Repositório de dados na U.Porto

Um fluxo de curadoria suportado numa extensão ao DSpace

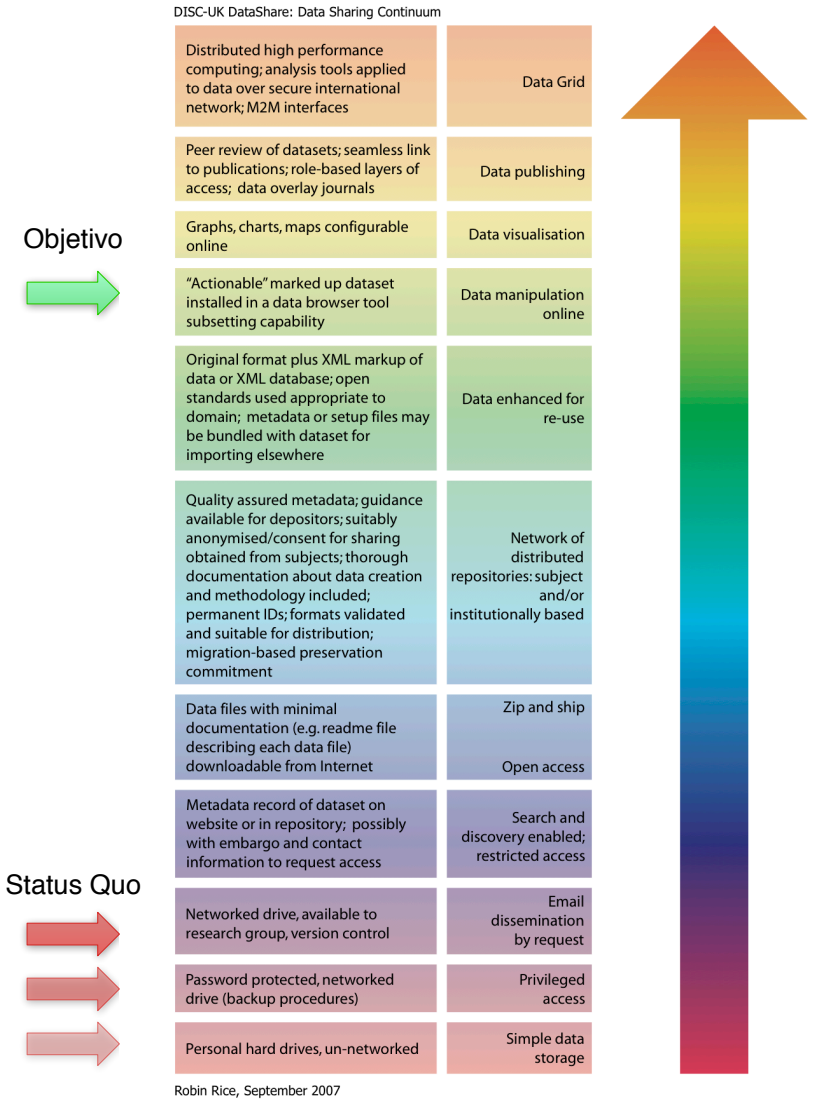
Cristina Ribeiro	DEI- FEUP/ INESC TEC
João Rocha da Silva	FEUP
Eugénia Matos Fernandes	Reitoria da Universidade do Porto
João Correia Lopes	DEI- FEUP/ INESC TEC

Conteúdo

- Objetivos
- Auditoria de dados na U.Porto
- Desenho de um workflow de gestão de dados
- Construção de um protótipo de repositório
- Conclusões e trabalhos futuros

