



Centrul Digital Humanities Transilvania



Studia Universitatis Babeş-Bolyai
Digitalia

Editorial Office:

Room 218, 31 Horea St., 400202 Cluj-Napoca, Romania

Web site: <http://digihubb.centre.ubbcluj.ro/journal/index.php/digitalia>

<http://studia.ubbcluj.ro/serii/digitalia/>

Contact: digihubb@ubbcluj.ro

Editorial Board

DIRECTOR and EDITOR-IN-CHIEF:

Corina MOLDOVAN, DigiHUBB, Babeş-Bolyai University, Cluj-Napoca, Romania

EDITORIAL BOARD:

Cristina FELEA, DigiHUBB, Babeş-Bolyai University, Cluj-Napoca, Romania
Christian SCHUSTER, DigiHUBB, Babeş-Bolyai University, Cluj-Napoca, Romania
Rada VARGA, DigiHUBB, Babeş-Bolyai University, Cluj-Napoca, Romania

ADVISORY BOARD:

Daniel ALVES, Universidade Nova de Lisboa, Portugal
Maria BARAMOVA, Sofia University "St. Kliment Ohridski", Bulgaria
Elisabeth BURR, Leipzig University, Germany
Claire CLIVAZ, Swiss Institute of Bioinformatics, Switzerland
Nicolae CONSTANTINESCU, Romania
Milena DOBREVA, Malta University, Malta
Jennifer EDMOND, Trinity College Dublin, Ireland
Øyvind EIDE, University of Cologne, Germany Alex GIL, Columbia University, USA
Elena GONZALEZ-BLANCO GARCIA, Director of the Digital Humanities Innovation Lab (LINDH),
Faculty of Philology, UNED, Madrid, Spain
Jean-Noël GRANDHOMME, Université de Lorraine, Nancy, France
Seth van HOOLAND, Université libre de Bruxelles, Belgium
Lorna HUGHES, University of Glasgow, UK
Patritsia KALAFATA, Academy of Athens, Greece
Jessie LABOV, CEU, Budapest, Hungary
Maurizio LANA, Università del Piemonte Orientale, Italy
Willard McCARTY, Kings College London, UK
Christian-Emil ORE, University of Oslo, Norway
Gábor PALKÓ, Petőfi Irodalmi Múzeum, Budapest, Hungary
Costas PAPADOPOULOS, Maynooth University, Ireland
Susan SCHREIBMAN, Maynooth University, Ireland
Marija SEGAN, Institute of Mathematics, Serbia
Liana STANCA, Babeş-Bolyai University, Cluj-Napoca, Romania
Adriana TUDOR TIRON, Babeş-Bolyai University, Cluj-Napoca, Romania
Aleš VAUPOTIČ, School of Humanities of the University of Nova Gorica, Slovenia

DTP:

Edit Fogarasi

CONTENT

Eloisa PAGANONI, Epigraphic Squeezes: Digitisation, Publication, and Enhancement 5

Gabriel BODARD and Polina YORDANOVA, Publication, Testing and Visualization
with EFES: A Tool for All Stages of the EpiDoc XML Editing Process 17

Adinel C. DINCĂ and Emil ŞTEŢCO, Preliminary Research on Computer-Assisted
Transcription of Medieval Scripts in the Latin Alphabet using AI Computer
Vision techniques and Machine Learning. A Romanian Exploratory Initiative 37

Radu NEDICI, DaT18 Database: A Prosopographical Approach to the Study of the
Social Structures of Religious Dissent in Mid-Eighteenth-Century Transylvania 53

BOOK REVIEW

Anna Wing-bo Tso, *Digital Humanities and New Ways of Teaching*, Springer, 2019
(Anisia IACOB) 71

Olivier Le Deuff, *Digital Humanities. History and Development*, London: ISTE,
Hoboken: Wiley, 2018 (Alexandru-Augustin HAIDUC) 75

Epigraphic Squeezes: Digitisation, Publication, and Enhancement

Eloisa Paganoni
Ca' Foscari University of Venice

Abstract: Epigraphic squeezes are a key tool for research and teaching. They also have historical and documentary value. They are reliable copies of inscribed text and become the only evidence that remains if inscriptions are lost or destroyed. This paper describes the *Venice Squeeze Project* for the preservation and enhancement of epigraphic squeezes in the Department of Humanities at Ca' Foscari University of Venice. For the initial phase of the project, the Ca' Foscari University collection of epigraphic squeezes was published in the digital *ektypothekē E-stampages*. The current phase involves developing a web application to digitise epigraphic squeezes according to the metadata architecture of *E-stampages*. The first part of this paper describes the background of the *Venice Squeeze Project* and methodological issues, which fostered the partnership with *E-stampages*. The second part describes the relational database that was set up to digitise the Ca' Foscari collection. The third part introduces the project initiatives to promote a network of Italian institutions interested in digitizing their collections of epigraphic squeezes.

Keywords: Greek epigraphy, squeezes, database architecture

1. Introduction

An epigraphic squeeze is a facsimile of an inscription. Nowadays, squeezes are generally made of acid-free chemistry filter paper but until some years ago, they were also made of plaster or latex (McLean 67-73). The squeeze is one of the most important tools for epigraphers. In providing an exact reproduction, it allows epigraphers to check the legibility and disposition of the text as well as the shape of the letters on the stone.

The squeeze is a *testimonium* to the preservation of an inscription at the time it was made. Squeezes produced in the late nineteenth to early twentieth century show a better state of preservation than those of the same inscriptions made today. Many

stones are not stored in museums but left at the place of discovery. They remain exposed to weather that washes away the surface, making inscribed signs harder to decipher. Stones are often lost both museums and at archaeological sites, especially if they are small fragments. They can also be destroyed during events such as wars or natural disasters.

These eventualities suggest that the squeeze could easily become the best (or the only) *testimonium* of an inscription. Squeezes are more than simply by-products of the activities of epigraphers, as they have long been regarded: they are heritage with an intrinsic documentary and cultural value. Several institutions promote projects to safeguard and enhance this form of documentation, and their projects often result in digital collections of epigraphic squeezes.¹ This paper presents the initiatives promoted by Prof. Claudia Antonetti (Department of Humanities, Ca' Foscari University of Venice) within the framework of the *Venice Squeeze Project* (VSP).

2. The Ca' Foscari collection and the *Venice Squeeze Project*

The Department of Humanities at Ca' Foscari University of Venice hosts a collection of 605 epigraphic paper squeezes. The collection was built up from the late 1970s to 2000s by Claudia Antonetti, who donated it to the Department of Humanities in 2019.² It thus reflects the development of Claudia Antonetti's research interests since the beginning of her career. She produced squeezes for her early study on Sicily. Over five hundred squeezes were made from inscriptions on the museums of Agrinion, Thermos and Thyrraeon, which now lie in the Aitolioakarnania region. They are connected to Antonetti's studies on the epigraphic culture of Central Greece, one of her primary research areas over time. Then, a small but significant group of squeezes were made from inscriptions preserved at the museums of Venice and Veneto region. A few squeezes were donated by scholars who collaborated with Claudia Antonetti.

This is the richest and most coherent collection of epigraphic squeezes in Italy. It has contributed to making Ca' Foscari a hub for teaching on Greek Epigraphy. Every year, undergraduate and graduate students practice reading epigraphic squeezes. Thanks to squeezes of inscriptions dating from the Archaic to Roman Imperial eras, students learn to recognise alphabets and lettering from different ages and regions. They acquire knowledge and skills that could otherwise only be gained through a visual autopsy of the original documents. Although Greek Epigraphy has been penalised in higher education programmes in recent years, both in Italy and elsewhere, the enthusiasm of participating students speaks to the lively interest in inscriptions (Paganoni *et al.*).

¹ See the digital collections of Princeton University (<https://www.ias.edu/krateros>), the CSAD of the University of Oxford (<https://www.csad.ox.ac.uk/squeeze-collection-0>), the British Institute of Ankara (<https://www.biaatr.org/squeeze>), and the Aleshire Center for the Studies of Greek Epigraphy of the University of California (<http://aleshire.berkeley.edu/holdings/images>).

² <https://www.unive.it/pag/16969/>.



Fig. 1. Claudia Antonetti and some of her squeezes. Photo by Paolo Della Corte. Courtesy of Claudia Antonetti

Certain that squeezes were an invaluable resource to both scholars and students, Antonetti embarked on a series of initiatives under the name of the *Venice Squeeze Project* (VSP) to make her collection widely available (Antonetti *et al.* “Digital epigraphy” 491-495; Antonetti *et al.* “Collezioni”).³ In 2012, she promoted a reorganisation of the collection. It was arranged according to geography, and squeezes were given an inventory number. She then launched the first digitisation of the collection, creating a FileMaker archive. The archive collected basic information on the squeezes and the inscriptions they were made from along with photos of the reverse and the obverse of the squeezes as well as the inscriptions themselves⁴. This was used in the research and teaching activities of the members of the Laboratory of Greek Epigraphy.

In 2017, Ca' Foscari University funded a second phase of the VSP to publish the university collection online and contribute to the enhancement of other Italian collections of squeezes. Online publication posed new challenges as it required a suitable method to reproduce and describe the squeezes in a digital environment. Claudia Antonetti contacted her colleague Prof. Michèle Brunet (University of Lyon 2), who was developing the *E-stampages* project with aims analogous to those of the VSP. In view of the shared aims and methods of these projects, they established an official partnership that led to publication of the Ca' Foscari squeezes in the *E-Stampages* digital collection.

³ <http://mizar.unive.it/venicesqueeze/public/frontend>.

⁴ Michela Socal designed and implemented the FileMaker archive.

3. E-stampages: a Digital Ektypotheke

Under the leadership of Michèle Brunet, the *E-stampages* project (Levivier *et al.*; Antonetti *et al.* “Collezioni” 42-48),⁵ aims to digitise a number of epigraphic squeezes of the Laboratoire HiSoMA (Histoire et Sources des Mondes Antiques, Maison de l'Orient et de la Méditerranée, Lyon)⁶, and the EfA (École française d'Athènes)⁷. In addition to the Laboratoire HiSoMA and the EfA, the original consortium of *E-stampages* included two other partners: the Pôle Système d'information et réseaux de la Maison de l'Orient et de la Méditerranée Jean Pouilloux (MOM),⁸ which provided the planetary scanner for the creation of digital images and supported the setup of the CMS (Content Management System) for online publication; and the Digital Epigraphy and Archaeology Project, which provided the 3D software.⁹ Thanks to the proposal of collaboration submitted by Michèle Brunet and Claudia Antonetti to the EfA, the VSP joined this consortium in 2017. The VSP then adopted the *E-stampages*' metadata architecture and published the Ca' Foscari collection on the French website. The collaboration between the VSP and *E-stampages* took their shared aim into account and recognised that *E-stampages* proposed the most sophisticated and innovative digitisation protocol for the collection of squeezes.

E-stampages intends to create a digital *ektypotheke*. This term is composed of the word *ektypon*, which means ‘squeeze’ in Modern Greek and derives from the Ancient Greek word for ‘object in relief’, and the suffix *-theke* (‘collection’). The word *ektypotheke* clarifies the aim and methods of *E-stampages* by stressing that the squeeze (and not the inscription or the text) is the focus of the digital resource. The metadata architecture derives from this as the squeeze is at the centre of the digital resource, while information about the stone, text, and images is arranged around it. *Ektypon* reminds us that a squeeze is an object in relief. This is the key feature that allows epigraphers to read squeezes. Thus, it had to be preserved in the digital environment via a 3D model (Antonetti *et al.* “Collezioni” 43).

E-stampages partners (HiSoMA, EfA, and VSP) follow the same protocol for dataset and image preparation. Michèle Brunet and Adeline Levivier developed the metadata architecture in 2016, which was then further refined during a meeting of the partners at the EfA in 2018.¹⁰ This metadata architecture is composed of five interrelated entities. The first entity, the squeeze, is linked to the entities of the 2D images, the 3D model, and the text. The text is connected to the final entity, the artefact (that is, the object). The text description includes a reference to the *editio princeps* or

⁵ <https://www.E-stampages.eu/s/E-stampages/page/accueil>. I thank Michèle Brunet for allowing me to introduce *E-stampages*.

⁶ <https://www.hisoma.mom.fr/>.

⁷ <https://www.efa.gr/index.php/fr/ressources-documentaires/les-archives/archives-estampages/le-programme-E-stampages>.

⁸ <https://www.mom.fr/les-services-de-la-federation/pole-systeme-d-information-et-reseaux/presentation>.

⁹ See below in this paragraph.

¹⁰ For HiSoMA and MOM: Michèle Brunet, Adeline Levivier, Richard Bouchon, Bruno Morandière and Hélène Vuidel; for the EFA: Marie Stahl, Louis Mulot, Anaïs Michel, Nicolas Genis and Julien Fournier; for the VSP: Claudia Antonetti and Eloisa Paganoni.

the reference edition; if needed, an entry from the *Supplementum Epigraphicum Graecum* is added to highlight a contribution that significantly adds to the reading of the reference edition. Bibliographical references are managed via Zotero.¹¹

Research groups within the HiSoMA, EfA, and VSP had earlier produced datasets that were already available, but these needed to be expanded and synchronised. Teams entered missing information and carried out a re-documentarisation of the datasets by adding metadata needed to describe digital items and make them function. Metadata was organised according to a hierarchy that is searchable, interoperable, and reusable according to the instructions of the 3W Consortium on the development of the Semantic Web.

Using either a planetary scanner or a camera, the teams produced a series of digital images for each squeeze¹²:

- 2 files TIFF (400 dpi) of the obverse and reverse for long term storage;
- 2 files PNG (400 dpi) of the obverse and reverse for online publication;
- 2 files PNG (200 dpi) of the reverse, rotating the light exposure by 90°, for the 3D model.

The 3D model was produced using DEA software. This software was developed by Eleni Bozia and Angelos Barmoutis at the University of Florida as part of the Digital Epigraphy and Archaeology Project (Bozia *et al.*).¹³ It produces a 3D model that can be embedded and displayed in web pages. Users may move, rotate, zoom in, zoom out, or change the shading and visualisation mode to see details that are otherwise invisible.

Datasets from the three partners (HiSoMA, EfA and VSP) are stored and distributed on the internet by The Huma-bum Box¹⁴ and the service Nakala¹⁵ of the Très Grande Infrastructure de Recherche nationale française (TGIR) Huma-Num.¹⁶ Collections of epigraphic squeezes are published on the *E-stampages* website, which is managed through the CMS (Content Management System) Omeka S.¹⁷ This CMS requires no advanced IT proficiency to arrange datasets for online publication. Furthermore, it allows for data exploration in several—and in some cases unprecedented—ways.

A beta version of the *E-stampages* website was presented at the Sixth ‘Seminario Avanzato di Epigrafia Greca’ held in Venice in January 2019. The first version of the *E-stampages* website, which hosted a sample of the HiSoMA and EfA squeezes of Thasos and Ca’ Foscari squeezes of the Museum of Agrinion, was released shortly after. The online collection has steadily expanded since then to include other squeezes from Thasos and the remaining squeezes of the Ca’ Foscari collection, which are now completely available online.

¹¹ <https://www.zotero.org>.

¹² Scanner models: Digibook Zeutschel (in Lyon), Copibook™ Cobalt (in Venice).

¹³ <https://www.digitalepigraphy.org>.

¹⁴ <https://humanum.hypotheses.org/2711>.

¹⁵ <https://www.huma-num.fr/services-et-outils/exposer>.

¹⁶ <https://www.huma-num.fr/>.

¹⁷ <https://omeka.org/s>.

The *E-stampages* website is organised into six sections. The first two sections introduce the project and describe the digitisation protocol. 'Les collections' describes the collections published on the website; from this section, users can explore each collection by provenance and number of squeezes. 'Les estampages par provenance' presents squeezes by a geographical criterion. The following section, 'Parcourir toutes les collections', allows exploration of collections by object and text types. The last section currently available, 'Bibliographie & liens web', lists the project contributions and the collections published on the *E-stampages* website. Another section currently under development, 'Portraits d'épigraphistes', will host temporary exhibitions on epigraphy with a focus on influential scholars that have shaped the discipline.

4. The Venice Squeeze Project database

Partnership between the VSP and *E-stampages* was facilitated by a common aim (the online publication of a squeeze collection), but it also responded to relevant methodological questions. *E-stampages* offers the most articulate means for the online publication of epigraphic squeezes. Its metadata architecture focuses on the squeeze itself and provides a full description of the text and inscription. A holistic concept of the squeeze thus underlies metadata architecture as it considers all features describing or attached to the squeeze. This descriptive care helps transform squeezes from by-products of the activities of epigraphers into pieces of heritage with historical and documentary value. The digital images and 3D models provide users with an accurate digital reproduction of the squeezes.

Furthermore, the partnership aims to overcome fragmentation featuring digital resources. It is hard to keep pace with the ever-increasing number of resources that have become available in recent years. Some resources (the smallest and most specific ones, above all) are inevitably overlooked or underexploited. Cooperation between similar initiatives may limit fragmentation and publishing small collections in one repository boosts their potential dissemination.

The VSP team created a new MySQL relational database to prepare metadata for publication in the digital *ektypotheke* of *E-stampages*. Luigi Tessarolo, the IT technician for the project, developed the software to manage the database—not via a CMS, but directly in PHP within the Zend Framework.¹⁸ Realised between the end of 2017 and spring of 2018, the VSP database was planned as a tool to prepare the dataset for *E-stampages* and to set up a searchable digital archive that could substitute for the FileMaker archive.

Designing a database offers several advantages. For one, it ensures maximum flexibility in modelling metadata architecture and export modes. A database allows data to be easily controlled, checked, and corrected. Input methods such as drop-down menus give access to controlled vocabularies so ambiguous terms and typos can be avoided. Upon finding a mistake, the user intervenes only once to correct wherever it appears. For another, the software managing the database enters and

¹⁸ <https://framework.zend.com>.

manages several data automatically. The software also supports re-documentarisation by generating record identifiers (ID) and connections among records (Antonetti *et al.* "Collezioni", 57-59). This serves to guarantee the functioning of the database and the structure of the dataset as the software automatically manages images, 3D models, and metadata. Finally, the database permits re-use of the dataset from the FileMaker archive, which has become a starting point for a new round of digitisation.

Accessible to project members only, the VSP database is hosted on the Mizar.unive.it server run by the ASIT (Area Servizi e Telecomunicazioni) at Ca' Foscari University.¹⁹ The database reproduces the *E-stampages* metadata architecture composed of the five entities: squeeze, text, artefact (that is, the stone), the 2D images, and the 3D model. The user inputs only information that the software cannot enter automatically. Accordingly, (s)he intervenes in some fields for the squeeze, text, and artefact entities. The implementation process moves from the artefact to the squeeze to ensure the correct relationship between the two. Relations may indeed be complex: an artefact may bear (and thus, be related to) more than one text; and text may be reproduced (and thus, related to) for more than one squeeze. The other entities (the 2D images and 3D models) are implemented automatically, as we will see.

As with the Ca' Foscari collection itself, the database is in Italian and arranged by region. Populated with the dataset from the FileMaker archive, the database generated all instances of the artefact, text, and squeeze entities necessary to describe the items of the Ca' Foscari collection. Then, any instances missing data were updated and completed. When accessing the database, the home screen shows a list of squeezes with some additional information (place of discovery, date of inscription, and key editions). Selecting an entry on this list, the user accesses three subsequent boards that allow him/her to enter data on the stone, the text, and the squeeze: a screen about the artefact leads to one about the text, and then one about the squeeze. This latter screen is divided into two parts that show fields for implementation on the right, and photos of the reverse and obverse of the squeeze on the left. The black-and-white, low resolution images were imported from the FileMaker archive and preserved to support the new digitisation. When available, the 3D model is displayed below the photos.

Three input methods are used to enter data: drop-down menus, checkboxes, and free-form fields. Most data entries are accessed via drop-down menus, and options are selected according to *E-stampages* controlled vocabularies. This method is also used to enter inscription origin and place of discovery. There is no point in establishing a default list of place names as it forms while the cataloguing of squeezes progresses. However, strict controls are needed for the names of places. The spelling of a place name may vary according to time and language. For example, the name Delphi is spelled Delphoi in Ancient and Modern Greek, Delfi in Italian, and Delphes in French. It is easy to create duplicates and in order to avoid this, the VSP database manages the names of places. To search for a place, the user first adds the name to the menu (*viz.* for place of discovery) and then selects the entry from the menu. Thereafter, the user can select the same place from the list of names already recorded as needed.

¹⁹ <https://www.unive.it/data/struttura/111575>.

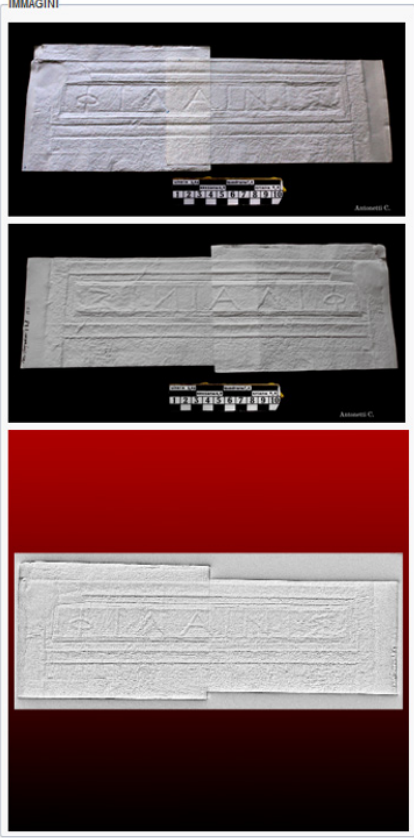
TITOLO
 THYRREION - Stele funeraria di Philainis

Identificativo locale: Thymeion 150
 Data di realizzazione: 1981-1985
 Autore del calco: Antonetti, Claudia
 Materiale: carta
 Larghezza: 34 Altezza: 12 in cm
 N° di linee: 1 N° di fogli: 2
 Calco parziale: ☐
 Stato di conservazione: buono -- Aggiungi --
 Annotazioni sul calco: numero di inventario -- Aggiungi --
 Luogo di conservazione: Università Ca' Foscari Venezia
 Cassetto: 4
 Busta: 2
 Note: Per l'editore, non per l'esportazione
 Esporta il calco in E-stampages ☒

IMMAGINI A BASSA RISOLUZIONE
 Carica recto Carica verso

IMMAGINI AD ALTA RISOLUZIONE

File	kB	Data
calco_thymeion_0150_3D.obj	99435	
calco_thymeion_0150_R.png	14354	5985 x 2394
calco_thymeion_0150_R.tif	10481	5985 x 2394
calco_thymeion_0150_V.png	14354	5985 x 2394
calco_thymeion_0150_V.tif	10032	5985 x 2394
calco_thymeion_0150_VL.png	3469	2817 x 1229
calco_thymeion_0150_VO.png	3589	2992 x 1197



[Fig. 2. VSP database. Sample of the screen of the squeeze. Thyrreion 150 – Funerary inscription of Philainis (CEGO 2, 269). Courtesy of Claudia Antonetti].

Checkboxes allow for selection between two options or the entry of commands managed through Boolean-typed variables. For example, the user ticks a checkbox to indicate whether a stone is complete, fragmented, or lost; or to indicate whether the script is regular or irregular. Databases can easily manage certain information that is lost or unknown. When a piece of information (such as the addressee of a text) is unknown in the VSP database, the user leaves the field empty and the software auto-fills it with the word *ignoto* (unknown). A piece of information can be uncertain as well, and encoding uncertainty is a great challenge to digital humanists. When a piece of information is uncertain in the VSP database, the user ticks the checkbox *incerto* (uncertain) beside a field. If a decree was probably passed in Athens, for instance, the user enters the word 'city' in the field 'issuing authority' and checks the box marked 'uncertain' next to it. When exporting data, uncertainty can be translated in several ways. According to *E-stampages* guidelines, uncertain data appear in square brackets. As checkboxes also serve to enter commands, they can be used to indicate whether a squeeze should be exported to *E-stampages*. In some cases, an inscription has more than one squeeze as there may be duplicates or reproductions of part of a

stone. These are included in the VSP digital archive as part of the Ca' Foscari collection, but not published on the *E-stampages* website. After entering data, the user ticks the checkbox 'export to *E-stampages*' if the squeeze needs to appear in the online collection.

