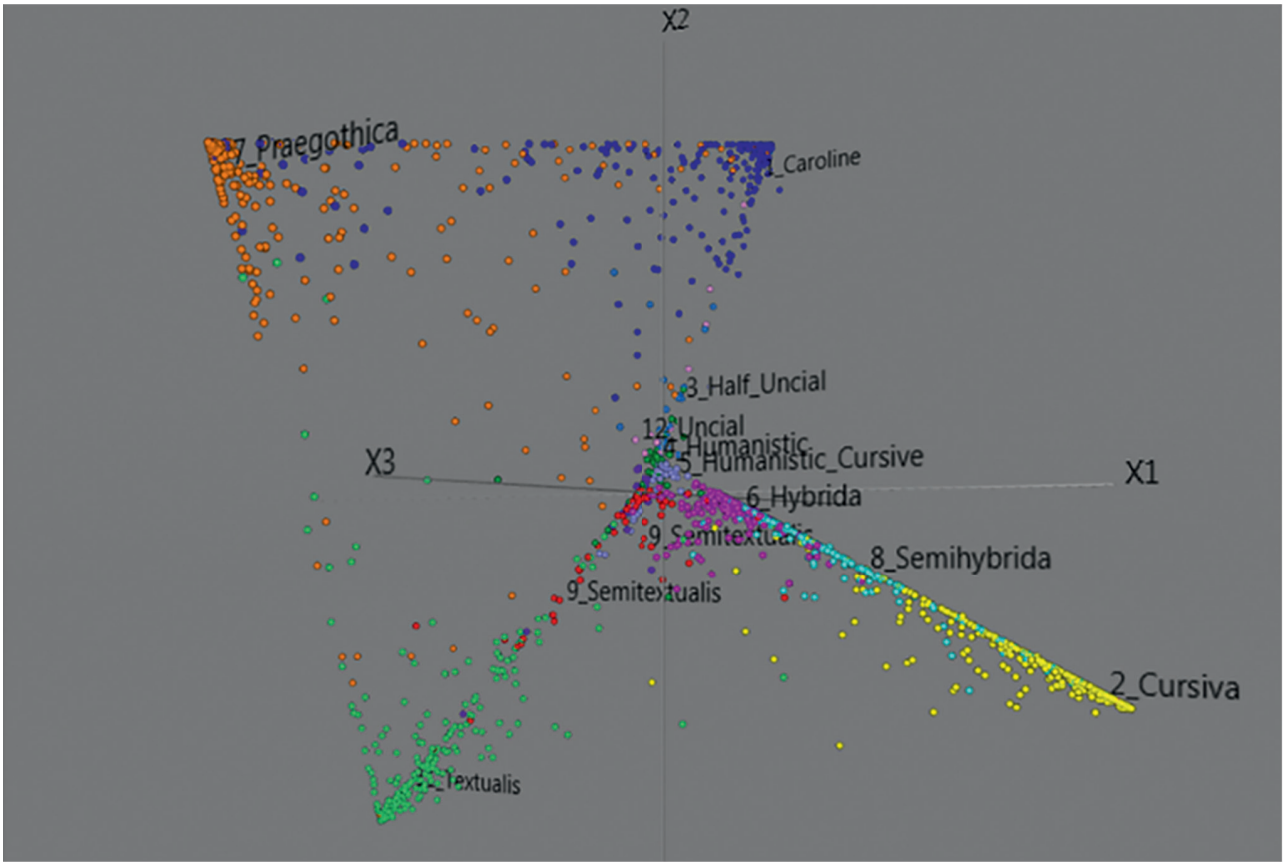
mc Nº 15 2020

manuscript cultures

Hamburg | Centre for the Study of Manuscript Cultures ISSN 1867–9617



Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG
CENTRE FOR THE STUDY OF MANUSCRIPT CULTURES
DFG

Publishing Information

Natural Sciences, Technology and Informatics in Manuscript Analysis
Edited by Oliver Hahn, Volker Märgner, Ira Rabin, and H. Siegfried Stiehl

Proceedings of the third *International Conference on Natural Sciences and Technology in Manuscript Analysis* and the workshop *OpenX for Interdisciplinary Computational Manuscript Research* that took place at the University of Hamburg, Centre for the Study of Manuscript Cultures, on 12–14 June 2018.