Objetivos

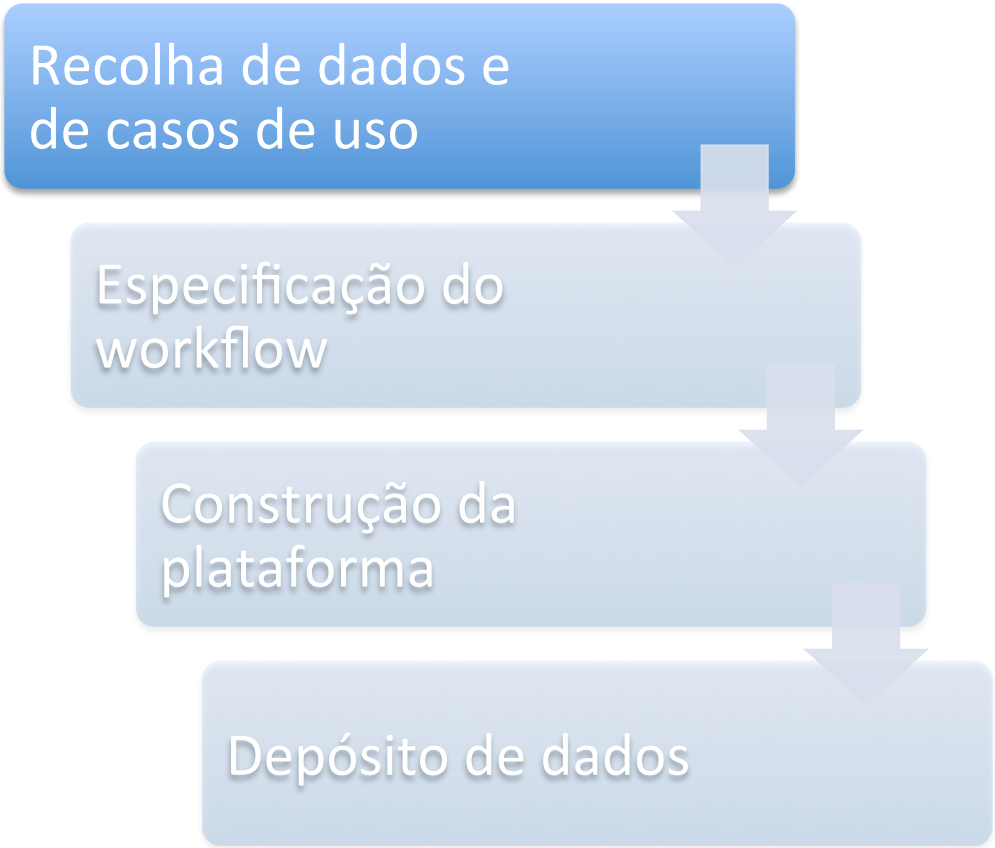
- Determinar as necessidades de gestão de dados dos investigadores da U.Porto
- Desenhar e implementar um repositório de dados para satisfazer estas necessidades
- Procurar uma solução que não esteja limitada às necessidades de um só grupo



Satisfazer objetivos

- Representar dados tabulares em formatos próprios para preservação
  - XML
- Interrogação online
- Reutilização de descritores nos metadados

Fase 1 : Entrevistas



Auditoria de dados

- Entrevistas com investigadores
  - Engenharia, ciências sociais, educação, ciências da terra, biologia, economia, …
- Recolha de amostras de dados
- Recolha de casos de uso
- Relatório e resultados submetidos a aprovação dos investigadores

Os investigadores dizem

- … a gestão de dados é complexa
- … a gestão de dados não deveria distraí-los do seu trabalho
- … precisam de apoio profissional na gestão de dados
- … “o que ganho em guardar os meus dados num repositório? Os discos externos são tão baratos!”

