From Normalizing Serials to Normalizing Ships: Improving Access to All Types of Digitized Resources

Peter McCracken
ShipIndex.org, peter@shipindex.com

Follow this and additional works at: <https://docs.lib.purdue.edu/charleston>

An indexed, print copy of the Proceedings is also available for purchase at:

<http://www.thepress.purdue.edu/series/charleston>.

You may also be interested in the new series, Charleston Insights in Library, Archival, and Information Sciences. Find out more at: <http://www.thepress.purdue.edu/series/charleston-insights-library-archival-and-information-sciences>.

Peter McCracken, "From Normalizing Serials to Normalizing Ships: Improving Access to All Types of Digitized Resources" (2010). *Proceedings of the Charleston Library Conference*.
<http://dx.doi.org/10.5703/1288284314853>

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

FROM NORMALIZING SERIALS TO NORMALIZING SHIPS: IMPROVING ACCESS TO ALL TYPES OF DIGITIZED RESOURCES

Peter McCracken, (peter@shipindex.com) - Co-Founder & Director of Content and Business Development, *ShipIndex.org*

I want to talk today about the challenges encountered in my new company, ShipIndex.org, and how they relate to the conversion of historical data to an online format. As I walked in here, I realized that at a previous Charleston Conference, some indeterminate number of years ago, I gave a presentation in this very room. I don't remember what it was about, but I'm sure it had something to do with serials. And one of the things I want to talk about today is how ships are like serials. I literally went from normalizing serials to normalizing ships, and that came as a bit of a surprise to me. But as I started thinking about the similarities between those two actions, I started thinking about the challenges that I've found in trying to make the ShipIndex database work for modern researchers. I realized that the challenges we've faced recently are applicable in just about every field of historical research, and probably in many other fields, as well. So I thought it would be useful to talk about the experience I've found in ShipIndex and think about how one could apply it in other areas of research, and about actions we can take to improve access to historical resources. It's clear to me that there's an enormous amount of information that's hidden in our collections, or in the case of online resources, they're hidden in plain sight. ShipIndex is an example, I think, of a way that new technologies can release that information.

I'd like to shift gear for a moment, and share some information about the names of different types of rigs on sailing vessels. **Figure 1** shows a variety of different vessel types, including a schooner (defined as a vessel with two or more masts, all of which have fore-and-aft sails – that is, sails that are essentially parallel to the keel of the boat); a barkentine (defined as having three or more masts, and square sails – ie, square, or perpendicular, to the keel – only on the first mast); a bark, which unlike the barkentine has square sails on the second mast; a topsail schooner; and a brigantine. And my favorite of these all different types of ships is the one known, technically, as a *ship*. In fact, it's like the linking ISSN – just as with an ISSN-L, when you don't know if you're talking all the different format versions of a particular journal, or an online version of a print serial, here you don't know if you're talking about a vessel that's larger than a boat, or a specific three-masted rig called a ship.

Figure 1. A variety of sailing rigs. Top row, left to right: schooner, barkentine (or barquentine), bark (or barque). Bottom row, left to right: topsail schooner, brigantine, ship.

In the past two years, I've found many ways that ships are like serials. I was thinking that the vessel type is similar to the publication pattern, or formats. There are certainly more vessel types than there are formats, but we still have many different publication patterns. And the serial title is very similar to the ship name; I'll explore this point more closely later. The ISSN is like the IMO, the International Maritime Organization number, though IMOs are only assigned to a relatively select group of vessels. (Perhaps the equivalent is that blogs don't get assigned ISSN?) But I want to emphasize that this isn't about ship research. Ship research is something that's particularly interesting to me and something that I enjoy, but I think that ship research is similar to any other type of historical research. So I hope to use this very small niche group to look at a much broader research universe.

The problems we have with ships are the same as the problems we have with names or with geographical descriptors, and they can apply to other concepts, as well, with a bit of extra work. Non-unique items, like a model of car, such as a Thunderbird, have a similar challenge, especially when you want to talk about a particular vehicle that was owned by a specific person. One example of where we see this is in personal names – for example, look at how much work has been written by Anonymous, who clearly had an inordinate amount of time on his or her hands. Names are not exact, obviously; they can be shortened from the full name to a much shorter version of a name, such as Pablo Diego José Francisco de Paula Juan Nepomuceno María de los Remedios Cipriano de la Santísima Trinidad Ruiz y Picasso, whom we thankfully refer to as just Pablo Picasso.

