Review of Middle East Studies, 52(1), 153–159
© Middle East Studies Association of North America, Inc. 2018
DOI:10.1017/rms.2018.13

PEDAGOGICAL PERSPECTIVE

# Text Encoding Innocents Meet the *Egyptian Gazette*

Will Hanley
*Florida State University*

## Abstract

*This short piece describes an experimental course at Florida State University in which undergraduates digitize issues of a turn-of-the-century Egyptian newspaper. Beginning students of Middle East studies can benefit from exposure to raw primary sources. Learning to read and process this newspaper using digital methods such as character recognition and text encoding, they generate a repository of text that can be of value to scholars in the field. Training students to do this work certainly offers them valuable, transferable technical skills. The combination of technical and conceptual work in a collaborative, laboratory setting is not easily accomplished, however. From an area studies perspective, it can be challenging to train students to discern what material is of significance, and how they might proceed to analyze it. Nevertheless, a great deal of knowledge in our field is produced incrementally, provisionally, and collectively. We can use institutional moves to re-imagine teaching to involve even inexpert students in this process of knowledge production.*

**Keywords:** newspapers, encoding, open access, TEI-XML, collaboration, primary sources

I'm a bad swimmer. I learned the front crawl first, and it is my default stroke, but when I'm in open water and struggling to breathe, I can't keep my face in the water. I have to roll over and do the backstroke. The backstroke lets me breathe, but not swim straight. In Germany, I've been told, most swimmers learn the breaststroke first. This default stroke allows bad German swimmers to breathe with their faces out of water when things are going badly and still see where they're going. I wonder how things would go for me if I'd been taught a different stroke first.

I wonder the same thing about history. What if, instead of starting with a narrative of significant events, we taught students first of all to wade through unprocessed primary sources, separating the wheat from the chaff? Would their work seem more like the work we do as professional scholars? Could beginners do such work? Might some of them enjoy it?

Since 2016 I have been leading a "lab" of undergraduates digitizing the *Egyptian Gazette*, an Alexandrian newspaper from the turn of the twentieth century. This semester, we will close in on 700 digitized issues. My course

https://doi.org/10.1017/rms.2018.13 Published online by Cambridge University Press

provides students with technical skills, including text encoding, XML coding, and data visualization. It also teaches them that knowledge production is often a grind, unfolding in uneven increments and small iterations. I ask the students to do this labor in collaboration with me and with each other, with the purpose of building a lasting resource open to future scholars. I aim to link this technical work to the conceptual work of understanding Middle East history. This short essay presents my approach and its successes and failures thus far.

## Opportunity

The curriculum tempests of the twenty-first-century public university throw up driftwood that can be repurposed. Where I teach, a well-intentioned reform a few years back introduced a set of “e-series” classes. It is almost inevitable that these initiatives disappoint to some degree. The carefully enumerated learning outcomes turn out to be pro-forma rather than opportunities for rethinking; the class size turns out to be fifty rather than fifteen; funding for teaching assistants disappears. For scholars of the Middle East, however, it is also possible to see that (with all their warts) these curricular cycles are opportunities to expand our educational mission and to try new approaches to our job.

I took the opportunity of the curricular reform to answer some of my own needs and interests by designing a course digitizing the ubiquitous English-language daily newspaper of British-occupied Egypt. I latched onto the “scholarship in practice” course-type, which is meant to engage students in the creation of scholarship. After years of teaching lecture courses on the Middle East, I had become disenchanted by my own resources and my ability to reach the greatest part of my students. Humanities and social science majors, heritage learners, veterans or active duty military, believers, and especially the curious often engaged readily with what I had to offer. I wanted to do a better job of reaching the needs of students who weren’t particularly interested in the Middle East before they enrolled, and were unlikely to be afterwards. To do this, I wanted to offer them more than just Middle Eastern history. I needed to offer them a variety of kinds of learning, set in a Middle Eastern context of course, that would allow them to find their own style.

## Structure of the Project

Here, in a nutshell, is how I run the class. (See the syllabus for full details). Each student is assigned a week of the newspaper, which means six issues, or about 40 broadsheet pages of about 2000 words each. She uses a microfilm reader to make high-quality images of each page, then begin the arduous process of converting the image into text.

https://doi.org/10.1017/rms.2018.13 Published online by Cambridge University Press

For about half of each issue, this means using optical character recognition (OCR) software. This software automatically produces text with accuracy rates in the high 90%s (so long as the students have made good images). The students then read and correct this text. The next stage is encoding the text. The students use XML tags to indicate its structures: where the headline is, where articles begin and end, where page, paragraph, and column breaks fall, and so on. To do this work, I teach them the widely-adopted Text Encoding Initiative (TEI) standard. I've found that the standard is not as well developed for newspapers as it is for other kinds of documents, but I am fortunate to have the support of a TEI specialist in our library (Sarah Stanley).

