From: [redacted]</O=ITHAKA/OU=FIRST ADMINISTRATIVE GROUP/CN=RECIPIENTS/CN=[redacted]>
Sent: Monday, October 25, 2010 1:41 PM
To: [redacted] <[redacted]@ithaka.org>
Subject: RE: Update: JSTOR & MIT

Ahhh, I see...OK, I'll respond to this. If they want it all rolled up, I can do that (somehow!@$) :)

[redacted]
[redacted]
ITHAKA
[redacted]
Desk - [redacted]
Cell - [redacted]

ITHAKA (www.ithaka.org) is a not-for-profit organization that helps the academic community use digital technologies to preserve the scholarly record and to advance research and teaching in sustainable ways. We provide innovative services that benefit higher education, including Ithaka S+R, JSTOR, and Portico.

-----Original Message-----
From: [redacted]
Sent: Monday, October 25, 2010 1:28 PM
To: [redacted]
Subject: FW: Update: JSTOR & MIT
Importance: High

See below...

-----Original Message-----
From: [redacted]
Sent: Monday, October 25, 2010 1:23 PM
To: [redacted]
Cc: [redacted]
Subject: RE: Update: JSTOR & MIT

[redacted],

Can you elucidate for me how the column called "Total Articles" is derived? I know that seems like an incredibly basic question about what should be obvious terminology, however, when I actually look down the list, many of the numbers seem suspiciously low. It's an important number because it drives the percentage loss calculation, so it is important that these are right. Just for one example, my heart palpitates when I see that the Bulletin of the College Art Association was among the victims of the 9/25 theft, and our calculation is that 50% of their articles were downloaded! But the actual numbers say that 2 articles were downloaded out of a total of 4. It just so happens that the Bulletin was a short lived title - it changed name shortly after it started, so I could imagine that the count only looks at the underlying title and doesn't roll up to the full history of that title (though that might be a more appropriate way for us to look at these numbers). But I still can't reconcile a count of 4

articles. I only looked at the ToC for issue #4, and it had 23 articles in it. Is it possible that this count is really referring to a number of issues, and not the number of articles?

Seems important that we have a better understanding of the numbers before we decide how to work with MIT about this.

Best,

[redacted]

-----Original Message-----
From: [redacted]
Sent: Monday, October 25, 2010 12:17 PM
To: [redacted]
Cc: [redacted]
Subject: RE: Update: JSTOR & MIT

This is horrible.

-----Original Message-----
From: [redacted]
Sent: Monday, October 25, 2010 11:43 AM
To: [redacted]
Cc: [redacted]
Subject: RE: Update: JSTOR & MIT

My Apologies All,

The sheet originally provided only contained the 10.9 incident (8510 downloads), not the 9/25 incident (400K downloads). Somewhere between Mac --> JIRA --> PC the sheet with the original incident got dropped.

Please find the revised document attached with one sheet for each incident.

Thanks,

[redacted]

-----Original Message-----
From: [redacted]
Sent: Thursday, October 21, 2010 9:09 AM
To: [redacted]
Cc: [redacted]
Subject: RE: Update: JSTOR & MIT

Thanks [redacted],

I took another look at the JIRA and [redacted] has already compiled that data. I threw it in Excel and sorted by percentage taken (see attached). If there is other information that would be of use in analyzing our approach, don't hesitate to ask.

[redacted]

-----Original Message-----
From: [redacted]
Sent: Wednesday, October 20, 2010 12:14 PM
To: [redacted]
Cc: [redacted]
Subject: RE: Update: JSTOR & MIT

Hi [redacted],

Can you provide info on the content involved here. 562 journals were affected. Were there instances in which the entire back run, or most of the back run, of a journal was affected? If so, which ones? 453,000 articles is approximately 7% of the articles in our database; this is significant, but beyond this, and depending on what specifically was downloaded, it also could be 97% of a publisher's -- or many publishers -- content....

