An Open Cultural Digital Content Infrastructure

Ioanna-Ourania Stathopoulou, Haris Georgiadis, Vangelis Banos, Panagiotis Stathopoulos, Nikos Houssos, Evi Sachini

National Documentation Centre / National Hellenic Research Foundation
Athens, Greece
Outline

• Objective and scope
• Approach and design choices
• Architecture and implementation
• Related work
• Status and future work
Context and objectives of the activity

• The funding environment
  – More than 70 digital cultural heritage projects / about 60 million Euros
  – Co-funding by Greece and EU structural funds
• Assist the funder to ensure the availability and quality of project output
  – Availability of the produced material (metadata and digital files) at a central, secure, enterprise-grade infrastructure
  – Infrastructure and mechanisms to check the quality of metadata and digital content generated through the projects
A suite of services for repositories

• Harvesting and aggregation of content
  – Single point of access to content / unified search and browse

• Validation (metadata + digital files)
  – Largely automated checking of compliance with specifications

• Safe-keeping of digital files in the highest available quality
Why validation?

• Problems identified in project output of past funding programmes (indicative list):
  – Inadequate quality of metadata records – use of custom data models and formats instead of standard schemata
  – Poor digitization quality, lack of basic features like OCR for text material
  – Lack of standard programmatic interfaces for continuous access to the material
  – Inadequate infrastructure to ensure availability and safekeeping

• Result: Low reuse and return on investment
Overview of the validation approach

• Publish interoperability specifications in advance
• Associate successful validation with funding cash flows
• Validate multiple times throughout project - first during initial stages to provide early feedback
• Validation of both metadata and digital files
• Validation is largely automated – essential for scalability, feasibility and sustainability
• Validation of live systems
  – The material is validated by directly retrieving it through programmatic interfaces (e.g. OAI-PMH)
• Modular and extensible validation infrastructure – gradual support of multiple schemata, formats
Interoperability requirements

• Specifications have been published in advance, before the beginning of the projects
• Available at http://hdl.handle.net/10442/8887
• Cover interoperability at the system, syntactic/structure, semantic levels
• Availability of the metadata as linked data is mandatory
• The funded institution is required to provide also information such as the detailed specifications of the primary cataloguing schema used, mappings to standard formats, the controlled vocabularies / thesauri utilised
System architecture
Data model and implementation aspects

• The aggregator uses the Europeana Data Model (EDM) as the main data model and metadata schema

• Infrastructure implemented using certain components of the Europeana open source software infrastructure

• A range of utilised technologies: Java, Python, REST APIs for interactions among components, Solr, MongoDB, PostgreSQL
Validation logic architecture and implementation

• Two interacting components
  – Front-end and back-end
• Front-end provides UI for authorized users to carry out validations / get results / produce reports, back-end provides the core validation logic
• Modular and extensible scheme for the definition and insertion of rules in the system
• Validation Domain-Specific Language (VDSL)
  – Rules and rule-sets, boolean operators, control flow
  – Simple JSON format
• Gradual support for checking compliance with various schemata and formats
Comparison with related work

• Validators exists in various international systems (e.g. OpenAIRE, ARIADNE)

• Distinct features of our approach:
  – Validation of both metadata (records, controlled vocabularies) and digital files
  – Extensive validation at the semantic level
  – Modularity and extensibility allows combined validation along multiple dimensions (e.g. check availability of metadata in multiple formats)
  – Validation Domain Specific Language
  – Support for connecting validation with administrative procedures in a decoupled fashion
Status of implementation and operation - further work

- The infrastructure is in production since Spring 2014
- Several validations have been already completed with the infrastructure
- Next steps:
  - Provide a user interface for repository managers to perform test validations before actually submitting their content
  - Refine validation rules and aspects of their implementation
  - Public operation of the aggregator portal
Acknowledgements

• The work presented in this article has been partly supported by the project
• "Platform for provision of services for deposit, management and dissemination of Open Public Data and Digital Content" (Ref No 327378)
• Co-funded by Greece and the European Union-European Regional Development Fund through the Operational Programme "Digital Convergence" (NSFR)
Thank you for your attention!

• More info:

http://www.epset.gr/en/SaaS_Services

iostath, hgeorgiadis, vbanos, pstath , nhoussos, esachin AT ekt.gr