Easy access to the version of record (VoR) could help combat piracy: views from a publishing technologist

In the 1990s many publishers saw the potential of the internet and started to move their content online. This consolidated the need for a shift in their business models from a focus on individuals to IP-mediated institutional access. Libraries were purchasing institution-wide subscriptions with access facilitated through fixed computers, in libraries and offices on campus. Over time, publishers added other institutional authentication mechanisms – trusted referrer URLs, library cards, EZProxy support, and so on – but we never addressed the poor user experience associated with off-campus access. Now, with the rise in mobile and tablet devices and increasing flexibility in work spaces, access control is failing.

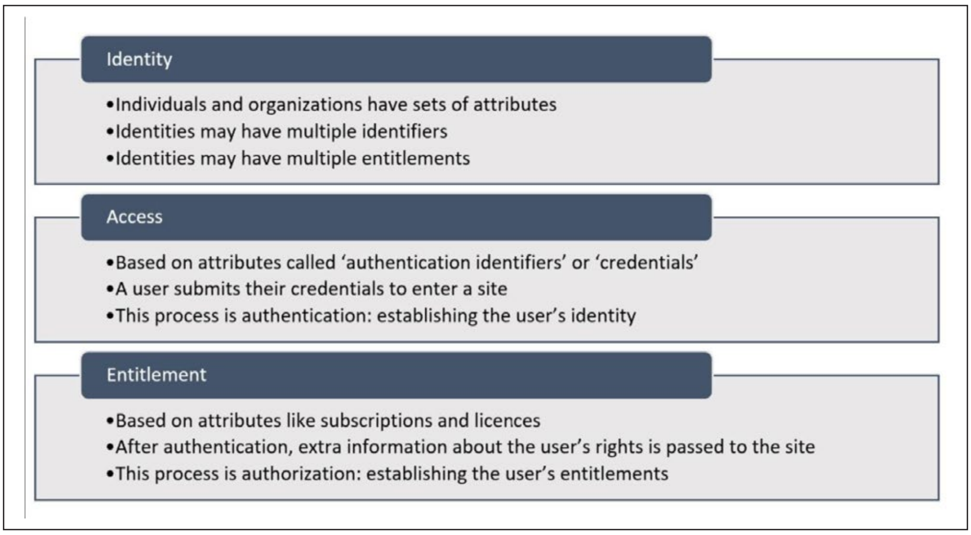
In this article, I argue that we need to find a balance between our desire for security and lowering barriers to access. As an industry, we can make use of technologies and initiatives which are already in place to help us to strike that balance, encouraging users to access versions of record instead of resorting to less legitimate copies through services such as Sci-Hub.

A little bit of jargon

By way of an introduction, I would like to familiarize you with the critical terms that are going to come up repeatedly – please see Figure 1. These include identity, authentication and authorization, the latter two often abbreviated to 'auth & authz'. You will also see the term 'IAMS', an acronym for 'identity and access management system'.

TASHA MELLINS-COHEN

Director of Publishing
Microbiology Society



Identity

- Individuals and organizations have sets of attributes
- Identities may have multiple identifiers
- Identities may have multiple entitlements

Access

- Based on attributes called 'authentication identifiers' or 'credentials'
- A user submits their credentials to enter a site
- This process is authentication: establishing the user's identity

Entitlement

- Based on attributes like subscriptions and licences
- After authentication, extra information about the user's rights is passed to the site
- This process is authorization: establishing the user's entitlements

Figure 1. Explanation of key terms used in this article

Identity and identifiers

As explained in Figure 1, individuals and organizations have sets of attributes, such as their names, their subscriptions, and so on. One type of attribute is the *identifier*.

I have a *lot* of identifiers: two personal e-mail addresses and a work one – never mind the e-mail addresses which are now defunct; a phone; instant communications through Skype and Slack; multiple social accounts, from LinkedIn, Twitter and Facebook, to Pinterest and – possibly less familiar to you – Ravelry. Then there are the shopping accounts with Amazon, eBay, Phase Eight, Etsy, and others … the list goes on. But they are all part of one identity: me!