Drop-down menus and checkboxes are used to enter most information as they allow for data control. The VSP database requires a few free-form fields as well: the title and the note fields. According to the requirements, each instance of the artefact, text, and squeeze entities must include a title that contains information needed to identify the entity record. The artefact title, for instance, indicates the place of origin, object type, text type, and key edition (*viz.* the item 'Thermos 16' in the Ca' Foscari collection has the artefact title 'THERMOS - Stele con dedica di una associazione religiosa a Dioniso : IG IX 1(2) 1, 117').²⁰ In the VSP database, the field 'notes' allows for additional remarks about specific aspects. While irrelevant to digitisation, these notes may record details significant to the study of the document.

The FileMaker archive includes oblique light photos of the squeezes. When the second phase of the VSP began, a new image archive needed to be produced according to the technical specifications of *E-stampages*. Most of the images for long-term storage, online publication, and 3D models were produced using the BAUM (Biblioteca di Area Umanistica) planetary scanner at Ca' Foscari University²¹. In October 2019, AMS Archive Services LTD (Athens) digitised approximately sixty squeezes that could not be captured by the scanner.²² Due to the large number and dimensions of the photos, the image archive is not stored on the server. Instead, it is stored on a hard disk along with a back-up copy on a cloud service (Google Drive). Images are organised in folders containing both the images and 3D model of each squeeze. Java software then reads and exports the image data to a CSV file. Finally, the file is imported into the VSP database via FTP (File Transfer Protocol).²³ This process creates the entities of the 2D images and 3D reconstructions.

Once the dataset is complete, the database is ready to prepare data for the *E-stampages* digital collection and to support the research and teaching activities of epigraphers at Ca' Foscari University. Exporting data from the database is easy and quick: the VSP database has an export function that can be applied to the whole dataset or just part of it (such as data on squeezes of inscriptions from Thermos). Data are exported in five CSV files as each file contains data on one entity. For images, the Java software copies files shared with *E-stampages*. An advanced search tool has been set up to browse the database. Users can set a range of one to ten search criteria (including inscription place of origin, discovery, and preservation; text type; object type; and edition) and explore data via a search bar.

²⁰ <https://www.e-stampages.eu/s/e-stampages/item/5595>.

²¹ <https://www.unive.it/baum>.

²² AMS Archive Services LTD: <http://www.scanning.gr/el/>. The digitisation of these squeezes was funded by the École française d'Athènes.

²³ FileZilla Client was used: <https://filezilla-project.org>.

Fig. 3. VSP database. The search tool. Courtesy of Claudia Antonetti.

5. Future developments

The VSP aims to promote recognition of epigraphic squeezes as historical and documentary evidence. In this sense, it supports the promotion of other collections of epigraphic squeezes. When the project began, a few Italian collections were known to exist.²⁴ No up-to-date information about the extent, preservation, and organisation of these collections was available. Due to a long tradition of study in Greek and Latin Epigraphy, they were supposed to make up only a small portion of the collections preserved in Italian institutions. For this reason, the VSP team surveyed scholars, universities and archaeological museums to bring ‘forgotten’ collections to light.

The results still preliminary, so remarks on the scope, state of cataloguing, and preservation of these collections are necessarily limited. So far, the survey has reported nine collections. The most significant include from 160 to around 250 squeezes, while the small ones contain just a few specimens. Four collections are entirely uncatalogued, and one is partly catalogued. For three collections, the number of items is unknown. These results are not surprising as squeezes often lie neglected in the closet of a department or museum for years. They are sometimes even forgotten, as illustrated by the Palazzo Altemps collection in Rome that was quite large (174 squeezes) and found by chance behind a closet a few years ago.²⁵ The surveyed collections are mainly comprised of paper squeezes, but two contain plaster and latex squeezes: La Sapienza has 120 plaster squeezes, while the Scuola Normale Superiore collection in Pisa consists of 49 latex squeezes.²⁶

Scholars recognise the importance of the epigraphic squeezes, but they lament the barriers facing initiatives to safeguard and enhance squeezes. It is difficult to find support from institutions as a cost-benefit analysis discourages the digitisation of small collections. Creating and managing an online collection is expensive. Publishing each collection on its own website increases costs and reduces visibility. Although a digital archive can be useful for collection management, digitisation without online publication cannot be taken into account.

²⁴ The collections of La Sapienza University (see Bevilacqua) and the Scuola Normale Superiore.

²⁵ Claudia Antonetti and I thank Silvia Orlandi for this information.

²⁶ Claudia Antonetti and I thank Francesco Camia and Francesco Guizzi for information about the collection of La Sapienza University, and Carmine Ampolo and Anna Magnetto for that on the collection of the Scuola Normale Superiore.

The existence of a network, tools, and metadata standards would foster new initiatives to safeguard and enhance the archives of epigraphic squeezes. Thanks to support from the Venice Centre for Digital and Public Humanities,²⁷ the VSP is developing a web application that will be made available to those who wish to create their own digital collection. A beta version of the web application in Italian and English is nearly ready. It will be tested by scholars from the Archeological Mission in Cyrenaica, and the Universities of Bologna, Macerata and Tor Vergata.²⁸

The new application consists of web-based software running a database that complies with *E-stampages* standards. It is modelled on the VSP database but overcomes its technical shortcomings. The most relevant novelty concerns image management. As detailed earlier, the VSP image archive is not stored on a server but on a hard disk; data transfer from the archive to the database requires local software and an FTP connection. In the future, images will be archived on cloud storage linked to the new application. The application will read and import image data into the database automatically. As with the VSP database, users will enter information. Once the data is entered, the database will serve as a digital archive that will contribute to the preservation and everyday use of the collection. Institutions that want to publish digital collections online will be able to export both the dataset and the images. Collections may be published on the *E-stampages* website as the data and images comply with the standards of the French programme. Alternatively, institutions may choose to publish collections on their own websites. This application is conceived as a flexible tool that complies with high-quality digitisation standards. It meets the needs of those who wish to digitise a collection for preservation and management purposes, and of those who wish to publish them.

The survey also brought the existence of plaster and latex squeezes to light. Their digitisation requires a protocol that takes their specific features into account. Descriptive metadata may remain unchanged—perhaps only an additional field for the thickness of the squeeze is needed. But the DEA software works only for paper squeezes, so a 3D model cannot be produced. The VSP team is going to establish a collaboration with IUAV University of Venice²⁹ to explore other 3D modelling techniques for squeezes made of paper and/or other materials.

6. Conclusion

Initiatives such as the VSP and *E-stampages* attest to fresh interest in epigraphic squeezes. The VSP intends to support the rise of an Italian network for the preservation and enhancement of epigraphic squeeze collections by offering institutions a web application to launch their digitisation programmes. Positive response to the survey of Italian collections indicates lively interest; hopefully, other

²⁷ <https://www.unive.it/pag/39287>.

²⁸ The beta version was scheduled for presentation at a two-day workshop in Venice held in May 2020, which was postponed indefinitely due to the 2020 health emergency. With participation from Michèle Brunet and the VSP team, the workshop would have presented the outcomes and future developments of the VSP and *E-Stampages*.

²⁹ <http://www.iuav.it/homepage>.

collections will be digitised in coming years. As with the VSP and *E-stampages*, these programmes will aim both preserve squeezes and make them available for scholars and students. They will contribute to safeguarding archives of epigraphic squeezes, recognising their documentary and historical value.

Ongoing and future programmes acknowledge epigraphic squeezes as heritage that should be protected, as through digitisation. This operation generates a digital reproduction of an object (a piece of heritage). In meeting high-quality standards, the digital reproduction itself becomes heritage. Accordingly, we now face a two-fold challenge: safeguarding squeezes as both tangible and digital heritage. Safeguarding archives of epigraphic squeezes is likewise essential for guaranteeing the long-term survival of current and future digital collections. In this sense, the creation of a network of institutions could be fundamental.

Bibliography

- Antonetti, Claudia, Michèle Brunet, and Eloisa Paganoni. "Collezioni di calchi epigrafici: una nuova risorsa digitale". *Axon*, vol. 3, no. 2, 2019, pp. 41–66, <https://edizionicafoscari.unive.it/media/pdf/article/axon/2019/2/art-10.14277-Axon-2532-6848-2019-02-004.pdf>
- Antonetti, Claudia, Stefania De Vido. "Digital Epigraphy at the Greek Epigraphic Laboratory, Ca' Foscari University of Venice". *Historika*, vol. 7, pp. 491–502, <https://www.ojs.unito.it/index.php/historika/article/view/2608/2442>
- Bevilacqua, Gabriella. "Da Federico Halbherr a Luigi Moretti, il percorso dell'epigrafia attraverso i calchi epigrafici della Facoltà di Lettere e Filosofia" dell'Università di Roma 'La Sapienza'. *Mediterraneo Antico*, vol. 16, 2013, pp. 563–582.
- Bozia, Eleni, Angelos Barmoutis, and Robert S. Wagman. "Open-Access Epigraphy. Electronic Dissemination of 3D-digitized Archaeological Material". *Information Technologies for Epigraphy and Cultural Heritage. Proceedings of the First EAGLE International Conference*, edited by Silvia Orlandi, Raffaella Santucci, Vittore Casarosa, Pietro Maria Liuzzo, Roma: Sapienza Editrice, 2014, pp. 421–435, <https://www.eagle-network.eu/wp-content/uploads/2015/01/Paris-Conference-Proceedings.pdf>
- Levivier, Adeline, Elina Leblanc, and Michèle Brunet. "E-STAMPAGES: archivage et publication en ligne d'une ectypothèque d'inscriptions grecques". *Les nouvelles de l'archéologie*, vol. 145, 2016, pp. 24–27, <https://journals.openedition.org/nda/3801>
- McLean, Bradley H. *An Introduction to Greek Epigraphy of the Hellenistic and Roman Periods from Alexander the Great down to the Reign of Constantine (323 B.C.-A.D. 337)*. Ann Arbor: The University of Michigan Press, 2002.
- Paganoni, Eloisa, Stefania De Vido, and Claudia Antonetti. "Il Laboratorio di Epigrafia Greca dell'Università Ca' Foscari. Una fucina didattica per l'epigrafia greca". *8th Annual Conference AIUCD 2019. Udine, 23 – 25 January 2019. Teaching and Research in Digital Humanities' Era. Book of Abstracts*. Udine, 2019, pp. 193–195, <http://aiucd2019.uniud.it/book-of-abstracts>.

Publication, Testing and Visualization with EFES: A tool for all stages of the EpiDoc XML editing process

Gabriel Bodard and Polina Yordanova
University of London, University of Helsinki

Abstract: EpiDoc is a set of recommendations, schema and other tools for the encoding of ancient texts, especially inscriptions and papyri, in TEI XML, that is now used by upwards of a hundred projects around the world, and large numbers of scholars seek training in EpiDoc encoding every year. The EpiDoc Front-End Services tool (EFES) was designed to fill the important need for a publication solution for researchers and editors who have produced EpiDoc encoded texts but do not have access to digital humanities support or a well-funded IT service to produce a publication for them.

This paper will discuss the use of EFES not only for final publication, but as a tool in the editing and publication workflow, by editors of inscriptions, papyri and similar texts including those on coins and seals. The edition visualisations, indexes and search interface produced by EFES are able to serve as part of the validation, correction and research apparatus for the author of an epigraphic corpus, iteratively improving the editions long before final publication. As we will argue, this research process is a key component of epigraphic and papyrological editing practice, and studying these needs will help us to further enhance the effectiveness of EFES as a tool.

To this end we also plan to add three major functionalities to the EFES toolbox: (1) date visualisation and filter—building on the existing “date slider,” and inspired by partner projects such as Pelagios and Godot; (2) geographic visualization features, again building on Pelagios code, allowing the display of locations within a corpus or from a specific set of search results in a map; (3) export of information and metadata from the corpus as Linked Open Data, following the recommendations of projects such as the Linked Places format, SNAP, Chronontology and Epigraphy.info, to enable the semantic sharing of data within and beyond the field of classical and historical editions.

Finally, we will discuss the kinds of collaboration that will be required to bring about desired enhancements to the EFES toolset, especially in this age of research-focussed, short-term funding. Embedding essential infrastructure work of this kind in research applications for specific research and publication projects will almost certainly need to be part of the solution.

Keywords: Text Encoding, Ancient Texts, Epigraphy, Papyrology, Digital Publication, Linked Open Data, Extensible Stylesheet Language Transformations

1. Background

The benefits of online publication are well understood and sought after by epigraphists.¹ EpiDoc is a tool-set and a community of practice for the digital encoding of edited ancient texts, including inscriptions, papyri, seals, coins, and related objects, in TEI XML. EpiDoc is used by dozens of projects relating to the classical world and beyond, and hundreds of people worldwide have been trained in the use of EpiDoc practices and tools.²

However, after having invested a substantial amount of effort and time in encoding their materials in a digital format, researchers are often met with another obstacle before publishing. The transformation of EpiDoc XML documents and their visualization online require a different set of advanced technical skills, or the support of a dedicated IT unit or a development team, which might ultimately render publication without substantial institutional support or project funding a daunting challenge.

The increasing demand from the epigraphic community for a tool that is free, specifically designed to reflect the research and encoding practices of the discipline, but also customizable to fit the particular requirements of the individual projects, and accessible to users without advanced technical knowledge, was the prompt behind the creation of the EpiDoc Front End Services (EFES) from 2017.³

EFES was aimed at facilitating the creation of a rich output similar to those of some of the prominent epigraphic projects such as the Ancient Inscriptions of the Northern Black Sea (IOSPE), Inscriptions of Roman Tripolitania (IRT), Inscriptions of Aphrodisias (IAph), and Roman Inscriptions of Britain (RIB), which are all large-scale projects with extensive technical support or in-house expertise.⁴ The platform was designed to provide the components that are theoretically a common interest for most publications of epigraphic material—display of individual texts in interpretive and diplomatic editions along with their historical and descriptive data, multiple indices, browse and search interface, and possible export of information through Linked Open Data, and make them accessible to the non-technical user via an “out-of-the-box” publishing platform.

¹ The authors would like to thank Elina Boeva, Martina Filosa, Tamara Kalkhitashvili, Jamie Norrish, Charlotte Roueché, Alessio Sopraca, Simona Stoyanova, Charlotte Tupman, Irene Vagionakis, Valeria Vitale and the journal's anonymous reviewer for invaluable comments on a draft of this essay.

² EpiDoc: Epigraphic Documents: epidoc.stoa.org.

³ EFES: EpiDoc Front-End Services (University of London, 2018–20): github.com/EpiDoc/EFES.

⁴ These corpora and other digital resources are listed at the end of the paper, alongside the bibliography.

As base for building a publication platform to match the desiderata of epigraphers, the developers chose the Kiln application, developed and used at the King's Digital Lab, King's College London.⁵ Since 2010, Kiln—and its predecessor under another name—has been used in the creation of over fifty digital humanities projects, among which the above-mentioned IOSPE.⁶ The platform is designed to work with content primarily in the form of XML and is itself largely written in XML, allowing users who are familiar with editing XML to perform customization, templating and project-specific code adjustments.

Kiln is built on Apache Cocoon and integrates several independent components—Apache Solr as a searching platform, Sesame 2 as an RDF store, built-in Jetty web server, XSLT-based templating system with inheritance infrastructure, and Foundation, a front-end framework based on HTML, CSS and JavaScript. Thanks to its specific file structure, which separates different components into distinct files and directories, Kiln provides a 'batteries included' experience, allowing the user to see an organized display of the XML content by simply placing the documents in the content directory, and easily to index and harvest the data through the online administrative panel. The general framework is put in place for more project-specific customizations, such as Schematron validation of XML files and infrastructure for multilingual sites. In its EFES incarnation, Kiln has been supplemented with specific display templates for EpiDoc XML, further search facets, and indexing features. Kiln comes with documentation on each of its main functional components, and a tutorial that guides the user through the process of setting it up and customising its various parts.⁷ A user guide has been created for EFES, covering its specific EpiDoc-related enhancements and additions (Yordanova).

With Kiln chosen as the foundation of the publication platform, we employed Jamie Norrish, one of the specialists involved in its creation and development, to build upon its core components a specification suited for the particular needs of epigraphists and papyrologists. In defining the parts shared by philological editions that needed to be provided with the EFES package as the base of the platform, Norrish worked in close collaboration with Yordanova, who brought in the perspective of the epigraphists regarding the *sine qua non* of an online publication. Yordanova had had previous experience in epigraphic projects applying the EpiDoc standards of encoding and conducted multiple consultations with EpiDoc developers and scholars who had expressed interest in working with the platform, in order to determine the most essential common features for projects with different backgrounds.

We used as the starting point several existing projects in the field, and discussed with a selected group of stakeholders what features of these projects they thought would be applicable to their material, what they would need to handle differently in accordance with the specifics of their data, and also what the crucial elements in terms of representation were for their respective projects. These features

⁵ Kiln (King's College London, 2012–2019): github.com/kcl-ddh/kiln. Kiln Documentation (2012): kiln.readthedocs.io/en/latest/.

⁶ Kiln Documentation (2012), "Projects": kiln.readthedocs.io/en/latest/projects.html.

⁷ Kiln Documentation (2012), "Tutorial": kiln.readthedocs.io/en/latest/tutorial.html.

were then added in iterations, using the responses we received from stakeholders, learning from feedback we received during the first training workshop in London in 2017 at which EFES was introduced to epigraphists, and building on user experience in order to improve the platform. This cycle was repeated with a second training event in Sofia, Bulgaria, later in the year.

Thus EFES was first equipped with discipline-specific customizations in the form of templates for some of the most common edition layouts, represented by the EpiDoc Reference Stylesheets (Elliott *et al.*, *Stylesheets*), code infrastructure for indexing with pre-existing matrices for some indices that could be populated with project data, potential for RDF export, and search facets based on categories shared by most EpiDoc projects. Further refinement with project-specific customizations was expected to be easily achievable by users who are well-versed in XML, but do not necessarily have other advanced technical experience, on the basis of training and documentation made available.

1.1 Status quo

The vast majority of the specified technical objectives of the EFES tool were therefore met during the first phase of the project: the platform can be easily downloaded and installed on any operating system, allowing users to add their own EpiDoc files into a specified folder and view a basic transformation of their texts in the web browser. Several customized templates for viewing the texts can be selected, via existing parameters in the EpiDoc Reference Stylesheets. If EpiDoc recommendations for tagging have been followed, several core features of the editions will be indexed, and search results will be filtered according to the selected facets. Instructions for selecting pre-existing templates, indices and facets for the core display, and for creating new ones, are well-documented and have proved easy to follow. A few active projects (Inscriptions of Roman Cyrenaica, Telamon: Greek Inscriptions from Bulgaria, SigiDoc: text encoding for Byzantine sigillography, Epigraphic Corpus of Georgia, Epigraphic Database Vernacular) are now using EFES as part of their pre-publication editorial workflow. They have in particular reported that one benefit of the system is its use to make evident and diagnose encoding errors in their electronic texts as they go along.

```
<div type="edition" xml:lang="grc" xml:space="preserve">
  <ab>
    <lb n="1"/><placeName ref="#lato" type="ethnic"><supplied reason="lost">Λάτιο</supplied>₁</placeName>
      <persName type="divine" key="Aphrodite">Ἀφροδίται</persName>
    <lb n="2"/><w lemma="νικάω">νικάσ<g type="Λ"/>αντες</w>
    <lb n="3"/><rs type="institution" subtype="tribe" role="eponym" key="Aischeis" ref="#lato">ἐπὶ τῶν
      <w lemma="Ἀισχέϊς">Ἀισχέων</w></rs>
    <lb n="4"/><placeName ref="#olous" type="ethnic">Βολωντίος</placeName>.
  </ab>
</div>
```

Figure 1: Example of encoded text from an EpiDoc file (© 2020 Vagionakis)

Divinità, ninfe ed eroi	
Le parentesi quadre indicano che il nome è o parzialmente o totalmente integrato.	
Nome	Occorrenze
Acheloios	ic1_17_7.1 [ic1_17_7.3]
Achilles	ic3_4_37.4
Amphitrite	chaniotis_61b.180 [ic1_16_5.74] [ic1_17_1.6] [ic1_17_1.18] ic1_30_2.5 [ic2_16_2.4] [seg_26_1049.83] [seg_33_638.3]
Aphrodite	chaniotis_54-56a.12 chaniotis_54-56d.1 chaniotis_61b.162 chaniotis_61b.182 ic1_14_2.3 ic1_16_18.7 ic1_16_25.10 [ic1_16_27.8] [ic1_16_5.70] [ic1_16_5.75] [ic1_18_9.C.6] [ic1_22_2.4] ic1_8_4.b.14 ic1_9_1.a.27 ic3_3_3ab.B.14 ic3_3_5.14 [ic4_171.15] [ic4_174.59] ic4_174.75 [ic4_183.21] [seg_23_547.B-C.52] [seg_26_1049.63] [seg_26_1049.84] [seg_33_638.4] [seg_41_743.13] seg_54_841.6 seg_60_996.A.3 [seg_61_722.B.9]
Apollo	chaniotis_55b-56b.11 [ic1_16_3.20] ic1_16_4.A.15 [ic1_19_2.8] ic1_23_1.17 ic1_5_20a.II.19 ic1_8_12.48 ic1_8_6.8 ic2_1_2.B.23 ic3_3_3ab.B.11 [ic3_3_3ab.B.12] [ic3_4_43.5] [ic4_3.2] ic4_51.2 [id_1602.4] ig_iv2_1_96.11 ig_ix2_3_783.2 seg_28_746a.4 [seg_45_1528.2] syll3_560.33 syll3_737.22

Figure 2: Example of index in EFES based on encoding in fig. 1 (© 2020 Vagionakis)

Where our aims were less well met, however, was in the area of the ease of customization and deployment by a non-technical user. Of the approximately forty people to whom we offered training in the processes of creating and customizing indices and search facets for EFES in that first year, three individuals (all exceptionally motivated and having access to personal support), plus the five project teams mentioned above, have gone on to deploy the platform and were able to master the skills required to create their own display, index, search and similar features. They all report finding EFES extremely useful as an editing and testing environment—viewing the results of the encoding in real-time—even if not all are quite ready to deploy it for publication. While this is a pleasing result in its own right, feedback from the vast majority of users, some of whom left the training workshops before completing the week of practice, was that the coding and technical skill required for creating new indices was too steep a learning curve for most editors, even those who have acquired facility in editing EpiDoc XML. On aggregate this was a disappointing outcome, and we came to the conclusion that the target user of the EFES platform (a non-technical editor who had created an EpiDoc corpus without institutional support) could not easily on her own install and customize the tool to create an online publication, even with the help of the very thorough documentation and user guide.

Our intention is now to mitigate this shortcoming in the usability of the EFES platform in two ways. Firstly we need to rethink the approach whereby EFES provides a bare-bones handful of example indices and search facets by way of demonstration, and users are expected to customize the platform themselves to add any further indexing functionality desired for their project. The core version of EFES included indices for: symbols, abbreviations, fragments, numerals, words (base form or lemmatised) and personal names. The *Cretan Institutional Inscriptions* (CII) project added indices for prosopography, divinities, toponyms and institutions. The hope is that other projects will further equip the distribution version of EFES with twenty or so

further indices, to include: bibliographical concordance; names, people and places all subdivided into multiple criteria and subgroups; as well as categories of text, object, material, offices, languages, calendars and the like. Between them, these indices should cover the majority of basic needs of most classical and other ancient corpora, as identified in the outreach phase of the initial EFES project.

