ESCIDOC
- A SERVICE INFRASTRUCTURE FOR CULTURAL HERITAGE CONTENT

N. Bulatovic [a*], U. Tschida [a], A. Gros [b]

[a] Max Planck Digital Library, Research & Development, 80337 Muenchen, Germany – (bulatovic, tschida)@mpdl.mpg.de
[b] Max Planck Digital Library, Digital Collections, 80337 Muenchen, Germany – gros@mpdl.mpg.de

KEY WORDS: Service-oriented architecture, Digital Repository, content model, semantic relations, persistent identification, preservation, eSciDoc, Dariah

ABSTRACT:

The eSciDoc project (http://www.escidoc.org) provides an open source infrastructure for scientists to store, preserve, retrieve and work with research-based data, including cultural heritage content. Since multiple requirements demand a broad set of re-usable, generic, and flexible components, it is designed and developed as a service-oriented architecture. In addition, community-specific solutions built around the services allow user- and research-focused working environments. The heterogeneity of research questions, tools, workflows, and primary data, as well as traditional forms of publication required us to focus on supporting multiple content models and descriptive metadata formats, together with common functionalities such as persistent identification, adequate versioning and management of primary data, aggregation of data, annotations, and access control. The underlying Fedora repository, enriched with the eSciDoc core services layer, allows for management and structuring of data at different levels of granularity - from basic items to complex aggregations of data. In the course of the DARIAH project, our institution and our partners aim at building a pan-European research infrastructure based on Fedora/eSciDoc. This infrastructure will enable transparent access to and management of digital content across all participating museums, libraries and other institutions in the arts and humanities sector. In our paper we will focus on the benefits of eSciDoc's data and content modelling and of building a sustainable technical infrastructure. In addition, we will describe the current project approach and the current requirements regarding cultural heritage content.

1.INTRODUCTION

eSciDoc (http://www.escidoc.org) as a joint project of the Max Planck Society (MPG) and FIZ Karlsruhe, is funded by the German Federal Ministry of Education and Research (BMBF). The aim of the project is to provide a scholarly communication platform for multiple disciplines, developed as open source to be re-used by any interested research organisation.

Due to the challenges of emerging eScience scenarios, services, solutions and tools developed will have to handle research results as well as primary data, in order to support distributed and interdisciplinary research questions. Designed in a modular and complementary way, the infrastructure will provide effective, reliable and comprehensive access to data and information. The solutions on top are designed as user-centered working environments, focusing on a specific research scenario and serve as tools for visualising, managing, publishing and enriching their artefacts. (Dreyer et al. 2007)

2.PROJECT APPROACH

The 80 institutes of the Max Planck Society act as main stakeholders in the development of the eSciDoc project. Given the variety of disciplines and hence the multiplicity of data formats and information entities, the set of institutes provides a vivid panopticum of current and latent research scenarios, and poses challenges to introduce corresponding support of information technology.

In a close and ongoing exchange with the institutes, our institution is actively monitoring and identifying current requirements from scientists, which vary from publishing digitized artefacts to adequate solutions for handling the scientific lifecycle of experimental data. Understanding the main research questions behind a scientific information lifecycle, identifying the level of dynamic interaction envisioned by scientists that is required as a basis for analyzing artefacts, and knowing about their formats and the processes within the scientific information lifecycle, that is the starting point for being able to identify generic components and processes for eSciDoc. Another important aspect is business process modelling, i.e. to describe organisational aspects, which is interdependent with identifying services and solutions needed, i.e. technical aspects.

This approach supports the overall project development, as with each new solution we gain insight into generic requirements for the infrastructure and into necessary "add-ons" that are specific to disciplines or certain types of research data.

Therefore, the complete development life cycle facilitates reflection and buildup of know-how on open and sustainable formats, semantic relations between artefacts and their constituent parts, as well as aspects regarding authorisation, persistent identification, data curation and long-term archiving.

