**Database Mining and Statistical Analysis of Governmental and Litigation Databases:**

**An Examination of Contemporary Legal Issues and Potential Solutions**

[I. INTRODUCTION 3](#_Toc362462010)

[II. THE AGGREGATION OF JUDICIAL DATA 6](#_Toc362462011)

[A. Federal Courts 7](#_Toc362462012)

[B. State Courts 9](#_Toc362462013)

[III. TRANSPARENCY AND PRIVACY 12](#_Toc362462014)

[A. Transparency Needs 12](#_Toc362462015)

[B. Privacy Concerns 15](#_Toc362462016)

[IV. CONTROLS 21](#_Toc362462017)

[A. Technological Solutions 21](#_Toc362462018)

[B. Organizational Solutions 23](#_Toc362462019)

[V. CONCLUSION 24](#_Toc362462020)

I. INTRODUCTION

There has been a lot of recent discourse about database mining and statistical analysis of commercial databases, especially as databases become larger, more aggregated, and more widely used. The discussion has focused on the transparency and privacy aspects of becoming a data society with the bulk of the reflection on commercial databases; however, governmental databases and, specifically, legal databases are also concerns as they grow and become as large, if not, larger than the private databases. While the majority of the contemporary discussion has been concerned with the growth of corporate databases, any concern that society has for the private sphere should not overlook the same concerns in the governmental sphere.

A very recent TIME magazine article by Joel Stein titled, "*Data Mining: how companies now know everything about you?"* echoed a current societal paranoia of commercial databases.[[1]](#endnote-1) Companies are leveraging the latest data-mining and statistical analysis tools to understand these databases and understand us.[[2]](#endnote-2) Stein describes how Google, Facebook, Intellidyn, BlueKai, and EXelate have large databases containing personal information of the public which can, and sometimes is, being sold to marketing companies to predict consumer patterns.[[3]](#endnote-3) Stein uses an allegory of the 18th century circular prison invented by Jeremy Bentham to describe the paranoia and chilling effect these large corporate databases might start having on society.[[4]](#endnote-4) In a circular prison, continual observation of the prisoner is not needed because the prisoner looks up and can't tell if he is being observed and therefore always assumes he is. Like the prisoner the modern internet user doesn’t know when he or she is being monitored and always assumes he or she is. As these corporate databases are leveraged more and more to predict our patterns, society may curb its own patterns of behavior to either limit or utilize these predictive analysis.

Aside from the privacy and transparency concerns, databases and analytics are finding positive uses, and there have been some commentators that have explored the influence of the coming era of mega-databases and data-mining on both governmental and judicial databases. Edward K. Cheng, reviewing Ian Ayres’ Super Crunchers, discusses the prevalence of statistical analyses of large databases and how successful techniques in the private sphere will lead to techniques utilized in the legal sphere.[[5]](#endnote-5) He notes successful uses in the private industry including film companies that predict blockbusters, the matchmaking of couples, and companies that offer palliatives to keep their statistically best customers placated.[[6]](#endnote-6) Cheng also describes some uses in the public sphere and, more specifically, the legal database sphere.[[7]](#endnote-7) For example, Super Crunchers gives the example of the Supreme Court Forecasting Project which ran an analysis on the Supreme Court.[[8]](#endnote-8) The project used the National Science Foundation’s Supreme Court Database.[[9]](#endnote-9) The purpose of the project was to compare the ability of statistical analyses of machines with the ability of well established human legal experts to predict the outcome of cases. The statistical model had 75% predictive accuracy where human legal experts only had 59.1%.[[10]](#endnote-10) The project showed the potential predictability powers of data and statistical analysis in both the governmental and legal realm. Cheng recommends that courts should be using more database analytics, not only because they promise greater accuracy, but also because they provide the consistency and transparency to which the law often aspires.[[11]](#endnote-11)

