

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

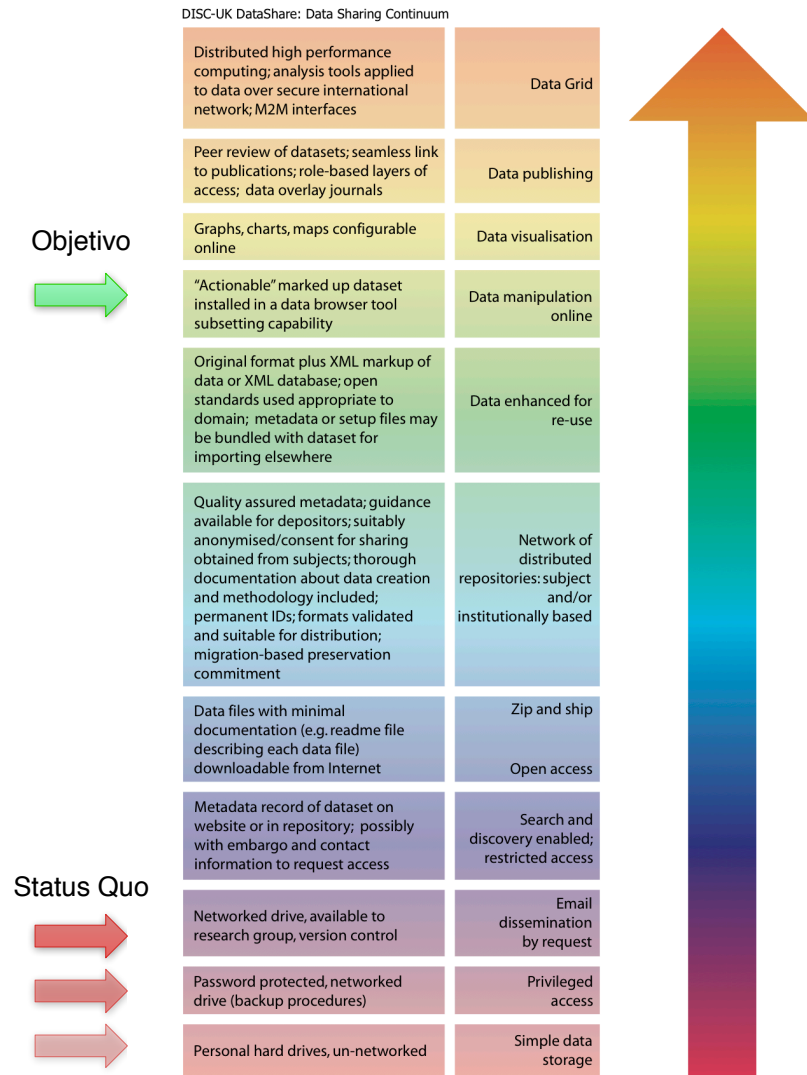
Conteúdo

- 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

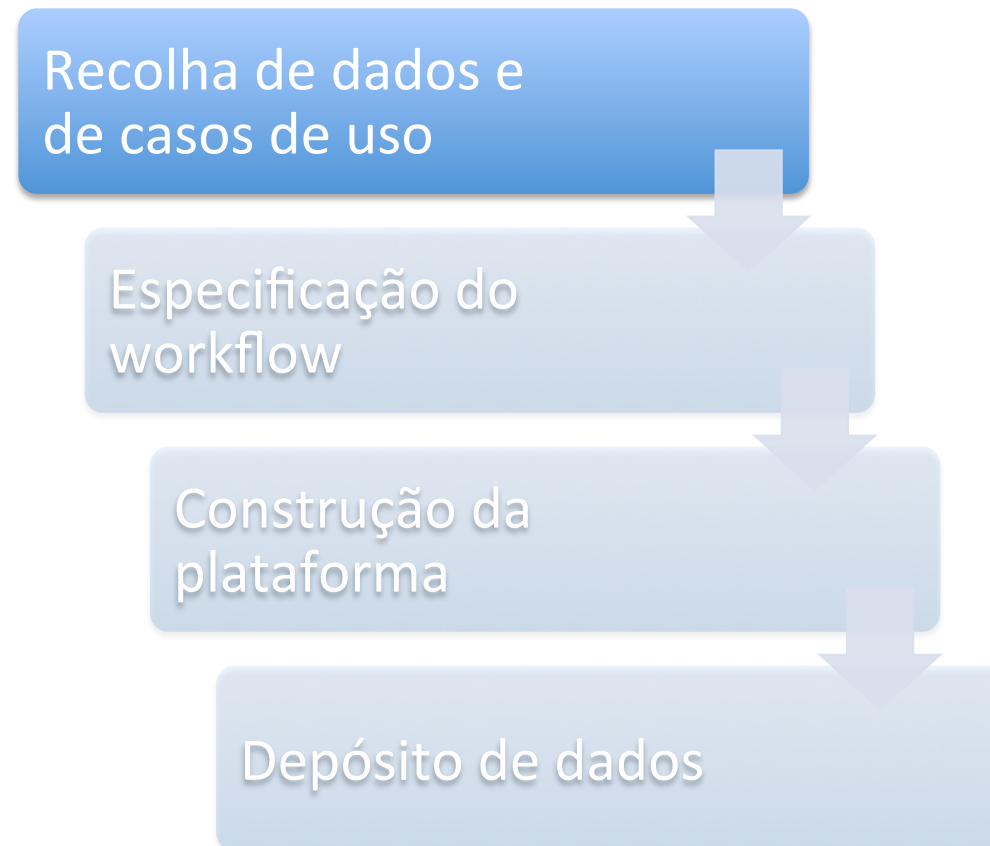
Satisfazer objetivos



Robin Rice, September 2007

- 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!”