3.REQUIREMENTS FOR CULTURAL HERITAGE

3.1 Context Max Planck Society

Within the MPG, several institutes are hosting distinguished holdings of digitized cultural heritage collections. The *Max Planck Institute for the History of Science* (http://www.mpiwg-berlin.mpg.de/) has done important work in organizing and providing a platform for more than 70 seed collections on European cultural heritage collections (ECHO http://echo.mpiwg-berlin.mpg.de/home). The *Kunsthistorisches Institut Florenz KHI* (http://www.khi.fi.it/) and the *Bibliotheca Hertziana* (http://www.biblhertz.it/default.htm) in Rome provide important holdings in their photo libraries as well as digitized engravings or architectural drawings. The *MPI for European History of Law* (http://www.mpier.uni-frankfurt.de/) provides digitized historic journals.

Most of the institutes have already undertaken considerable efforts – on their own or together with partners – for digitization of their material, to some extent in high-resolution quality, and to some extent even within very sophisticated viewing environments. However, most solutions are lacking sustainable and interoperable environments to store and extend their collections. The provision of reliable and open access to valuable resources largely depends on funding, personal enthusiasm and technical expertise, which is not easy to find and retain. The transition from mere publishing to provision of interactive working environments is often limited up to now with respect to annotating, semantic linking, individual compilations or adding new media types.

The eSciDoc project tries to address these aspects by providing an infrastructure which can be used to support and/or substitute locally developed solutions. Consequently, the eSciDoc infrastructure and its services can be used for cooperative working environments and management of content, as well as a mere archive solution for content managed in external systems. In addition, some services can be integrated into external applications, independently from the overall eSciDoc infrastructure. Staying agnostic to data, technologies, and data structures in the design of the logical infrastructure enables us to react quickly to new content, data formats, and functionalities.

Thus, the aim of the project is not only to provide solutions for cooperative working environments, but in addition to create a stable infrastructure to advance accessibility, dissemination, as well as re-purposing and mashing-up of data across disciplines.

3.2International context - DARIAH

The DARIAH project aims at creating a pan-European research infrastructure for the arts and humanities. DARIAH intends to provide seamless access to distributed data holdings across Europe and aims at enabling integration into national research infrastructures. In this respect, eSciDoc represents a core reference technology for archiving, managing and disseminating scientific publications, primary sources and secondary documents in the humanities. This in turn poses several requirements on eSciDoc. Firstly, eSciDoc has to provide comprehensive access control mechanisms across distributed data holdings, e.g. via Shibboleth. Secondly, resources in the arts and humanities can be highly diverse with respect to content and data formats and are often interlinked. Therefore, a content and relation model that allows capturing scientific and scholarly work processes as well as modeling their complex webs of documents and data types has to be provided. Thirdly, DARIAH intends to capture also the semantic connections between the digital objects within the connected repositories. Hence, eSciDoc has to provide interfaces for services that link between representations of repository object models and corresponding ontologies. Fourthly, eSciDoc has to be interoperable with national research infrastructures. One crucial predicament about offering a corresponding interoperability service layer for external initiatives is the agreement about – and utilization and documentation of – standards for data-exchange and access.

3.3 Current eSciDoc solutions in the context of cultural heritage

The solution VIRR ("Virtueller Raum Reichsrecht") is an example for the common scenario of publishing the corpus of digitized artefacts like images and texts. VIRR provides a digital compilation and working environment for various artefacts of the period of the Holy Roman Empire, and is developed together with the MPI for European History of Law. (Please visit http://colab.mpdl.mpg.de/mediawiki/ViRR:_Virtueller_Raum_Reichsrecht for more information. The first prototype is available under http://faces1.mpdl.mpg.de:8080/virr_presentation/). In the first phase we focus on the publication of the artefacts. Structural navigation through the scanned textual resources (collection – multivolume – volume – chapter – page) with respective metadata is provided; the current viewing environment will be improved, the improvement will be based on the DigiLib* tool. In a second phase we will focus on developing an interactive working environment, while the collection will be indexed, edited, and enlarged cooperatively by scientists and trained staff.

