ABSTRACTS

FATHER COLUMBA STEWART, HILL MUSEUM & MANUSCRIPT LIBRARY

The Digital Dawn of Comparative Manuscript Studies: How Authority Control Has Become the Critical Link

Before 2003, HMML worked in Europe and Ethiopia and had a rudimentary online catalog based on standards developed for western medieval manuscripts. As we made the transition to a fully digital preservation process and expanded into several new eastern Christian manuscript cultures and then into Islamic traditions, the limitations of our previous cataloging practices became obvious, especially as we were encountering so many texts that moved across linguistic and even religious boundaries. Although it was clear that we needed to be much stricter about authority control, many of the traditions we work with lack internationally recognized name authorities. There were also cases of multiple authorities because catalogers did not recognize a name that had been transliterated or adapted when a text was translated into a new language. Given that our work provides the basis for truly comparative manuscript studies, attention to these challenges became imperative. This led to the HMML Authority File (HAF) project, where both new authorities and established forms could be included in a searchable database that would be used by HMML catalogers and freely shared with projects around the world. This presentation will provide the broader context and then tell the story of this evolution at HMML.

JÜRGEN KETT — DEUTSCHE NATIONALBIBLIOTHEK

Overview of current developments in standardisation in the GND

In the web of data, the needs of the various areas of culture and science converge. Authority data as stable virtual bridges (consisting of persistently addressable authority data on actors, corporations, geographical areas, events, intellectual creations and subject terms) form a decisive ingredient for the backbone of a cross-domain knowledge graph. To meet this challenge, in 2017 the partners of the Integrated Authority File (GND) started a process of opening up and modernization. We aim to offer our growing community of cultural and scientific institutions in the D-A-CH area a reliable shared vocabulary for semantic linking of their data. This fundamental renovation is coordinated and managed by the Standardization Office (AfS) at the German National Library. In the lecture the basic concepts and the achieved status will be presented. In the context of the subsequent discussion, we will approach the special challenges and opportunities with regard to the indexing of manuscripts based on authority data.

CAROLIN SCHREIBER AND ROBERT GIEL — HANDSCHRIFTENPORTAL

Handschriftenportal: Types of authority data and their use

This paper will give an introduction to the new German national portal for Western manuscripts and its use of different types of authority data. The portal will reference the usual entitites provided by the GND such as persons, place names, or corporate bodies. A newly introduced and especially important entity concerns the physical objects, the manuscripts themselves, which are represented by so-called cultural object documents and their counterpart in the GND, the "written heritage documents" (Schriftdenkmäler). All authority data will be maintained and curated in an internal authority data module, where they can be augmented and exchanged with the GND or other repositories. In addition, subject-specific controlled vocabularies and thesauri will be developed, which will be based on data from the GND and other authority data providers. The technical basis for this module will be a graph database to ensure international connectivity via Linked Open Data. Intersections between the two major manuscripts projects are to be expected especially in the area of the thesauri for codicology and material analysis which therefore should be developed in close collaboration.

GERHARD MÜLLER — STAATSBIBLIOTHEK ZU BERLIN

Multilinguality in the GND

The Integrated Authority File (GND) is a network of relations between entities, including different types of entities, such as relations between persons (500), corporate bodies (510) or subject terms, e.g., professions, instruments, topics, or fields of study. Internationally aggregating services have long since supported the interconnection of metadata by clustering potentially identical entities. Through International collaboration and internationalisation of information supply, the question of multilinguality becomes increasingly relevant. Apart from a short description in English, Wikidata also offers descriptions in German and French. Subject terms in the GND are linked to the Library of Congress Subject Headings (LCSH) and the Répertoire d'autorité-matière encyclopédique et alphabétique unifié (RAMEAU). So far, however, terms that are used to identify and describe a person (such as profession) have only been available in one language, German or English. This presentation will contribute to discussing the further development of the GND through two scenarios that elucidate the potential for multilingual authority files.

