

I am here today to talk to you about Portico and our completeness analysis tools.

In brief background, Portico preserves digital journals, books and collections of primary sources for the scholarly community, libraries, and publishers.

I am going to talk about journals and completeness today. Books have a different set of problems and collections of primary sources their own set of problems and I'll come back and talk about those another day.





PORTICO

Amy Kirchhoff
Archive Service Product Manager

Amy.Kirchhoff@ithaka.org
@AmyPlusFour



Before we begin, let me talk a bit about how we get content.

At Portico, we get content directly from the publishers, their platforms, and sometimes even their prepress services.

It comes to us packaged up in all sorts of different ways.



If we are lucky, the package of content sent to us has a packing label on it. Sometimes we can identify vol/iss information in the packaging, but not always and not reliably.



The thing is, we don't just get one package. We have nearly 30,000 journals and we process between 2 and 17 million articles a years. On average in 2016 we have received 31,000 packages a day.



At these rates, we don't much look at these packages when they come in. The packages of content land on our door (or FTP site) and we shove them onto a conveyor belt of processing that migrates, normalizes, and repackages the content and puts it in the archive. A whole lot goes into the tools along that conveyor belt and what we do when content is so problematic it gums up the works, but that's another presentation.

Completeness

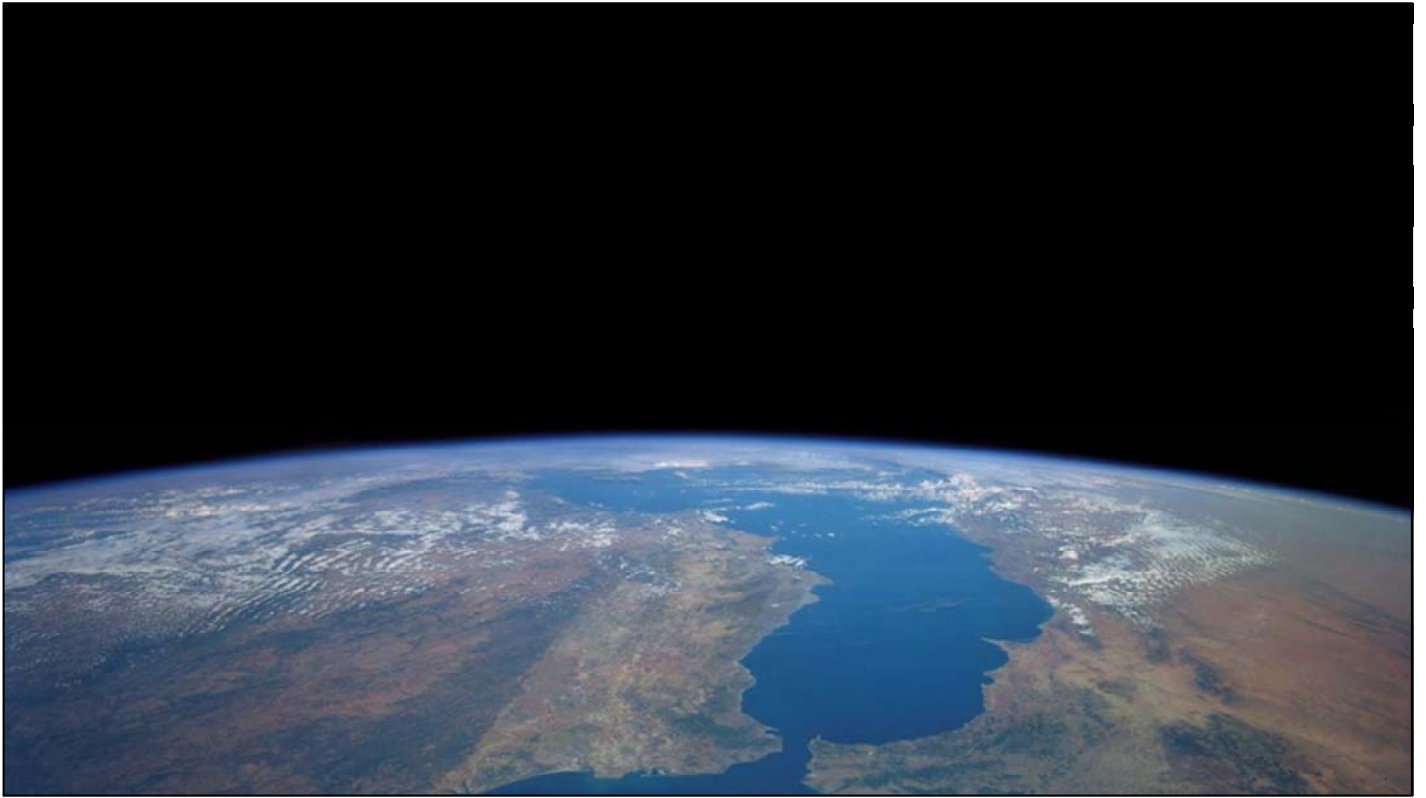
We are going to talk about what I call completeness. As content moves through the conveyor belt, we check that it is well formed and that we aren't missing any files and so forth. At Portico, we preserve the scholarly record as it was published or at least, how it was sent to us. ☺ We need at least a publication year and something to let us identify the journal (and we've been known to grant exemptions to both of those requirements). At the point that we put the article into the Portico archive we pull out some MD – what we can – out of the articles and put it into an Oracle DB for easy access.

So, we've placed all this content in our archive ... now we have to think about ... what do we have and what are we missing?

Most of the time we can answer the question: what do we have?

Answering the question of: What should we have and don't?

Is very difficult. There isn't data out in the wild to tell us what we should have at the volume, issue, or article level.



We do a bunch of things to try and assess completeness.

We can do a very high analysis, looking for patterns in the data.

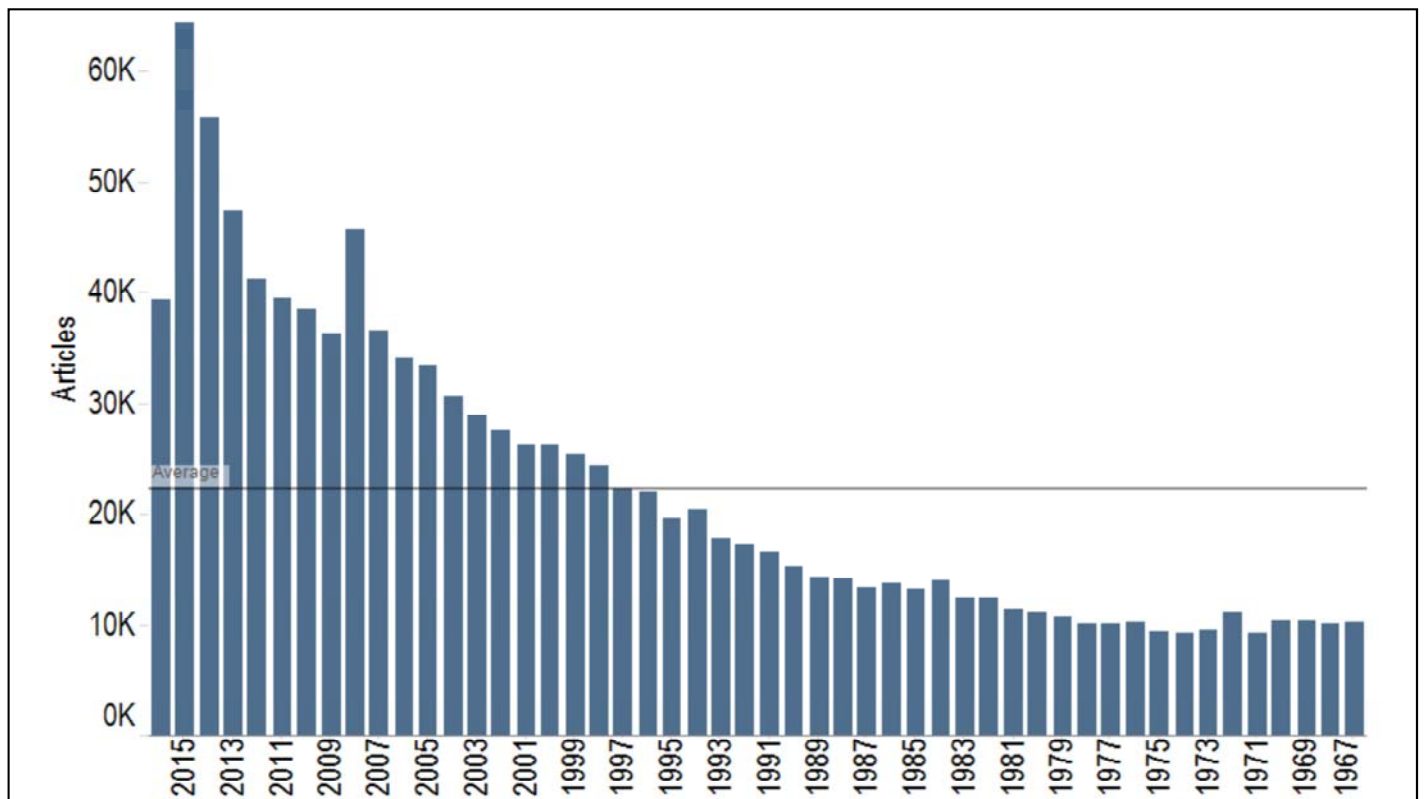
The easiest way.



To do this we use Tableau which is really cool data visualization software. Very powerful. Not easy, but very powerful. I have become a big groupie over the past couple of years.

We also use some scripting and Excel reports, but I'm not going to talk about that as I am sure you are all familiar with scripting to count and average things and reporting those results out in Excel. 😊

Tableau on the other hand lets us visualize the data ...



