Cultural Heritage: From Library Shelves to Network Residents

Friederike Kleinfercher, Kristina Koller
Max Planck Digital Library
kleinfercher@mpdl.mpg.de, koller@mpdl.mpg.de

Abstract: In the context of the eSciDoc project, the Max Planck Digital Library and the FIZ Karlsruhe are building an e-research environment for multi-disciplinary scientific research organizations. Based on the eSciDoc infrastructure, several solutions for the end-user will be developed and provided as open source software. One of them is ViRR (Virtueller Raum Reichsrecht), a solution to support collaborative and interdisciplinary research on text resources like manuscripts or books. A user-centred approach was applied to define necessary functionalities and adequate graphical user interfaces. ViRR provides several smaller flexible tools in one web interface for the creation and enrichment of metadata, for the modelling of the structure of a work and for the enhancement of the collection with related resources such as annotations and transcriptions. One of them is a configurable online editor for defining the structure of the digitized work in accordance with the structure of the original resource.
This paper will give an overview of the ViRR solution which was developed to support researchers from different backgrounds working together on text resources. Additionally, we will outline eSciDoc, the underlying infrastructure of the ViRR solution.

Keywords: eSciDoc, digitized text resources, collaborative workbench, online editor

1. Introduction

In the context of the eSciDoc project (http://www.escidoc.org) the Max Planck Digital Library (MPDL) has developed a web based solution for different user groups (researchers, librarians) to make their textual holdings online available. ViRR [1] enables the enrichment, dissemination and preservation of digitized cultural heritage like manuscripts or books. Its aim is mainly to support scholars in the humanities in the analysis and evaluation of text resources.
The MPDL is a scientific service unit within the Max Planck Society (MPG), which consists of about 80 institutes from various scientific disciplines, and therefore the development of services and solutions has to deal with requirements from diverse research contexts. During the development of the ViRR solution, the general approach was to start with specific requirements from a pilot community, and then identify generic services, which can be re-used by other disciplines. The aim is to develop a solution which can fulfil most of the diverse requirements of working with digitized text resources within the MPG.
The name ViRR derives from the content of the first collection, which consists of about 20.000 scans of legal artifacts from the period of the Holy Roman Empire provided by the Max Planck Institute for European Legal History (http://www.mpier.uni-frankfurt.de).

2. Working with Digitized Text Resources

Solutions, which support scholars in their work with digitized text resources, differ in their focus and quality as working instruments. The very basic level is the mere digital representation of a single text resource with basic browsing functions and without any sophisticated user management or re-use options.
A more enhanced level offers functionalities to intellectually enrich digitized text resources. Hereby, the scholars and librarians are able to uncover the “hidden” information, which cannot be provided by a mere digital representation. Some of these functionalities imply the

capturing and enhancement of structural metadata and semantics, ideally in different standard formats like METS [2], MODS or TEI. Detailed information about the composition of a resource might be gathered, such as the pagination (logical and physical) or the structure of a work (see e.g. [3, 4]). Standardized interfaces support the re-use of this additional information in other contexts, such as library catalogues, aggregated viewing environments or mash-up services, and allow the integration of external knowledge bases, such as dictionaries or viewing tools.

Having the resources and the related information on the web, the logical consequence is the support of collaborative scenarios from various disciplines, which might assist the creation of knowledge related to the artifacts [5]. The possibility to describe different entities of a resource on a semantic, lexical, etymological or pragmatic level, and to describe the relations of these entities to other resources such as annotations, transcriptions, images or dictionaries, enables a real workbench scenario for scholars in the humanities.

To provide a sustainable solution for supporting these different aspects of a workbench, we have chosen a gradual approach in the development: providing an online editor for the enrichment of structural information, at the same time developing robust content models, to enable future interlinking to other artifacts.

3. The eSciDoc Solution ViRR

The eSciDoc solution ViRR combines a set of tools (components) for publishing scientific content in one user interface. This includes the two key features, the electronic modeling and editing of the original source material (ViRR Editor, see Fig. 1) and its online representation in a digital library (ViRR Viewer, see Fig. 1). These features are often separated from each other and realized in different tools, so that data transformations between these tools become a necessary drawback. The integrated design of ViRR allows users to perform all working steps within one software solution.

