Guidelines for legacy repository migration

José Carlos Ramalho
KEEPS/University of Minho

Miguel Ferreira
KEEPS

Luís Faria
KEEPS

Archiving 2013
Washington DC, USA, 2-5 April 2013

Who are we?

- 6 years ago a research group inside an university computer science dep.:
  - 15 years experience with archives: developing solutions, migrations, …
- Last 6 years: a software company
  - Solutions for archives: finding-aids, digital assets management (images, audio, video, etc), electronic balcony, interventions management, etc;
  - Solutions for libraries: catalog, lowns, etc;
  - Solutions for museums;
  - Solutions for agreggation portals: harvesters supporting, etc;
  - Solutions for Digital Preservation: RODA, CRiB;
  - Open-source contributors: RODA, Koha
  - Commited to research:
    - 2 European projects: SCAPE and 4C;
    - Many other national projects;
    - Undergoing Phds and Msc thesis.

Find out more:
KEEP SOLUTIONS Lda
http://www.keep.pt

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Motivation

- Today we have a 30-40 years information systems legacy;
- Technological evolution has never been so fast and with such big steps;
- We have a rapid data growth giving rise to new problems and paradigms: “Big Data Problems”;
- Data or system migration is not an option, it is a necessity and it is potentially complex.

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Why shall we migrate?

- Repository system does not cope well with current business needs;
- Budget cuts: more financially sustainable repository is adopted;
- The repository vendor or supporter ceased to exist;
- Repository vendor or supporter does not provide a satisfactory level of support services;
- The technological environment needed by the repository system is no longer supported.

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

What is migration?

x = old system

y = new system

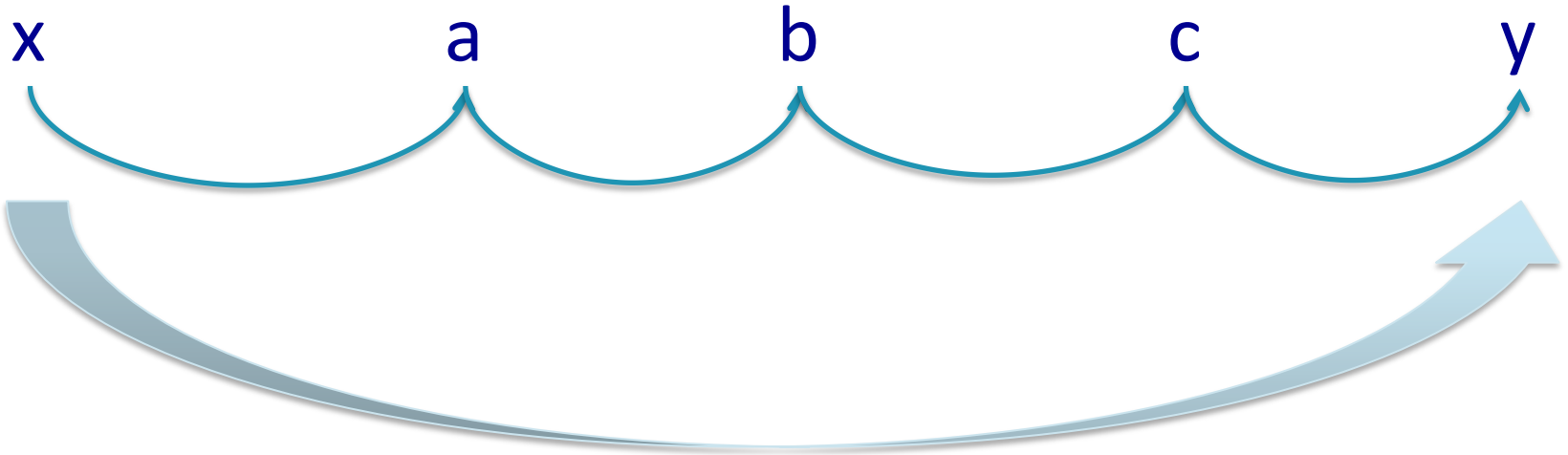
f(x) = y

About f:

- Complete or partial?
- How long does it take?
- Does time matter?
- Is it reliable?
- …