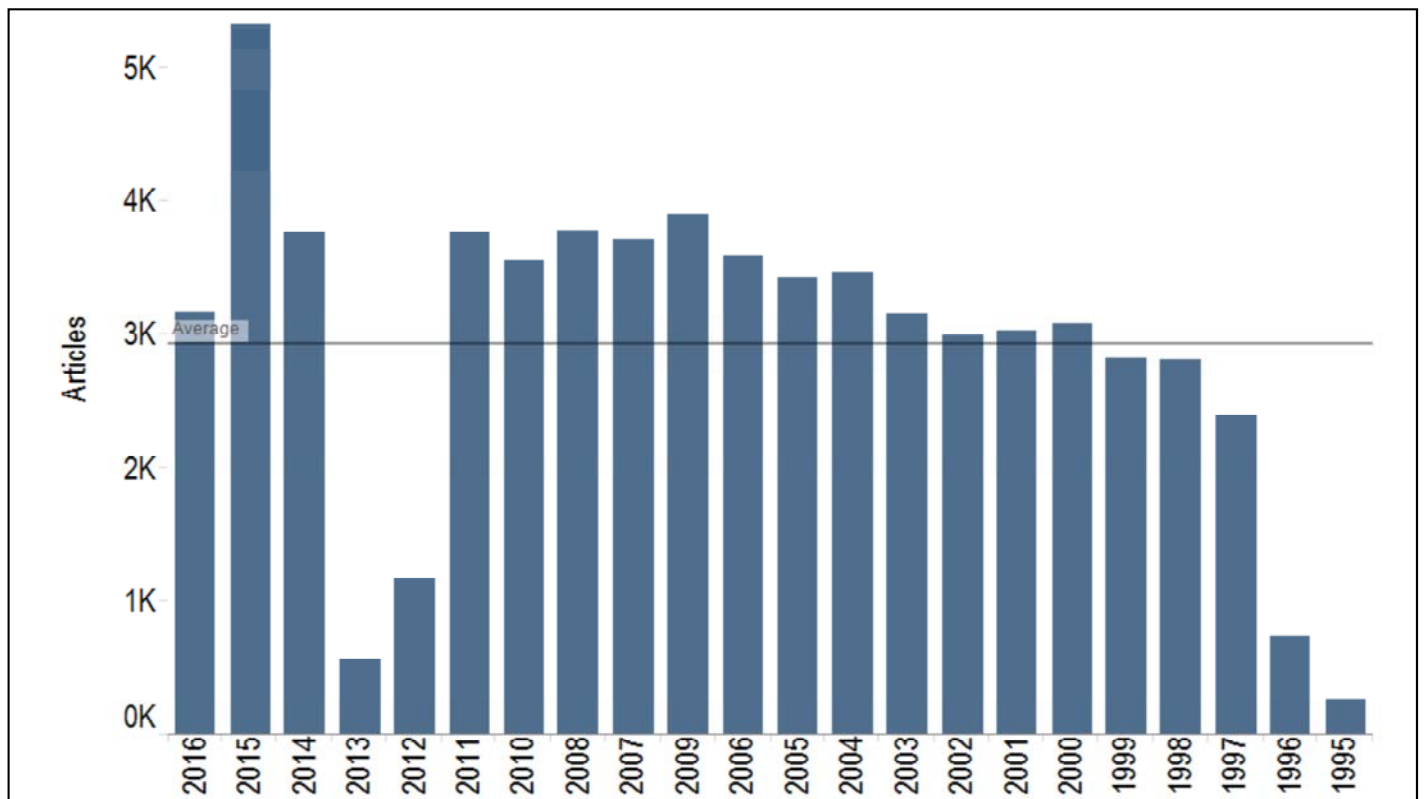
So, here is what I would consider a pretty normal pattern for a publisher.

What we have is a chart counting articles by publication year for a publisher. The X axis is year of publication and the Y axis is number of articles.

This is a pretty normal pattern, a slow increase in articles published every year for about 50 years, with some recent jumps. Another pretty common pattern is to see a peak about 5 years ago and a bit of a slow decline in recent years.

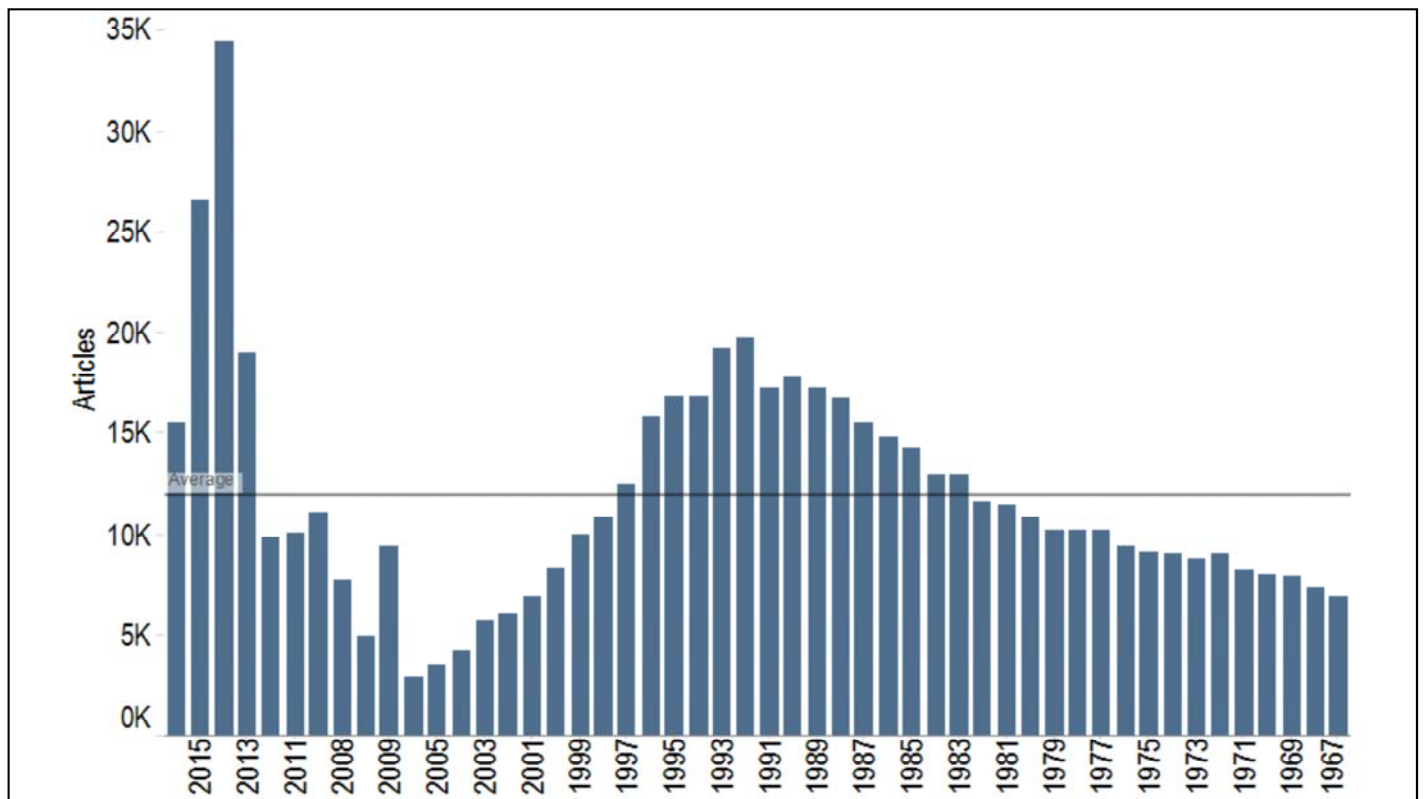
So, given that this is a normal pattern ...

ACS - normal



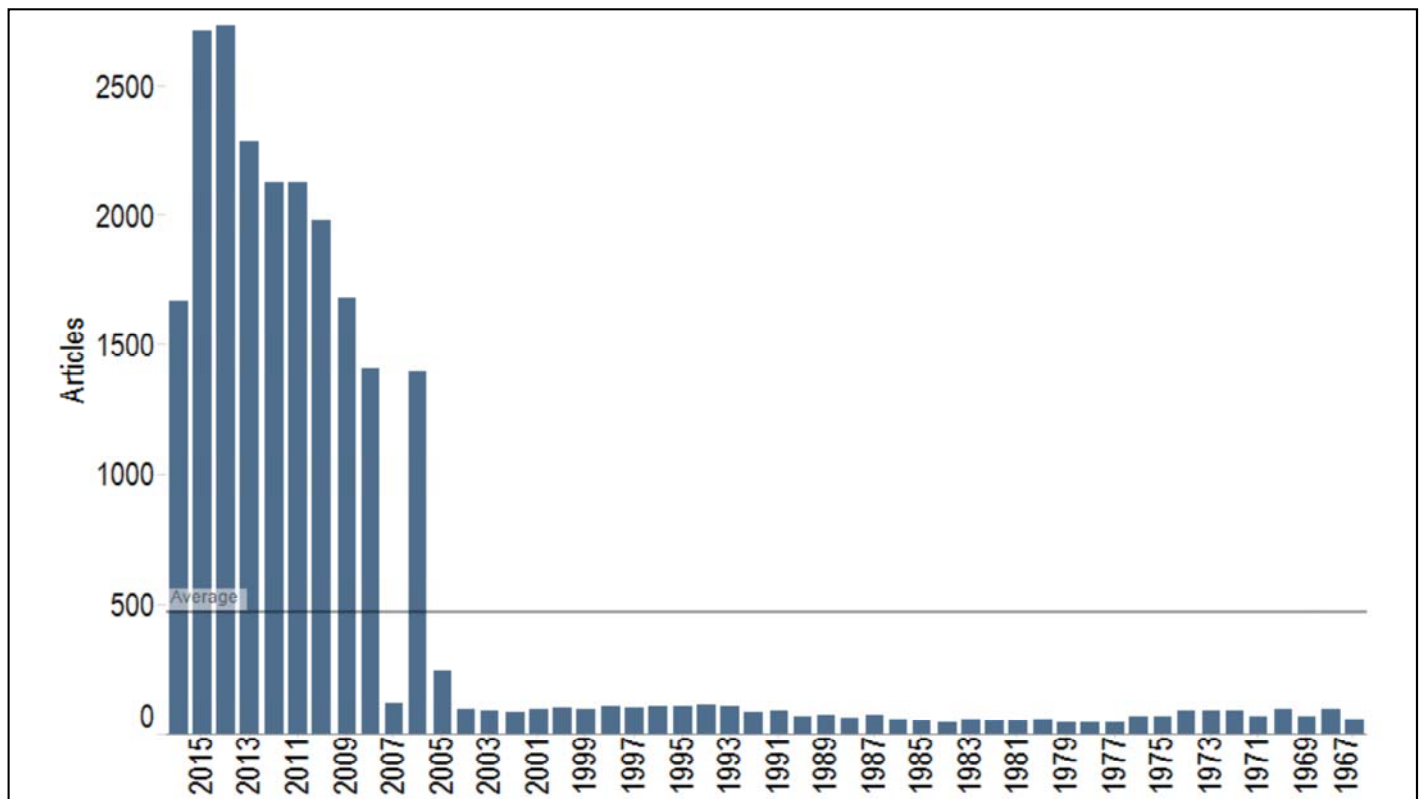
Who can spot the problem with this publisher?