The key takeaway here is that when someone claims to have a lot of identities, what they really mean is that they have a lot of *identifiers*.

Authentication and access

So what do all of those identifiers give me, apart from a headache when someone tries to get hold of me in twenty different ways? Access to services, of course. To get the best out of Amazon, I want to see recommendations which are tailored to my interests, special offers on items I have on my wish list, and details of my order history. In order to get all of that stuff, I need to log in with my e-mail address and password (my 'authentication identifiers' or 'credentials').

A note on security: while I use the same e-mail address as the username component for many services, every single one of them has a different 13-character password. Remembering them can be a pain, but it is worth it!

Authorization and entitlement

After Amazon's IAMS finishes authenticating me, it checks my list of entitlements – things like my Kindle library and Prime music. The IAMS uses that entitlements list to authorize me to view things that I have purchased, as well as my account details and so forth.

The (simplified) authentication and authorization process

Passive auth & authz

In the context of the scholarly communications space, there are some identifiers which permit passive auth & authz of users – that is, the process of granting access and entitlements happens automatically when the user hits a scholarly website, without them having to take any action. IP addresses and trusted referrer URLs fall into this category.

Figure 2 walks us through the workflow of passive auth & authz:

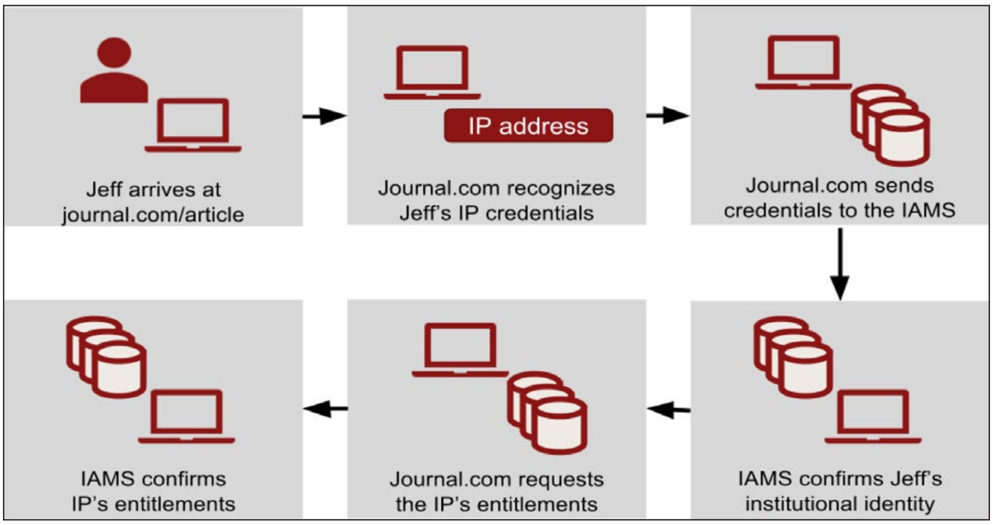

Jeff arrives at journal.com/article
IP address
Journal.com recognizes Jeff's IP credentials
Journal.com sends credentials to the IAMS
IAMS confirms Jeff's institutional identity
Journal.com requests the IP's entitlements
IAMS confirms IP's entitlements

Figure 2. Passive auth & authz

- the user arrives on a journal website
- the website passes the user's IP address – his identifier – across to the IAMS
- the IAMS confirms that the IP address is a valid identifier for an institutional identity
- the website then requests the institution's entitlements
- the IAMS confirms the institution's entitlements, and the user has access.

On campus, this entire process happens without the user's knowledge, or – critically – his action. That is the key to a good IAMS: an auth & authz process which is as quick and invisible as possible. IP recognition is so well used (anecdotally, most publishers still see around 90% of usage from this credential) precisely because it requires no effort on the part of the user. Off campus, IP authentication can of course be achieved through the use of VPNs and proxy servers: this does require some action on the part of the user, but once he is set up on the proxy server or VPN, the authentication process is seamless.

