Tools for NextGen Scholarly Communication:

Work underway at the University of Minho

29/Jan/2007

Ana Alice Baptista
http://www.dsi.uminho.pt/~analice

(this document is available in Open Access at http://hdl.handle.net/1822/6180)
Contents

Scope:

- UMinho and DSI
- SDUM and RepositóriUM
- Odisseia
- CRiB
- Tea for Two
- Other Projects
- Future work
Contents

Scope:

- UMinho and DSI
- SDUM and RepositóriUM
- Odisseia
- CRiB
- Tea for Two
- Other Projects
- Future work
• Since 1974
• 2 campi
• 11 schools/ institutes (faculties)
• 30 research centres
• 13500 undergrad
• 1500 grad (MSc + doctorate)
• 1116 FTE
Information Systems Department

http://www.dsi.uminho.pt

- School of Engineering
- One of the two computer departments: **technology in context**
- Main interface areas: management and information science
- 2 graduation programs, 3 MSc programs and 1 doctorate program
- 37 teachers and 12 hired researchers
- 114 MSc students - 43 doing thesis
- 42 PhD students
SDUM
http://www.sdum.uminho.pt

- Documentation Services
- 4 main libraries
- + 21 other
- Several digital services (catalog, search portal, digital publications, repositóriUM,...)
RepositóriUM
http://repositorium.sdum.uminho.pt

- Developed in the scope of the e-UM Virtual Campus project;
- Implemented in DSpace
- Start date: April 2003
- Open to the public: November 2003
- Provided a framework for R&D activities:
  - 2 + 4 add-ons for DSpace (2 included in version 1.4)
RepositóriUM
http://repositorium.sdum.uminho.pt

<table>
<thead>
<tr>
<th>Consultas</th>
<th>Downloads</th>
</tr>
</thead>
<tbody>
<tr>
<td>470184</td>
<td>531118</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Docs</th>
<th>Consultas</th>
<th>Downloads</th>
</tr>
</thead>
<tbody>
<tr>
<td>5151</td>
<td>91.3</td>
<td>103.1</td>
</tr>
</tbody>
</table>

Since 01/Jan/06
RepositóriUM
http://repositorium.sdum.uminho.pt

Since 01/Jan/06

<table>
<thead>
<tr>
<th>Origem</th>
<th>Consultas</th>
<th>Downloads</th>
</tr>
</thead>
<tbody>
<tr>
<td>Portugal</td>
<td>194413</td>
<td>250505</td>
</tr>
<tr>
<td>Brazil</td>
<td>103878</td>
<td>133603</td>
</tr>
<tr>
<td>Universidade do Minho</td>
<td>70553</td>
<td>32484</td>
</tr>
<tr>
<td>United States</td>
<td>50456</td>
<td>18026</td>
</tr>
<tr>
<td>India</td>
<td>3165</td>
<td>8954</td>
</tr>
<tr>
<td>N/A</td>
<td>6086</td>
<td>8802</td>
</tr>
<tr>
<td>China</td>
<td>3156</td>
<td>8208</td>
</tr>
<tr>
<td>Spain</td>
<td>4026</td>
<td>5682</td>
</tr>
</tbody>
</table>
Odisseia
http://odisseia.dsi.uminho.pt

- Research group on Scholarly Communication
- Scientific sub-area: Information Society
- Started in 2004 with some projects, including the “DSpace Dev@University of Minho”
Currently:

- CRiB - digital preservation
- Tea for Two - informal communication + social networks
- plus some other satellite projects

Odisseia
http://odisseia.dsi.uminho.pt
Contents

• Scope:
  • UMinho and DSI
  • SDUM and RepositóriUM
  • Odisseia
  • CRiB
  • Tea for Two
  • Other Projects
  • Future work
CRiB

- Basically a PhD and satellite projects
- Start date: Jan/2005, End date: Dec/2007
- Problem: Repositories are
  - excellent at archiving and disseminating materials
  - poor at preserving those materials in the long run
CRiB

- Service Oriented Architecture (SOA)
- Preservation approach: migration
- Recommendation service - Recommends an optimal migration strategy taking into account
  - Requirements of each client institution
  - Behavior/quality of each migration service
- Migration services
  - Service composition
- Evaluates the outcome of each migration
  - Performance, data loss, format characteristics
  - Produces an evaluation report (authenticity)
CRiB - Scenario

• A collection of digital objects of a certain format
• e.g. JPEG files collected from a digital camera
• e.g. A collection of text documents
CRiB - Scenario