In some cases names have strange limitations. For example, Michael J. Fox, the actor, doesn't have 'J' as a middle name, but when he went to Hollywood, there was already a "Michael Fox" in the Screen Actors Guild. SAG actors are like racehorses, in that they have to have a unique name. Actors' names are unique, while others of course are not.

Geographic names are a similar challenge – of course there are numerous different versions of the name of that large German industrial city on the Rhine, known in English as Cologne, in German as Köln, and in other languages as Keulen, Colonia, Colònia, Kolín nad Rýnem, Cwlen, Kolonjo, and many other variations.

At ShipIndex.org, we have created a database of citations – currently at 1.42 million¹ – from over 200 different resources in the premium database, and there are over 140,000 citations in the freely-available database. Like Serials Solutions’ original goals, we focus on metadata alone. I believe this huge database changes how one does maritime research, because it makes it possible to find so much more information than one could previously. It locates large tracts of previously unknown and unfindable information, particularly for folks who don’t know how to do maritime research. If you don’t know the maritime literature – and there’s a very limited number of reference librarians who really do, as it’s not a popular topic – then you don’t know where to start. And looking in the online catalog isn’t going to help if you actually want to get to the books’ or journals’ indexes.

Basically, we’ve scanned in all these indexes, or collected citations from sites online, and we’ve put them all into this great big online database. I should point out that, with regard to the idea of resources hiding in plain sight, there are many freely-available resources in maritime history, as in every other area, but one usually doesn’t know what’s in them, to say nothing of knowing they exist in the first place. There’s an Australian Shipwreck Database, for example, and if you don’t know that it exists, then you don’t know to go look in it.

So, how are we going to organize this? How do we manage this? The first thing we need is an access point. There are many different access points that one might use, but only one that makes any sense. There’s the vessel name, of course. There are also vessel numbers, but they have many limitations. For example, the IMO number was only introduced in 1996, and only applies to vessels that are larger than 100 gross tons, and are ocean-going. Many exceptions exist, as well. The captain could be an identifier, but they change between the voyages and die during them, so that’s not very good. The rig or the vessel type could be one option, but those can change, as well, and we have seen how complicated they can be, particularly for those not familiar with them. Owners occasionally changed rigs, sometimes to better match new uses for the vessel, or to deal with labor issues – having more masts would mean that sails were smaller so one could operate the ship with fewer sailors, for instance. Plus, of course, there’s the “ship” problem – a ship is either something bigger than a boat, or it’s a three-masted sailing vessel with square sails on all three masts. One could also look at the port of registration, but that’s obviously meaningless when you look at how many vessels are registered in Liberia but have never been there.

Obviously, the vessel name is the only one that makes sense. That should be easy: it’s written on the back of the boat. One can just check to see what it says there. **Figure 2** shows one example, from the last American whaling ship. (Actually, it’s a “bark”, a vessel with three or more masts, and no square sails on the last mast.) It clearly says “Charles W. Morgan.”

¹ At the time of this manuscript’s submission, the number of citations had grown to 1.52 million in the premium database; it continues to grow.

However, in the 1872 *American Lloyd's Register* it was the “Chas. W. Morgan”. That’s fine, we see it and know that’s simply an abbreviation, to save space in a printed resource, but it matters 130 or 140 years later when we’re trying to do online searching and the digitized version aims to be an exact transcription of the print version. And, actually, a couple of years earlier it was listed as the “Chas H. Morgan.” In WorldCat it is represented five different ways – some have different descriptors, such as including or excluding its date of launch or describing various rigs the ship has carried, but one reads just “Charles Morgan”. Clearly, it’s referring to the same vessel, but it is represented in many different ways. And this is a very famous ship; it’s the last existing American whaling ship, built in 1841. It is moored at Mystic Seaport and right now is out of the water undergoing an enormous restoration. When the restoration is completed, the Seaport plans to sail it – to New Bedford, and New York, and other locations.

Figure 2. The stern of the *Charles W. Morgan*, photo from May 2010, by the author. Also, two 19th century entries from *American Lloyd's Register* that show different names for the same vessel.

But when we’re looking at a system where we’re trying to process lots of data very quickly, and use specific and standardized access points, we’ve discovered innumerable errors. That’s just life; it’s not a big deal, but we need to identify them, and find or create solutions for fixing them as quickly as possible. In fact, this is a core part of what people pay for when they subscribe to a metadata-based service. I always felt that the difference Serials Solutions had to offer was the accuracy of its metadata – anyone could create an A-to-Z list or an OpenURL resolver, but if the journal titles were not normalized together, and the data behind the knowledgebase wasn’t accurate, then the results wouldn’t be, either. When I see these metadata accuracy issues with ships, I am certain it occurs in many other historical subjects, as well.