> 'the key to a good IAMS: an auth & authz process which is as quick and invisible as possible'

Active auth & authz

Active identifiers are becoming more common as increasing numbers of users shift to off-campus and mobile research. They include username/password combinations, Shibboleth/OpenAthens (Figure 3) and voucher codes, as well as some more esoteric identifiers. The challenge here is that users must (a) be aware of their active identifiers, (b) know which active identifiers are valid on a specific publisher site, and (c) care enough to jump through that hoop instead of taking the easy route to Sci-Hub,[1] ResearchGate,[2] or another scholarly communication network which offers free access to content.

> 'Active identifiers are becoming more common'

One challenge with the Shibboleth workflow outlined in Figure 3 is the 'WAYF' ('where are you from') page, more properly called a discovery service. As well as introducing another step into the authentication flow, the WAYF requires a user to know exactly what their institution is called in the context of this particular website: as we all know, institutions tend to have a proliferation of names, and it is not always easy to know which to look for.

Some organizations make use of federated access solutions such as Shibboleth internally, to control access to everything from e-mails and e-learning to payslips. Within such organizations we can expect a greater degree of awareness of federated access options, but even here, not every member of staff or student will be aware that their credentials will gain them access to publisher content.

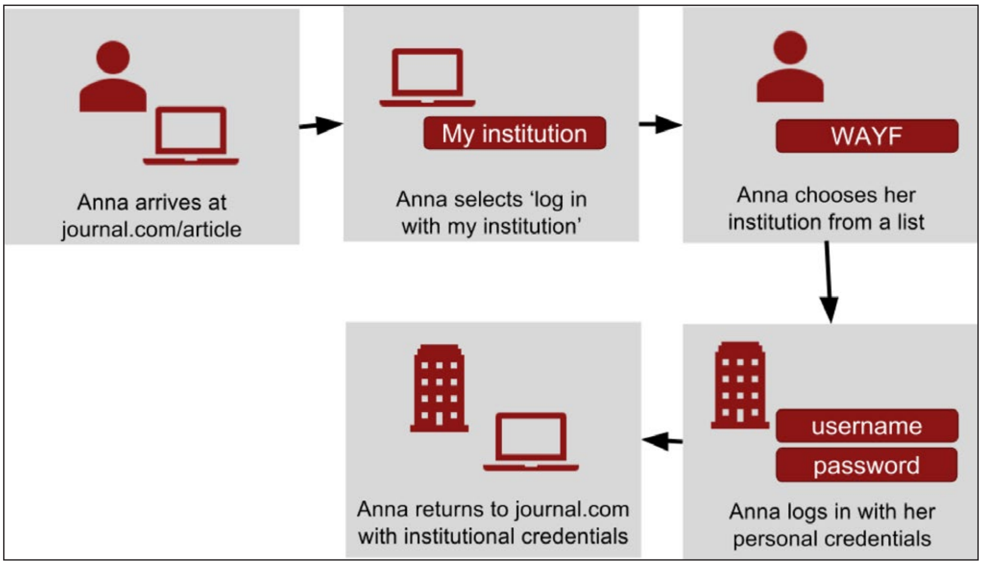

Anna arrives at journal.com/article
My institution
Anna selects 'log in with my institution'
WAYF
Anna chooses her institution from a list
username
password
Anna logs in with her personal credentials
Anna returns to journal.com with institutional credentials

Figure 3. Auth & authz using Shibboleth

The balancing act

The publishing industry is facing a lot of pressure to ease access to published research. While open access (OA) and the desire to eliminate paywalls is a part of this pressure, much of it also comes from the challenges associated with off-campus access, as described in the section on active auth & authz. The difficulties encountered in accessing content off campus, even where an organization has a subscription, are not negligible, as Roger Schonfeld described so well in his talk at STM Frankfurt in 2015 (video available).[3] As the music world discovered many years ago, making it difficult for individuals to gain access to legitimate content simply encourages them to look elsewhere. We need to think long and hard about learning from their experience and simplifying access to paywalled research, while securing our content against piracy.

