CHAPTER 5

“Pirates of Metadata”

[Or, The True Adventures of How One Journal Editor and Fifteen Undergraduate Publishing Majors Survived a Harrowing Metadata-Mining Project]

Cheryl E. Ball
Illinois State University

Introduction

In this chapter, I discuss the use of metadata in digital publishing as both a necessary means for creating accessible and sustainable scholarship and a method of promoting information literacy in students. To make this point, I argue that information literacy extends beyond technical competence and into a critical understanding of the contexts and ecologies in which information is created and used. That is, while understanding metadata, as a concept, is a functional part of information literacy, understanding the role metadata plays in information communication, such as scholarly publishing, requires far more rhetorical and critical understanding, which enhances information literacy practices. The study that showcases this practice centers on a digital publishing class during which I asked undergraduates to mine metadata from an open access scholarly journal that publishes exclusively hypertextual and multimedia scholarship.

Setting the Scene: The Precarious Scholarly Landscape of the World Wide Web

The Internet was built for scholarly communication, and the Web made its distribution that much friendlier. While the military and the sciences had been using the Internet for decades, the Web’s arrival in 1994 allowed aficionados in the digital humanities to take better advantage of this technological and scholarly infrastructure. Within a year of the Web’s debut, online journals proliferated (Hitchcock, Carr, and Hall 1996), and a group of graduate students from around the United States decided to start their own scholarly journal in the interdisciplinary areas of rhetoric, technology, and pedagogy—a field then known as "computers and composition," populated primarily by college writing instructors who also happened to be techies. The journal is now called *Kairos: A Journal of Rhetoric, Technology, and Pedagogy*. The field of computers and composition (sometimes more recently known as digital writing studies or digital rhetoric) researches how writing functions and is taught in networked digital writing environments. This research overlaps with information literacy. As digital rhetorician Stuart Selber (2004) aptly explained in his book, *Multiliteracies for a Digital Age*, computer literacy programs often overemphasize technical skills to the disservice of students (and teachers) who need to engage in higher-level literacy practices. It is generally agreed in digital writing that although information literacy practices necessarily include functional, practical computing skills (e.g., one needs to know how to use a word processing program in order to write in it), these lower-level skills should be incorporated only into teaching and learning practices that frame learning within contextually driven spaces that focus on higher-level rhetorical and critical-literacy practices. Thus, the approach to information literacy practices that I espouse in this chapter is akin to what the Association of College and Research Libraries (ACRL 2000) states as the goal of information literacy (versus information technology) in reporting on the differences between these two terms: "Information literacy's focus [is] on content, communication, analysis, information searching, and evaluation; whereas information technology 'fluency' focuses on a deep understanding of technology and graduated, increasingly skilled use of it" (para. 5). In the next section, I describe how the journal I edit, *Kairos*, served as an experiment in information literacy learning for students in an undergraduate publishing class.

Rising Action: Everything Seems under Control until ... Metadata!

It won't be news to librarians, information literacy specialists, and digital communication scholars that 1996 was a time filled with both promise and peril for any new publication starting on the Web. *Kairos*'s editorial staff knew that and planned well, filling a niche in scholarly publishing that was made for the Web: hypertextual and (as Web-based design technologies matured) media-rich scholarship. In the spirit of the academic discipline that the journal calls home—one in which writing is composed and taught as a collaborative process between multiple authors in networked computing environments—

Kairos has always peer-reviewed submissions collaboratively and has always been an open access journal, making itself freely available to anyone with Internet access.[1] It is the longest-running journal of its kind in the world.

From the first issue in January 1996, the editors had the foresight in planning for the future of the journal to find in-kind server space, plan editorial collaboration via e-mail distribution lists, create sustainable information architecture for the journal's twice-a-year publication, distribute the workload through co-editors who work virtually with each other, and commit to being an independent publishing venue so that the mission and vision of this experimental journal could remain strong. However, what the original staff didn't know was how crucial metadata would be to finding information on the Web in five, ten, or (as of the writing of this chapter) seventeen years later—or what kinds of metadata would be important to capture the history and exponentially growing future of scholarly multimedia, or how expensive the process of creating metadata after the fact could be, a particular problem for an independent journal with a no-money-in/no-money-out business model.

In 2008, there was a brief lull in the action when the editors began implementing a small metadata schema in every newly accepted webtext (*Kairos*'s term for scholarly multimedia articles). Using a version of Dublin Core, the editorial staff began copying, pasting, and tweaking a dozen or so lines of metadata into the header of every HTML page the journal would publish, starting with the Fall 2008 issue. Keeping in mind that all webtexts are built with a series of linked, interactive webpages, media files, and file directories, the process of pasting, tweaking, and also copyediting and proofreading the metadata for *every* HTML page in an issue is no small undertaking. As an example, the Summer 2012 issue of *Kairos* had 128 HTML pages across fifteen folders and subfolders, and the metadata had to be pasted into and changed to match the unique data (such as URI) of each page, never mind all of the other editorial production work that the staff completes to ensure the highest quality scholarship possible.

