February 4, 2021

Hon. Ron Wyden
U.S. Senate
221 Dirksen Senate Office Bldg.
Washington, D.C., 20510

Hon. Rob Portman
U.S. Senate
448 Russell Senate Office Building
Washington, DC 20510

Hon. Hank Johnson
U.S. House of Representatives
2240 Rayburn House Office Building
Washington, DC 20515

Hon. Jim Jordan
U.S. House of Representatives
2056 Rayburn House Office Building
Washington, DC 20515

Subject: Suggested changes to the Open Courts Act to shorten the timeframe, reduce the cost of the transition, and more effectively meet the vision of this legislation, which is to make the workings of the U.S. courts much more accessible to the public.

Dear Senator Wyden, Senator Portman, Congressman Johnson, and Congressman Jordan:

We write on behalf of a group that has extensive experience building large public sites on the Internet. The purpose of this letter is to advance action on improving public access to federal court records, which are presently offered by the government through an outdated PACER system.

We have extensive experience putting large government databases on the Internet and then working with public officials to help government do this work better. Our experience includes making available federal databases such as the U.S. Patent and Trademark database, the Securities and Exchange EDGAR database, the IRS Form 990 database, 14,000 hours of Congressional video from hearings posted at the request of the Speaker of the House, and over 6,000 government videos from the U.S. National Archives posted in cooperation with the Archivist of the United States. We have extensive experience working with legal information, and operate some of the largest sites for access to federal court filings, as well as the U.S. Code, the Code of Federal Regulations, the regulations of all 50 states, and much more.

1. The Bill Does Not Achieve the Goals Set For It

We congratulate members of the House and the Senate for their leadership and their efforts to make the PACER database more broadly accessible. Today, PACER is the most significant federal database of public importance that is not freely available. Transparency in our judicial process is not possible when access to the courtroom depends on access to a credit card. We must do better, and your leadership has been instrumental in moving this ball forward.

While H.R.8235, the Open Courts Act of 2020, as passed by the House during the last Congress aims to make public access to public filings a reality, the bill as written would not accomplish that goal. Under the bill, the Administrative Office of the U.S. Courts ("AO") would have up to five years to implement a new system. During those five years, the costs for PACER would go up as "power users" would be taxed at an even higher rate. At the end of the five years, PACER would not be free; it would still include charges for bulk data to large users based on "Service Level Agreements."

Under the bill, funding for this new system would most likely include, in addition to retrieval fees for "power users," an increase in court filing fees. However, the way that the bill currently reads, it appears that the AO has in mind a surcharge on initial filings, with a cap of 15% of the filing fee. Today, for example, an action in civil courts costs $350 for the filing fee and a $52 administrative charge. Under the current bill, the AO will be allowed to have a surcharge of up to 15%, this surcharge being capped at a maximum of $52 in the case of a civil action. This one-size-fits-all supplemental tax does not recognize that some litigants file far more documents than others. Litigants, the users of the courts, should pay based on their usage.

2. Litigants, the Users of the Courts, Should Pay Based On How Much They Use

We applaud shifting the burden for paying for the PACER system from the public to litigants, the actual users of the court system. We also applaud the provisions that would carve out exceptions for litigants who cannot afford these increased fees. However, a flat tax on each filing does not make any sense. For example, as the American Bankers Association pointed out, as currently drafted, there is the potential to put an undue burden on those that wish to file a simple Proof of Claims (POC) in a bankruptcy case.

We believe filing fees are the proper place for the AO to raise funds, but a one-size-fits-all tax is not the right approach. In many cases, the filings are simple and not voluminous. A complaint is filed, a response is filed, and there is a prompt judgment or settlement. In other cases, dozens of lawyers will file hundreds or even thousands of documents in litigation that can often stretch for many years. It is those voluminous users who are submitting thousands of pages that have to be read by clerks and judges who should be paying more, again with appropriate carveouts for financial need and the interests of justice.

