Amalia
Integrated access to digital data and documents in the humanities and social sciences

Jean-Philippe Magué
26/10/2011
Ecole Normale Supérieure de Lyon

• A High School
• Similar to a university

• Several teams conducting research in the humanities
• Many projects in the field of digital humanities
Context

Digital Humanities at the ENS Lyon

- Digital Humanities
  - eScience for the humanities
  - Using tools in humanities: an epistemologic shift

- ENS Lyon: The Digital Humanities Workshop
  - A place for inter-projects communication
  - Common trainings
  - A Cyberinfrastructure
A cyberinfrastructure for Digital Scholarly Editions

- Primary sources
- Digitization
- Images
- Transcription
- OCR
- Text
- Natural Language Processing
- Encoding
- Publication
- Analysis
Example: the Hyperdonat project

- Terence (195/185–159 BC): an ancient Roman playwright
  - Wrote 6 plays
- Donat (4th century AD): a Roman grammarian:
  - Wrote a commentary of Terence’s plays
  - A school book during middle age and renaissance
  - Never translated in any modern language
- Hyperdonat
  - An electronic edition of the translation
Example: the Hyperdonat project

A standard workflow

Primary sources:
Wesner edition (1902)
Manuscripts

Transcription

Annotations

Publication

XML

TEI

HTML

BaseX

Query

TL.ODT
TF.ODT
DL.ODT
FD.ODT
The hyperdonat project

Manual Management

File exchange through email
Manual application of XSTL stylesheets
File renaming as version management

... Problems, errors, conflicts, ...

A cyberinfrastructure for collaborative corpus management
A cyberinfrastructure

**Required functionalities**

- A central repository
- Authentication and authorization
A cyberinfrastructure

**Required functionalities**

- **Version management**
- **Documents Lifecycle**
- **Workflow**
A Cyberinfrastructure

Using eSciDoc?

<table>
<thead>
<tr>
<th></th>
<th>eSciDoc</th>
</tr>
</thead>
<tbody>
<tr>
<td>Repository</td>
<td>✔️</td>
</tr>
<tr>
<td>Authentication &amp; authorization</td>
<td>✔️</td>
</tr>
<tr>
<td>Versioning</td>
<td>✔️</td>
</tr>
<tr>
<td>Lifecycle</td>
<td>✔️</td>
</tr>
<tr>
<td>Workflows</td>
<td>❌</td>
</tr>
</tbody>
</table>
Amalia

An eSciDoc solution for collaborative corpus management

- Modeling projects workflows
  - Definition of actions
    - Eg. a call to some web service
    - Parameters
    - authorization
  - Definition of chains of actions
- Leverage on eSciDoc
  - Central repository (OAI PMH)
  - Authentication & authorization
  - Versioning
  - Document lifecycle
Amalia

*An eSciDoc solution for collaborative corpus management*

- One size doesn’t fit all
  - A service oriented architecture
Amalia: past, present and future

Past: Amalia v.1

- Designed for the Hyperdonat project
- 6 months of development during an internship
- A GWT application
  - Java framework
  - Uses the Java client library
  - Designed for eSciDoc 1.2
- Web service connection
  - An XSLT processor
  - The BaseX XML database
Amalia: past, present and future

Past: Amalia v.1

• Problems
  – Static organization of data in eSciDoc
    • Plays of Terence
  – Static item types
    • ODT, XML TEI and HTML files
  – Static project workflow
    • ODT to XML TEI
    • XML TEI to HTML
    • XML TEI to BaseX

• No longer maintained
Amalia: past, present and future

Present: Amalia v.2

- New features
  - Let the user design her item types
  - Let the user design her data organization
  - Let the user design and pipe her actions
  - Use eSciDoc to manage resources (configuration files, xslt, css,…)

- A second internship, 6 more months
- eSciDoc 1.3.2
Amalia: past, present and future

Present: Amalia v.2

Structure of the repository

- [Organizational Unit]
  - [Context]
    - [Container]
  - [Project]
    - [Container]
- [Configuration]
  - [Container]
- [Resources]
  - [Container]
- [Data]
  - [Container]
Amalia: past, present and future

- *Present: Amalia v.2*

- Contains the definition of item types
  - The different representations an item may have

```xml
<ItemTypes>
  <ItemType name="hyperdonat">
    <rep name="TEI" ext="xml"/>
    <rep name="HTML" ext="html"/>
    <rep name="ODT" ext="odt"/>
  </ItemType>
</ItemTypes>
```

- Will contains the definition of actions and pipes
Amalia: past, present and future

Present: Amalia v.2

Structure of the repository
Amalia: past, present and future

- Present: Amalia v.2

- Designed to contain resources needed by the actions
  - XSLT style sheets
  - CSS style sheets
  - Configuration files
  - Scripts
  - ...
Amalia: past, present and future

Present: Amalia v.2

Structure of the repository

[ Organizational Unit ]
[ Context ]
[ Container ]

project [ Container ]

Configuration [ Container ]

Resources [ Container ]

Data [ Container ]
Amalia: past, present and future

Present: Amalia v.2
Amalia: past, present and future

**Future**

- Implements actions and pipes
- Make the most of eSciDoc features
  - Rethink the structure of the repository
    - One Organizational Unit per project
  - Use content models
    - Our item types are redundant
  - A better user management
    - Project based roles definition
    - Shibboleth
  - Set up an OAI-PMH repository
    - Define actions to create DC metadata
Amalia: past, present and future

- **TGE Adonis**
  - A grid computing facility for the humanities
- **Deploy eSciDoc and Amalia on that grid**

![Diagram](image)
Conclusion

Our point on view on eSciDoc

• Issues
  – Complex
  – Slow learning curve
  – Distributed documentation!

• Strengths
  – Incredible number of features bundled in one piece
  – Fits perfectly in a SOA
  – Central piece of our cyberinfrastructure