Open Library of Humanities

The Dragoman Renaissance Research Project in Library/DH Linked Data Partnerships

E. Natalie Rothman, University of Toronto, CA, natalie.rothman@utoronto.ca

Kirsta Stapelfeldt, University of Toronto, CA, kirsta.stapelfeldt@utoronto.ca

The Dragomans Renaissance Research Platform represents a long-standing collaboration between the authors. The project explores the role of dragomans (diplomatic interpreter-translators) in mediating relations between the Ottoman Empire and its European neighbours from ca. 1550 to ca. 1730. The project leverages surrogates and transcriptions from multiple archival and secondary sources to produce a structured data set that can be queried using SPARQL. The Dragomans use case illuminates the potential of library-partnered DH projects, like other unique collections, to become machine-readable research data sets made available through generalized, sustainable architectures that also suit researcher needs for UI-driven query and display.

The Dragoman Renaissance Research Project (Dragomans, 2024) represents a long-standing collaboration between a Digital Humanities Research team and the library's Digital Scholarship unit at the University of Toronto, Scarborough Library. The project employs surrogates derived from various archival and secondary sources to investigate the pivotal role of dragomans (diplomatic interpreter-translators) in facilitating diplomatic interactions between the Ottoman Empire and Venice (and, to a lesser extent, other European counterparts) during the period spanning approximately 1550 to 1730. The project seeks to illuminate the trans-imperial dynamics inherent in Mediterranean diplomatic chancery production by examining the intricate interplay between Ottoman and Venetian genres, textual artifacts, practices, and practitioners active in early modern Istanbul. The Dragoman Renaissance Research Platform is the technical infrastructure of the project; a web application and repository that facilitates research into the personal and professional trajectories and textual and visual practices of dragomans employed by the Venetian bailate (permanent residency) in the Ottoman capital and beyond. Our project aims to investigate how canonical genres of diplomacy (such as diplomatic reports) and statecraft more broadly (including decrees, petitions, memoranda, and letterbooks), did not merely evolve from singular, uninterrupted textual and evidential traditions—whether Ottoman or Italianate. Instead, we seek to understand how these genres emerged within specific circulatory regimes that connected various practitioners, including dragomans, secretaries, and scribes. Each of these practitioners embodied diverse modalities of knowledge production and belonged to distinct rhetorical communities, with unique positionality and institutional affiliations.

At the time of writing, The Dragoman Renaissance Research platform has curated a wealth of resources, including digital surrogates of archival materials from multiple archives and special collections, such as the Venetian State Archives and the Ottoman Prime Minister's Archives, among others. The project makes available transcriptions, transliterations, and English summaries of Ottoman and Italian-language archival records. The project also provides a bio-bibliographical dataset, which chronicles the lives and careers of hundreds of dragomans and other diplomatic personnel. This bibliographic dataset enables researchers to trace dragomans' kinship, financial, and professional networks, as well as their written output and artistic patronage. Site visitors can also find metadata and digital representations of 57 figures that are cited in Rothman's open access monograph, *The Dragoman Renaissance: Diplomatic Interpreters and the Routes of Orientalism* (2021). The list of figures contains portraits, oil paintings, genealogical charts, manuscript codices, and interactive data visualizations that delve into dragomans' kinship networks. The project also includes data on 84

bi- and multilingual Ottoman dictionaries and grammars in Latinate languages up to 1730 and provides citations for these works and for an extensive body of secondary literature on early modern Ottoman lexicography. Further enriching the collection are citations for 55 works authored, compiled, or translated by early modern dragomans, offering insights into their linguistic and cultural influence. Finally, the platform hosts a collection of multimedia presentations by Rothman and other team members since 2014. These presentations span a variety of topics related to the Dragoman Renaissance project and are tailored to both academic and general audiences.

