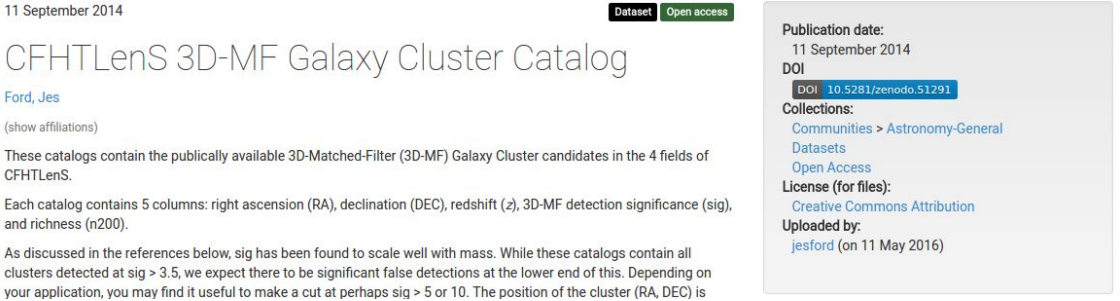
- Bacteriology
- Infectious Diseases
- Microbiology

TAGS

bioluminescence, Phenotypic microarray results, animal models, bioluminescence imaging, lux, luciferase, enteric pathogens, reporter genes, biophotonic imaging

zenodo

- DOIs
- Suporte para comunidades
- Metadados são pesquisáveis



11 September 2014 Dataset Open access

CFHTLenS 3D-MF Galaxy Cluster Catalog

Ford, Jes

(show affiliations)

These catalogs contain the publically available 3D-Matched-Filter (3D-MF) Galaxy Cluster candidates in the 4 fields of CFHTLenS.

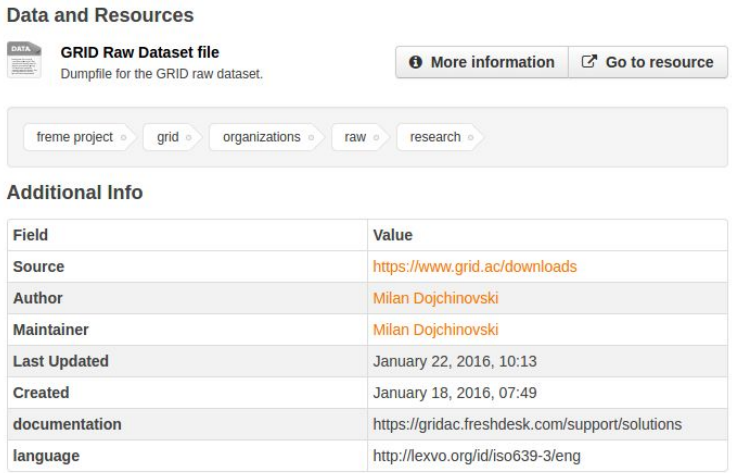
Each catalog contains 5 columns: right ascension (RA), declination (DEC), redshift (*z*), 3D-MF detection significance (sig), and richness (n200).

As discussed in the references below, sig has been found to scale well with mass. While these catalogs contain all clusters detected at sig > 3.5, we expect there to be significant false detections at the lower end of this. Depending on your application, you may find it useful to make a cut at perhaps sig > 5 or 10. The position of the cluster (RA, DEC) is

Publication date: 11 September 2014
DOI DOI 10.5281/zenodo.51291
Collections: Communities > Astronomy-General Datasets, Open Access
License (for files): Creative Commons Attribution
Uploaded by: jesford (on 11 May 2016)

ckan

- Instalável localmente
- Pode ser amplamente personalizado
- Tem amplas referências no domínio governamental



EUDAT

- Módulos de colaboração, armazenamento e processamento de dados
- Apoio de agências europeias

- Módulos que cobrem parcialmente o workflow de RDM:
  - Gestão e comunicação de equipas
  - Plataforma de cloud computing
  - Disseminação de dados em motores de pesquisa
  - Preservação a longo prazo
- Colaboração RDM@FEUP - EUDAT
  - **DataPublication@UPorto**: *multi-disciplinary data description and deposit linking the Dendro staging platform with the EUDAT European Infrastructure*

Repositórios de dados

- São repositórios finais
  - Preservam os dados
  - Garantem a sua publicação
  - Guardam estatísticas sobre acessos e citações
  - **Mas não tentam encorajar a sua descrição tão cedo quanto possível**

Dados podem nem chegar a estas plataformas!