Editors
Prof. Dr Michael Friedrich
Universität Hamburg
Asien-Afrika-Institut
Edmund-Siemers-Allee 1/ Flügel Ost
D-20146 Hamburg

Tel. No.: +49 (0)40 42838 7127
Fax No.: +49 (0)40 42838 4899
michael.friedrich@uni-hamburg.de

Prof Dr Jörg Quenzer
Universität Hamburg
Asien-Afrika-Institut
Edmund-Siemers-Allee 1/ Flügel Ost
D-20146 Hamburg
Tell. No.: +49 40 42838 - 7203
Fax No.: +49 40 42838 - 6200
joerg.quenzer@uni-hamburg.de

www.csmc.uni-hamburg.de

Editorial Office
Dr Irina Wandrey
Universität Hamburg
Centre for the Study of Manuscript Cultures
Warburgstraße 26
D-20354 Hamburg
Tel. No.: +49 (0)40 42838 9420
Fax No.: +49 (0)40 42838 4899
irina.wandrey@uni-hamburg.de

Layout
Miriam Gerdes

Cover
Image of ICDAR2017 Tensmeyer's distance matrix (axes 2 and (1 and 3)), see article by Dominique Stutzmann, Christopher Tensmeyer and Vincent Christlein in this volume.

Translation and Copy-editing
Amper Translation Service, Carl Carter, Fürstenfeldbrück

Print
AZ Druck und Datentechnik GmbH, Kempten
Printed in Germany

ISSN 1867–9617

CONTENTS

ARTICLES

Article

Creating Workflows with a Human in the Loop for Document Image Analysis

Marcel Gygli (Würsch), Mathias Seuret, Lukas Imstepf, Andreas Fischer, Rolf Ingold | Windisch, Fribourg

Abstract

Workflows in document image analysis have special requirements compared with workflows in most other scientific fields. As many methods are semi-automatic, the user needs to be kept in the loop and interact with the workflow, change parameters and investigate results in order to generate the best results. However, most traditional workflow management systems do not allow for this and can execute only a fully automatic workflow from beginning to end.

With DIVA-DIP, we introduce a new workflow management system that focuses on the special needs of document image analysis workflows. DIVA-DIP makes it possible to design workflows visually and for processes to have a specific state in the execution phase, in which they wait for user interaction, thus keeping the user integrated in the process.

To test the application on existing document image analysis methods, we integrated methods that are offered through DIVAServices as Web Services. This enables DIVA-DIP to execute methods without the need to install them on the local machine.

Introduction

Document image analysis (DIA) has a long history in computer science research. The goal of DIA is to process digital document images and transform the information contained, such that computers can use it. Methods developed by researchers are used to analyse the layout elements on a document page or to recover the written text by means of either optical character recognition (OCR) if it is printed text or handwritten text recognition (HTR) if it is handwritten text.