COLINDA LINDERMANN AND CHRISTOPH RAUCH — PROJECT ORIENT-DIGITAL

Authority records in Qalamos: Practices and perspectives

Since the start of the project in 2020, "Orient-Digital" has established connections between the portal (Qalamos: Connecting Manuscript Traditions) and GND/LoC authority files by importing authority records for names (mainly authors, scribes, and owners) and linking them to our manuscript records. One of the challenges we are encountering in this process is the question of how to establish an efficient workflow for exchanging, adding, and improving data in the GND and whether to re-import GND records after having modified them locally. This is a question that pertains to names, but increasingly also to work titles: as we aim to map and eventually also visualize relations between authors and works as well as between works (master text, commentary, gloss, etc.), we start to think of a workflow for GND title records. The current situation is not yet satisfactory, and its improvement will require closer collaboration between libraries and research institutions. In this presentation, we will discuss Orient-Digital's current practice with regard to GND name records and then take the example of Ibn al-Ḥājib's *Kāfiyah* and its commentaries and glosses to discuss the role GND title records may have in the future.

WERNER SCHWARTZ — STAATS- UND UNIVERSITÄTSBIBLIOTHEK GÖTTINGEN

Allowing for multiple standard forms: The CERL Thesaurus

The Consortium of European Research Libraries (CERL) is focussing on historic printed books and manuscripts, the written heritage of Europe. Its HPB database is a pool of catalogue records from more than 60 research institutions. From the start it has been CERL's aim to solve the problem of standardisation of names in its main database. As an instrument helping to achieve this the CERL Thesaurus (CT) has been created. This presentation will describe the basic idea behind the CT, its creation and main technical features.

YASMIN FAGHIHI — CAMBRIDGE UNIVERSITY LIBRARY

The role of a coordinator of data and content

Aggregation of data requires coordination - the role of a coordinator is an on-going one in response to the use of the data over time. How is this reciprocal relationship between the creation of data and the use of data managed across a distributed network? We will draw on examples from established union catalogues such as FIHRIST and eCodices and also look at the impact of innovative approaches such as IIIF and ISMI. How is the fluctuating scene of data transformation and re-use being managed in sync by curators and data owners?

DAVID CALABRO, JOSHUA MUGLER, CATHERINE WALSH — HILL MUSEUM AND MANUSCRIPT LIBRARY

Titulus and HMML Authority File: Creating and sharing authorities from cross-cultural manuscript traditions at the Hill Museum & Manuscript Library

The Hill Museum & Manuscript Library (HMML) has been involved in manuscript preservation and description since 1965, first with microfilm in Western Europe and Ethiopia. In 2003, projects shifted ground to begin digital preservation of first Eastern Christian and then Islamic manuscripts across the Middle East, then ranging geographically and culturally as far as Western Africa, India, and Southeast Asia. Because of the duration and cultural breadth of HMML's work, the use and creation of authorities has centered around three central issues:
1) the shift from initial descriptions lacking any authority control to a purely internal system, and more recently to the development of the Titulus and HMML Authority File (HAF) databases to share authorities both internally and externally;
2) the need to create records that bridge the gap between Western forms of reference and other scholarly traditions, with texts translated and retranslated across linguistic and geographical boundaries; and
3) the difficulties inherent in representing authors and texts for which the scholarly infrastructure is less established, based on manuscripts where information is idiosyncratic and sometimes incomplete.
This presentation will provide an overview of HMML's recent authority projects and the working methods that inflect them, as well as presenting some of the challenges and solutions that have arisen throughout this work.

NATHAN GIBSON — BIBLIA ARABICA, LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN

Authority records at the frontiers of research: How can small projects incorporate large providers into their workflows?

Many small projects enthusiastically participate in the linked-open data ecosystem. Some incorporate permanent URIs from Virtual International Authority File (VIAF, viaf.org) or their national library into the infrastructure of the project. Some create their own URIs for entities, and then link them to these resources. Some even fiddle with RDF to make their link assertions friendly to machine processing.