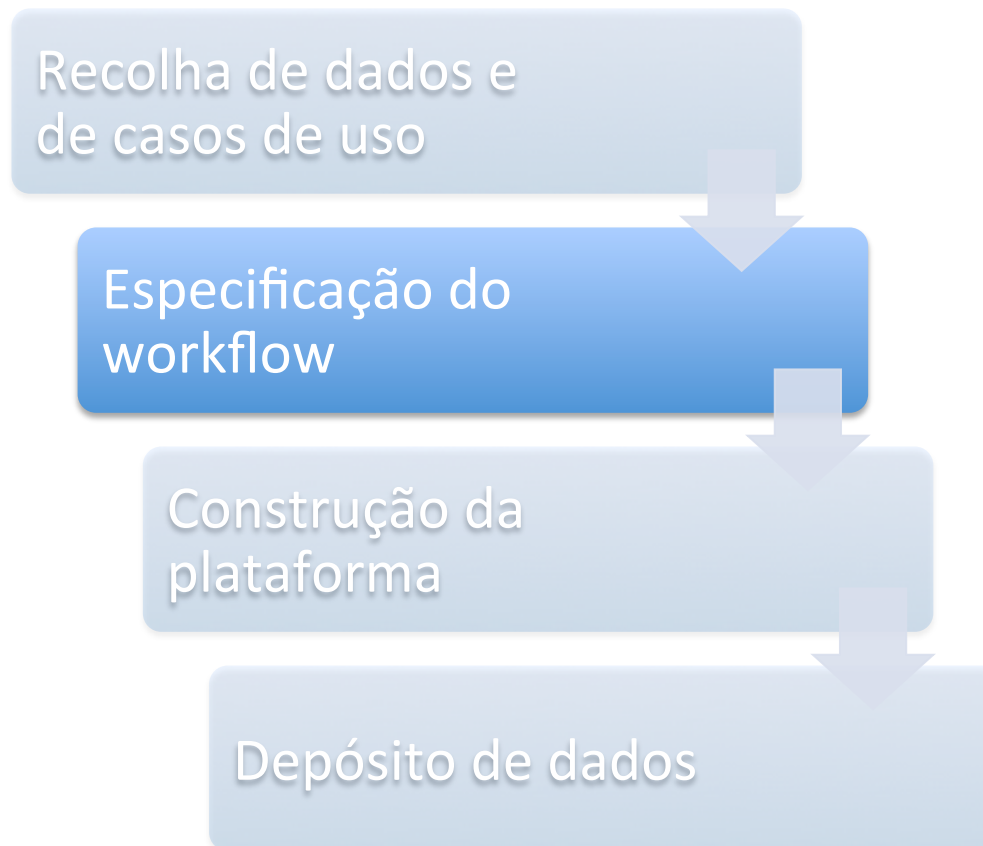
Os investigadores dizem

- ... 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