Research focused on text-based publications requires the possibility of enriching publications with information on text-inherent structures. We therefore have to support generation and storage of XML transcriptions of the fulltext (e.g. TEI). For an adequate working environment, the user needs to have the option to combine the viewing environment with an editing environment (e.g. zooming in on a part of a scan and enriching the focused part with annotations). In this context, versioning, persistent identification, and the possibility to annotate publications and their constituent parts must be supported.

Complementary to VIRR, the solution FACES offers a web-based collection of image data, containing facial stimuli, and is developed in conjunction with the MPI for Human Development. (Please visit http://colab.mpdl.mpg.de/mediawiki/Faces for more information. The initial prototype is available under http://faces1.mpdl.mpg.de:8080/faces/). The images have corresponding metadata and attributes and can be used for individual or project-specific research questions, for example by filtering and sorting them into individual subsets, which can be published and shared. As the images are the basis for multiple

* DigiLib is a web based client/server technology for viewing and working with images. This open source software was jointly developed by the Max-Planck-Institute for the History of Science, the University of Bern and others. More under http://colab.mpdl.mpg.de/mediawiki/Digilib

research questions (e.g. rating studies in the context of human development or neuro-scientific research), the collection might be extended with new images, and images might be enriched with new metadata and attributes. Appropriate authentication and authorization mechanisms have to be provided to support potential legal constraints. In addition, the viewing environment must enable detailed analysis of the facial stimuli. In the meantime, the solution has raised interest at some other institutes with a need for an image-handling solution (and with quite diverse content, such as digitized photographs or microscope images).

4.THE ESCIDOC INFRASTRUCTURE

4.1Data and content modelling

Understanding the structure and the nature of the content resources was essential for the ability to meet the emerging requirements within the project. The first step was to define a general data model to support functionalities such as versioning, persistent identification, relating and annotating resources, and authorization. The second step was to refine the model by analyzing various types of resources to support characteristics specific to certain disciplines, institutions or resources. (Tschida, Bulatovic 2008) Content resources are defined by two generic object patterns: Item and Container. An Item resource consists of metadata records (e.g. SISIS MAB record, MODS record, DublinCore record) and optionally of components that represent the actual content (e.g. PDF file, JPEG file, XML file). A Container resource is an aggregation of other resources that allows grouping of other items or containers. Like the Item resource, Container can be described by multiple metadata records. In addition, each resource is maintained in a single administrative Context. Context resources are created by organizations (e.g. a project group) in accordance with their needs to express rules for content creation, update, quality assurance of metadata, dissemination, preservation, authorization policies, submission policies, etc.

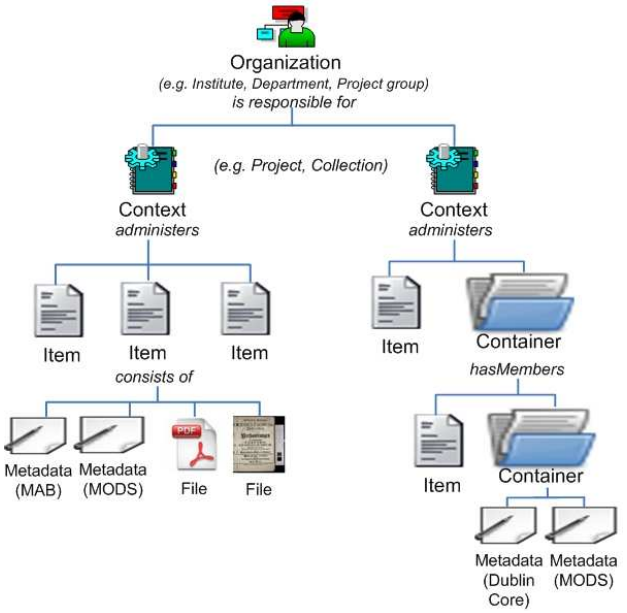

Figure 1: eSciDoc data model

Content models are formal representations of discipline-specific data models such as an integrated image and text view of primary sources or a precisely documented collection of images. For example, a digitized book can be expressed as a container of book page items and related transcription items. The book container has bibliographic metadata based on the MODS metadata schema. Each page item consists of the digitized image of the page and a metadata record. The metadata record is inherited from the book container metadata. In addition, it has metadata containing the page number (e.g. 1,2, 3, 4 or I, II, III, IV), chapter information etc.