Further customization will always be necessary for less common indexing and searching features, but this is no longer likely to be a blocker to the start of a project, or even in some cases to publication itself. The indices covered will include some features that would be challenging for even a confident developer to implement independently, such as lemmatized word searching, sorting of search results, and the handling of multiple search terms in a single XML tag (such as a word with two simultaneously valid lemmata). There are already (as of 2020) concrete proposals to include elements of this future development in project applications for SigiDoc (Sopracasa-Filosa) and the Epigraphic Corpus of Georgia (Doborjginidze). We are also working with a newly identified community of scholars working on Hittite and other Cuneiform inscriptions in EpiDoc XML to develop new desiderata and needs in this related subdiscipline;⁸ this initiative should additionally help to track the usability of the new indices and documentation going forward.

1.2. A workflow tool

One of the barriers to the adoption of digital editions, including but by no means exclusively epigraphic corpora, as the standard for philological publication has been the lack of suitable software that supports the workflow developed by editors who were trained in the compilation and publication of printed editions (Rosselli Del Turco 227; Burghart §1.1). EFES aims not only at providing an easy-to-use solution for publishing the final digital corpus, but also at addressing some of these steps in the process of preparing the edition.

The steps involved in the traditional editorial workflow include, not necessarily in this order and certainly with multiple iterations of some aspects: fieldwork for assembling materials; outlining the questionnaire required to complete individual records; critical treatment and addition of information; data management; discovering patterns and preparing historical commentary; checking for consistency and completeness; and finally publication. The production of a digital edition in EFES is able to mirror several of these steps, while the separation of labour made possible by the specific file structure of the platform allows for these processes to be running not only as series, but also in parallel.

Data management in EFES is available directly out-of-the-box by placing the EpiDoc XML files into the ‘content’ directory in the files system and indexing them through the admin panel. This gives immediate access to the list of documents in the project and reports any obvious technical issues, since the platform will return errors for ill-formed documents and other unexpected features. The platform can also be

⁸ HittiDoc (ed. Gabriel Bodard, İlgi Gerçek, Katherine Shields). 2019-20. “Hittite and Cuneiform in EpiDoc.” github.com/EpiDoc/HittiDoc.

customized to validate the files against a specialist schema to ensure consistency of the annotation throughout the project. Since XML content, controlled vocabularies, EpiDoc transformation stylesheets, and CSS files and libraries are all kept separately, the EpiDoc encoding of the documents, creation and maintenance of controlled vocabularies, editorial decisions for organizing and displaying the data, and graphic design of the site can all be performed simultaneously or in parallel, perhaps by different team members, and will complement one another.

Although EFES is therefore highly customizable, allowing users to shape and specify it locally according to the particular needs of each project, the platform provides a default structure based on EpiDoc encoding practice. Thus, projects making their first steps into digital epigraphy can have a solid frame around which to create their content. Many of the pre-made indices in EFES are generated on the basis of authority lists, controlled lists of terms or names, which are XML documents that give all items unique identifiers and may include other information, variant spellings or localisation. These lists may be defined *ad initio*, but can also be built retrospectively or iteratively throughout the preparation of the edition, helping maintain the consistency of vocabulary and terminology. The information stored in the authority lists can be exported to RDF and shared with other projects. This has the potential to encourage the development and maintenance of community standards, as well as working towards data compatibility, sharing and reuse.

The possibility to visualize each document and the indices generated from the markup and through references to the controlled vocabularies, proves useful in keeping an overview of the corpus as new texts are added. It is also a tool for validation and “sanity checking” while assembling the individual edition. The process of annotating and indexing may for example aid the editor in finding patterns and structuring the content, say into chapters or sections, by grouping and visualising relevant criteria.

Most editors will appreciate that access to both philological and data querying and visualisation tools are an essential part of the editing process, not just an added feature for the final publication. As an editor of an early EpiDoc project (who shall remain anonymous) complained, “Our readers can now go and find all the errors in my work, thanks to the tools we’ve created for them, to which I didn’t have access while writing it!” Using a tool like EFES as part of the workflow of editing a corpus both enables the editors to verify their own analysis and output at every stage in the process, and potentially offers the opportunity to pre-publish and make provisional results available for peer review at an early stage, so that the final publication both benefits from community contributions and more closely suits the needs of scholarly readers and users.

2. Visualization possibilities

One of the core priorities for EFES is to produce user-friendly display of indexed information using tools that are both generic and robust enough to be deployed by a non-technical editor. This includes not only static indices that reproduce the display from a conventional print book, and faceted search filters, but more novel and dynamic

visualizations, such as maps, sliders and customisable tabular data. For those indices that are generic enough to be deployed “out of the box” for a majority of corpora, the editor would ideally be able to select, activate and deactivate options from a graphical user interface, without needing to edit stylesheets, XML templates or Cocoon sitemaps in most cases.

The main visualisation tools that we would like to see added to the EFES platform, which are likely to be of value to the largest number of text corpora, ancient or otherwise, include:

1. a date slider that both displays the distribution of search results as a histogram, and allows the selection of a data range as a search modifier;
2. geographical visualisation to display search results on a map, with links back to individual text editions, and identifiers from a standard gazetteer;
3. the ability to export search results and other data in tabular or graph form, in as many open standards as possible, so that external tools that are better suited to the handling of data or particular user needs may be employed to analyse or display results as needed.

Even with all of the enhancements suggested above, from a larger menu of indices and search facets, to visualisation tools and export formats, EFES would still be a relatively bare-bones platform. Inevitably, any range of features implemented will be but examples of the possibilities enabled by digital publication of textual corpora, and its greatest strength remains the extensibility of the platform so that editors can create new features to serve their unique needs.

2.1 Date slider

One of the most widespread and useful means to organize, sort and filter an epigraphic corpus, or any other large collection of historical texts, is by date of creation. To a lesser degree, dates of discovery, accession, publication, and for that matter other dates in the history of the text or its support may be used to filter or navigate a corpus. Conventional epigraphic or papyrological editions are very often sorted by date, perhaps within other criteria such as locations or categories of text, and almost every historical corpus is explicitly delimited by chronological extent in some way, whether by editorial choice or as a result of the occupation or epigraphic habit of a site.

Dates as an item of metadata for a textual edition may be expressed in several fundamental ways (and arguably one edition might choose several of all of these forms of data expression in parallel, for different purposes), including:

- Textual expression of the date: “Late first century,” “presumably within *circa* fifty years of the death of Hadrian,” “third century, but not before the citizenship edict of Diocletian in 212,” etc. For human-readability by scholars, such an expression contains information, including argument, that cannot be replaced by any numerical or computational formula.
- Dates gathered into named periods, defined by more or less granular start and end dates: “Hellenistic,” “Byzantine,” “Julio-Claudian,” “Middle Bronze Age” and so forth. There are authority lists and ontologies of such named

periods, which reflect the contingency of naming, geography, scholarly tradition, and other features on the dating of said periods. See for instance the PeriodO project and ontology (Rabinowitz).

- The original date expression in the text: “the 23rd day of Hathur, year 3,” “in the archonship of Phainippos the second,” “the 14th year of the reign of Antoninus,” “after the death of Sekhemre Shedtawy.” These may or may not be resolvable to Gregorian dates by modern scholars. For instance GODOT project supplies identifiers for ancient date and calendar expressions (Grieshaber 2017).

Arguably the most powerful format for capturing dates for searching, sorting or filtering historical documents, is to express the dates (where possible, of course) as numerical data, capable of having mathematical functions performed upon it. A date expressed numerically can be sorted, can be used to calculate periods (“how long did this reign last?”), to find objects within a date range (“-0100 – 0100”) or within a certain number of years of a fixed point (“0212±25 years”), and so forth. Once all of these dates are in a single format—Gregorian dates, including proleptic Gregorian years for events in the ancient world, are the conventional rendering—such operations and calculations are computationally trivial. One of the most elegant ways both to express the parameters and represent the results of such a date search or browse is via a visual slider with start- and end-points that can be moved along the range of available dates in the corpus. The standard search interface of projects built using the Kiln tool, upon which EFES is based, includes a basic date slider, coded with CSS and Javascript. Although this date slider does not recalibrate and scale when the time-range of the search results is limited by the application of other facets in the search interface, making it misleading as a visualisation of date, it is still highly intuitive to use and easier than entering numerical dates in text fields.⁹

The industry standard for the date slider view and input field, however, is exemplified by the dynamic search results visualisation in the Pelagios Project’s geographical search and visualization tool, Peripleo.¹⁰ In this view, the results of a search are shown, in addition to the weighted geographical distribution on a map, by a simple histogram in the date slider; that is, a series of bars showing the relative proportions of results by chronological period. When the endpoints of this date slider are manipulated by the user, the search results and geographical display dynamically update, showing only those results within the new date criteria.¹¹ The advantages of this implementation include the fast-updating dynamic display, and the visually pleasing and intuitive histogram. One could criticise the lack of a horizontal scale on

⁹ See for example the Ancient Inscriptions of the Northern Black Sea (IOSPE) project search page iospe.kcl.ac.uk/search/en/-500/1800/ and the People of Mediaeval Scotland project search page poms.ac.uk/search/.

¹⁰ *Peripleo: The Pelagios Exploration Engine*. (Rainer Simon *et al.* Austrian Institute of Technology, Exeter University, The Open University, University of London, 2016–2019). peripleo.pelagios.org. The “Timerange Selector” is a free-standing open source tool published in 2017, with code at github.com/pelagios/timerange-selector.

¹¹ An example search showing off the features of the Peripleo date slider for results for the query string “defixio” can be seen at: peripleo.pelagios.org/ui/#q=defixio&filters=true.

the date slider, other than the endpoints, and therefore the unclear granularity of the bars in the histogram, but it remains a very elegant, usable visualisation widget.

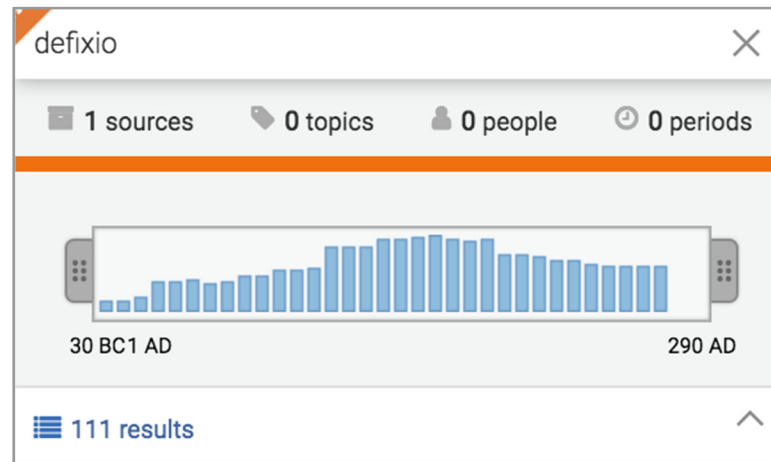


Figure 3: Peripleo time range selector, showing results for “defixio” (© Pelagios Project)

It is a clear priority to explore extending the EFES code-base with the implementation of a Peripleo-like data slider visualisation, especially since the existing code in a partner open source project makes it somewhat low-hanging fruit. We believe that a project intending to customize EFES for their own publication could implement this rather quickly, and ideally contribute the code back to the EFES Git repository for the benefit of the community.

2.2 Geographic organization and visualization

Geographical or spatial distribution of information is arguably the most universal and compatible criterion for organising and dividing bodies of ancient materials, including but not restricted to objects bearing texts, such as epigraphic, papyrological or numismatic editions. It is inevitable that place and space will be among the most important features of any corpus (as arguably for any historical dataset).

It is certainly not the aim of EFES to include full GIS database functionality; many existing tools will perform these functions more effectively, and they may be integrated with the online presentation of a text database or XML corpus in a variety of ways. Given that geographical or spatial data form such an integral part of the organisation of most corpora, and will accordingly be highly desirable as indexing and search features of a digital publication, the visualisation of space and place should form part of the features of the generic platform.

An ancient epigraphic corpus typically may contain several categories of information about the geographic context of the texts and text-bearing objects within it, perhaps including:

- The original or other attested or assumed ancient location of the object, such as a gravestone presumably originally erected in the necropolis found re-used in a late antique fortification wall. A text may also have (one or more) ancient places of origin independent of its object, as in the case of a letter authored in Rome and copied in the provinces, or a copy of a contract placed in an archive.
- The modern place of finding, which although often a proxy for or the best evidence for the original location, is technically a different category of information—and a different category of place, being a modern location usually with coordinates as opposed to a historical settlement or place of known or unknown original name.
- Other modern places and locations in the provenance of the object, including repositories that may hold or have held it, and its current or last known location.
- Places explicitly named or otherwise referenced in the text, whether evidence for the location of the object or further context for people and events mentioned in it.

Some of these places may be principally expressed in a digital edition by means of geographic coordinates in a database, in particular if place of finding or other attested modern locations are captured by a GPS-enabled camera or similar technology. More often, however, a controlled list or gazetteer of named (or otherwise) settlements, monuments and other relevant places is likely to be preferred. Disambiguating places and locations in a text edition or its associated data record to an internal authority list or public gazetteer helps to avoid redundancy and potential errors in the data, and allows common information to be stored in a central or shared resource. Most fundamentally, this would mean that the text records and index files only need to supply the place identifier, while the gazetteer will be the source of coordinates, extent and other geographic information needed to display the items on a map or other visualisation.

In addition to structured text indices and search facets, any or all of these geographical entry-points into the epigraphic data may be sorted, visualised or queried via some form of map interface. The range of texts or objects may be displayed on a map showing distribution, weighting, network of relationships, movement or time. Points or regions on the map may be grouped by chapter, century, type of text, object, or essentially any other facet that is also used to index or search the concept. Ideally the map would be an entry-point into the individual text editions, as well as an overview and visualisation: each item on the map would include a link to the page for the text, as would an entry in an index.

It is our aim to avoid the reinvention of the wheel as far as possible in the development of features in EFES, and of course when it comes to geographical visualisation of ancient and historical data, there is much prior art to draw upon, of which we shall mention only a small selection here.

1. As the gazetteer of record for ancient places, Pleiades (originally founded at Ancient World Mapping Center at UNC but now run by the Institute for the Study of the Ancient World at NYU) captures spatial, historical and philological information, including relationships between places, and provides visualisations of some of these features and relationships.¹² As well as being the key example of a digital gazetteer, which provides the canonical identifiers and key spatial information about ancient places, Pleiades has been at the centre of a community of practice for the querying, visualisation and sharing of such information (Elliott-Gillies).
2. Another project originating at the Ancient World Mapping Centre at UNC is *Antiquity à la Carte*, a custom, web-based map-building tool targeted at scholars wanting to create, customise and publish maps of the ancient world; the platform enables the selection of layers, features and labels to build maps for different delivery media, and is a very useful and user-friendly, entry-level visualisation tool.¹³
3. The Digital Atlas of the Roman Empire (DARE) is a parallel project to, and close collaborator with, Pleiades.¹⁴ As well as providing canonical records and identifiers for ancient locations and places, DARE is the source of a set of widely used historical map tiles showing landscape features such as coastlines, rivers and woodland as they were in the Roman period.¹⁵
4. The World Historical Gazetteer is a collaborative project to build a database of historical placenames, aligned to existing and new gazetteers, along with a web platform for querying and displaying the records, and an application programming interface for access to shared and interoperable data.¹⁶
5. The Pelagios Project ran from 2011-19, and built a large community of practice around the representation of historical spatial information as Linked Open Data. Among its key achievements of relevance to this discussion are the Linked Places Format (on which more below), the Recogito geographical annotation and lightweight visualisation tool, and Peripleo (mentioned above), a search and visualisation tool for historical geographical information. As of 2020 the Pelagios Project has been succeeded by a Pelagios Network of independent partners who commit to taking the goals of the community forward (Simon *et al.*).¹⁷

In practice, the EFES implementation of a geographical indexing and visualisation features will need to be relatively light-weight and highly generic and customisable, while working out-of-the-box for the core constituency of classical epigraphy. The backbone of the spatial information will be an index of the Solr component that drives EFES search functions. This index will contain label, type, and

¹² Pleiades: pleiades.stoa.org.

¹³ Antiquity à la carte application, awmc.unc.edu/wordpress/alacarte/.

¹⁴ Johan Åhlfeldt's DARE: imperium.ahlfeldt.se.

¹⁵ Klokantech's Roman Empire Map Tiles: github.com/klokantech/roman-empire.

¹⁶ World Historical Gazetteer: whgazetteer.org/about/.

¹⁷ Pelagios Network: pelagios.org.

coordinate information for each geographical identifier, along with the individual texts from which the place is referenced. It is expected that as a rule the EpiDoc corpus will store only the references and identifiers, and geographical and spatial information will be retrieved from the record in the gazetteer.¹⁸



Figure 4: Periplo place distribution visualisation, showing results for “defixio” (© Pelagios Project)

The default implementation of this index will be based on a current dump of the Pleiades information, retrieved by the user and copied into the EFES file structure. Instructions will be made available for editors from other epigraphic traditions, to transform the gazetteers of relevance for their corpus into the form expected by the EFES indexer, following the practice of the Linked Places Format developed by the Pelagios Project and the World Historical Gazetteer (Grossner 2018).

The resulting index will then be displayed in an embedded map interface such as the Leaflet Javascript library for interactive maps,¹⁹ using the historical map-tiles developed for DARE. This will enable various views of the data, such as:

- all places (mentioned, findspots, etc.) in the whole corpus;
- all places in the current search or filter results from the search page;
- all places mentioned in or associated with a single inscription.

¹⁸ The process will need to be aware of place records in the gazetteers that have no associated coordinates, and that can therefore probably not easily be mapped onto a traditional visualisation.

¹⁹ Leaflet: leafletjs.com.

These places may be represented as pins or icons on the map, which may be selected to pop-up an “infobox” containing information, description or images relating to the text or texts that refer to this place, and a link back to the page for the digital edition of each text. The function of the map for users of the corpus will therefore be exactly comparable to an index of place information, but in a more useful and interactive interface.

This mapping functionality will still be generic and limited to features that the designers have considered and that the software available provides. We will not, for example, attempt to implement full faceted search options within the geographic visualisation interface, enabling users to adjust date or other facets in the search page and see the map re-generated or re-centered on the screen in response; or moving or zooming the map interface and having facets updated based on which texts are represented by the area of the world on the current map. One might in the long term envisage incorporating the entire Peripleo tool as a plugin to EFES to achieve this, but it is more likely that an editor wanting this functionality would be advised to use Peripleo or some other GIS library to deliver their corpus, alongside (or perhaps even instead of) EFES.

As noted, we do not intend to reproduce most of the functions and potential of GIS databases and visualisation tools, but all geographic data (including the indexes containing spatial data imported from gazetteers) we be made available for export in common formats such as comma-separated values (CSV), GeoJSON or RDF for use in third-party tools for visualisation or analysis.

2.3 Open data export and connectivity

Ultimately all of the features described here—indexes, date selection tool, geographic display and other visualisations—can only be implemented in a generic (albeit extensible) way, in a platform designed for many users and corpora with diverse needs. For advanced digital and quantitative analysis and visualisation of data, more technical users and editors alike will need access to the underlying data, in a portable format, using open standards, and licensed for unrestricted reuse and republication, to process and study using their own or published tools and algorithms. In addition to the underlying EpiDoc XML and other formats that drive the online publication, EFES will make available for download the processed indexes, vocabularies and other structured data that feed the internal visualisations.

From any given index, search or view of the data, we plan to enable the download, at a single click or via command to an API, the tabular data behind the index in a common format such as CSV, JSON or RDF, with a clear indication of license. Users may then visualise this data, display it on a digital map, process it using social network analysis, or remix it in other ways in combination with related or enhanced data, and redistribute or publish the results. Such re-use is both essential for reviewers wanting to assess and comment on the published work, and the next generation of editors building on and citing the work of earlier authors, as is and has always been scholarly practice.

Open data is a fundamental component of electronic publishing, and especially when fully findable, accessible, interoperable and reusable (FAIR), it enables new forms of collaboration (Wilkinson *et al.*). This interchange and sharing of digital information comes into its own when relationships within and between datasets, vocabularies, entities and classes of data are expressed in semantic and web-accessible formats. In addition to simply being available online, using open licenses and formats, good Linked Open Data should be expressed in semantic web formats such as RDF, and should be linked to other datasets and collections that also use dereferenceable identifiers—*i.e.* HTTP URIs that point to online information about those specific entities or properties (Kim-Hausenblas). In this model, the entities under discussion (inscriptions, places, people, objects, publications, events or scholarly claims), the predicates or relationships that define or comment on them, and the properties or classes used to categorise and describe them, are all represented by HTTP URIs that reference concepts and entities defined in ontologies, taxonomies, vocabularies and related knowledge systems.

The ontologies and LOD communities of particular interest for exposing and sharing data from an epigraphic or other text corpus include:

1. The Linked Places Format, discussed above, developed by the Pelagios and World Historical Gazetteer projects to share and align gazetteers, relevant for exposing geographical data.
2. Guidelines and ontology made available by the Standards for Networking Ancient Prosopographies (SNAP), for the sharing and alignment of person-databases and exposing references to persons and names (Bodard *et al.*; Bodard).
3. Identifiers and online tools relating to ancient dates and calendars that occur in primary texts, made available through the Graph of Dated Objects and Texts (GODOT) project (Grieshaber “Godot”).
4. The ChronOntology Project’s ontology of dating and types of dates (Schmidle *et al.*), and PeriodO’s taxonomy of named historical periods used in scholarship (Rabinowitz), for exposing and sharing various kinds of dating information, as discussed above.
5. A growing community of epigraphic bodies and projects proposing ontologies and vocabularies for sharing data about inscriptions in particular. Early work exemplifying the needs and potential of epigraphic LOD accompanied the release of Epigraphische Datenbank Heidelberg (EDH) content as open data, and has since been taken forward by the Epigraphy.info community, and in particular the Epigraphic Ontology working group; among the specialised vocabularies available, the EAGLE Network published canonical lists of identifiers under headings including

Inscription Type, Object Type and Material (Grieshaber “Epigraphic Database”; Granados García, §19–21; Tupman).²⁰

In order for any of these types of information to be exported as LOD, the underlying EpiDoc files would need to contain references to the relevant identifiers in standard elements, of course. A generic publication tool such as EFES should enable export in a few of the most useful ontologies, but an editor will have to decide which information to encode and therefore make available in this form, and further customisation may be required to add connection to new ontologies and data ecosystems to the platform.

3. Future research

In a very real sense, the services and functionalities offered by a visualisation and publication tool like EFES also serve a prescriptive role in the practice of EpiDoc encoding (itself a set of guidelines for narrowing down the application of TEI XML to ancient texts). The implementation of search, index, visualisation and export features that rely on internal authority lists and consistent encoding in the EpiDoc files, demands a more constrained encoding scheme than the flexible options in the EpiDoc Guidelines (Elliott *et al.*, *Guidelines*) themselves. While a certain amount of flexibility is desirable, given the breadth of disciplinary practices and epigraphic cultures EpiDoc and EFES need to support, there is a demonstrable need—and frequent requests—for prescriptive guidance on encoding features (Bodard-Stoyanova).²¹ Just as the TEI Guidelines create a narrow practice for existing academic practice, EpiDoc limits TEI to even narrower practices, the expectations of a tool like EFES further delimit EpiDoc, and a single project will no doubt define their own, extremely consistent internal encoding scheme.

It is a truism of the Digital Humanities that open standards are essential for responsible and sustainable online publication of digital editions and other research data. It is equally true however that the tools used to create, prepare, analyse, communicate and disseminate digital data and editions are part of the research process, and so ought to be available, well documented, and reusable by readers and reviewers (Turska; Liuzzo 48-9). Since electronic publication is in effect a form of data visualisation, the open distribution of “source code” should include not only EpiDoc XML files, but also XSLT stylesheets and the customised version of the platform such as EFES that was used to generate indices, search and visualisations. It is recommended that editors who create their publication using EFES do so by “cloning” the Git repository into a new fork of the platform, and make all customisations and

²⁰ See also Epigraphy.info: A Collaborative Environment for Digital Epigraphy: epigraphy.info; Epigraphic Ontology Group: groups.google.com/forum/#!forum/epont; EAGLE Vocabularies: eagle-network.eu/resources/vocabularies/.