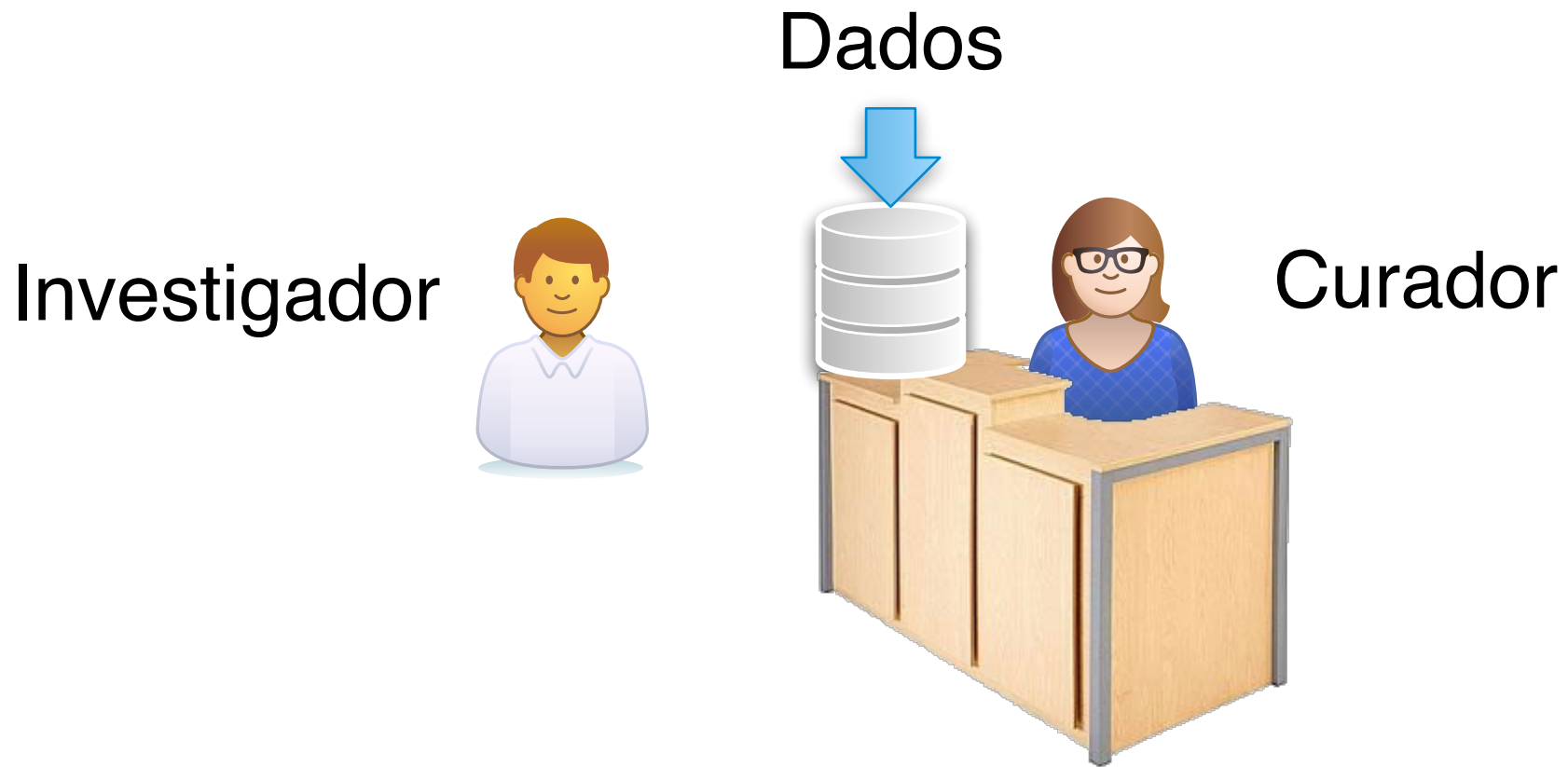
Preocupações e necessidade dos investigadores