Note that in the sections which follow, I am discussing only paywalled content – OA is a separate matter which is more than adequately covered elsewhere!

> 'making it difficult for individuals to gain access to legitimate content simply encourages them to look elsewhere'

Lowering the barriers to access for the version of record

While publishers clearly have a commercial driver around increasing subscriptions, removing the barriers to access for legitimate use can help to reduce the amount of usage cannibalization from scholarly collaboration networks and piracy. After all, if you can stream a film in HD on Netflix for a few pennies, why bother scrounging around the internet for a low-resolution pirate copy?

Peer-to-peer content sharing

The first of my three types of barrier reduction is simplified sharing among peers of individual content items, mostly at the article level.

Howcanishareit[4] came out of the STM consultation on what to share and how to share it, back in 2015, and is a great little site which uses DOIs to look up sharing policies. I believe that publishers could do much better when it comes to promoting this service, from signing up to the site and promoting it on journal home pages, to adding a widget to each article page outlining the sharing options which are permissible under the publisher's policy.

A similar, commercial initiative is the Springer Nature sharing link,[5] delivered through ReadCube, which allows authors to share read-only versions of their articles, as well as allowing media links for promoted/press released articles.

> 'these initiatives promote good behaviour on the part of the reader'

Both of these initiatives promote good behaviour on the part of the reader. After all, if users can simply and legitimately share an interesting article, they are less likely to seek out alternatives such as ResearchGate or ICanHazPDF.[6] They also encourage more use of the version of record (VoR), which leads me nicely to point two…

Improve discoverability of the version of record

As an industry, we could make big inroads into the access issue just by taking advantage of technology which is already available.

For starters, we absolutely need to clean up our metadata to improve discoverability. The new MetaData 2020 initiative[7] from Crossref is well worth checking out! As part of this effort, items which are free should be tagged as such – both Google Web Search and Google Scholar prioritize items known to be free ('world readable') in their search results. This can be done at the level of the site (e.g. 'this whole platform is OA'), the journal, the issue and the article. It is even possible to classify content and make rules such as 'all editorials are free' and have your metadata created to reflect that. As already mentioned, this use of metadata is not new, it is just best practice – speak to your typesetter for more information!

> 'we absolutely need to clean up our metadata'

The second thing we can do here is work with partners who will drive traffic to the VoR. For example, we can collaborate with Google Scholar via the Subscriber Links initiative,[8] in which publishers share identifier and entitlement information with Google Scholar. It is a counterpart to the Library Links initiative,[9] and in conjunction the two mean that when on-campus users arrive on a Google Scholar search results page, articles that match the entitlement information are highlighted, which drives traffic to the VoR. Similarly, ScienceOpen[10] indexes metadata and diverts users to the publisher VoR for full-text access.

Libraries also have a part to play here, in ensuring that they optimise their knowledge base data. Recent sector initiatives to address this have included GoKb[11] and KB+.[12] Clean data enables more accurate and comprehensive OpenURL linking from library search, again directing users to VoRs via a simple user experience.

Finally, we can and should be promoting existing alternatives to subscriber access, such as inter-library loans, as well as non-VoR fall-backs such as green OA repositories.

All of these activities help to reduce the profile of versions of our content stored on scholarly collaboration networks such as ResearchGate and Academia.edu[13] – some of which are perfectly legitimate, others less so.

Off-campus access

Last, but most definitely not least, in this barrier reduction exercise is true off-campus access.