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Legacy system migration

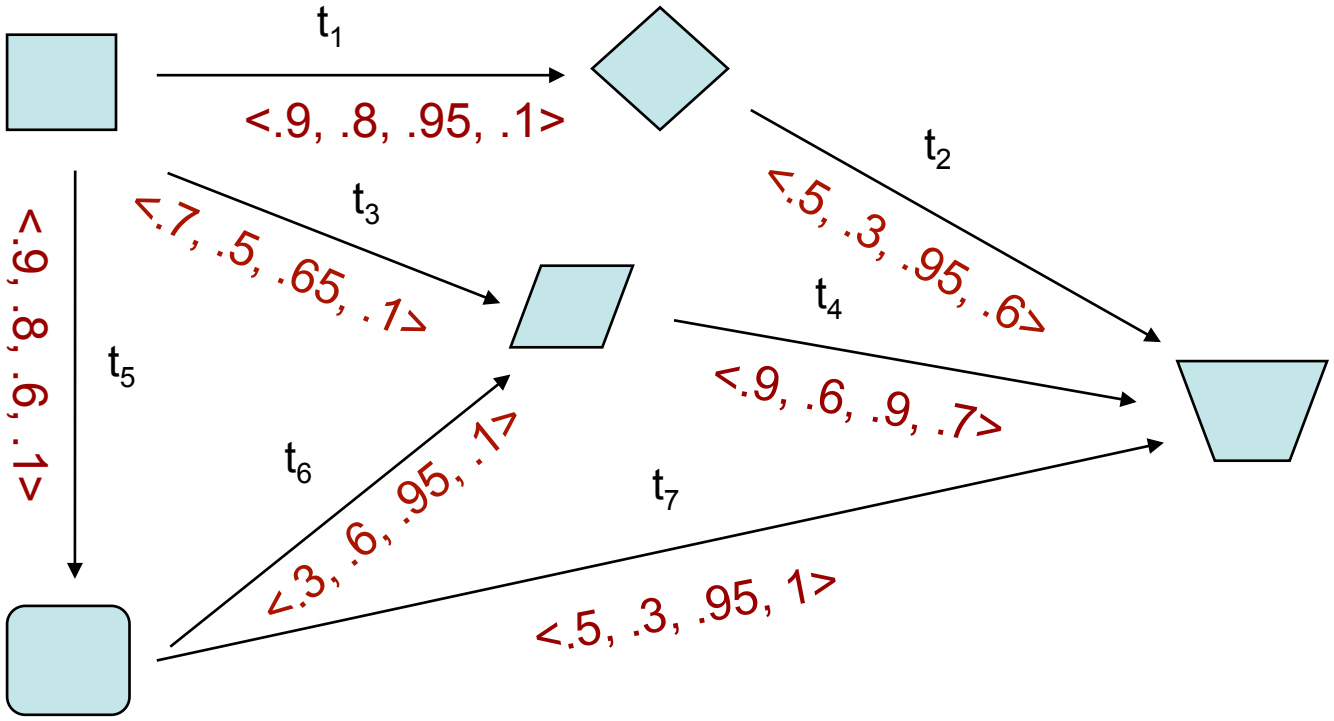


$$y = f_{c2y} \circ f_{b2c} \circ f_{a2b} \circ f_{x2a}$$

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Migration Service

CRiB project: http://crib.dsi.uminho.pt

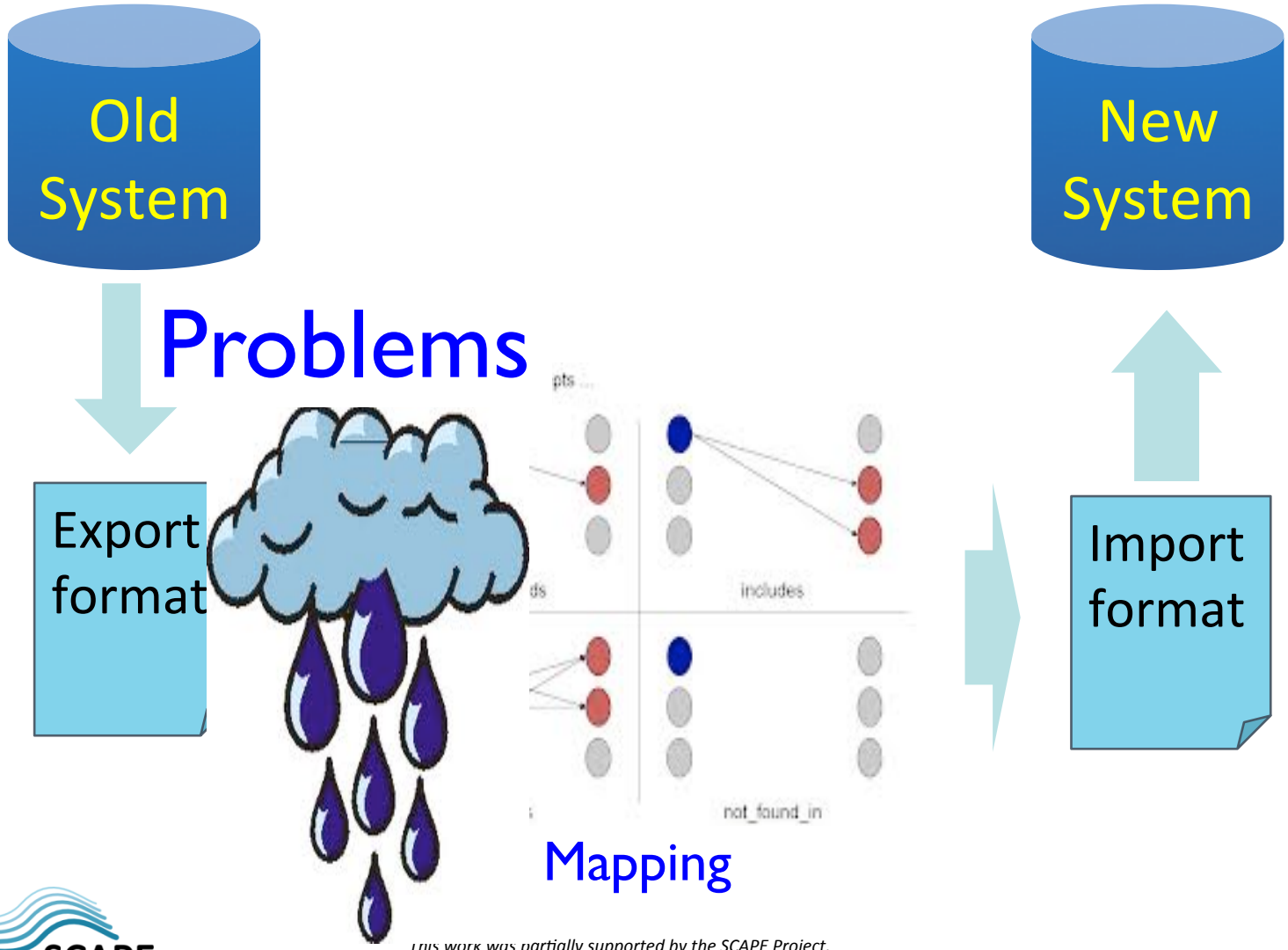


History

- Migration (CALM): all Portuguese National Archives document funds - several millions of records, a few thousand funds and associated digital representations, …
- Migration (ARQBASE): all Azores document funds: Ponta Delgada, Angra do Heroísmo e Faial;
- Several migrations from other archive platforms;
- Migration: several libraries from medium to large: ISCTE, UBI, Lusíada, etc;
- Migration/Upgrade: several Dspace repositories .