- “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

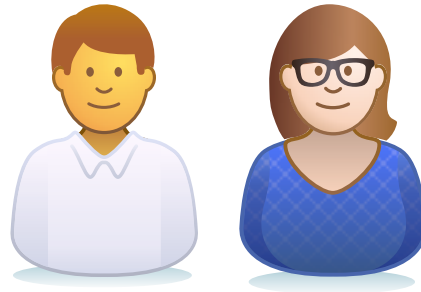
Fase 2 : Modificar o workflow



O Papel do “Curador de Dados”



Reunião de curadoria



Dados
Curados

Anotação de dados

Elementos do XML Schema

do domínio como
descritores
e **colunas**

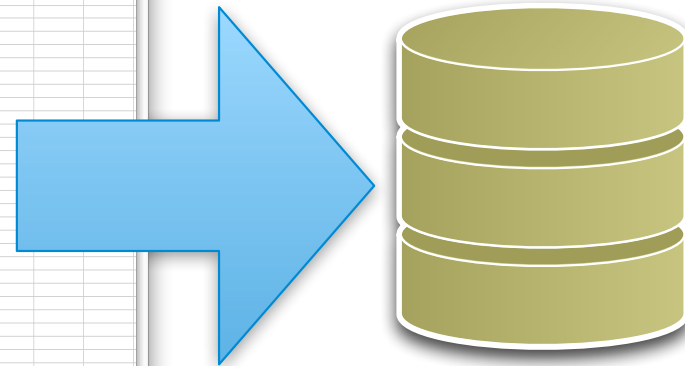
dc:contributor.author	Silva, João Rocha	} Table-level metadata		
dc:lastModified	01-01-2011			
dc:title	Azores GPS Run			
dc:rights	License: CC ShareAlike			
END_METADATA				
time.gps_sow	latitude	longitude	gravity.specific	} Dimensions
488496.999194	38.760267507	-27.084113730	-53.750371	} Data
488497.999193	38.760267485	-27.084113744	-67.168032	
488498.999192	38.760267506	-27.084113739	-80.584969	
488499.999191	38.760267489	-27.084113743	-93.994527	
488500.999190	38.760267493	-27.084113746	-107.391006	

Terceira Flores

Depois da reunião

The screenshot shows an Excel spreadsheet with the following data:

dc.creator	Bastos, Luisa; Deurloo, Richard		
dc.date.issued	1992		
dc.rights	open access		
dc.title	Aerial Gravimetry Run (GPS Processed Data for Terceira Island - Sensor: Airplane Back		
dc.description	Processed GPS coordinates for the airplane, for the Terceira Island (airplane back)		
dc.type	Numerical Data		
END_METADATA_SECTION			
grav.gstime	grav.latitude	grav.longitude	grav.height
488743.9999	38.76026894	-27.08411221	112.833
488744.9999	38.76026892	-27.0841122	112.838
488745.9999	38.76026894	-27.08411221	112.834
488746.9999	38.76026893	-27.08411219	112.836
488747.9999	38.76026891	-27.08411219	112.837
488748.9999	38.76026894	-27.0841122	112.834
488749.9999	38.76026893	-27.0841122	112.836
488750.9999	38.76026889	-27.08411219	112.84
488751.9999	38.76026889	-27.0841122	112.839
488752.9999	38.76026889	-27.08411219	112.838
488753.9999	38.76026886	-27.0841122	112.841
488754.9999	38.76026889	-27.0841122	112.837
488755.9999	38.76026889	-27.0841122	112.842
488756.9999	38.76026889	-27.0841122	112.838
488757.9999	38.76026886	-27.08411218	112.84
488758.9999	38.76026887	-27.08411218	112.839
488759.9999	38.76026885	-27.08411218	112.84
488760.9999	38.76026885	-27.0841122	112.837
488761.9999	38.76026888	-27.08411219	112.836
488762.9999	38.76026888	-27.0841122	112.835
488763.9999	38.76026887	-27.08411219	112.837
488764.9999	38.76026889	-27.0841122	112.838
488765.9999	38.76026887	-27.08411218	112.841
488766.9999	38.76026889	-27.08411219	112.835
488767.9999	38.76026889	-27.0841122	112.834
488768.9999	38.76026889	-27.08411219	112.839
488769.9999	38.76026889	-27.0841122	112.838
488770.9999	38.76026888	-27.08411219	112.839
488771.9999	38.76026888	-27.08411221	112.842
488772.9999	38.76026889	-27.08411222	112.84
488773.9999	38.76026889	-27.0841122	112.835
488774.9999	38.76026888	-27.0841122	112.838
488775.9999	38.76026887	-27.0841122	112.838
488776.9999	38.76026886	-27.0841122	112.843
488777.9999	38.76026886	-27.0841122	112.843
488778.9999	38.76026887	-27.08411222	112.845
488779.9999	38.76026887	-27.08411221	112.841
488780.9999	38.76026888	-27.08411221	112.842
488781.9999	38.76026889	-27.08411221	112.838
488782.9999	38.76026886	-27.08411219	112.842
488783.9999	38.76026889	-27.08411222	112.84
488784.9999	38.76026889	-27.08411221	112.839
488785.9999	38.76026887	-27.08411222	112.838
488786.9999	38.76026887	-27.08411221	112.836