RA21 (Resource Access for the 21st Century)[14] is a joint STM/NISO project to explore alternatives to IP authentication, with the mission 'to align and simplify pathways to subscribed content across participating scientific platforms'. The project aims to identify solution(s) to the off-campus access problem which meet these principles:

- The user experience for researchers will be as seamless as possible, intuitive and consistent across varied systems, and meet evolving expectations.
- The solution will work effectively regardless of the researcher's starting point, physical location, and preferred device.
- The solution will be consistent with emerging privacy regulations, will avoid requiring researchers to create yet another ID, and will achieve an optimal balance between security and usability.
- The system will achieve end-to-end traceability, providing a robust, widely adopted mechanism for detecting fraud that occurs at institutions, vendor systems and publishing platforms.
- The customer will not be burdened with administrative work or expenses related to implementation and maintenance. The implementation plan should allow for gradual transition and account for different levels of technical and organizational maturity in participating institutions.

RA21 is a fantastic idea, and we should all be watching this space intently. At present, the project is focused on federated access solutions such as Shibboleth, and in particular on solving the problems associated with determining where a user is from. One concern which I have, personally, is that there is no mention in the principles of the *publisher* not being overburdened with administrative work or expenses related to implementation and maintenance!

'RA21 is a fantastic idea, and we should all be watching this space intently.'

Another interesting project in this space, predating the announcement of RA21 and just about to pilot, is CASA[15] (Campus-Activated Subscriber Access), a joint Google Scholar/HighWire Press idea which takes the existing Subscriber Links initiative and extends it. Under CASA, when on-campus users arrive on a Google Scholar search results page, as well as their entitlements being highlighted, a cookie will also be placed on their device which records their institutional identifier. When that user returns to Google Scholar off campus at a later date, their institutional entitlements will again be highlighted in the results list. On clicking through to read the article, Google Scholar will pass over an encrypted CASA token to the journal website, which the journal website will then decrypt in order to passively authenticate and authorize the user. (Figure 4.)

Increasing security through good practice

We have seen that there are many options for us around reducing barriers to access. On the other side of the balance beam, there is much less that we as an industry can do to increase security without driving more users to sites like Sci-Hub. However, there are some things we should all be doing as part of our business hygiene routines that will help to protect our content.