Beyond offering a repository of archival surrogates and rich data sets available for researchers to pursue, download, and query, the project also aims to facilitate a range of ongoing research endeavors concerning translation and chancery practices. To that end, it focuses on a series of some 40 bound registers known as *Carte turche*, or Turkish Charters. These registers feature roughly 2,000 copies of sultanic orders and other Ottoman official records which were matched with facing contemporaneous Italian translations produced, and often signed, by specific bailate dragomans. These pairs of copies and translations are organized chronologically, spanning the period 1589–1785, with some notable gaps, and constitute the backbone of the bailate archives.

Why Linked Data?

Using computational approaches to elucidate multiple entanglements that shaped this part of the early modern world has required encoding a complex, uneven, and multilingual documentary record. For digital humanists, references to 'encoding' may evoke the Text Encoding Initiative (TEI), in use since the 1980s as a way of making texts machine readable by representing them in an XML-based digital form. However, encoding in the context of the Dragoman Renaissance Project has sought to elevate historical figures and institutions in the data model to the same conceptual level as the documents that they wrote and in which they are represented. Linked data appeals precisely because of the ability to atomize and aggregate knowledge across multiple sources, evoking the people and places represented in the dataset with greater lucidity and lending them a primacy away from the source texts that describe them. Linked data is particularly suited to what are called 'Open World' assumptions in knowledge base systems, based on the belief that no single observer has complete or absolute knowledge, which allows for a broader and more pluralistic approach to the production of a queryable dataset (Keet, 2013; CIDOC CRM Special Interest Group, 2024). We believe an open world approach is better suited to the open-ended questions that animate the project. Information entities can preserve their ambiguity more easily, and collect other traces, such as differing time systems, systems of pagination, and

archival organizational principles (sometimes for the same document). Linked data principles allow for machine-readable statements about the world that are more open to ambiguity and partial evidence. As we learn more, we can widen the set of classes, attributes, relationships, and corresponding vocabulary. Linked data, designed to be used outside of the confines of a single system (Horster, 2024), is more porous to other sources of knowledge and available to multiple technologies—in other words, a linked data approach can be more generally useful to the wider world of information about the early modern Mediterranean, and can facilitate wider and potentially more generative research communities without requiring as much bespoke programming and maintenance; costs that are ultimately untenable for any digital humanities project. There is growing evidence of interest in the use of linked data approaches in the cultural heritage space (Giovannetti, 2022). Project data has a strong reuse value in other projects and allows for iterative understanding of the subjects being described and relationships that may be uncertain or even contradictory. Linked data is not a panacea for the multiple challenges that occur in collaboration and the sharing of research data: it cannot solve the tension between a scholar's individual goals and the compromise (and additional labour) required to create assets with greater reuse potential. However, it is a gesture towards a future in which negotiations between the two become more possible.

A Dragoman Ontology

Ontology is a formal representation of the concepts and relationships within a particular domain through a set of classes, properties, and relationships. But ontology is also the philosophical study of existing, and the ontology represented in The Dragoman Renaissance postulates a theory of being that is fundamentally and foundationally social. In the project, our focus is on an object-oriented ontology, where archival records, the documents they contain, and, especially, the textual practices that shape them, are at the centre. This choice is made in part to decentre a still pervasive idealist approach to the history of diplomacy that focuses on what high-level diplomats said/ wrote (rather than on the material forms in which diplomatic practices were grounded, and the myriad of institutions and often-invisible practitioners that enabled them). It is also an analytic to allow us to identify long-term patterns in the corpus of translated records that forms the heart of the project, and to correlate such patterns with, e.g., temporal shifts, the sociological profiles of different dragomans, the genres and addressees they were intended for, etc. While focusing on events would be possible, the Dragomans model shifts the focus from events and why they happen to practices and their evolution. If a document is an imprint of an event, it is a tangle of processes

and actors that can be untangled and organized to better reflect how communicative (documentary, textual, diplomatic) practices evolved and shaped (and were shaped by) the social conditions of the early modern Mediterranean.