In addition, this metadata work is done manually, which isn't at all surprising given the editorial workflow the journal has always used. In fact, every step in the publishing process—from soliciting and reviewing submissions, to copyediting and design-editing webtexts, to publishing an issue—has been performed manually by staff members for the entirety of the journal's history. This means staff members employ functional information literacies such as downloading zip files or folders using a free secure file transfer protocol (SFTP) program,[2] copyediting those files in an HTML editor, uploading those files to another virtual server using SFTP, and e-mailing the staff distribution list to indicate that the text is ready for the next stage of copyediting. Those functional, technical skills support the rhetorical and critical information literacies they practice as editorial and disciplinary specialists in digital media composition, technical communication, user experience design, and so on. Of course, the problem is that it's not 1996 anymore, and *information literacy* no longer refers simply to functional skills but also incorporates the higher-level rhetorical and critical skills. The journal's own communicative practices needed to adapt.

In early 2010, the senior editors knew that the journal's workflow needed to change to keep up with the proliferation of submissions as well as the technologies and technical standards required of Web-based scholarship. We needed a system that would help us automate and sustain this otherwise functional process, a system that would be technologically well beyond our current practice of relying on one editor's personal e-mail archives and "Type-A" approach to publication timelines (myself, as editor) and another editor's extensive server knowledge (Douglas Eyman, senior editor of *Kairos*). This system should also allow us to set up a workflow that didn't rely on our institutional memories so that new editors could step into these roles without problems. But there were no editorial management systems on the market (either open-source or commercial) that, out of the box, could handle the kind of multimedia content *Kairos* publishes. And with no budget, we couldn't afford to buy a commercial system and request tweaks, nor could we implement an open-source system and pay a programmer to make changes suitable for us. So, we applied for a National Endowment for the Humanities (NEH) Digital Humanities Start-Up Grant, which would allow us to pay a programmer to build multimedia-specific plug-ins for the open-source software Open Journal Systems (OJS).

OJS has been around since 2001 and is distributed for free by the Public Knowledge Project (PKP).[3] *Kairos* chose OJS as the foundation for its editorial-system grant because PKP's founder, John Willinsky, is extremely dedicated to open access scholarship and to making OJS open-source, including opening its codebase to programmers, which was an important requirement for completing the grant project on time and on budget. In addition, with OJS we wouldn't have to maintain our own system; we would just have to build plug-ins that work with the existing system and offer those plug-ins back to the open-source community for others to make use of and improve. We were lucky to get the NEH grant on the first try, and the *Kairos*-OJS plug-ins should be available to the public by the fall of 2012. But, in our excitement at getting the grant, we'd forgotten one thing.

The Climax: The Specter of Metadata Returns!

OJS runs on a database, and in order for that database to work, it needed data that we didn't have. We had only a very limited set of Dublin Core metadata for two years of the journal's then fourteen years of publication, and that data was not easily extracted from the code in which it was embedded. So the first order of business was to create a metadata schema that would capture the data *we* wanted to capture within OJS, which uses a limited variation of Dublin Core. Doug Eyman, myself, and Kathie Gossett, *Kairos*'s associate editor, spent three months creating a crosswalk comparing Dublin Core, OJS, and *Kairos*'s unique metadata needs specific to its multimedia content. (Unfortunately, space precludes me from detailing the outcomes of that process in this chapter.)

In the process of discussing schemata, we realized we wanted to capture data not only at the webtext (or article) level, but also at the level of the media element, such as a path-specific URI that identifies where a media element falls within the architecture of a specific webtext (e.g., /images/header.gif). With this goal in mind, we ended up with twenty-nine fields to capture at the webtext and media element levels, including Title, Creator, Keyword, Description, Designer, Status, Genre, FileType, and others. We could use a metadata field such as Title to refer not only to the title of a webtext but also to the title of an HTML page, since each page in a webtext functions as a discrete, nonlinear unit in our publications. In addition, a metadata field such as Description could stand in for a webtext's abstract but also, at the media element level, as the alt text for an image used in the webtext. This level of granularity would allow us to provide more comprehensive and more finely tuned research opportunities for readers and potential authors, eventually allowing us to tag every media element in a webtext so that it would be independently searchable and remixable and could be cited appropriately. This granularity would also allow us to better describe and preserve, if only through metadata, some of the webtext components that become technologically obsolete with age. A good example is *Kairos*'s most-cited webtext, "a bookling monument" by Anne Wysocki. It's a Shockwave piece from 2002, designed in Macromedia Director (when, alas, there was such a program), that only occasionally still runs, depending on whether browser companies decide to keep the Shockwave browser plug-in up-to-date. For several years in the late 2000s, the piece was completely inaccessible, but people still cite it because it is one of the most cutting-edge and unusually designed pieces in the journal's history. Metadata would help us preserve the import of the Shockwave piece for archival and research purposes, even if the medium—or, more specifically, file format—in which the piece is delivered becomes inaccessible again in the future.

