"Future Philology: Digitization and Beyond":

A Two-day Online Symposium organized by the Invisible East Programme, University of Oxford:

Thursday 30 September – Friday, 1 October 2021

Background

At present, philology--the old art of reading slowly and carefully--is reinventing itself to become a discipline equipped with database technology. Scholars are setting up digital corpora of documents and manuscripts of various periods and regions. These digital corpora not only made precious historical sources much more accessible to academia, but also opened up new avenues of research on their content as well as their materiality. The Invisible East Programme at Oxford University is contributing to this field by creating, for the first time, a multilingual digital corpus of documents from the pre-Mongol Islamicate East. This region includes Iran, Afghanistan and Central Asia, in which documents were written in, inter alia, (Judeo-)Persian, Arabic, Middle Persian, Bactrian, Sogdian, and Khotanese. We are now inviting scholars working on databases of documents from the ancient and medieval world to share their experiences and insights.

Programme summary

The online symposium will last two days and be divided into four sessions of three to four presentations each. Each presenter will deliver a **20-minute** lecture followed by a **10-minute** discussion. A concluding session on the second day will wrap up the symposium.

DAY ONE: Thursday, 30 September 2021

All the times in PM BST (UTC +1)

Opening Remarks

2:00 - 2:15 IE Programme Director Arezou Azad and ERC Post-Doctoral Researcher Zhan Zhang, University of Oxford

Session 1 The Back-End 1 - Assembling and interpreting manuscripts

How do we select or collate manuscripts? To what extent is computerized reading of ancient manuscripts feasible in the current state of AI tools, and how do existing specialized tools contribute to paleographical studies? What kind of development is still needed? How do we piece together fragments from the manuscripts? What is better left to be done in "the old-fashioned way"?

Chair: Columba Stewart, Hill Museum & Manuscript Library

2:15-2:45 "Beyond Neural Networks: eScriptorium as an Edition Environment"
Daniel Stoekl Ben Ezra, École pratique des hautes études

2:45-3:15 "Tocharica digitalia"
Hannes Fellner, University of Vienna

3:15-3:45 "The Audition Certificates Database"
Konrad Hirschler, Universität Hamburg
Thomas Efer, Universität Leipzig

3:45-4:15 "Gandhari.org and READ: Digital Philology for Gāndhārī and Beyond"
Stefan Baums, LMU Munich
Stephen White, Ca' Foscari University of Venice

4:15-4:30 Session 1 wrap-up and general discussion

4:30-5:00 Coffee break

Session 2: The Back End 2 - Corpus structure, organization and search

What is the best practice of meta-data and data-tagging? What can tagging do that was inconceivable or infeasible before? In what ways is our perception of the data shaped by the metadata?

Chair: Marina Rustow, Princeton University

5:00-5:30 "Trismegistos: Connecting Texts from the Ancient World"
Mark Depauw, KU Leuven

5:30-6:00 "As simple as possible, but no simpler: basic principles for corpus organization"
Maxim Romanov, University of Vienna

6:00-6:30 "Just about everything pre-1517: Set-up, contents, and new features of the Arabic Papyrology Database (APD)"
Ursula Hammed, LMU Munich

6:30-7:00 "Digital Library Organization and Accessibility"
Émilie Pagé-Perron, University of Oxford

7:00-7:15 Session 2 wrap-up and general discussion

DAY TWO: Friday, 1 October 2021

Session 3: The Front End - Application and Visualization of Data

How do we apply and show data analysis results, e.g. in morphological and/or lexical studies, and visualized prosopographies.

Chair: Mark Geller, Freie Universität Berlin

2:00-2:30 "Digitization and Text Database: from the case of Old Tibetan Studies"
Kazushi Iwao, Ryukoku University

2:30-3:00 "Visualization and Search of Zoroastrian Manuscripts: The Avestan Digital Archive and the Corpus Avesticum Berolinense"
Alberto Cantera, Freie Universität Berlin
Miguel Ángel Andrés-Toledo, University of Salamanca

3:00-3:30 "Data-Driven Modeling: Building Data Models from Ancient 'Archives"'
Adam Anderson, UC Berkeley

3:30-4:00 "Digital Etymology on Screen: Some Basic Principles"
David Calabro, Hill Museum & Manuscript Library