The rest of the paper consists of advertisements and financial tables. I give my students templates for these elements (currently there are about 400 of them). They copy and paste these templates, updating the contents to reflect the particularities of their own issues of the paper. By completing these tasks, they are mimicking the work of compositors of the paper a century ago, who also used boilerplates to do their work. Transcribing figures into templates can be tedious, though some report finding the repetition "relaxing."

As a final project, each student analyzes a "serial question" over the breadth of the newspaper. For instance, she might look at concert programs from the Alexandria bandstand, which appear almost every week. Using the XPath query language, she can extract the name of every piece and every composer (providing that it has been properly encoded), then represent this information in a data visualization of some sort and write a 1500-word discussion of the question in its historical context.

I grade this course on a completion basis. I spell out the requirements of every part of the encoding work; when each portion is complete, I award full marks. The result is that students have an incentive to complete each phase of their work, and those who do receive a high grade in the course. I run the course through a public website, and aim to offer clear and complete instructions for every part of the project. Documentation is not a great strength for most historians, and I am trying to make this project methodologically transparent.

Part of transparency is admitting incompleteness. For instance, I encourage students to tag persons, places, and events that come up in the text, but this work has only begun, and is inconsistent from week to week and issue to issue. This incompleteness is an essential part of what I want students to understand about how history is produced. Research is an imperfect and tentative endeavor, and scholars' explanations of the past are mutable assertions rather than objective facts. I encourage them to see their own

https://doi.org/10.1017/rms.2018.13 Published online by Cambridge University Press

work as "scholarly and subjective" (borrowing a phrase from Sarah Ketchley). It becomes clear to them, over the course of the semester, that their work will never be complete. Together, they come to see our knowledge about the newspaper itself as iterative rather than definitive. They experience frustration with tedium and doubts about relevance. This direct experience analyzing primary sources, I believe, tells some truth about how history is made.

## Collaboration, Labor, and Benefits

Like many scholars in the humanities, I feel a sharp disconnect between the work I do as a researcher, which is technical and philological, and the work I do as a teacher, which is general and often skims the surface of my field. The part of history that excites me—reading primary documents and telling stories about what I discover—often seems inaccessible to my students, or available to them in only the most curated versions. I very rarely give them something with which they can surprise me.

This class attempts to subvert that dynamic by engaging students with an unread primary source in relatively large volume. The *Egyptian Gazette* was published in English, so there is no language barrier preventing my mostly monoglot students from conducting research. This course also offers new transposable technical skills that students might not otherwise acquire: basic understanding of coding and familiarity with some of the primary building blocks of digital work (plain text editing, Github, regular expressions, data visualization). I also hope to give them the chance to publish their editions under their own names, and to contribute to the common scholarly enterprise via our open-access repository. It is also my hope that my students have the chance to discover material no one had discovered before, and to tell the world why they think it is interesting. This is the most rewarding part of scholarship.

While students enrol for required academic credits, it is my intention to train them in the values of humanities research: edification about the Middle East, critical thinking, analytical and writing skills, and active curiosity. All of this is a normal part of any class I teach, and also a common goal of all liberal arts instruction. What is unusual about this class, however, is that I harness their labor in a project that dovetails with my own research challenges. When I have attempted to use the *Egyptian Gazette* in my own research, its scale has been daunting. Hoping to read a complete data series, I was unable to make my way through the entirety of the publication. I was unwilling to invest months of work reading a source many would consider passé: an English-language, elite, colonial source that could only corroborate stories I

https://doi.org/10.1017/rms.2018.13 Published online by Cambridge University Press

discovered in sources deemed more "worthy." But if I could make the *Egyptian Gazette* searchable, I would use it all the time.

Is it exploitation to use my students to create this resource? In the interest of transparency I should note that certain students voiced (and voice) this concern, as do some of my colleagues. My answer is that all coursework requires labor, and it is exploitation to demand unproductive, unstimulated labor. It is a real waste for a student to write a paper without joy or inspiration, a paper I may then read without pleasure or interest. That is a truly unfair transaction. Of course, not all classes or students experience wasted labor in this way. But I came to realize that the concern about exploitation was triggered precisely because I considered what the students in the digitization class might produce to be of value. According to some perverse accounting system, futile exercises could be non-exploitative; reusable products could be exploitation. This class is an attempt to retool these assumptions, and to expose myself to an economy of labor in which I need my students' work as much as they need mine.