²¹ See instances of such requests on the Markup list (lsv.uky.edu/archives/markup.html) or the EpiDoc Feature Requests tracker (sourceforge.net/p/epidoc/feature-requests/) at any given time.

further enhancements available to readers; some of these improvements may be offered back to the core EFES code-base by means of “pull requests.”²²

The description of the EFES platform offered in this paper, and the proposed enhancements to its functionality, are by no means intended to suggest a complete, ideal, stable and perfectly sustainable publication tool for the future. No software package, least of all free and open source software, can make that claim. There remain several issues that EFES needs to address in the longer term, in particular complex dependencies on several other open source tools, some of which are no longer regularly updated and supported. There are occasional and unpredictable issues with installation on certain versions of certain operating systems (which of course are not unique to this tool), and a current dependency on a legacy version of the Java Virtual Machine, that will need to be resolved sooner or later. Although EFES is in principle deployment-ready, with a built-in Jetty server instance allowing all the other components to be displayed in a web browser, further support and guidance is currently needed for a user intending to deploy it independently, for instance on a commercial web host without institutional and technical support.²³

It is our hope and intention that the community of EpiDoc editors and developers will take ownership of EFES as part of the EpiDoc tool set (alongside the Guidelines, Schema and Reference Stylesheets), and that future improvements to the platform will come about as a result of funded project-specific requirements fed back into the code base. Several funding applications in the pipeline have already included EFES development in their budgets, and new scholars request advice or training in the use of the platform on an almost weekly basis. Collaboration, both with individual scholars and with related large publication and infrastructure projects, is essential for the continued development and sustainability of any open source platform such as EFES, and the engagement and generosity of interested scholars and developers in contributing to our collective work has so far been exemplary.

Corpora and Web Resources

- CII. *Cretan Institutional Inscriptions*. Irene Vagionakis, Venice Centre for Digital and Public Humanities, 2020. cretaninscriptions.vedph.it.
- ECG. *Epigraphic Corpus of Georgia*. Demo version. Nino Doborjginidze *et al.*, Ilia State University, 2019. epigraphy.iliauni.edu.ge.
- EDV. *Epigraphic Database Vernacular*. Nadia Cannata, Luna Cacchioli, Alessandra Tiburzi. Sapienza University of Rome, 2020. edv.uniroma1.it.
- I Aph. *Inscriptions of Aphrodisias*. Joyce Reynolds, Charlotte Roueché, Gabriel Bodard. King's College London, 2007. insaph.kcl.ac.uk/iaph2007/.
- IOSPE. *Ancient Inscriptions of the Northern Black Sea*. Irene Polinskaya, Askold Ivantchik, *et al.*, King's College London, 2015–2017. iospe.kcl.ac.uk.

²² An example of a pull request that integrated dozens of improvements by Irene Vagionakis into the EFES code-base can be found at github.com/EpiDoc/EFES/pull/54.

²³ Current web hosting documentation to be found in Vagionakis, “Host server setup” (in Yordanova): github.com/EpiDoc/EFES/wiki/Host-server-setup.

- IRT. *Inscriptions of Roman Tripolitania*. Joyce Reynolds and John Bryan Ward-Perkins, enhanced electronic reissue by Gabriel Bodard and Charlotte Roueché. King's College London, 2009. inslib.kcl.ac.uk/irt2009/.
- RIB. *Roman Inscriptions of Britain*. Robin G. Collingwood and R. P. Wright, enhanced electronic reissue by Scott Vanderbilt. 2014–2020. romaninscriptionsofbritain.org.

Bibliography

- Bodard, Gabriel. "Linked Open Data for Ancient Names and People." *Linked Open Data for the Ancient World: A Practical Introduction*, edited by Paul Dilley, Ryan Horne and Sarah Bond. *ISAW Papers* 20, 2020 (forthcoming). dlib.nyu.edu/awdl/isaw/isaw-papers/20/.
- Bodard, Gabriel and Simona Stoyanova. "Epigraphers and Encoders: Strategies for Teaching and Learning Digital Epigraphy." *Digital Classics Outside the Echo-Chamber: Teaching, Knowledge Exchange and Public Engagement*, edited by Gabriel Bodard and Matteo Romanello. London: Ubiquity Press, 2016, 51–68. dx.doi.org/10.5334/bat.d.
- Bodard, Gabriel, Hugh Cayless, Mark Depauw, Leif Isaksen, K. Faith Lawrence, Sebastian Rahtz. *SNAP Cookbook*. King's College London, 2014. snapdrgn.net/cookbook.
- Burghart, Marjorie. "The TEI Critical Apparatus Toolbox: Empowering Textual Scholars through Display, Control, and Comparison Features." *Journal of the TEI* 10, 2016. doi.org/10.4000/jtei.1520.
- Elliott, Tom and Sean Gillies. "Digital Geography and Classics." *Digital Humanities Quarterly* 3.1, 2009. digitalhumanities.org/dhq/vol/3/1/000031.html.
- Elliott, Tom, Gabriel Bodard, et al. *EpiDoc Guidelines*. Version 9. Stoa Consortium, New York University, 2008–20. epidoc.stoa.org/gl/latest/.
- Elliott, Tom, Zenata Au, Gabriel Bodard, Hugh Cayless, Carmen Lanz, Faith Lawrence, Scott Vandebilt, Raffaele Vigiante, et al. *EpiDoc Reference Stylesheets*. 2008–20. github.com/EpiDoc/Stylesheets.
- Granados García, Paula Loreto. "Hesperia" (review). *RIDE* 7, 2017. ride.i-d-e.de/issues/issue-7/hesperia/.
- Grieshaber, Frank. "GODOT. Graph of dated objects and texts, building a chronological gazetteer for antiquity." *Epigraphy Edit-a-thon. Editing chronological and geographic data in ancient inscriptions, April 20–22, 2016*, edited by Monica Berti, University of Leipzig, 2017. nbn-resolving.de/urn:nbn:de:bsz:15-qucosa-221532.
- Grieshaber, Frank. "Epigraphic Database Heidelberg – Data Reuse Options." University of Heidelberg, 2019. doi.org/10.11588/heidok.00026599.
- Grossner, Karl. "Contributing in Linked Places format." *World-Historical Gazetteer*, 2018. whgazetteer.org/2018/09/11/lp-format/.
- Kim, James G. and Michael Hausenblas. "5 ★ OPEN DATA." 2015. 5stardata.info.
- Liuzzo, Pietro Maria. *Digital Approaches to Ethiopian and Eritrean Studies*. Aethiopica Supplement 8, Harrassowitz Verlag, 2019. doi.org/10.2307/j.ctvrnfr3q.

- Rabinowitz, Adam *et al.* "Making Sense of the Ways We Make Sense of the Past: The PeriodO Project." *Bulletin of the Institute of Classical Studies* 59.2, 2016, 42–55. onlinelibrary.wiley.com/doi/full/10.1111/j.2041-5370.2016.12037.x.
- Rosselli Del Turco, Roberto. "The Battle We Forgot to Fight: Should We Make a Case for Digital Editions?" *Digital Scholarly Editing: Theories and Practices*, edited by Matthew Driscoll and Elena Pierazzo, Open Book Publishers, 2016, 219–238. doi.org/10.11647/OBP.0095.
- Sahle, Patrick. "What is a Scholarly Digital Edition?" *Digital Scholarly Editing: Theories and Practices*, edited by Matthew Driscoll and Elena Pierazzo, Open Book Publishers, 2016, 19–40. doi.org/10.11647/OBP.0095.
- Schmidle, Wolfgang, Nathalie Kallas, Sebastian Cuy and Florian Thiery. "Linking Periods: Modeling and Utilizing Spatio-temporal Concepts in the ChronOntology Project". *Computer Applications and Quantitative Methods in Archaeology (CAA)*, 2016. Oslo.
- Simon, Rainer, Leif Isaksen, Elton Barker and Pau de Soto Cañamares. "Peripleo: a Tool for Exploring Heterogeneous Data through the Dimensions of Space and Time." *Code4Lib* 31, 2016. journal.code4lib.org/articles/11144.
- Sopracasa, Alessio and Martina Filosa. "Encoding Byzantine Seals: SigiDoc." *Atti del IX Convegno Annuale AIUCD. La svolta inevitabile: sfide e prospettive per l'Informatica Umanistica*. Università Cattolica del Sacro Cuore, 2020. aiucd2020.unicatt.it/aiucd-Sopracasa_Filosa.pdf.
- Tupman, Charlotte. "Where can our inscriptions take us? Harnessing the potential of Linked Open Data for epigraphy." *Epigraphy in the Digital Age: Opportunities and Challenges in the Recording, Analysis and Dissemination of Epigraphic Texts*, edited by Isabel Velázquez Soriano and David Espinosa Espinosa. Archaeopress, forthcoming 2021.
- Turska, Magdalena, James Cummings and Sebastian Rahtz. "Challenging the Myth of Presentation in Digital Editions." *Journal of the TEI* 9, 2016. doi.org/10.4000/jtei.1453.
- Wilkinson, Mark D., Michel Dumontier, Barend Mons, *et al.* "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific Data* 3, 2016, 160018. doi.org/10.1038/sdata.2016.18.
- Yordanova, Polina *et al.* *EFES User Guide*. University of London via Github, 2017–18. github.com/EpiDoc/EFES/wiki/User-Guide.

Preliminary Research on Computer-Assisted Transcription of Medieval Scripts in the Latin Alphabet using AI Computer Vision techniques and Machine Learning.

A Romanian Exploratory Initiative

Adinel C. Dincă and Emil Şteţco

Babeş-Bolyai University

Zetta Cloud, Cluj-Napoca

Abstract: The objective of the present paper is to introduce to a wider audience, at a very early stage of development, the initial results of a Romanian joint initiative of AI software engineers and palaeographers in an experimental project aiming to assist and improve the transcription effort of medieval texts with AI software solutions, uniquely designed and trained for the task. Our description will start by summarizing the previous attempts and the mixed-results achieved in *e-palaeography* so far, a continuously growing field of combined scholarship at an international level. The second part of the study describes the specific project, developed by Zetta Cloud, with the aim of demonstrating that, by applying state of the art AI Computer Vision algorithms, it is possible to automatically binarize and segment text images with the final scope of intelligently extracting the content from a sample set of medieval handwritten text pages.

Keywords: Middle Ages, Latin writing, palaeography, Artificial Intelligence, Computer Vision, automatic transcription.

A. Introduction

In 1971 György Granasztói (1938-2016), a Hungarian historian and demographer, teamed up with a Russian technician, Valentin A. Ustinov (1938-2015), and together used the computer – inputting data via punched cards – to analyse from a quantitative perspective (Granasztói 1971) a tax register compiled in 1475 in Braşov (an account book regarding the tax revenue of *Portica* neighbourhood in Braşov,

preserved at Serviciul Județean Brașov al Arhivelor Naționale [Brașov County Service of the National Archives], „Primăria orașului Brașov”, Oficiul impozitelor oraș Brașov, seria III Da, no. 1/1, 1475). The authors were following a fashionable ideological trend of that moment, which pointed out that, by using mathematical and computer-oriented tools, historians could move toward a fundamental scientific objective: the development of a common language for interdisciplinary work that could solve complex social problems (Putnam 1971, 24). Beyond the political principles that motivated this endeavour, some elements still valid 50 years later need to be highlighted: first, the research was based on an **experiment**, which was conducted with the use of an **innovative** machine at that time. Secondly, it involved a Hungarian historian, a Russian computer analyst and a Romanian archival item which reflected information concerning a medieval Saxon town in Transylvania – thus, the operation did not only have an **international** character, but was also validated by a **cross-field** level investigation. These key-words: “Experimenting” – “Innovating” – “International” – “Cross-field level” still define today the interaction of humanities and computer sciences, even though the results have moved greatly from a simple quantitative report produced half a century ago towards the next phase, simply put as the transcription of medieval writing with the help of an intelligent machine.

The history of handwriting, **palaeography**, is a specific field within medieval studies. According to the definition:

“this discipline studies the appearance and development of different types of script and their uses by diverse social groups across time and in diverse documents (books, records, charters, etc.). Further, it analyses the transmission and staging of a written message, attending not only to the text but also to its form, script, layout and support.” (Stutzmann 2016)

In these circumstances, the transcription of medieval manuscript texts sets itself apart as a complex, highly specialized, and time-consuming activity. Experts in palaeography need to be trained in the language of the text, the historical usages of various styles of handwriting, common writing customs, and scribal/notarial abbreviations. Thus, one may spend many hours transcribing rather short medieval original documents to offer researchers ways of indexing, finding, and consulting such evidence. Complete editions of extended works usually take years to complete, a reason why modern scholars, confronted with the issue of “short-termism”, avoid engaging (too often) in such endeavours or prefer to do partial editions (Bertrand 2005, Słóń 2015). Some religious texts such as sermons or saints’ lives, for example, were completely neglected by editorial projects, due to their large number of individual copies transmitted in variants copied over centuries in a plethora of writing solutions and contexts.

More than any other disciplines, palaeography is most dependent on visual evidence. Thus, the invention of photography in the 19th century compelled Ludwig Traube (1861-1907), the palaeographer who held the first chair of Medieval Latin at the “Ludwig Maximilian University” of Munich, to address the 1900s as ‘the age of photography’ in the history of this particular field of study (Lehmann 1909, 57). Visual contexts provided palaeographers with ways to broadcast and compare manuscripts

while slowly linking researchers into interconnected networks around the digitized images (Widner 2017) that became valuable tools for consulting important manuscripts to which access could be hard to obtain, for reasons of conservation or value (Wakelin). The conversion of texts and pictures into a digital form that can be processed by a computer produced a *digital representation* or, more specifically, a *digital image* in the form of binary numbers, which facilitated computer processing and other operations. This process of **digitalization** has set a milestone in the study of automated transcription, yet, in order to be “read” by either the human eye or the artificial intelligence, the image had to be further “prepared” for browsing.

The classical problem in computer vision, image processing and machine vision has been that of **recognition**, from the first **Computer Vision** projects that were meant to mimic the human visual system. Experimenting began in the late 1960s, at universities in the West, as a task connected to artificial intelligence, and the pioneering studies in the next decades formed the foundations for many of the computer vision algorithms that exist today. The next phase, bringing further life to the field of computer vision, was the so-called deep-representation learning based on neural networks. **Deep Learning techniques** have currently the accuracy and performance close to that of humans, relying on **convolutional neural networks (CNNs)**. This evolution, not only in terms of sophistication of the learning process, has been paralleled with the progression of the performed tasks: while

“the first wave of digital humanities’ work, from the late 1990s to early 2000s, was described as quantitative, mobilizing the search and retrieval powers of the database, the second wave emphasizes the keywords: qualitative, interpretive, experiential, emotive, generative. [...] It harnesses the digital toolkits in the service of the Humanities’ core methodological strengths: attention to complexity, medium specificity, historical context, analytical depth, critique and interpretation.” (*Digital Humanities Manifesto 2.0*)

Experts in palaeography or the “auxiliary sciences of history” – such as epigraphy, codicology, numismatics, sphragistics to name just a few – were always a scarce resource, and currently the changes in the educational system make them even more difficult to find. As it has been noted, “palaeographic skills can be acquired only through intensive training and prolonged exposure to rare artifacts that can be difficult to access” (Kestemont et al. 2017, S87). Recent opinions even suggested that such experts are obsolete due to their subjective, dogmatic and authoritarian perspective (Stokes 2009, 313-317), and plead for objective criteria in palaeography, by generating a set of measurements which can ultimately be statistically analysed and compared by computers. The result of these divergent perspectives may lead to a paradoxical situation: humanity will soon have an exhaustive digital memory of its medieval evolution, but it will not have access to it, due to the missing corpus of trained specialists able to decode old forms of writing.

The variability of the handwriting, the complexity of the vocabulary and styles, the multi-linguistical and multi-graphical aspects, the difficulty of automatically isolating the characters and segmenting the text lines make the currently systems based on recognition, the Optical Character Recognition (**OCR**) and the Handwritten Text

Recognition (**HTR**), unable to automatically recognize and transcribe medieval texts. While OCR is considered a closed problem in computer vision, handwritten text recognition still presents an open challenge. Yet, the need to provide efficient solutions to time-consuming and laborious palaeographic tasks pushes all academic fields towards multi-disciplinary collaboration in the search for new options. The creation of the Text Encoding Initiative (**TEI**) (<https://tei-c.org/>), developed in the 1990s, has offered humanities scholars common standards for encoding electronic texts and representing texts in digital form, TEI **guidelines** being used around the world by libraries, museums, publishers, and individual scholars for online research, teaching, and preservation.

B. Computer-assisted palaeography. An overview of European projects

Medieval scripts developed in the Latin cultural area are characterized by different handwriting styles from diverse places and periods of time over one thousand years, from the 6th to the 15th century. Moreover, the typological diversity of such graphical solutions for human communication (Derolez 2003) is increased by the intention laying behind a certain text. For example, a notarial document will never be similar to a liturgical text, as much as a university handwritten textbook could never be mistaken for a formal papal bull or a royal charter. Different sets of letterforms were specifically designed for individual contexts of use and destinations, each context being closely connected to the issuer and the beneficiary, respectively with their intention regarding the text. Therefore, notions like **hierarchical structure of information**, **formality**, **standardization** or **level of execution** play an important role in classifying and describing a certain sample of medieval handwriting. All these factors create the uniqueness of every handwritten record produced during the Middle Ages.

The current scholarly field of “computer-assisted” or “artificial palaeography” represents the small tip of an iceberg: its massive base assembles the digitized heritage of “European” handwriting, that is, based on the Latin alphabet, with initiatives that also take into consideration Greek and Coptic (<https://d-scribes.philhist.unibas.ch/en/home/>), or Hebrew (<http://www.erabbinica.org/>), with all their quirks and features.

A large part of the medieval texts preserved in archives or libraries all over the world has been intensively digitized during the last two decades. Considerable institutional or private funding was invested in preserving the textual memory of the medieval past, aiming at the same time an improved accessibility to the source material for international scholarship, as usually repositories with medieval books or documents are in various countries across the continent(s). Today, virtually every single larger library and archival collection from Austria, France, Germany, Italy (with the Vatican state), Switzerland, to name only a few, has its own digitization project, on various levels of development but comprising tens of thousands of scanned units with uncountable number of pages. However, even though such digital libraries – BVMM (<https://bvmm.irht.cnrs.fr/>), Gallica (<https://gallica.bnf.fr/>), e-Codices (<https://www.e-codices.unifr.ch/en>), Manuscripta Mediaevalia (<http://www.manuscripta-mediaevalia.de/#1>), DigiVatLib (<https://digi.vatlib.it/>) – and archives – Monasterium (<http://monasterium.net:8181/mom/home>) – are amassing reproductions of medieval manuscripts and archives, they offer scarce metadata. The

construction of such impressive digital libraries required so far technical solutions and skills and usually neglected or undervalued the traditional scholarship based on palaeography and textual criticism.

As briefly stated, digitization projects represent at this point an accomplished task for most European archives, Romania included (see, for instance, the digitization project of the Romanian National Archives, <http://arhivamedievala.ro/>). Database creation and management is also a fruitful ongoing activity; however, this approach does not cover the specific needs of digital palaeography. Beyond the first stage of acquiring images, the interaction of palaeography with computerized tools seeks to provide efficient solutions to the consuming palaeographic tasks. Various attempts have produced mixed-results in **e-palaeography** over the last two decades, to cite just a few: the **System for Paleographic Inspections (SPI)** was the first software dedicated to digital palaeography developed in 1999 by a group of researchers in the History Department, University of Pisa, for the inspection of ancient Roman manuscripts – from this project derived in 2004 the first attempts at automatically clustering scripts, which led to Arianna Ciula's coining of the term "digital palaeography" (Aiolfi et al. 1999; Ciula 2005; Aiolfi & Ciula 2009; Aussems & Brink 2009).

The **Monk** system (<https://www.ai.rug.nl/~mrolarik/MLS/>; <http://monkweb.nl/>) is a continuous project, developed at the University of Groningen in 2005 by a research group at the Institution of Artificial Intelligence and Cognitive Engineering (ALICE), under the supervision of Lambert Schomaker. The **GRAPHEM** (*Grapheme based retrieval and analysis for palaeographic expertise of medieval manuscripts*) research project was funded from 2007 to 2011 by the French National Agency for Research, under the supervision of Dominique Stutzmann, on the basis of automated image analysis without the need to select individual letters (Cloppet et al. 2007; Muzerelle & Gurado 2011).

The research programme **ORIFLAMMS** (*Ontology research, image feature, letterform analysis on multilingual medieval scripts* – <http://oriflamms.teklia.com/>), 2013-2016, tackled the same issues with new methods that combined letterform analysis with Computer Vision for script classification. ORIFLAMMS, coordinated by Dominique Stutzmann, aimed at analysing the evolution of writing systems and graphical forms during the Middle Ages and according to their production contexts (informal, documentary, book scripts) and languages (Latin or vernacular) (Stutzmann 2016; Oriflamms 2017).

DigiPal (*Digital Resource for Palaeography* – <http://www.digipal.eu/>) developed between 2010-2014 under the supervision of Peter A. Stokes, a scholar at King's College, London (Stokes 2009; Stokes 2011). The project's aim was to catalogue, describe and, where possible, source digital images of about 1200 scribal hands. Data was organized with the help of **Archetype** (<http://archetype.ink/>), an integrated suite of web-based tools for the study of medieval handwriting, art and iconography. Peter A. Stokes also proposed a different approach (Stokes 2007), consisting of two stages, the first known as "feature extraction", which involves generating the numerical measurements, and the second, "data mining", finding similarities and classifying handwriting based on these measurements.

tranScriptorium project (<http://transcriptorium.eu/>) was active between 2013-2015; its aim was to develop innovative, efficient and cost-effective solutions for the indexing, search and full transcription of historical handwritten document images, using an enhanced version of Handwritten Text Recognition (HTR) technology and Layout Analysis. The result, the **Transkribus** software (<https://transkribus.eu/Transkribus/>), is currently hosted by the “Digitisation and Digital Preservation group” (DEA) at the University of Innsbruck. It is also used by the EU-funded e-infrastructure project **READ** (Recognition and Enrichment of Archival Documents), coordinated by Günter Mühlberger of the University of Innsbruck (<https://read.transkribus.eu/>), an undertaking that concluded just some months ago.

Another recent project was **HIMANIS** (*H*istorical *MAN*uscript *I*ndexing for *user-controlled Search* – <https://www.himanis.org/>), again under the coordination of Dominique Stutzmann, aimed at developing cost-effective solutions for querying large sets of handwritten document images, more precisely the textual heritage of the French Royal Chancery compiled in the 14th and 15th centuries: charters, registries and formularies. To this end, innovative keyword spotting, indexing and search methods have been developed, tested, adapted and/or scaled up to meet the challenges of more than 83.000 digitized pages (Stutzmann 2018).

The **Scripta-PSL: The History and Practices of Writing** programme at Université Paris Sciences et Lettres (<https://escripta.hypotheses.org/>) aims at integrating the fundamental sciences that deal with written artefacts (palaeography, codicology, epigraphy, history of the book, etc.), with other disciplines in the humanities and social sciences (linguistics, history, anthropology, etc.), together with digital and computational humanities, around the study of writing. The digital component, **eScripta**, intends to mash software programmes (such as **Archetype**, developed by Peter Stokes and the team at King’s College London, which allows deep annotation and extensive palaeographic study of writing, an open-source OCR software called **kraken**, developed by Benjamin Kiessling (PSL research engineer) and **Pyrrha**, a tool for post-correction of POS tagging and lemmatization with the help of CNNs to isolate objects from their background and distinguish between main writing, decoration (illuminations, drop capitals, etc.), and interlineal or marginal annotations (Stokes et al. 2019).