Thanks,
[redacted]

-----Original Message-----
From: [redacted]
Sent: Friday, October 15, 2010 9:29 AM
To: [redacted]
Cc: [redacted]
Subject: RE: Update: JSTOR & MIT

Thanks [redacted],

We are agreed, the number is staggering and real theft. I also agree that we should pursue a much more extensive resolution and want to be a facilitator here as best I can, until others are ready to liaise with our contacts there as needed. Given their response thus far, and given historical cases, it resembles language and approach that suggest they have not as yet and perhaps cannot identify the user. It is not terribly uncommon that this is the case, but usually big infrastructure translates into a much more exacting reply, so I am a bit surprised by where it stands at present. You'll see in my correspondence with them, I am trying to get clarification on that issue. If they are able to identify the user, we can discuss what steps seem appropriate based on their response.

I have not made them aware of the numbers from the 9/25, 9/26 or 10/9 incidents as they just came in, but am seeking to do so shortly in the hopes that the urgency and gravity translates directly.

[redacted]

-----Original Message-----
From: [redacted]
Sent: Friday, October 15, 2010 9:04 AM
To: [redacted]
Cc: [redacted]
Subject: Re: Update: JSTOR & MIT

Thanks for this update, [redacted]. This is all exactly right. In addition, I'd like to understand from the university how it manages these type of incidents on a broader scale. This is an astronomical number of articles -- again, real theft (and one can assume willful malfeasance given the use of a robot, etc.). Does the university contact law enforcement? Would they be willing to do so in this instance?

[redacted]

----- Original Message -----
From: [redacted]
Sent: Friday, October 15, 2010 08:45 AM
To: [redacted]
Cc: [redacted]
Subject: Update: JSTOR & MIT

Good Morning,

We have received word back from MIT. I am including their email and my response as well (see below). That is where we stand at the moment.

For those of you following OPS-1843, Quantify MIT Abuse Cases, you'll see that some startling numbers came in last night. The good news is that the latest incident was contained much more quickly. That said, some significant work to be done yet. Summarizing here.

Incident on 9/25 & 9/26

IP = 18.55.6.215
Start = 25-SEP-10 05.06.49.109524 PM
End = 26-SEP-10 04.24.54.297995 AM
Total Sessions = 1,256,249
Total Articles Downloaded = 453,570
Total Journals Affected = 562

Comments: This is an extraordinary amount and blows away any recorded abuse case that I am aware of since

the CASS days.

Incident on 10/9

IP = 018.055.005.100
Start = 2010-10-09 14:53:18 from
End = 2010-10-09 19:08:01
Total Sessions = 8,515
Total Articles Downloaded = 8,422
Total Journal Affected = 714

Comments: Noticed quicker, dealt with quicker.

Correspondence as of yesterday...

Hello [redacted] and [redacted],

Our investigations here point to the same guest that was involved in the 9/27 incident. We don't have enough information to follow the trail completely, but the signs suggest that the same guest user was responsible for this latest activity. To pursue this further, our IS&T group would need more information. Specifically, they are wondering if you are seeing any robotic activity from MIT currently and if so, whether you have any information about the IP addresses involved.

Given that it appears all of this excessive use was caused by a guest visitor at MIT, we have been considering next steps, and would like to suggest that we move to a new access model that will eliminate use by guests. We have recently developed an additional authorization layer that we can apply to particular products to prevent access by guests/walkins. We've tried this approach with one or two publishers where we had seen repeated excessive use, and it has stemmed the problem in those cases.

We would orchestrate this change by changing the proxy configuration on this end, and then we'd ask you to change the list of acceptable MIT IPs to only our proxy server's address -- a single IP.

If this sounds like an acceptable approach, let's discuss the next steps. To carry out the change, I'd have JSTOR work with [redacted], copied here.

Best,

[redacted]

Thank You [redacted],

I appreciate your response here. It appears we still have a ways to go to reach resolution, but I am glad to assist.

First, this activity is not continuing at the moment. Given that we saw it twice in two weeks, starting on a Saturday, I will hazard a guess that if this does recur, it will begin again on a Saturday. That said, if an when it does recur, we will be denying IP ranges significant enough to prevent it from continuing, while hopefully avoiding the need to block the entire range again. Internally, we are agreed on this point.

Second, we typically follow each case of excessive downloading with a three step process for considering the incident resolved...

1. Is it continuing? Not at the moment, but the jury is still out and will be for a few weeks.

2. Did the institution take the necessary steps to prevent recurrence? I see your suggestions here and have some thoughts on it as a follow on conversation. At present however, it is very important for us to understand if the users password has been changed and if the user has been contacted directly to address this issue. As a guest user, and likely the same user involved previously, using an efficient robot to grab lots of content, this is paramount to solve at the individual user level. If it is a shared account or used by multiple users, this is even more critical.