We were so wrapped up in what data we wanted to collect in our redesigned version of OJS, however, that we forgot we would need to collect data for all of our back issues as well. To populate the impending OJS database, we would need to *create* metadata for what was, on early counts, over 500 webtexts and 25,000 media elements that the journal had already published. Worse, having already spoken with several supercomputing experts on data mining, we knew there was no way to do this algorithmically with our multimedia content. (In fact, those experts are only now, two years later, starting a project where this work *might* be possible.) At the time, not a few tears were shed during the confrontation with this massive metadata-mining challenge, which we knew could be completed only with human labor and a ton of perseverance. But how? The journal staff consists of around twenty-five PhD students and tenure-track scholars who volunteer a few hours a week (and sometimes much more) on top of their high teaching loads (the average is four classes per semester) to put out two or three issues a year. The additional workload would have been an undue burden for them, and the documentation I would have had to prepare to make this project work at twenty-five different sites (since the staff is distributed) would have been an undue burden for me. And if I took this project on myself, what could I learn from it? Better yet, I realized, my students could learn from mining metadata from scholarly, open access multimedia?

Yes, I would have a captive audience of fifteen undergraduates in my digital publishing class the following semester. All of them would be seniors in my department's publishing studies sequence, the most difficult sequence to get into (due to the number of seats available). Thus, the sequence has the highest standards for students—standards that, in my experience teaching in this sequence, the students surpass on a weekly basis. They are the best of the best. On the one hand, I admit feeling guilty about throwing them into such a massive project, and one that I would see professional benefit from. On the other hand, students in this sequence crave real-life and practical publishing experiences, and this project was unlike any they would work on in their other publishing classes. Most students wanted something "digital" in this sequence, and many waited a semester to take this class with me because it dealt specifically with digital topics. This class opportunity was the perfect solution to my metadata problems, and I vowed from the beginning to credit the students' data-mining work, either through acknowledgements or co-authorship, as the case warranted.[4]

Falling Action: Teaching Metadata to Make the Journal Sustainable

To collect this data—in what turned out to be over 800 webtexts from *Kairos*'s then fifteen years of publication—I created a syllabus for my senior-level publishing class that included a ten-week sequenced assignment of mining the metadata, which I discuss in more detail below, and a reading list on metadata, open access and digital publishing, and nontraditional scholarship. Some of these readings included Baca's (2008) *Introduction to Metadata*, Fitzpatrick's (2010) *Planned Obsolescence*, Borgman's (2007) *Scholarship in the Digital Age*, and Willinsky's (2009) *Access Principle*; I purposefully used the open access versions of these texts when they were available. Based on those readings, we discussed issues such as these:

- What is scholarship, and why is peer review important?
- What role does peer review play in your professors' lives?
- What does open access mean?
- How does being open access impact the sustainability of digital scholarship?
- What are these "webtexts" we're working with?
- What is metadata, and why is it important to digital publishing and to webtexts in particular?

The students were eager to discuss these topics in detail since most of the concepts were brand-new to them, and all directly related to their major. For instance, the students had no idea what tenure, or the tenure track, was, even though these concepts pervade their university lives through their professors.[5] Tenure relates directly to the ideologies and processes of scholarly publishing, and so to be better editors and publishers, these publishing studies students would need to know as much about this form of scholarly communication as they could. We had long discussions—in relation to reading the peer-review sections of Fitzpatrick's (2010) book supplemented by my personal experiences and research regarding the use of digital and open access, peer-reviewed scholarship in applying for tenure[6]—about why professors have to research, what the outcomes of that research look like in different humanities fields, where and how it gets published, who reviews it, what editorial reviewers get paid, and what getting a peer-reviewed article published in a scholarly journal means in relation to their teaching effectiveness and tenure. All of this information was crucial for students to know so they could better understand why an author or an editor might face certain institutional and disciplinary challenges when choosing to publish in an open access journal, never mind in a medium—such as webtexts—that differs from traditional forms of scholarly communication.