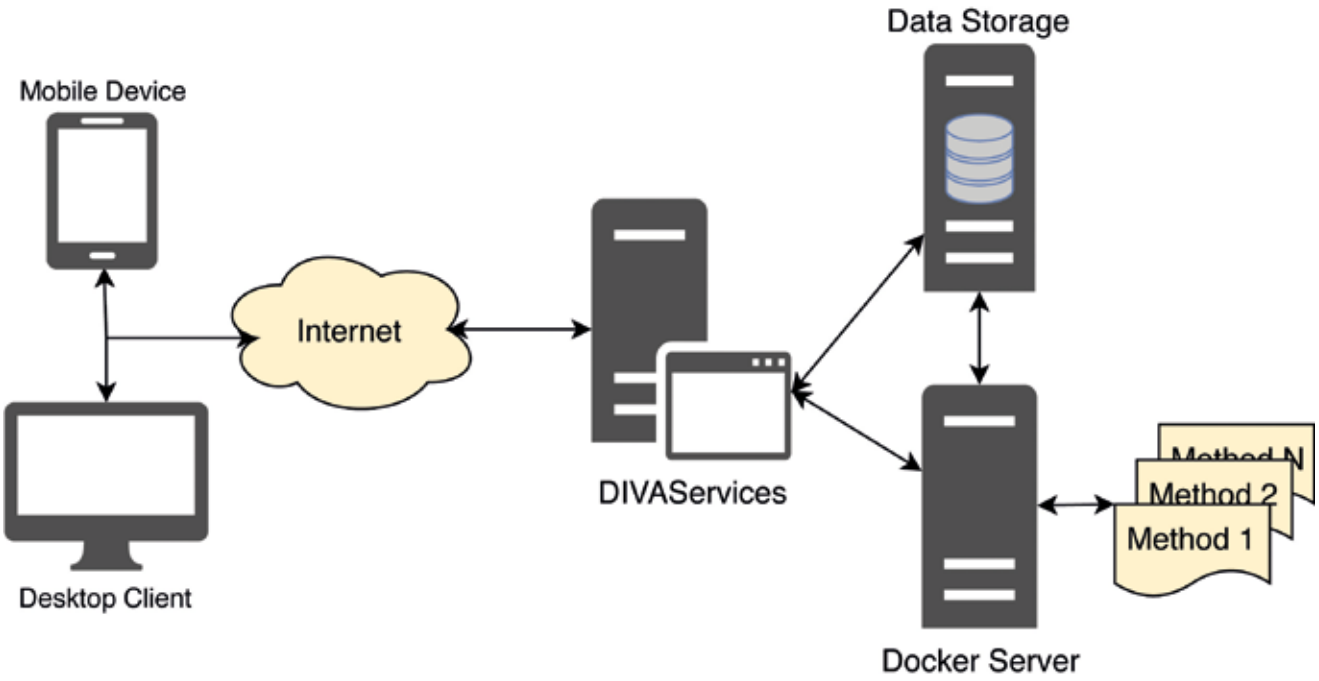


Fig. 1: Architecture overview of DIVAServices. At DIVAServices, we use three different servers, one for the general handling of requests (DIVAServices), one for storage of the data (Data Storage) and one on which the actual methods are installed and executed (Docker Server). This architecture setup allows DIVAServices to scale in case of usage growth. (Figure taken from Würsch, Liwicki, and Ingold 2018.)

Creating and executing workflows for document image analysis is not a simple task. One has to find suitable DIA methods, typically fine-tune the parameters of such methods, adjust them to work together and then execute them on a specific dataset. With DIVAServices, Würsch et al. introduced a solution for one part of this problem, by introducing a framework for providing access to DIA methods as RESTful Web Services (Würsch, Ingold, and Liwicki 2016).

In this paper, we introduce the DIVA Document Image Processor (DIVA-DIP), a workflow management system (WMS) that can be used to design and execute complex workflows using methods provided by DIVAServices. We take special care to address the special needs of researchers in DIA. This means that, in comparison with most traditional WMS, DIVA-DIP allows the user to interact and adapt the workflows while they are using it.

DIVA-DIP makes use of the Human-in-the-Loop paradigm (Karwowski 2001). This paradigm allows users of a system to directly interact with it, change its behaviour and therefore have more control over the behaviour of the application. Additionally, we also make use of visual programming (Myers 1986). This will allow users to create their workflows by connecting very simple building blocks. Users will therefore not require programming knowledge, but can interact with the user interface that allows them to build their workflows.

The remainder of this article is structured as follows. In the next section we introduce relevant related work, focusing on other WMS from the literature. In Section 3, we provide an overview of DIVAServices, since DIVA-DIP will make use of methods provided by it. This is followed by the explanation of DIVA-DIP and how it works. The article closes with a summary and a look at possible research areas for future work.

Related work

In recent years, various WMS have been introduced: Pegasus (Deelman et al. 2015) workflow solution for scientific experiments with a focus on exploiting distributed computing infrastructures, and Taverna (Wolstencroft et al. 2013), a domain-independent WMS that is used mainly in the life sciences. Neither of these tools found relevant adaptation in the DIA community. We believe that this is due to the special nature of the domain. Most of the WMS are designed to execute workflows with zero interaction. In DIA, however, it is often the aim to keep the human in the loop, meaning that the user has the possibility to interact with the workflow as it is running.

State-of-the-art, fully automatic DIA systems are still prone to errors when applied to difficult cases, most notably historical documents. This applies across various tasks such as optical character recognition (OCR), handwriting recognition, writer identification and manuscript dating. Considering the sheer amount of different historical scripts and languages, the goal instead should be to provide human experts with interactive DIA tools that support them in their work.

Examples of such tools include the CATTI system for computer-assisted transcription of historical documents (Romero, Toselli, Rodríguez, and Vidal 2007), Aletheia for annotating historical prints (Clausner, Pletschacher, and Antonacopoulos 2011), PhaseGT for the binarization of historical manuscripts (Nafchi, Ayatollahi, Moghaddam, and Cheriet 2013) and GraphManuscribble for intuitive interaction with digital facsimiles (Garz, Seuret, Fischer, and Ingold 2016), to name just a few. Over time, annotations provided by humans can be used to train methods based on machine learning, which in turn can improve the suggestions made by DIA tools, thus closing the loop between human experts and the semi-automatic systems.

DIVAServices – Making DIA methods available as Web Services

Our goal is to build a WMS that can be used as a workflow designer for methods demonstrated in the DIVAServices framework (Würsch, Ingold, and Liwicki 2015). This framework allows DIA researchers to provide access to their own methods through a unified API and for DIA methods to be executed over the Internet without the need to perform any local installation. For a thorough technical explanation of how DIVAServices works and how to interact with it, we refer the reader to our publications on the topic (Würsch et al. 2016, 2018), as well to as our online documentation.[1]

Figure 1 provides an architectural overview of DIVAServices and all the components it interacts with. DIVAServices handles all incoming requests and performs the necessary actions. This includes storing data on our own data storage server and the execution of methods on our own Docker server, where all methods are installed. A good introduction to Docker for reproducible research is

[1] See *DIVAServices: A RESTful Web Service Framework for Document Image Analysis Methods* <https://diva-dia.github.io/DIVAServicesweb/>.