AAS



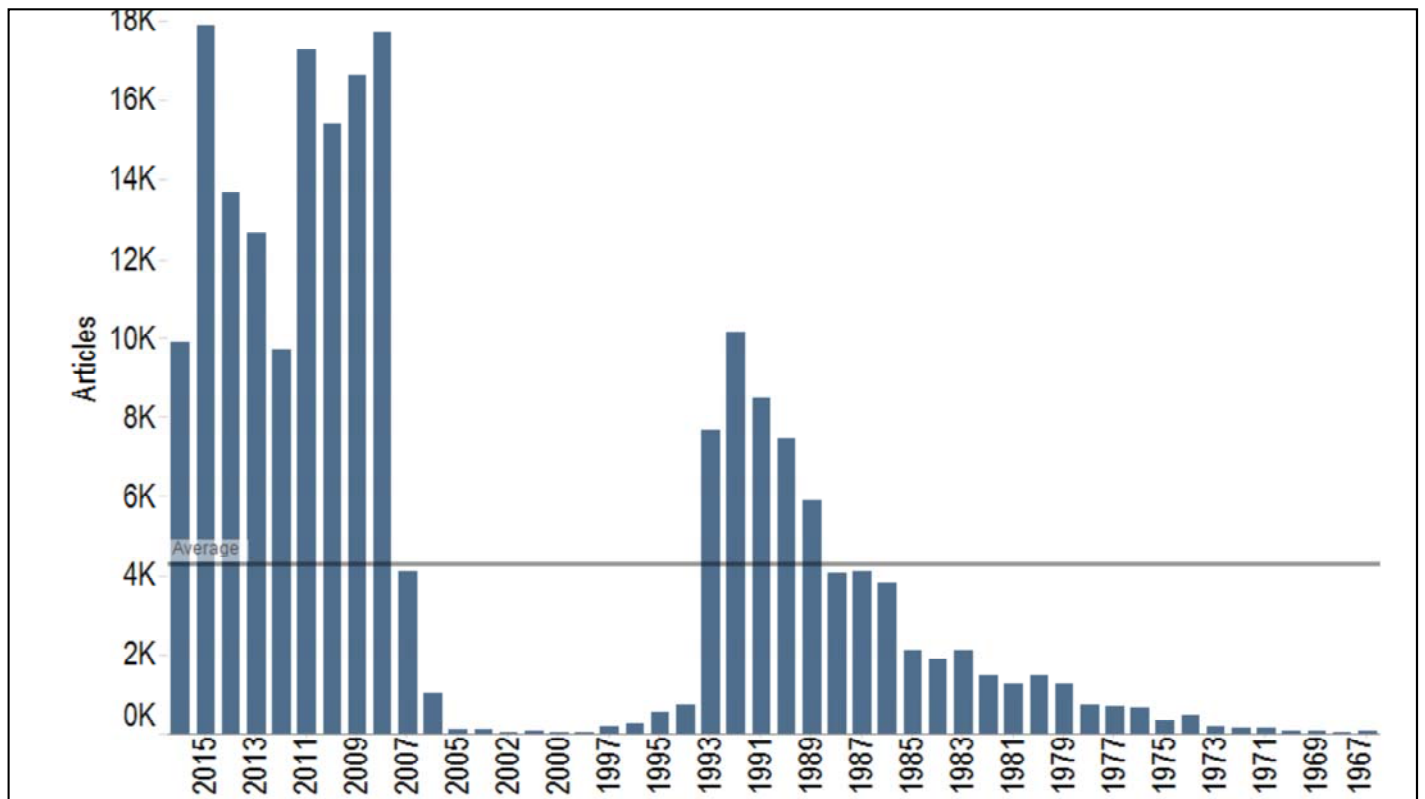
And, this publisher?

CUP



And this publisher?

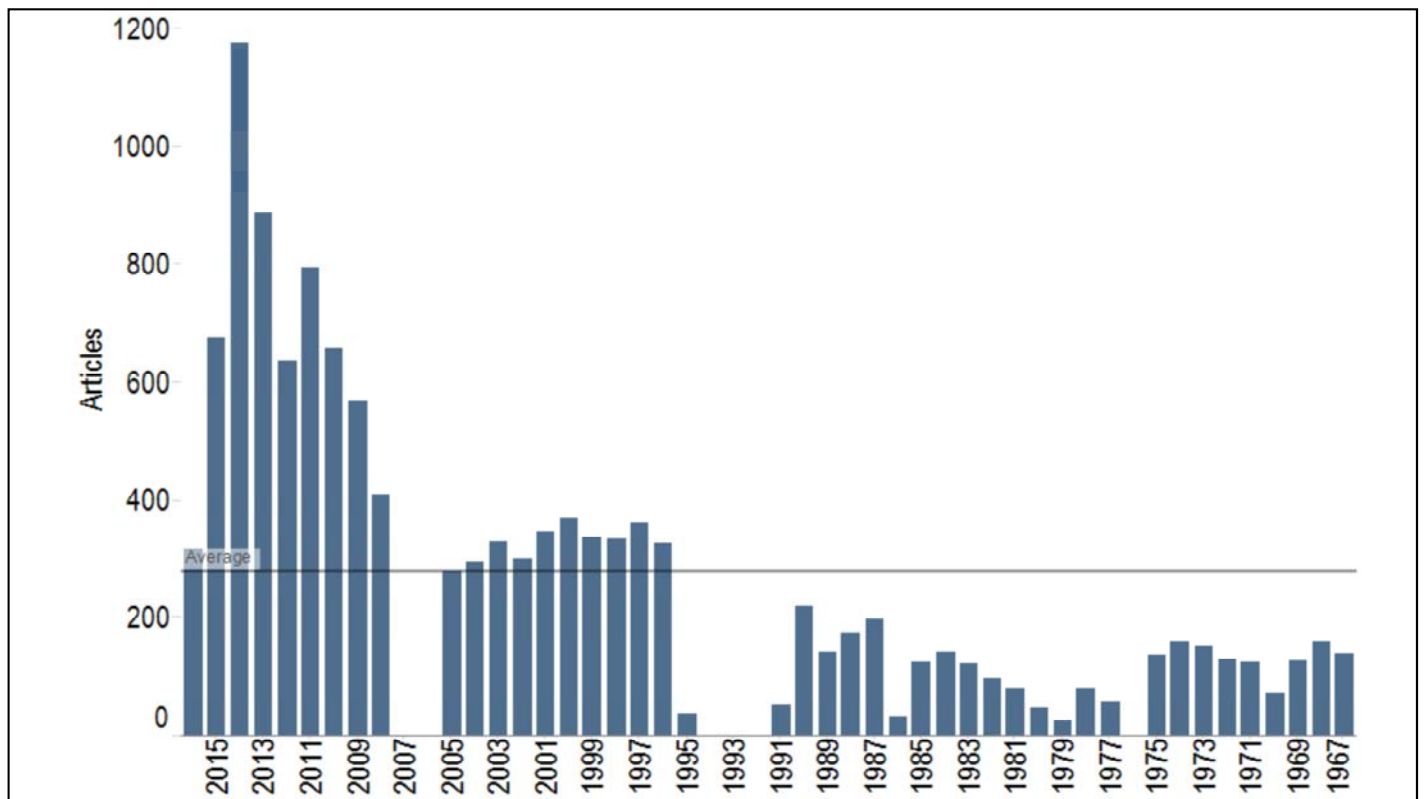
Duke



And here?

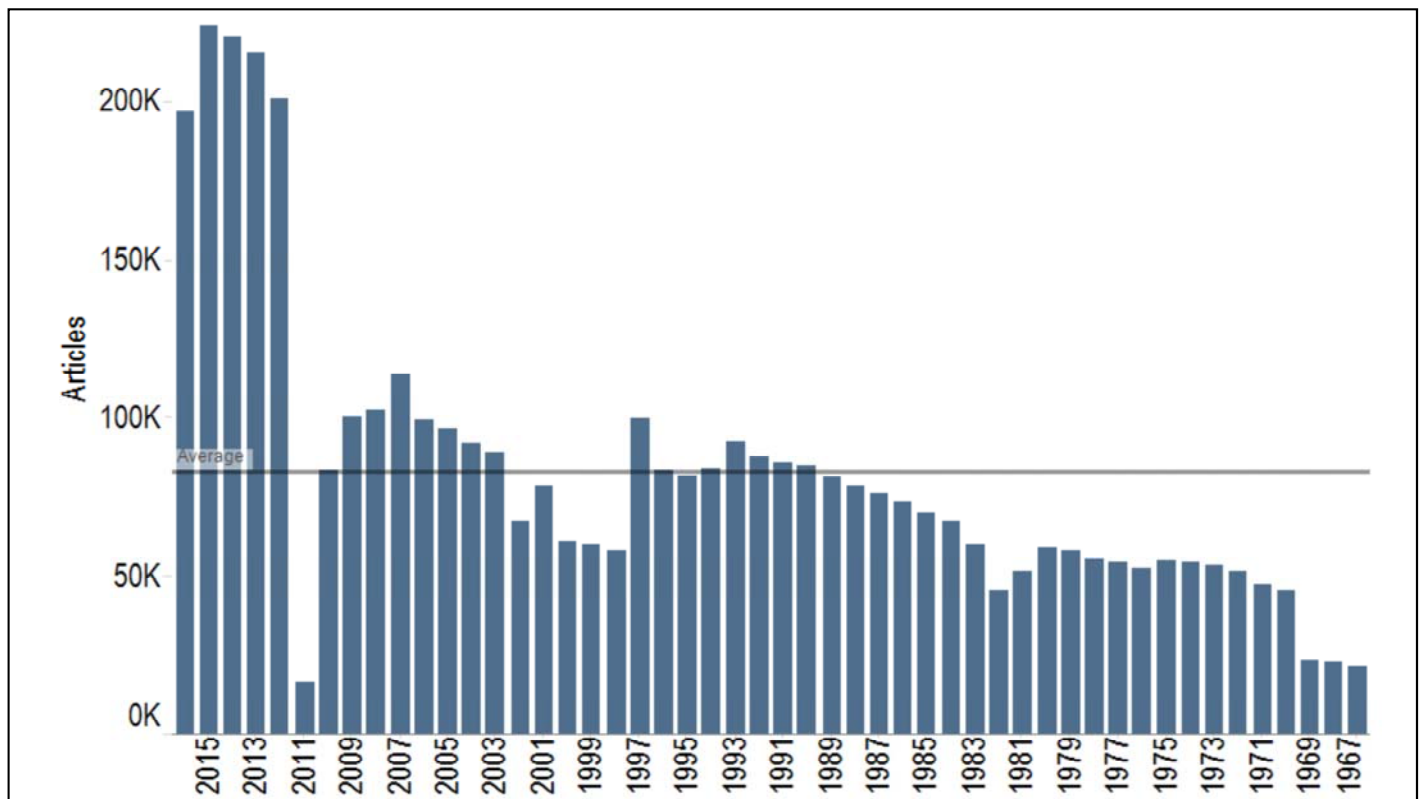
--

SPIE



And here?