Inevitably, they face a problem. Small projects usually exist for a very specialized purpose, such as conducting frontier research or preserving the cultural heritage of an otherwise overlooked community. By their very nature, they are involved in areas where secondary literature or archival metadata, let alone reference resources, are sparse. The lack of such sources makes it likely that many persons, works, places, and other entities in the research area have not yet been included in authority files like viaf.org or The Integrated Authority File (GND, dnb.de), or even in crowd-sourced databases like Wikidata (wikidata.org). Where such records do exist, the lack of resources may result in duplicate or conflated records. Thus, small projects may find themselves having to create their own authority records rather than benefiting from existing infrastructure.
Additionally, they face the problem that their size and funding does not allow them to devote many person-hours into converting their data into formats that large providers can ingest, nor into opening communication channels with multiple large institutions, however interested and approachable those institutions may be. This is not to mention that their staff may not be trained in archival standards.
Finally, there is the problem of mismatched priorities. Authority file providers are sometimes only interested in creators rather than, for example, owners mentioned in the marginalia of a manuscript. Wikidata, too, has a "notability" criterion. For research projects, in contrast, these may be the very people, works, or places they want to investigate.
In sum, (1) obscurity, (2) size, and (3) priorities all complicate the ability of small projects to incorporate large providers into their workflows. At the same time, it is precisely their frontier nature—as discoverers of new entities and as subject experts for messy data—that can make their data especially valuable to those providers.
While I have no all-encompassing solutions to these issues, I will reflect here on experiences representing some possible strategies to address them: notably, Syriaca.org's partnership with VIAF's "Scholars' Funnel" and the informally organized "Historical Middle East Data Alliance."

THEODORE BEERS AND KHOULOUD KHALFALLAH — ANONYMCLASSIC, FREIE UNIVERSITÄT BERLIN

Flexibility vs. interoperability in manuscript metadata: Reflections from the AnonymClassic project

As a precursor to the goal of constructing a synoptic digital edition of the Arabic versions of *Kalīla and Dimna*, the AnonymClassic project at the Freie Universität Berlin has so far gathered copies of around one hundred manuscripts of the work. These manuscripts are held by a diverse set of libraries across Europe, the Arab world, and beyond, and the catalogue data available for them varies accordingly. It has been necessary for the AnonymClassic team to develop its own system for fleshing out manuscript metadata and recording it in a consistent format. This is included as part of the new digital editing platform that is being built for the project.
The natural priority of researchers in our team is to describe the manuscripts as comprehensively as possible. We make note of features as basic as colophon dates, scribes' names, and handwriting styles, as well as more detail-oriented points such as readers' notes and other marginalia, the inclusion of illustrations and the specific motifs chosen, and the presence of "non-standard" orthography and syntax. While this type of cataloguing is useful and suits the needs of the project, there is a question of ensuring that our idiosyncratic body of metadata will be usable by other scholars and institutions in the future. I would like to speak about some of the efforts that we are making in this direction. Our aim is for the project to leave behind (among other things) an easily consumable database of key details about *Kalīla and Dimna* manuscripts.

MOHAMED MASLOUH — MMP-II, GHENT UNIVERSITY

MARC21, RDF, VIAF, and fifteenth-century Arabic historiography

My presentation will briefly introduce our ERC-Consolidator Grant Project "The Mamlukisation of the Mamluk Sultanate II: Historiography, Political Order and State Formation in 15th-Century Egypt and Syria (MMSII) (2017-2021)" (PI: Prof. Jo van Steenbergen- University of Ghent, Belgium). After a general introduction of the project's goals and status/achievement, it will move to discuss one of the key aspects of the project that is "creating a reference database of metadata for the production, reproduction, and consumption of all Arabic historiographical texts from the period 1410-1470".
The presentation will detail our experience in creating our databases with focusing on how we standardized our metadata and authority files, and how we facilitated the exchange and the manipulation of our data through our Islamic History Open Data Platform (IHODP). The main three key points will be:

1) Introduction of IHODP, how it is used to integrate and connect different projects, and how it can facilitate further collaboration between other projects in the field.
2) The transition from a MARC21 environment to the current RDF platform.
3) Our choices to adopt VIAF and OpenITI's CTS URNs as stabilized data links.