Over the years, the project has generated multiple vocabularies and relationship types that give equal primacy to the concept of a person, a transcription (document) either produced by, addressed to, or referring to that person, and the material archival record containing it (a fascicle, a bundle, a register, etc.). **Figure 1** presents a high-level view of our current data model, its data entities, relationship entities, and controlled vocabularies.
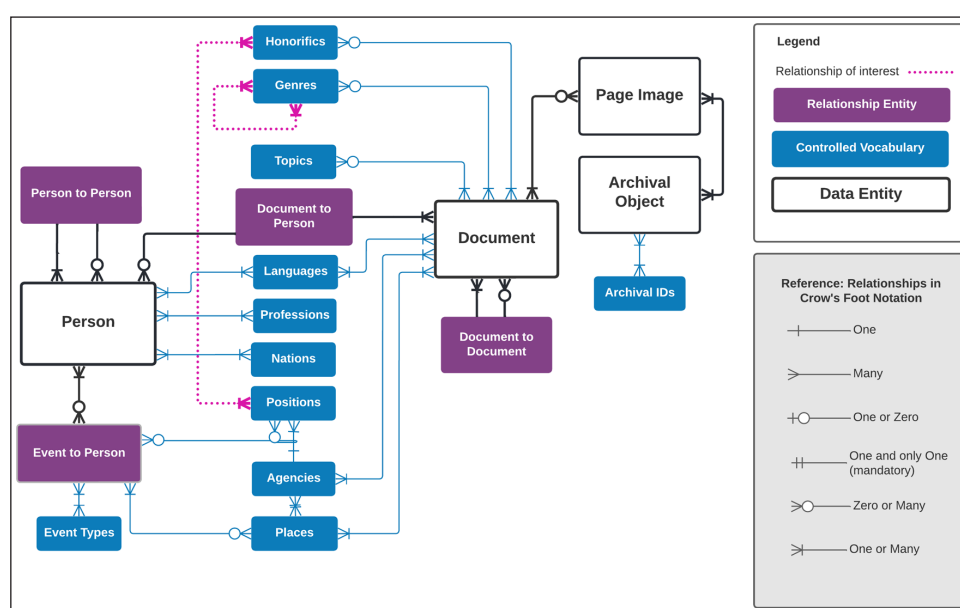

Figure 1: Dragomans Data Model. This model reflects the relationships, entities, and controlled vocabularies represented in the project as well as their interrelationships and was first published in Rothman et al., 2023. This diagram is continually evolving to represent our highest-level understandings of the project.

Persons, documents, archival objects, and their corresponding page images (depicted in white) are interconnected through relationships (shown in purple) and are organized using controlled vocabularies (represented in blue). This modeling allows us to visualize and infer diverse types of relationships depending on the project's evolving research questions. This granularity extends to the individual entries of the controlled vocabularies. Control is light in this context, as we limit the level of orthographic standardization performed during encoding and cataloguing. As discoveries are made about the variance and use of terms, these can be further described in the vocabulary record, itself a fielded entity capable of being interrelated with other terms.

Most recently, the project has been focused on the undetermined relationship between Ottoman and Venetian genre terms, exploring connections organically through cycles of cataloguing and querying. **Figure 2** outlines how a graph of uses may emerge from this work, while leaving for further investigation the critical questions of temporal, regional, and institutional variations in terminological usage.
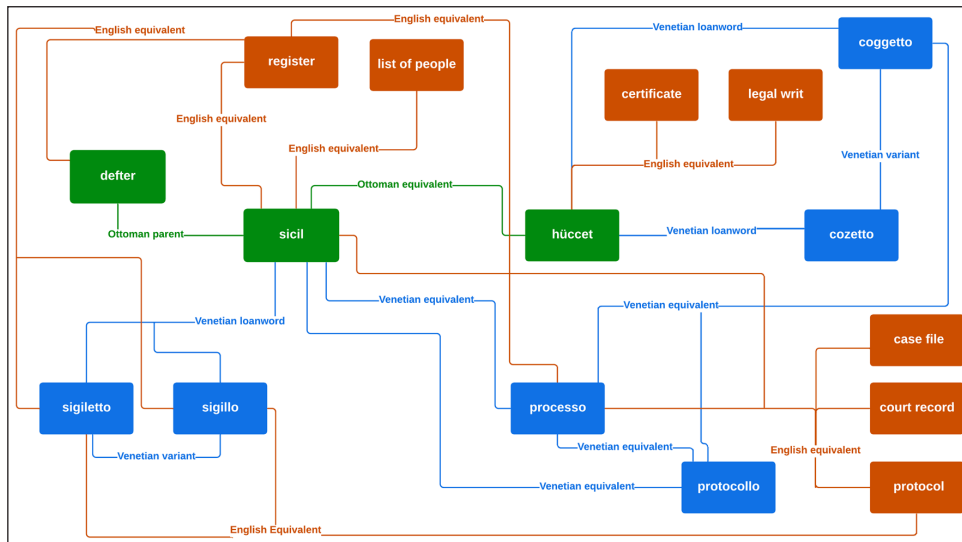