3. Was the content acquired deleted? This can be tricky, we understand, but if you can identify the user, in combination with adjusting their credentials, we must request that the best effort be made to insure that the content acquired is deleted from the storage device or web space in which they are storing it.

We can give you very granular log files from our end if identifying the user is problematic, but not identifying the user and assuring that the content is deleted, especially on a incident of this size, is a sizeable barrier to bringing this incident to a close.

As for your suggestion, we would gladly adjust the IPs that have access to JSTOR at your request. Note that some of our very large institutions do authenticate in this way. Also note that most very large institutions that do use proxy servers, use 2 or 3 to meet their bandwidth and access control needs. That said, I want to make sure we are on the same page here. Adjusting your configurations to prevent future occurrences is separate from bringing resolution to this incident.

If your IS&T group need additional information for activities between the time frames already provided, please do let me know what kind of information they are looking for and how much. Like, logs for at least 30 consecutive actions from an MIT IP between the times of 16:00 and 16:30 on Saturday, and we'll be happy to provide them.

Thanks,

[redacted]

-----Original Message-----
From: [redacted]
Sent: Tuesday, October 12, 2010 2:06 PM
To: [redacted]
Cc: [redacted]
Subject: Update: JSTOR & MIT

Just a quick update...

[redacted] is compiling the last of the stats surrounding these two incidents. All IP addresses have been restored for access to JSTOR at MIT with [redacted] keeping a watchful eye for recurrence. I have been in contact with our contacts at MIT and they are very helpful. Once we have the IPs and date stamps from our logs, I will be requesting a summary from their side, an outline of steps taken and passing along our summary to you all.

[redacted] at MIT is very appreciative of our efforts here and was not upset that their IPs had been blocked, but seeking, as we all are, to have full reinstatement and activity return to normal with the requisite accountability. We will continue working together toward that end.

[redacted]

-----Original Message-----
From: [redacted]
Sent: Tuesday, October 12, 2010 10:39 AM
To: [redacted]
Cc: [redacted]
Subject: RE: Update: JSTOR & MIT

Thanks [redacted],

First, let me take the opportunity to clarify the two versions of this that occur...

1. An institution trips one of our abuse threshold (300 PDFs in one session, 5000 sessions in one hour), there individual IP is blocked for 30 minutes.

a. Users from that IP address (sometimes a proxy serving the whole campus, sometimes just one IP address) will see the standard error page that was created last August as we implemented abuse tools...

Access Suspended

Access to JSTOR from your current IP address ([redacted] has been suspended. We will be in contact with the administrators at your institution directly and will work to have access restored as quickly as possible. For more information, please contact JSTOR Support.

...If the activity occurs just once, we consider the issue resolved and the message effective in outlining the Terms & Conditions of Use for the end user. If the blocking recurs for that institution, we typically get a hold of the institution and seek correspondence and resolution. Long term cases at institutions are fairly rare and usually don't persist day in and day out, but occur a few times over the course of a few weeks until the institution can get it resolved. Each block basically = 300 PDFs, which means a small amount of the archive is leaking out, never en masse.

b. This particular case highlights that our 5000 session limit (implemented as a response to MIT on 9/29) is calculated per IP AND per server. We were under the impression that it would be applied per IP only, which would have caught this 2nd incident. We will use the data derived from this incident to put a limit in place that accounts for the per IP, per server metric.

2. In the MIT case, the Class A range was blocked, at [redacted] request, at the firewall level. This was necessary because the traffic itself, even if denied the ability to download PDFs, was so intense it would have had the same effect on our server stability. In this case, users are seeing...

"Server not found. Firefox can't find the server at www.jstor.org."

...because it is not implementing the Literatum abuse tools, but is blocked at the firewall.

In summary and answering your questions directly. I can only recall one other time that [redacted] blocked an IP at the firewall. It wasn't abuse, but it was a robot gone haywire, downloading the same PDF at a wild rate and beginning to threaten our capacity to serve the public site on some servers. We can alter the message that users see when IPs are blocked, but it is a one size fits all solution. We cannot alter what users see when their IP is blocked at the firewall.