One man recently got a printed copy of *Lloyds List* from 1812, which was a four-page newspaper published in London, several times a week. He digitized every page, then transcribed the whole thing, making a list of all of the captains and all of the ships. He planned to continue through all the years of the War of 1812, but Carpal Tunnel Syndrome apparently caught up with him and he quit. The work he committed to this project was amazing. What’s interesting to me is that he didn’t look at his end result and address the typographical errors introduced by the original publishers. He didn’t say, “Oh, geez, when I see ‘Adolph & Fredricky’, I notice there are tons of ships named ‘Adolph & Fredericka’, so that must be what they meant.” In another instance, from different pages of the same issue, one entry says “A-I-R-L-E-Y Castle” and another spells it “A-I-R-L-Y Castle”. Now, I don’t mean to complain about the original

publishers; this was a newspaper that came out two or three times a week, and they were undoubtedly rushing to put it together.

In fact, I think there's some very interesting material regarding how this data was recorded. There seem to be some clear conventions on how they made abbreviations or how they saved space. And in many cases, they probably didn't know the correct spelling. Perhaps someone just ran in and yelled, "Oh, Airley Castle has just arrived with Capt. Annett" and someone else scribbled it down and didn't ask about the spelling. (The captain's name, in fact is spelled as both "Annett" and "Annet" in different instances.) Then there are other challenges: In one case we find entries for both "Elliota" and "Ellieta", with the same captain, just a few weeks apart. It must be the same ship, but there's no way to know which spelling is correct. It's an interesting artifact of the creation of this newspaper, but if we want to make it usable to modern researchers through modern technology, we need to address and resolve these issues.

My intent here is simply to identify the challenges that are faced here. In addition to errors introduced by those who created the newspaper, more were introduced by the man who did all the transcriptions, which is not a big surprise. "Earnets" appears for "Earnest"; "Elizaneth" for "Elizabeth" – it's no coincidence that the 'n' and the 'b' are next to each other on the keyboard. Those are easy to fix. The guy who did the Lloyd's List is pretty incredible; I certainly don't begrudge him introducing some of these errors.² It is important to not get too upset at the folks who commit large amounts of time, energy, and resources, to digitization project like this. I do, however, have one exception – I found a transcribed file of 100,000 entries from a late 19th century shipping directory, and if the folks who had transcribed it had just spent a little extra time to do some sanity checking, they'd have found that some of the ships apparently date from the 28th century, tonnage numbers are completely off, spelling is occasionally an interesting new style of fiction, and they've introduced so many errors that the database is barely useable. We certainly won't introduce this data into our database until we can correct at least most of these, and that could take years.

We figured out solutions to most of these problems, but some big ones remain. If you search for "Elizabeth" in the database you find nearly 1900 citations. If you search for "Mary" you find 2600 citations. There are clearly multiple citations for a single name. How are we going to narrow those down so we know we're talking about the *Elizabeth* from the early 20th century, and not the *Elizabeth* from the late 18th century or 17th century or earlier? We could try limiting the results sets, perhaps by time, or by rig, though we don't always have that information. Nationality or geographical location might be possible, but the whole point of maritime studies is that it's between places – it's the history of that space between places. And what happens when a vessel changes its name? If someone is researching a specific vessel, they want to know about the entire life of that vessel.

² About six weeks after presenting this paper, however, I did find a reason to begrudge the work done by the individual working on *Lloyd's List*; although he had digitized thousands of pages of *Lloyd's List* for 1812, he apparently didn't create a backup of the site. We discovered that our links to his site were not working, and after some email followup we learned that he'd experienced a hard drive crash and had lost much of what he'd had on his website, including all of the images of Lloyd's List pages. I feel that the failure of him to create a backup – especially after ShipIndex had offered to provide hosting support to him – does allow for a bit of begrudging.