The **Digital forensics for historical documents** project, funded by the Royal Netherlands Academy of Arts and Sciences (https://en.huygens.knaw.nl/projecten/digital-forensics-for-historical-documents/?noredirect=en_GB), is currently ongoing (2018-2021) under the supervision of Mariken Teeuwen. The research is developing the results of previous projects that have explored automatic methods for handwritten text analysis: **Monk**, **Transkribus**, **DigiPal** and **Artificial Palaeography** (developed recently by Dominique Stutzmann e.a., see Kestemont et al. 2017). The project Digital Forensics aims at creating a bridge in between two different ways of handwriting analyses: the forensic method (graphanalysis) and the palaeographical method (the study of the development of scripts through space and time), combining the two methods in a single ‘deep learning system’.

Before concluding this brief overview of the blooming **Digital Palaeography** arena, it is our intention to share some thoughts regarding how Artificial Intelligence, Computer Vision techniques and Machine Learning processes could contribute to a more efficient and accurate transliteration of historical texts issued within the cultural area of the Latin Middle Ages. First, the technology is not yet mature, and it poses a series of obvious issues. Difficulties in fluid communication between palaeographers and computer scientists is a prevailing problem, another is restriction of property: licensed vs. open source software (Hassner et al. 2012, 15, 24). Reiterating the opening paragraph of this presentation, digital humanities work best when a fruitful collaboration is established between scholars, computer scientists and, last but not least, cultural heritage institutions – in other words, when the roles and competence of each party have been identified and acknowledged. Integration of user feedback (participatory engagement) is another aspect that needs to be taken into consideration when discussing the software development strategies. Our view on the matter is therefore not to replace the expert's eye and hard acquired training (as it is sometimes casually suggested and expected), but to equip palaeographers with new tools and solutions that would help them to face with improved efficiency the overwhelming mass of medieval manuscripts available now in the digital environment (see a parallel experiment in Fischer et al. 2009). Interpretation of the results would have to derive from an interdisciplinary discussion across fields of expertise, elaborated by scholars of both Humanities and Computer Science disciplines. After all, the final goal of computer scientists is to develop an expert system that will emulate human expertise, in this particular case that of palaeographers' competences, methodologies, taxonomies and "ground-truth" (Stutzmann 2015).

C. A computer-assisted transcription project for medieval text images

In our specific project, developed by Zetta Cloud (a SME entity), the goal was to demonstrate that, by applying state of the art AI Computer Vision algorithms, we are able to automatically binarize and segment text images with the final scope of automatically extracting text from a sample set of medieval handwritten text pages. For applying AI algorithms used in Computer Vision to the handwriting transcription task, we trained, in an experimental manner, a Recurrent Neural Network (RNN) to recognize text by feeding it with enough previously humanly transcribed text image segments. For this purpose, a sample was selected from the 12th century handwritten Latin manuscript from Engelberg, Stiftsbibliothek: *Vitae Sanctorum et passiones Martyrum. Pars aestivalis* (<https://www.e-codices.unifr.ch/de/list/one/bke/0002/>). The selected manuscript was copied in the second half of the 12th century within the premises of Engelberg Abbey, Switzerland, a Benedictine foundation of 1120. The script engaged for the transmission of the widely spread text (Lives of the Saints, organized according to the liturgical calendar, the summer part) is a Caroline Minuscule in its late development before the spreading of Gothic script (<http://carolinenetwork.weebly.com/>). This particular form of Caroline Minuscule reflects accurately the calligraphic mastery of the Swiss Benedictine Convent

(Bruckner 1950), the script achieving at that time an almost typographic regularity. There is virtually no variation of the individual letter forms and the use of elements of abbreviation, as well as the diacritical signs, are remarkably constant. All these features (further details in Dincă 2011), combined with the exceptionally good digital copy at hand and an almost optimal state of preservation, advocate the above-mentioned manuscript as a suitable sample for our research. It must be added that the book comprises hundreds of folios copied by the same hand, thus providing enough reliable material to be worked with. Minor, subjective variations of forms (dimensions, rightward slanting, marginalia etc.), if needed, can be inserted into recognizing and identification patterns used in Computer Vision.

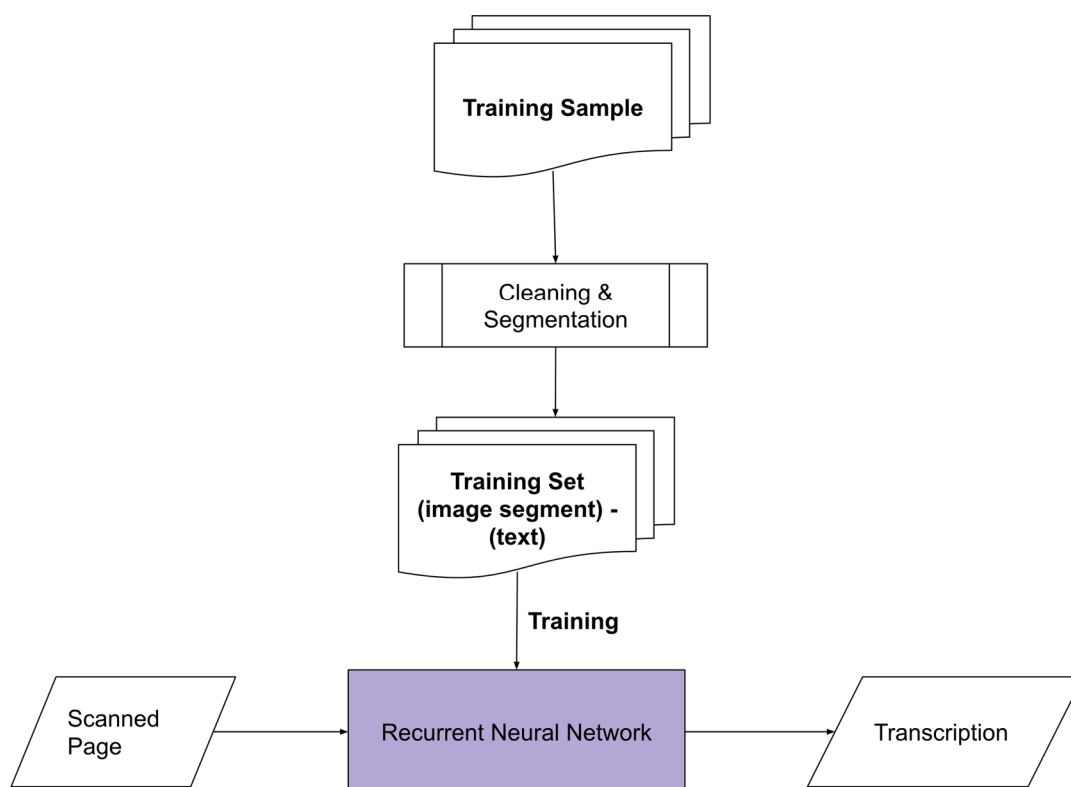
We have done extensive research to find the best implementation of an RNN for our purposes. Also, previous research carried out by well-known universities were analysed in order to gain as much knowledge as possible. In search for the best RNN implementation, our goal was to find an Open Source library that allows the specification and use of different Recurrent Networks architectures with ease and, if possible, with a no-code approach. Kraken implements a dialect of the Variable-size Graph Specification Language (VGSL), enabling the specification of different network architectures for image processing purposes using a short definition string. For our project purposes, the need arose for a library that could be used to train lightweight AI Computer Vision models where the recognition is done not on the character level but rather on image pattern level, this to allow expansions of abbreviations used by the medieval writer. On the other hand, Kraken implements the full Handwritten Recognition Pipeline one would expect, meaning: Page Segmentation followed by Line Segmentation and Handwriting recognition. At the end of our research, we have decided to use the Kraken Open Source Library as the foundation for our implementation.

We applied three training phases, each followed by testing phases carried out using previously humanly transcribed text image segments that were not used during the training phase to assess the achieved model accuracy.

Once the image acquisition issue was solved, the first training stage, consisting of a set of examples or training samples, could be put together. In order to obtain this training set, a pre-process of preparation was required: images were cleaned from noises but still preserved the text and its features, while an automatic binarization process was applied on all training sample pages. The set was then split into (1). a training set and (2). a testing set (typically, the ratio was 80% of the set for training and 20% of the set for testing), followed by training sessions of the Machine Learning Algorithm with this set; accuracy was automatically determined using the testing set. Essentially, the run experiments used intuition and statistics to make the machine able to understand what needs to be done to solve the specific handwriting transcription problem.

In addition to the automatic validation at the end of a training phase, human expert evaluation was absolutely needed during both training and experiment phases. The training corpus consisted of pairs of images and their “caption” transcriptions, and

the initial aim was to prepare a training data set comprising at least 50 text images with their corresponding transcriptions. Computer scientists have tailored and then used algorithms for automatic binarization of original text images. The same procedure was followed with implementing and using AI algorithms for automatic segmentation of the original text images. The result of this phase was used by the expert in palaeography during the transcription phase. Segmented versions of the original documents were manually transcribed for the purpose of creating the training and testing sets used at training the AI computer vision algorithm. The manual transcription has been undertaken by a qualified expert in Latin palaeography, to ensure the accuracy of the data set.



A typical Machine Learning Flow

For the first training phase, 2972 segments were used, extracted from the first 38 pages ([1r](#) to [19v](#)) of the sample manuscript. Training the RNN AI algorithm with these 2972 segments (keeping just 80% of the set for training and 20% for testing and control purposes) generated an automatically computed accuracy of 0.9461962161730662 (94.61%). This is an excellent result for a first phase of training. We have also analyzed automatic transcriptions of pages not seen by the AI algorithm by now (any pages starting from [20r](#) outwards) and the results were extremely positive.

cessit ab ea. Cumq̃ p̃uenisset addomū illius uri nobilissimi bonifacius nomine q̃ fuit mutus. uidens eū iacentē & mutū ait. Dñe ih̃u aperi os eus et tuum nom̃ quod ē be- nedictū inuocet & credat. quia tu es d̃s ui- uens in secl̃a s̃cl̃oꝝ. Cumq̃ dictū fuisset xp̃i anis amen. eadem hora soluta ē lingua eius & laudabat dñm dicens. n̄ est alius dominus nisi quem hic beatissimus predicat apollinaris. In eodem loco amplius quā quingenti homines cre- diderunt in ih̃m. agentes gr̃as dō. Non p̃t multos dies inflati paganorū quidē asp̃u immundo tenuerunt eū. & celum fustib⁹ prohibebant eū ne loqueretur in nomine ih̃u. Qui iacens in terra testificando fortiter clama- bat de nomine ih̃u. Non ferentes paga- ni hoc testimoniu. nudis pedib⁹ sup̃ pru- nas stare eū fecerunt. Et ille in cessante de nomine ih̃u p̃dicabat. Crescebat popl⁹ xp̃ianorū maxime nobiliū. Erat autē quidā rufus patricius. cui filia infirmabatur. & uis- sit ep̃m ad domū suā uenire. Et cū p̃ue- nisset defuncta ē. Qui dixit patri puellē si ducialiter age. & iura in quod p̃mittas pu- ellam sequi saluatore suū. & modo cogno- scis uirtutē dñi. Rufus dixit. Scio qđ mor- tua ē puella. tamen si uidero eā stantē & loquentē. Laudabo uirtutē dñi tui. & di- mittā eā sequi saluatore suū. Apollina- ris accessit & tetigit puellā dicens. dñe ih̃u	aliū xp̃ianū credider. xp̃o. Et p̃ h̃c accusatus apaganis apud c̃sarem. ducebat ad tormenta. & multa pati ens c̃fitebat dño. Quidā uero xp̃paga- nis qui senior erat infamulo di. arrip- tus ademonio subito exsp̃rauit. Viden- tes autē xp̃iani tantā inuiriā famuli di cōmoti sunt. & irruentes sup̃ paganorū ducentos homines occider. Et iudeus uis- sit eum in carcere claudi cū grauissimo pondere ferri. & in ligno extendi. & nichil illi ministrare ut deficeret. Angelus dñi nocte uenens ad eū. ui- denib⁹ custodib⁹ pauit eū. & c̃fortans eū abiit. Et post temp⁹ cuiusdā primi & magni uiri fr̃. lepius effectus ē. Cūq̃ uidisset cū apollinaris dixit ei. Ihs̃ fieri sanus. Qui respondens ait. Volo. Apol- linaris dixit. Crede in dñm ih̃m xp̃m. Respondit. Sime sanū fecerit. ip̃e erat d̃s n̄s. Et inuocans apollinaris nom̃ ih̃u xp̃i. tetigit eū. & statim sanus fac- tus ē. Qui renuntiatis simulachris ere- didit in ih̃m. & baptizatus ē in subur- bano p̃ncipis senatoris. Subito orta ē sedicio in ciuitate de paganis de no- mine apollinaris. Et irruentes ppli sup̃ eum. ligatum p̃duxerunt cedentē res ei & uulnerantes. Quē uidentes pontifices capitolii. in dignati s̃t di-	cessit ab ea. Cumque peruenisset ad domum illius uri nobilissimi bonifacius nomine qui fuit mutus. uidens eum iacentem et mutum ait. Domine ihesu aperi os eius et tuum nomen quod est be- nedictum inuocet et credat. quia tu es dominus ui- uens in secula seculorum Cumque dictum fuisset a christi anis amen. eadem hora soluta est lingua eius et laudabat dominum dicens. non est alius dominus nisi quem hic beatissimus predicat apollinaris. In eodem loco amplius quam quingenti homines cre- diderunt in ihesum. agentes gracias domino. Non post multos dies inflati paganorum quidem a spiritu immundo tenuerunt eum et cesum fustibus prohibebant eum ne loqueretur in nomine ihesu. Qui iacens in terra testificando fortiter clama- bat de nomine ihesu. Non ferentes paga- ni hoc testimonium. nudis pedibus super pru- nas stare eum fecerunt. Et ille incessanter de nomine ihesu predicabat. Crescebat populus christianorum maxime nobilium. Erat autem quidam rufus patricius, cuius filia infirmabatur. et ius- sit episcopum ad domum suam uenire. Et cum perue- nisset defuncta est. Qui dixit patri puellam si ducialiter age, et iura nihilo quod permittas pu- ellam sequi saluatorem suum. et modo cogno- scis uirtutem domini. Rufus dixit. Scio quod mor- tua est puella tamen si uidero eam stantem et loquentem laudabo uirtutem domini tui et di-
--	--	--

As planned, the goal was to understand how would the accuracy of the algorithm evolve when the training set was increased to 5427 segments, extracted from approximately 68 pages (1r to 34v). At this point, a second RNN Training phase was conducted, starting with an increased effort to transcribe further pages in a traditional manner, up to page 34v. By segmenting all these pages, a total of 5427 segments were generated for this second training phase. The computed achieved accuracy was of 0.9509186465708205 (95.09%), consequently the accuracy gain was of only 0.48%.

Therefore, by almost doubling the manual transcription effort a gain of only 0.48% of computed accuracy was achieved. Nonetheless, by conducting expert analyses of the automatic transcriptions generated by this second trained AI model, we have concluded that the 0.48% accuracy gain was worth the effort since the AI algorithm seemed to improve the recognition of abbreviations and spaces between words. This second experimental phase allowed for a consolidation of information both from the perspective of palaeography and computer science: that is to better understand how the AI algorithm managed to differentiate and recognize on an improved level various details of palaeographical and linguistic nature: abbreviating solutions, words as linguistic units, double letter forms ("s"-long, "s"-round, varieties of "a"), or letter similarities ("c" vs. "t" confusions). Such details are not always easy to determine even for the human reader at a medium level of expertise in this field. It can be concluded that an increased effort of traditional transcription may not bring a significant advance in accuracy of automatic reading but could lead to a sensible improvement of the transliteration quality. More experiments should be conducted to verify such hypotheses.

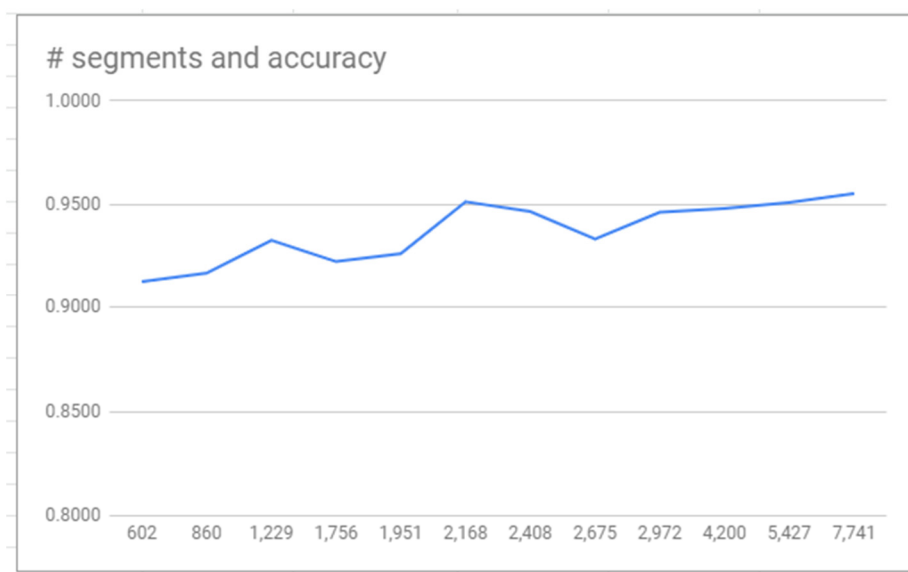
Even if it seemed at this point that the results were indicative of the Pareto principle (Newman 2005) applied to machine learning, the fact that the microscopic

accuracy gain had an abundance of meaning for the preciseness of the transcription, made the Zetta Cloud team decide to continue increasing the set of segments for the last RNN training phase.

# pages	# segments	accuracy	
8	602	0.9128	Experiment 8
11	860	0.9168	Experiment 7
16	1,229	0.9327	Experiment 6
22	1,756	0.9224	Experiment 5
25	1,951	0.9263	Experiment 4
28	2,168	0.9512	Experiment 3
31	2,408	0.9466	Experiment 2
34	2,675	0.9334	Experiment 1
38	2,972	0.9462	Phase I
53	4,200	0.9481	Experiment 9
68	5,427	0.9509	Phase II
100	7,741	0.9553	Phase III

To find the **effort inflection point**, the minimum number of manually transcribed pages (number of segments) needed to achieve sufficient transcription accuracy, we have operated several training and testing phases using sets of reduced number of segments, starting from the number of segments used by the Phase I - 2972 segments. Experiments were conducted with 4200 segments, as well. This is to plot the accuracy obtained by training the RNN AI algorithm with a number of segments between the number used for Phase I training and the Phase II training.

Here are all observations plotted on a simple graph:



Effort Inflection Point is at 2168 segments (28 pages)

Experiment 8, conducted by using just 602 segments (approximately 8 pages, manually transcribed) shows that 91.28% accuracy could have been achieved with a minimum effort. To conclude, the **effort inflection point** is at **2168 segments** (approximately **28 pages**), meaning that the algorithm could have been trained with **just 28 pages** of manual transcription to gain an accuracy value that is very close to what was achieved by manually transcribing **100 pages** and training the algorithm with them.

For the purpose of testing the transcription AI algorithm that was trained during the project (the result of Phase III – 100 manually transcribed pages for training), we've developed and deployed a very simple UI interface that one could use to upload one or more page images to obtain an automatically generated transcription. The interface is available here: <https://catmeti.intellidockers.com/>.

We have also put together a quick video tutorial that shows how to use the interface to obtain automatically generated transcription for your page images. The tutorial is available <http://bit.ly/catmetipoc>. (NOTE: You should use pages starting from 51r of the manuscript. These pages were never seen by the algorithm and they will show you how the AI computer vision algorithm is doing when put in front with completely new pages to be transcribed).

D. Conclusion

An innovative collaboration between two apparently opposing fields of study, palaeography and computer science, has led to a pioneering Romanian experiment aiming to demonstrate that by applying state of the art AI Computer Vision algorithms, text images can be automatically binarized and segmented for the systematic extraction of content from a sample set of medieval handwritten text pages. Beyond the theoretical challenges pertaining to the computational aspects of the endeavour, the Zetta Cloud computer scientists auspiciously tackled the larger scope of such an enterprise: to respond to the complex needs of the end-users, from experienced palaeographers interested in maximising the effort of transcribing a larger volume of handwritten text to the young enthusiasts observing the contents of a historical source that had been, so far, mediated by printed editions. However, making palaeography accessible to non-experts does not directly imply that the assessment of experienced scholars – perceived as academics existing in an ivory tower – is overrated and inconsequential, but rather that a highly restricted field of study can open up the scientific enquiry to young researchers, by facilitating access to the raw historical source and speeding up the formative stage of a specialised work force and academic career.

A handful of factors have contributed to the favourable outcome of this experiment: by carefully selecting the writing samples to be transcribed (i.e. Carolingian minuscule for continental manuscripts from the 9th/10th to the 12th/early 13th, as well as for the Humanistic calligraphic iterations from the 15th to 16th century) an immense chronological and geographical area of employment for this highly advanced palaeographical tool can be achieved. In this context, the next logical step

would be to apply the same workflow for obtaining AI computer vision models to other handwriting styles and complexities.

Previous and ongoing developments in the field of e-palaeography make the present project even more viable: potential collaborations and exchange of experience are currently heading towards an international community of practice, suitable for open dialogue and scientific progress. Leveraging crossover skills in today's academic environment means, more than any other time, pushing one's versatility and adaptability in the direction of new interdisciplinary projects and, last but not least, of prospective funding opportunities from both the public and the private sectors. Romanian memory institutions and their holdings may be the first to benefit by such computational advancement, while researchers could explore the historical material in new ways, from putting together editions of previously unpublished documents / account books / letter collections – to name just some of the potential archival records that would make suitable candidates for such an approach – to solving larger “puzzles” arising from the emerging field of “fragmentology” (as applied to the situation in Romania, where the study of medieval manuscript fragments has an explored potential, see Dincă 2011 and Dincă 2017). Thus, the long-term goal of this collaborative team is to develop a platform aiming to assist the experts in the transcription process, automating the tasks, speeding up the rendition of documents and improving its capabilities along the way. The automatic, Computer Vision based, handwriting recognition system should cooperate with the human expert to generate the final, highly qualitative text transcription.

This platform, with the implementation of a full Machine Learning Loop, would become smarter with every human expert intervention. Machine Learning Loop or Human-in-the-loop (HITL) is the process where the machine is unable to solve a problem with great accuracy initially starting with a small amount of annotated samples and it needs human intervention for creating a continuous feedback loop allowing the algorithm to give every time better results. Mainly HITL approach is used, when there is not much data available initially so that human experts can produce just enough annotated samples to obtain the best accuracy possible. The AI Computer Vision models produced using this approach can then be applied to the high volume of manuscripts to obtain the best accurate transcription. In this respect, our experiment shows that by initially annotating just 602 segments (approximately 8 pages, manually transcribed) we could obtain a model with 91.28% accuracy that is very near to the 95.53% accuracy that we obtained with training the same RNN neural network with 100 annotated paged. Our envisioned system would allow human experts to start with a very few annotated samples and reach a just enough accuracy level in a quick as guided manner through the HITL process.

Together with further AI technology like **Named Entities Extraction**, **Automatic Summarization** and **Forensic Author Identification**, the collaborative team will be able to do content extraction and various summaries which will allow any researcher to focus on the most relevant parts of the verified texts. Eventually, any sort of text in any documentary form will be available for further research purposes with a moderate transcription effort.