- Using the recommendation service
- Preservation format (i.e. The target format)
- Migration service (or combination of services)
CRiB - Scenario

- Using the conversion services
- Check for data loss and generate a migration report
- Store the report
- Return the converted file and the report back to the
CRiB

General Architecture

Application layer
- Custom applications
  - Web client
  - Rich client
- Digital repositories
  - DSpace
  - Fedora
  - Eprints

Business layer
- Metaconverter
  - Service Registry
  - Migration Broker
  - Object Evaluator
  - Format Evaluator
  - Migration Advisor

Data layer
- UDDI
- Format Knowledge Base
- Evaluations Repository
- Conversion Services
- External Services
CRiB
General Architecture

Handles all communication between the client and the CRiB system
It's a web service
Orchestrates the communication within the system
CRiB
General Architecture

Carries out format migrations
Invokes all the necessary conversion services
Measures the performance of the conversion process
CRiB

General Architecture

Determines the amount of data loss involved in the migration
Detects the similarity of each significant property of the digital object
Produces evaluation reports in PREMIS format (eventOutcomeDetail)
Generates recommendations of optimal migration choices
Uses information provided by the client to determine the best available option
Confronts those requirements with the accumulated knowledge about the behavior of each conversion service
CRiB
General Architecture

Application layer
- Custom applications
  - Web client
  - Rich client
- Digital repositories
  - DSpace
  - Fedora
  - Eprints

Business layer
- Metaconverter
- Service Registry
- Migration Broker
- Object Evaluator
- Format Evaluator
- Migration Advisor

Data layer
- UDDI
- Format Knowledge Base
- Evaluations Repository
- Conversion Services
- External Services
CRiB
Recommendation Engine

Evaluation criteria
<table>
<thead>
<tr>
<th></th>
<th>WS1</th>
<th>WS2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Availability</td>
<td>95%</td>
<td>90%</td>
</tr>
<tr>
<td>Stability</td>
<td>80%</td>
<td>90%</td>
</tr>
<tr>
<td>Scalability</td>
<td>1</td>
<td>1.5</td>
</tr>
<tr>
<td>Throughput</td>
<td>1 Kb/s</td>
<td>2 Kb/s</td>
</tr>
<tr>
<td>Cost</td>
<td>0.01 €</td>
<td>0.05 €</td>
</tr>
<tr>
<td>Number of chars</td>
<td>70%</td>
<td>95%</td>
</tr>
<tr>
<td>Ordering of chars</td>
<td>60%</td>
<td>85%</td>
</tr>
<tr>
<td>Page layout</td>
<td>80%</td>
<td>20%</td>
</tr>
<tr>
<td>Market share</td>
<td>30%</td>
<td>70%</td>
</tr>
<tr>
<td>Support level</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Open standard</td>
<td>yes</td>
<td>no</td>
</tr>
</tbody>
</table>

Weights
<table>
<thead>
<tr>
<th></th>
<th>WS1</th>
<th>WS2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Availability</td>
<td>40%</td>
<td></td>
</tr>
<tr>
<td>Stability</td>
<td>30%</td>
<td></td>
</tr>
<tr>
<td>Scalability</td>
<td>0%</td>
<td></td>
</tr>
<tr>
<td>Throughput</td>
<td>20%</td>
<td></td>
</tr>
<tr>
<td>Cost</td>
<td>10%</td>
<td></td>
</tr>
<tr>
<td>Number of chars</td>
<td>40%</td>
<td></td>
</tr>
<tr>
<td>Ordering of chars</td>
<td>40%</td>
<td></td>
</tr>
<tr>
<td>Page layout</td>
<td>20%</td>
<td></td>
</tr>
<tr>
<td>Market share</td>
<td>30%</td>
<td></td>
</tr>
<tr>
<td>Support level</td>
<td>30%</td>
<td></td>
</tr>
<tr>
<td>Open standard</td>
<td>40%</td>
<td></td>
</tr>
</tbody>
</table>

Normalised evaluations [1-5]

<table>
<thead>
<tr>
<th></th>
<th>WS1</th>
<th>WS2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Availability</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Stability</td>
<td>5</td>
<td>5</td>
</tr>
<tr>
<td>Scalability</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Throughput</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Cost</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Number of chars</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Ordering of chars</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>Page layout</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>Market share</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>Support level</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Open standard</td>
<td>5</td>
<td>1</td>
</tr>
</tbody>
</table>