O projeto TAIL

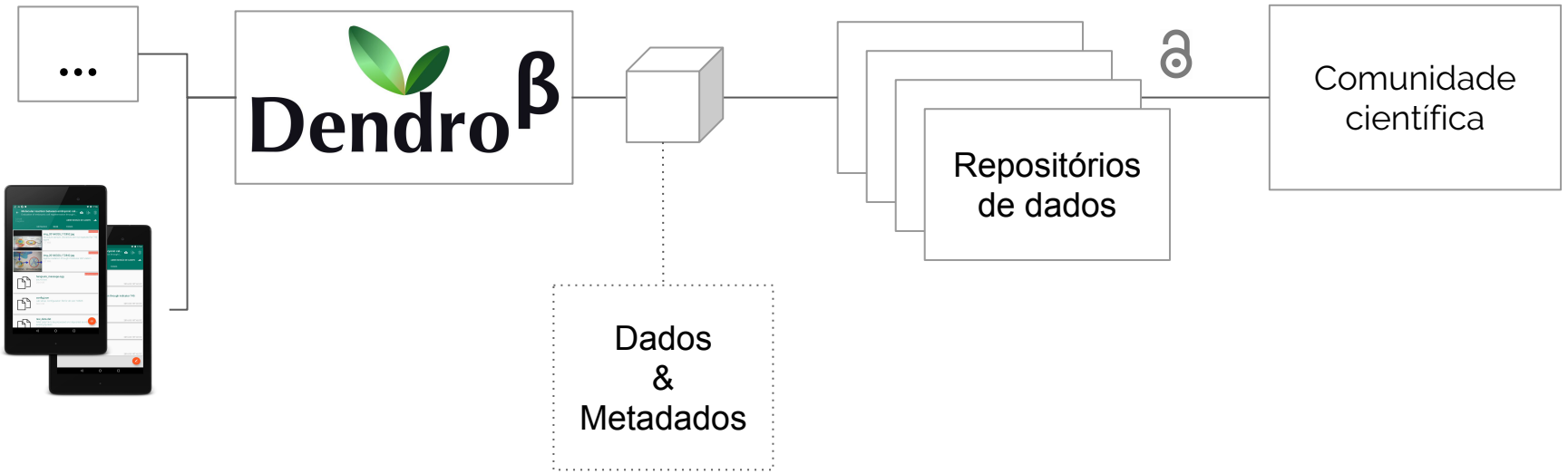
Objetivos

- Capturar dados e metadados **cedo** para **posterior depósito** em qualquer repositório
- Assegurar acesso a dados após inevitável obsolescência
- Suportar o processo de investigação com RDM desde o início
- Promover recolha de metadados específicos do domínio