Open Access

The first major lesson of the class centered on understanding the importance of open access. Students in the publishing sequence are trained primarily in print-based, literary and nonprofit (grant-funded and subscription-based) publishing, and they know how to edit, design, market, and distribute literary texts. But prior to this class, they hadn't considered what access they'd have to these texts, or to any of the scholarship professors require them to cite in their own papers, once they graduate. John Willinsky's (2009) book provided a great and easy-to-read (so said the students) introduction to the principles of open access. For instance, Willinsky bluntly says:

> What is clear at this point is that open access to research archives and journals has the potential to change the public presence of science and scholarship and increase the circulation of this particular form of knowledge. What is also clear is that the role that open access will play in the future of scholarly publishing depends on decisions that will be made over the [next] few years by researchers, editors, scholarly societies, publishers, and research-funding agencies.
>
> This is a book that lays out the case for open access and why it should be a part of that future. It demonstrates the vital and viable role it can play, from both the perspective of a researcher working in the best-equipped lab at a leading research university and that of a history teacher struggling to find resources in an impoverished high school. (ix–x)

To drive these points home, and perhaps much to the chagrin of my university's library officials and information technology staff, the students and I had frank conversations about the purchase of proprietary software for creating bibliographies when dozens of open-source versions existed, which students could learn now, for free, and continue to use long after they graduate. We also discussed the difference between open access and open source, and the fact that some open-source programs, like Zotero, could capture and store open access and openly available documents on the Web. To clearly demonstrate the levels of access that students would have after they graduate, I asked them to look up the CV of their favorite professor, find an article he or she had written, and see whether they could access the full text of that article online without going through our library's website. In every case, the answer was no. Yet they were, or would be, that high school teacher (or nonprofit editor) Willinsky referred to.

To compound Willinsky's point, I relayed the news of the National Institute of Health's decision to require scholars receiving NIH grants to publish their results in a venue that is open to the public.[7] In reading Borgman's (2007) book on digital scholarship, with its particular emphasis on e-science, we had already discussed the salary and grantfunding disparities between the sciences and the humanities and the fact that the sciences usually build paying for open access publishing into their grants, so the NIH's decision wasn't that big a deal, whereas open access in the humanities could be a financial hurdle as well as an ideological one. As a counterpoint to the NIH example, I told them an anecdote that Brett Bobley (2010), Chief Information Officer for the NEH, shared at a conference once:

> I get a little Google alert whenever various things occur, and I saw a little article about the fact that [a big-name scholar has published an article]… And I click on it and what comes up? A pay wall. It's printed in some journal, and that means I'll never get to read it. Ever. And I work for the NEH! I fund this stuff! Scholars all the time say to me, "Hey, Brett, did you read that article I published?" I go, "Did you publish it open access?" No. I never read it. I can't afford journal subscriptions.

Bobley reminds academics that if scholarship is not published open access, neither the funders nor the general public will ever see it. Given this information paired with the class discussions about tenure and peer-reviewed scholarship, the students could easily see why open access was an important point along the publishing and information communication spectrum. And, although this publishing course was not a special topics class in open access scholarship, most of our discussions came back to this issue throughout the semester, including why and how we were to collect metadata for *Kairos*.

Metadata for Webtexts

The connection between information literacy and the production of the metadata was implicit in the class, but I hope to make that connection explicit for readers in this section. The point is that technical tasks, such as metadata creation, should not exist outside of the critical, rhetorical contexts in which they are being performed if a full sense of information literacy is to be expected. In this case, the critical and rhetorical *topoi* include digital scholarship, peer review, open access venues, copyright, and other issues within the scholarly communicative landscape.

To prepare for the metadata-mining project, we spent the first few weeks of class reading about open access, peer review, and the kinds of digital media scholarship that *Kairos* publishes. We read Baca's (2008) *Introduction to Metadata*, which put into larger context some of the instruction sets on mining metadata from *Kairos*, which I provided students on a weekly basis. Based on the great questions raised by

Baca's book, such as why metadata is important, I wrote lengthy contextual explanations into the instruction sets for students as a way to reinforce the scholarly and publishing importance of creating metadata for digital texts. Their first handout explained several reasons why we were collecting metadata from *Kairos* and what that data would be used for:

> 1. It will be used by *Kairos* editors to populate a database they are creating. This database, which will interact with Open Journal Systems (a scholarly publishing platform) to allow readers, editors, and authors to better search for useful digital media scholarship in the journal.
> 2. It will allow for more accurate citation practices of the digital media elements within *Kairos* webtexts.
> 3. It will make previously published webtexts more accessible for more users—both for scholars doing Web-based researchers [*sic*] and for users who are differently abled.
> 4. It will serve as a prototype for metadata in all future *Kairos* submissions, so that authors will begin to create their own metadata upon submission to the new database/system; thus making the gathering of metadata more sustainable in the future, based on your experiments and workflow recommendations.
> 5. It will be used by editors and researchers to discover new information (e.g., relationships, visualizations, search patterns, reading patterns, media-types, etc.) and to create new knowledge about digital media scholarship.
> 6. Once the metadata terms we are using have been conceptualized through your work and proven to be successful (or not), the metadata terms will be distributed to other digital media publications so as to become a standard for this kind of scholarly publication. (Ball 2011)