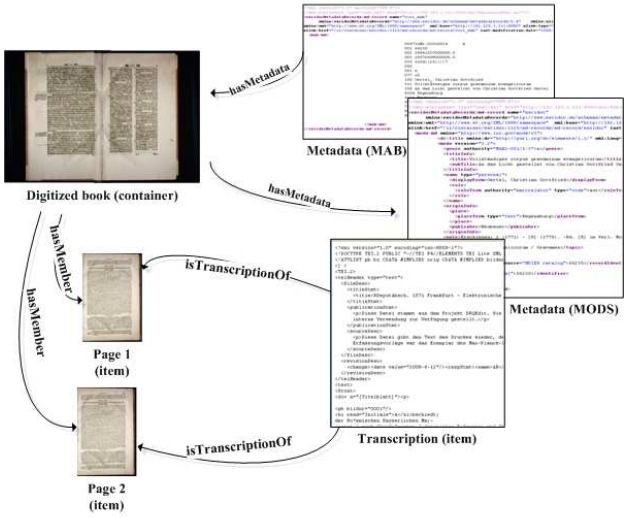

Figure 2: Visualization of the Digital Book content model

Relations between resources (e.g. structural relations within collection, kinship relations) are defined with respective relation ontologies. They are applicable to any kind of content managed in the digital repository. The content model defines which relation ontologies (and types of relations) can be used to relate the resources. For example a digitized book allows relations such as hasMembers (between book container and each page item), isTranscriptionOf (between transcription item and a page item).

By applying these mechanisms the eSciDoc infrastructure is able to easily adopt new types of resources in a standardized manner. In addition, this approach fosters the exchange of data and the reuse of resources for different purposes.

4.2Services and service layers

The eSciDoc infrastructure is designed as a service-oriented architecture (SOA). Services are grouped into three service layers: core services, intermediate services and application services.

Core services implement basic CRUD operations on resources managed by the infrastructure. In addition they support the basic lifecycle of resources by a set of task-oriented operations such as submit, release or withdraw. Core services add value as they implement the functionalities such as versioning and persistent identification. Services involved in this group are Object Manager (Item, Container and Context handlers), Organizational Units Handler, Semantic Store Handler, Role Manager, Content Model Handler and User Account Handler.

Intermediate services represent both a service requestor and a service provider within the eSciDoc SOA. They act as adapters and façades to the basic services or provide additional functionality. Intermediate services can additionally manipulate their own data. Services involved in this group are Validation Service, Duplication Detection, Technical Metadata Extraction, PID Handler, Statistics Manager and Digilib.

Application services may combine services from all layers and implement business logic from a solution-specific domain. They are candidates for future process-centric services enabling service orchestration. Services involved in this group are Depositing Service, Publishing Service, Citation Style Manager, SearchAndIndexing, SearchAndOutput and Export Manager.

Authentication and authorization. The infrastructure implements distributed mechanisms for authentication and authorization. It is based on Shibboleth* and XACML** to define authorization policies. This approach enables integration with external identity management system such as LDAP. Service requests are protected by a security interceptor which forwards the request further to the Policy Decision Point (PDP) engine. The PDP engine evaluates the requests and authorizes clients, or prevents clients from performing the service operation requested. Policies can be defined for various levels of resource granularity, e.g. Organizational Unit, Context, Item, Container or File.