Cheng has a point. Transparency and consistency are valued policies in our judicial system. Transparency values are echoed in the 6th amendment of the Constitution which grants the right of the accused a public and speedy trial, which defines the right to keeping the trial open to the public eye.[[12]](#endnote-12) The founders wanted to keep trials open to the public because open trials are fairer. Justice Hugo Black said that open judicial proceedings are “a safeguard against any attempt to employ our courts as instruments of persecution. The knowledge that every criminal trial is subject to contemporaneous review in the forum of public opinion is an effective restraint on possible abuse of judicial power.”[[13]](#endnote-13) The use of statistics and database mining can uncover hidden knowledge by pointing out information that was obscured from normal view. Statistics and data mining can also take a lot of information and put it in a form that is easily digestible to the human mind. Statistics and data mining can help legal practitioners navigate difficult decisions and make sound and informed decisions.

Privacy concerns have slowed access to and analysis of governmental and judicial databases. What once considered open and public data, now has safeguards to prevent violations of privacy. Factors creating these concerns are the ubiquity and permanence of electronic storage over traditional paper based storage.[[14]](#endnote-14) The paper based systems allowed for a protective barrier of privacy because retrieval of data has a higher cost.[[15]](#endnote-15) Electronic data is accessible in a cheap and easy fashion and the risk of personal information leaking out is exponentially higher in an electronic system. The old forms of legal protection in regards to privacy and transparency will have to be examined and change. Access to and mining of governmental and judicial databases will require a delicate balance of transparency and privacy.[[16]](#endnote-16)

We are definitely moving into the era of large aggregated databases with powerful tools that statisticians and data-miners can access and use in ways never before foreseen. These questions will become more pressing in the coming years as more aggregated databases are being housed, operated, and controlled by both private and state organizations. Unless we live as complete hermits, there is no way to exist in modern society without leaving information traces, wherever we go.[[17]](#endnote-17) These databases and the analytical programs are powerful tools for us, and there is not much we can do to prevent this era; it is here, but society can examine and study this phenomenon in order to protect ourselves within the current legal framework. While the issue of databases and analytics can be applied to both governmental and judicial database, in this paper, I will focus the growth and aggregation of judicial data, the privacy and transparency tensions, and two possible solutions to this tension.

II. THE AGGREGATION OF JUDICIAL DATA

As their private and governmental counterparts, legal databases are growing in this new era, being driven by a need for more efficiency and access. Different agencies within federal and state governments are looking to increase their own access to judicial databases and create interoperability between agencies, which is driving this growth. Federal and State governments have created their own initiatives to create and aggregate judicial data for their own use. The Federal government and courts have created a system called PACER to store and access electronic data, and state courts have used several different types of models.

# Federal Courts

The system that the Federal Court’s have used to aggregate judicial data is the PACER (Public Access to Court Electronic Records) system. Federal Level court data has been largely taken over by PACER. Since 1998, PACER has been fully accessible on the web.[[18]](#endnote-18) PACER has largely functioned as a system for electronic document upload and retrieval.[[19]](#endnote-19) The electronic documents are in the form of images or PDFs. It should be noted that electronic documents do have some of the protections of old paper documents because the data is buried within the computer code of the document. However, the data from electronic documents can be extracted from the documents through Optical Character Recognition, digital scrapping the document, or processes pulling text from the documents.

PACER has been quickly moving to a system that breaks document information into variables stored in large databases. For example, basic docketing information has been extended to include the names of the parties, attorneys, general type of action, and other variables for document retrieval.[[20]](#endnote-20) The variables populate the documents when the user retrieves them. Over the past several years, PACER has been experimenting with a few court documents that have fillable PDF documents that write all the documents fields and variables to a database.[[21]](#endnote-21) PACER has expanded its scope and the number courts it covers; it now even includes digital audio recordings of court proceedings.[[22]](#endnote-22) There is no doubt that the scope and size of PACER’s data is increasing, and also the ability to run analytics across the data.