JBJS-B



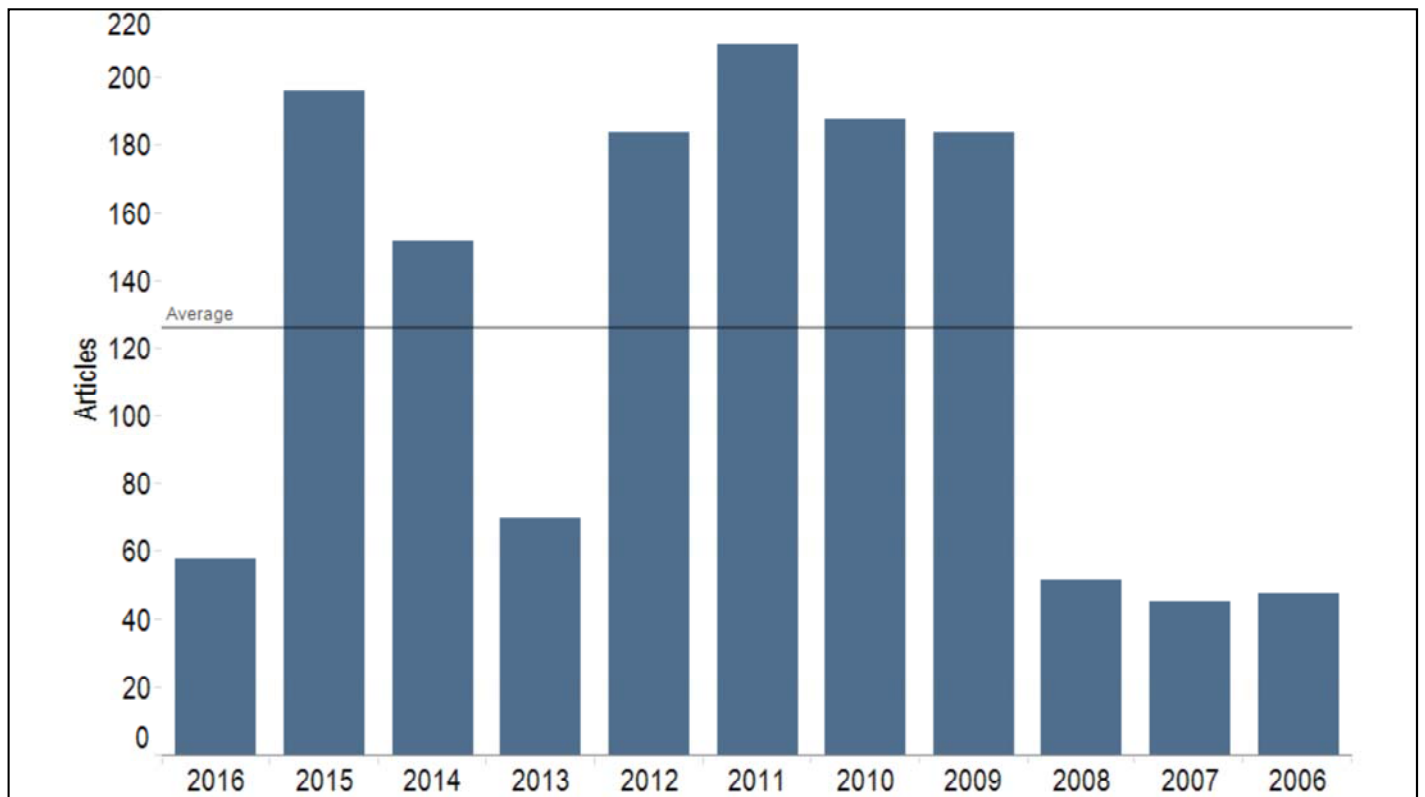
And this publisher?

(Please note we are addressing all of these problems and working on them with the publishers ... in fact I wouldn't be surprised if in the past 5 days since I grabbed this screen shot, this one has improved, I think we actually have this content on hand.)

When we see a pattern like this, we reach out to the publisher and ask them to deliver the swaths of missing content to us. If we are lucky, it comes to us in packages our code already understands and can manage. Unfortunately, sometimes this content that we request outside of the regular flow comes to us packaged differently.

--

Wiley



How about this publisher?

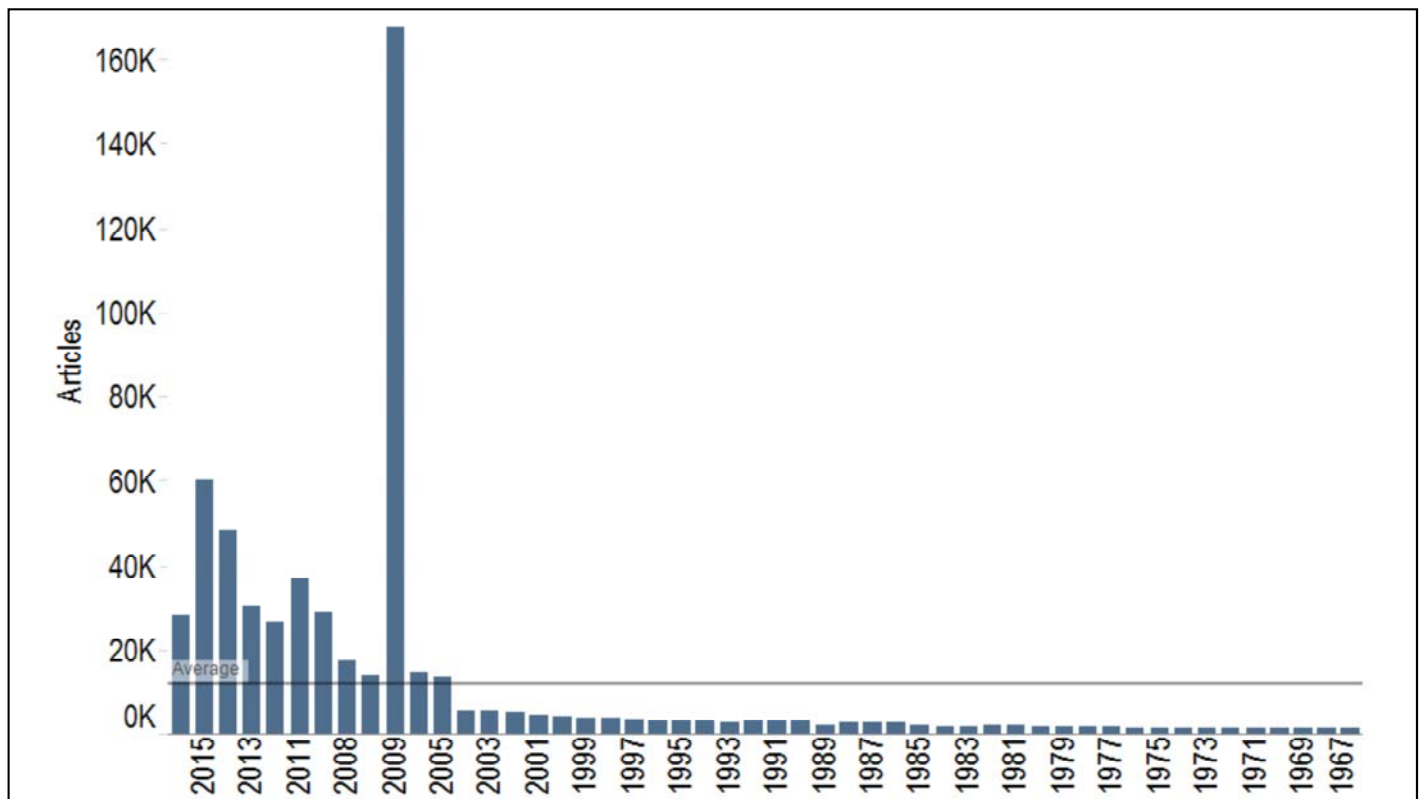
This one is actually a bit of a trick question.

This is a small publisher. I don't know how well the numbers work on the screen, but this publisher maxes out at about 215 articles a year, with earlier years about a quarter of that amount.

So, is 2013 actually a problem? To be honest, I don't know. I haven't looked into it yet. ☺ This is one where I would not be comfortable reaching out to the publisher without confirming that we are actually missing content.

--

Vilnius – small pub



Who can spot “the problem”?

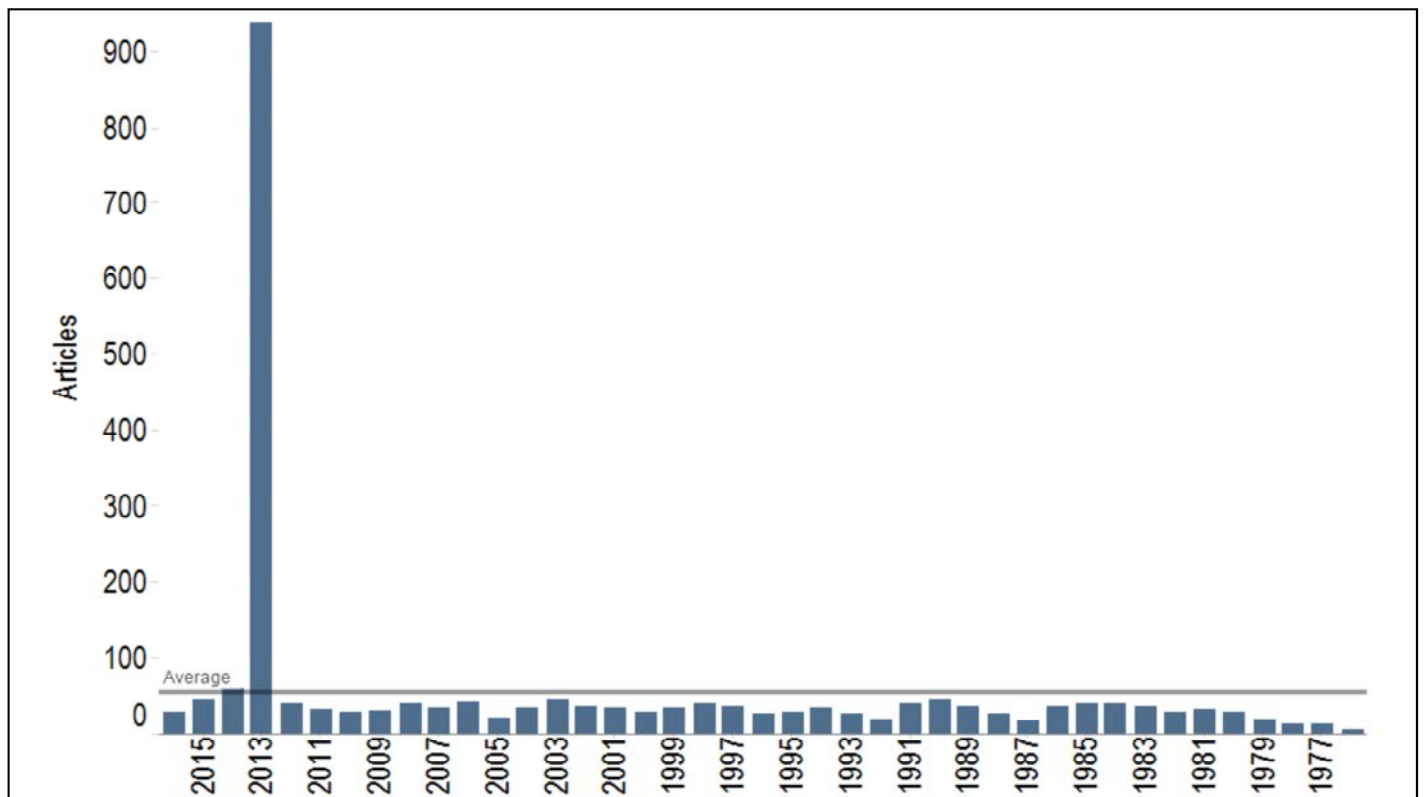
This slide and the next are just to show how MD can sometimes be difficult to interpret.

it turns out, this isn't exactly a problem from a completeness point of view. This particular publisher send us a whole bunch of digitized back content ... ummm, probably about 140,000 articles worth. Sadly, we use the copyright year as our publication date with this publisher and all those copyright years were 2009 – I assume the year the content was digitized.