It would not be difficult for the AO to start charging for filings by the page or by the document, with such a charge starting after a certain of pages have been filed. Those users that consume more judicial resources should be asked to pay more. This is also a sustainable business model, whereas charging for promulgation of primary legal documents will always be subject to attack as a barrier to access to justice. The AO should embrace this shift.

It is absolutely vital that the filing fees do not impose undue burdens on litigants that impose barriers on access to justice. The large litigants with resources should pay more. Those without the resources should not. Any filing fee schedule must take this principle into account. We believe it would be useful in the bill itself to clearly outline these principles, including who would be exempt from fees and that filing fees should be progressive not regressive when it comes to the resources litigants possess.

3. The Goal Should Not Be A Better PACER Site, It Should Be Many Better PACER Sites

The bill specifies a number of features for a new PACER site, such as full text search and permanent, linkable addresses for documents. The bill assumes that the problem is that the PACER site needs those improvements to be better. That is indeed true, but that is not the core problem.

The core problem is not PACER as it exists, but that it is very, very difficult for other sites to build databases of court records and become an alternative to PACER. This is unique when it comes to federal data. When you access a patent document today, perhaps you will go to the U.S. Patent and Trademark Office, but you are just as likely to use IBM's Patent Service or Google Scholar. If you want Securities and Exchange filings, you will likely access them on your broker's site or a financial service. If you want weather data, you might go to the National Weather Service, but you are just as likely to go to AccuWeather or Weather.Com.

One of the lessons we have learned in decades of working on large government databases is that you don't start with a fancy new web site, you start with the core bulk data. Then, you put in place an Application Programming Interface (API). An API allows one to navigate the bulk data. In the PACER example, an API request might be for the identifiers of all cases on a particular date that had new filings. A second API request might be for a list of the documents filed in a specific case, with a link to the documents themselves.

Only after bulk data is available, and an API is in place, does it make sense to think about building a web site on top of it. If there is bulk data and an API, other groups are able to build web sites. This decentralized approach treats government data as the raw materials of our democracy, information that should flow throughout society. If government wants to build a better web site, using the API, that is to be applauded, but it cannot take the place of better access for all to bulk data.[1]

As currently drafted, the bill has the order of actions reversed, which is why it would take 5 years and cost an inordinate amount of money. And, even if the AO is able to carry out this 5-year plan, it will have erected an even higher wall around "its" data and it would be exceedingly difficult for other groups to provide comprehensive access to court dockets.

4. This Process Should Take One Year, Not Five Years

There is no reason this process should take five years. There are two core tasks that need to be carried out:

1. Get ready for revised filing fees and improve privacy protections.
2. Provide a system for bulk access.

These tasks can be accomplished in one year, not five. The AO should focus on these two tasks which will lead to demonstrable change quickly, meeting a pressing need. These suggestions are meant to refocus the bill to set achievable goals, not merely tactics towards those goals.

4.1 The AO Should Focus on Establishing Filing Fees and Better Privacy Protection

The AO has some work to do to get its system ready to charge litigants for filings. It needs to establish a set of fees and a process to validate those fees as currently specified in the bill. Getting ready for new filing fees is not a difficult task. Use of PACER requires a credit card on file. There is already a procedure for waiver of fees. PACER has an extensive system for billing users in place. Much of the work that needs to be done has already been done. Under this proposal, the AO would have a sustainable source of funding for the future based on the use of the courts, instead of trying to enforce artificial restrictions based on dissemination fees.

[1] This principle was explained very well in Felten and Yu, "Government Data and the Invisible Hand," Yale Journal of Law & Technology, Vol. 11, 2009.

Indeed, we believe under our proposal, the AO will be able to maintain and even increase the resources it is able to bring in, helping it meet many pressing needs in a much more straightforward manner. We believe our proposal will be good for the AO.

In addition, the AO needs to do a much better job of enforcing privacy rules for court filings. It is the responsibility of lawyers who prepare filings to observe the guidelines established by the Judicial Conference to protect private information, such as names of minor children, social security numbers, and a variety of other categories of Personally Identifiable Information (PII).