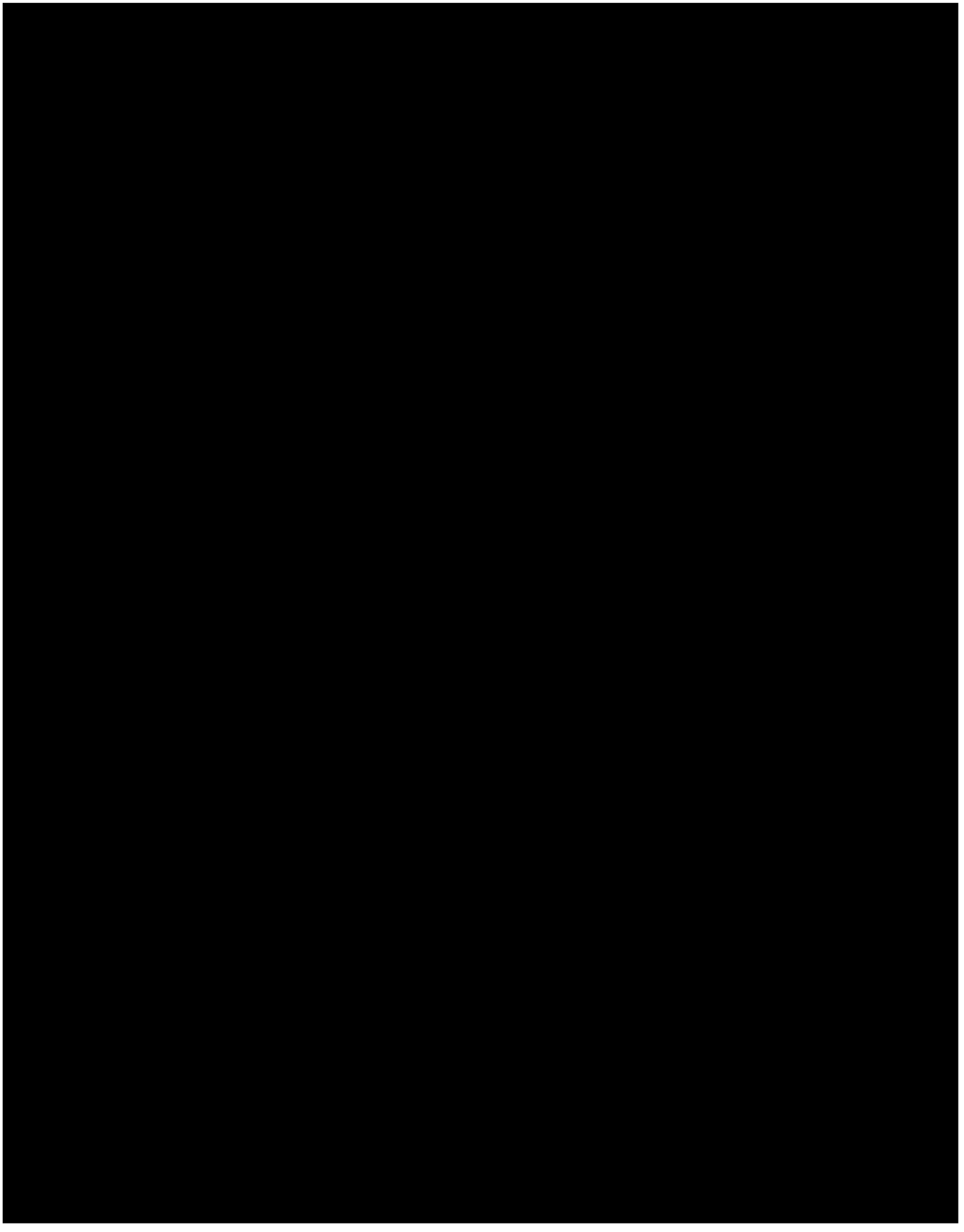
It is perhaps useful to note that the librarians we are in contact with are rarely defensive or irritated, and almost always shocked, embarrassed and apologetic. These are also the same librarians that we sell our content to. Our basic approach is to leave them with the impression that we are simply being good stewards of the content and using reasonable means to do so. Blanket IP range blocks and excessive force are to be avoided when possible and are not necessary 99% of the time. Once the librarian understands the different pieces of the abuse puzzle, they are very cooperative and looking to help.