Because of the scope of this project, I knew it would be crucial to remind students that it was equivalent to an internship and would be a useful résumé line for them. (Although I wasn't expecting it, two students went on to get jobs where their primarily responsibilities were to work with metadata in digital publishing venues.) I translated to layperson's terms the 25 items that we would capture over the eight weeks of hands-on class time spent on this assignment: Authors, Designers, Creators, Author/Designer Affiliation, Academic Rank, Author/Designer (current) Emails, Webtext Title, Abstract, Publisher, Volume/Issue, Date Published, Section, Language, Peer-Reviewed Status, Peer-Reviewers, DOI, Rights, File Name, File Size, MimeType, Dublin Core Metadata Initiative (DCMI) Type, URI, Page Title, Alt Text, and Genre. These fields crossed three categories of data we wanted to collect: at the level of the webtext, at the micro-level of the media elements within a webtext, and contact information for authors. (There were thirty-five fields of metadata total, some referenced below and repeated across the webtext level and the media-element level, but I had to cut back based on what the students would be physically and emotionally able to accomplish during the term, so we ended up with nineteen. Space prevents me from detailing all of these fields.) I parsed the mining project into the assignments shown in Table 5.1, which I thought would create a workflow that made the most sense given the concepts, locations in the webtexts, and technologies students would need to find them.

Table 5.1
Metadata Elements Presented During the Semester

Week 1 [Feb 9]	Fields requiring Little Instruction
Week 2 [Feb 16]	Fields requiring Simple Lists (not MimeTypes) + DOI
Week 3 [Feb 23]	Rights + Affiliation, Rank, Email
Week 4 [March 2]	Abstract, Keywords + Notes [spring break]
Week 5 [March 16]	MediaID + FileName, FileSize, MimeType, DCMI Type, URI
Week 6 [March 23]	Page Title, Alt Text, Creator + [Webtext] DCMI Type, File Size, URI
Week 7 [March 30]	Genre [Webtext & Media tabs], Creator
Week 8 [April 6]	Update Rights & Affiliations fields

Every week, students would get another multipage handout describing in detail how to collect or create some grouping of this metadata. These handouts always included brief discussions about why fields as seemingly simple as Author, Title, Publisher, and Date might be difficult to find and might even be contested. For instance, the handout "Fields Requiring Little Instruction" included directions for finding authors, webtext titles, volume and issue, language, designers, and peer reviewers and was five single-spaced pages with four images—two each to demonstrate how to find designers and peer reviewers (information that is rarely included in webtexts). The description for finding authors alone included the following details (which probably won't make sense to readers, but did make sense to students since we'd spent a good deal of time looking at the journal before starting the project):

> Authors:
>
> 1. To find the authors for a webtext, look at the Table of Contents (TOCs) for each issue of *Kairos* or on the "home" page for each individual webtext. To access the back issues, go to the *Kairos* home page (http://kairos.technorhetoric.net) and click on the tab at the top for "Issues." The TOC is on the main page of the journal, EXCEPT for the following issues: 7.3, 6.2, 5.2, 4.1, where the TOC for the "CoverWeb" section has to be accessed by clicking on the themes or the hyperlinked title to the CoverWeb.
> 2. Once you find the authors, copy them from the webtext and paste them into the Authors column in the Webtext tab of the Excel spreadsheet. Authors should be listed just like they appear in the webtext, including any middle initials, but NOT including any degrees or ranks (e.g., PhD, if it follows a name).
> 3. If there are multiple authors for a single webtext, they should be listed in the order they appear on the webtext, with commas separating each full name. BUT MAKE SURE TO DELETE the "and" which will usually be included in the TOC.
>
> EXAMPLE:
>
> Author listing in the TOC: Christopher Dean, Will Hochman, Carra Hood, and Robert McEachern
>
> Author listing in the spreadsheet: Christopher Dean, Will Hochman, Carra Hood, Robert McEachern

Date of publication (another not-simple entry) would have been easier if the journal hadn't changed its publication schedule halfway through its history, or its name (a third entry that required choosing from multiple options) from *Kairos: A Journal for Teachers of Writing in Webbed Environments* to its current name in 2004.