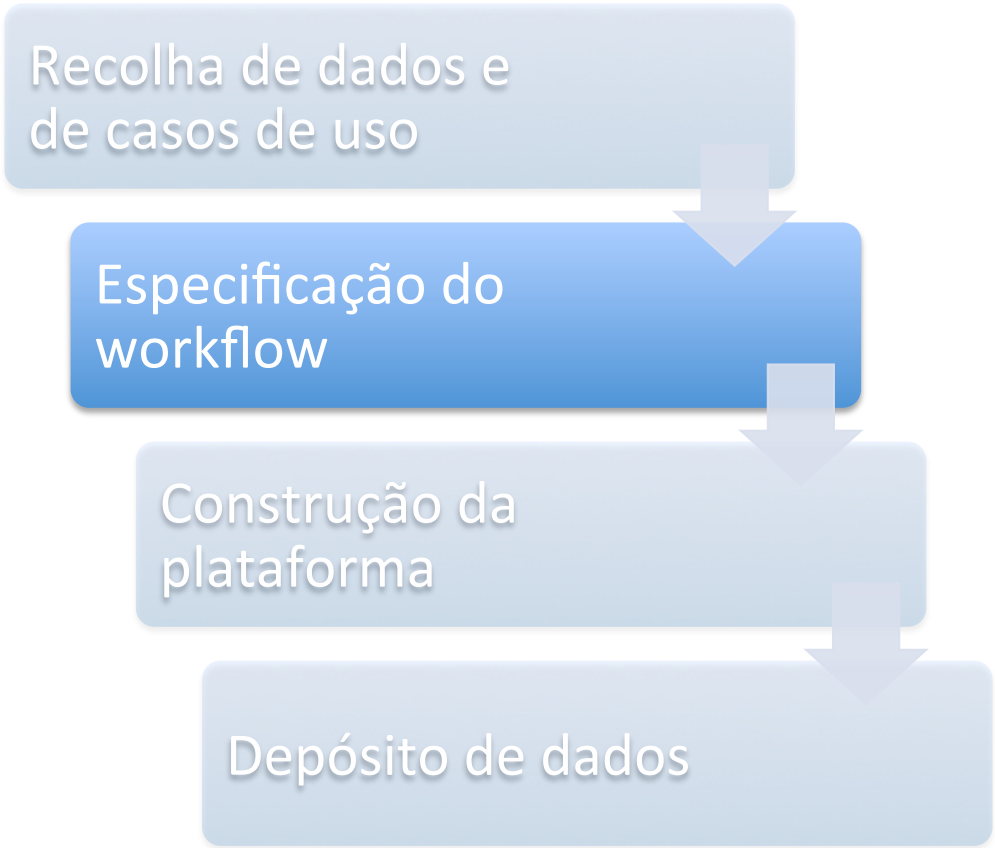
Os investigadores dizem

- … já perderam dados devido a formatos que foram abandonados
- … precisam de sítio para partilhar dados com parceiros, em vez de usarem o email
- … precisam de ferramentas para manipulação de dados online

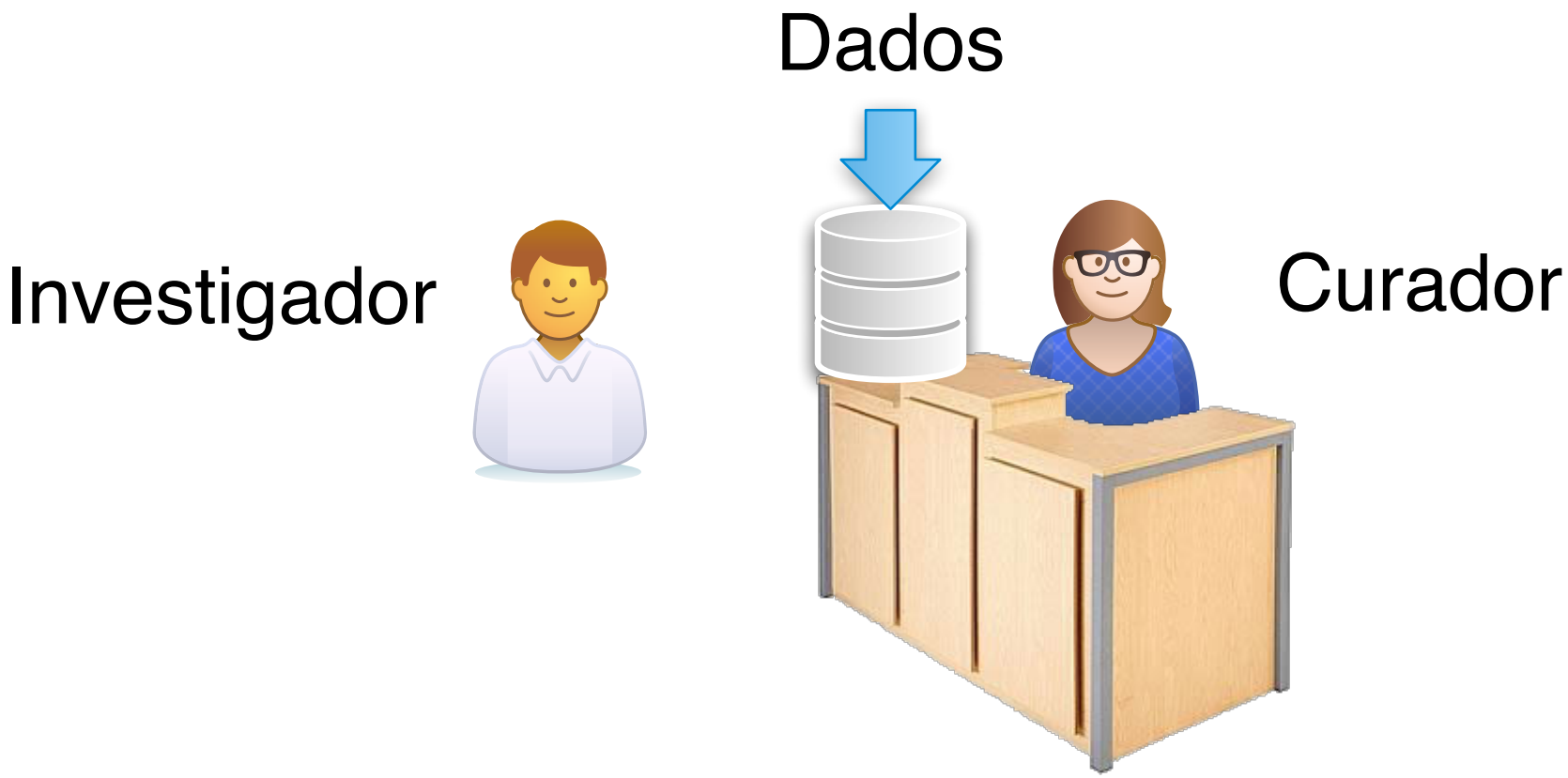
Preocupações e necessidade dos investigadores