While agencies and courts have access to this data, one barrier for the public on the PACER system is the costs of retrieval, where it is currently about $0.08 per page.[[23]](#endnote-23) To do an accurate, data analysis on judicial proceedings in PACER, a statistician would need to run an analysis across many cases. The amount of data and documents needed to make such a statistical analysis would require a high fee for the pulling of many documents and exclude those individuals and entities with little resources.

Several third-party organizations have been going around the cost barrier by harnessing group processing. One company that follows this formula is RECAP, which provides a Firefox browser plug-in for users of the PACER system, tracking PACER use while they are searching PACER.[[24]](#endnote-24) Once the user has retrieved a document or data from PACER, he can submit the data and documents into the RECAP repository.[[25]](#endnote-25) This data will be available for future users, and RECAP will indicate that they are in the repository and can be pulled for free.[[26]](#endnote-26) The group processing method will become more common across other forms of judicial and governmental data due to the cost barriers of public access.

Individual organizations have been recreating the governmental and judicial data. This paper previously noted the Supreme Court Database that predicted judicial outcomes. Volunteers to the Supreme Court Database take court opinions and translate them into variables that can be used for data-mining and statistical analysis. They are essentially recreating the court and litigation data. With the expansion of technology a lot of this work is being automated and more of these organizations will create their own judicial data. In the legal field, larger firms will be able to finance the creation of court and litigation databases, creating organizations that retool the public available judicial data for their own private use. The freeing and creation of judicial databases will probably be driven more by market demands than a real need for public access without some kind of legislative controls on access.

Federal Legislation has looked into the public access to judicial databases like PACER, but legislation has largely been inadequate.[[27]](#endnote-27) For example, the E-Government Act of 2002 was an attempt to take government data and create requirements that would facilitate public access.[[28]](#endnote-28) The act gave support requirements and spelled out specific guidelines for government agencies using the internet as a means to deliver information.[[29]](#endnote-29) The act has been successfully in certain areas at getting government data to the public. However, the act has been inefficient in freeing electronic data because it was more focused on areas with public documents and only devoted one small section to electronic data. [[30]](#endnote-30) The legislatures will need to step in further to facilitate public access to governmental and judicial databases.

# State Courts

States have created some comprehensive judicial databases but have largely been a hodge podge of solutions because of issues of cost, access, and size. The state courts deal with more cases, more court dockets, and more agencies; which make the hurdles higher. Like PACER, many states have created their own database and document storage regimes.[[31]](#endnote-31) New York’s model, which created a commission report in 2004 and then passed legislation in 2005 to launch its pilot program, created a centralized electronic filing system at the state trial level.[[32]](#endnote-32) The goal of the New York system was to have a central storage that provided interoperability to the many judicial branches and agencies. The original electronic storage was limited to only 15 counties, but, due to its success, it has been expanding annually.[[33]](#endnote-33) New York is a promising strategy for agency aggregation and interoperability, since it was done at the state level.[[34]](#endnote-34) New York, having a legislative directive at the state level, has been relatively successful in trying to integrate the vast number of court and government agency databases into one central location.[[35]](#endnote-35) Any public analytics on the New York system could be done since it is centralized, but, like Pacer, public access has been limited.

California, on the other hand, created the Esources initiative to provide incentives to the individual courts to build their own electronic storage.[[36]](#endnote-36) California requires agencies to institute the databases and interoperability themselves by providing incentives for data connectivity and aggregation. This is not the most desirable system for database aggregation because the courts in the state’s 58 counties have more than 200 case management systems.[[37]](#endnote-37) Most of these case management systems have very little cross-system compatibility or interoperability.[[38]](#endnote-38) Having the data in disperse databases with little cross compatibility, makes it hard to mine or do statistical analysis with both internal and public analytics. The Administrative Office of the California Judicial branch and the state Judicial Council have tried to reign the system in but the control of the database systems still lie with individual courts. [[39]](#endnote-39)