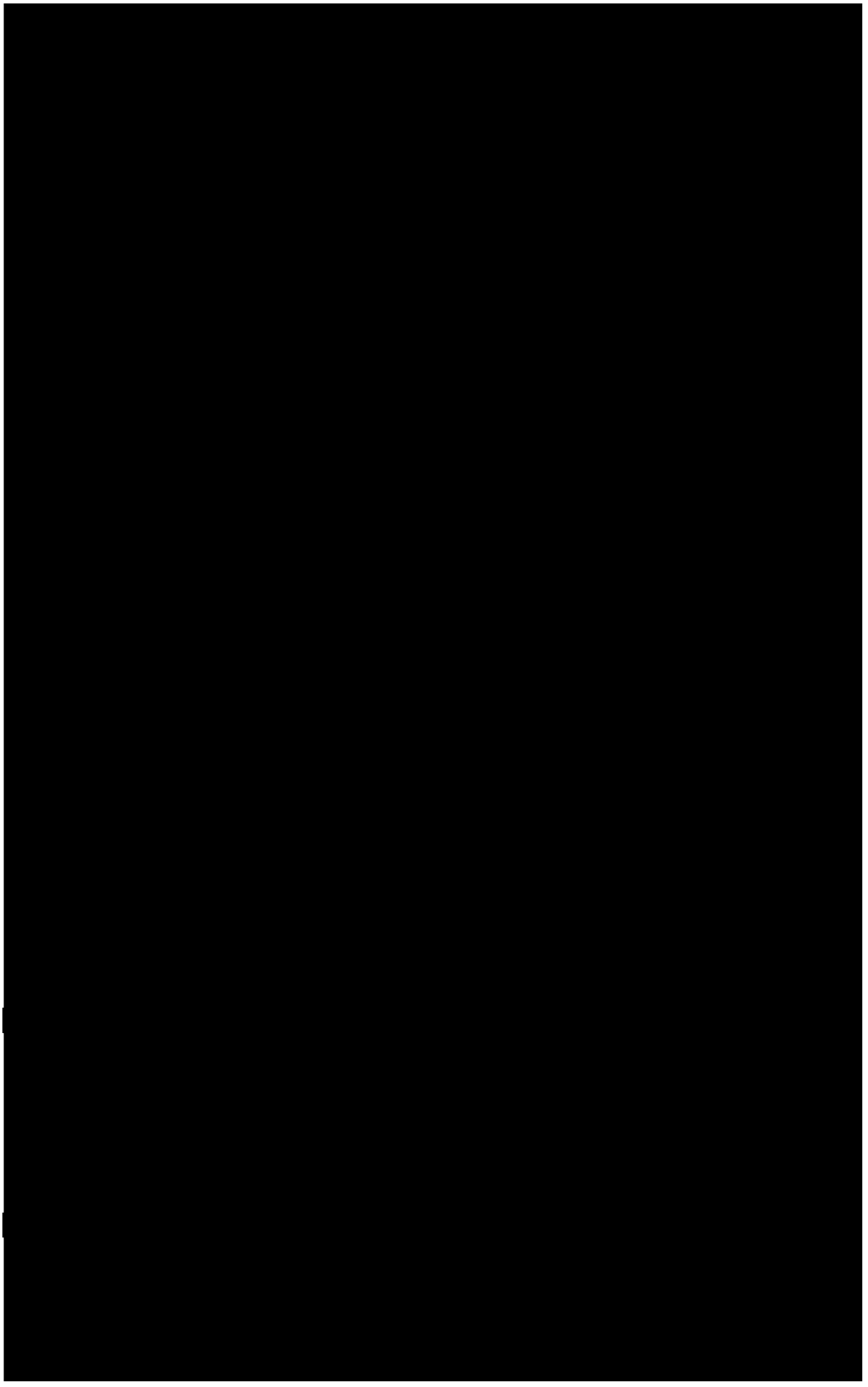
That said, it is a useful exercise to understand the nature of the problem here. By doing a simplified Chinese language Google search on "EZProxy password", you will find numerous lists with valid authentication information for hundreds if not thousands of schools. I copied the contents from a random site on the first page of results found using this search below just now. The number of sites like these are legion. So it's not that the librarian or technical staff are able to stem this tide either and we need to understand their position as well. We need to be level headed and even handed. This particular MIT case is extremely abnormal.