- "Os repositórios não podem ser *cemitérios* de dados"
- "O principal objetivo na preservação de dados é a partilha/ reutilização/ citação"
- "Os dados têm de ser bem anotados ou não podem ser usados para validar resultados

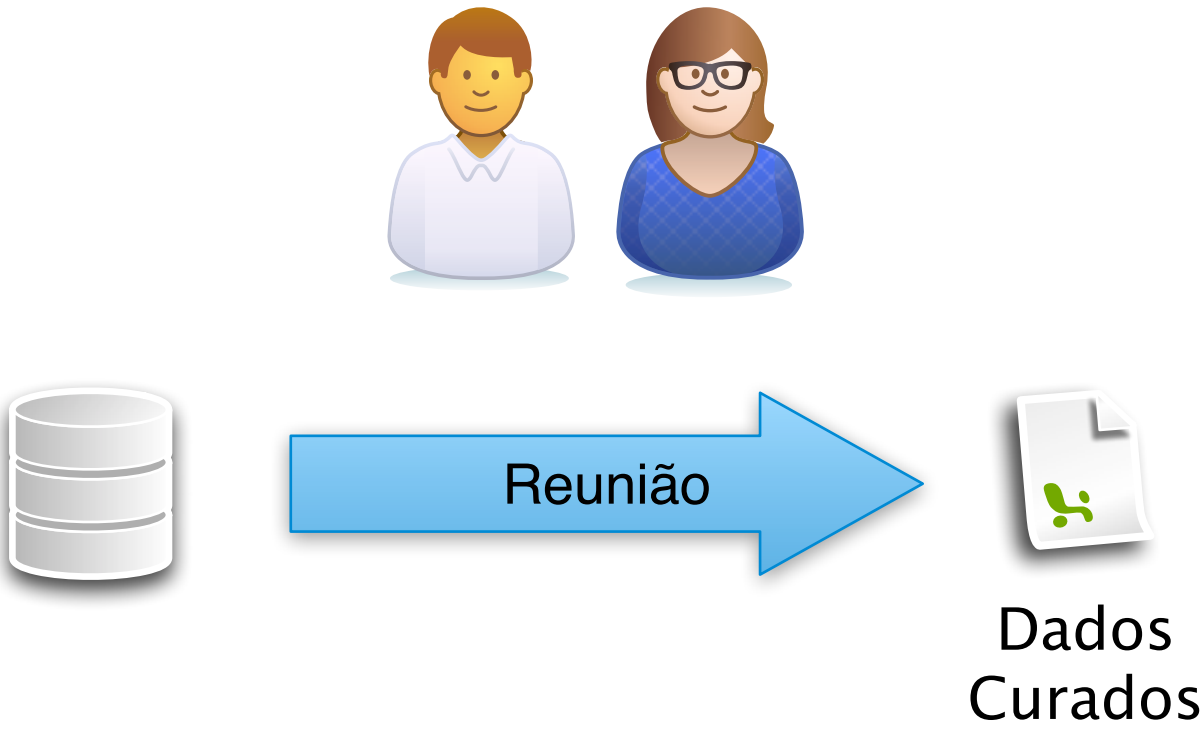
Fase 2 : Modificar o workflow



O Papel do “Curador de Dados”