Acknowledgment

This work was supported by a grant from the Romanian National Authority for Scientific Research, CNDI–UEFISCDI, project PN-III-P4-ID-PCCF-2016-0064: “The Rise of an Intellectual Elite in Central Europe: Making Professors at the University of Vienna, 1389-1450” (<https://rise-ubb.com/>). We thank our two anonymous reviewers for their comments.

Works cited

- Aiolfi, Fabio & Ciula, Arianna. “A case study on the System for Paleographic Inspections (SPI): challenges and new developments”. *Proceedings of the 2009 Conference on Computational Intelligence and Bioengineering: Essays in Memory of Antonina Starita*, IOS Press, 2009, pp. 53-66.
- Aiolfi, Fabio & Simi, Maria & Sona, Diego & Sperduti, Alessandro & Starita, Antonina & Zaccagnini, Gabriele. “SPI: A System for Paleographic Inspections”. *AIIA Notizie*, vol. 4, 1999, pp. 34-48.
- Aussems, Mark & Brink, Axel. “Digital palaeography”. *Codicology and palaeography in the digital age 2*. Edited by Rehbein, Malte & Sahle, Patrick & Schassan, Torsten, BoD, 2009, pp. 293-308.
- Bertrand, Paul. “La numérisation des actes: evolutions, révolutions. Vers une nouvelle forme d’édition de textes diplomatiques?”. *Vom Nutzen des Edierens. Akten des internationalen Kongresses zum 150-jährigen Bestehen des Instituts für Österreichische Geschichtsforschung, Wien, 3.-5. Juni 2004*. Merta, Brigitte & Sommerlechner, Andrea & Weigl, Herwig Böhlau, 2005, pp. 171-176.
- Bruckner, Albert. *Scriptoria medii aevi Helvetica: Denkmäler schweizerischer Schreibkunst des Mittelalters*, vol. VIII. *Schreibschulen der Diözese Konstanz, Stift Engelberg*, Genf, 1950.
- Ciula, Arianna. “Digital palaeography: Using the digital representation of medieval script to support palaeographic analysis”. *Digital Medievalist*, 1, 2005. www.digitalmedievalist.org/journal/1.1/ciula/. [20.06.2020].
- Cloppet, Florence & Daher, Hani & Églin, Véronique & Emptoz, Hubert & Exbrayat, Matthieu & Joutel, Guillaume & Lebourgeois, Frank & Martin, Lionel & Moalla, Ikram & Siddiqi, Imran & Vincent, Nicole. “New Tools for Exploring, Analysing and Categorising Medieval Scripts”. *Digital Medievalist*, vol. 7, 2011. journal.digitalmedievalist.org/articles/10.16995/dm.44/. [20.06.2020].
- Derolez, Albert. *The palaeography of Gothic manuscript books from the twelfth to the early sixteenth century*. Cambridge University Press, 2003.
- Dincă, Adinel C. “Datarea manuscriselor medievale latinești. Evaluari metodologice”, *Anuarul Institutului de Istorie «George Barițiu» din Cluj-Napoca, Series Historica*, tome L, 2011, pp. 295-306.
- Dincă, Adinel C. “The Medieval Book in Early Modern Transylvania. Preliminary Assessments”, *Studia UBB, Historia*, vol. 62, Issue 1, 2017, pp. 23-34.

- Fischer, Andreas & Wüthrich, Markus & Liwicki, Marcus & Frinken, Volkmar & Bunke, Horst & Viehhauser, Gabriel & Stolz, Michael. *Automatic Transcription of Handwritten Medieval Documents*. Conference paper at Proc. 15th Int. Conf. on Virtual Systems and Multimedia (VSMM'09), 2009. DOI: 10.1109/VSMM.2009.26.
- Granasztói, György. "Computerized Analysis of a Medieval Tax Roll. *Acta Historica Academiae Scientiarum Hungaricae*, vol. 17, no. 1/2, 1971, pp. 13-25.
- Hassner, Tal & Rehbein, Malte & Stokes, Peter A. & Wolf, Lior. "Manifesto from Dagstuhl Perspectives Workshop 12382. Computation and Palaeography: Potentials and Limits". *Dagstuhl Manifestos*, vol. 2, issue 1, 2012, pp. 14-35, drops.dagstuhl.de/opus/volltexte/2013/4167/pdf/dagman-v002-i001-p014-12382.pdf. [20.06.2020].
- Hassner, Tal & Sablatnig, Robert & Stutzmann, Dominique & Tarte, Ségolène. "Report from Dagstuhl Seminar 14302. Digital Palaeography: New Machines and Old Texts". *Dagstuhl Reports*, vol. 4, issue 7, 2014, pp. 112-134. www.researchgate.net/publication/269168418_Digital_Palaeography_New_Machines_and_Old_Texts_Dagstuhl_Seminar_14302 [20.06.2020].
- Kestemont, Mike & Christlein, Vincent & Stutzmann, Dominique. "Artificial Paleography: Computational Approaches to Identifying Script Types in Medieval Manuscripts". *Speculum*, vol. 92 (S1), 2017, pp. S86-S109. DOI: 10.1086/694112.hal-01854939. [20.06.2020]
- Kiessling, Benjamin. *Kraken – a Universal Text Recognizer for the Humanities*. Paper presented at Digital Humanities Conference 2019 (DH2019), Utrecht, the Netherlands. doi.org/10.34894/Z9G2EX, dev.clariah.nl/files/dh2019/boa/0673.html [20.06.2020]
- Lehmann, Paul (ed.). *Ludwig Traube, Zur Paläographie und Handschriftenkunde*, Beck, 1909.
- Muzerelle, Denis & Gurrado, Maria, (eds.), *Analyse d'image et paléographie systématique : travaux du programme "Graphem": communications présentées au colloque international "Paléographie fondamentale, paléographie expérimentale: l'écriture entre histoire et science" (Institut de recherche et d'histoire des textes (CNRS), Paris, 14-15 avril 2011)*. Association Gazette du livre médiéval, 2011.
- Newman, M. E. J. "Power laws, Pareto distributions and Zipf's law". *Contemporary Physics*, vol. 46, no. 5, 2005. DOI: 10.1080/00107510500052444. arxiv.org/PS_cache/cond-mat/pdf/0412/0412004v3.pdf [20.06.2020]
- Oriflamms. *Compte-rendu final du projet ORIFLAMMS / ORIFLAMMS Final report*. 2017. oriflamms.hypotheses.org/files/2017/04/Oriflamms-Compte-rendu-final.pdf. [20.06.2020].
- Putnam, George F. "Soviet historians, quantitative methods, and digital computers", *Computers and the Humanities*, vol. 6, Issue 1, September 1971, pp. 23-29.
- Schnapp, Jeffrey & Presner, Todd & Lunenfeld, Peter & Drucker, Johanna. *Digital Humanities Manifesto 2.0*, jeffreyschnapp.com/wp-content/uploads/2011/10/Manifesto_V2.pdf. [20.06.2020]

- Słoń, Marek. "Pryncypia edytorstwa źródeł historycznych w dobie rewolucji cyfrowej [Principles of Editing Historical Sources at the Time of the Digital Revolution]", *Studia Źródłoznawcze/Studies in Historical Sources*, vol. LIII, 2015, pp. 155-161.
- Stokes, Peter A. & Kiessling, Benjamin & Tissot, Robin & Stökl Ben Ezra, Daniel. *EScripta: A New Digital Platform for the Study of Historical Texts and Writing*, paper presented at Digital Humanities Conference 2019 (DH2019), Utrecht, the Netherlands. hal-02310781. dev.clariah.nl/files/dh2019/boa/0322.html [20.06.2020].
- Stokes, Peter A. "Computer-Aided Palaeography, Present and Future". *Kodikologie und Paläographie im digitalen Zeitalter / Codicology and Palaeography in the Digital Age*, BoD, 2009, pp. 309-338, kups.ub.uni-koeln.de/2978/. [20.06.2020].
- Stokes, Peter A. "Computer-Aided Palaeography, Present and Future". *Kodikologie und Paläographie im digitalen Zeitalter / Codicology and Palaeography in the Digital Age*. Edited by Rehbein, Malte & Sahle, Patrick & Schaßan, Torsten, BoD, 2009, pp. 309-38.
- Stokes, Peter A. "Digital Resource and Database for Palaeography, Manuscripts and Diplomatic". *Gazette du livre médiéval*, vol. 56-57, 2011, pp. 141-142; www.persee.fr/doc/galim_0753-5015_2011_num_56_1_1991. [20.06.2020].
- Stokes, Peter A., "Palaeography and Image-Processing: Some Solutions and Problems". *Digital Medievalist*, vol. 3, 2007. <http://doi.org/10.16995/dm.15>. [20.06.2020].
- Stutzmann, Dominique & Kermorvant, Christopher & Vidal, Enrique & Chanda, Sukalpa & Hamel, Sébastien & Puigcerver Pérez, Joan & Schomaker, Lambert & Toselli, Alejandro H. "Handwritten Text Recognition, Keyword Indexing, And Plain Text Search In Medieval Manuscripts". Conference paper at *Digital Humanities 2018 Puentes-Bridges. Book of Abstracts*, p. 298-302. dh2018.adho.org/wp-content/uploads/2018/06/dh2018_abstracts.pdf. [20.06.2020].
- Stutzmann, Dominique. "Clustering of medieval scripts through computer image analysis: Towards an evaluation protocol". *Digital Medievalist*, vol. 10, 2016. DOI: <http://doi.org/10.16995/dm.61>. [20.06.2020]
- Wakelin, Daniel. "«An anthology of images»: DIY digital photography in manuscript studies", *DIY Digitization*. diydigitization.org/contributed-papers/wakelin/ [20.06.2020]
- Widner, Michael. "Toward Text-Mining the Middle Ages: Digital Scriptoria and Networks of Labor". *The Routledge research companion to digital medieval literature*. Edited by Boyle, Jennifer & Burgess, Helen J., Routledge, 2017, pp. 131-144.

DaT18 Database: A Prosopographical Approach to the Study of the Social Structures of Religious Dissent in Mid-Eighteenth-Century Transylvania

Radu Nedici

University of Bucharest

Abstract: Drawing on the many records created by the Habsburg state during the confessional troubles in Transylvania from the 1740s to the 1760s, the DaT18 project merges digital instruments and prosopography to arrive at sketching the social pattern of the Orthodox leadership. This article briefly discusses the technical choices involved in building the relational database that my approach centres on, before talking in more detail about the challenges faced when transposing the information in the primary sources into digital format. First, the question of making use of structured vs. unstructured data, as most of the documents I work with already present some form of tabular layout, while the more narrative ones require different strategies to mitigate losses when converting them. Secondly, the difficult process of record linkage, with many of the persons only mentioned by their first name and no surname to help label each individual entered in more than one source. Lastly, the daunting task of estimating economic resources, since there was no reliable standard in an age that saw four different fiscal systems in use and many regional flavours within the same scheme.

Keywords: prosopography, relational database, clerical careers, data structuring, Greek Orthodox Church.

Early modern Transylvania was a land full of contrasts, from social to political and religious, that fuelled one another, leading in turn to the establishment of stricter normative identities. Generally, divergence was kept in check by a system of mutual toleration and an equal distribution of power between the political nations and confessions. Its roots were medieval, but the major elements were set in place during the sixteenth century and they

were further reinforced once the Habsburgs took over the province at the end of the seventeenth century (Keul; Roth). Conflicts were solved through negotiation and, before the rise of modern nationalism, they rarely broke out into open confrontation. The clashes that involved the Orthodox and Greek Catholic Romanians between the 1740s and the 1760s thereby offer a fertile research ground to expand existing knowledge on the strategies of strife and cohabitation in this part of Europe.

In spite of their number, Transylvanian Romanians were at the margins of the above framework, as they were allowed to have their own religious institution, headed by a recognized bishop of Byzantine rite, but were denied participation in the political life as a distinct estate (Hitchins 14–17). Efforts to increase the standing of the entire community received a boost with the Habsburg rule, since the emperor in Vienna needed to secure collaborators that would help legitimize the Catholic party's claim to power. In exchange for acknowledging religious union with Rome, the Romanians were promised access to the same privileges already enjoyed by the members of the political estates (Bernath 73–82). Both sides pinned high hopes on the resulting Greek Catholic Church, but not much changed after the act of union in 1701. Initially rejected by only a minority, the compromise with Rome came under attack more frequently as decades went by. With the involvement of the Orthodox archbishop of Sremski Karlovci, the conflict entered a new radical stage after 1744. Over the following 17 years internecine violence was rampant, affecting mostly southern Transylvania, where villagers ousted their Greek Catholic priests and replaced them with Orthodox clerics, while petitioning the authorities for the free exercise of religion. The Habsburgs' response, which centred on persecuting the leaders of dissent, failed to temper the spirits and induced the Orthodox Romanians to look for support to their cause outside the borders of the Monarchy, notably to Russia. By the late 1750s, at the height of the Seven Years War, Vienna finally conceded that it lacked the required resources to impose a return to earlier conformity. Tolerance for the Orthodox in the principality was officially proclaimed in 1759, although they remained the object of discriminatory regulations for at least another century. A separate bishop – Dionisiје Novaković – was appointed in 1761 to head the newly created diocese in Transylvania, which by that time totalled almost four fifths of all the Byzantine-rite Christians in the province, signalling a glaring defeat for the competing Greek Catholic Church (Nedici, "Religious"; Nedici, "Rethinking"; Nedici, "Cum să pornești").

These two decades of confessional unrest are at the core of my ongoing research project, *Dissent and toleration in Habsburg Transylvania: A socio-political history of the Orthodox protests (1740s–1760s)*, which has received national funding from the Executive Unit for Financing Higher Education, Research, Development and Innovation and is hosted at the University of Bucharest (<https://www.dat18.ro>, hereafter DaT18). It aims to study the religious troubles in mid-eighteenth-century Transylvania by turning away from the confessional consequences of the split within the Romanian community and on to the social underpinnings and political meanings of the opposition movement itself. Part of the results are now made available in an online database on the project's website, which for now consists of a catalogue of all

the Orthodox priests active in Transylvania from 1761 to 1767, with plans to further develop the application to include data concerning the career of leading lay persons. This article will briefly discuss the technical choices involved in building the database, before talking in more detail about the challenges of transposing the information in the primary sources into digital format and readying it for analysis.

Database overview

Because it is implicit, the behind-the-scenes construction of the new Orthodox eparchy in the 1740s–1760s goes largely unacknowledged in most history texts. To this day, little is known about the men and women behind the spread of dissent and those that emerged as leaders in each village during the troubles and once toleration was announced. Since those at the helm of the protests were mostly common people, able, at best, to just write their names and doing everything they could to avoid suspicion from the Habsburg authorities, their existence is documented in only minimal forms. While there is a tone of information still to be found in the archives, it is scattered and fragmentary, and it usually consists of nothing more than mere names.

The whole rationale behind the project was to somehow arrive at piecing together the many instances when the hundreds and even thousands of anonymous characters were mentioned in order to gain better insights into the social structures of religious dissent. A relational database was best suited for the purpose, as the only instrument capable of exploiting to the fullest the existing sources by linking complementary facts and emphasizing overlapping information, where available. Also, by breaking the data into smaller segments, it becomes possible to selectively and creatively rearrange it an infinite number of times to answer new sets of questions one might ask and to reveal details that the original documents alone did not comprise. Prosopography would in turn allow to move between individual careers and the collective biography of the group to grasp connections and interactions (Keats-Rohan), while at the same time bridging quantitative and qualitative approaches (Cohen et al.).

From a technical standpoint, the database was designed and information was entered in Microsoft Access. This dataset was later imported into Zoho Creator, where the application for querying it was created. I will not go in more detail over the reasons why I opted for such a workflow since I have already addressed them in a conference paper that is forthcoming (Nedici, “Spre o istorie”). The structure of the database was modelled on the equation that each person is mentioned as participating to a historical event only once by the same source, except in the case of obvious errors. This does not impede accounts where the same people were present in more than one circumstance or those where for one separate incident there could be numerous attendants, nor indeed does it deny that different sources may provide contrasting reports on identical facts. However, contextualizing the information passed to the three tables at the core of the database – Person, Event, & Source – provides equal weight to all the statements and ensures that the accuracy of conflicting details is dealt with at a later stage, instead of being inferred during data entry itself. This is further enhanced by the rule that a record is created only when the person and event are mentioned by a contemporary source, meaning that only first-hand information is

converted into entities in most cases, while indirect accounts will, at best, provide attributes that are ascribed to something explicitly stated. An appointment for a parish priest, for instance, which also reveals when and where he had been ordained, does not lead to a distinct entry on ordainment alone, but rather this data is placed among the details of his nomination. Additional tables hold values on the standard place names for the individuals' birth and residence and for mapping the facts, as well as on the economic status of those for which the fiscal censuses provide that sort of input. Three further tables are used to link the various entities together, first by creating unique identifiers for the persons and the events, which are then correlated with one another according to existing references (Fig. 1).

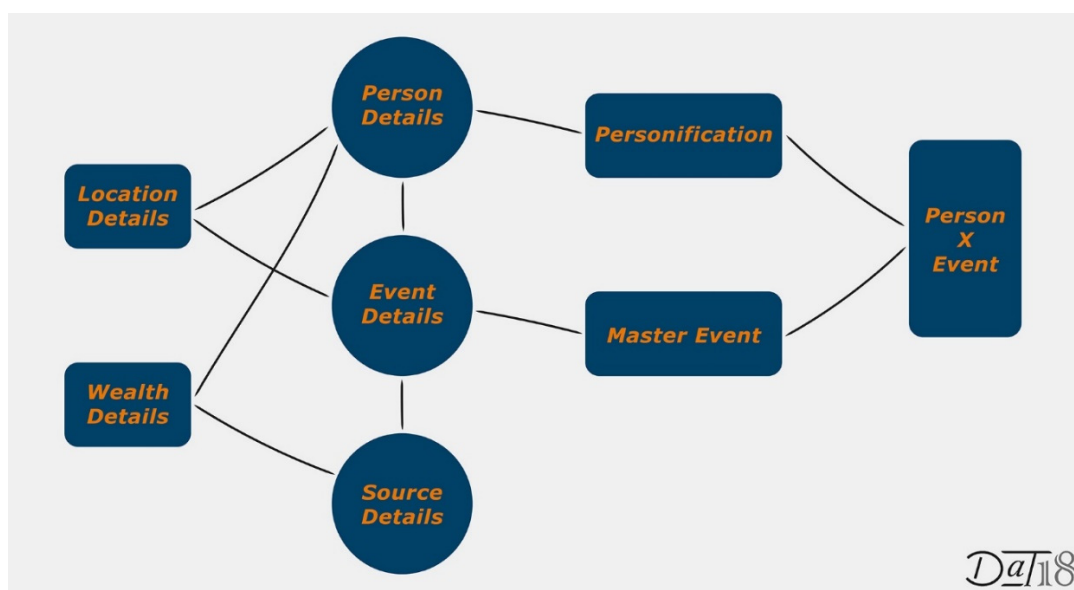


Fig. 1. DaT18 Database logical structure

At the time of writing, the database is at version 3.1, the most recent update being published online in April 2020 ("DaT18 Database"). In line with the objectives set for this stage, it now consists of a complete catalogue of all the parish priests that served in the Orthodox eparchy in Transylvania according to the successive statistics compiled by Bishop Dionisije Novaković from 1761 to 1767. Access to the database is free for every researcher and does not require any registration before using it. The search interface brings forth two possibilities for querying the metadata, either by using a keyword or by scrolling through the tables that make up the database. The quick search option enables the user to interrogate certain fields in the tables, such as ID, name, event, and location, retrieving all instances that contain the specific term (Fig. 2). Scrolling through each of the main tables allows for a more advanced search, including the use of Boolean operators to narrow the range of results, with only add, edit, and delete functions omitted for obvious reasons (Fig. 3). From the default list view it is possible to navigate to underlying data in other tables by clicking on the highlighted links or to open a detailed view of the object, which already includes relevant information from the related tables (Fig. 4).

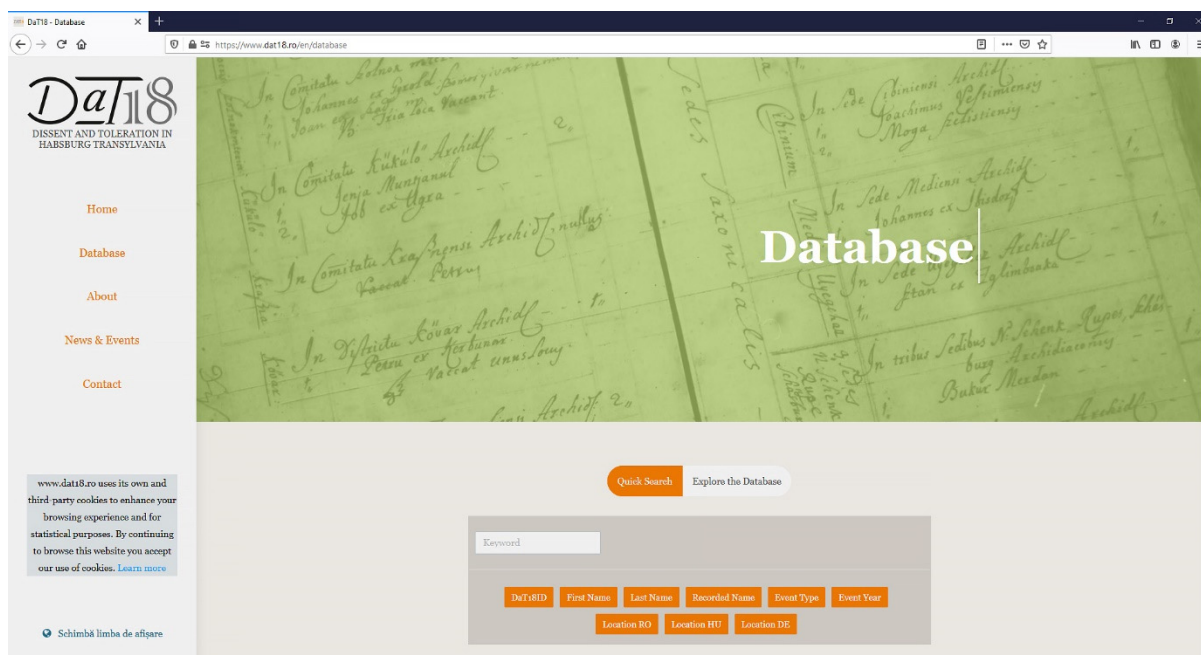


Fig. 2. DaT18 Database quick search interface

tb11 Person Details Report

Person ID	DaT18 ID	Event ID	Recorded Name	Birth Place Location	Recorded Birth Place Name	Residence Location	Recorded Residence Name	Age	Gender	Denomination	Marital Status
2601	2601	2601	Mihail Samoilescu	33.6.2	Balotir	33.6.2	Balotir		M	ortodox	
2600	2600	2600	Vasile Fetul	41.2.1	Făgăraș	32.2.6	Turdaș		M	ortodox	
2599	522	2599	Iosif Vlad	33.2.6	Turdaș	32.2.6	Turdaș		M	ortodox	
2598	479	2598	Ioan Tărtărie	33.3.9	Tărtărie	13.1.0	Gelnar		M	ortodox	
2597	696	2597	Ioan Mereșescu	13.1.3	Bîrlint	13.1.3	Bîrlint		M	ortodox	
2596	1254	2596	Avram Pescar	33.6.1	Șibot	33.6.1	Șibot		M	ortodox	
2595	495	1377	George	33.6.2	Balotir	33.6.2	Balotir		M	ortodox	
2594	465	2594	George Putrașcu	41.2.1	Făgăraș	32.3.1	Cioara		M	ortodox	
2593	1260	2593	Avram Uret	33.6.3	Vinerea	33.6.3	Vinerea		M	ortodox	
2592	1261	2592	Petru Nemeș	99.0.1	Sîrmeș	32.6.3	Vinerea		M	ortodox	
2591	2591	2591	Ioan Popescu	33.6.7	Vaidet	33.6.7	Vaidet		M	ortodox	
2590	1269	2590	Petru Ciuci	33.6.7	Vaidet	33.6.7	Vaidet		M	ortodox	
2589	210	2589	Constantin Drăgoescu	33.6.6	Romoș	33.6.6	Romoș		M	ortodox	
2588	1266	2588	Ioan Popescu	33.6.5	Romoșel	33.6.5	Romoșel		M	ortodox	
2587	433	2587	Nicolae Popescu	32.2.4	Sibișel	11.8.15	Tămășasa		M	ortodox	
2586	1265	2586	Ionascu Adamescu	33.2.4	Sibișel	33.2.4	Sibișel		M	ortodox	
2585	313	2585	Daniel Moștescu	33.3.1	Cioara	32.2.2	Castău		M	ortodox	
2584	363	2584	Ilie Dragomir	33.2.1	Beriu	33.2.1	Beriu		M	ortodox	
2583	1263	2583	Ioan Popescu	33.2.2	Castău	32.2.5	Sereca		M	ortodox	
2582	206	2582	Silvestru Roman	11.8.3	Orăștie	11.8.3	Orăștie		M	ortodox	
2581	1264	2581	Radu Petrovici	99.1.41	(Valea)	32.2.3	Prizac		M	ortodox	
2580	2580	2580	Ioan Popovici	99.1.51	(Moldova)	32.8.1	Orăștie		M	ortodox	
2579	314	2579	Grigore Calugăr	33.3.1	Cioara	33.8.1	Orăștie		M	ortodox	
2578	573	2578	Petru Dirzu	33.4.2	Căpâlnă	33.8.1	Orăștie		M	ortodox	
2577	2577	2577	Ioan Marcu	43.7.6	Rucăr	43.7.6	Rucăr		M	ortodox	
2576	1188	2576	Bucur Matei	43.7.6	Rucăr	43.7.6	Rucăr		M	ortodox	
2575	1185	2575	Ioan Anghel	43.7.1	Feldioara	43.7.1	Feldioara		M	ortodox	