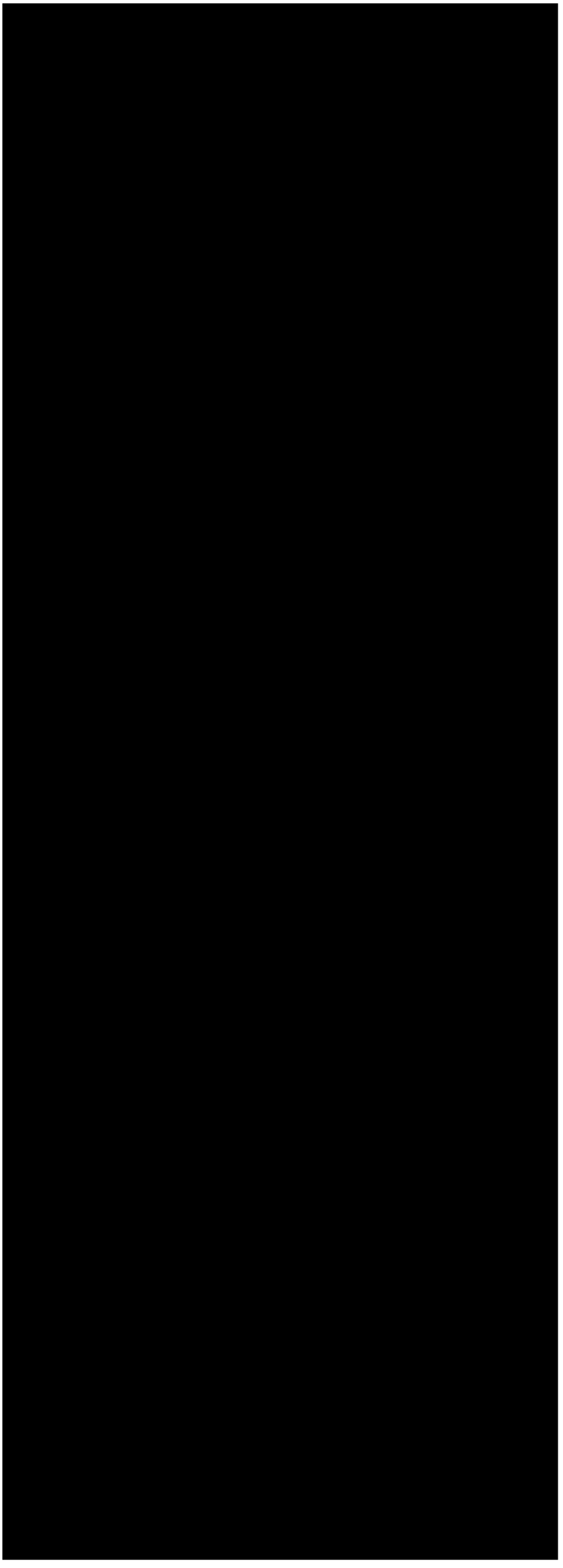
All that said, with CSP on our doorstep, it would be a valuable enterprise to understand our partner's expectations for protection of their content and to help them understand our capabilities and limitations as well. In some cases, we will be doing more to protect the content than they have historically, in others, because our usage is so high, it will be hard to match their efforts because the abuse tools don't scale particularly well to both prevent excessive downloading and maintain access for legitimate users. Proxied access is especially hard in this regard. That is, you could easily imagine a larger school having 200 unique sessions from one IP (proxy) in an 5 minute span (a professor assigning one article in a large lecture could hit this mark in isolation), whereas

200 sessions in a 5 minute period from the same IP at the UC Press website might look like an onslaught.

In case, once MIT is resolved, we will have to circle back and at least breakdown what our protocols should be going forward and begin to scope the CSP engagement with regards to abuse at JSTOR.







[redacted]

-----Original Message-----
From: [redacted]
Sent: Tuesday, October 12, 2010 8:05 AM
To: [redacted]
Cc: [redacted]
Subject: RE: Update: JSTOR & MIT

For the future, what happens when we deny an entire site (from an end user perspective) -- what message do users receive? Is there any opportunity to customize? How frequently do we have to take action at this scale?