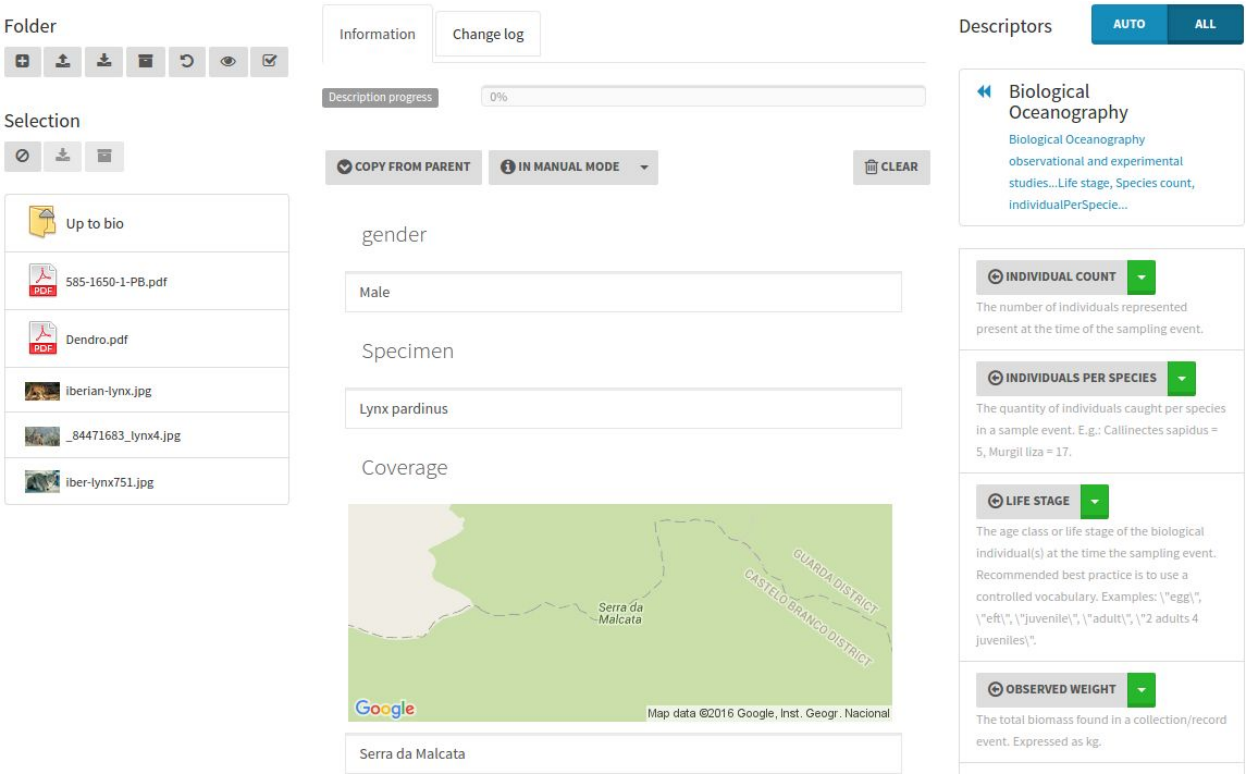
TAIL @ FEUP


Compatibilidade com protocolos de disseminação
Sob controlo da instituição
Metadados de domínio
Avaliado junto dos investigadores
Baseado na web semântica
Dendro β
Gestão colaborativa
Exportação para qualquer repositório
Cliente móvel (caderno de laboratório eletrónico)
Gestão de permissões
Armazenamento local e privado