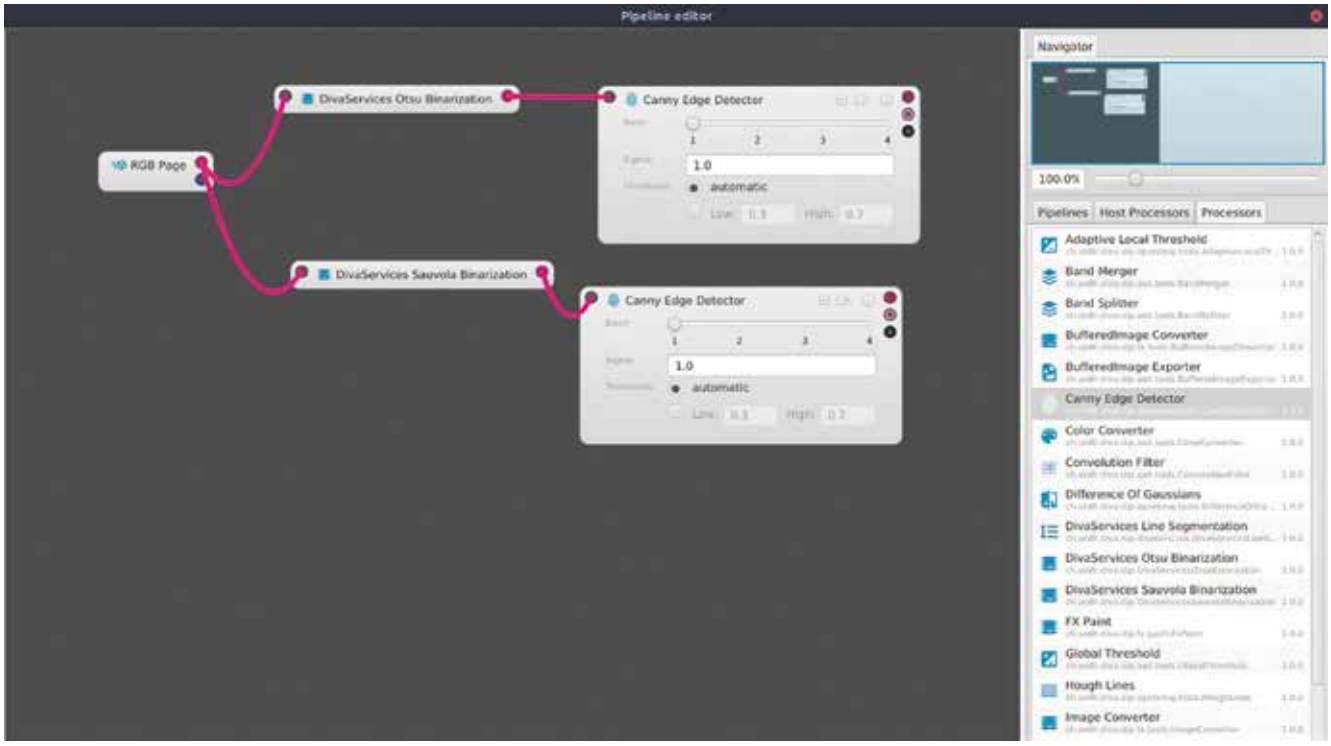


Fig. 2: Workflow designer of DIVA-DIP. The various inputs and outputs are colour-coded for simplicity. Users can change available parameters for each method.

available in Boettiger 2015.When using DIVAServices, the data must be transferred onto its servers to be processed. For this, we require all data that are processed with DIVAServices to be released under a Creative Commons[2] license. We are aware that this is a problem in many humanities institutions where data is locked behind very restrictive licenses. For this reason, we provide the complete DIVAServices framework under an Open Source LGPL v2.1[3] license. This allows everyone to install their own instance of DIVAServices on their local infrastructure, making use of the same principles without losing access to their data.

Currently, when working with DIVAServices, the only way to design and execute complex workflows is by interacting and by programming it. This requires a developer who writes the necessary code that brings together the inputs and outputs of the various methods and chains them together. Some of the users of DIVAServices, like scholars from the humanities, sometimes lack these skills. In this work, we aim at building a WMS that allows them to design their workflows without the need to write a single line of code.

The DIVA Document Image Processor (DIVA-DIP)

DIVA-DIP is a novel WMS for designing DIA workflows of methods offered through DIVAServices. The goal is to provide a tool that allows non-technical experts to design and execute workflows without the need to programme them. With the direct integration of DIVAServices, DIVA-DIP is able to offer state-of-the-art DIA methods without the need to install any of them on the local computer.

Figure 2 shows the most important interface of DIVA-DIP, the workflow designer. In the designer, the user can see all currently available methods (called processors). To start, each workflow needs to have at least one input processor that represents the input page (digital image). Processors can be added to the workspace and the current workflow by dragging them onto the workspace. DIVA-DIP supports the user in connecting processors together using a colour-coding schema. Each input and output has a specific colour and connections can be made only between colours that match. This should help non-technical users in particular to find out what inputs and outputs work together. Additionally, if a method takes additional arguments, they can be provided for the method but can also be changed later when executing the workflow.