MAXIM ROMANOV — UNIVERSITÄT WIEN

Being Pragmatic: Data and Authority Files in the OpenITI corpus

Our field lags behind others. This is neither inherently bad nor good; this is how things are. The digital turn is not an exception: fields like the Classics already had long-established academically curated digital libraries when we made our first steps into the world of the digital. And when we did, there already were standards, formats, and frameworks that we had to face as de-facto best practices; we had to fit into the digital space that had been already preformatted by others, even if it did not fit our needs. Despite these negative aspects, joining the digital turn late also had its advantages. For example, while TEI XML has become the de-facto standard for textual data, we do not have a critical mass of our textual data in TEI XML, so we do not have to stick to it if it does not work sufficiently well for our purposes. In fact, our field does not have any kind of high-quality digital data that could be efficiently reused in different projects and whose critical mass could effectively start dictating its "standards" to the rest of the field. What this means in practical terms is that every digital project that deals with any things Islamicate has an opportunity to experiment with existing formats, standards, and technologies in order to figure out the most efficient and suitable solution. Despite seeming idiosyncrasies, the OpenITI adheres to a series of principles that ensure its longevity and compatibility with other projects (even if some algorithmic transformations might be necessary). These principles include: 1) machine readability vs. adherence to well-established but bulky standards; 2) automated procedures whenever possible vs. manual work; 3) federated open data vs centralized closed data. Thus, following these principles, all data is openly available in a federated manner through GitHub, where anyone can suggest changes (through pull requests). All texts and relevant metadata files (authority files) are organized following simplified principles of canonical text services (CTS). Texts are formatted into OpenITI mARkdown that provides the required minimum of structural annotation (the work is still in progress). Metadata files—on authors, books, and editions—are stored using expandable YML format, where "Arabic" values are keyed in using the OpenITI betacode that allows one to enter values only once and have them automatically converted into Arabic script as well as any desirable transliteration system. The paper will illustrate these main aspects of "being pragmatic" within the OpenITI project.

LORENZ NIGST — KITAB PROJECT

Categories not made for their content

Premodern Islamicate texts hold ready numerous challenges with regard to authority records as they often come with pieces of information that (currently) only with difficulty can be accommodated within these records. If we narrow these challenges down, they seem to be linked with two major issues: on the one hand, many of them relate to specific premodern scholarly practices that do not follow the logic that one author wrote one book (such as *takhrīj, tartīb,* etc.); on the other hand, the specific name components in Arabic have to be fitted in the "first name" "last name" scheme which is alien to them.
In this presentation, we will take a look at concrete examples for both problems and we will suggest that while some loss of information perhaps is inevitable, it nonetheless seems worthwhile highlighting the problem, making a collective effort at documenting respective cases, and trying to make items of metadata specific of the Islamicate textual tradition with which we are dealing fully machine-readable.

TILL GRALLERT — OPEN ARABIC PERIODICAL EDITIONS, PROJECT JARĀʾID AND SIHAFA

Who is ʿAlī Pasha? Modelling personal names from the late Ottoman Eastern Mediterranean (c. 1850–1920)

My work on the discursive space of the predominantly Arabic-speaking periodical press of the late Ottoman Eastern Mediterranean aims at establishing intellectual networks of people, works, and places. It therefore relies on extensive authority files to identify and disambiguate such entities across the growing corpus of currently 7 journals with more than 600 issues and 7 million words.
This short presentation will focus on modelling personal names with TEI XML. Such modelling is needed to address two related aspects for the disambiguation of entity references.
The fluidity and heterogeneity of historical practices. During the period of interest to my research, people made use of at least three reference systems for personal names: the Islamicate/Arabic tradition, Ottoman practices, and the "modern" western style. They also used multiple alphabets (Arabic, Latin, Hebrew, Armenian, Syriac …) and transcriptions between all of those. This, of course, is in addition to nicknames, pseudonyms, etc. If we want to track people across texts from various domains, we must be able to translate between alphabets/transcriptions and reference systems.
The conceptual rigidity of existing online-authority files. While some rigidity is certainly necessary for claiming "authority", narrow modelling of entities based on the cultural preferences of the twentieth-century Global North represents epistemic violence that we have to address, if we want to link our datasets to these authorities (which we obviously do). One way is to automatically "guess" a normalisation of fluid historical practices into standard patterns, such as "surname, forename(s)".
With reference to a generic *علي باشا*, we might be lost and unable to establish any certainty without extensive study of the context in which he is mentioned. Yet, we might be able to link a certain *علي باشا ابن محمد الجزائري* to the entry for "ʿAlī ibn Muḥammad ar-Rašīdī" in the DNB—even though the latter does not contain the name in Arabic script.