More broadly, it is important to acknowledge that there are no digitized sources that do not impose costs. Many of the best digitized primary sources for modern Middle East history are beyond the budgets of all but a few elite institutions. Freely available resources, meanwhile, are often the product of "invisible labor." Education and knowledge are commodified today in ways that make it normal to ask questions of value. But the open access movement is, I believe, an important effort to describe a different system of knowledge exchange, and I want my students, through their contributions to our public repository, to have the chance to participate in it.

## Results

To date, the technical side of the class is the part that is working best. I have a clear sense of the process of transforming microfilm to encoded text, and manage to guide the students through it and impart the technical skills they need. I try to manage their anxieties and expectations. As a collectively-produced repository of texts for research and researchers, the project is beginning to look like a success.

Admittedly, as a class teaching about history and about the Middle East, it is thus far less successful. The technical challenge of the course takes most of my instructional time and requires concrete explanations. When I designed the class, I expected my students' curiosity to motivate self-directed learning about history. The *Egyptian Gazette* contains a lot of material that I find funny and sad and intriguing, and I assumed that its variety would ring bells for my students as well. This has not been the case.

https://doi.org/10.1017/rms.2018.13 Published online by Cambridge University Press

I suspect that because my students belong to a post-newspaper world, the medium is more opaque than familiar. While I had thought they might understand its structures intuitively, this is rarely the case. As students do not consume news and information through newspapers anymore, the newspaper itself is almost as unusual as the microfilm. But there is a larger concern at work here, and one that may be revelatory more generally for humanities instruction. Raw primary sources are difficult for those with untrained curiosity or untrained instincts. In traditional lectures we pre-select material for significance without training students to begin selecting these materials. We thus forget that historians develop a sense of a shared canon of significance over time as part of our own disciplinary training. This we learn and later impart as much through imitation and osmosis as instruction. And so it is an unfair expectation for my students to exercise native curiosity in a way that corresponds to my sense of disciplinary significance. They are eager to give me what I want; I do not wish to tell them what they ought to find in a primary source I have not read; therefore they are anxious. This explains the preponderance of analysis of murders, crime, and disease in the "serial analysis" essays they propose, and the lack of analysis of commodity prices and shipping schedules (which is what I think is most interesting in the paper). The students lack an example of why such data might be interesting.

I have tried to provide some of these alternative indicators for "interesting" historical details by using guest speakers, but this has revealed to me that political economy approaches to history, for instance, do not readily resonate with most of my students—their instincts are not yet attuned to these terms of analysis. The analysis projects are meant to engage my students' own curiosity. A student who sails studied the results of all of Alexandria's regattas, looking for the best sailor in town. One student studied "Gum Arabic" prices because he likes to chew gum. As noted above, the dramatic elements of murder and violence tend to attract interest. Even the economics majors who take the class, and may be predisposed to reflect on more quotidian affairs, are struck by how little their training prepares them to analyze the data the newspaper provides.

The course also reveals that collaborative work is no simple task. In order to do their serial question analysis, students rely on the texts digitized by their colleagues. Of course, the students often find that their colleagues have not completed the encoding of the text that they need. Initially, their response to this finding tends to be moralistic. They are surprised that someone would fail to do the work that they need. This reaction is tempered when they contemplate their own week of the paper, however. They see just

https://doi.org/10.1017/rms.2018.13 Published online by Cambridge University Press

how difficult it is to do work that meets everyone's needs. It seems to me that this is a useful lesson both in the production of history and in the post-classroom world.

One of the great surprises in this class has been the students who come back to see me after the semester is over. None of this (self-selected) group wants to talk about the Middle East. All of them want to talk about the process of encoding the newspaper. It seems to satisfy in them (as it does in me) some sort of perfectionist nit-picking impulse ("OCD," one of them calls it). This impulse is not sufficient for the joint production of scholarship, of course, but it is necessary and—for some of us—satisfying work. Unfortunately (and in contrast to the sciences), this kind of primary source work is something we in Middle East studies can rarely offer to our undergraduate students. Yet it could be a vital component of our own work as scholars, and if students can do it, it satisfies them too. It remains to be seen how many of these coders might ultimately be interested in history. I am trying, with each passing semester, to find ways to make the connections clearer.

https://doi.org/10.1017/rms.2018.13 Published online by Cambridge University Press