Figure 2: Representing 'sicil'. First published in Rothman et al., 2023, this diagram envisions the possible interrelationships between genre terms across the corpus. These types of creative exercises guide the development of data enrichment as well as queries in the project.

The methodology for organizing and encoding is documented in diagrams such as these as well as a 'data dictionary'—an ever-developing manual that is collaboratively maintained by team researchers, and versioned using Google Drive's native feature. This manual supports uniformity in data organization and expresses the rationale behind the determinations made by the research team.

This data-driven approach allows for tracking relationships over time, space, genres, and institutional sites of knowledge production and is supported by the project's technical infrastructure.

Technical Infrastructure: Islandora

The first formal documents related to the Dragomans ontology were published seven years ago, emerging in a different technical environment, and designed to be utilized in a different technical infrastructure. Just as ongoing research has shaped and reshaped the project's ontology and increased its volume of structured information, the project's technical infrastructure (the software and systems that facilitate the

creation, structuring, and publication of digital expressions of the corpus) and data modeling (the understood interrelationship between information concepts, such as people and genres, documents and artifacts) have accordingly undergone substantial and continual evolution. First modeled in what is now called 'Islandora Legacy', the platform is currently supported by a version of 'Modern Islandora' which has both stronger native linked-data capabilities as well as a closer relationship with the highly flexible, open-source, enterprise application software Drupal. Drupal, in addition to being a robust open project for which one can always hire developers, is used by the majority of top post-secondary institutions in North America (Joseph, 2024). Islandora advertises itself as an

> extensible, modular, open-source digital repository ecosystem focused on collaborative authorship, management, display, and preservation of digital content at scale. Islandora adheres to widely adopted best practices and open standards and frameworks used in information practice (Islandora, 2024).

Each information object in Islandora (vocabulary term, relationship, or data 'entity') can be modeled in RDF (a standard method of representing linked data) using the RDF module. The process involves editing a YAML file to map fields in entities to RDF predicates. Entitles are exposed as RDF through a representation in JSON-LD that can be consumed by other applications, notably (in the project's case) Blazegraph, which is a widely adopted, open technology used by groups such as Wiki Data. The platform facilitates open access to the linked data serialization of the project, though the latest iteration is not yet exposing an open query endpoint. A key learning from the last eight years has been that of all the ways the linked data can be accessed, humanities scholars are not overeager to write SPARQL queries. Multiple other serializations are available in Islandora, including an OAI-PMH endpoint serving MODS and DC metadata, Archival Information Packages, and an Apache Solr index. The library-friendly features of the platform naturalize the research data from the project in a wider ecosystem of information practice. Similarly, the project benefits from adherence to emerging standards for data display, including IIIF manifests—an increasingly significant exchange format for cultural data. This approach is designed to bring longevity and greater reuse potential to the project's data.