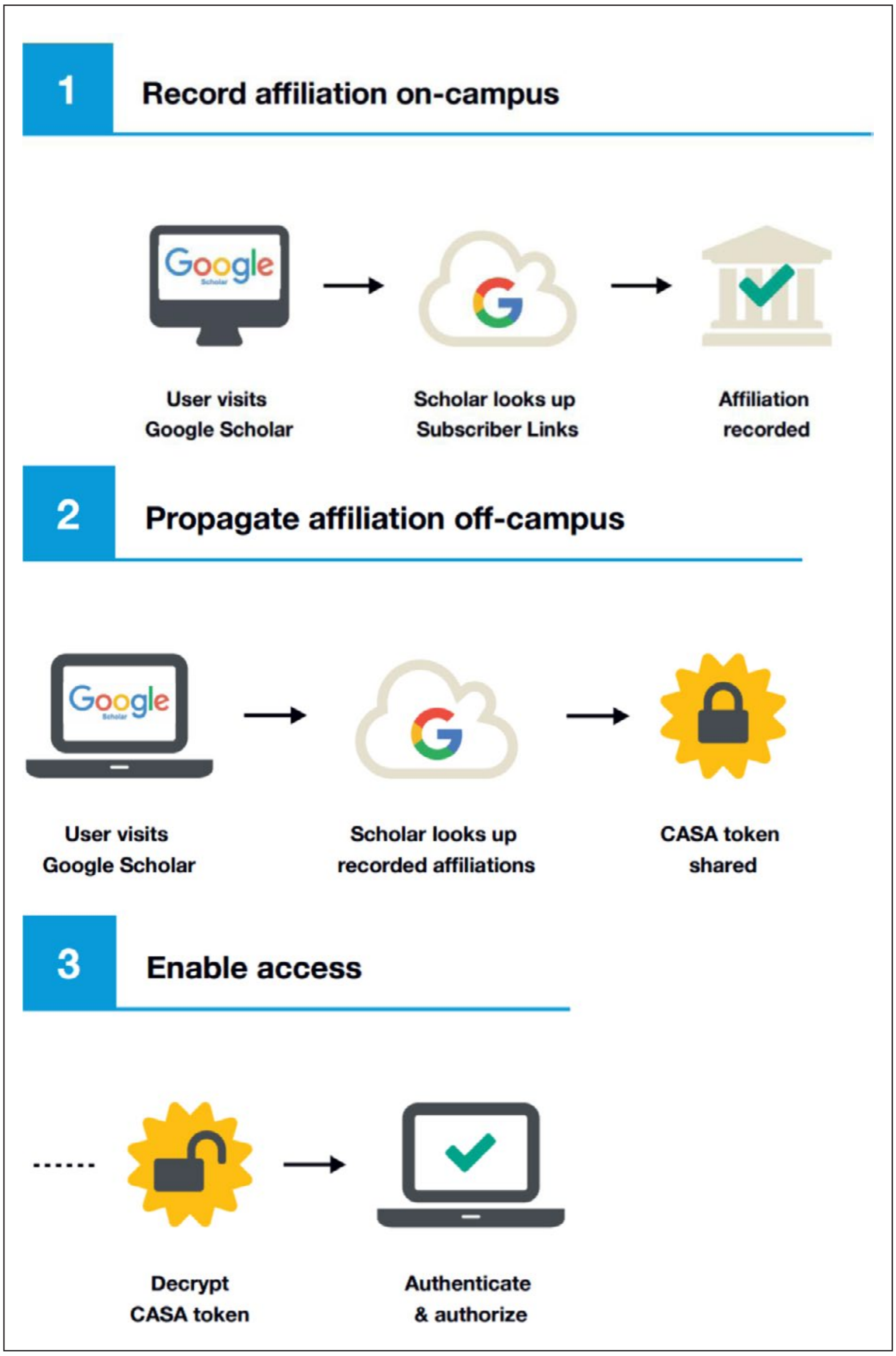

1
Record affiliation on-campus
Google Scholar
User visits Google Scholar
Scholar looks up Subscriber Links
Affiliation recorded
2
Propagate affiliation off-campus
Google Scholar
User visits Google Scholar
Scholar looks up recorded affiliations
CASA token shared
3
Enable access
Decrypt CASA token
Authenticate & authorize

Figure 4. CASA auth & authz flows. (Diagram reproduced with permission.)

First, publishers can undertake a simple audit of IP ranges to clean up institutional subscription IPs and prevent content being opened up to the wrong customers without payment. There are a whole range of services available which will help in this, such as the IP Registry from Publisher Solutions International.[16] Some of the IP ranges identified during such an audit may be perfectly valid, for example remote campuses, while others are less so. A plea: if and when you do identify a dubious IP – from a pirate, hacker, crawler, or robot – please do log and share that information with fellow publishers and with other institutions.

Second, monitoring usage activity and setting alerts for unusual spikes or surges in activity can help to spot piracy. The first action on spotting a spike would usually be a simple investigation, and possibly suspension of the account if the usage is not in line with the subscription. If you can do this in real time, great! However, reviewing per-institution usage patterns every so often is good practice. In my experience, analytics reviews like this can help to spot the source of piracy. In one case at Semantico, a small but sustained increase in usage over several days triggered an investigation. We identified that a specific institutional username/password combination had been shared with Sci-Hub, allowing them to scrape content under the guise of legitimate usage. The publisher contacted the institution concerned and persuaded them to switch to a more secure authentication mechanism, shutting down the ongoing scraping of the site.

There have been instances where extremely wide IP ranges meant that two complete institutions, in different parts of the world, were able to access content under the same subscription. By spotting the unusual activity (two periods each day of activity, separated by a lull), the publisher was able to investigate and resolve the discrepancy.

Finally, undertaking appropriate due diligence for all sales agents ensures that publishers adhere to the legal compliance requirements of the UK Bribery Act of 2010,[17] the US Foreign Corrupt Practices Act,[18] and similar international laws concerning anti-bribery and corruption, as well as identifying and preventing rogue agents profiting from the unauthorized resale of personal rate publications and membership subscriptions.