User assigned weights of importance

<table>
<thead>
<tr>
<th></th>
<th>WS1</th>
<th>WS2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Availability</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>Stability</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Scalability</td>
<td>0.0</td>
<td></td>
</tr>
<tr>
<td>Throughput</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>Cost</td>
<td>0.1</td>
<td></td>
</tr>
<tr>
<td>Number of chars</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>Ordering of chars</td>
<td>0.4</td>
<td></td>
</tr>
<tr>
<td>Page layout</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>Market share</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Support level</td>
<td>0.3</td>
<td></td>
</tr>
<tr>
<td>Open standard</td>
<td>0.4</td>
<td></td>
</tr>
</tbody>
</table>

Ranking process

\[
\begin{bmatrix}
1.6 & 2 \\
2 & 1 \\
0.6 & 1.2 \\
1.5 & 1.2 \\
2 & 0.4
\end{bmatrix}
\times
\begin{bmatrix}
2 \\
1.5 \\
1.5 \\
0.5 \\
0.1
\end{bmatrix}
\text{ = }
\begin{bmatrix}
12.5 \\
11.8
\end{bmatrix}
\]

Overall rating

12.5
11.8

(1) Average evaluations per service
(2) User assigned weights of importance
(3) Normalised evaluations [1-5]
(4) Ranking process
(5) Overall rating
CRiB - Who

Other people in a non-regular basis.

+ PhD co-supervisor

Miguel Ferreira
mferreira@dsi.uminho.pt

Ana A. Baptistana
analice@dsi.uminho.pt
Contents

Scope:

- UMinho and DSI
- SDUM and RepositóriUM
- Odisseia
- CRiB
- Tea for Two
- Other Projects
- Future work
DSpace Dev@University of Minho

http://dspace-dev.dsi.uminho.pt

- Started and Finished in 2004
- Implementation of some “old” ideas
- Participants: Ana Alice Baptista and Miguel Ferreira
- Add-ons developed:
  - Controlled Vocabulary
  - Commenting
  - Web of Communication
  - Recommendation
Controlled Vocabularies add-on

- Goal: to facilitate and improve resource discovery by means of controlled vocabularies
- First Implementation: papadocs with ACM CCS
- Additional deliverables: Specific XML Schema; Official XML version of the ACM CCS
- Integrated in version 1.4 of DSpace
- Being upgraded in a joint project with FAO (Food and Agriculture Organization - United Nations)
Metadata Based and Combined Information Services

OAISTER  BASE  Google Scholar

Repository Formal Communication Layer

Citation Relationship
Controlled Vocabulary Specific Relationship
Data Flow
Selezione os conceitos a pesquisar

Filtro: ACMCCS98
- General Literature
- Hardware
- Computer Systems Organization
- Software
- Data
- Theory of Computation
- Mathematics of Computing
- Information Systems
  - GENERAL
  - MODELS AND PRINCIPLES
  - DATABASE MANAGEMENT
  - INFORMATION STORAGE AND RETRIEVAL
  - INFORMATION SYSTEMS APPLICATIONS
Commenting add-on

• Goal: to bring informal communication to DSpace in a way that it would be directly related to formal communication.

• First Implementation: papadocs

• Specific issue: graph (and not tree) based threads
Tree-based comments' network

Graph-based comments' network

---

Comment Relationship
Todos os registos no repositório estão protegidos por leis de copyright, com todos os direitos reservados.
WoC add-on

- Goal: to provide a way to determine and depict the existing relationships between scholars that would otherwise be invisible.
- A Web of Communication is a network of weighted communication relationships between persons.
- First implementation: papadocs.
Recommendation add-on

- A kind of WoC in text mode
- Visible relationships:
  - related people
  - related formal documents
  - related comments
Papadocs → Comissão Científica → Propostas de Doutoramento

Título: Preservação de Longa Duração de Informação Digital no contexto de um Arquivo Histórico

Autor(es): Ferreira, Miguel (n/a)

Orientador(es): Baptista, Ana Alice Ramalho, José Carlos

Submetido por: Miguel Ferreira

Data submissão: sem data

Resumo: Certo tipo de documentação produzida por determinados ramos da Administração Pública (e.g. Registo Civil, Cartórios Notariais e Tribunais) é obrigatoriamente incorporada e preservada em Arquivos Históricos. Acontece que, uma grande parte dessa documentação é, actualmente, produzida em formato digital, o que levanta sérios problemas no que diz respeito à sua preservação. Os documentos digitais estão intrinsecamente relacionados com o software que os produziu, no entanto, o software possui um tempo de vida limitado sendo a sua descontinuidade um facto incontornável. O futuro previsível resume-se a um número variável de documentos prisioneiros do seu próprio formato de codificação. Diversas estratégias têm sido apontadas no sentido de solucionar o problema, no entanto, todas elas apresentam lacunas variadas e não abordam o problema.
Tea for Two - Rationale

- Knowledge communities are built having formal documents as a basis
- People belong to one or more subject related communities
- There are persons especially gifted to perform the role of information mediators within and across communities
- The knowledge these persons share is not properly stored and accessible
- If it was, it could help to discover the “undiscovered public knowledge (Swanson, 1986)
Tea for Two - Aims

• To provide more favorable means of uncovering undiscovered public knowledge. Creating the necessary conditions for the increase in informal interactions, storing them, and making them more visible, thus facilitating the access to them.