Sigh.

--

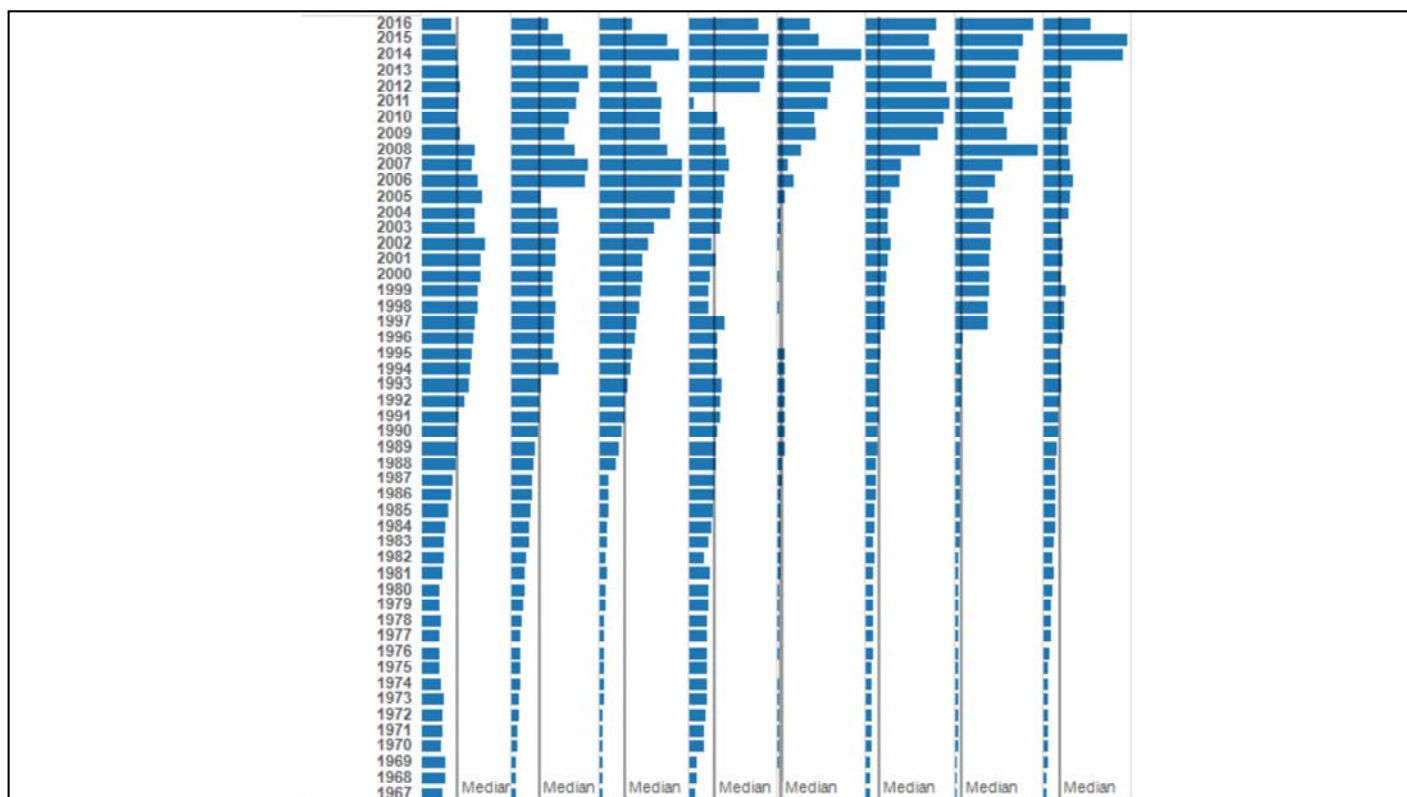
De Gruyter



And here is another one. I just discovered this one while looking for screen shots for this presentation, but I'd guess this is a similar situation. Though, it is possible that we just got the 2013 content 10 times. 😊

I will eventually look into this, but it isn't urgent, as it doesn't look like we are missing swaths of content.

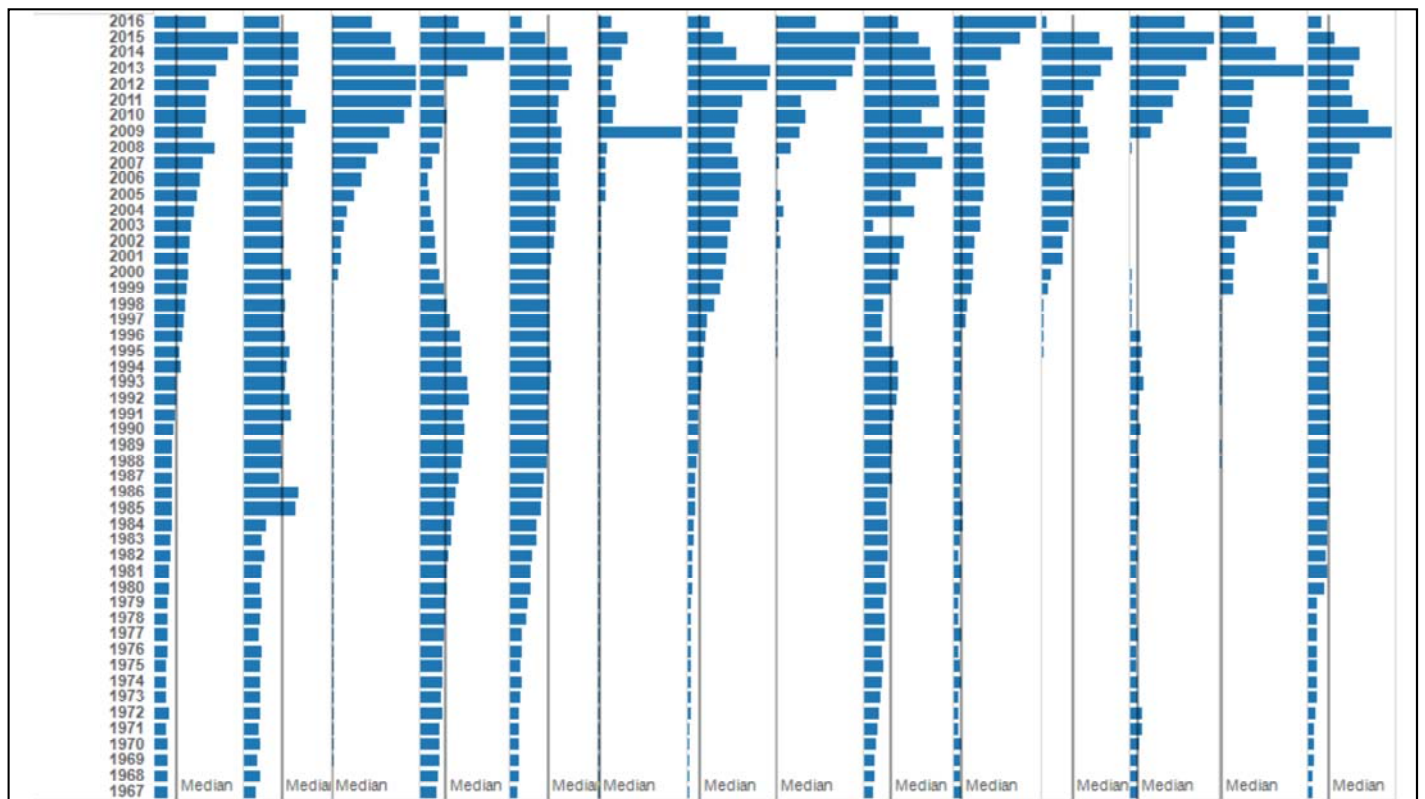
New Prairie Press



Now, I don't want to flip people out that our archive is half empty. 😊 I cherry picked some odd looking publishers for purposes of display. To even things out, here are our largest 8 publishers, they all publish between 100,000 and 2 million articles a year on average. As you can see, things are pretty steady. 😊