[2] See *Creative Commons* <https://creativecommons.org/choose/> for the different type of Creative Commons license. We suggest using CC BY-NC 4.0, which allows DIVAServices to create adaptations of the original work without any problems.

[3] See *GitHub* <https://github.com/DIVA-DIA/DIVAServices>.

Fig. 3: The User Interface of DIVA-DIP. The focus is on providing much space to the actual document. All workflow related information is available on the right.

Fig. 4: Visualization of a computed result. The user can switch between the various results using the radio buttons on the right side.

Unlike traditional WMS that only execute the workflow, DIVA-DIP offers additional features. The most important distinction is that, in DIVA-DIP, processors can enter a specific state where they wait for user input. This makes it possible to design interactions with semi-automatic methods without having to build specific applications. An example of where this could be used is in correcting recognized text line bounding boxes when performing layout analysis. A second difference of DIVA-DIP is that it allows for inspecting the results computed at each step. This enables a user to inspect each result individually and make changes to the workflow based on this visual feedback. Based on changes made after

this visual inspection, the user can then re-run only the necessary parts of the workflow without having to re-run the complete workflow.

Figure 3 provides an overview of the regular user interface. This interface focuses on providing as much space as possible to the document that the user is currently working on, pushing all information related to the current workflow to the right of the screen. There the user can see all the individual steps of the current workflow and can execute them.

After a computation successfully finishes, the user can visualize the result of this computation, as shown in Figure 4. The status of the workflow step (the small square box next to its name) will turn green when the computation, in this case a binarization, is successfully executed. Using the radio buttons enables the user to select the result of the step they would like to have visualized on the workbench.

In our continued effort to open science and reproducibility, DIVA-DIP is fully Open Source released under LGPL v2.1.[4] This makes it possible for everyone to add processors to the application, offering methods that do not necessarily need to be hosted on DIVAServices but could be served either as local applications or through other Web Service frameworks.

Future Work and Conclusions

DIVA-DIP is a new WMS that focuses intensely on the specific differences of workflows in DIA, by keeping the user in the loop. For this, we introduced a new state that processes can enter when they wait for user input. Additionally, DIVA-DIP offers access to state-of-the-art DIA methods through DIVAServices and also eliminates the need to perform any local installation for the methods.

We believe that these additional features make it a great tool for scholars in the humanities, as it keeps them engaged in the process and enables them to investigate the impact of each step involved in the workflow. DIVA-DIP also completely removes the need for any programming knowledge to design the workflows, but offers a visual interface that is easy to understand.

Our hope is to evolve DVA-DIP into the main workflow design tool for DIVAServices. In order to get there, we need to add some more features in the future, namely: develop a way that automatically provides access to all methods currently offered on DIVAServices. Currently, a developer needs to programme and add each method as a processor in DIVA-DIP. We would like to change this, such that at start-up DIVA-DIP would automatically check which methods are available on DIVAServices and provide access to these.

We would also like to engage with the community at large and invite its members to provide access to their own methods through DIVA-DIP as well, so that it could become a WMS providing access to as many DIA methods as possible.

[4] See *Github* <https://github.com/DIVA-DIA/dip>.

REFERENCES

Boettiger, Carl (2015), 'An Introduction to Docker for Reproducible Research', *ACM SIGOPS Operating Systems Review*, 49.1: 71–79. doi: <10.1145/2723872.2723882>.

Clausner, Christian, Stefan Pletschacher, and Apostolos Antonacopoulos (2011), 'Aletheia: An Advanced Document Layout and Text Ground-Truthing System for Production Environments, *Proceedings of the International Conference on Document Analysis and Recognition*, 48–52. doi: <10.1109/ICDAR.2011.19>.

Deelman, Ewa, Karan Vahi, Gideon Juve, Mats Rynge, Scott Callaghan, Philip J. Maechling, Rajiv Mayani, Weiwei Chen, Rafael Ferreira da Silva, Miron Livny, and Kent Wenger (2015), 'Pegasus: A Workflow Management System for Science Automation', *Future Generation Computer Systems*, 46: 17–35. doi: <10.1016/j.future.2014.10.008>.

Garz, Angelika, Mathias Seuret, Andreas Fischer, and Rolf Ingold (2016), 'GraphManuscribble: Interact Intuitively with Digital Facsimiles', *Second International Conference on Natural Sciences and Technology in Manuscript Analysis, CSMC, Book of Abstracts*, 61–63.

Karwowski, Waldemar (ed.) (2001), *International Encyclopedia of Ergonomics and Human Factors* (London: Crc Press).

Myers, Brad A. (1986), 'Visual Programming, Programming by Example, and Program Visualization: A Taxonomy, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 59–66. doi: <10.1145/22627.22349>.