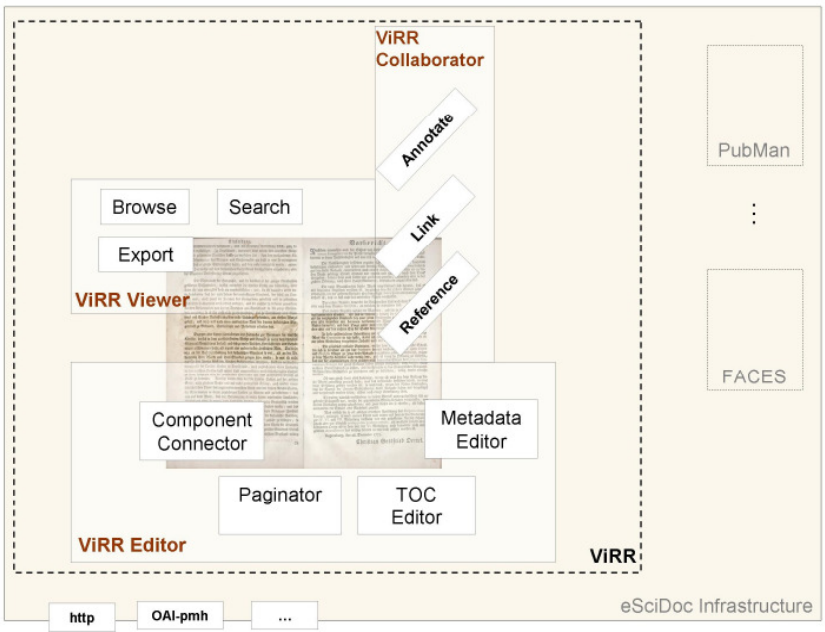

Figure 1: Overview of the different components of ViRR, embedded in the eSciDoc infrastructure

The core of ViRR is the online editor for the creation of electronic representations of cultural artifacts. While browsing through the scans (Fig. 2), several independent working steps are supported: the semi-automatic recording of the logical pagination next to the already

available physical one (Fig. 3), the gathering of the structure via building a hierarchical tree based on different structural elements like, for example introduction, chapter or paragraph (ToC editor, Fig. 4) and the assignment of corresponding scans and descriptive metadata to these structural elements (metadata editor, Fig. 4). All of these working steps are presented in one complex, but flexible workspace. This design was chosen due to different user groups (e.g. librarians, scientists) with various working methods. It allows every user to configure the editor workspace based on his focus of work by providing relevant and hiding distractive information for each working step separately. Further on, all working steps can be performed in any order or can be mixed up depending on the individual needs of the user.
Created data (structure, pagination, metadata) can be published online at any time during the editing process and therefore immediately be reused by other users.

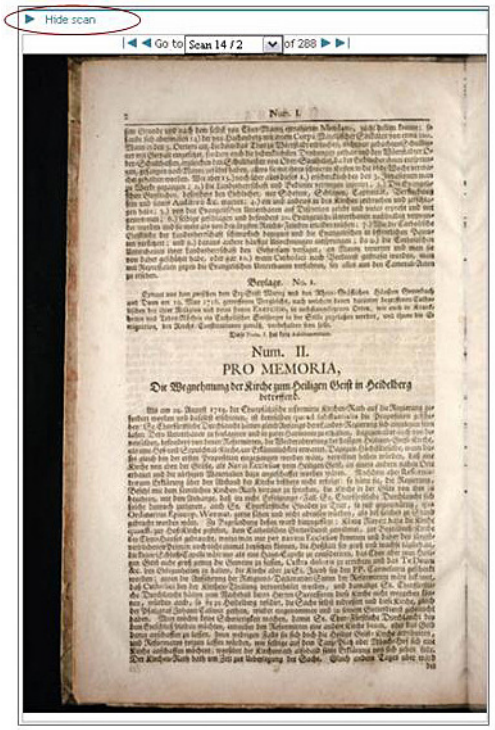

Figure 2: Parallel browsing during the editing process

Figure 3: Paginator

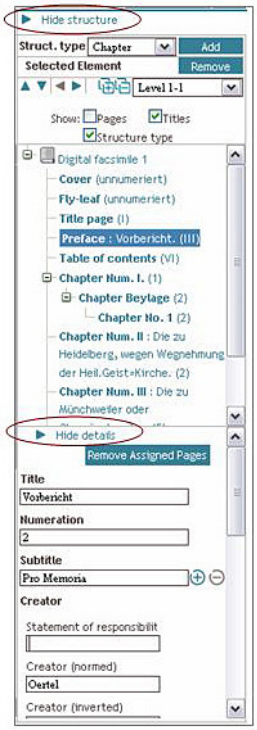

Figure 4: ToC and metadata editor

Within the ViRR Viewer, the content of the collection (multivolumes, volumes and monographs) is navigable via a browsing tree. Each work can be browsed separately in a configurable workspace where the user himself can decide whether he wants to see the bibliographic metadata, the logical structure in form of a table of contents or some parts of it, the scans, or a mixture of all of them. The offering of such customizable viewing sections provides each user an optimized environment to focus on his special interest.