--

100,000 to 2 million articles a year on average

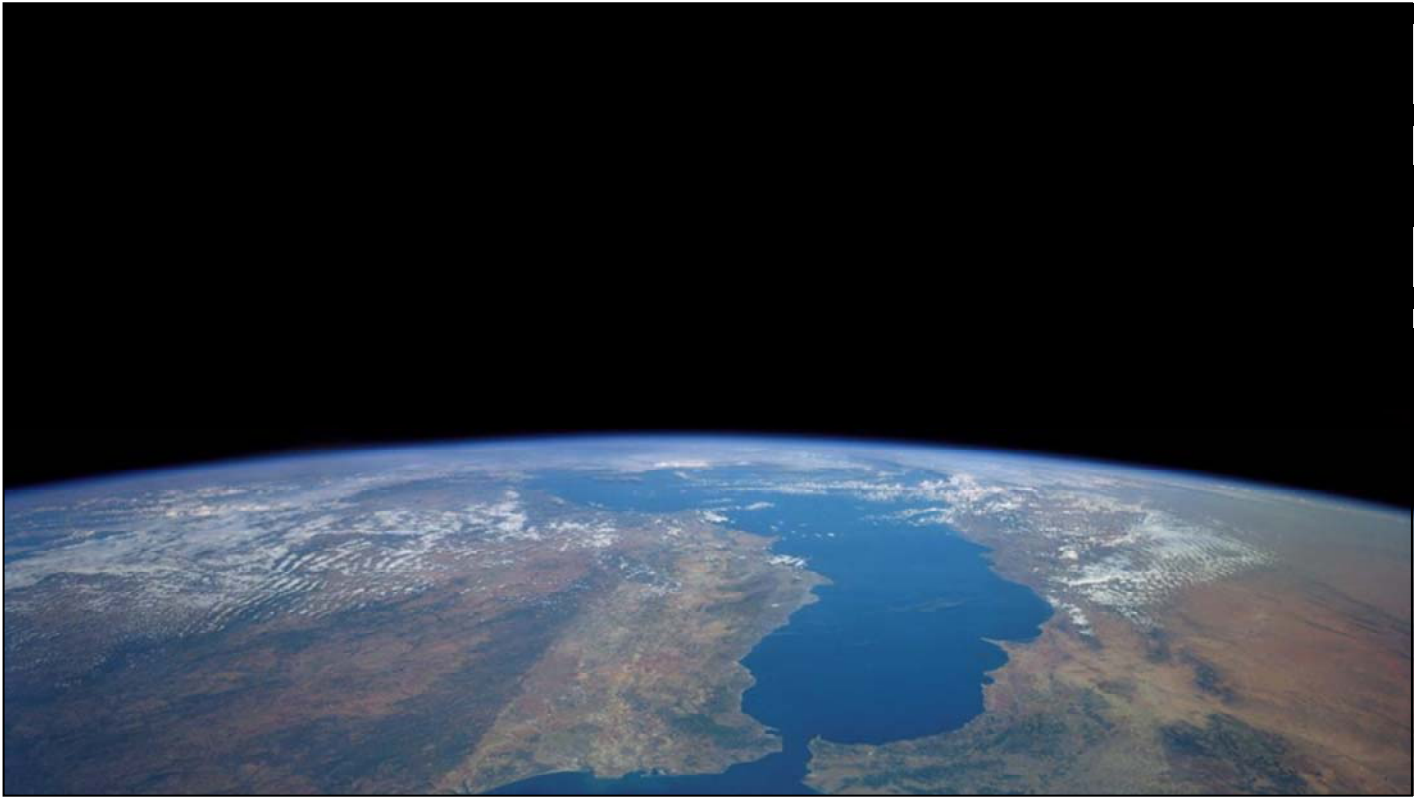


And here are the next 14.

These publish between 25,000 and 99,999 articles per year.

Mostly steady (though a couple of these showed up in my earlier slides).

To put this in context, we are working with about 300 publishers and well over 250 of them publish less than 25,000 articles a year.



We do a bunch of things to try and assess completeness.

We can do a very high analysis, looking for patterns in the data.



... let's talk in the weeds, next.

6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,8,

And of course, the weeds are full of numbers.

When we are missing huge swaths of content, we are very comfortable going back to the publisher and asking for it without any details ... please send us all of 2006, it seems to be missing from the archive.

However, very often we are missing an issue here or an issue there.
an article here and an article there.

We'd really like to track these down and it is really difficult.

We have made some progress, it isn't perfect, but it is progress.

Describe it here ...

Analyze what we have to guess the publication pattern.

Analyze what we have to identify things that are missing or incomplete.

Abdominal Imaging -

http://audit.portico.org/Portico/#!/journalLOVIView/cs=ISSN_09428925?ct=E-Journal%20Content

1997 - 2015

Completeness		Years Preserved: 1997-2015 (v.22-40)					
as of Oct 28, 2016		Portico Holdings: 1997 - v. 22 (1-6), 1998 - v. 23 (1-6), 1999 - v. 24 (1-6), 2000 - v. 25 (1-2, 4-6), 2001 - v. 26 (1-6), 2002 - v. 27 (1-3), 2003 - v. 28 (1-3, 5-6), 2004 - v. 29 (1-6), 2004/2005 - v. 30 (1-6), 2006 - v. 31 (1-6), 2007 - v. 32 (1-6), 2008 - v. 33 (1-6), 2009 - v. 34 (1-6), 2010 - v. 35 (1-6), 2011 - v. 36 (1-6), 2012 - v. 37 (1-6), 2013 - v. 38 (1-6), 2014 - v. 39 (1-5), 2015 - v. 40 (1-2, 6-8)					
		Publication Pattern: 6,6,6,6,6,6,6,6,6,6,6,6,6,6,6,8,					
Status		Journal incomplete					
	1	2	3	4	5	6	7
v.22 (1997)	n.1: Unknown Pub Date: 1997-07-01 Articles: 29	n.2: Complete Pub Date: 1997-02-01 Articles: 45	n.3: Complete Pub Date: 1997-03-01 Articles: 59	n.4: Complete Pub Date: 1997-07-01 Articles: 44	n.5: Complete Pub Date: 1997-09-01 Articles: 48	n.6: Incomplete Pub Date: 1997-11-01 Articles: 37	No more issues
v.23 (1998)	n.1: Incomplete Pub Date: 1998-01-01 Articles: 86	n.2: Complete Pub Date: 1998-03-01 Articles: 57	n.3: Incomplete Pub Date: 1998-05-01 Articles: 44	n.4: Incomplete Pub Date: 1998-07-01 Articles: 66	n.5: Incomplete Pub Date: 1998-09-01 Articles: 50	n.6: Complete Pub Date: 1998-11-01 Articles: 54	No more issues
v.24 (1999)	n.1: Unknown Pub Date: 1999-01-01 Articles: 20	n.2: Complete Pub Date: 1999-03-01 Articles: 24	n.3: Complete Pub Date: 1999-05-01 Articles: 79	n.4: Complete Pub Date: 1999-07-01 Articles: 23	n.5: Complete Pub Date: 1999-09-01 Articles: 71	n.6: Complete Pub Date: 1999-11-01 Articles: 42	No more issues
v.25 (2000)	n.1: Incomplete Pub Date: 2000-01-01 Articles: 43	n.2: Complete Pub Date: 2000-03-01 Articles: 27	n.3: Missing Pub Date: 2000-03-01	n.4: Unknown Pub Date: 2000-03-01 Articles: 23	n.5: Complete Pub Date: 2000-09-01 Articles: 48	n.6: Complete Pub Date: 2000-11-01 Articles: 27	No more issues
v.26 (2001)	n.1: Complete Pub Date: 2001-	n.2: Incomplete Pub Date: 2001-	n.3: Incomplete Pub Date: 2001-	n.4: Incomplete Pub Date: 2001-	n.5: Incomplete Pub Date: 2001-	n.6: Complete Pub Date: 2001-	No more issues

With all of that data we can do things.

One thing we can do is make it available to our auditor librarians like this. We really want to be transparent, so for any publisher where we don't think our data looks too whack-a-doodle, we will display it in the audit interface like this. As you might expect, sometimes our analysis comes out with some crazy guesses (I'll show you examples) and in those instances although we keep the data for our own analysis, we don't present it to the auditors.

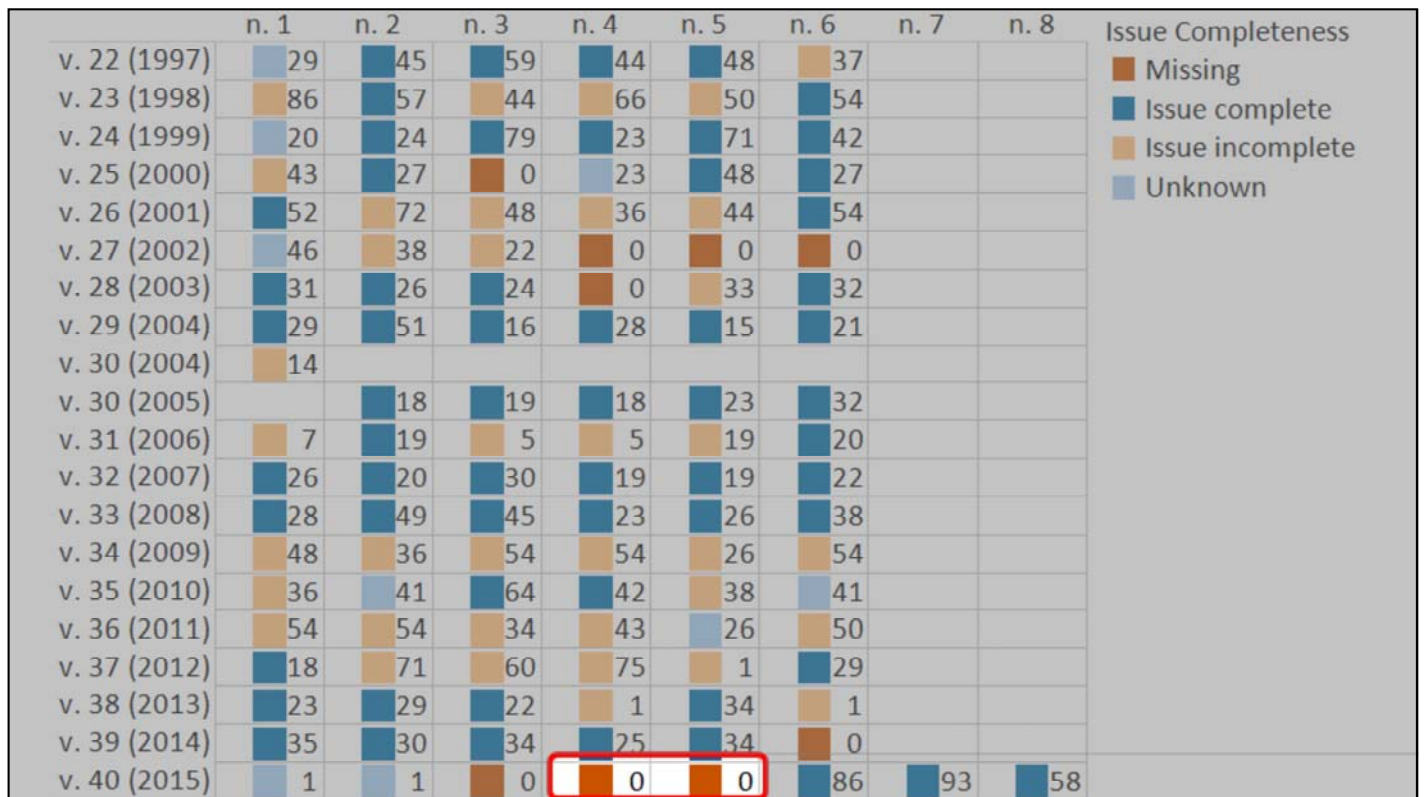
	n. 1	n. 2	n. 3	n. 4	n. 5	n. 6	n. 7	n. 8	Issue Completeness
v. 22 (1997)	29	45	59	44	48	37			Missing
v. 23 (1998)	86	57	44	66	50	54			Issue complete
v. 24 (1999)	20	24	79	23	71	42			Issue incomplete
v. 25 (2000)	43	27	0	23	48	27			Unknown
v. 26 (2001)	52	72	48	36	44	54			
v. 27 (2002)	46	38	22	0	0	0			
v. 28 (2003)	31	26	24	0	33	32			
v. 29 (2004)	29	51	16	28	15	21			
v. 30 (2004)	14								
v. 30 (2005)		18	19	18	23	32			
v. 31 (2006)	7	19	5	5	19	20			
v. 32 (2007)	26	20	30	19	19	22			
v. 33 (2008)	28	49	45	23	26	38			
v. 34 (2009)	48	36	54	54	26	54			
v. 35 (2010)	36	41	64	42	38	41			
v. 36 (2011)	54	54	34	43	26	50			
v. 37 (2012)	18	71	60	75	1	29			
v. 38 (2013)	23	29	22	1	34	1			
v. 39 (2014)	35	30	34	25	34	0			
v. 40 (2015)	1	1	0	0	0	86	93	58	

Here is that same journal, only I'm showing you the completeness data in a Tableau visualization. I am doing this just because I can fit the whole journal on a single skinny screen at once.