‘We need to take action as an industry to make it easier for researchers to gain access to legitimately purchased content’

To summarize

We need to take action as an industry to make it easier for researchers to gain access to legitimately purchased content, wherever they are and whatever device they are using, without forcing them to jump through hoops (like VPN log-in to the library management system), or risk driving them to piracy. There are options for subscription-based publishers – we just need to get better at using them!

Acknowledgement and statement
The author would like to thank Andrea Fallas of HighWire Press Ltd for the CASA graphics.
This is a personal opinion piece, and does not necessarily reflect the opinion of the author’s employers or previous employers.

Abbreviations and Acronyms
A list of the abbreviations and acronyms used in this and other *Insights* articles can be accessed here – click on the URL below and then select the ‘Abbreviations and Acronyms’ link at the top of the page it directs you to: http://www.uksg.org/publications#aa

Competing interests
The author has declared no competing interests.

References

1. Sci-Hub:
https://sci-hub.io/ (accessed 12 June 2017).

2. ResearchGate:
https://www.researchgate.net/ (accessed 12 June 2017).

3. Schonfeld R, 13 November 2015, Dismantling the Stumbling Blocks that Impede Researcher Access to E-Resources, The Scholarly Kitchen:
https://scholarlykitchen.sspnet.org/2015/11/13/dismantling-the-stumbling-blocks-that-impede-researcher-access-to-e-resources/. (accessed 6 June 2017).

4. How Can I Share It:
http://www.howcanishareit.com/ (accessed 12 June 2017).

5. Shared It:
http://www.springernature.com/gp/researchers/sharedit?countryChanged=true (accessed 12 June 2017).

6. I Can Haz PDF:
https://en.wikipedia.org/wiki/ICanHazPDF (accessed 12 June 2017).

7. Hendricks G and Lammey R, MetaData 2020: Presentation given at the Crossref-THOR outreach meeting: Much more than infrastructure: working together to connect research.
https://zenodo.org/record/571824#.WT69aBgrLrc (accessed 12 June 2017).

8. Google Scholar Subscriber Links:
https://scholar.google.com/intl/en/scholar/publishers.html#otherpolicies (accessed 12 June 2017).

9. Google Scholar Library Links:
https://scholar.google.com/intl/en/scholar/libraries.html (accessed 12 June 2017).

10. ScienceOpen:
https://www.scienceopen.com/ (accessed 12 June 2017).

11. Global Open Knowledgebase:
http://gokb.org/ (accessed 12 June 2017).

12. Knowledge Base+:
https://www.kbplus.ac.uk/kbplus/ (accessed 12 June 2017).

13. Academia.edu:
https://www.academia.edu/ (accessed 12 June 2017).

14. RA21 (Resource Access for the 21st Century):
https://ra21.org/ (accessed 12 June 2017).

15. Mellins-Cohen T, Campus Activated Subscriber Access. Poster at the UKSG Conference, Harrogate, April 2017.

16. The IP Registry:
http://theipregistry.org/ (accessed 12 June 2017).

17. Bribery Act 2010:
http://www.legislation.gov.uk/ukpga/2010/23/contents (accessed 12 June 2017).

18. Foreign Corrupt Practices Act:
https://www.justice.gov/criminal-fraud/foreign-corrupt-practices-act (accessed 12 June 2017).

Tasha Mellins-Cohen
Director of Publishing,
Microbiology Society, GB

E-mail: tashalouiza@gmail.com

ORCID ID: http://orcid.org/0000-0001-6229-9675

To cite this article:
Mellins-Cohen T, Easy access to the version of record (VoR) could help combat piracy: views from a publishing technologist, *Insights*, 2017, 30(2), 44–51; DOI: https://doi.org/10.1629/uksg.360

Published by UKSG in association with Ubiquity Press on 10 July 2017