4:00-4:30 "Analysing and editing medieval astronomical tables: metadata and front-end in DISHAS"
Matthieu Husson, CNRS, SYRTE-Observatoire de Paris-PSL
Ségolène Albouy, CNRS, SYRTE-Observatoire de Paris-PSL

4:30-4:45 Session 3 wrap-up and general discussion

4:45-5:15 Coffee break

Session 4: The Outside World - Participation and Cooperation

New methods of enabling wider academic audiences, and non-academic audiences to participate in deciphering texts and developing databases, such as crowdsourcing, are emerging. What are the pros and cons of such public approaches? What are the lessons learned? Is this something we should encourage/develop? There is a growing number of important digital corpuses of texts in the world that are publicly available but not inter-searchable. How can we and should we interlink these?

Chair: Sarah Savant, Aga Khan University

5:15-5:45 "The Giza Project at Harvard University: Can public outreach and scholarly access co-exist?"
Peter Der Manuelian, Harvard University

5:45-6:15 "The Princeton Geniza Project: Crowdsourcing and Interoperability"
Marina Rustow, Princeton University

6:15-6:30 Session 3 wrap-up and general discussion

6:30-6:45 Concluding remarks by IE team members

Abstracts

Session 1: The Back-End 1 - Assembling and interpreting manuscripts

Daniel Stoekl Ben Ezra, École pratique des hautes études
Beyond Neural Networks: eScriptorium as an Edition Environment

eScriptorium currently is the only fully open-source infrastructure for automatic analysis of handwritten documents. Based on the open source kraken AI system, the user can train the computer to automatically segment the layout and transcribe the text of complex documents in any sequential script. Advanced users have full control of the neural network architectures and a rich API. However, in addition to that, eScriptorium is becoming quite a comfortable environment for editing texts with complex layouts. The presentation will focus on its use for Genizah fragments in the new HTR4PGP project and also refer to our work on Dead Sea Scrolls, Greek papyri and Medieval Hebrew and Arabic manuscripts

Hannes Fellner, University of Vienna
"Tocharica digitalia"

Tocharian is a comparatively understudied branch of the Indo-European language family. The two Tocharian languages, Tocharian A and B, were spoken on the northern route of the Silk Road in the Tarim Basin in today's Xinjiang Uyghur Autonomous Region of the People's Republic of China in the first millennium CE. They are mainly attested in the form of manuscript fragments written in Tarim Brahmi, a Central Asian variant of the Brahmi writing systems, dating from around the 4th to the 10th century CE. There are several challenges for traditional approaches to philology, paleography, and linguistics presented by Tocharian. The Tocharian corpus is utterly fragmented. There are almost no complete manuscript leaves and therefore almost no coherent texts. Furthermore, Tocharian B – which accounts for two thirds of the corpus – shows a high degree of variation both linguistically and paleographically on both the synchronic and diachronic axes. This talk will address some of these challenges and show how digital tools – currently implemented in the Tocharian databases in Vienna – allow us to tackle them. The general aim of the talk is to emphasize that digital philology not only provides

the means of answering traditional questions, but paves the way for new avenues of research that would otherwise be impossible in the sphere of Tocharian and beyond.

Thomas Efer, Universität Leipzig
Konrad Hirschler, Universität Hamburg
"The Audition Certificates Database"

Arabic manuscripts contain an outstandingly large number of notes that owners and users wrote on them, among them audition certificates *(samāʿāt)*. These certificates documented the public reading of a text during which students received the right to transmit this text further. On account of their wide dissemination and central position in transmitting scholarly authority these certificates represent a treasure trove of prosopographical, topographical and other data. So far, the field of Middle Eastern history has used this unparalleled source for medieval history only for individual manuscripts or small corpora. The Audition Certificates Database (ACD) aims at remedying this situation by offering the first large-scale tool for future scholarship.

The data acquisition for the database begins with the digitization of the manuscript notes in a machine-readable full text format that is faithful to the source material. Building on that, the relevant passages mentioning people, places, institutions, dates and other so-called entities can be marked in place. Those annotations can further be augmented, for example with scholarly reconstructed renderings and transliterations of their names. They can also be completed using auxiliary sources and reference works, contextualized by connecting them throughout different notes across manuscripts and finally linked to authority files. In this manner they form an ever-growing knowledge base that allows for detailed query of the contexts of specific books and actors. Furthermore, they enable scholars to uncover larger trends and networks through the use of digital humanities techniques.