Showing 1000 of ###

Search

Person ID

DaT18ID

Event ID

Recorded Name

Contains

Is

Is Not

Is Empty

Is Not Empty

Starts With

Ends With

Like

Not Contains

Age

Gender

Denomination

Marital Status

Kinship

Social Status

Wealth Estimate

Education

Occupation

Office

Search

Fig. 3. DaT18 Database advanced search options

tbl4 Location Details Report

Location ID	Location Name RO	Location Name HU
8115	Dealul Mare	Dădăfalva
8114	Bărașu Mare	Nagyborsz
8113	Băba	Băba
81124	Vina Mare	Tordavina
81123	Sălișca	Szelecske
81122	Glod	Szamosdomező
81121	Polana Blenchi	Blenkemező
81120	Ileanda	Nagyfonda
8112	Căseiu	Alsókossály
81119	Muncel	Kishavas
81118	Magura	Kishegy
81117	Coroian	Karúfalva
81116	Căpâlna	Căicăkpölina
81115	Căplean	Kappon
81114	Călcău	Kackó
81113	Guga	Guga
81112	Gostila	Csicsgombás
81111	Gălgău	Galgó
81110	Frânceni de Piatră	Kőérőfalva
8111	Utrior	Alór
7379	Ocnita	Mezőfalna
7378	Pericu	Százszételek
7377	Posmus	Paszmos
7376	Stupini	Mezősolymos
7375	Vila Tecl	Kolozsnagyida
7374	Comlod	Komlód
7373	Șeuș	Kissajó

Showing 1000 of 222

Overview

Location ID	7376
Location Name RO	Stupini
Location Name HU	Mezősolymos
Location Name DE	
Part Of	Cluj
Location Type	sat

tbl2 Event Details

Event ID	Master Event ID	Location ID	Recorded Location Name	Event Year	Event Month	Event Day
429	429	7376	Stupini	1763	octombrie	3
1708	1377	7376	Stupini	1767		

tbl1 Person Details

Person ID	DaT18ID	Event ID	Recorded Name	Birth Place Location ID	Recorded Birth Place Name	Residence Location ID
429	429	429	Popa Vasile din Stupini	7376	Stupini	7376
1708	429	1708	Vasile Popovici	2174	Gimbuț	7376

Fig. 4. DaT18 Database detailed record view

Reformatting the sources

The structuring of digital information was deeply impacted by the sources themselves. While my aim was to always replicate as much of the original data, the heterogenous type of historical evidence I worked with meant that this was not always realistically viable or even possible. Adjustments to the principles described by Townsend et al. (ch. 3) and Mandemakers and Dillon (34–38) had to be made at various stages from data entry to its public release. From the outset of the project then, the purpose has never been to arrive at delivering a digital textual edition of the documents themselves, but rather to develop a structured resource, capable of holding and connecting information from various backgrounds, which would end up enriched by the presence of interdependent data.

Over the first two years of the project I focused on recovering the clerical careers of the some 1,500 Orthodox priests active throughout the 1760s, in an attempt to explain the birth of this social group and its possible roots in the earlier times of religious confrontation. My choice reflected the preservation of significant serial data concerning the clergy, unlike most of the lay leaders of dissent, thus making for more predictable outcomes in the initial stages. Equally important, the problems posed by working with early modern statistical records of church personnel have been the object of close scrutiny by the team behind the *Clergy of the Church of England Database*, which I reckoned as a distant model (Burns et al.).

Design decisions in building the database have been informed by the direct knowledge and experience of the types of documents susceptible to converge in the final application. In the end, it was all about reaching a compromise between filling-in the details to a general questionnaire that mirrored the avouched intentions of my future analysis and preserving the particularities in every source that was to be so

recast. This has hopefully resulted in a flexible frame that allows for a systematic extraction of data from the primary sources in order to quantify the spread of dissent, but, at the same time, maintain those details on the life of each individual that should not be overrun by generalizations.

As mentioned previously, much of the historical evidence targeted by the DaT18 project already comprised some form of structured data in the shape of either religious or fiscal censuses. The fields existing or deduced from the common pattern of the records have largely found their way intact into the new tables and their values have been transcribed literally to the database. Breaking down the information into its smallest components posed no problem for certain documents, for instance the staff register used by Bishop Novaković around 1761–1763 to keep track of the ordained clergymen, which only lists the parish and the priest's name, or the few local *urbaria* that have been included to assess the platform. Nevertheless, in this pre-statistical age census-like records were much less standardized than their modern counterparts, which raises significant issues that go beyond the frequently encountered spelling variations. Even in such a straightforward source as the above-quoted staff list, there are instances when the same person was registered twice and the error left uncorrected, while at other times individuals were first written in and later crossed out of the manuscript (e.g. "DaT18 Database," DaT18ID 660, 1021). In keeping with the requirements of source integrity, I opted to also include them as separate entries into the database, mentioning the anomaly in the notes.

Similarly, the various registers compiled during the eighteenth century still contain much verbiage even when the records conform to a pattern. Take the case of another book covering the period 1762–1767, in which at least three different hands inscribed the priests ordained, appointed, and confirmed by Bishop Dionisije Novaković. Here they no longer employed a tabular form, but rather added complete phrases on every single character, signalling his name, birthplace, parish of residence, ordination and appointment dates, etc. On entry into the database I standardized information except for people and place names and only quoted fragments from the original text in the notes when it said something significant regarding the life events or commented upon the person of the priest. This obviously induced some minor losses compared to the primary source, e.g. in converting relative to absolute chronology, but the trade-off was worth it in terms of data standardization and normalization (Fig. 5). In fact, more alterations were probably due to using the published edition (Hitchins and Beju) instead of the manuscript for harvesting data, although I did run parallel checks to the original and tacitly corrected the more flagrant transcription errors encountered.

The screenshot displays a web application for the DaT18 Database. On the left, a search bar is set to 'Person ID is "204"'. Below it, a table lists search results with columns for Person ID, DaT18 ID, Event, Recorded Name, Birth Place, Location, and Recorded Birth Place. The first result is for Person ID 204, DaT18 ID 204, Event 204, Recorded Name 'Popa Petru din Culcea', Birth Place 'Culcea', and Recorded Birth Place 'Culcea'.

On the right, a modal window titled 'tb11 Person Details Report' is open, showing a table with fields for Ordained By, Ordination Date, Ordination Place, Confirmed By, Confirmation Date, Presented By, Morality, Office Income, Personage, Glebe, and Notes. The 'Notes' field contains the text: "...având drept a servi cât timp este surd".

Below this, another modal window titled 'tb15 Personification' is open, showing a table with fields for DaT18 ID, Uniform First Name, Uniform Last Name, Active Lo, Active Hi, and Notes. The 'Notes' field contains the text: "Numerele lui lipsește din conscripția anului 1767."

Fig. 5. The use of the notes field in the DaT18 Database

Great disparities exist between information provided by the sources even within the same category and at a few years distance in time, which made it necessary that some details be inferred for the needs of statistical analysis. None of the three major lists that date from the 1760s talk about the gender of the parish priests, for instance. We know that in the Byzantine tradition only male candidates could be ordained, so there is no risk in assuming that the registered clerics were in fact men, in order to distinguish them eventually from the few situations when women were involved in the religious disputes. Likewise, denomination was mostly inferred from the presence on the official roll of Orthodox priests or from the charges brought against the laymen by the authorities, while occupation, which attempts to group the persons in the database in general categories, was filled based on their office, if lacking any positive statements. However, when in doubt, my choice has always been to leave the field empty rather than try and extrapolate something for which there was not enough proof.

Finally, transferring the narrative texts into table format implied an interpretative approach that came with its own norms and limitations. The most detailed sources are the transcripts of the questioning of presumed leaders of dissent, which offer a first-person account of their earlier life and recent actions, albeit these are to be read with utter consideration for the warnings in the literature (Davis; Farge). They range in length from a couple of pages to a dozen or more and the biographical information is scattered among the many answers. A literal transcription of an entire document would not have been feasible both from the perspective of the time consumed, as well as the formulaic nature of some of the questions and answers that would add nothing to the general knowledge. Besides, it would still not replace the filling-in of values in the predefined fields of the database, which was the absolute requirement. The imperfect solution was to handpick just those bits of information that fitted in the tables, such as

the names, age, dates, occupation, etc., and insert the detailed descriptions of events in abridged form only. Also, since questioning referenced past occurrences in addition to the latest and offered an unique insight into the perception of the Orthodox themselves on their opposition movement, these have generated multiple entries by exception to the above rule of solely using contemporary material ("DaT18 Database," EventID 1359, 1361, 1365). Supplementing this data with extended citations from the primary source seemed unnecessary at this point, because it would have been unsuitable for taking the research further anyhow. Inputting the information into fixed fields breaks internal relationships within the original text and fundamentally alters its coherence. The succession of questions and answers might seem unimportant for now, but it might pick up relevance later, which is why I decided against placing large inserts from the documents into the database itself.

Personifications

Record linkage operations were for the most part semi-automatic and simultaneous to data entry, when each person, event, location, and source received a unique ID number that helped establish instant relations between the different entities in the tables. The process was not always that simple and it did require specific skills beyond the reach of artificial intelligence. Due to language and spelling variations in the sources, pinpointing the precise location of places of residence and the events proceeded from information in historical gazetteers, but also involved a basic understanding of philology and geography, aside from some sheer strikes of luck for the most baffling of them. However, once the settlement was identified and the corresponding code attributed manually, the linkage to all other instances in the database was automatic, as was the assignment of standard names and dependencies.

More complicated and delicate resulted the outlining of individuals from the separate entries of persons in the database, for which I borrowed the term employed by Burns et al. to describe what they suggestively penned a "God-like process in which 'people' are created" (732). Personification turns empty names and cold facts into (once) living persons. Although inspired by a common set of goals and principles, the main sources that contribute to the database were not concerned with establishing an absolute identity for the priests they registered, instead focusing on single events that had recently happened or simply attesting their existence. The persons are unique only within the boundaries of the same source and their details are reshuffled for all others. Thus, names employed have a certain meaning for the compilers of the staff register in 1761–1763 but acquire a different value in the census of 1767. For the purpose of research, down to the basic query based on entering a name as keyword, they had to be structured so that different people could be discerned regardless of them sharing a common name and, likewise, references grouped together when these point to one person under many recorded names. By linking the recurring instances of the same character and providing a unified account of his undertakings, flat pieces of information become three dimensional and restore depth to the clerical careers.

Personification is subsequent to the input of significant data and constitutes a distinct operation altogether. Individuality is ascribed manually one case at a time by creating a standardized first and last name, by entering the interval of active years based on the earliest and latest records available, supplemented by a notes field that gives a reason if their career terminates before 1767, and also by assigning a unique ID which is to be replicated every time a different record of the same person is encountered. This has allowed for the indexing of 1,676 people distributed over 2,599 personal details' records, of whom 1,619 parish priests. The latter number is still some 150 to 200 individuals above the estimated total clerical figure for the diocese in the period under question, but further identifications are contingent upon unravelling new evidence on the fate of those persons for whom no connections could yet be determined.

The creation of identities rested on two fundamental elements: the recorded name of the priest and the locations he was linked with, whether just his residence or the residence and birthplace together. A name alone is not a strong enough argument for personification, particularly given the specific case of Romanian naming conventions of the time, and even in this combination they do not always work to eliminate all uncertainties. Nevertheless, considering that the database comprises of three different sets of data from a mere seven-year interval and that in the eighteenth century the clergymen rarely transferred from one parish to another, it is within a reasonable margin of error to assume that two or three records a few years apart that document people bearing the same name, functioning in the same parish, and possibly born in the same place are, in fact, multiple instances of one person (Fig. 6).

The screenshot displays the DaT18 Database interface. On the left, a search bar shows 'DaT18ID is 313'. Below it, a table lists person details for DaT18ID 313, with the first name 'Daniel'. On the right, an 'Overview' section shows a summary of the person's data:

DaT18ID	313
Uniform First Name	Daniel
Uniform Last Name	Mogilescu
Active Lo	c. 1760
Active Hi	1767
Notes	ok

Below the overview, a 'tblt Person Details' table shows multiple records for the same person, illustrating name entries variation:

DaT18ID	Event ID	Recorded Name	Birth Place Location ID	Recorded Birth Place Name	Residence Location ID	Recorded Residence Name
313	313	Popa Daniel din Castlu	33.2.2	Castlu	33.2.2	Castlu
313	1258	Daniel			33.2.2	Kasztu
313	2585	Daniel Mogilescu	33.31	Olcara	33.2.2	Castlu

Fig. 6. Name entries variation in the DaT18 Database

Names or rather the lack of them posed the greatest problems. Priests' surnames only appear regularly in the census of 1767, while earlier records give the Christian name alone or accompany it with the sobriquet "Popa," i.e. the priest, which efficiently marked the distance from laymen. When a positive identification was possible within the database, the standard last name was created on the pattern of that registered in 1767. This latest form took precedence over earlier records of the surname when they existed, except when previous patronymics or nicknames proved more suitable to distinguish between the bearers of generic surnames such as "Popovici," which is a derivative of "Popa" (e.g. "DaT18 Database," DaT18ID 113, 487, 613).

With only first names to work with, personification would be a nearly hopeless task. More than one fifth of the priests were baptized John or something similar, which hinders any attempt at discriminating between them if the location does not provide sufficient clues. It was already troubling those who entered information into the staff registers more than two and a half centuries ago, which is why they accommodated by either using generational suffixes to distinguish between younger and elder members of the family (e.g. "DaT18 Database," PersonID 1295, 1296) or by alternating the spelling of the names when they repeated for the same parish ("DaT18 Database", PersonID 1193, 1194). However, these distinctions are operational only within the framework of the primary source itself and they do not translate precisely into the next, meaning that such variations cannot prevent personification dilemmas on the whole. Furthermore, if between successive records one of the two persons ceased to exist and none of the previous details were kept with regard to the other, it is impossible to determine which of the parties was still mentioned. Issues with names account thereby for most of the uncertainties that have prevented a higher rate of record linkage.

Locations to which people were connected present obstacles of their own, which add to the already discussed question of spelling and language variants. For one thing, only two of the three main sources concerning the clergy contain a field on the birthplace, which drastically limits the reliability of the assumptions made on those people that share a common name and are known only through their parish of residence. Fortunately, almost half of those concerned by this situation were ordained by Bishop Novaković from 1764 to 1767 and have had little time and reason to change parish until the census in 1767, which confirmed their presence and logged more complete details. On a closer look, the mentioning of the place of origin does not seem an unfailing criterion for judging identities, at least not in the context of the fluctuating records. Far from being substantial, there are nonetheless a worrying number of cases in which the same person is credited with two distinct places of birth, one in the staff register of the early 1760s, the other in the census of 1767 (e.g. "DaT18 Database," DaT18ID 488). Even more doubtful appears the situation of the implicit provenance, suggested by the inclusion in the surname of a settlement designation, which is afterwards contradicted by reference to a different birthplace in the same source (e.g. "DaT18 Database," PersonID 1584). Finally, there are also parish priests who do change residence in the course of time. Again, this does not appear to be a widespread practice, but it is indeed hard to quantify with exact precision the proportion of those involved. While very few of them applied for the permission of the hierarchy to do so,

many others elected to transfer between neighbouring villages over which they probably already exerted authority. It has been possible to operate personifications in the case of those priests that had less common names (e.g. "DaT18 Database," DaT18ID 34), but obviously most escaped any identification since their residences no longer coincided.

All in all, these departures from the bureaucratic practices of record keeping and from the expected behaviour of those who had made the subject of such entries warn us over setting up personifications solely on conjectures. Inconsistencies and equivocal data are intrinsic to early modern sources and the DaT18 project did not try to run counter to that. In this sense, no record linkage was established whenever there was a doubt that could not be resolved. At most, plausible connections were suggested in the notes field of each individual, so that future research would not ignore, but arrive at verifying them.

Fiscal censuses, taxes, and wealth

Biography alone tells only half the story and it would not take us much further compared to traditional history writing on its own. By electing to focus on the social structures of dissent, the DaT18 project vowed to depart from the common narrative of who did what and when, in order to sketch a fuller and compelling picture of the background against which certain people rose to prominent positions within the opposition movement and the administrative structures of the reborn Orthodox Church. The economic resources of the individuals have been reckoned as an indicator of utmost importance for the prosopography of dissent, since they offer a more in-depth view of inequalities and status in the rural communities than the handful of official categories that segmented the Transylvanian society. Integrating this data is one of the long-term goals of the project and, although they were not a priority at this stage, the database was built to accommodate and make full use of them.

Driven by the quest for reforms of the Habsburg state or owing to the landlords' concerns for managing their estates, censuses became quite common in the second half of the eighteenth century. Apparently at least, the historian is spoilt for choice when it comes to the range of fiscal and economic material at his disposal, mostly still in manuscript in the archives. The final tables of the 1750 general fiscal census alone, the first successful initiative of its kind in Transylvania, span more than 33,000 pages (Gyémánt et al, 1: vii; 2, part 1: xiii), an overwhelming reality that makes the coverage of the entire province impossible for a single researcher. A different strategy will be put in place instead, which involves the careful selection of representative samples, both geographic and in terms of population, for which information will be added to the database. On the other hand, the preservation of the sources is problematic, for the data in the local censuses was made redundant with each new update of the statistics, meaning that they could be more easily disposed off, introducing as a consequence large gaps in chronology and the spatial distribution of these documents. With better survival rates, the state-sponsored censuses were, nevertheless, few and far between, providing only snapshots at the beginning and the end of the period under question,

which makes obtaining comparable sets of information quite difficult. Plus, given that data collection was to observe the existing exemptions, there is also a high probability that many of the community leaders may have been left outside of the records wittingly, due to their position as parish priests or village elders.

The biggest concern, however, was the lack of any standardized measurement of wealth across the sources. Some focused on the land plots held by each family and on the crops and their potential annual yield, while others registered mainly the livestock. Supplying a monetary value of either tax or income was an exceptional occurrence. Furthermore, since the census takers came inevitably from the ranks of the local administration, their records diverge substantially from one another based on jurisdiction, even when referencing the same categories. Further confusion reigns with regard to the units of measurement, as, for instance, two contemporary records in the Saxon seat of Sibiu alternatively used both shocks and bundles to approximate the harvest of flax and hemp (“DaT18 Database,” SourceID 11, 12). While data would have to be entered into the dedicated table in the database as it were in the primary source, it also had to be somehow converted using a common denominator for it to make sense. Comparing two cows and ten buckets of wine with one acre of land and three pigs to determine which one defines the richest is basically impossible. Hence the idea to use the methods employed by the Habsburg bureaucracy itself to assess a global figure for the taxes owed by the households and rely on it for any subsequent calculation.

There was no definitive model to start with, for in little less than three decades Transylvania experienced four different fiscal systems, passing from the traditional division on fiscal *gates* and *calculi* of pre-1753 to individual taxation according to the scheme proposed by Gábor Bethlen in 1754 and later revised by Adolf Nikolaus von Buccow in 1763 and Samuel von Brukenthal in 1769 (Trócsányi 590–596; Ionaș). The later three methods exchanged various formulas for calculating the due contributions that comprised of both a poll tax and money levied on assets and income. The poll tax grouped taxpayers into fiscal classes, which varied according to their social and economic status as well as their residence, but proved difficult to implement fairly and consistently, leading to corrective interventions and to further segmentation of the initial categories.

While revisions were also the norm for the taxes on assessable property and revenue, the definition of specific goods that were subject to taxation stayed constant for longer, with only their value changing over time. Table 1 below offers an overview of these differences across the three fiscal systems that went in successive use in the 1750s and 1760s. To some degree the naked figures are misleading, for they fail to disclose the entire complexity of the transformations, which impacted significantly how the landed property was taxed. The arable land was classified into four categories according to productivity and a rate was set for each, with further differentiation for vineyards and meadows. At first, the charge was on the actual crop, but it soon emerged that this practice forced the peasants to abandon some of their fields in order to pay less, and so, starting with the implementation of the Buccow- and afterwards the Brukenthal-system, the rate was linked to the surface of the plot, changing not only the amount, but the very substance of what was recorded in the tax rolls.

Table 1

Comparative overview of the tax amounts levied on property and income in Transylvania after 1754 (all figures in kreuzer).

Fiscal system	Land tax / ⅓ acre	Oxen / head	Cows / head	Sheep / head	Pigs / head	Wine / bucket	Hay / cart	Revenue / 1 Rfl.
Bethlen	8–20	30–40	20–27	3–4	3–4	0.75–1	1.5–2	3–4
Buccow	9–18	24	20	3	5	3	5	5
Brukenthal	8–20	24	20	3	5	3	6	6

Source: Author's own elaboration based on data in Bozac and Pavel (395–398).

The introduction by Bethlen of a tax-unit called *cubulus*, although soon abandoned by the following systems, was in hindsight a sensible choice when trying to equate objects that did not lend themselves easily to comparison. Except for crops, everything else from wine, hay, animals, to even revenue obtained from running mills, lending out properties, or selling brandy were each considered to be worth a certain amount of *cubuli* based on their market and approximate relative value, so that, for instance, 2 oxen were the equivalent of 3 cows or 20 sheep, pigs or bee hives, and of 40 wine buckets (Ionaş 81). A conversion rate of 3 kreuzer (4 in the case of the Saxons) to *cubulus* was then applied to calculate the tax liabilities of any given household. Later systems improved upon this model, as they dropped the *cubulus* and made all calculations directly in guildens, while also levelling the taxes between the inhabitants of the counties and those in the Saxon seats and altering at times radically their base value for better balancing between the various goods.

Despite its ultimate rejection by contemporaries, the *cubulus* makes arriving at meaningful values relatively straightforward for simple estimates of wealth. Not only did it cover an entire decade right in middle of the timeframe targeted by the DaT18 project, but this artificial tax-unit has also the merit of expressing economic resources into decimal format, thus sparing the trouble of working in guildens and its subdivisions, the equally factitious money of account of the Monarchy.