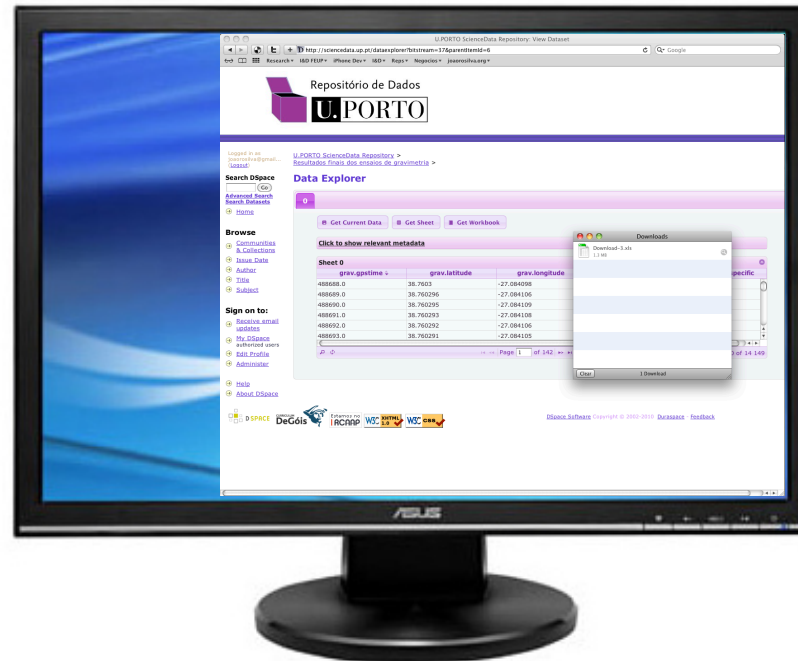
Repositório

Dados+Metadados em formato Excel

Dados disponíveis

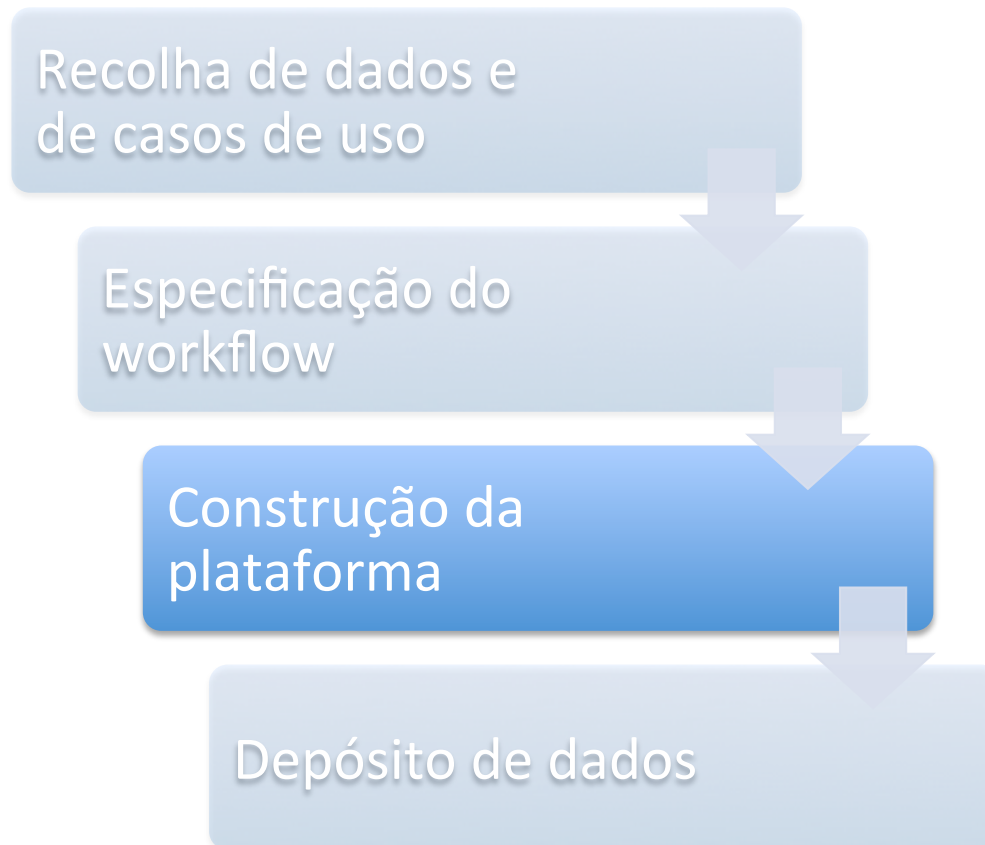
Repositório de dados de investigação

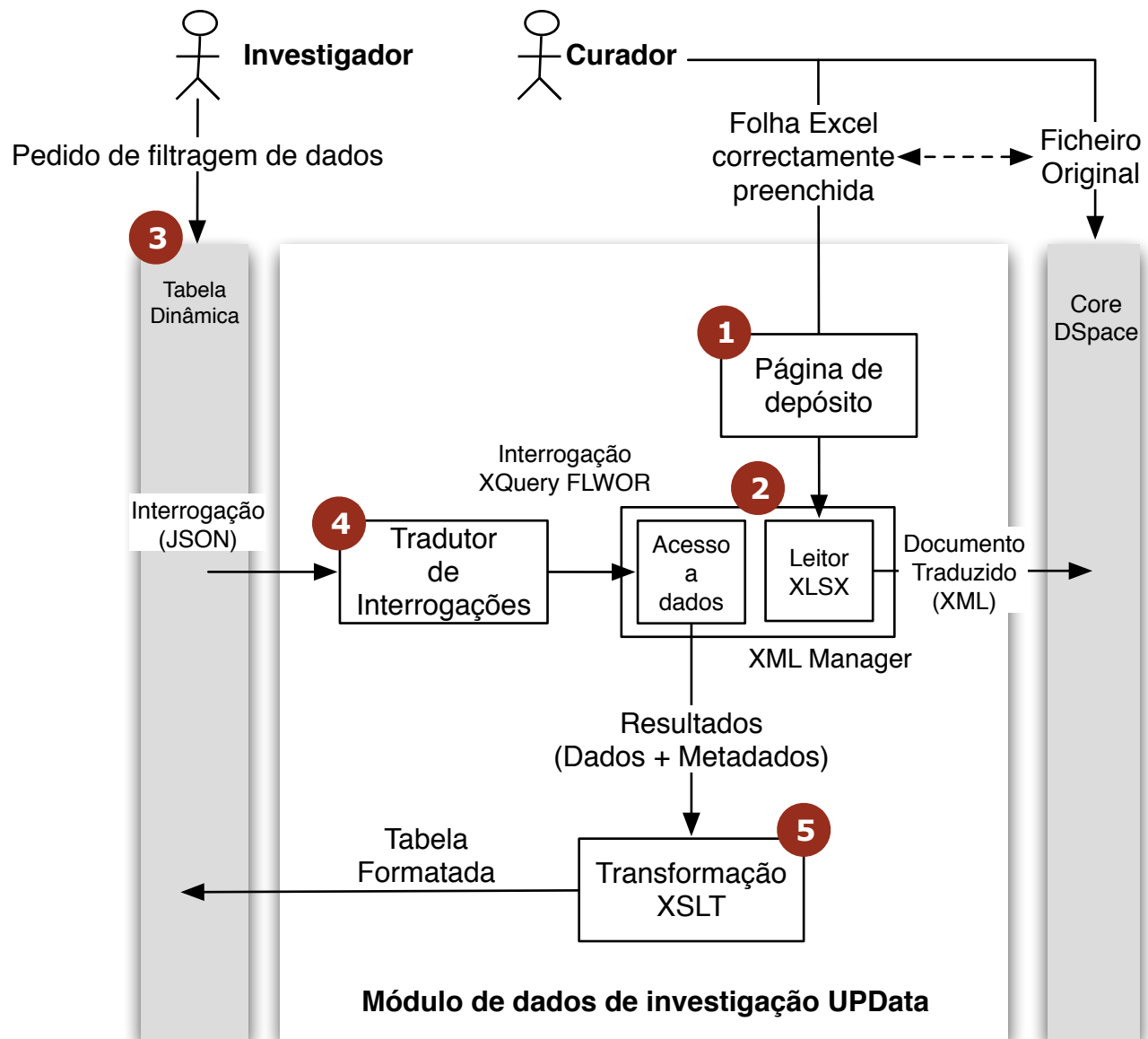