Stefan Baums, LMU Munich
Stephen White, Ca' Foscari University of Venice
"Gandhari.org and READ: Digital Philology for Gāndhārī and Beyond"

Starting in 2002 and reaching completion in 2014 (with ongoing updates), Stefan Baums and Andrew Glass compiled a complete corpus of all manuscripts, inscriptions and other documents in the Gāndhārī language, as a basis for their Dictionary of Gāndhārī and presented on the website Gandhari.org (https://gandhari.org). In 2013, development began (with Stephen White as lead programmer) on a comprehensive software package for digital philology called READ (Research Environment for Ancient Document; https://github.com/readsoftware/read), which now serves as a new backend to Gandhari.org and as tool in a wide range of philological projects on languages and scripts as diverse as Sanskrit, Pali, Tibetan, Egyptian, Latin and Mayan. This presentation will give an overview of the features and uses of READ for studying the material and textual aspects of Gāndhārī and other documents, and for managing their diverse interpretations, with a special focus on three tasks for which it provides semi-automated so-

lutions: (1) the creation of paleographic reports on the basis of linked images and transliterations; (2) the reassembly of manuscript or epigraphic fragments into reconstructed documents; and (3) the guided interpretation of written symbols with computer-vision methods. For each of these solutions, it will evaluate the amount of support they can provide for different kinds of source material, and the human philological judgement and labor still needed in conjunction with them.

Session 2: The Back End 2 - Corpus structure, organization and search

Mark Depauw, KU Leuven
"Trismegistos: Connecting Texts from the Ancient World"

Trismegistos (www.trismegistos.org) is a platform aiming to bring different disciplines of the ancient world together, by integrating basic information about texts of any genre, on any writing surface, and in any language or script. It currently focuses on the area from Scandinavia to Ethiopia and from the Canary Islands to the Indus Valley. For the moment, chronological limitations are 800 BCE - 800 CE, although some earlier and later texts are included as well.

Trismegistos' ambition is not to replace other databases: it is obvious that a small project such as ours cannot keep abreast of evolutions in so many disciplines dealing with the ancient world. Our partners provide us with enough information to identify the texts and assign a persistent identifier. We make the documents retrievable through searches with basic criteria, but for the text itself of the ancient document or a photograph, users of Trismegistos are led to other databases by links. Access to the sources is thus facilitated for scholars from other disciplines, and the unambiguous identification greatly enhances the exchange of information between projects, including GLAM institutions.

On top of the text database, Trismegistos has built sections such as People, Places, Collections, Authors, Editors, Time, and others. The persistent identifiers for each of these are a further contribution to setting up a Semantic Web for the ancient world.

Maxim Romanov, University of Vienna
"'As simple as possible, but no simpler': basic principles for corpus organization"

Data collection and organization are an important part of any research cycle, especially when it comes to long-term projects. Working with research data, we all strive toward the same goal—to ensure that our process of data encoding is efficient and our data is sufficiently robust not only for the current project but also for future uses. There is a variety of approaches to how research data should be collected, structured, and organized. The most common ones, however, (like, for example, relational databases and TEI XML) are either overly complicated, inflexible or may confine our data to isolated silos that will be of limited use both to ourselves and the field at large. The paper will present

a robust yet low-tech approach to annotating texts, organizing them into a machine-readable corpus, and supplying them with relevant metadata. At the core of this approach are the following general principles: 1) Linked Open Data (LOD), mainly the use of Uniform Resource Identifiers (URI) that are the key to linking any kind of information objects; 2) markdown, an approach to text annotation that is intended to be as easy-to-read and easy-to-write as possible; and 3) Arabic betaCode, a method of representing Arabic text with ASCII characters that allows for automatic conversion between transliteration and Arabic text.