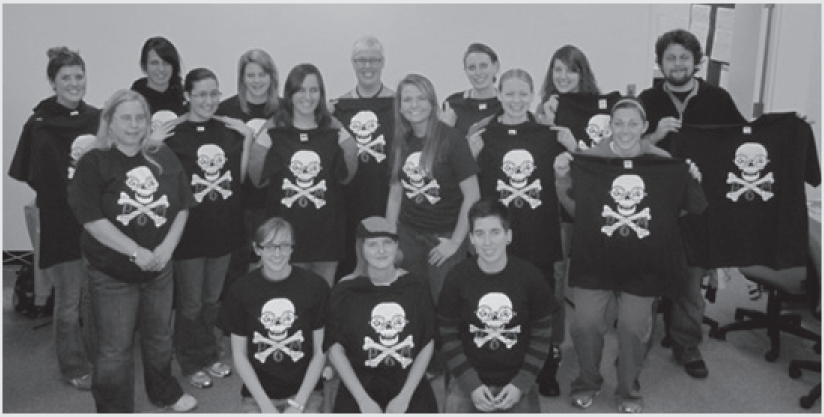
In the schedule shown in Table 5.1, there was a definite split between the work completed before spring break and the work completed afterwards. After break, students had to move from browser-based mining to code-based and file-directory-based mining. That is, before spring break, they had been searching through the interfaces of the journal and webtexts to find the information they needed, using web browsers such as Firefox—technologies they were familiar with. After spring break, they had to use FTP programs and web-editing programs like Dreamweaver to download and search through the code, in some cases. The major hurdle here was not necessarily the difficulty level of teaching students what a DCMIType was and when a GIF is not a StillImage but a MovingImage.[8] The difficulty was that most of the students had never before made a webpage or put it on a server; they had to be taught how to search for, download, and install web-editing and publishing software on our lab computers and their laptops, then to complete intricate and extended searches in HTML code or file directories for the metadata information they needed. For instance, the most efficient and least technologically complicated way I could figure out how to mine for alt tags on all images was to have students search for the alt tag code in an entire issue of *Kairos*. For most of the students, this was their first introduction to HTML code, so the instructions on just this one part of the week's assignment were three and a half pages long, and that didn't include the definitions for terms such as *file directory, HTML tag,* and *SFTP program* (which had previously been covered). The instructions included definitions for nearly every step in setting up a site in Dreamweaver, including what Dreamweaver was and what open-source programs students could use if they didn't have Dreamweaver at home.

Each set of instructions also included Mac- and PC-compatible keyboard shortcuts or menu names. Most of each three-hour studio class had us working hands-on to start that week's mining assignment, troubleshooting the instructions when students inevitably ran into interface, architecture, or technology issues that didn't match every possible combination I could think of in advance. The instruction sets, initially created for a student with learning disabilities in the class, quickly became the reference for all students as we collaborated as a class on how to use and improve them. In and of themselves, they were a perfect example of how access for one can mean better access for all, a macrocosmic example of what alt tags do for each microcosm of a webtext. Finally, this course was a prime lesson in what Stuart Selber (2004) has termed the functional, critical, and rhetorical literacies inherent in being multiliterate in a digital age. Without the critical literacies of understanding digital scholarship, peer review, and open access publishing; without the functional literacies of tinkering with file directories in Dreamweaver, Firefox or Safari, and Filezilla; without the rhetorical literacies of applying typical units of analysis to webtexts (e.g., who is the audience, what is the text's purpose, in what context is it published, etc.), these students could not have *begun* to compete this project. But they did. And their data was, as much as could be expected, clean and excellent.

Denouement: The Pirates of Metadata Are Salvagers Extraordinaire!

This was a massive project—too big—which the students and I coped with in different ways. Students would come to class excited to tell me how they explained this metadata project to their history or biology major roommates. At the same time, they were exhausted by its menial orientation, not surprising given the cut-and-paste tasks at the heart of this project. The students completed the semester by producing an Excel spreadsheet for each of the journal's issues they were assigned to mine. On average, each spreadsheet contained 35,000 cells of data, and each student had at least two spreadsheets. My rough count is that students collected over a million cells of data. On its own, the data has the potential to shape the way scholars research and think about *Kairos* webtexts as representative of the history *and* future of design on the World Wide Web. This makes their work no small feat (the outcomes of which I discuss more below). The students coped by expressing how they felt week after week of mining metadata: Arrgh!!! It drove them crazy, but they also love-hated it. They started calling themselves the Pirates of Metadata and made their own logo and T-shirts, covered with metadata jokes only they would appreciate (see Figure 5.1). One of the jokes was a riff on our DOI schema—volume.issue.section.authorLastName-et-al—which was transformed into 5.11.kairos.arrgh-et-al for a tagline on the shirts (5.11 was for May 2011, when the class ended). They'd twisted the functional literacy of a DOI naming schema into a rhetorically appropriate parody—a

Figure 5.1
The Pirates of Metadata, Proudly Sporting their T-shirts



transfer that showcases, even in a minor and fun way, their critical-information literacy learning.