The result, I believe, is that we need a unique vessel identifier. There's no other way to know which *Elizabeth* became *Hogwart's Belle*. As I mentioned earlier, many vessel identifiers currently exist, but each has its own problems. The Coast Guard Documentation Number is only used within the US, and only for vessels larger than five net tons, with a number of significant types of vessels excepted; the IMO number, as previously mentioned, only started in 1996, and only for ocean-going vessels of over 100 gross tons, again with exceptions; naval identifiers do not work because they can certainly change - CVN-65 (the aircraft carrier USS *Enterprise*) had a previous hull number of CVA(N)-65, for example. And when a vessel leaves the US Navy and joins another navy, of course its hull number changes. There are many different identifiers, but they all have issues that make them inappropriate for our needs.

Instead, we need to use a unique, unchanging, identifier. It has to be "meaning-less" – denoting that nothing is loaded into the identifier. The identifier should have a one-to-many relationship between the identifier and its data. That is, one identifier can have multiple hull numbers. Each hull number, if it is recorded as more than just a free-text field, should be associated with only one identifier. In addition, the identifier itself doesn't change, but the information associated with it can. If you try to encode any information into that identifier, you're destined to screw it up over time, because that encoded information could – and likely will – change.

In today's world, identifiers need to be quickly and automatically assigned, and they need to be assigned by anybody working on a relevant project. A person should be able to say, "OK, give this ship name an identifier, because I need one now. I know something about it. I know this ship right here became that one over there, and I want to connect them. I need the identifier – now – so that I can record and preserve that information."

It could be something done through a standards process, though the time involved in bringing that to fruition scares me a bit. A better solution, perhaps, is that I and three other people could build something and just get it out there. And then we would actually put it to use and get it used. As I said before, individuals should be able to assign it right away. We need to give people – anyone – the ability to split or combine or modify any records in the database. And I want to repeat that this applies in just about any area where there is historical information that we want to be able to use in different places; while I am discussing ships here, I believe the concept applies globally, to any field of interest. It's worth noting that I imagine we would only touch maybe 10% of what's in ShipIndex – if we're lucky. We'd only look for the very obvious entries, such as every entry for *Constitution*, and figure out which citations are talking about that big US Navy ship that's still floating in Boston Harbor, and which ones are not, and make sure that the latter are *not* given the identifier for the one in Boston.

The next step is to decide what fields would be appropriate in providing additional information about the ship. For example, we need to include other naval vessel identifiers, IMO numbers, and identifiers from others systems. Obviously, we need to offer multiple fields for each identifier. And then we need to think about all the other pieces of information that one might want to include, as well. This would include, at a minimum, building date, launch date, length (overall, and on deck, and other measures of length), width, displacement (there are even more options for displacement than for length), how much it draws, hull material (ie, wood, metal,

fiberglass, reeds, etc.), builder, construction location, construction yard, and much, much more. There would be a drop-down list of all the common types of rig for sailing vessels: ship, bark, brigantine, barkentine, hermaphrodite brig, schooner, sloop, ketch, yawl, topsail schooner, skipjack, lugger, and on and on – plus an “other” field, for rigs not listed. We would want to include fields for the dates that the ship had that particular rig, if it changed over time. Of course, nearly any sailing vessel today would also have an engine. So, perhaps we ask, “Does it have sails?” and also “Does it have an engine?” We want to record information about military usage, including for which country or countries, and over what dates. All these areas would grow with time; perhaps we would allow anyone with interest to add new data fields, in addition to adding new data.

A voyage identifier would be quite useful, as well. It would be associated with a specific ship, and of course each ship would have many voyage identifiers – except for *Titanic*, I suppose. One could indicate who the captain was and who the crew were, when it mattered – so you’d note that for *Acushnet*, on its voyage in 1841, that Herman Melville was aboard as a greenhand. Or that Jimmy Carter served on board USS *Pomfret*, among other vessels.

If you had a link to some sort of geographic place identifier, you could easily generate a list of what vessels came to a particular port in a particular year. For vessels where you know when they left one port and arrived in another, you could start to average the amount of time it took to sail from one place to the other. Or if you identify a particular ship in a certain painting, you could connect the ship and the painting. A system for identifying content in painting exists – though I know almost nothing about it – so there’s no reason why the two systems could not be connected with each other, further enhancing the ability of individuals to locate paintings of the ships that interest them. As these systems are created in many different areas, we can find new ways of drawing out incredible amounts of new information and insights from this data.

The key point that I want to make is that it will take a lot of work, in many different areas, but it’s something that will be very useful throughout future historical research. I believe that the experiences I’ve had from Serials Solutions, such as using ONIX for Serials, to what we see now as a need for unique vessel identifiers, occurs in every field. And once you’re able to start growing on those resources, the opportunities for research and study expand dramatically. I very much look forward to that future.