Now, you are probably thinking, this is great, so you can ask the publisher for all those incomplete and missing issues, right?

Unfortunately, know. We really need to be confident an issue is missing or incomplete before we go to the publisher. We are going to get the attention of a sliver of a person. Our publishers are working flat out to get their content to your users, they just don't have a lot of cycles to give us and we need to take advantage of those cycles wisely. We don't want to ask them for content we might have or doesn't actually exist.

Let's look at two issues specifically...



Here is that same journal, only I'm showing you the completeness data in a Tableau visualization. I am doing this just because I can fit the whole journal on a single skinny screen at once.

Now, you are probably thinking, this is great, so you can ask the publisher for all those incomplete and missing issues, right?

Unfortunately, know. We really need to be confident an issue is missing or incomplete before we go to the publisher. We are going to get the attention of a sliver of a person. Our publishers are working flat out to get their content to your users, they just don't have a lot of cycles to give us and we need to take advantage of those cycles wisely. We don't want to ask them for content we might have or doesn't actually exist.

Let's look at these two issues specifically.



Remember how I said we were feeling lucky if we got content with a packing label ...

batches

Summary

Batch List (6)

Search

Clear Search

%00261%2015%

Filter(*)

Filter()

ALL

ALL

Batch Name	Lifecycle State	Last Activity	Executor Queue	Queue Priority	Hold
ftp_PUB_15-03-25_05-14-21_JOU=00261_VOL=2015.40_ISU=4	Problem	Resolve File References		Normal	No
ftp_PUB_15-03-25_05-14-22_JOU=00261_VOL=2015.40_ISU=4	Problem	Assign Content Set Name		Normal	No
ftp_PUB_15-05-30_05-17-16_JOU=00261_VOL=2015.40_ISU=5	Problem	Resolve File References		Normal	No
ftp_PUB_15-05-30_05-17-17_JOU=00261_VOL=2015.40_ISU=5	Problem	Resolve File References		Normal	No
ftp_PUB_15-05-30_05-17-18_JOU=00261_VOL=2015.40_ISU=5	Problem	Resolve File References		Normal	No
ftp_PUB_15-05-30_05-17-19_JOU=00261_VOL=2015.40_ISU=5	Problem	Resolve File References		Normal	No

Well, we are in luck. <click>

Ingest v 2.0.2

atches

Summary Batch List (6) Search Clear Search

%00261%2015% Filter(*) Filter() ALL ALL

Batch Name		Lifecycle State	Last Activity	Executor Queue	Queue Priority	Hold
ftp_PUB_15-03-25_05-14-21_JOU=00261_VOL=2015.40_ISU=4		Problem	Resolve File References		Normal	No
ftp_PUB_15-03-25_05-14-22_JOU=00261_VOL=2015.40_ISU=4		Problem	Assign Content Set Name		Normal	No
ftp_PUB_15-05-30_05-17-16_JOU=00261_VOL=2015.40_ISU=5		Problem	Resolve File References		Normal	No
ftp_PUB_15-05-30_05-17-17_JOU=00261_VOL=2015.40_ISU=5		Problem	Resolve File References		Normal	No
ftp_PUB_15-05-30_05-17-18_JOU=00261_VOL=2015.40_ISU=5		Problem	Resolve File References		Normal	No
ftp_PUB_15-05-30_05-17-19_JOU=00261_VOL=2015.40_ISU=5		Problem	Resolve File References		Normal	No

We have a label. ☺ And there are at least some of the articles in issue 4 and 5. This is a screen shot of our content processing system. We call it ConPrep and it is the primary piece of that conveyor belt.

For whatever reason, this set of packages (or batches in our parlance), didn't make it into the archive. We will work on them and either get them into the archive by tweaking our tools or tweaking the data or we'll go back to the publisher and ask them to fix whatever is wrong. It would not be helpful, though, for us to go back to the publisher and ask for these two issues as "missing" this week and then in a month go back to the publisher and ask for them to fix these two issues.

On to some more examples.

	n. 1	n. 1-4	n. 2	n. 3	n. 4	n. 5	n. 6	n. 7	n. 8	n. 9	n. 10	n. 11	n. 12	n. 13	Issue Completeness
v. 13 (2015)	0		0	0	0	0	0	0	0	0	0	0	0	35	Unknown
v. 14 (2015)	0		0	0	0	0	0	0	0	0	0	0	0	0	Missing
v. 15 (2015)	0		0	0	0	0	0	0	0	0	0	0	0	0	Missing
v. 16 (2015)	0		0	0	0	0	0	0	0	0	0	0	0	0	Missing
v. 17 (2015)	0		0	0	0	0	0	0	0	0	0	0	0	0	Missing
v. 18 (2015)	0		0	0	0	0	0	0	0	0	0	0	0	0	Missing
v. 19 (2015)	0		0	0	0	0	0	0	0	0	0	0	0	0	Missing
v. 20 (2015)	0		0	0	0	0	0	0	0	0	0	0	0	0	Missing
v. 21 (2015)	0		0	0	0	0	0	0	0	0	0	0	0	0	Missing
v. 22 (2015)	0		0	0	0	0	0	0	0	0	0	0	0	0	Missing
v. 23 (2015)	0		0	0	0	0	0	0	0	0	0	0	0	0	Missing
v. 24 (2015)	0		0	0	0	0	0	0	0	0	0	0	0	0	Missing
v. 25 (2015)	0		0	0	0	0	0	0	0	0	0	0	0	0	Missing
v. 26 (2015)	0		0	0	0	0	0	0	0	0	0	0	0	0	Missing
v. 27 (2015)	0		0	0	0	0	0	0	0	0	0	0	0	0	Missing
v. 28 (2015)	0		0	0	0	0	0	0	0	0	0	0	0	0	Missing
v. 29 (2015)	0		0	0	0	0	0	0	0	0	0	0	0	0	Missing
v. 30 (2015)	0		0	0	0	0	0	0	0	0	0	0	0	0	Missing
v. 31 (2015)	0		0	0	0	0	0	0	0	0	0	0	0	0	Missing
v. 32 (2015)	0		0	0	0	0	0	0	0	0	0	0	0	0	Missing
v. 33 (2015)	0	31				0	0	0	0	0	0	0	0	0	Missing
v. 46 (2015)	28														Issue complete
v. 96 (2014)			47												Issue complete
v. 97 (2015)	0														Missing
v. 97 (2016)			39												Unknown
v. 98 (2016)	38		0												Missing

































































Ah, love this one. Here we can see that in the more recent volumes, we have v. 33 (1-4), v. 46, v. 96 (2), v. 92 (2), v. 98 (2).

And look, we have v. 13 (13).

And of course with that v. 13 (13) in there, we think we are missing v. 14 (1) through v. 33 (13).

Dollars to donuts, that is a data problem and the publisher swapped the numbers. I'd bet good money that that issue should be v. 31.

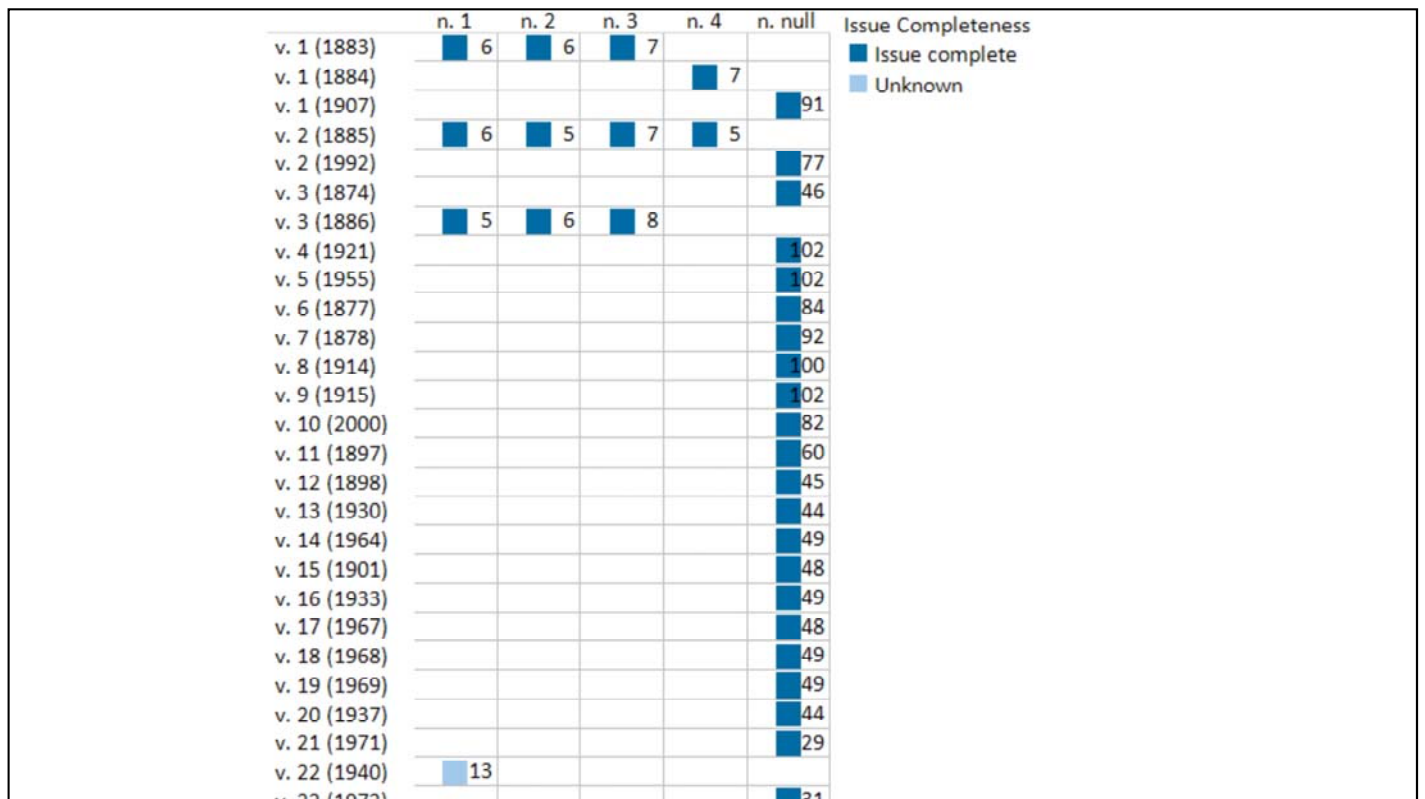
If we went back to the publisher and asked for v. 14 (1) through v. 33 (13) they'd look at us like idiots, because most of those issues probably don't exist.