4. ViRR Collaborative Aspects

In a next step, the ViRR solution will be enhanced with a new component, the ViRR Collaborator (as presented in Fig. 1), with the aim to improve the scientific value of the digitized collections by revealing hidden semantics and relations between various disciplines.
The provision of adequate collaboration tools is especially of interest when dealing with different research contexts: investigating textual aspects focus on certain details of a collection (e.g. transcriptions or the identification of text fragments) whereas studies on visual aspects focus on e.g. high resolution scans and referencing of image parts. Others might be interested in the collection as such by e.g. browsing through the scans and

investigate the metadata. The challenge is to identify the generic functionalities for annotating and sharing, and to provide a working environment adaptable to the requirements of different holdings. Different collaboration tools can be applied like graphical annotations, e.g. by integrating the enhanced viewing environment DigiLib (http://digilib.berlios.de), or textual annotations. Further on, transcriptions of the original text corpora will be included to improve the semantically exploitation and retrieval of the digitized works. For easy creation and quality assurance of metadata, we will aim to integrate discipline specific authority data, either stored externally or provided by the eSciDoc service CoNE (Control of Named Entities [6]).
For supporting collaborative work around different collections we would like to enable users to invite others to co-work on a collection by assigning fine granular access rights to private content.

5. The eSciDoc Infrastructure

The collaborative refinements of the ViRR solution are mostly enabled by its underlying technical infrastructure. The eSciDoc infrastructure [7, 8] is designed as a service-oriented architecture. It is an open source joint development of the Max Planck Society and the FIZ Karlsruhe, funded by the German Federal Ministry of Education and Research (BMBF).
A service-oriented architecture fosters the reuse of existing services; therefore an eSciDoc service may be reused by other projects and institutions and become a building block within a broader e-Science infrastructure [9]. The data storage system for the eSciDoc infrastructure is based on the Fedora Commons platform (http://www.fedora-commons.org).
The eSciDoc content model primarily consists of two generic objects called item and container. An item object, in case of ViRR, is the digital representation of a cultural artifact (e.g. scanned page) and contains metadata (such as MAB, MODS) and optionally components (such as jpeg, pdf). A container object is an aggregation of objects (items or containers) such as a journal issue which aggregates several articles. Using this content model, ViRR specializes item and container objects into volume, multivolume, monograph, ToC, and scan (see Fig. 5).