 Some states like Colorado have outsourced a lot of their data design, aggregation, and interoperability to commercial service providers like LexisNexis or ICON.[[40]](#endnote-40) Colorado’s system is named the Colorado Integrated Criminal Justice Information System (CICJIS).[[41]](#endnote-41) Like the New York system, Colorado is trying to centralize its data for efficiency and access reasons. However, Colorado has created centralization with a “virtual dataset” that utilizes middleware and data exchanges between the various courts and state agencies.”[[42]](#endnote-42) The best way to describe the Colorado model is to think of it as a cross between New York and California. It has individual data information at the local court level but has created a middleware to connect these individual courts. Colorado collaborates between five state agencies: Colorado Department of Public Safety/Colorado Bureau of Investigation (CBI), Colorado District Attorneys Council (CDAC), Colorado Department of Corrections (DOC), Colorado Department of Human Services/Division of Youth Corrections (DYC), Colorado Judicial Branch.[[43]](#endnote-43) The programming is the responsibility of the individual agency.[[44]](#endnote-44)

Colorado’s model is a promising method for internal data analytics; however, it has public access problems as the state agencies have wide latitude and control over the data, while the public and its needs are being largely left out of these design considerations. There is little access to the database by the public that would allow for comprehensive data-mining or statistical analysis. For example in the Colorado model, certain agencies can request to gain access through a separate web portal; however, the web portal doesn’t allow for large dataset retrieval or for access to all variables.[[45]](#endnote-45) The portal only gives its data in an HTML based web page, which could only be extracted by inefficient means like screen scraping. External agencies need to request access to this portal. If the agency or individual can’t access the portal, they can still obtain the court information through a public portal located at the State Court’s website. However, the public portal has very limited data and, like the PACER system, requires payment for documents, which makes large statistical analysis virtually impossible.[[46]](#endnote-46) The public, as opposed to the state, is largely limited in its access to court and litigation data especially data that can be mined or statistically analyzed.

III. TRANSPARENCY AND PRIVACY

While the legal system understands the need for transparency, what is driving this lack of access to the Federal and State judicial databases is a concern for privacy.

# Transparency Needs

The goals of transparency are to shed light on governmental activities.[[47]](#endnote-47) Lynn M. LoPucki’s law review article, *Court-System Transparency,* enumerates the benefits of transparency as being (1) Exposure and Reduction of Corruption, (2) Enhancement of Legislative Control over the Courts, (3) Popularization of the Law, and (4) the Prediction of Litigation Outcomes.[[48]](#endnote-48)

Exposure and reduction of corruption happens when the government makes public its transactions allowing the public to discover corrupt transactions. The fear and threat of discovery from the public also deters corrupt governmental transactions.[[49]](#endnote-49) In relation to database analytics, statistics on court databases can highligh corrupt and biased judicial opinions creating fairer and more just outcomes. [[50]](#endnote-50)

Furthermore, transparency provides enhancement of legislative control over the court's system. Legislatures will be able to measure the judicial outcomes in a statistical manner to help determine if the results of statutes are following the original policies they desired. The legislature can get insightful information on how the courts interpret their laws through the data analytics.[[51]](#endnote-51) For example, between 1991 and 1996, Delaware required that businesses filing for chapter 11 confirm to the courts that the bankruptcy would not be followed by liquidation or other debtor related reorganization.[[52]](#endnote-52) The judges continued this policy following the requirements of the legislature. A statistical analysis of court proceedings and the businesses had been run in 1996, which found that 54% of the companies that had made the confirmation had filed a second bankruptcy or had been liquidated in contradiction to what the judges had predicted following the legislative guidelines.[[53]](#endnote-53) Delaware judges were able to heighten the confirmation requirement based on this data.[[54]](#endnote-54) Both the legislature and judicial branch in Delaware had to reconsider the confirmation requirement.