Our extensive audits of privacy violations in PACER has shown that the AO could do much more. On modern computer systems, there are many options available that enable one to find PII in large document stores. On Google Drive, for example, enterprise users get periodic reports indicating if they have potential PII stored in their directories. There is no reason the AO could not scan every document as it gets filed and looks for the presence of PII.

If such information is detected, that does not require rejection of the document, but it is an opportunity to present a screen to the filing attorney indicating where the match for PII was and asking them if they are sure that they want to continue to file that document. If an attorney says "yes," and it turns out there is indeed PII submitted in violation of the Judicial Conference rules, then the judge always has the option of imposing sanctions for violations of those rules.

4.2 The AO Should Provide for Bulk Access

The second thing that needs to happen is to provide bulk access. There are two steps here. First, there needs to be a copy of the existing data, the "back file." Second, there needs to be a mechanism (an API), to query for new documents.

Getting the back file is not difficult. It could be as simple as providing a single account with no fees to a coalition that uses the existing system and copies what is there. We know how to do this already, in fact many of the signatories to this letter have extensive collections of documents from PACER. We would happily use a single account into the existing system and systematically bring in all the documents. It would be a bit slow and a bit cumbersome, but would not be hard to do that. We would share the data among ourselves and put copies of the data for open access on systems such as the Internet Archive, Dropbox, Google Drive, and Amazon's cloud storage.

If a coalition of groups, such as some of the signatories to this letter, got the data that way, one of the things we would do before posting any documents would be to use the same PII detection mechanisms outlined above, and move any flagged documents to the side.

We would like to stress that this option—give us an unbilled account to provide for better public access—could begin immediately.

There are, of course, more systematic medium-term approaches to the bulk data than simply giving us an account. We recommend that the AO put up a separate system that contains a copy of the back file. This "side car" system could be used to run the PII detection software. The bulk data back file would contain only public PACER documents, not those under seal. After the PII scans have been completed, the AO can place an API on the system that provides access to not only the back file, but to new documents.

The AO has not been forthcoming on statistics about PACER, but we have heard several times that PACER contains "one billion documents." How big is PACER? The Internet Archive has documents for 7,891,497 PACER cases. This collection includes dockets, PDF filings, and various derivative files (such as full text for each document obtained through the use of optical character recognition). This collection uses a total of 3.9 terabyte of disk space and contains 9,777,179 PDF files. The average size of each original PDF file is 399 kilobytes. We get similar results on systems such as Court Listener, which has determined that 11,327,691 PDF documents have an average size of 379 kilobytes.

If PACER really is "one billion documents," that indicates a database size of approximately 400 terabytes. In today's world, this is not a big deal. If the AO wanted to put a computer in place to hold the back file, a system with 34 18-tbyte drives, 2 tbytes of RAM, 48 CPUs, and 2 4-tbyte SSD drives for the operating system would cost approximately $55,000. Alternatively, one could put 400 tbytes of data on Amazon's S3 service for $8,800/month for frequent access. Once the back file is copied, the cost for infrequent access to a database of this size would $5,000/month on Amazon. These are not expensive systems for government IT efforts.

The next question is daily updates. Our experience shows approximately 100,000 new documents per day listed in the "RSS" feeds (a "news feed" that indicates new items available each day). However, those RSS feeds don't cover all the federal courts (and many courts only have partial feeds), so we believe a better estimate is 150,000 documents/weekday. For 200 weekdays in a year, we expect approximately 30,000,000 documents. Again, this is a very manageable amount of data.

There is no reason that the AO could not, in one year, put a system in place containing the back file, create an API on that system for access, scrub the back file for privacy issues, and then put in place a mechanism that takes any new public documents filed and places a copy on that system in addition to placing the file on the existing PACER system.

We are here to offer our assistance to the AO on making this process happen. If instead, the AO wishes to use services inside the federal government, it can do no better than to avail itself of the services of the General Services Administration or the U.S Digital Service, both of which are widely admired throughout the world for their ability to create state-of-the-art systems in a timely and cost-effective manner.