Persistent identifiers. Resources can be persistently identified at different levels of granularity. An item that consists of an image and metadata records can be persistently identified as a single resource. In addition, the image itself can be persistently identified. The infrastructure allows for assignment of persistent identifier to a resource at different stages of the internal lifecycle of a resource: during creation, modification or publishing. A persistent identifier can be assigned to a selected version, or to all versions of a resource separately. The choice of a persistent identification system (PID system) to be used is left to the user, i.e. the infrastructure is designed to provide an interface to a PID system preferred by a user. By now support of the Handle System (see http://www.handle.net/) has been implemented.

4.2.1 Example: Validation Service

Content models define the overall structure of the resources and the metadata schema that can be used to describe them. Validation Service goes a step beyond and allows for the definition of additional sets of constraints a resource must conform to in a specific context and at a particular stage during its lifecycle. Examples of such constraints are:

a) A transcription cannot be created without at least one author
b) A transcription cannot be published if the page item is unpublished.
c) If a publisher name is given, the publication year must be given as well.
d) If a resource is a journal article, the journal name must be added.
e) All publications of a project group collection must have at least one author affiliated to that project group.

Validation Service introduces the concept of a validation schema to group a set of constraints that belong together. The validation schema includes validation points to define when a certain constraint should be checked (e.g. during creation, for a user-defined event or when publishing a resource). Validation Service allows for multiple validation schemas for a content model. This enables different organizations to have different criteria for, for example, the quality of the metadata while still reusing the content models defined once. For instance, project group X may define constraint b) from the example above as restrictive during content creation. Project group Y may completely ignore this constraint. In this sense, project group X and project group Y work with different validation schemas for transcriptions. The selection of a validation schema to apply is a matter of configuration of the administrative Context (see paragraph 4.1) within the eSciDoc infrastructure.

At present two types of constraints are supported: informative or restrictive. For input it accepts a representation of a resource and produces a validation report in XML format as output.

If during the validation process a constraint is not fulfilled, the service provides an exception message with an identifier and an XPath pointer (see http://www.w3.org/TR/xpath) to the exact position where validation failed. The use of the identifier for exception messages enables providing user-friendly (internationalized) messages to end users via a graphical user interface.

Validation Service can be used completely independently from other services in the infrastructure. Internally it uses Schematron (http://www.schematron.com/) for definition of validation schemas. The service is offered via three interfaces: EJB3 (Enterprise Java Bean), SOAP and REST.

4.2.2 Example: SearchAndOutput

SearchAndOutput is a service that supports searching and exporting search results in a specific format, such as Endnote, APA citation style, eSciDoc native format enriched with a citation style output, or comma separated values (CSV) format. It enables, for example, exporting a digital collection of FACES images (at a resolution selected) together with their metadata in a single request. It is built as a service composition of three other services: SearchAndIndexing, Citation Style Manager, Export Manager.

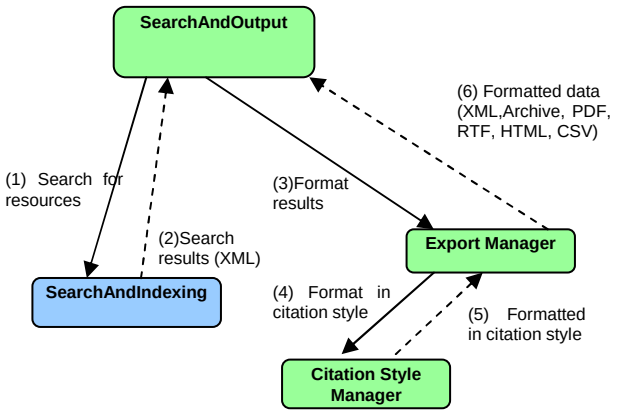

Figure 4: SearchAndOutput and service composition

* see http://shibboleth.internet2.edu/ for more details
** see http://wiki.oasis-open.org/xacml/ for more details

Services are called in the following manner: A client issues a service request to SearchAndOutput specifying its search criteria, desired export format (APA style, EndNote format, CSV format), desired output format (PDF, HTML, RTF) and optionally an archival format (zip, gzip). SearchAndOutput invokes the search operation of SearchAndIndexing with the specified search criteria and output format parameters. Once it has retrieved the search results, it sends them to Export Manager (including the desired output format) as parameters. Export Manager formats the results in accordance with the parameters conveyed, and sends the formatted data back to SearchAndOutput. The latter then forwards the formatted data to the service requester. Optionally, Export Manager may send a request to Citation Style Manager (if the data is requested to be formatted in a defined citation style). The service interface of SearchAndOutput is available via REST and EJB3 interfaces.