Transparency popularizes the law by revealing to the public the rules that govern them.[[55]](#endnote-55) The statistical analysis via transparency would make the public more aware of what actually rules or governs them. An example found by Lynn M. LoPucki, in a paper on transparency and data, would start with a regression analysis on the body of cases challenging clickwrap contracts. Clickwrap contracts are the ubiquitous internet contracts that users click on to gain access to online websites. Having access to judicial data a regression analysis could be done to uncover which clickwrap terms were being enforced and the statistical likelihood of enforcement.[[56]](#endnote-56) They could be published in an academic journal, provided to the media, displayed on the internet, or actually embedded in the software to make them readily available to users.[[57]](#endnote-57) The software end-user could more intelligently decide whether or when to enter into the agreement.[[58]](#endnote-58) As click-wrap contracts become larger and more ubiquitous, many users don’t pay them much attention, but making statistically important sections more popular, the end-user can verify those sections before signing into the contract.

Transparency also helps lawyers and laypeople in the prediction of litigation outcomes.[[59]](#endnote-59) Oliver Wendell Holmes’s defined law when he wrote that “[t]he prophecies of what the courts will do in fact, and nothing more pretentious, are what I mean by the law.”[[60]](#endnote-60) Delaware courts again provide an example, with the discovery of high re-filing and failure rates among reorganized large, public companies.[[61]](#endnote-61) The law entitles public companies to choose their bankruptcy courts, and in the early 1990s, companies began using the court in Delaware.[[62]](#endnote-62) From 1990, the Delaware court’s docket for large, public-companies increased from zero to eighty-seven percent in ten years.[[63]](#endnote-63) Delaware gained the reputation from bankruptcy professionals and academics as the most sophisticated in the nation, and other courts began emulating their methods.[[64]](#endnote-64) In March 2000, a statistical analysis was run on the companies during this period using the Bankruptcy Research Database.[[65]](#endnote-65) The analysis found that the outcomes to these filings in Delaware failed at rates two to seven times as high as those filing in all other courts.[[66]](#endnote-66) The experts and the industry had predicted that the Delaware and New York courts were having the best outcomes for corporate reorganization within their judicial system when in fact, in one aspect, they were at the bottom.[[67]](#endnote-67)

Some legislative requirements at the federal level have been enacted to provide for more transparency to the public. For example, the Federal Freedom of Information Act is an attempt to provide transparency.[[68]](#endnote-68) The purpose of FOIA was to require disclosure for information that would shed light on the conduct of any Government agency or official for public use.[[69]](#endnote-69) However, the vast majority of FOIA requests are made by businesses for commercial purposes.[[70]](#endnote-70) The FOIA has turned many government agencies into “information brokers” rather than “a examining and monitoring how the government works.[[71]](#endnote-71) In order for court and litigation data to be accessed in a meaningful way and public way, legislatures are going to need to do more than require open access.

# Privacy Concerns

In the push for more transparent government and judicial data, the primary concern is that we may inadvertently create privacy problems. With the ubiquity of data, this loss of privacy may affect the judicial process by putting a chill on openness. Witnesses and juries that participate in the judicial process may change their behavior in the courtroom for fear that personal information may leak out. Parties to law suits will also have to worry about their information being used to harm them by ill intentioned individuals using it as blackmail or other harms.

The Constitution does not explicitly mention the right to privacy; however, the Supreme Court has repeatedly recognized that a right of personal privacy does exist under the Constitution.[[72]](#endnote-72) Privacy is a person’s right “in avoiding disclosure of personal matters” and the individual’s right to control these disclosures.[[73]](#endnote-73) The shift from a paper based system to an electronic system, as discussed previously, is causing concerns because paper documents gave an added step to protect private information. This barrier can be seen in Professor Harry Surden’s article, *Structural Rights in Privacy*, where he describes the physical costs of obtaining paper documents and the difficulty of searching through large amounts of paper documents for one particular piece of information.[[74]](#endnote-74) Searching all the public records and paper filings would be like searching for the proverbial needle in the haystack.[[75]](#endnote-75) However, with electronic databases, that extra barrier is being whittled away because one never needs to leave the office to gain access to a wider number of internet databases. Using data-mining software and other software allows an individual to search through all the hay in the hays stack at astounding speeds. We have tools that process these haystacks very easily. Furthermore, the permanency of data allows for easier access since it can easily be transferred to a new medium once the old medium becomes out of date without degrading the data[[76]](#endnote-76). In the paper based era, the data was a lot harder to transfer and would usually evaporate with the demise of the original medium. Digital data can move easily from medium to medium even as these physical mediums begin to degrade creating a longer perpetual shelf life in the new electronic system.