Islandora also supports direct collaborative editing and creation of data entities via a GUI, as well as workflows associated with content creation. Discussion is ongoing of the possibility of using some of these features to support transcription of documents related to the project via a curated crowdsourcing workflow that would further ease the burden of contributing to the project for a multilingual audience. Data can also

be exported into filtered .csv formats and then updated en masse using spreadsheet software and a tool called Islandora Workbench. The use of Islandora, which is a system centrally managed by the University of Toronto Scarborough Library's Digital Scholarship Unit, has been key to the project's ongoing sustainability. One benefit of Islandora is that if the data is appropriately modelled, we can serialize and reserialize based on different complementary ontologies. CidocCRM appeals because of its breadth and depth, as well as the large number of collaborative/interoperable models to which we might map our classes and predicates. However, any sufficiently broad and extensible ontology may suit depending on the research questions and conclusions of the moment.

Institutional Symbiotics and Library Partnerships

The University of Toronto Scarborough Library's Digital Scholarship Unit is staffed by two developers, a systems administrator, and a dedicated collections coordinator. This team, with guidance and support from librarians and the broader open-source communities, develops Islandora as a 'deployment' based on the Islandora starter site, injecting institutionally specific configurations such as a custom theme. Changes made to this base 'deployment,' including custom modules, are versioned in a separate repository and the data is subject to backup as part of the institutional backup strategy, as well as the specific digital preservation workflows of the library. All aspects of this process are designed to facilitate scale in partnerships for scholarship and increase the sustainability of individual projects by maximizing cooperation across the institution. The unit's 'deployment', as well as the staffing, documentation, and workflows of the unit are designed to integrate with existing systems, reducing the burden of maintenance. Even the Dragoman Renaissance project itself maps on to accepted definitions for a 'special collection' from the library's perspective. The various descriptive fields used for the project extend a base descriptive model maintained and used in multiple other projects. The logic behind this approach is that while the research agenda is unique, there is much that can be generalized and reused between projects. Where custom features are needed for the Dragoman Renaissance project, they are developed in a generalized way to find other uses and avenues of support in the wider open-source community. See, for example, the Recogito annotation module (2024), which allows users in any Drupal site to facilitate annotation of HTML content—an important evolution for the project's next phase.

Our approach to application evolution requires both patience for a researcher who must wait out longer development cycles and some flexibility around working style, but also promises researchers some relief from the grind of grant cycles. Specifically, this approach has relieved us of the need to assign limited and time-bound resources

to the expensive pursuit of custom software development. Through a combination of library-acquired grants, institutional funding, and project-specific research grants (most recently: from the University of Toronto's work-study program and the Jackman Humanities Institute's Scholars-in-Residence program, as well as a multi-year Insight Grant from the Canadian Social Sciences and Humanities Research Council) we have funded a broader research team working within and beyond the library, involving dozens of undergraduate and graduate students, postdoctoral research associates, and occasional consultants and collaborators across disciplines and institutions. This process is rewarding for all involved and supports an overcoming of disciplinary solipsism and speciation, as well as a wide range of opportunities for mentorship, professionalization, and cohort building.

One result of our approach has been partnering with local co-op programs to create positions for computer science students. These students work with existing staff in the Digital Scholarship Unit to create and publish features for the Dragoman Renaissance project as part of a broader 'Emerging Professionals' program. This program also supports students from the University of Toronto's Faculty of Information, performing routine data maintenance and cleanup tasks that build their knowledge of specific standards. Undergraduate students often participate in special projects that use or extend the project's data and sources, developing research and digital literacy skills. Throughout the duration of the project, student research assistants have consistently contributed a diverse array of linguistic and paleographic proficiencies, from early modern Ottoman Turkish and Italian to Serbo-Croatian, in addition to a wide range of digital capabilities, which include scripting, mapping, and data analytics. Leveraging and enhancing students' diverse skillsets is also in keeping with strategic institutional goals related to experiential learning. The project's transdisciplinary teams, emerging out of a transdisciplinary partnership, are multilingual, multi-disciplinary, and multi-generational; a product of the unique institutional context of the University of Toronto Scarborough. The relative stability afforded by strong institutional backing has also allowed the project to develop a variety of collaborative sub-projects with subject-area experts (Ottomanists, translation studies scholars, historians of archives and knowledge production, data scientists, and open-source software developers) beyond the University of Toronto.