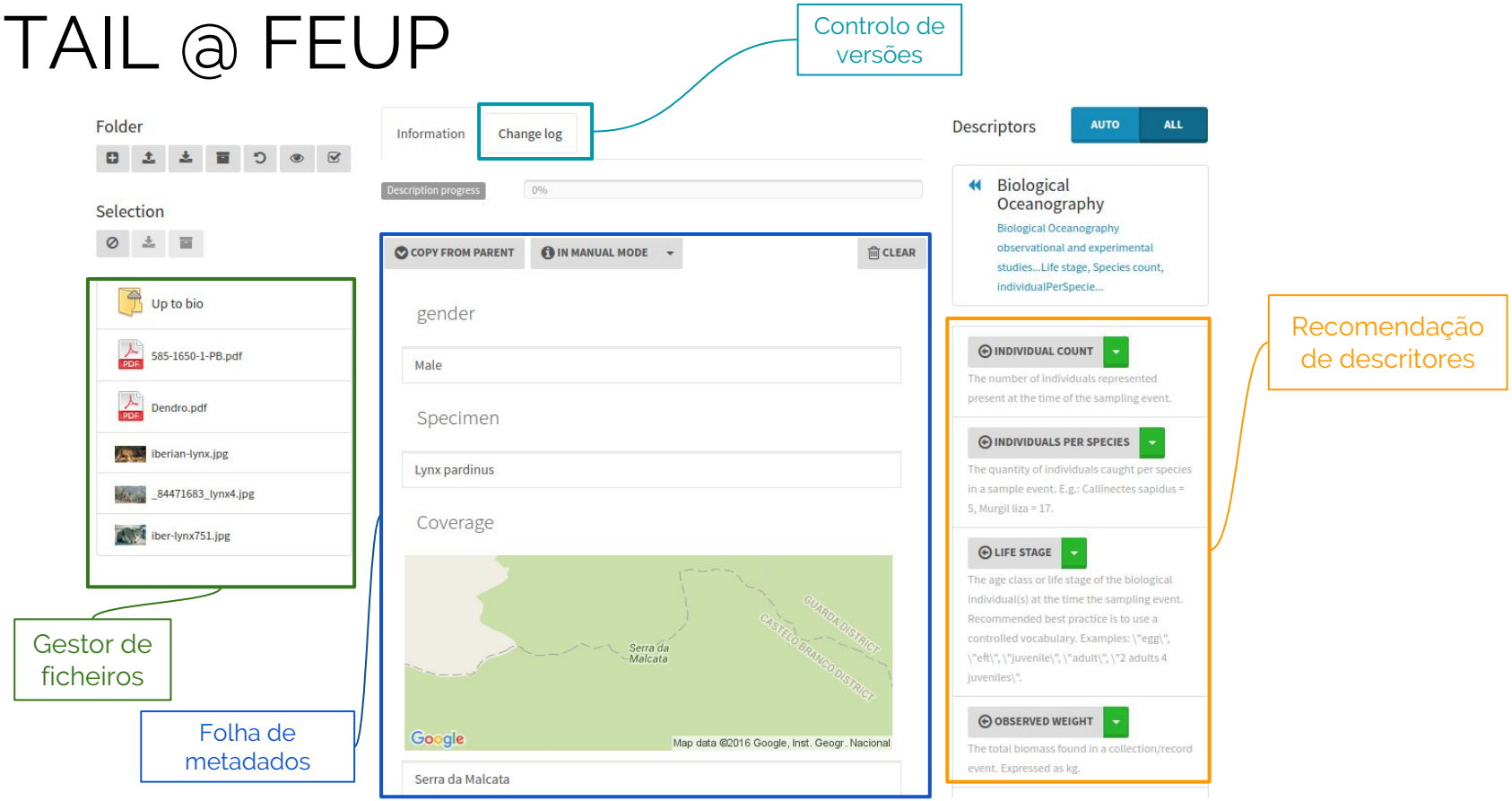
TAIL @ FEUP


...
Dendro β
Repositórios
de dados
Comunidade
científica
Dados
&
Metadados

TAIL @ FEUP

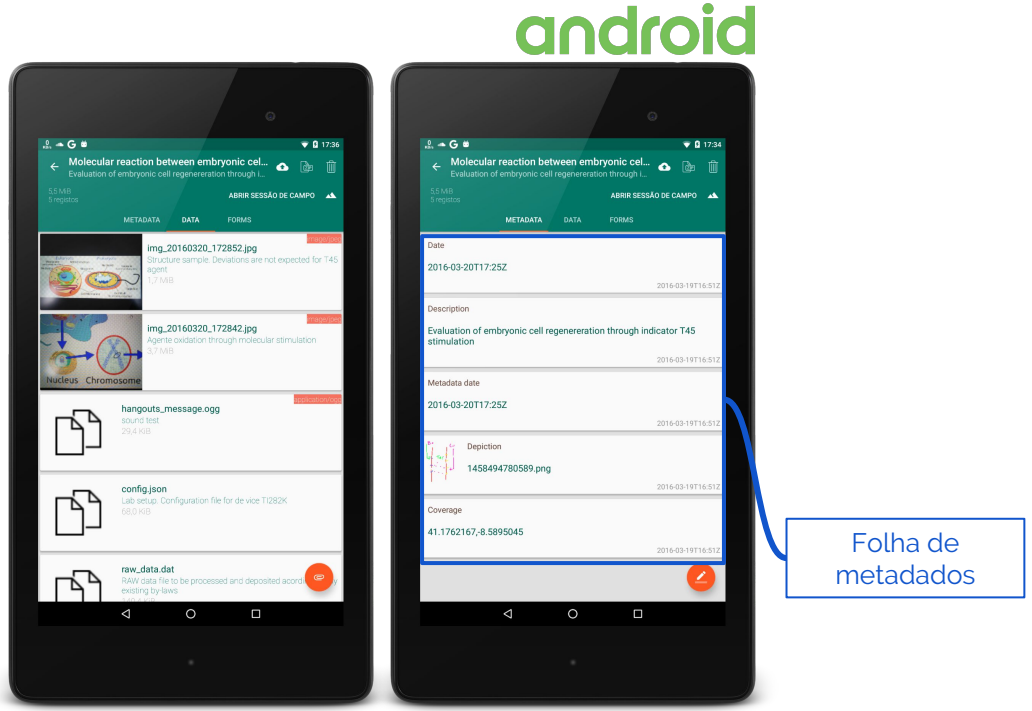

Folder
Selection
Up to bio
585-1650-1-PB.pdf
Dendro.pdf
iberian-lynx.jpg
_84471683_lynx4.jpg
iber-lynx751.jpg
Information
Change log
Description progress
0%
COPY FROM PARENT
IN MANUAL MODE
CLEAR
gender
Male
Specimen
Lynx pardinus
Coverage
Serra da Malcata
GUARDA DISTRICT
CASTELO BRANCO DISTRICT
Google
Map data ©2016 Google, Inst. Geogr. Nacional
Serra da Malcata
Descriptors
AUTO
ALL
Biological Oceanography
Biological Oceanography observational and experimental studies...Life stage, Species count, individualPerSpecie...
INDIVIDUAL COUNT
The number of individuals represented present at the time of the sampling event.
INDIVIDUALS PER SPECIES
The quantity of individuals caught per species in a sample event. E.g.: Callinectes sapidus = 5, Murgil liza = 17.
LIFE STAGE
The age class or life stage of the biological individual(s) at the time the sampling event. Recommended best practice is to use a controlled vocabulary. Examples: \"egg\", \"eft\", \"juvenile\", \"adult\", \"2 adults 4 juveniles\".
OBSERVED WEIGHT
The total biomass found in a collection/record event. Expressed as kg.

TAIL @ FEUP


Controlo de versões
Folder
Selection
Information
Change log
Description progress
0%
Descriptors