Ursula Hammed, LMU Munich
"Just about everything pre-1517: Set-up, contents, and new features of the Arabic Papyrology Database (APD)"

The basic aim of the Arabic Papyrology Database is to provide access to pre-1517 Arabic documents on different writing materials (except coins and inscriptions). Full text documents from published research on the one hand, and thousands of unpublished documents with metadata on the other hand, form in itself a research tool, but also lay the basis for further functions of the APD.

Among the most important features are the lexicon function, in which all expressions found in Arabic documents have been included as lemmata. Accessibility to non-Arabists is facilitated by the use of transliteration.

Apart from giving full text of editions, the database also provides a comprehensive overview on the literature published in the field and in neighbouring, relevant disciplines (BIBLIO tool). Moreover, it highlights re-editions or variant readings of previously published material, and the emendations and additions therein are fed directly into the full text display.

During the last years, the APD was extended to accommodate a tool for typology and one for palaeography. The TYPO tool uses a novel approach of categorizing Arabic documents into formal types, including descriptions of the individual "typical" structures, and examples.

The newest feature, which is still being refined, is the PALEO site. So far, it presents examples of paleographic shapes and line inclinations, measured along a combination of several angles. All these functions make the APD an indispensable instrument for papyrologists, who widely use it, but also for other researchers not acquainted with the field, for whom it makes Arabic documents accessible.

Émilie Pagé-Perron, University of Oxford
"Digital Library Organization and Accessibility"

This talk discusses the structure and organization of the data, metadata and visual assets which the Cuneiform Digital Library Initiative (CDLI) manages, and how the accessibility of these data and assets is intimately linked with their organization. From its in-

ception 25 years ago, CDLI has seen its data structure evolve as the digital landscape changed. While its transliteration format endured, its linguistic annotations have migrated from XML to inline annotations and finally to the CoNLL format. Its metadata started as a few string type data fields, have continued as a complex interlinked database, and now to Linked Open Data formats using shared ontologies and vocabularies. CDLI data has been organized around specialists' needs but over time has expanded to formats and classifications that are widely used outside the field of Assyriology. This not only makes the data more accessible to a wider audience, it renders possible the approach of our data from other scholarly perspectives and use this same data quantitatively as part of wider reaching studies, e.g. on specific materials or linguistic elements across artifacts types and languages, well outside the scope of the initial research field. Finally, the way to navigate the data has also changed over time: from spreadsheet-like presentation to complex search queries including the usage of regular expressions, while now also including multi-layer annotations querying; CDLI has renewed with simple and accessible multiple-entry browsing, has added fuzzy searching, and filtering search results in order to lower the access barrier to the data, including for non-specialists.

Session 3: The Front End - Application and Visualization of Data

Kazushi Iwao, Ryukoku University
"Digitization and Text Database: from the case of Old Tibetan Studies"

Discoveries of Central Asian documents in the 19th century drastically changed the understanding of the early history of Central Asia. Dunhuang manuscript is a good case: since the accidental open of the hidden cave which was packed with several ten thousand old manuscripts at the beginning of the 20th century, many researchers have studied these and found new facts even now.
However, as the manuscripts are dispersed all over the world and kept at various libraries, it was not easy to look over the whole collection of manuscripts. The International Dunhuang Project (IDP: http://idp.bl.uk), the project of the digitization of manuscripts kept in these libraries, started in the British Library in 1994, and soon changed the situation: researchers can now check high-resolution photos of manuscripts from all over the world.
On Old Tibetan studies, one more web project was going on: the Old Tibetan Documents Online (OTDO: https://otdo.aa-ken.jp/). It aims at the construction of the text database of the Old Tibetan materials from Dunhuang, Central Asia, and Tibet. Therefore Tibetologists have now both photo and text databases. Two tools currently work in cooperation to run the study of manuscripts very well, but we are seeking new features to progress studies such as philology, morphology, and codicology. In this talk, I will mainly introduce the history of the study of Old Tibetan documents and the current challenge of the OTDO project.

Alberto Cantera, Freie Universität Berlin
Miguel Ángel Andrés-Toledo, University of Salamanca

"Visualization and Search of Zoroastrian Manuscripts: The Avestan Digital Archive and the Corpus Avesticum Berolinense"
Zoroastrian manuscripts are not always easily accessible to scholars, and even less to the general public. The Avestan Digital Archive (ADA) project aimed to solve this problem by digitizing, indexing and making online available and searchable Zoroastrian manuscripts preserved in the most important collections all over the world. In this presentation, I will show how the researchers of the ADA project transferred the digital image of these Zoroastrian manuscripts to an open-access online searchable database, in which paratextual information and different display modes are combined for visualization.