Investigador



- Explorar, filtrar e descarregar só o necessário

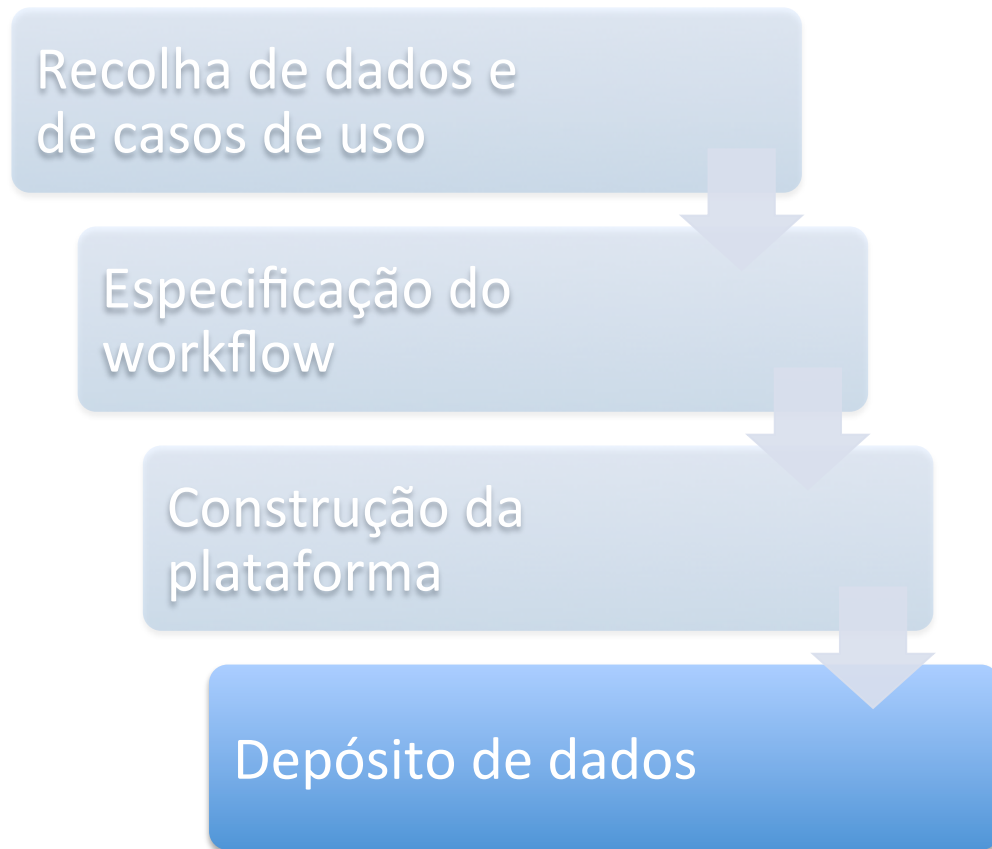
Fase 3 : Construir ferramentas de suporte ao workflow





```
<?xml version="1.0"?>
<record>
  <metadata>
    <dc.creator>Bastos, Lu&#237;sa; Deurloo, Richard</dc.creator>
    <dc.title>Aerial 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>
    </rows>
  </data>
</record>
```

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>