It was through the students' information literacy learning via this metadata project that they were able to make a significant contribution to digital publishing studies. And vice versa: because I reinforced weekly that the students were contributing to scholarship in digital publishing studies, they understood that their work had reach far beyond the classroom and were willing to push themselves harder to make that impact successful. To reach this outside-the-classroom audience, I asked students to write a report outlining their methods of data mining (particularly if they deviated from the instructions I provided) and include observations about their dataset and recommendations for stakeholders. Their audience was editors, librarians, information literacy scholars, and others who might implement a similar project with a scholarly multimedia journal in the future. Goals of the assignment included reflecting on what they learned from the metadata project in relation to the theoretical contexts of digital publishing studies and to summarize outcomes of that learning (via Findings, Discussion, and Recommendations sections) by providing succinct examples from their metadata sets. For instance, in the Findings section of the report, I suggested some kinds of data they might report on:[9]

- the kinds of genres they ended up using
- the number of media files they ended up with
- a short list of examples of how media files were named by the authors
- the sections their Volume.Issue covered
- the number of alt text or page titles (or not) used in their webtexts

This data was typical of those we spent more time discussing in class, as opposed to the more (but not exclusively) functional cut-and-paste fields such as Author, Volume.Issue, and URI. The question about which *sections* appeared in students' particular Volume.Issue, however, would elicit information critical to the historical changes in sections that *Kairos* has undergone (e.g., the first issues had a section called Pixelated Rhetorics, which morphed into *Kairos* Meet the Authors, which morphed into two different sections: Interviews and Praxis). Changes in sections sometimes indicated the peer-reviewed status (another metadata collection point) of webtexts, which has repercussions for authors' tenure and promotion. Although students wouldn't always know these larger issues that I, as editor, could easily interpret from the data, we discussed these issues in class, and if I suggested the impact factor of section changes, students could easily grasp its import to publishing studies as a whole. A quick overview of the students' outcomes and recommendations from this study shows the following import of seemingly functional topics such as *genre*, *media files, naming conventions,* and *alt text*, which the students and I have presented elsewhere (Ball et al., 2011):

- Web architecture has changed dramatically in fifteen years, with a noticeable shift between volumes 1–10 and 11–15. Journal architecture as a whole is messier than it should be (particularly in older issues), but individual webtexts have become more "deep" in their folder structures.
- File-naming conventions have become slightly more rhetorical (e.g., named according to rhetorical function, such as header.gif) and more technologically sustainable (e.g., fewer filenames in ALL CAPS or weird spaces).
- Genres and DCMITypes change dramatically as the journal grows.
- The number of webtexts published per issue has been halved.
- Accessibility elements such as alt text and page titles are missing from most early issues and are inconsistently used in later issues.

This is just a small sample of the observations students made in their reports about *Kairos* based on the metadata project. And a major observation that nearly all the students had was that mining metadata retroactively is costly and prone to human error. Some of the problems that students encountered in trying to mine metadata—such as finding accurate affiliations, ranks, and e-mail addresses for authors, particularly those in earlier issues—are already part of OJS or were already planned as part of the Kairos-OJS version. But the students came up with other recommendations or requirements that were incredibly insightful and that *Kairos* plans to implement in future iterations of our metadata-collection schema in OJS, such as:

- Webtexts need technology descriptions in abstracts that also describe the interactive designs of each piece.
- Accessible documentation (alt text, transcripts, reading instructions, etc.) should be a mandatory part of any webtext submission.
- A controlled vocabulary (if that's possible?) for webtext and media genres should be provided so that authors can tag their own elements from this set list.

Finally, students recommended that the labor of metadata be shifted to authors. This is not a surprising recommendation given the state of digital scholarly publishing at the moment. Calls for open access and collaboration are often accompanied by calls for crowdsourcing, which essentially means re-envisioning the labor structure of publishing. Students who were brand-new to digital publishing could, after only one semester of study, see this and agree that this discussion needs to take place. Their recommendations are important—and exciting, knowing that these students are the next generation of critically, rhetorically, and functionally literate editors of scholarly communication.

Acknowledgements

I owe thanks to Ryan Trauman for the idea of being able to remix and better cite media elements through metadata tagging. I also thank the editors and Isaac Gilman for great advice in shaping this chapter. Finally, the most thanks goes to the Pirates of Metadata: Sarah Chance, Lyndsey Eagle, Meghan Engel, David Gaudio, Susan Grogin, Melissa Hermann, Laura Patrick, Kira Plotts, Valerie Romack, Constance Ruholl, Kali Shevlin, Ameliah Tawlks, Amy Thomas, Kirsten Van De Veer, and Jessica Wosniak.