The AO has indicated that big users will require special Service Level Agreements that specify the technical parameters for bulk access, a subscription service that specifies parameters such as how long it takes to transfer documents, response time to queries, the percentage of uptime, and other parameters. It also indicates that to provide this assured level of service requires big bulk access fees. We do not believe big users need a Service Level Agreement for access to bulk data. Indeed, we don't know of any government databases that are used extensively by private industry that have such commercial Service Level Agreements in place. Instead, the government provides a no cost, best effort service, but one based on a modern architecture and adequate capacity.

5. The Existing PACER System Does Not Need to Change Immediately

Under our proposal, the AO would not need to modify the existing PACER system as the first order of action. Indeed, we have no problem if the AO wants to keep the current charging regime in place for the time being. We don't think that is good public policy, but if there are alternative sites to access an always-current PACER, we don't care.

Under this proposal, many lawyers would continue to access the PACER system, particularly for access to sealed documents. The current system of petitioning a judge for free access to PACER would stay in place. Judges and clerks who wish to use the current system, and the public using courthouse access terminals, could continue to use the system.

What would be different however, is that the public, and especially the bar, would have access to a number of alternatives, such as Lexis, West, Fastcase, PACER Pro, UniCourt, Casetext, Justia, the Internet Archive, and Court Listener that would now have real-time access to court filings. Especially important is that free public access sites would have the same real-time comprehensive access to the data as commercial legal services, as opposed to the current "economy class" position the general public enjoys.

There is another important side effect of making the bulk data available for free, and that is modern computer science can be used by researchers to help make the data more accessible. For example, researchers could use advanced techniques for looking for PII and use that information to notify lawyers that they may have inadvertently filed documents that violate the Judicial Conference rules. Likewise, the courts could be presented with reports that inform them about how well lawyers are observing their privacy obligations.

Working with "big data," using techniques known as "machine learning" or "text and data mining," is at the forefront of modern computer science. You can see those techniques in action if you make a Google search and the documents you really want are at the top of the search results. Likewise, "big data" makes possible machine language translation into hundreds of languages, and it allows for the automatic recognition of speech as closed captions or transcripts.

In the field of legal informatics, there has been much exciting research on data, but only when it becomes available. Researchers have analyzed citation graphs on court cases to show which cases are important and in what contexts. For example, researchers analyzing Marbury v. Madison, a case we all know as a bedrock of Supreme Court jurisprudence, determined that the case did not really become important in federal jurisprudence until the late 1800s, with the advent of the industrial age and in particular the increased regulation of industry.

These same techniques can be used to scrub files for PII without removing the primary responsibility for ensuring compliance from the litigants.. The techniques can be used to look for disparities in access to justice, such as different results in similar circumstances based on the ethnicity of the litigants. The techniques can identify how judges rule, and which lawyers are successful in which kinds of cases. There is so much potential.

6. If the Existing PACER System Does Change, the Act Should Include More Specificity

We recognize that the AO has long wanted to build a better web site for access to PACER documents. We believe strongly that this process should begin only after bulk data has been provided for and after the "business model" for the system has changed from a significant barrier to public access to one where the users of the courts pay for their work.

Under our proposal, we suggest that the AO first focus on those two key points, at which time it can begin the construction of a new web site, using its own API as a tool. However, we believe the bill, as currently drafted, is not specific enough on some key features that need to be in a "next generation PACER." In particular, the current PACER system is not one system, it is 200 separate systems. The act should specify that a single system should be built.

With a single system, perhaps located in a commercial or government cloud, rather than one for each court, a number of things can be accomplished. In particular, the system can be more effectively secured, something that is absolutely imperative given the recent breach of sealed documents. A single system would also remove innumerable differences between courts in the presentation of records, with significant benefits for those trying to understand federal law.

A more modern architecture would also require much better data standards and schemas, and enforce modern accessibility standards. A more modern architecture can also integrate modern coding standards, so that the system can scale properly to handle spikes in usage and to integrate facilities such as PII screening. A modern architecture also allows the AO to quickly roll out new features and bug or security fixes in days instead of over several months as is now the case.