Nafchi, Hossein Ziaei, Seyed Morteza Ayatollahi, Reza Farrahi Moghaddam, and Mohamed Cheriet (2013), 'An Efficient Ground Truthing Tool for Binarization of Historical Manuscripts', *12th International Conference on Document Analysis and Recognition*, 807–811. doi: <10.1109/ICDAR.2013.165>.

Romero, Verónica, Alejandro H. Toselli, Luis Rodríguez, and Enrique Vidal (2007), 'Computer Assisted Transcription for Ancient Text Images', *Image Analysis and Recognition* (Berlin, Heidelberg: Springer), 1182–1193. doi: <10.1007/978-3-540-74260-9_105>.

Wolstencroft, Katherine, Robert Haines, Donal Fellows, Alan Williams, David Withers, Stuart Owen, Stian Soiland-Reyes, Ian Dunlop, Aleksandra Nenadic, Paul Fisher, Jiten Bhagat, Khalid Belhajjame, Finn Bacall, Alex Hardisty, Abraham Nieva de la Hidalga, Maria P. Balcazar Vargas, Shoaib Sufi, and Carole Goble (2013), 'The Taverna Workflow Suite: Designing and Executing Workflows of Web Services on the Desktop, Web or in the Cloud, *Nucleic Acids Research*, 41: W557–W561. doi: <10.1093/nar/gkt328>.

Würsch, Marcel, Rolf Ingold, and Marcus Liwicki (2015), 'DIVAServices: A RESTful Web Service for Document Image Analysis Methods, *Digital Scholarship in the Humanities*, 32.suppl.1: i150–i156. doi: <10.1093/llc/fqw051>.

Würsch, Marcel, Rolf Ingold, and Marcus Liwicki (2016), 'SDK Reinvented: Document Image Analysis Methods as RESTful Web Services', *12th IAPR Workshop on Document Analysis Systems*, 90–95. doi: <10.1109/DAS.2016.56>.

Würsch, Marcel, Marcus Liwicki, and Rolf Ingold (2018), 'Web Services in Document Image Analysis: Recent Developments and the Importance of Building an Ecosystem', *13th IAPR Workshop on Document Analysis Systems*.

Written Artefacts as Cultural Heritage

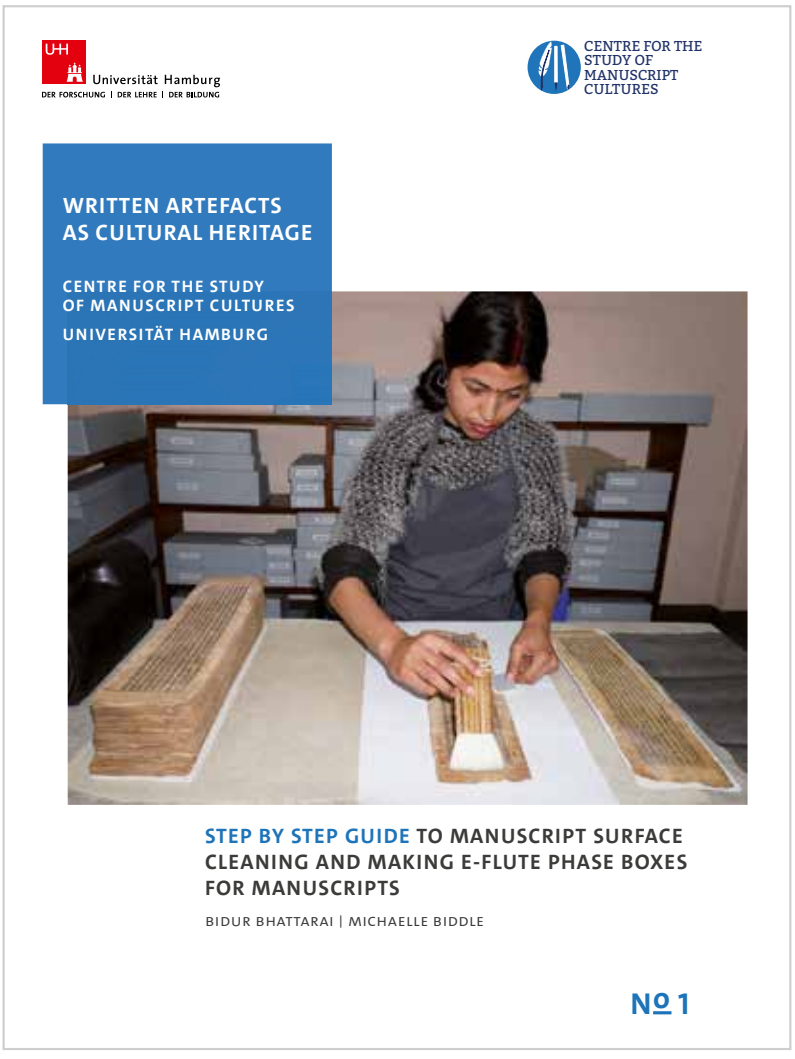
Ed. by Michael Friedrich and Doreen Schröter