Notes

1. As of last count, the journal has over 45,000 unique hits a month, with readers in 180 countries (Eyman 2006).
2. *Kairos* advocates open-source software such as Cyberduck.
3. For more information about the Public Knowledge Project, see http://pkp.sfu.ca/history.
4. We have already co-authored a poster session on the outcomes of their mining workflow: see Ball et al. 2011.
5. Illinois State University is a second-tier school in the Normal tradition, well respected for its faculty teaching and its teacher-education programs, and the English department faculty typically teach two to three classes per semester, but the university still has strong research expectations, with peer-reviewed articles and scholarly books making up the bulk of what's expected prior to tenure.
6. See Ball 2009.
7. For more information, read the National Institutes of Health Public Access Policy Details at http://publicaccess.nih.gov/policy.htm.
8. There are twelve terms—or descriptors for "the nature or genre of the resource"—in the DCMI Type controlled vocabulary: Collection, Dataset, Event, Image, InteractiveResource, MovingImage, PhysicalObject, Service, Software, Sound, StillImage,

and Text (DCMI Usage Board 2012). *Kairos* uses only a subset of these Types (e.g., the journal doesn't publish PhysicalObjects, Events, or Services). All webtexts in *Kairos* are considered InteractiveResources under DCMI's definition, but not all GIFs are StillImages because animated GIFs *move*, which makes them MovingImages instead. In this case, information technology skills (e.g., knowing that .gif represents an image file) don't help a metadata miner understand the context in which that GIF is used. Instead, a student needs to understand the rhetorical context of the GIF by viewing it on the webpage in which it was published (e.g., what's the GIF *doing* and in what context) in order to evaluate its function within the webtext and thus tag it appropriately in the metadata.

9. One goal of this assignment was to teach students how to write business reports, a genre that publishing majors would need in their jobs. For the full assignment (and links to other assignments on the syllabus), see http://ceball.com/classes/354/spring11.

References

ACRL (Association of College and Research Libraries). 2000. *Information Literacy Competency Standards for Higher Education.* Chicago: ACRL, January 18. http://www.ala.org/acrl/standards/informationliteracycompetency.

Baca, Murtha, ed. 2008. *Introduction to Metadata,* 2nd ed. Los Angeles: Getty Research Institute.

Ball, Cheryl. E. 2009. "Tenure Letter." http://www.ceball.com/tenure/intro/tenure-letter.

———. 2011. "Metadata Project Description Sheets [English 354]." http://www.ceball.com/classes/354/spring11/wp-content/uploads/2011/02/spreadsheet-descriptions1.pdf.

______. 2011a. "Metadata Instruction Set: Fields Requiring Little Instruction." http://www.ceball.com/classes/354/spring11/wp-content/uploads/2011/02/metadata-instructions-LITTLE.doc

Ball, Cheryl E., and The Pirates of Metadata. 2011. "Learning through Leading: Digital Media Scholarly Publishing." Poster presented at New Media Consortium conference, Madison, WI, July 19.

Bobley, Brett. 2010. "Opening Up the Ivory Tower? Access and Academic Publishing." Fora.tv. YouTube video. 2:56. From a discussion at the conference The Digital University, New York, NY, April 21, 2010. http://www.youtube.com/watch?v=7mRFRe4DxdM.

Borgman, Christine. 2007. *Scholarship in the Digital Age: Information, Infra-*

structure, and the Internet. Cambridge, MA: MIT Press.

DCMI Usage Board. 2012. "DCMI Metadata Terms." Dublin Core Metadata Initiative. June 14. http://dublincore.org/documents/2012/06/14/dcmi-terms.

Eyman, Douglas. 2006. "The Arrow and the Loom: A Decade of *Kairos*." *Kairos: A Journal of Rhetoric, Technology, and Pedagogy* 11, no. 1 (Fall). http://Kairos.technorhetoric.net/11.1/binder.html?topoi/eyman/index.html.

———. 2007. *Digital rhetoric: Ecologies and economies of digital circulation.* Dissertation. Michigan State University, Lansing, MI.

Fitzpatrick, Kathleen. 2010. *Planned Obsolescence: Publishing, Technology, and the Future of the Academy.* MediaCommons Press edition. http://mediacommons.futureofthebook.org/mcpress/plannedobsolescence.

Hitchcock, Steve, Leslie Carr, and Wendy Hall. 1996. "A Survey of STM Online Journals 1990–95: The Calm before the Storm." The Open Journal Project. Last updated February 14. http://journals.ecs.soton.ac.uk/survey/survey.html.

Selber, Stuart A. 2004. *Multiliteracies for a Digital Age*. Studies in Writing and Rhetoric. Carbondale: Southern Illinois University Press.

Willinsky, John. 2009. *The Access Principle: The Case for Open Access to Research and Scholarship*. Cambridge, MA: MIT Press.

Wysocki, Anne. 2002. "Bookling Monument." *Kairos: a Journal of Rhetoric, Technology and Pedagogy,* 7, 3 (Fall). http://kairos.technorhetoric.net/7.3/coverweb/wysocki/.