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Perfect scenario



This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

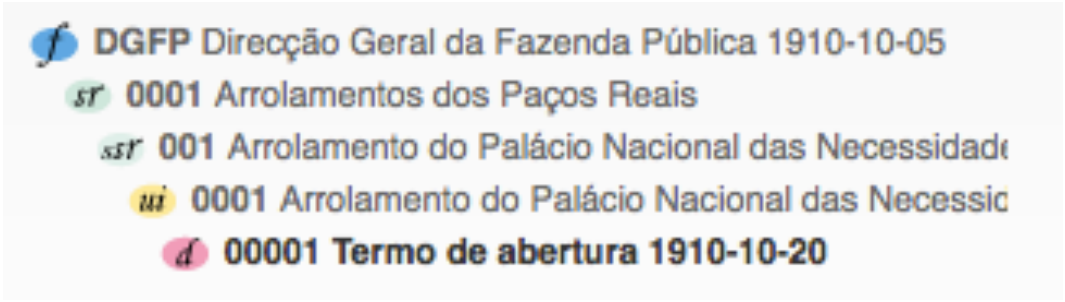
Detected Problems

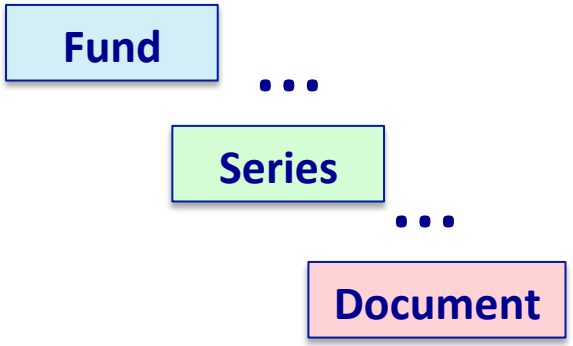
Import/Export features:

- Not implemented or not working properly;
- Rarely tested during acquisition (no data, lack of experience);
- When they are needed it is too late…

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Detected problems (Archives)





- Today's restrictions did not apply yesterday…
- Example: hierarchical structural invariants
  - Not supported by the platforms from which we migrate;
- Exporters functioning with problems;
- Lack of information: data creators are no long available.

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Example: Portuguese National Archives

```
<calmreg>
  <Date>1894-1910</Date>
  <CatalogueStatus>Valida��o</CatalogueStatus>
  <UserText2>Estante 1, prateleira 1, volume 4</UserText2>
  <AltRefNo>Arquivo das Congrega��es, liv. 4</AltRefNo>
  <ArchNote>Descri��o elaborada por Luisa Dias (ANTT), a partir
  <UserText3>Julho de 2008</UserText3>
  <Repository>ANTT</Repository>
  <Title>"Col�gio de Santa Clara"</Title>
  <Extent>1 liv.</Extent>
  <RefNo>PT-TT-AC/L4</RefNo>
  <Level>Documento Composto</Level>
```

- 2 millions of records:
  - More than 20 date formats;
  - Lots of free fields (that should have had controlled vocabularies);
  - 70.000 structural errors:
    - Intermediate level records did not exist;
    - Many orphan registers

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Example: Portuguese National Archives

```
<calmreg>
  <Date>1894-1910</Date>
  <CatalogueStatus>Valida
  <UserText2>Estante 1, pr
  <AltRefNo>Arquivo das Co
  <ArchNote>Descri  o ela
  <UserText3>Julho de 2008
  <Repository>ANTT</Reposit
  <Title>"Col gio de Santo
  <Extent>1 liv.</Extent>
  <RefNo>PT-TT-AC/L4</RefN
  <Level>Documento Composto
```

erros.txt

```
290467; UI; PT/TT/AOS/D/E/12/2/19; Imposto de selo. Alterações à tabela aprovada pelo
Decreto-Lei n.º 21916/32, de 28 de Novembro.; Inconsistência hierárquica
290554; UI; PT/TT/AOS/D/E/17/2/2; Baixela Germain; Inconsistência hierárquica
290099; DC; PT/TT/AOS/D/E/3/3/1; Correspondente do Banco de Portugal em Santa Comba Dão.;
Inconsistência hierárquica
290110; DC; PT/TT/AOS/D/E/3/3/12; Negociações com a Alemanha para a liquidação de
pagamentos.; Inconsistência hierárquica
290111; DC; PT/TT/AOS/D/E/3/3/13; Recondução de Álvaro Pedro de Sousa e de Carlos de
Barros Soares Branco nos lugares de vice-governadores do Banco de Portugal.;
Inconsistência hierárquica
290112; DC; PT/TT/AOS/D/E/3/3/14; Suspensão da convertibilidade do franco francês.;
Inconsistência hierárquica
290113; DC; PT/TT/AOS/D/E/3/3/15; Pontos de vista e opiniões recolhidas por Carlos de
Barros Soares Branco em Londres.; Inconsistência hierárquica
290114; DC; PT/TT/AOS/D/E/3/3/16; Pagamento de material de guerra alemão a fornecer a
Portugal.; Inconsistência hierárquica
290115; DC; PT/TT/AOS/D/E/3/3/17; Acordos de compensação celebrados entre o Governo
Português e os vários governos estrangeiros.; Inconsistência hierárquica
```