While Export Manager and Citation Style Manager can be used as completely independent services, SearchAndOutput is loosely coupled with the SearchAndIndexing. To relate it to another search service one needs to change the configuration properties and provide a transformation of input data into a format (XML) supported by SearchAndOutput.

4.3Technical architecture

The eSciDoc infrastructure is based on the FedoraCommons (http://www.fedora-commons.org/) platform and other open-source software packages, such as PostgreSQL, JBoss application server, Lucene, Tomcat, JHove. The FedoraCommons platform provides the repository, preservation and semantic services the eSciDoc infrastructure services are built on. The eSciDoc infrastructure adds value by contributing object patterns, a content model, a versioning model, implementation of the basic lifecycle of content resources, referential integrity, interfaces to persistent identification systems, fine-grained access control as well as the possibility to define new relation ontologies on top of the Fedora base relation ontology.

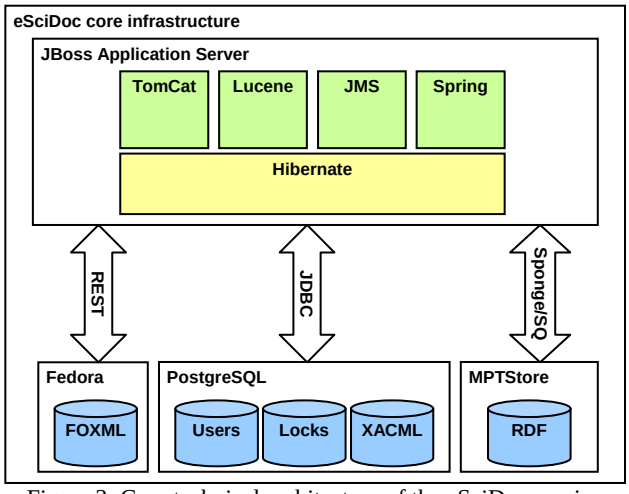

Figure 3: Core technical architecture of the eSciDoc service infrastructure

Figure 3 depicts the core technical architecture of the eSciDoc infrastructure. Being designed as a service-oriented architecture, the infrastructure does not prescribe the technology in which a new service or a solution can be developed. Service operations are offered via REST and/or SOAP interfaces (depending on the service). Some services are exposed as Enterprise Java Beans (EJB3) to accelerate the internal development where necessary. However a new service or solution that needs to use other services can be added independently of the programming language in which other services are built. Also the internal data structures of a programming language are not a limiting factor since the service input and output messages are formatted as XML data. Services within the eSciDoc infrastructure are built keeping this notion of independence from specific technologies in mind, even when they communicate among one another.
Another advantage of the service-orientation approach is the possibility to modify the service implementation in a process transparent to service users. This is especially useful when adding new functionality or extending the existing one. Service users need not be overburdened with the internal implementation of the service. They may rather focus on the functionality offered by the service, on input and output service messages, i.e. on service input and output data.

5.DIGITAL PRESERVATION ASPECTS

Preservation of digital content is a basic requirement that has to be fulfilled by any system managing this type of content. The eSciDoc infrastructure has been designed and implemented as a system that strongly emphasized issues related to the digital preservation.