In a further effort to reduce the variables, I set out to only consider the animal stock as the most common indicator of wealth. At a time when much of the rural population only possessed a house with a scant inventory and a plot of land to feed themselves, the animals could count as the real differentiator. More draft animals made farming easier and meant that additional land could be claimed for cultivation by members of the same household, while an above average number of livestock signalled a diversification of income from selling the surplus products to the market. Regardless of the fiscal system in use or the purpose for taking the census, whether general or local, animals were almost never absent from the records. They were already understood as a distinct tax category by all three systems of the age. And, unlike in the case of wine and hay, which become harder to quantify in the 1760s,

counting them could hardly lead to massive differences of scale, though tax evasion did pose a serious threat to any realistic estimates, as eighteenth-century authorities were well too aware.

To sum things up, the table on wealth translates exactly the information in the primary sources, whether it concerns the estimated crop, the animal heads in one's possession, the structure of income, and even debts and exemptions. It then employs a simple formula for calculating the *cubulus* value of the animal stock and exports this figure to a field in the personal details page that is used to assess the economic value of the individual in relation to his peers in the database. By focusing on a limited amount of data and transposing it in a common format that is both consistent with the period and with our modern needs, the whole estimate is kept within some manageable limits.

Final considerations

The digital medium changes little in terms of the research routine of the historian. While the range of programming skills is ultimately optional in an industry that is more and more collaborative, the decisions I was faced with during the development of the database were not at all different from those I would have confronted in a more traditional setting. As always, the main dilemma centres on how to make the original sources answer questions on topics they were not initially intended to speak about, yet which they encompass. The DaT18 Database took shape as one possible solution to break the silence that the men and women at the helm of the religious protests in Transylvania kept on themselves for too long.

At the same time, prosopography and the relational database that provides the underlying data for it is just one piece in a wider approach aimed at interpreting the Orthodox dissent in Transylvania using the methods of social history and, equally important, reassessing the political practices of the subordinate categories involved in the opposition against the Greek Catholic Church. Constant improvements and updates are for certain needed to ensure the viability of the platform, but so is the actual investigation that motivated it in the first place. The database is nothing more than an instrument of research, the digital content assisting to open up new possibilities of interpretation. At the end of the day though, it remains the task of the historian to employ the best methods of his craft to arrive at a compelling scholarly analysis, which is not replaced, but rather augmented by the new medium.

Acknowledgement

This work was supported by a grant of the Ministry of Research and Innovation, CNCS – UEFISCDI, project number PN-III-P1-1.1-PD-2016-0296, within PNCDI III. Special thanks go to Oana Sorescu-Iudean and to Vlad Popovici for all their critical comments at various stages of my project, as well as to the anonymous reviewers who provided helpful suggestions on a first draft of the manuscript.

Works Cited

- Bernath, Mathias. *Habsburgii și începuturile formării națiunii române*. Translated by Marionela Wolf, Editura Dacia, 1994.
- Bozac, Ileana, and Teodor Pavel, editors. *Călătoria împăratului Iosif al II-lea în Transilvania la 1773 / Die Reise Kaiser Josephs II. durch Siebenbürgen im Jahre 1773*. Vol. 1, Institutul Cultural Român – Centrul de Studii Transilvane, 2006.
- Burns, Arthur, et al. "Reconstructing Clerical Careers: The Experience of the Clergy of the Church of England Database." *The Journal of Ecclesiastical History*, vol. 55, no. 4, 2004, pp. 726–737.
- Cohen, Gidon, et al. "Towards A Mixed Method Social History: Combining quantitative and qualitative methods in the study of collective biography." *Prosopography Approaches and Applications: A Handbook*, edited by K. S. B. Keats-Rohan, Occasional Publications of the Unit for Prosopographical Research, 2007, pp. 211–29.
- "DaT18 Database." *Dissent and toleration in Habsburg Transylvania: A socio-political history of the Orthodox protests (1740s–1760s)*, Version 3.1, 2018–2020, <https://www.dat18.ro/en/database>. Accessed 12 May 2020.
- Davis, Natalie Zemon. *Fiction in the Archives: Pardon Tales and Their Tellers in Sixteenth-Century France*. Stanford University Press, 1987.
- Farge, Arlette. *Le goût de l'archive*. Éditions du Seuil, 1989.
- Gyémánt, Ladislau et al., editors. *Conscripția fiscală a Transilvaniei din anul 1750*. Univers Enciclopedic, 2009–2016. 2 vols in 5 parts.
- Hitchins, Keith. *The Idea of Nation: The Romanians of Transylvania, 1691–1849*. Editura Științifică și Enciclopedică, 1985.
- Hitchins, Keith, and Ioan N. Beju, editors. "Documente privitoare la trecutul Bisericii Ortodoxe Române din Transilvania după 1761." *Mitropolia Ardealului*, vol. 19, nos. 1–3, 1974, pp. 13–46.
- Ionaș, Vasile. "Reformele fiscale din Transilvania în secolul al XVIII-lea." *Annales Universitatis Apulensis, series Historica*, vol. 4–5, 2000–2001, pp. 79–90.
- Keats-Rohan, Katharine. "Prosopography and computing: a marriage made in heaven?" *History and Computing*, vol. 12, no. 1, 2000, pp. 1–11.
- Keul, István. *Early Modern Religious Communities in East-Central Europe: Ethnic Diversity, Denominational Plurality, and Corporative Politics in the Principality of Transylvania (1526–1691)*. Brill, 2009.
- Mandemakers, Kees, and Lisa Dillon. "Best Practices with Large Databases on Historical Populations." *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, vol. 37, no. 1, 2004, pp. 34–38.
- Nedici, Radu. "Cum să pornești o revoltă în veacul al XVIII-lea: Activism, adunări publice și propagandă în comunitățile ortodoxe din Transilvania (1740–1760)." *Revista Istorică*, vol. 28, nos. 5–6, 2017, pp. 479–98.
- . "Religious violence, political dialogue, and the public: the Orthodox riots in eighteenth-century Transylvania." *Economy and society in Central and Eastern Europe: Territory, population, consumption*, edited by Daniel Dumitran and Valer Moga, LIT Verlag, 2013, pp. 87–100.

- . “Rethinking religious dissent in mid-eighteenth-century Transylvania: Political practices and the plebeian public sphere.” *Analele Universității București: Istorie*, vol. 63, no. 1, 2014, pp. 101–24.
- . “Spre o istorie socială a disidenței religioase: Un inventar al surselor referitoare la protestele ortodocșilor din Transilvania de la mijlocul secolului al XVIII-lea.” *Sursele unei istorii (pre)moderne românești în Moldova și Valahia*, 21 September 2018, Institutul de Istorie ‘Nicolae Iorga’, Bucharest. Conference Presentation.
- Roth, Paul W. “Das Diploma Leopoldinum: Vorgeschichte, Bestimmungen.” *Siebenbürgen in der Habsburgermonarchie. Vom Leopoldinum bis zum Ausgleich (1690–1867)*, edited by Zsolt K. Lengyel and Ulrich A. Wien, Böhlau Verlag, 1999, pp. 1–11.
- Townsend, Sean, et al. “Digitising History: A Guide to Creating Digital Resources from Historical Documents.” *AHDS Guides to Good Practice*. University of Essex, 1999, http://hds.essex.ac.uk/g2gp/digitising_history/index.asp. Accessed 12 May 2020.
- Trócsányi, Zsolt. “A New Regime and an Altered Ethnic Pattern (1711–1770).” *History of Transylvania*, vol. 2, edited by László Makkai and Zoltán Szász, Social Science Monographs, 2002, pp. 515–618.

Book Review

**Anna Wing-bo Tso, *Digital Humanities and New Ways of Teaching*,
Singapore: Springer, 2019, 249 p., ISBN 978-981-13-1277-9.**

As countries across the globe face the COVID-19 crisis, the topic of the digital humanities might be of special interest. Undeniably, the use of digital means to teach across the globe increased due to this, making the digitization an important process that offers useful resources. According to *Britannica*, humanities are traditionally seen as the disciplines that deal with human activity and culture, having their roots in the ancient Greek concept of *paideia*. Today, the digital humanities stand for the combination of digital tools and means with traditional humanist disciplines. This book is part of the Digital Culture and Humanities series by Springer and aims to provide a broad picture on how humanities are impacted, expanded and challenged in the age of technology in Asia and the Pacific region.

The editor of this volume is Anna Wing-bo Tso and she is currently an Associate Professor of English and Comparative Literature at the School of Arts and Social Sciences where she supervises the Master of Arts in Applied English Linguistics. She is also Director of the Research Institute for Digital Culture and Humanities and is the editor for the series of books titled *Digital Culture and Humanities*, published by Springer. Anna Wing-bo Tso has a keen interest in children's literature, gender studies and studies concerning translations and literature.

Digital Humanities and New Ways of Teaching is divided into four parts that contain about three papers each. The papers are grouped around the theme of each part and deal with archives and cultural heritage, the current situation regarding research on digital humanities in Asia and the Pacific area, teaching and digitization, and lastly discussing some future directions of the digital humanities in the mentioned area. They all allude to the same idea and that is that the practices of digital humanities can have positive results when they are used inside the university and other academic-related institutions. By giving examples and showing the results from various conducted studies, the goal of this book is to argue for the use and study of digital humanities on a larger scale in the Asian academic environment. The set goal is well achieved through the selection of papers the editor, Anna Wing-bo Tso, did as most

of the authors back up their claims with some type of evidence, have an analytical approach and try to make use of quantitative, qualitative and also mixed methods that help to emphasise certain ideas or aspects.

The first segment explores a situation beyond the Eastern scope, while also noting some very useful and more general applications. By reading this part (Monika Gänßbauer) one finds that the German-speaking world shares different opinions over the methods employed by digital humanities, having a general disinclination towards them. Even though some call Germany a country that is in hibernation in terms of digitization, the author effectively argues otherwise. The second chapter (Jack Hang-tat Leong) focuses on Chinese Canadian studies and how digital means are used by librarians and scholars for the study of heritage materials. The first presented project employs digital means to show how the Chinese people helped to build North America through railroads, the second shifts to the entire Chinese diaspora from Toronto, while the last two projects provide resources on Chinese migrants from around 1960's up to 2015. The third chapter (Sir Anril Pineda Tiatco, Bryan Levina Viray, and Jem Roque Javier) presents the fragile condition of material archives and documents, suggesting that this could be changed with a digital archive. The presented project deals with the cultural heritage of cultural performances through an online archive created to house all the data on such performances from Philippine.

The second part informs the reader on current research on digital humanities, from individual experiences to collective efforts. The fourth paper (Andrew Parkin) presents the experience of author Andrew Parkin from the type-writing where no mistake was allowed and the possibility to modify the document was not an option, up to the laptop and internet that allow a writer to easily get his/her work published. The fifth and sixth papers (Andy Chi-on Chin and Chaak-ming Lau) deal with a common topic of Cantonese related projects and studies. Author Andy Chi-on Chin starts by outlining the importance of Cantonese in Hong-Kong, continuing by launching a description of the construction and logic behind a Cantonese corpus. He then describes the two phases of the project, with emphasis on the latest as it involves the use of video segments and provides an ontological type of information. The last chapter of this segment details the need for constructing a thesaurus online dictionary, with an open data policy. Chaak-ming Lau explains that this project relies on the work on voluntaries that submit, edit and review content so that the users can enjoy verified and processed information. He details the process of selection, the use and role of social media, while the reader is presented with an analytical review of the project's strong and weak points that can help to guide similar endeavours.

Teaching in conjunction with digital humanities is introduced next. In chapter seven (Helena Hing-wa Sit, Sijia Guo) the flipped class method is presented to enhance the acquisition of a second language. As an extension of the *Developing Online Capacity in Introductory Chinese Language Units*, the project consists of a variety of activities that have both video and audio resources for learning. The research methods used are both quantitative and qualitative, with fifty-six entries in an online survey to provide data that prove flipped class to have positive results. The next

paper (Noble Po-kan Lo, Billy Cheuk-yuen Mok) tackles the language used by gaming, suggesting that it might prove to be a resource for teachers. After introducing gaming literacy, a set of basic characteristics follows that enable the identification of gaming language and terms. After conducting a survey, the research provides a table with the best-known gamers terms and in-depth results. The topic of e-literature teaching in the Pacific is explored in the ninth paper (John Paolo Sarce), as it's both contested and explored in the East. John Paolo Sarce launches a critique of post-colonialism and its relationship with technology. He argues on the uneven distribution of digital humanities centres and its possible factors and proceeds further to a critical approach towards school policies against technology. In the end, the author supports and highlights the novel forms of e-literature from the Philippines, the text tula or the social serye.

The last part casts itself upon the future directions of digital humanities and so do the three papers that form it. The first (Anna Wing-bo Tso, Janet Man-ying Lau) tackles the multimodal approach for the "Claude Monet: The Spirit of Space" exhibition from the Hong Kong Heritage Museum. The authors discuss the digitization from the last decade, explain multimodal literacy and practices, and state their aim to investigate the effectiveness of multimodal methods in exhibitions. Research questions, a part that details the Monet exhibition, the methodology used, its findings and the answers given by the visitors that joined the study are provided. The last two papers (Dora Wong and Winnie Siu-yee Ho) deal with digital literacy in different age groups. Groups of students join a creative project where they have to learn how to use both digital and traditional means to create a story through a digital video that is the result of the project. Three case studies show different means of achieving the same end, describing how the particular experience impacts the participants. The last piece of research argues that digital literacies can be found and developed on social media. The study has as target adult users of the Hong Kong Air Cadet Corps, via Facebook and the aim of author Winnie Siu-yee Ho is to show how online literacies are a resource for the volunteers. She makes use of semi-structured and unstructured interviews with the users, trying to uncover the history of digital literacy of the user and their purpose for using Facebook.

As can be seen from the summary of the book, the papers manage to follow the same silver thread and keep their discussion around education and teaching. The evidence used to support each author's claim is well selected and well used, in the forms of surveys, interviews, with quantitative and qualitative methods. Keeping this in mind, the selection tries in most cases to engage with its audience. It has to because the audience is made up of people that will be impacted directly by the use of digitization in various aspects of education. The results of this book show clear, palpable evidence on how digitization affects, or not, the lives of these individuals.

However, in most cases, the number of survey participants is quite small and this may count as a detriment to the bigger picture. If one particular reader is looking to get a wider and yet general view, the number of the participants and their opinion might not be of much help. Despite this, the merit of this book lies in the various examples it offers and the potential of serving as an example for countries that are looking to develop their digital humanities practices in connection to teaching. Even if the focus is on Asia, the intended audience has a larger scope, that of teachers and

students alike. The obstacles and achievements of the authors that are mostly engaged in the pedagogical area may inspire and guide teachers and researchers that find themselves in similar situations.

The present volume is part of a series that proves a growing interest with digital humanities and new teaching methods, painting an honest and detailed image of this in Asia. Touching and discussing a wide range of situations and projects, it effectively presents an inspiration for both practical and theoretical future approaches. By employing a rich bibliography, the authors provide numerous means for a researcher to read further on the subject. All the papers have an analytical way of questioning and investigating the matter at hand, often using combined methods to support the arguments.

All in all, one should read this book and even consult other titles from the series if one is looking to find out how individuals deal with rather unwelcoming opinions and views regarding the digital humanities and digitization, as it provides ways of effectively employing the digital means in teaching practices. Based on the surveys conducted by the authors, a teacher can use the information to adapt his ways of teaching to the possible needs and capacities of the students. Using and processing the given info, this teaching process can be taken and improved further or it can be successfully introduced in countries that find themselves in a digital slumber. The use of such a book is evermore growing in the present context when it seems as if teaching becomes dependent on technology. By being unable to hold classes face-to-face, teachers have to find new and innovative ways of employing technology in their teaching process. This book can be a great example and inspiration which can provide a number of strategies that are so crucial for this current predicament. The flipped classroom, the multimodal approach, the many online platforms and projects highlighted are inspirational and outlined well enough to be replicated. The theories on the various approaches on the conjunction between digital humanities and teaching can have a practical use, starting from the theoretical framework laid out by this series.

Anisia Iacob

Book Review

Olivier Le Deuff, *Digital Humanities. History and Development*, London: ISTE, Hoboken: Wiley, 2018, 165 p. ISBN 978-1-78630-016-4

The contribution proposed for analysis, published in 2018 by ISTE and Wiley, is included in ISTE's Intellectual Technologies Set coordinated by Jean-Max Noyer and Maryse Carmes. Rather than presenting an exhaustive history of digital humanities, the author, lecturer in Information and Communication Sciences at Bordeaux Montaigne University, manufactures a sketch that traces the origins of digital humanities to the birth of modern science. The paradigm postulated by the author breaks away from the vision of congruence between computer technologies and digital humanities and suggests that, alternatively to the generation of digital humanities by the computing tools, digital humanities precede computer technologies. The thesis proposed by Olivier Le Deuff, however valiant, is also audacious, as the precedents he sets to prove the existence of digital humanities prior to the Digital Revolution may seem strained. As the author himself states, more than a presentation of the history and development of digital humanities, the book ought to be treated as an archeology of knowledge and methodologies, a presentation of the antecedents of the current directions in digital humanities.

The book is devised as a diptych; on the one hand, tracing the genealogies of the digital humanities and, on the other hand, deliberating on the evolution of several fields of the digital humanities. Each part of the book comprises five chapters and the perspective proposed is chronological, but also thematic. The first part distinguishes and evaluates five nodal points, precisely five genealogies that determined the evolution of digital humanities, while the second part analyses current directions in digital humanities and their precedents. The perspective constructed by the author in the first five chapters is similar to a stair, which leads from the origins of digital humanities, at the dawn of modern science, to the current directions that mark this field of science, each chapter representing a tread.

The first step regards the emergence and the evolution of the Republic of Letters into the Republic of Sciences. The Republic of Letters is not to be perceived as a circuit of letters and communications between savants, but as a "learning circle",

a concept theorized by the author in the *Introduction*, which represents the milieu of the scientists, intellectual and social alike. The growth of knowledge, concurrent with its differentiation, leads, in turn, to the need of rationalization, one of the rationales of digital humanities. Thus, the information must be operationalized and one of the instruments that enables us to operationalize it is the index, instrument that determined, according to Olivier Le Deuff, the formation of the digital humanities. Prior to considering the index's role in the evolution of digital humanities, the work proposed for analysis regards digital humanities as a science of writing, emphasizing the need for interdisciplinarity. Returning to the problematic of the index, the infobesity, term used frequently by the author, is considered the main factor which shaped the need to handle large amounts of information. The author links the index with computer technologies, highlighting its role as a precursor of the electronic means. Were we to summarize the author's theory this far, we would be able to notice a network, the "learning circle" of the Republic of Letters, and an instrument, the index, both perceived as being part of incipient digital humanities. However, the index was not sufficient to handle the information overload, because it treated only a book or an author, not all the printed works. As a result, other instruments were called upon, instruments such as notebooks and library catalogues. Not only the network and the instruments used by the scientists were symptomatic of a new field of knowledge, the digital humanities, but also the adjustments of the scientists' habitat or workstation, as Olivier Le Deuff regards it. The transformations suffered by the scientist's habitat were generated by the same condition that generated the instruments above-named, infobesity, precisely the need to access multiple documents. The author suggests that the transformations marking the researchers' workstation, ultimately, lead to the development of the Web.

The thesis postulated by the author in the first five chapters of his work implies that the formation of the Republic of Letters caused an information overload, which, in turn, determined the inception of digital humanities. However, we believe that the theory above-mentioned is strained, because, rather than digital humanities, it handles something that, indulgently, may be referred to as pre-digital humanities. We consider that it is difficult to discuss about digital humanities prior to the 20th century, as the technical means were analogue. Although the evolutions mentioned by the author represent solid precedents to digital humanities, we appreciate that they are a far cry from what digital humanities constitute.

As the Nietzsche quote used as a motto for the sixth chapter suggests, the last five chapters treat several methods used by the digital humanities and their evolution, from their pre-digital traces to contemporaneity. Alternatively to the perspective of the first half of the book, which places each field of the pre-digital humanities, for instance, the indices, in a diachronic evolution of the entire phenomenon, the second half of the work proposed for analysis deliberates on each field separately, autonomously of the formation of the digital humanities as a scientific phenomenon.

Similarly to the approach of the previous chapters, the author traces the origins of each method and field regaled to the dawn of modern science. For instance, the quantitative methods currently used by the digital humanities originate in statistics,

perceived, in a Foucauldian manner, as a pillar of the modern state, as a means used by the emerging modern governments to control their individuals. However, the current quantitative trends used by digital humanities are determined by the broadening of statistics and quantification at the turn of the 19th century, reflected in Émile Durkheim's works and, later, in the directions of the Annales School. In the *Conclusion*, the author pleads for an alliance between quantitative and qualitative methods, because digital humanities consist in documentation and production as well. Another field of digital humanities illustrated by Olivier Le Deuff in a diachronic approach is that of automatic processing. The author traces the origins of automatic processing to Jean-Claude Gardin, rather than Robert Busa, and he also identifies a distinct period of pre-digital humanities, that of the *Humanities Computing* age. A nodal point between the *Humanities Computing* and the digital humanities, as the author states it, is the success of the Web. Therefore, the transition from pre-digital humanities to digital humanities was assured by the advent of the Web. Metadata is also a source of the digital humanities, the book detecting its origins in the days of Mesopotamia, its formalization in the creation of catalogs and, its current use with the markup languages. Scientometrics' antecedents, as a field of digital humanities, are discovered in bibliometrics. The perspective postulated by Olivier Le Deuff is not that of simple quantifications of articles, but that of a field which perceives science as a system whose evolution can be analyzed and even predicted. Finally, the author analyzes the use of maps in digital humanities and he highlights the fact that maps illustrate more than territories, but different aspects of the reality and of the Internet infrastructures as well. A methodological issue that we must raise is that of the criteria used by the author to select and order the above-mentioned fields and methodologies of the digital humanities. We believe it would have been appropriate for the author to name the criteria based on which he selected and ordered the contents of the last five chapters.

According to Oxford's Lexico, digital humanities represent an academic field concerned with the application of computational tools and methods to traditional humanities, disciplines such as literature, history, and philosophy. The book proposed for review suggests a broader perspective, one which traces the origins of digital humanities beyond computational tools and follows the phenomenon in the long term. The theory envisaged by Olivier Le Deuff is tributary to an evolutive perspective, rather than presenting the inception of digital humanities as a result of an intellectual revolution. However, even if the rediscovery of scientists such as Emanuel Goldberg and Paul Otlet, which places the advent of digital humanities prior to the contributions of Vannevar Bush and Robert Busa, may seem legitimate, tracing the origins of digital humanities to the birth of modern science is strained, because, rather than the origins of digital humanities, it reveals the origins of humanities. Nevertheless, Olivier Le Deuff's work is valuable because, as he suggests, it sketches an archaeology of knowledge, a tableau of the distant sources that precede digital humanities. A compelling division is introduced by the author in the chapter regarding automatic processing, as he introduces the notion of *Humanities Computing*, to define an age that precedes that of the digital humanities. The criterion that divides *Humanities Computing* and digital

humanities is that of the succes of the Web. Consequently, the development of the digital humanities was determined by the emerge of the Web, prior to the Web existing a pre-digital humanities aeon, with a distinct final age, the age of *Humanities Computing*.

Following Alan Liu's approach, Olivier Le Deuff succeeds in sketching a synthesis of the history of digital humanities in the *longue durée*, dismissing a revolutionary logic, although his reasoning concerning the distant origins of digital humanities may be questionable. One of the author's conclusions is that digital humanities tend to transform into a broader topic, encompassing mankind and changing its characteristics to a digital humanity. For instance, in the field of science the researchers who embrace digital humanities are compared by Alan Liu to dragonflies that abandon their chrysalis stage. Hopefully, the evolution in store for this academic field, will not be as short lived as the lifespan of a dragonfly.

Alexandru-Augustin HAIDUC