• To create an overall view of the communication relationships between the scholars, i.e., a cross repository, using the (formal and informal) communication interactions.
I - Controlled Vocabulary

– To use a rich standard description language to express controlled vocabularies.
2 - Comment – To turn the add-on into a cross-repository service.
3 - Comment – To add mechanisms for comprehensive metadata description of informal documents, including subject, author and type information;
4 - **Comment** – To add an intermediate informal communication layer between documents and comments
5 - **WoC** – To include an easy to use way to view subject information in a relationship between two persons;
6 - WoC – To combine existing relationships between two persons in a single one and depict it;
7 - WoC – To calculate the measure of the tightness of a combined relationship between two persons and find a way to depict it;
8 - WoC – To include other information (citation, co-authorship relationships) to calculate the tightness of combined relationships;
9 - WoC – To develop user configurable filters for the WoC depiction;
10 - WoC – To depict the direction of the communication between two authors.
1 - **Controlled Vocabulary** – To use a rich standard description language to express controlled vocabularies.

2 - **Comment** – To turn the add-on into a cross-repository service.

3 - **Comment** – To add mechanisms for comprehensive metadata description of informal documents, including subject, author and type information;

4 - **Comment** – To add an intermediate informal communication layer between documents and comments

5 - **WoC** – To include an easy to use way to view subject information in a relationship between two persons;

6 - **WoC** – To combine existing relationships between two persons in a single one and depict it;

7 - **WoC** – To calculate the measure of the tightness of a combined relationship between two persons and find a way to depict it;

8 - **WoC** – To include other information (citation, co-authorship relationships) to calculate the tightness of combined relationships;

9 - **WoC** – To develop user configurable filters for the WoC depiction;

10 - **WoC** – To depict the direction of the communication between two authors.
Tea for Two - Who

Ana A. Baptista
analice@dsi.uminho.pt

Miguel Ferreira
mferreira@dsi.uminho.pt

Eva Oliveira
eva.oliveira@gmail.com

Simone Costa
scosta@reitoria.uminho.pt
Contents

Scope:

- UMinho and DSI
- SDUM and RepositóriUM
- Odisseia
- CRiB
- Tea for Two
- Other Projects
- Future work
Other Projects

- Educational Contents Repository
- User Studies (waiting for approval)
- Social Tagging
- Informattica Online (standby)
Other Projects - Who
(Educational Contents Repository)

Ana Dias
anadias@tecminho.uminho.pt

Ana A. Baptista
analice@dsi.uminho.pt

José Carvalho
jcarvalho@tecminho.uminho.pt

Elisabete Cardoso
elisabete@dsi.uminho.pt
Other Projects - Who
(Social Tagging)

Ana A. Baptista  
analice@dsi.uminho.pt

Elisabete Catarino  
ecatarino@dsi.uminho.pt

Anabela Novais  
anabalanovais@gmail.com

Susana Pinheiro  
susana-pinheiro@hotmail.com
Contents

Scope:

- UMinho and DSI
- SDUM and RepositóriUM
- Odisseia
- CRiB
- Tea for Two
- Other Projects
- Future work
Future Work (CRiB)

• to change the way the overall rating (recom. engine) is calculated
• to finish the integration of evaluators
• to validate the prototype
• “Real” test as a service for digital preservation
Future work (T4.2)

- to do more research on social networks
- to get together with other teams working on same/similar subjects
- to develop and test the new prototype
- to perform user studies in order to evaluate its achievements - validate the system
Future work (Odisseia)

- Pursue the “informal way” to “open” channels for scholarly communication to occur and

- to share knowledge and

- to allow information mediators to do their job so that they can

- help to discover the “undiscovered public knowledge”.
Contents

Scope:

- UMinho and DSI
- SDUM and RepositóriUM
- Odisseia
- CRiB
- Tea for Two
- Other Projects
- Future work
End...

- ... so far.

THANK YOU!

analice@dsi.uminho.pt