

# Europeana

*Technical Review Europeana Technology Platform*



**europeana**  
think culture

Capgemini  
**December 13<sup>th</sup> 2010**





# Agenda

## Introduction: our Capgemini approach

- Understanding the question
- Initial findings and ideas
- Next steps



# What we DID and DID NOT do

- **What we DID do**

The findings in this report are based on 3 different input sources:

1. An conference call with Jill Cousins - Director of Europeana Foundation and The European Library
1. Extensive desk research into the documents and website references delivered by Jill Cousins (Description of work ECP 2008 DILI 558001, Vancis hosting proposal, Vancis received hosting proposal) + a number of website references. For this the Capgemini evaluated this inputs using the standard Capgemini checklist (see annex 1)
2. Based on these inputs, an extensive interview was conducted with Jan Molendijk, Technical and Operations Director of Europeana



# What we DID and DID NOT do

## ■ What we DID NOT do

1. Our standard Capgemini approach in these cases is to conduct a number of interviews/workshops with relevant stakeholders, both on the business/client side as well as on the technology side. For budget reasons we limited ourselves to the two interviews, one with Jill Cousins and another one with Jan Molendijk
2. Our standard Capgemini approach is to conduct a number of *on premise* technical stress tests in order to acquire first hand management information about the systems performance under duress. In this instance we have taken the input from the two Vancis proposals and the verbal clarification of Jan Molendijk - Technical and Operations Director - on the performance of the systems as a basis for our assessment and have not conducted any hands on testing ourselves
3. We have not been able to look into the specifics of the software production process since no valid data were made available to us



# Agenda

- Introduction

## Understanding the question

- Initial findings and ideas
- Next steps



# Understanding the Question

- From our information, these are the major questions to be answered:
- the Comité would like to be certain that the technology on which Europeana has been developed can be relied upon for further developments and extensions. The Comité has the following questions in mind :
  1. Is the technology used to develop European a state-of-the-art and up-to-date technology or is there a risk that it will be rapidly outdated by technological evolutions ?
  2. Is this technology able to deal with a dramatic increase in terms of the number of objects indexed ? Is this technology able to deal with a steep increase in the number of unique visitors ?
  3. What would be your main recommendations in order to make sure that Europeana has the most efficient technology and organization in order to deliver its mission ?

The volumes for the current system are calculated at 4000 concurrent users and to allow for 1 million unique visitors per week. Since the launch of the prototype no marketing has been done towards end users. This will recommence now and we expect to build to around 500,000 unique visitors per month by April 2011.



# Agenda

- Introduction
- Understanding the question

## **Our findings and suggestions**

- Next steps



# Our findings and suggestions

- *Is the technology used to develop European a state-of-the-art and up-to-date technology or is there a risk that it will be rapidly outdated by technological evolutions ?*
- Based on our study of the separate documents delivered to us and both the interviews conducted, it is our understanding that the Europeana project is based on a robust technology foundation that includes Open Standards, architecture, security. We suggest to look into the possibilities of cloud computing (see next page)
- *Is this technology able to deal with a dramatic increase in terms of the number of objects indexed ? Is this technology able to deal with a steep increase in the number of unique visitors ?*
- To answer this question we would rather have done our own hands on stress testing. Based on our desk research into the two Vancis documents and the interview with mr. Molendijk, it is our perception that the system should be able to deal with 500.000 unique users per month as planned. Yet this is something that Capgemini would rather test hands on by our own standard testing protocols. We expect that there is room for Database Optimization, but we can only judge that after thorough testing and hands on investigation



# Our findings and suggestions

- *What would be your main recommendations in order to make sure that Europeana has the most efficient technology and organization in order to deliver its mission ?*
- The absence of an overall **Software Architecture Document (SAD)** describing the current and the future overall architecture of the platform is a crucial limitation. Most ingredients appear to be there. They need to be collected and written down into one coherent document that is accepted by both the Europeana technical team as well as the representatives of the business side.
- The absence of **Software Measurement Metrics** - although we have not found data that would suggest the software development process is underperforming, the absence of solid benchmark data makes it quite difficult - if not impossible - to verify the overall quality of the software development process in terms of productivity
- In absence of an overall SAD, the (future) possibilities of **cloud computing have not carefully been assessed**. Our view is that cloud computing is an important overall IT trend that should be incorporated into Europeana's future architecture and therefore described in the SAD already mentioned. Cloud is more than just virtualization (Infrastructure as a Service), it also encompasses Platform as a Service and Software as a Service. The latter two are even more relevant to Europeana than the IaaS dimension of cloud technology which appear to be the prevailing perception of cloud technology within the Europeana team. See the last slide with our suggestion for a "journey to the cloud" workshop



# Agenda

- Introduction
- Understanding the question
- Initial findings and ideas

## Annex



# Checklist Questions (1)

- **Software Architecture Document:**
  - How easy is it to upscale when data and usage increase?
  - Analytics: what do we know about current performance and stress test results?
  - How can future growth be simulated / predicted / reviewed?
- **Infrastructure:**
  - Is current set-up (with load balancers and firewall) prepared for easy upscaling apps/db servers?
- **Software / applications / business logic:**
  - Search caching: possible?
  - API strategy: how will it impact performance?
  - Software architecture set-up to minimise database performance?
- **Content & data:**
  - Image resizing / rendering – why and how? What is impact on server performance?
  - Multi-linguality: how has it been set-up?
  - Taxonomy: who owns this?
  - Document oriented database – already analyzed?
- **User Experience:**
  - Next to API-strategy, can search be offloaded with fixed structure (fixed/cached queries)?
  - Social: what is the roadmap for social media usage and UGC?



# Checklist Questions (2)

## ■ Questions

- What are current (known) architecture bottlenecks for technology platform upscale requirements?
- Current stress-test execution: what are the results and what are the test scenarios?
- Database: what are the ideas with replacing current db with a document oriented version?
- What are the ideas with the destination sites?
  - Guided navigation to decrease amount of queries?
  - Fixed queries?
- What will be the API strategy (how are musea involved?) and how is it planned with the roadmap for the destination sites? Are there any predictions for impact on site traffic?
- What is the roadmap for user generated content (we understand that some experimentation is done) and use of social?



# Topics as discussed on 16-11 with Jan Molendijk

- Productivity: software development productivity is not measured, but expected to be good. Learnings from factory approaches are welcome. The current model for Europeana is not a factory model, but more a collaborative approach between multiple project partners (musea, universities, TNO, etc.). Standards and frameworks for data enrichment and quality assurance are installed and working.
- Optimization on database server is running at the moment, possible performance improvement with document based model.
- Currently there is no SAD, but most ingredients are there. It has not been collected and written down into one document.
- Performance testing is executed on a regular basis, and new scenario's can be tested thoroughly based on logfiles.
- The API approach is used already with best practice examples, but also with examples that cause a lot of load. For this, guidelines are under development to optimize.
- Family sourcing is used for translations.
- Social monitoring is partly implemented by the Marcom team.
- Semantic search / natural languages search will be implemented in the next versions.
- Sugar CRM is used for CRM towards partners.
- There are doubts on a positive business case for cloud. Discussing the impact on a strategic level seems not effective at this stage. However, it might be good to discuss future possibilities from the content council perspective (discuss specific challenges and opportunities and how to solve them with all partners involved).



# Journey to the Cloud Engagement