AUTO
ALL
Biological Oceanography
Biological Oceanography observational and experimental studies...Life stage, Species count, individualPerSpecie...
Up to bio
585-1650-1-PB.pdf
Dendro.pdf
iberian-lynx.jpg
_84471683_lynx4.jpg
iber-lynx751.jpg
COPY FROM PARENT
IN MANUAL MODE
CLEAR
gender
Male
Specimen
Lynx pardinus
Coverage
Serra da Malcata
GUARDA DISTRICT
CASTELO BRANCO DISTRICT
Google
Map data ©2016 Google, Inst. Geogr. Nacional
Serra da Malcata
INDIVIDUAL COUNT
The number of individuals represented present at the time of the sampling event.
INDIVIDUALS PER SPECIES
The quantity of individuals caught per species in a sample event. E.g.: Callinectes sapidus = 5, Murgil liza = 17.
LIFE STAGE
The age class or life stage of the biological individual(s) at the time the sampling event. Recommended best practice is to use a controlled vocabulary. Examples: \"egg\", \"eft\", \"juvenile\", \"adult\", \"2 adults 4 juveniles\".
OBSERVED WEIGHT
The total biomass found in a collection/record event. Expressed as kg.
Recomendação de descritores
Gestor de ficheiros
Folha de metadados

TAIL @ FEUP - LabTablet

- Caderno de laboratório eletrónico
- Suporte para metadados de domínio
- Modo de campo para acompanhar investigadores
  - Fácil de usar
  - Registo de percurso feito ou cobertura geográfica
  - Aproveitamento dos sensores disponíveis no equipamento (temperatura, ...)