Solution:

- Error reporting tool development;
- Several task forces were created locally;
- 3 month iterative work:
  - correct->report->correct->…

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Example: Azores (ARQBASE)

```
<RECORD>

<Tag_99>5.0. </Tag_99>
<Tag_101><Tag_101_a>BPARPD</Tag_101_a></Tag_101>
<Tag_201><Tag_201_a>FAM</Tag_201_a></Tag_201>
<Tag_301><Tag_301_a>ADCM</Tag_301_a><Tag_301_b>Arquivo Dias do Canto e Medeiros</Tag_301_b></Tag_301>
<Tag_501><Tag_501_a>Lv</Tag_501_a><Tag_501_b>001</Tag_501_b></Tag_501>
<Tag_601><Tag_601_a>1</Tag_601_a></Tag_601>
<Tag_603><Tag_603_a>Ponta Delgada</Tag_603_a><Tag_603_b>Carta de inquirição</Tag_603_b><Tag_603_c>António Jacinto
Mendes (Escrivão)</Tag_603_c><Tag_603_e>Carta de inquirição cível de testemunhas, relativa à sucessão da
instituição de Luís do Canto de Vasconcelos que opôs João Silvério Vaz Pacheco de Castro a António de Faria do
Canto e Maia e a Jacinto Pacheco de Castro. Contém inquirição do embargante Jacinto Pacheco de Castro, a
descendência de D. Isabel do Canto, de André Diogo Dias do Canto e Medeiros, e de Luís do Canto de
Vasconcelos.</Tag_603_e><Tag_603_f>Pública-forma</Tag_603_f></Tag_603>
<Tag_607>1830. </Tag_607>
<Tag_609>25 (p.1-43). António de Faria do Canto e Maia tb. é designado noutros docs. por António de Faria do
Canto Machado e Maia ou António de Faria do Canto e Medeiros. <Tag_609_a>25 (p.1-43). <Tag_609_d>Bom.
<Tag_609_e>Datas do doc. orig.: 1830.03.22-1830.04.14. António de Faria do Canto e Maia tb. é designado noutros
docs. por António de Faria do Canto Machado e Maia ou António de Faria do Canto e Medeiros.</Tag_609_e></Tag_609>
</RECORD>
```

- Export format was bad XML;
- Data creators unavailable;
- Lots of structural errors, semantic errors and normalization errors.

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Example: Azores (ARQBASE)

```
13 <RECORD>
14
15 <Tag_99>5.0. </Tag_99>
16 <Tag_101><Tag_101_a>BPARPD</Tag_101_a></Tag_101>
17 <Tag_201><Tag_201_a>FAM</Tag_201_a></Tag_201>
18 <Tag_301><Tag_301_a>ADCM</Tag_301_a><Tag_301_b>Arquivo Dia
19 <Tag_501><Tag_501_a>Lv</Tag_501_a><Tag_501_b>001</Tag_501_
20 <Tag_601><Tag_601_a>1</Tag_601_a></Tag_601>
21 <Tag_603><Tag_603_a>Ponta Delgada</Tag_603_a><Tag_603_b>Ca
21 Mendes (Escrivão)</Tag_603_c><Tag_603_e>Carta de inquiriçã
21 instituição de Luís do Canto de Vasconcelos que opôs João
21 Canto e Maia e a Jacinto Pacheco de Castro. Contém inquiri
21 descendência de D. Isabel do Canto, de André Diogo Dias do
21 Vasconcelos.</Tag_603_e><Tag_603_f>Pública-forma</Tag_603_
22 <Tag_607>1830. </Tag_607>
23 <Tag_609>25 (p.1-43). António de Faria do Canto e Maia tb.
23 Canto Machado e Maia ou António de Faria do Canto e Medeir
23 <Tag_609_e>Datas do doc. orig.: 1830.03.22-1830.04.14. Ant
23 docs. por António de Faria do Canto Machado e Maia ou Antó
24 </RECORD>
```