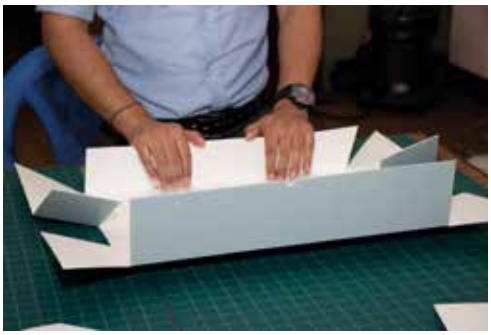
Written Artefacts as Cultural Heritage was established in 2020. The series is dedicated to the double role of written artefacts as representations and generators of humankind's cultural heritage. Its thematic scope embraces aspects of preservation, the identity-defining role of artefacts as well as ethical questions.

The mix of practical guides, colloquium papers and project reports is specifically intended for staff at libraries and archives, curators at museums and art galleries, and scholars working in the fields of manuscript cultures and heritage studies.

Every volume of *Written Artefacts as Cultural Heritage* has been peer-reviewed and is openly accesible. There is an online and a printed version..

If you wish to receive a copy or to present your research, please contact the editorial office:
https://www.csmc.uni-hamburg.de/publications/cultural-heritage.html





manuscript cultures (mc)

Editors: Michael Friedrich and Jörg B. Quenzer
Editorial office: Irina Wandrey

CSMC's academic journal was established as newsletter of the research unit 'Manuscript Cultures in Asia and Africa' in 2008 and transformed into a scholarly journal with the appearance of volume 4 in 2011. *manuscript cultures* publishes exhibition catalogues and articles contributing to the study of written artefacts. This field of study embraces disciplines such as art history, codicology, epigraphy, history, material analysis, palaeography and philology, informatics and multispectral imaging.

manuscript cultures encourages comparative approaches, without regional, linguistic, temporal or other limitations on the objects studied; it contributes to a larger historical and systematic survey of the role of written artefacts in ancient and modern cultures, and in so doing provides a new foundation for ongoing discussions in cultural studies.

Every volume of *manuscript cultures* has been peer-reviewed and is openly accessible:
https://www.csmc.uni-hamburg.de/publications/mc.html

If you wish to receive a copy or to present your research in our journal, please contact the editorial office:
irina.wandrey@uni-hamburg.de

manuscript cultures 14

manuscript cultures 13

manuscript cultures 11

manuscript cultures 10

manuscript cultures 9

manuscript cultures 8

manuscript cultures 7

manuscript cultures 6

manuscript cultures 5

manuscript cultures 4

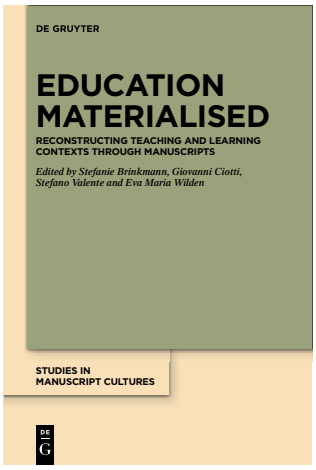
Studies in Manuscript Cultures (SMC)

Ed. by Michael Friedrich, Harunaga Isaacson, and Jörg B. Quenzer

From volume 4 onwards all volumes are available as open access books on the De Gruyter website:
https://www.degruyter.com/view/serial/43546
https://www.csmc.uni-hamburg.de/

DE GRUYTER

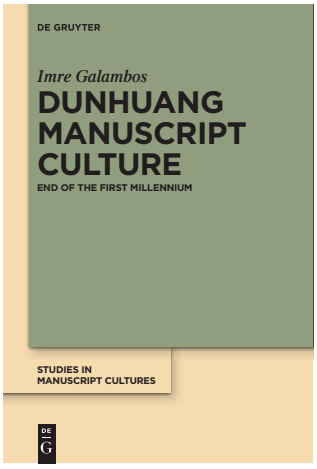
Forthcoming



23 – *Education Materialised: Reconstructing Teaching and Learning Contexts through Manuscripts,* edited by Stefanie Brinkmann, Giovanni Ciotti, Stefano Valente and Eva Maria Wilden

Manuscripts have played a crucial role in the educational practices of virtually all cultures that have a history of using them. As learning and teaching tools, manuscripts become primary witnesses for reconstructing and studying didactic and research activities and methodologies from elementary levels to the most advanced.
The present volume investigates the relation between manuscripts and educational practices focusing on four particular research topics: educational settings: teachers, students and their manuscripts; organising knowledge: syllabi; exegetical practices: annotations; modifying tradition: adaptations.
The volume offers a number of case studies stretching across geophysical boundaries from Western Europe to South-East Asia, with a time span ranging from the second millennium BCE to the twentieth century CE.