Frontiers — Artificial Intelligence

Considerable attention has been recently devoted to the potential impact of artificial intelligence (AI) in research. The Dragoman Renaissance project, thanks to its use of linked data, may benefit on two fronts: The first having to do with the processing of source data and the production of accurate data (a persistent bottleneck in recent

years). Artificial Intelligence, particularly machine learning, has rapidly advanced the capabilities of text recognition in surrogates of historical texts. Previous attempts to use Handwritten text recognition (HTR) on images of manuscripts have not been successful, but we continue to explore evolving models of text training and recognition as this would dramatically speed up the project's encoding. The second benefit is the potential for AI to produce visualizations and insights about the project precisely because of the linked data modeling and multiple serializations available through the platform. We are eager to discover what additional rewards there may be for taking a future-facing approach to data encoding in the Dragoman Renaissance project.

Conclusion

The Dragoman Renaissance Research project has emerged out of a unique partnership between the research team and the library, and uses a linked data approach to encode sources from multiple archives as well as secondary sources. Focusing on linked data has enriched the project in numerous ways, such as allowing inquiry to shape the structure of the current dataset, enabling the data to be reused in other contexts and facilitating multiple contributors. The Islandora-based technical infrastructure of the project reflects efforts by the University of Toronto Scarborough Library to produce infrastructures that are generalized and informed by researcher needs but also adhere to standards for information architecture and practice that facilitate longevity and visibility of information resources. With the institutional commitment made to the project, research grant funding can be used (in combination with the library's 'Emerging Professionals' program) to expand the pool of research assistants, drawn from a truly transdisciplinary body of students. Linked data represents a future-facing architecture for this expansive research project, and the time spent on this approach may bear additional fruit as a data set particularly legible to artificial intelligence algorithms.

References

CIDOC CRM Special Interest Group 2024 CIDOC conceptual reference model (Version 7.1.3) [Definition]. Retrieved from: https://www.cidoc-crm.org/Version/version-7.1.3 [Last Accessed Oct 4, 2024].

Dragomans 2024 The Dragomans Renaissance Research Platform. https://ark.digital.utsc.utoronto.ca/ark:61220/utsc73539. [Last Accessed Oct 4, 2024].

Giovannetti, F and **Tomasi, F** 2022 Linked Data from TEI (LIFT): A Teaching Tool for TEI to Linked Data Transformation. *Digital Humanities Quarterly*, 16(2). Retrieved from https://www.digitalhumanities.org/dhq/vol/16/2/000605/000605.html [Last Accessed Oct 4, 2024].

Islandora Foundation 2024 RDF Repository – Islandora. Retrieved from https://islandora.github.io/documentation/concepts/rdf/ [Last Accessed Oct 4, 2024].

Joseph, V 2024 *Drupal in Education: Data CMS Usage in World's Top 300 Universities. The Drop Times*. February 1. https://www.thedroptimes.com/37122/drupal-in-education-data-cms-usage-in-worlds-top-300-universities [Last Accessed Oct 4, 2024].

Keet, C M 2013 Open World Assumption. In: Dubitzky, W, Wolkenhauer, O, Cho, K H, Yokota, H (eds) *Encyclopedia of Systems Biology*. New York, NY: Springer. https://doi.org/10.1007/978-1-4419-9863-7_734

Recogito Integration 2024 Drupal.org. https://www.drupal.org/project/recogito_integration [Last Accessed Oct 4, 2024].

Rothman, E N 2021 *The Dragoman Renaissance: Diplomatic Interpreters and the Routes of Orientalism*. Ithaca, NY: Cornell University Press. https://doi.org/10.7298/fxrs-fn65

Rothman, E N, Stapelfeldt, K, Idil, E, McCarthy, V and **Karim, Q** 2023 Toward an Ontology of Trans-Imperial Ottoman Chancery Genres. *Journal of the Ottoman and Turkish Studies Association*, 9(2): 77–83. https://doi.org/10.2979/tur.2022.a902164