	n. 1	n. 2	n. 3	Issue Completeness
v. 25 (1996)	 1			
v. 26 (1997)	 0	 0	 0	 Missing
v. 27 (1998)	 0	 0	 0	 Issue complete
v. 28 (1999)	 0	 0	 0	 Unknown
v. 29 (2000)	 0	 0	 0	
v. 30 (2001)	 0	 0	 0	
v. 31 (2002)	 0	 0	 0	
v. 32 (2003)	 0	 0	 0	
v. 33 (2004)	 0	 0	 0	
v. 34 (2005)	 0	 0	 0	
v. 35 (2006)	 0	 0	 0	
v. 36 (2007)	 0	 0	 0	
v. 37 (2008)	 0	 0	 0	
v. 38 (2009)	 0	 0	 0	
v. 39 (2010)	 0	 0	 0	
v. 40 (2011)	 0	 0	 0	
v. 41 (2012)	 0	 0	 0	
v. 42 (2013)	 0	 0	 0	
v. 43 (2014)	 0	 0	 0	
v. 44 (2015)	 0	 0	 0	
v. 45 (2016)	 7	 6	 9	

Ooo, here is a pattern we see alot.

This kind of pattern happens for publishers where we get the current content very regularly, but have not yet gotten a bulk deposit of back content.

And then, the publisher corrects something in an old issue. Their system sends us anything updated, so we get that old issue. And now our calculation shows us missing 20 volumes of content. In this case, we probably are missing them, but we prefer to go at this type of problem from the high level direction. We are working with all of our publishers to get back content and that is best handled as a one-off project and not bit-by-bit.



On the face of it, this looks pretty good. Lots of blue in there.

Look closely at the dates, though. We have a v. 1 in 1883 and 1884 and then again in 1902.

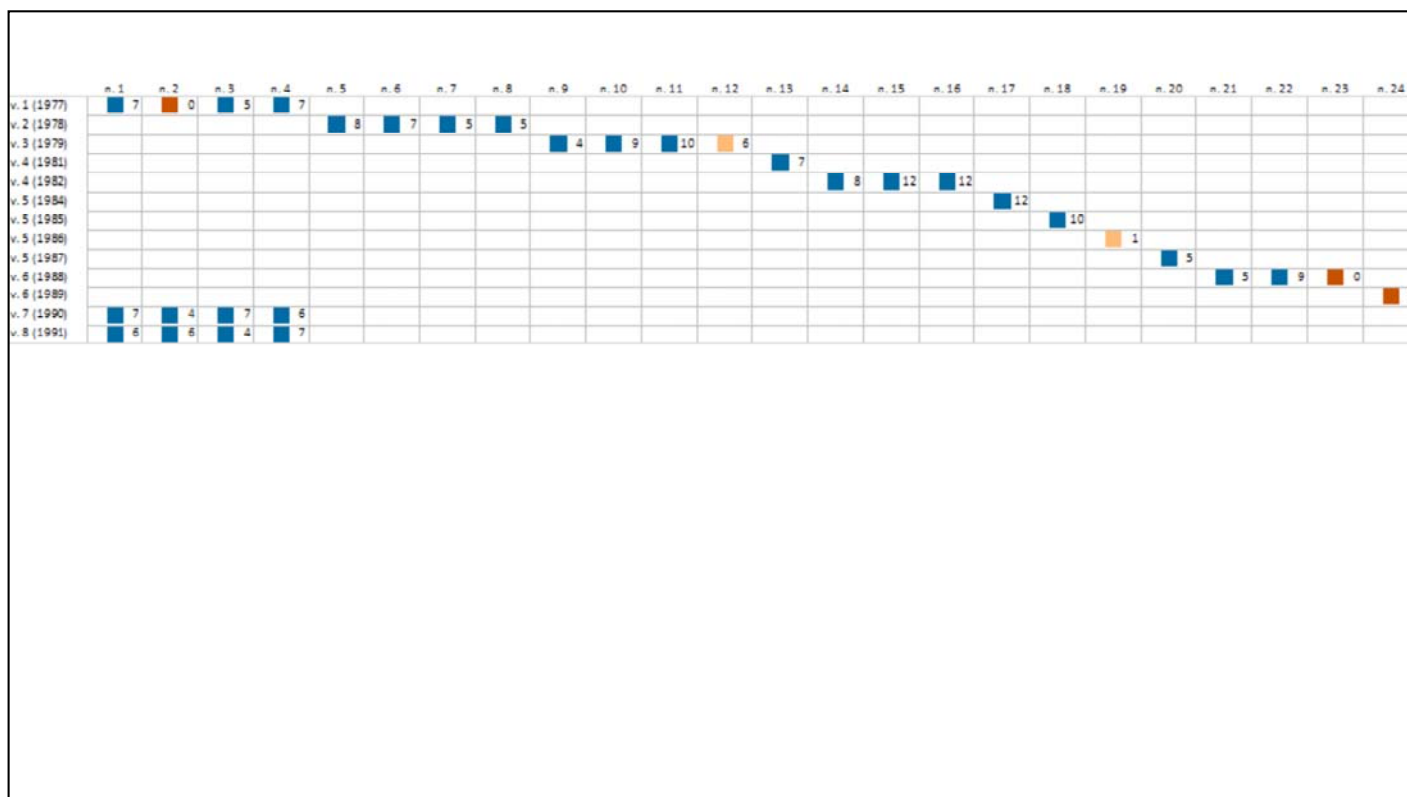
A v. 2 in 1885 ad 1992.

Any guesses as to the problem?

I selected the images blind. Kind of randomly going through a list of Portico title IDs. Which is to say I can't know for sure, but my guess is this is title with 2 or more series. MD series information comes to us very erratically. Usually it is just a part of the volume information. Sometimes it is a separate tag. Either way, we don't have a good way to capture it in our preservation metadata.

Thus, we can get some skewed analysis.

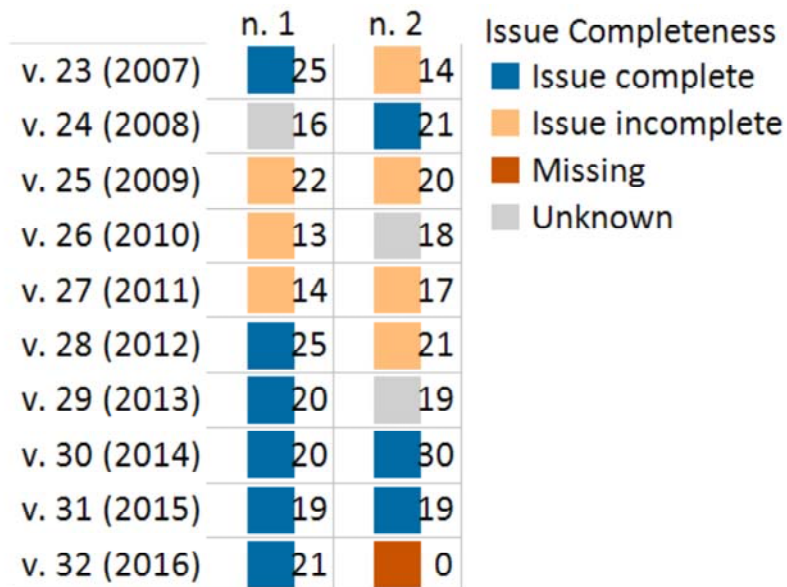
Fortunately, this is really limited to just two publishers.



Here is another interesting pattern.

I was excited to see how well the analyser did with this. The publisher had been publishing sequentially numbered issues and then swapped over to more regular issue numbering. There was a time when the Analyser would have anticipated a whole lot of missing issues in this title, but it has been modified to understand about sequential numbering and swapping to regular numbering and account for it.
















































































As you might imagine, we could spend all our time improving the analyser. Any serial librarians in the room? Yeah, you'd understand.



Here is a pretty plain one.

From looking at this, I think it likely that those issues we identified as incomplete are actually complete. It looks like the problem is that they aren't hitting the average, but I suspect that a rolling average might be a better assessment tool here than a flat average.

Something to consider.

	n. 1	n. 2	n. 3	n. 4	n. 5	n. 6	Issue Completeness
v. 7 (1972)	 14	 12	 17	 14	 12	 13	 Issue complete
v. 8 (1973)	 15	 15	 13	 8	 9	 15	
v. 9 (1973)	 14	 10	 12	 11	 13	 17	
v. 10 (1974)	 9	 13	 9	 16			
v. 11 (1974)	 12	 15	 16	 14			
v. 12 (1974)	 14	 12	 13	 19			
v. 13 (1975)	 14	 16	 13	 19			
v. 14 (1975)	 11	 12	 15	 11			
v. 15 (1975)	 12	 15	 10	 15			
v. 16 (1976)	 12	 13	 12	 15			
v. 17 (1976)	 14	 16	 12	 17			
v. 18 (1976)	 7	 15	 15	 11			
v. 19 (1977)	 13	 11	 18	 14			
v. 20 (1977)	 18	 12	 10	 18			
v. 21 (1977)	 15	 10	 14	 10			
v. 22 (1978)	 14	 14	 14	 17			
v. 23 (1978)	 15	 15	 13	 16			
v. 24 (1978)	 14	 12	 12	 19			

Gotta love all the blue.

We have determined that for some of our publishers, we can automatically build URLs to their journal issues and then count the number of articles on the issue table of contents at the publishers website.

This is probably one of those publishers. When we can do that, we can confirm the existence of an issue and completeness of an issue. We can't do it for all publishers, but we can for some.



PORTICO

Amy Kirchhoff
Archive Service Product Manager

Amy.Kirchhoff@ithaka.org
@AmyPlusFour