Finally, any system the AO builds must provide for vendor neutral citation. Today, the best one can do is take a case identifier and a mdocket number, but that is not how legal citation works. As a result, many judges wait until a company such as West put a judgment or order into F. Supp or Lexis puts it on their system and then use that citation. The Judicial Conference should be mandated to develop a system of citation for all docket entries on cases that can be used by the legal profession to precisely indicate which specific document they are referring to.

7. Summary of Key Points

In summary, our proposal has the following key points:

- The "business model" of the courts should be based on filing fees, not dissemination fees. Such a shift will preserve, indeed will likely increase, the resources that are available to the AO. This is a sustainable and justifiable business model for the service, one where the users of the system pay in a judicial process which the public may observe.
- Shifting filing fees to the users of the courts should be based on a progressive system, not a regressive one. Large users should pay more, and explicit exemptions for fees should be specified to maintain access to justice.
- The AO should provide a system for bulk access to data as a first priority, including a copy of the back file. This can be as simple as providing a single account that we can use to provide public access immediately.
- The AO needs to do a much better job screening incoming documents and the back file for PII violations.
- Going forward, the AO should provide a modern API that provides for programmatic access to filings.
- When the AO builds a "next generation" web site, it needs to be based on a modern and scalable architecture instead of the current system of 200 customized small sites. Accessibility, vendor neutral citation, far better security, and other specific features should be mandated.

8. Access to PACER Database is Crucial to the Functioning of our Judicial System

From Magna Carta and back further to the Twelve Tables of Rome, the principle that one cannot have the rule of law unless the law is promulgated has been foundational to our system of government. Our courts must function in the light of day, and a carefully guarded single system with no bulk access is the antithesis of the light of day in our modern Internet world. The windows to our courthouses must be opened.

Government data should not be hoarded. PACER data is the very lifeblood of our federal court system. That data should be promulgated widely so that the works of courts are viewable to our citizens. Government data is the basis of public access and is also the basis for so many companies that use this raw data as the fuel for their information industries. Imagine our modern stock market without ready access to Securities and Exchange filings. How could one possibly promote the progress of science and useful arts if the U.S. Patent database was behind a paywall?

We urge you to consider modifications to the Open Court Act to make this data truly accessible. The vision and motivations behind the Act as currently drafted is inspiring. The nonpartisan coalition of members that came together in the House and is now evident in the Senate is equally inspiring. By modifying this bill to truly unlock public access at an affordable price, your vision will become reality. The courts will have a sustainable source of fees based on filing. The public will have access to our courts. The legal information industry will provide better tools for the legal profession. Data scientists will be able to analyze the workings of our courts. None of that is possible today.

We thank you again for your leadership. Please don't hesitate to let us know what we can do to help.

/digitally signed/

Carl Malamud, President, Public Resource
Brewster Kahle, Librarian, Internet Archive
Corynne McSherry, Legal Director, Electronic Frontier Foundation
Jimmy Wales, Founder, Wikipedia
Tim Stanley, CEO, Justia
Ed Walters, CEO, Fastcase
Thomas R. Bruce, Co-Founder, Cornell Legal Information Institute
Josh Blandi, CEO, UniCourt
Pablo Arredondo, Co-Founder & Chief Product Officer, Casetext
Stephen J. Schultze, Technologist and Attorney
Mikey Dickerson, Founding Administrator, United States Digital Service
Aneesh Chopra, Former Chief Technology Officer of the United States
Todd Park, Former Chief Technology Officer of the United States
Megan Smith, Former Chief Technology Officer of the United States
Nick Sinai, Former Deputy Chief Technology Officer of the United States
Beth Simone Noveck, Former Deputy Chief Technology Officer of the United States
Jennifer Pahlka, Former Deputy Chief Technology Officer of the United States
Alexander Macgillivray, Former Deputy Chief Technology of the United States
Thomas Kalil, Former Deputy Director for Technology and Innovation, OSTP
Cori Zarek, Former Deputy Chief Technology Officer of the United States
Nicole Wong, Former Deputy Chief Technology Officer of the United States
Michael B. Toth, Former Chief Technology Officer, United States Intelligence Community
D.J. Patil, Former Chief Data Scientist of the United States