-----Original Message-----
From: [redacted]
Sent: Monday, October 11, 2010 7:56 PM
To: [redacted]
Cc: [redacted]
Subject: RE: Update: JSTOR & MIT

Done.

Dear [redacted] & [redacted],

Good evening. I am hoping to hear additional news from you about the status of this weekend's block of IPs for JSTOR access at MIT. We are beginning to receive feedback from MIT users on our Facebook page and via direct email and we would like to be able to let them know the current status of the IP denial and an expected timetable for resolution. We are reticent to do so having not heard from you. A progress report on this incident would be helpful to assist us in better serving our mutual patrons.

Again, please do let me know if I can assist further from our end and I'll be glad to do so.

Best,

[redacted]
[redacted]
JSTOR
[redacted]@ithaka.org

-----Original Message-----
From: [redacted]
Sent: Monday, October 11, 2010 7:36 PM
To: [redacted]
Cc: [redacted]
Subject: RE: Update: JSTOR & MIT

I would let our MIT contacts know immediately that we are hearing directly from end users and how they would like us to respond. We don't want this discussion to go viral on Facebook, etc., so my advice is to try to avoid direct responses about robots and such. This could result in criticism in both directions that could be hard to stop.

[redacted]

-----Original Message-----
From: [redacted]
Sent: Monday, October 11, 2010 7:32 PM
To: [redacted]
Cc: [redacted]
Subject: RE: Update: JSTOR & MIT

Good Evening,

By way of an update, we have one email and one Facebook post referencing the outage at MIT, both are from end users and are of the wondering what's up and giving us an FYI variety. Having not heard from MIT officially today, I am suggesting we respond to both users with the following...

Thanks for alerting us to the issue with JSTOR access and MIT. Over this past weekend, robotic activity was noticed at JSTOR that is in violation with our Terms & Conditions of Use. The scope of this activity required us to deny access to JSTOR for all of MIT until it can be resolved.

We are in communications with the library and technical staff at MIT and expect resolution shortly. Please accept our apology for any inconvenience this may have caused. We are working to restore JSTOR access to MIT as quickly as possible and anticipate a resolution shortly.

... but welcoming suggestions. We can also refer them to their librarian, but note that this can be seen as a passive aggressive step from their end, though it would provide additional pressure on them, and is usually reserved for the completely non-responsive official contacts.

No doubt, the correspondence thus far from them would seem to be direct and agreeable, but no word from them today. From the incident on 9/25 and 9/26, they confirmed resolution on the 29th, so it might be expected to take a day or two, but that was only denying a small subset of their range and this is much, much larger.

I will reach out again, directly, first thing tomorrow morning, just to make sure they are in receipt on their end and action is being taken. Without additional word directly from MIT or anyone on this email chain, I will respond to the two users others going forward as stated above by 10pm EST.

Best,

[redacted]

-----Original Message-----
From: [redacted]
Sent: Monday, October 11, 2010 2:05 PM
To: [redacted]
Cc: [redacted]
Subject: RE: Update: JSTOR & MIT

Thanks, [redacted].

Does sound quite probable that this is an open proxy issue. I suggest we also ask MIT to scan for other open proxies, given that we had a situation with them a couple of weeks ago as well. If it's not an open proxy (that is, if the infringer is on-site or locatable/identifiable), I'd like -- as you already note -- confirmation of deletion of harvested content. I'd like to understand with some specificity how they go about obtaining this confirmation and ascertaining its veracity. And, how do they "deal with" these situations, beyond requesting confirmation of deletion? Are they able to tie the activity to a former "visiting scholar" or other individual? If so, are they willing to work with us to pursue more stringent law enforcement efforts (I'm not saying that we would in this circumstance, but I'm not necessarily satisfied with letting things go simply because the activity "stopped"; again, this is industrial theft and it's happening on a large scale or organizations all over). Also, open proxy is one risk and we should consider what if any follow up is possible re tracking down the content stolen from locations far away, but I also have real concerns about our content being downloaded more locally to hard

drives or exported elsewhere. So, there may be different follow up depending on the type of infringement occurring.