Conclusões

Conclusões (1/3)

- Dados de investigação são diferentes das publicações
  - Políticas atuais são herdadas das publicações
- RDM deve ser introduzida no processo de investigação
  - Repositórios surgem demasiado tarde
  - Investigadores devem participar na descrição dos seus dados
- Necessário prever obsolescência
  - Soluções devem assegurar sobrevivência dos dados

Conclusões (2/3)

- Dendro: uma solução para gestão de dados dentro dos grupos de investigação
  - Metadados específicos do domínio
  - Publicação dos dados em múltiplos repositórios

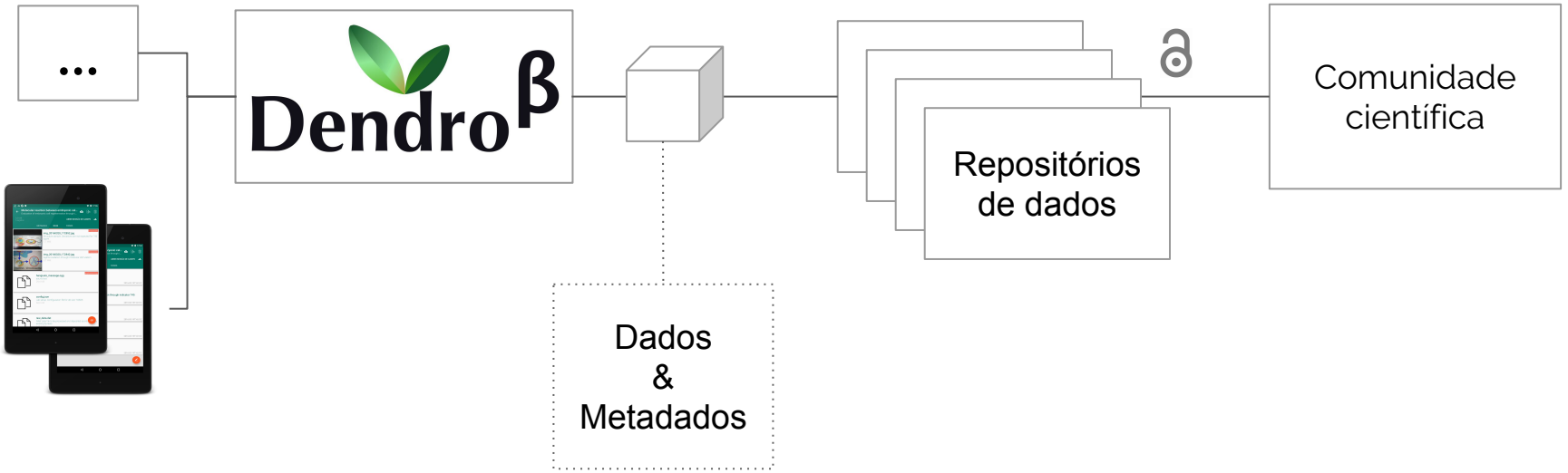
- LabTablet: um caderno de laboratório eletrónico
  - Suporta investigadores nos laboratórios e saídas de campo
  - Integrado com o Dendro para depósito de metadados e dados de ocasião

Conclusões (3/3)

- Projeto TAIL: acompanhamento dos investigadores desde a produção de dados até ao seu depósito
  - 10 a 50 grupos de investigação
  - Workflow integrado de ferramentas

- Resultados previstos
  - Envolvimento dos investigadores na descrição de dados
  - Conjuntos de dados depositados nos principais repositórios
  - Satisfação dos requisitos de DMPs dos grupos de investigação

TAIL @ FEUP



{joaoaguiarcastro, joaorosilva}@gmail.com
{rcamorim, mcr}@fe.up.pt