JONATHAN ROBKER — THE DIGITAL ORIENTALIST, UNIVERSITÄT MÜNSTER

Digitizing biblical manuscripts: Best practices and desiderata from extant exemplars

Digitized biblical manuscripts provide amazing opportunities, but also present distinct challenges to biblical studies. On the positive side, digital access to biblical manuscripts scattered throughout the world has increased substantially and continues to grow as more institutions housing manuscripts make them available online. That affords more users than ever before ever greater access to an increasing number of manuscripts from the trustworthy institutions housing the manuscripts. However, the caveat of such decentralized approaches simultaneously leads to the negative knock-on effect of a lack of normalisation: disparate forms of presentation, inconsistency in the types of data and metadata made available from each

institution, and even differences in the data from manuscripts within an institution. The range extends from projects that present only digitized photographic facsimiles of individual folios to online presentations of digitized photographic and typographical facsimiles with some hyperlink indexing and even search functionality. For this presentation, I will offer examples of some of the current best practices in the field of digitized biblical manuscripts, while also noting places where more reflection, organization, and consistency are commendable. I will conclude by offering some specific desiderata for normalising the data from the digitization of biblical texts in order to provide it as the most usable resource for scholars of the field.

DANIEL KINITZ, THOMAS EFER, TARIQ YOUSEF — BIBLIOTHECA ARABICA, SÄCHSISCHE AKADEMIE DER WISSENSCHAFTEN ZU LEIPZIG

Pragmatic approaches to authority control of premodern Arabic personal names and book titles

In our bio-bibliographic collection, we are faced with thousands of premodern, non-standardised Arabic names. We employ alignment algorithms and visualisation for this dataset to establish semi-automatic authority control. The data is taken from several Arabic and Persian manuscript catalogues and al-Ziriklī's bio-bibliographical dictionary *al-Aʿlām* as a reference work. Since these sources are very heterogeneous, special emphasis will be placed on preprocessing, using combined prosopographical and technical domain knowledge. The results can be implemented as a flexible recommendation system in databases that process great quantities of premodern names from the Islamicate world, such as Bibliotheca Arabica's research platform.

MICHAEL BECKER AND SARAH WINKELMANN — UNIVERSITÄTSRECHENZENTRUM LEIPZIG

Handling large quantities of person data: Opportunities and challenges

In the project Orient Digital, besides manuscript data, we have a great amount of authority data representing persons including both imported data from the GND catalogue and local data. Orient Digital incorporates several different manuscript collections, each with varying numbers of person data. This results in new opportunities for working with manuscripts, but we also face various challenges concerning authority data. On the one hand, a unified person index covering multiple collections is possible. On the other hand, we cannot eliminate the risk of person duplicates which were not uncovered during data import. In our presentation, we show a method and a software tool for identifying person duplicates based on person names and including additional metadata. We will present necessary preparatory work and data cleaning activities as well as preliminary results from several test runs. With the help of the tool, we were able to identify hundreds of possible person duplicates. Presenting possible duplicates and a probability, editors can unify clear duplicates while focusing their work on unclear edge cases. We discuss current results and provide an outlook on possible other contexts for using the tool.

HUW JONES — CAMBRIDGE UNIVERSITY

How reusable is your data?

What are the drivers for and the obstacles to the reuse of data? In discussing this question. I will draw on examples gained from working with Fihrist and with Cambridge Digital Library. I will talk about reuse by researchers and also reuse in systems, both those local to Cambridge and beyond, touching on areas such as formats, identifiers, vocabularies, licensing and interfaces.

CORNELIS VAN LIT — THE DIGITAL ORIENTALIST, UTRECHT UNIVERSITY

Create Locally, Share Globally: Customizing automation for building authority records out of existing catalogs

Digitization is only worth as much as we impose sensible orders on the vast quantity of data it produces. For Islamic manuscripts, we have seen two diverging paths. On one hand, institutions have digitized indiscriminately and at most have used their existing catalogs to expose their holdings through a website. On the other hand, there are several projects in an advanced stage that offer digital linked data and authority files relevant to islamic cultural heritage. Such initiatives include al-Thurayyā Gazetteer for place names, Onomasticon Arabicum for personal names, the British union catalog Fihrist with its freely offered XML-based catalog and authority files, and the closed-source Diamond-ILS cataloging system developed by IDEO with the FRBR-system in mind. Other projects loom at the horizon, such as the HMML Authority File project. How can we join those two diverging paths? Theoretical and technical examination will draw to the main message of this contribution: create locally, share globally. It seems not the time yet to come to an overarching solution. It is neither clear how that solution would look like and who would have custody over it. It is much more preferable to use external resources to enrich one's own resource than to have it dictate the form and structure. This comes at the penalty of having to stop and think how each of these resources could be plugged into one's own resource, but that is a small price to pay in order to have an automated solution for data amelioration. It further requires a tight integration within projects of domain experts and software engineers, projects in which the tedious work of manual data entry will become more and more rare.