- Internal format of data: All content resources are stored as Fedora FOXML documents. Thus the structure and description of each resource is readable by any human or machine user.
- Contextual information on what the data actually represents, for example, a publication or transcription of a scanned book page, is provided with a content model. Each resource references a content model resource. A content model is itself defined as a FOXML document.
- Resources are validated in accordance with rules that are not defined by the implementation logic but described by Schematron which is itself an ISO standard (ISO/IEC 19757-3:2006 2006).
- Even though the eSciDoc infrastructure does not impose a metadata schema in use, it strongly recommends using open metadata standards wherever possible. As a minimum, content resources have a Dublin Core metadata record associated with them.
- Persistent identification of resources is eventually enforced when the resource is published. Users may decide what to identify (resource or resource version) and when to assign the persistent identifier (during creation, updates, or publishing). Additionally, if a resource already exists on an external system and if it has a persistent identifier assigned, the eSciDoc infrastructure keeps track of it.
- Technical Metadata Extraction Service based on the JHOVE - JSTOR/Harvard Object Validation Environment extracts the technical metadata from a set of text, audio and image formats (see http://hul.harvard.edu/jhove/). In addition it creates a PRONOM identifier for files (see http://www.nationalarchives.gov.uk/pronom/) associated with the content resources.
- Each resource is associated with PREMIS metadata (see http://www.oclc.org/research/projects/pmwg/) that holds reacher information on the version history and events in addition to the natively provided Fedora audit trail.
- All software components used are open source. Software delivered by the eSciDoc project is open source itself, and

is available under the CDDL license (see http://www.escidoc.de/license). Documentation is also available to the public.

Digital preservation becomes increasingly complex when it has to deal with the management of content that is continuously updated. This is especially true for research and collaborative environments, where there is a tendency towards publishing results as early as possible. The eSciDoc infrastructure defines a basic workflow for all content resources, to enable better selection and general quality of resources that need to be preserved. In this manner it supports the OAIS reference model (OAIS 2002).

By using open standards and open source components, the eSciDoc service infrastructure benefits from community developments in the area of digital preservation. The work in general is not finished and there are many initiatives in the arena.

6.CONCLUSIONS

The provision of digital artefacts relevant for the cultural heritage requires an understanding of their potential of being re-used and re-purposed. Publishing digitized resources is not the aim, but the medium to facilitate new research and discourse – however, only if relevant structural, semantic, and cognitive entities can be modelled and described in a machine-readable way. We are approaching the development of a content model ontology and corresponding knowledge bases of resource descriptions. That allows for interoperability between various repositories, i.e. internal object models and the cross-validation of related resources. In that context, we will have the first proof of concept during the DARIAH project, where the eSciDoc infrastructure will serve as demonstrator platform.
In addition to sustainable data and content modelling, the requirement for sustainable technical infrastructures is of high interest in the humanities to avoid “buried treasures” and isolated collections in an increasingly interdisciplinary and globalized environment. Furthermore, the advance in virtual organisations leads to increasingly sophisticated requirements for handling users and their respective rights to access and enrich shared artefacts.
We will continue to work on solutions, services and tools to support sustainable access to the data holdings of the Max Planck institutes, and we are convinced that the overall design facilitates the support of multiple types of cultural heritage content.

7.REFERENCES

Dreyer et al. 2007. eSciDoc – a scholarly information and communication platform for the Max Planck Society. Conference-Paper for the GES 2007, Baden-Baden, Germany. http://www.ges2007.de/papers/ (accessed 24 Aug. 2008)

Tschida, Bulatovic 2008. eSciDoc – a flexible infrastructure for management and storage of cultural heritage data. Abstract for Poster session at TEI Member Meeting 2008, London. http://edoc.mpg.de/367779 (accessed 24 Aug. 2008)

OAIS 2002. Reference Model for an Open Archival Information System (OAIS). Consultative Committee for Space Data Systems. http://public.ccsds.org/publications/archive/650x0b1.pdf (accessed 24 Aug. 2008)

ISO/IEC 19757-3:2006 2006. Information technology — Document Schema Definition Languages (DSDL) — Part 3: Rule-based validation — Schematron. http://standards.iso.org/ittf/PubliclyAvailableStandards/index.html (accessed 24 Aug. 2008)

8.ACKNOWLEDGMENTS

We would like to thank to all eSciDoc project members for their work on enabling this paper. We would also like to thank Erik Altmann for assisting with the editorial work.