Reunião de curadoria



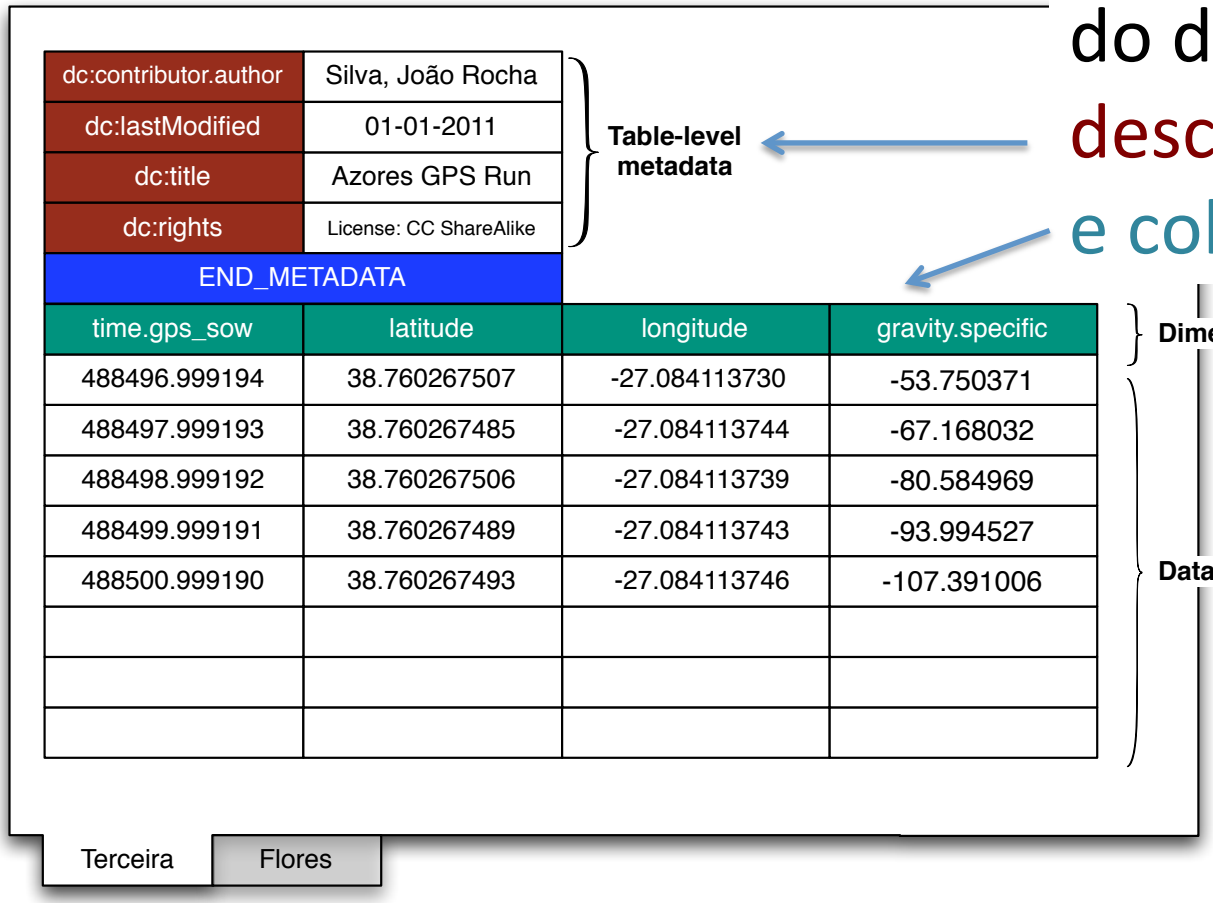
Anotação de dados

Elementos do XML Schema

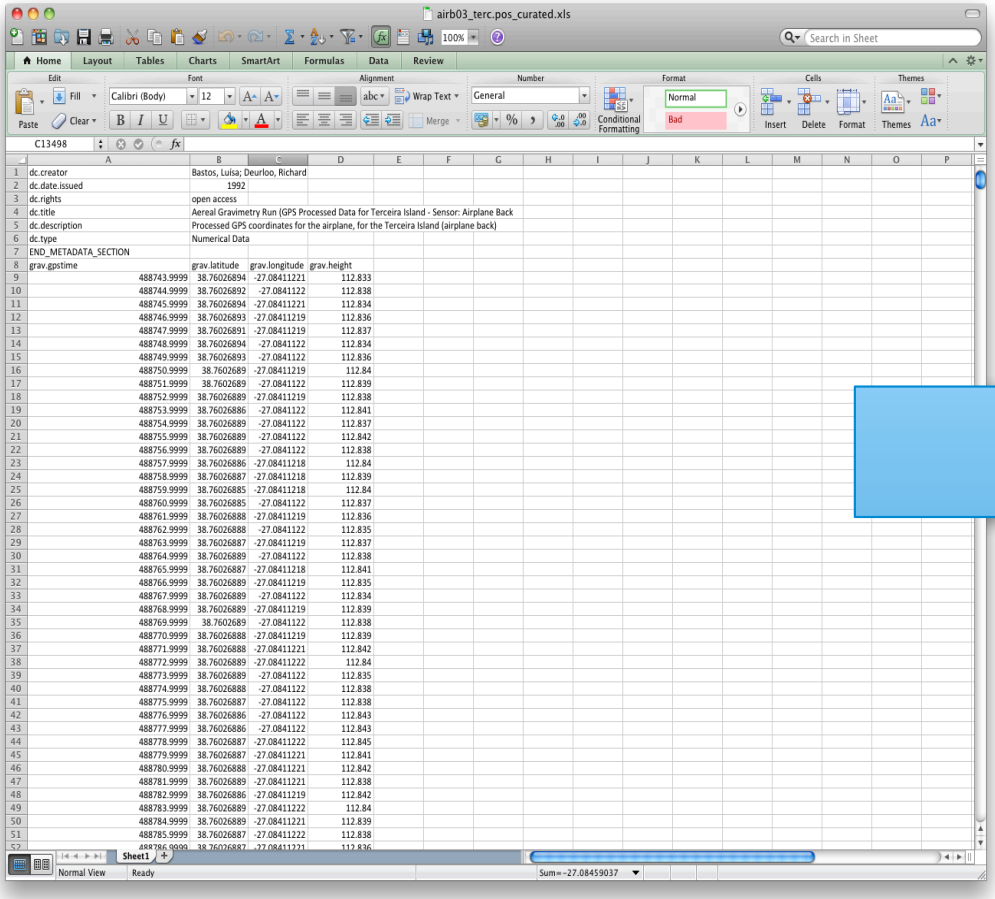
do domínio como

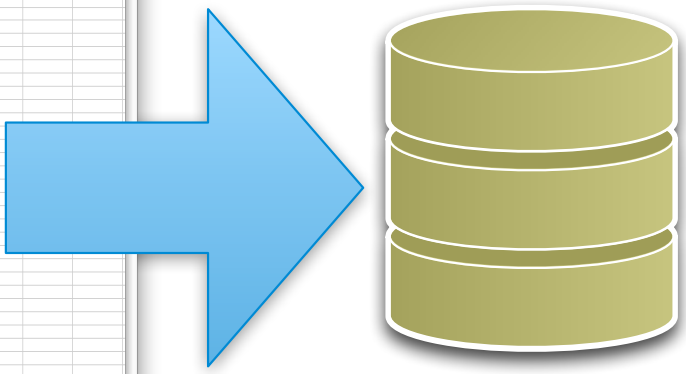
descritores

e colunas



Depois da reunião



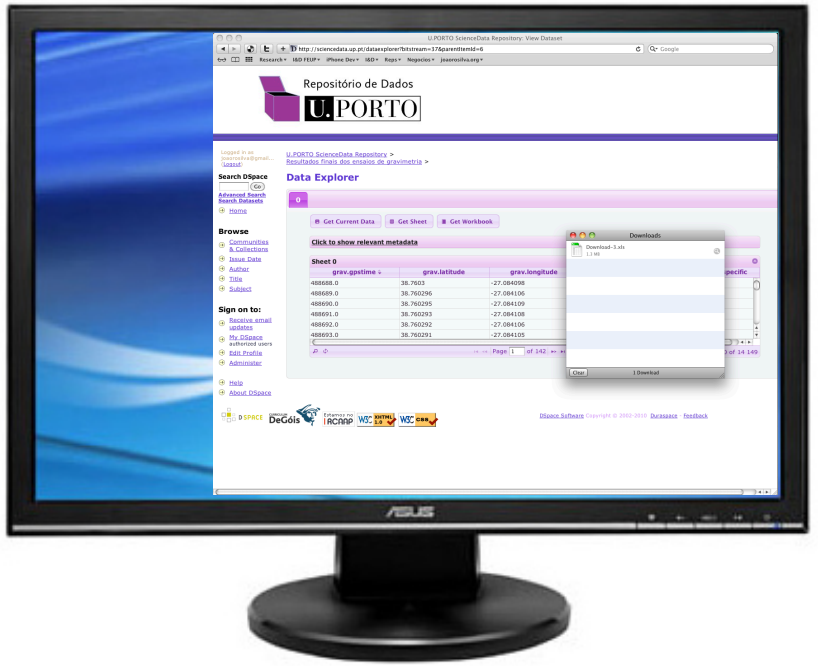


Repositório

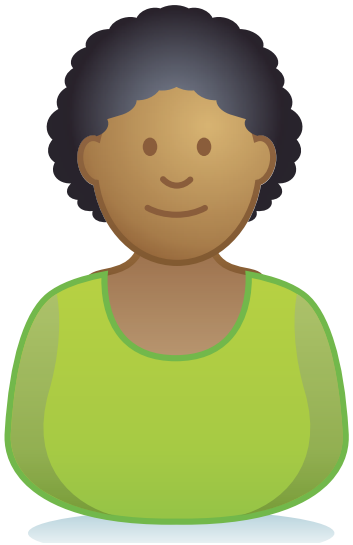
Dados+Metadados em formato Excel

Dados disponíveis

Repositório de dados de investigação

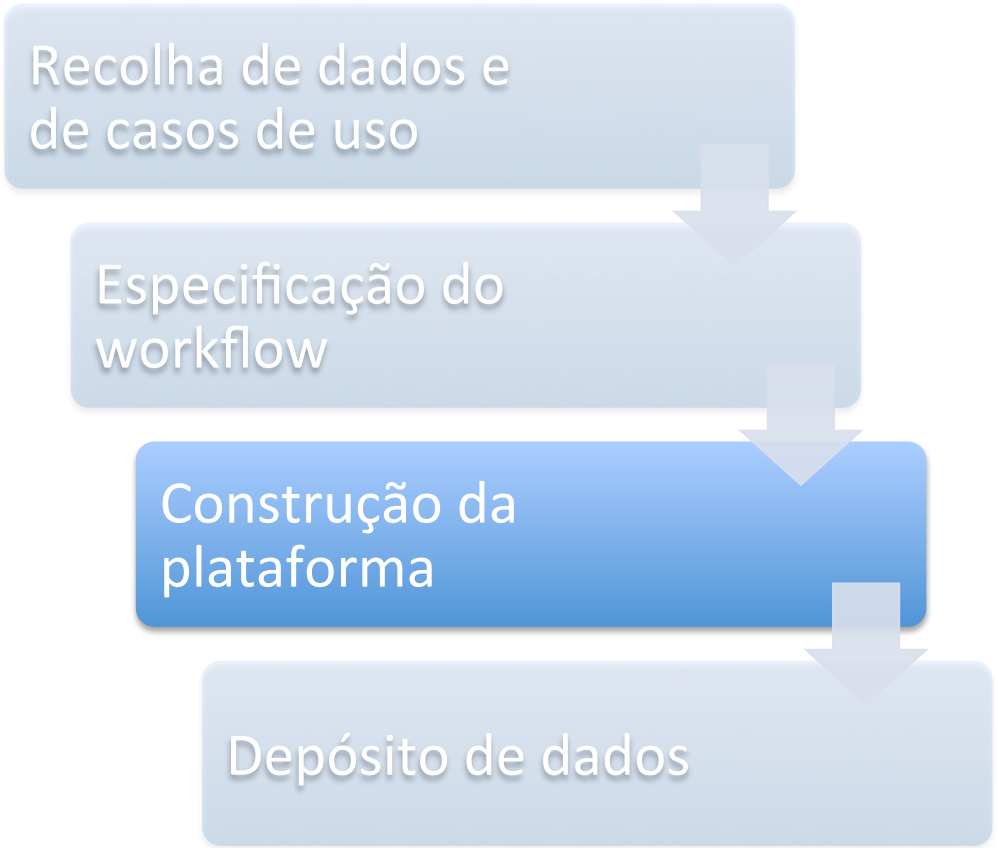


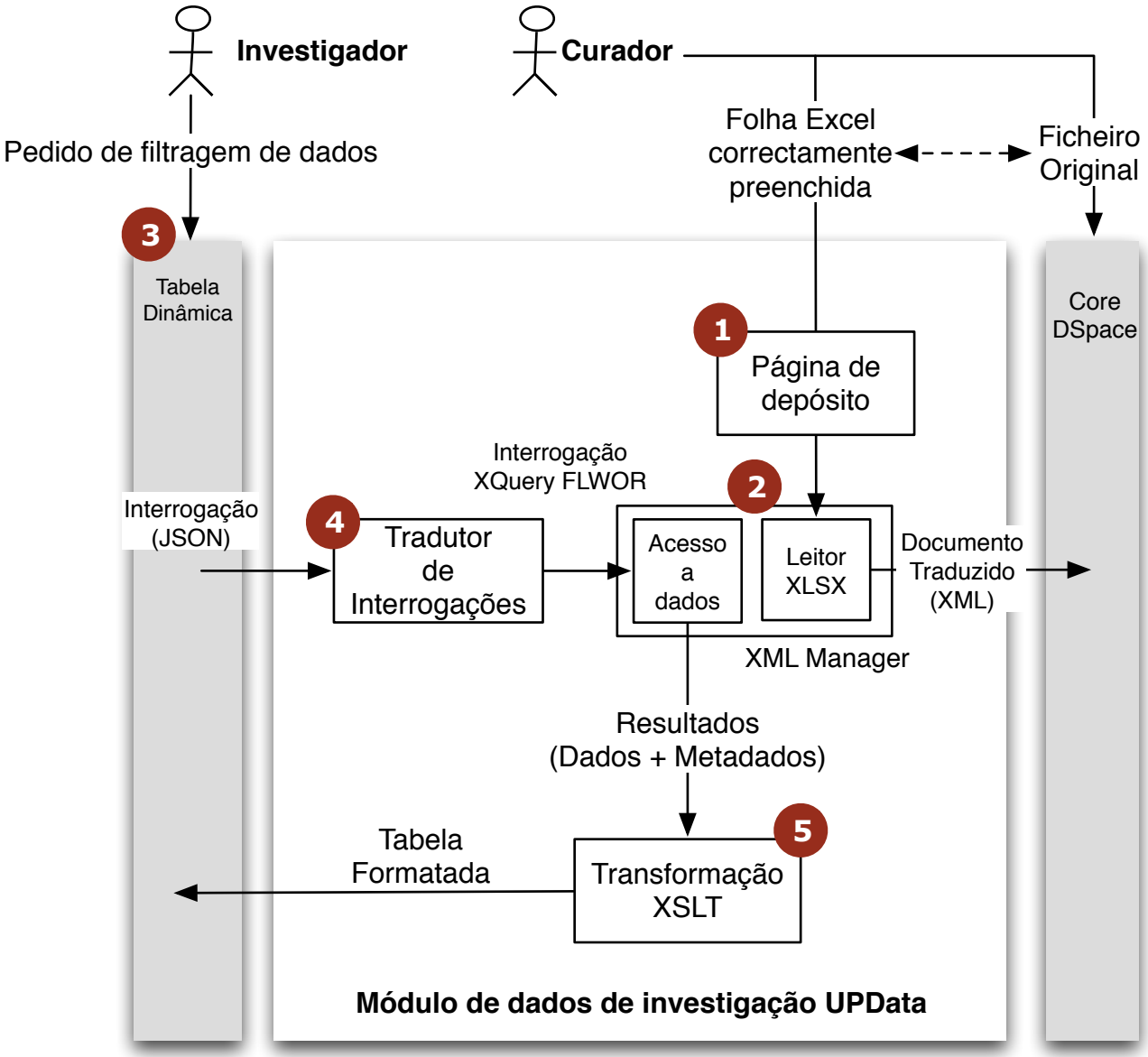
Investigador



- Explorar, filtrar e descarregar só o necessário

Fase 3 : Construir ferramentas de suporte ao workflow




Investigador
Curador
Pedido de filtragem de dados
Folha Excel correctamente preenchida
Ficheiro Original
3
Tabela Dinâmica
Core DSpace
1
Página de depósito
Interrogação XQuery FLWOR