New release



22 – *Dunhuang Manuscript Culture: End of the First Millennium*, by Imre Galambos

Dunhuang Manuscript Culture explores the world of Chinese manuscripts from ninth–tenth century Dunhuang, an oasis city along the network of pre-modern routes known today collectively as the Silk Roads. The manuscripts have been discovered in 1900 in a sealed-off side-chamber of a Buddhist cave temple, where they had lain undisturbed for for almost nine hundred years. The discovery comprised tens of thousands of texts, written in over twenty different languages and scripts, including Chinese, Tibetan, Old Uighur, Khotanese, Sogdian and Sanskrit. This study centres around four groups of manuscripts from the mid-ninth to the late tenth centuries, a period when the region was an independent kingdom ruled by local families. The central argument is that the manuscripts attest to the unique cultural diversity of the region during this period, exhibiting – alongside obvious Chinese elements – the heavy influence of Central Asian cultures. As a result, it was much less 'Chinese' than commonly portrayed in modern scholarship. The book makes a contribution to the study of cultural and linguistic interaction along the Silk Roads.

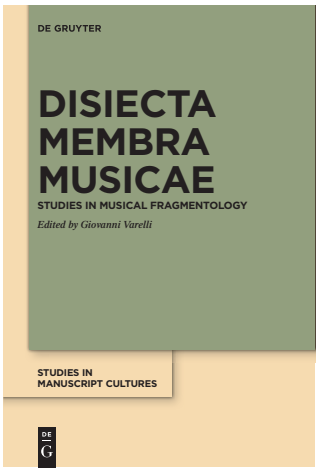
Studies in Manuscript Cultures (SMC)

Ed. by Michael Friedrich, Harunaga Isaacson, and Jörg B. Quenzer

From volume 4 onwards all volumes are available as open access books on the De Gruyter website:
https://www.degruyter.com/view/serial/43546
https://www.csmc.uni-hamburg.de/

DE GRUYTER

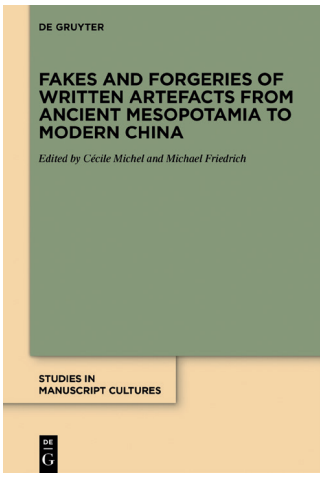
New release



21 – *Disiecta Membra Musicae: Studies in Musical Fragmentology*, edited by Giovanni Varelli

Although fragments from music manuscripts have occupied a place of considerable importance since the very early days of modern musicology, a collective, up-to-date, and comprehensive discussion of the various techniques and approaches for their study was lacking. On-line resources have also become increasingly crucial for the identification, study, and textual/musical reconstruction of fragmentary sources. *Disiecta Membra Musicae. Studies in Musical Fragmentology* aims at reviewing the state of the art in the study of medieval music fragments in Europe, the variety of methodologies for studying the repertory and its transmission, musical palaeography, codicology, liturgy, historical and cultural contexts, etc. This collection of essays provides an opportunity to reflect also on broader issues, such as the role of fragments in last century's musicology, how fragmentary material shaped our conception of the written transmission of early European music, and how new fragments are being discovered in the digital age. Known fragments and new technology, new discoveries and traditional methodology alternate in this collection of essays, whose topics range from plainchant to *ars nova* and fifteenth- to sixteenth-century polyphony.

New release



20 – *Fakes and Forgeries of Written Artefacts from Ancient Mesopotamia to Modern China,* edited by Cécile Michel and Michael Friedrich

Fakes and forgeries are objects of fascination. This volume contains a serie of thirteen articles devoted to fakes and forgeries of written artefacts from th beginnings of writing in Mesopotamia to modern China. The studies empha sise the subtle distinctions conveyed by an established vocabulary relating t the reproduction of ancient artefacts and production of artefacts claiming t be ancient: from copies, replicas and imitations to fakes and forgeries. Fake are often a response to a demand from the public or scholarly milieu, or ever both. The motives behind their production may be economic, political, reli gious or personal – aspiring to fame or simply playing a joke. Fakes may be revealed by combining the study of their contents, codicological, epigraphi and palaeographic analyses, and scientific investigations. However, certain fa mous unsolved cases still continue to defy technology today, no matter hov advanced it is. Nowadays, one can find fakes in museums and private collec tions alike; they abound on the antique market, mixed with real artefacts tha have often been looted. The scientific community's attitude to such objects calls for ethical reflection.

ISSN 1867–9617

Centre for the Study of Manuscript Cultures
Universität Hamburg
Warburgstraße 26
D-20354 Hamburg

www.csmc.uni-hamburg.de