Adam Anderson, UC Berkeley
"**Data-Driven Modeling: Building Data Models from Ancient 'Archives'"**

Collections of cuneiform tablets, including the Old Assyrian private archives and the Ur III royal / admin. archives, serve as prime examples of artifacts that were initially organized into archives in antiquity, but through looting and early excavation, the archival organization of the artifacts has been lost. This talk will describe the best practices and methods for building data-driven models, which allow the texts of an archive to speak for themselves, rather than through theoretical interpretive frameworks. The resulting workflows have been shown to be able to restore groups of texts to their respective archives, and make each text more accessible for querying within databases and statistical analysis using data models. Results are visualized in a series of figures and interactive network graphs which are available for scholarly engagement on GitHub (https://github.com/admndrsn/).

David Calabro, Hill Museum & Manuscript Library
"Digital Etymology on Screen: Some Basic Principles"

This paper outlines basic principles for the application and display of the results of data analysis in a digital etymological lexicon. I will discuss principles to address four issues: (1) audience and the choice of individual languages versus language families as alternative entry points, (2) audience and the use of transliteration versus native script, (3) etymological path versus range of cognates, and (4) the representation of responsibility through alternative interpretations and bibliography. I will provide examples for this discussion from an Afroasiatic etymological database that is currently a work in progress.

Matthieu Husson, CNRS, SYRTE-Observatoire de Paris-PSL
Ségolène Albouy, CNRS, SYRTE-Observatoire de Paris-PSL
"Analysing and editing medieval astronomical tables: metadata and front-end in DISHAS"

The Digital Information System for the History of Astral Sciences (DISHAS) is an international enterprise aiming at providing state of the art Digital Humanity tools for the

edition and analysis of the main types of content encountered by historians of the medieval astral sciences: tables, diagrams and texts. Since 2017, and as a first step, the project focuses on numerical tables. As a vehicle of astronomical and mathematical information in manuscript form, numerical tables have special features (e.g. grid format, numerical content, specific mode of transmission…). This led us to define our data structure and interfaces (admin and users) in creative ways allowing historical analysis to rely on these features to foster new research on these too often neglected sources, to weave together resources from different traditions, and to give a wide audience access to research data in a user-friendly manner.

Session 4: The Outside World - Participation and Cooperation

Peter Der Manuelian, Harvard University
"The Giza Project at Harvard University: Can public outreach and scholarly access co-exist?"

For over twenty years, the Giza Project (http://giza.fas.harvard.edu) has been assembling the largest dataset of archaeological documentation for one of the world's most important sites, the Giza Pyramids and surrounding cemeteries just west of modern Cairo. The carved and painted inscriptional material on Giza tomb walls and objects forms a critical aspect of this documentation, contributing to all facets of Egyptological research. This talk summarizes some of the Project's digitization efforts, from intelligent online databases to immersive visualization with 3D modeling. Over the years the Project's focus has shifted from pure scholarship to also include a wider public outreach, as well as some crowd-sourcing experiments with students in Harvard classes. Some hits and misses, pros and cons, and possible future directions for accessing Giza textual data will be described.

Marina Rustow, Princeton University
"The Princeton Geniza Project: Crowdsourcing and Interoperability"

The Princeton Geniza Project has not relied on crowdsourced information in its 35-year history. But we've dipped a toe into crowdsourcing, or maybe two toes: we advised the Zooniverse Scribes of the Cairo Geniza website, a crowdsourcing/citizen science platform that yielded (and is still yielding) mountains of data about geniza fragments, some usable, some less so; and we regularly vet aggregated handlist descriptions from the Friedberg Genizah Project, high-level crowdsourcing the results of which nonetheless require considerable editorial sifting. The first half of this paper will reflect on crowdsourcing as a way of generating information about a large unedited or uncatalogued corpus.

The second half will discuss interoperability, its promises (asking better questions), limits (if scribal traditions differed, databases will differ perforce) and possibilities (building simple tools to search across corpora).