2
Interrogação (JSON)
4
Tradutor de Interrogações
Acesso a dados
Leitor XLSX
Documento Traduzido (XML)
XML Manager
Resultados (Dados + Metadados)
5
Tabela Formatada
Transformação XSLT
Módulo de dados de investigação UPData

```xml
<?xml version="1.0"?>
<record>
    <metadata>
        <dc.creator>Bastos, Lu&amp;amp;#237;sa; Deurloo, Richard</dc.creator>
        <dc.title>Aereal Gravimetry Run (GPS Processed Data for Terceira Island - Beach) Sensor - tail of airplane</dc.title>
        <dc.type>Numerical Data</dc.type>
        <dc.rights>open access</dc.rights>
        <dc.date.issued>1992.0</dc.date.issued>
        <dc.description>Processed GPS coordinates for the airplane, for the Terceira Island (Beach)</dc.description>
    </metadata>
    <headers>
        <header>grav.gpstime</header>
        <header>grav.latitude</header>
        <header>grav.longitude</header>
        <header>grav.height</header>
    </headers>
    <data>
        <rows>
            <row>
                <grav.gpstime>488496.999194</grav.gpstime>
                <grav.latitude>38.760267507</grav.latitude>
                <grav.longitude>-27.08411373</grav.longitude>
                <grav.height>112.989</grav.height>
            </row>
            <row>
                <grav.gpstime>488497.999193</grav.gpstime>
                <grav.latitude>38.760267485</grav.latitude>
                <grav.longitude>-27.084113744</grav.longitude>
                <grav.height>112.995</grav.height>
            </row>
            <row>
                <grav.gpstime>488498.999192</grav.gpstime>
                <grav.latitude>38.760267506</grav.latitude>
                <grav.longitude>-27.084113739</grav.longitude>
                <grav.height>112.992</grav.height>
            </row>
```

Fase 4 : Testar ferramenta com dados reais

Conclusões e Trabalho Futuro

- Recolhemos requisitos e casos de uso dos investigadores da U.Porto
- Casos de uso mais importantes foram implementados em repositório DSpace
- Utilizadores podem navegar sobre dados online no repositório e descarregar subconjuntos selecionados
- Futuro:
  - Validação de ferramentas com utilizadores
  - Métodos mais simples de interação entre investigadores e repositório

Contactos e ligações

Cristina Ribeiro	mcr@fe.up.pt
João Rocha da Silva	joaorosilva@gmail.com
Eugénia Matos Fernandes	efernand@reit.up.pt
João Correia Lopes	jlopes@fe.up.pt

Repositório: http://sciencedata.up.pt/

Documentos: http://sciencedata.up.pt/doc