ARQBASE Reporting Utility (v1.0

General Data

Total number of records	59
Funds	1
Sections	3
Series	8
SubSeries	0
Units	24
Documents	23

Number of Registers by Fund

VN 59

Integrity Testing

Series

PT/BPARJJG/PSS/VN/AI/01

Tabela 3.2: Registos criados por fundo

Fundo	Nível	Designação
APRT	Secção	VDC
DFDPRT	Secção	RFPRT FRSTS
TCPRT	Secção	DC
	Série	068 158 160 161 162 163 184

Solution:

- Error reporting tool development;
- Automatic correction (there was no one to validate…).

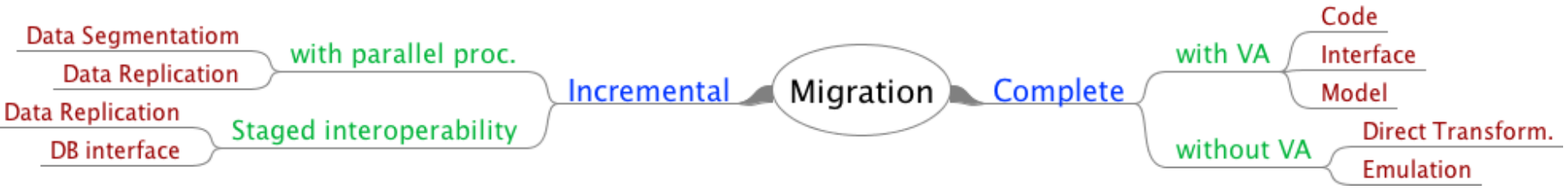
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Detected problems (Libraries)

- Information access:
  - Exporters do not work properly;
  - Suppliers don't give access to information;
- Format Problems: Marc21, UniMarc, XML;
- Encoding Problems: representation format hard to identify;
- Lots of structural changes:
  - Field folding;
  - Field unfolding;
  - Structural changes.

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Migration Strategies



This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Migration anatomy (part I)

1. Analysis & consultation

1A
Characterization of legacy environment

1B
Characterization of target environment

1C
Data analysis

1D
Strategic planning

1E
Requirements Specification

Expected outcomes:

- preservation plan;
- business goals;
- preservation requirements;
- evaluated alternatives;
- decision process results.

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Migration anatomy (part II)

2. Planning & design

2A
Project planning

2B
Design of migration routines

2C
Design of test plan

Expected outcomes:

- Project plan:
  - Who?
  - What?
  - When?

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Migration anatomy (part III)

3. Development

3A
Development of migration routines

3B
Development of testing routines

Expected outcomes:

- Code…

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Migration anatomy (part IV)

4. Setup & testing

4A
Target environment provisioning

4B
Rehearsal & testing

Expected outcomes:

- Installed target system;
- Verified and tested system.

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Migration anatomy (part V)

5. Execution

5A
Execution of migration routines

Expected outcomes:

- Migrated System .

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Migration anatomy (part VI)

6. Validation

6A
Execution of testing routines

6B
Reporting

6C
Cut-over

Expected outcomes:

- Reports from test scripts.

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Migration anatomy (part VII)

7. Wrap-up

7A
Training

7B
Documenting

7C
Supporting

Expected outcomes:

- All reports related to the process:
  - migration;
  - tests;
  - training materials;
  - manuals, …

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Summary

1. Analysis
2. Planning
3. Development
4. Setup & testing
5. Execution
6. Validation
7. Wrap-up

If you follow these steps things will run…

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Conclusion

1. Migration is complex,
2. Migration consumes resources,
3. Migration can go wrong,
4. Migration without planning is caotic,
5. Migration is inevitable.

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

SCAPE
SCAlable Preservation Environments

Questions?

José Carlos Ramalho
KEEPS/University of Minho

KEEPSOLUTIONS
University of Minho SPIN-OFF

http://www.keep.pt