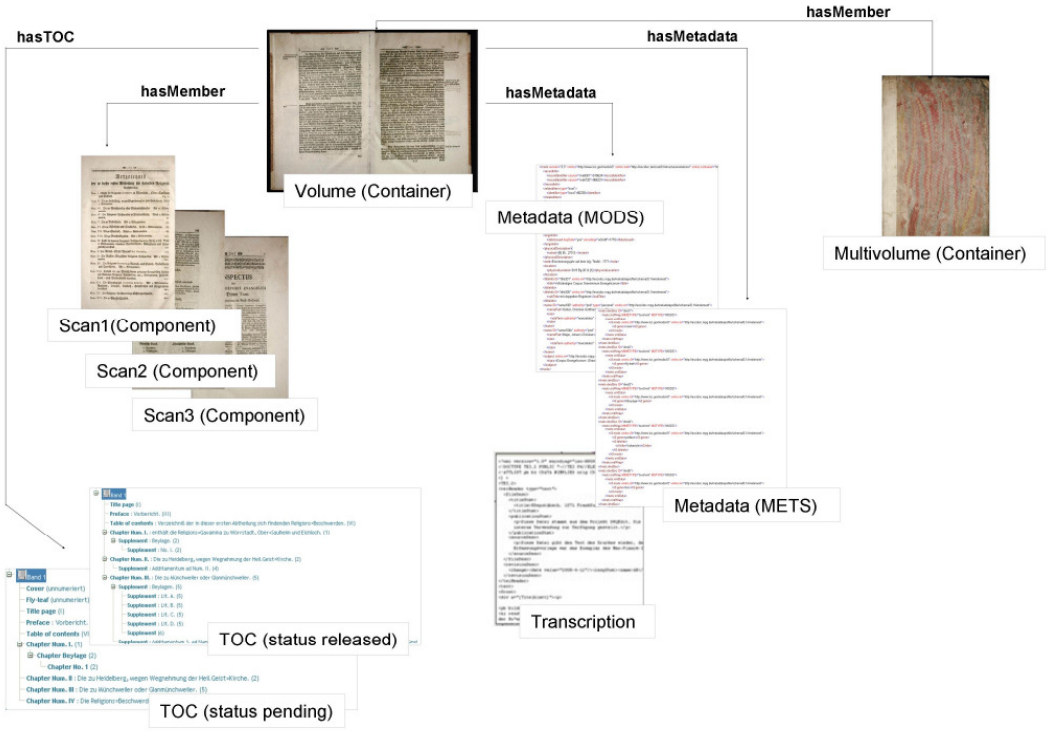

Figure 5: Digitized book content model

For example a digitized book is expressed in eSciDoc as a container, consisting of multiple items such as scans, transcriptions, and structural metadata. This container additionally holds the bibliographic metadata of the book (expressed in MODS). Additionally the generic data model of eSciDoc enables the integration and representation of data from diverse disciplines. The definition of new content models for other research data such as digitized journals or collections of images with discipline specific attributes can easily be integrated into the infrastructure, by defining a new content model with corresponding metadata profile.
As ViRR is fully embedded in the eSciDoc infrastructure it can profit from all existing eSciDoc services. Especially persistent identification (CNRI Handle or other), versioning, preservation (incl. PREMIS metadata) or the support of multiple metadata profiles (Dublin Core, MODS, custom profiles) would require, without the availability of eSciDoc, complex and time consuming development efforts for each new type of data.
eSciDoc is an open source project, setting a high priority in the implementation of standardized interfaces like oai-pmh, sword (http://www.swordapp.org) or RSS. Such an orientation fosters the integration of eSciDoc and eSciDoc-based solutions and their exploiting by other projects like the German national standardized viewing platform DFG Viewer (http://dfg-viewer.de). eSciDoc solutions are also evaluated in the context of other national or European initiatives like TextGrid (http://www.textgrid.de) or DARIAH (http://www.dariah.eu).
ViRR itself, besides offering functionality to process and disseminate data, provides as well services such as on-the-fly transformation of data. These can be used by other solutions, forming together an open accessible net of research data.

6. Conclusion

A range of requirements from different research disciplines exists for the handling of digitized cultural heritage on the web. Based on our experiences, this range can not be fulfilled by a monolithic software alone. One possibility to handle this range is to use an extensible infrastructure like eSciDoc, which focuses on standardization to support interoperability and therefore allows data exchange with services from other providers. So the data of eSciDoc solutions can be further re-used by external tools.
With the approach of using an underlying extensible infrastructure for the development of the ViRR solution, we are confident to fulfil most of the requirements arising from diverse disciplines concerning the work with digitized text resources, which is especially important in a heterogeneous research organization like the Max Planck Society.

7. References

[1] http://colab.mpdl.mpg.de/mediawiki/ViRR:_Virtueller_Raum_Reichsrecht

[2] McDonough, J. P.: METS: standardized encoding for digital library objects. In: International Journal on Digital Libraries (2006)

[3] Gow, J., Buchanan, G., Warwick, C., and Rimmer, J.: Document Structure in Humanities Collections. http://www.cs.ucl.ac.uk/staff/J.Gow/papers/DocumentStructure.pdf (accessed 08 Sept. 2009).

[4] Bainbridge, D., Thompson, J., and Witten, I.H.: Assembling and enriching digital library collections. In: Proceedings of the 3rd ACM/IEEE-CS joint conference on Digital libraries. IEEE Computer Society, Washington (2003) 323-334.

[5] Hyman, M. D., and Renn, J.: Toward an Epistemic Web. Unpublished manuscript prepared for the Dahlem Workshop on Globalization of Knowledge and its Consequences. (2007) http://archimedes.fas.harvard.edu/mdh/epistemic-web.pdf (accessed 09 Oct. 2009)

[6] http://colab.mpdl.mpg.de/mediawiki/Service_for_Control_of_Named_Entities
[7] Bulatovic, N., Tschida, U., and Gros, A.: eSciDoc - a service infrastructure for management of Cultural Heritage content. In: Proceedings of the 14th International Conference on Virtual Systems and Multimedia, (Eds.) Ioannides, M.; Addison, A.; Georgopoulos, A.; Kalisperis, L. ARCHAEOLINGUA, Budapest (2008) 138-143.
[8] Dreyer, M., Bulatovic, N., Tschida, U., and Razum, M.: eSciDoc – a Scholarly Information and Communication Platform for the Max Planck Society. In: German e-Science Conference, Seq. No.: 315471.0 (2007).
[9] http://colab.mpdl.mpg.de/mediawiki/ESciDoc_SOA_AtGlance