In any event, this is one of the reasons for wanting to implement discrete watermarking or identifiers, should we in time find our content re-purposed by other sites.

-----Original Message-----
From: [redacted]
Sent: Monday, October 11, 2010 12:47 PM
To: [redacted]
Cc: [redacted]
Subject: Update: JSTOR & MIT

Afternoon Update,

Still no word from MIT, but I suspect it will come shortly. That said, and wanting to be prepared, if there are any details or contingencies for reinstatement, we should be developing those now. They will likely come back and say it's taken care of again. They may or may not offer a reason. An immediate recurrence is highly unlikely, whether they have truly taken care of it or not, so it will be hard to solicit proof.

If I were forced to guess, I think they will report back that they identified a compromised User Name and Password and a bunch of referring access from IPs around the globe (typically some combination of China, Russia, and a smattering of Eastern European, Asian and South American origins). Some schools think that blocking those referring IPs is sufficient, which it is not, but isn't a bad addition. Hackers generally use Open Proxies to fake their actual location and can find an alternate Open Proxy to use quite readily. Only changing the password or disabling the offending Username and Password is an acceptable solution.

In cases like these, we ask them to confirm that the identity responsible has been dealt with, we also ask that they confirm deletion of harvested content, but if it is from a referring IP abroad, this user could be anyone/anywhere.

Anyway, if there are special requests or requirements to gain reinstatement, we should have them at the ready.

Thanks,

[redacted]

-----Original Message-----
From: [redacted]
Sent: Monday, October 11, 2010 11:04 AM
To: [redacted]
Cc: [redacted]
Subject: Re: Extreme robotic activity of JSTOR at MIT

Thanks [redacted],

There was one Facebook post at midnight, a normal user from MIT (at least via his profile he lists the MIT Network in Facebook), having trouble. I have not responded, wanting to give MIT at least the morning to touch base. Still no word from MIT.

Looping in [redacted] and [redacted]. I brought then up to speed last night.

[redacted]
[redacted]
JSTOR | Portico

[redacted]@ithaka.org
[redacted]

On Oct 11, 2010, at 10:40 AM, [redacted] <[redacted]@ithaka.org> wrote:

> Good to see this response. I fully understand our need to be down until this is remedied, but I'm also mindful of the potential loss of goodwill from innocent MIT users who rely on us. Has [redacted] received any inquiries on this front?

>
> -----Original Message-----
> From: [redacted]
> Sent: Sunday, October 10, 2010 9:43 PM
> To: [redacted]
> Subject: Fw: Extreme robotic activity of JSTOR at MIT
>
> Fyi
>
> ----- Original Message -----
> From: [redacted] [mailto:[redacted]@MIT.EDU]
> Sent: Sunday, October 10, 2010 08:15 PM
> To: [redacted]
> Cc: [redacted]@mit.edu>; [redacted]@mit.edu>
> Subject: RE: Extreme robotic activity of JSTOR at MIT
>
> Thank you, [redacted]. Your action was entirely appropriate, and I appreciate your courtesy in letting me know. It is infuriating that MIT's security appears unable to stop this pattern. We will redouble our efforts to solve the problem. [redacted]

>
> _____
> From: [redacted]@ithaka.org]
> Sent: Saturday, October 09, 2010 11:15 PM
> To: [redacted]
> Subject: Extreme robotic activity of JSTOR at MIT

>
> Dear[redacted]
>
> I wanted to let you know about an extreme step we have taken this evening. Our staff have blocked access to JSTOR from MIT. This is a highly unusual step and one we do not take lightly. We have had to do so because someone is systematically attempting to download large parts of the JSTOR database from within MIT's IP range. They use robots to open a session, download a PDF, open a new session, download another PDF, and keep repeating at a high rate. Not only is this a problem because it is beyond the terms of the license, but the downloading is so extensive that it impacts other users and has even brought some of our servers down. We worked through a similar incident at MIT three weeks ago and thought that the activity was being done by a visiting scholar who had left. But it has started again at an even faster rate. I am not writing you to complain about the activity; I just wanted you to be aware of the extreme step we have taken and why.

>
> Our staff have communicated with your staff and will be working to get MIT access back up just as soon as possible.
>
> I'll keep you posted as I hear more.
> Best regards,
>
>[redacted]