Technology allows for the comparing of databases, and data-miners can combine these sources to find personal information. There is a good reason for the privacy concern as people’s information has been used to steal identity, blackmail an individual, or intentionally disclosing harmful information. However, the government is already selling our data, purchasing data, and displaying private data.

The government has already been selling our data. In 1981, ten companies made long term arrangements with the Census Bureau to require the Bureau to track special data by zip code, where the companies would have exclusive access to this data.[[77]](#endnote-77) The problem goes the other way as the government is also purchasing data from private organizations to enhance their databases. Currently, the FBI and IRS purchase data from organizations that have been gathering and aggregating data from other private sources.[[78]](#endnote-78) A private company called ChoicePoint, Inc. has amassed a database of ten billion records from public records around the country and then combines it with other data.[[79]](#endnote-79) They have had contracts with at least thirty-five federal agencies; including the Justice Department which signed an $8 million contract and the IRS that signed a contract between $8 and $12 million.[[80]](#endnote-80) Our privacy is already being degraded, and ignoring the transparency concerns to protect privacy will hurt both privacy and transparency. By accepting a position that the public has a right to the data, the public, government, and the courts can begin to establish criteria and controls to allow access to protect privacy of this data.

The government has also already started to display private information to the public in other ways as well. The reasons for displaying private data have been expanding, individuals have been permanently labeled by this data, powerless subsections of the population are being disproportionately affected by this release, and third parties are finding ways to circumvent the privacy protections. Most states have adopted Megan’s Laws that mandate the disclosure of personal information of convicted sex offenders.[[81]](#endnote-81) In *Russell v. Gregoire*, the ninth circuit court of appeals upheld a Washington’s sex offender statute, which provided for the electronic, internet disclosure of the offenders personal information.[[82]](#endnote-82) The court held that the information was not private because it was “already fully available to the public.”[[83]](#endnote-83)

This creates another problem, a lot of times data can be inaccurate.[[84]](#endnote-84) Fully releasing the data, we can correct it, but if we don’t know about the data we can’t correct. Decisions about citizens could be made on this false information without the individual citizen knowing. Furthermore, governments have had a long history of social control and this use of permanent data on private individual could be used as a societal control.[[85]](#endnote-85) Many states have laws to prevent felons from participating in juries or voting and an individual’s felony status is stored in these court and litigation databases. As the state is more able to control and create data, it could easily start reducing rights for other things. It has been argued that using data to track felons and prevent them from rights has created a new era of “Dred Scott,” since the population effected by this activity is largely minorities and lower income.[[86]](#endnote-86) Both at the Federal and State level, these initiatives to permanently track citizen data and have created citizens that are barred from public housing, barred from food stamps, forced to “check the box” on employment applications, denied licenses to a wide range of professions, and locked out of the main stream society.[[87]](#endnote-87) It has been argued that electronic data is creating a permanent caste system out of certain sections of the population.[[88]](#endnote-88) Since the 1980s of more than 31 million people in the United States that have an arrest for a non-violent drug offense, the large majority are poor or minorities.[[89]](#endnote-89) Those sectors of the population that are labeled don’t have influence, power, or money to protect their own data, individually.

Several federal mandates have imposed data collection and access requirements on state courts and agencies including the National Child Protection Act of 1993, the Brady Handgun Violence Prevention Act, the Jacob Wetterling Act (including Megan’s Law), the Pam Lychner Act, the 1994 Violent Crime Control and the Law Enforcement Act.[[90]](#endnote-90) Initiatives and agency rules regulating immigration have added database interoperability between state agencies, the FBI, and ICE like the Secure Communities of U.S. Immigration and Customs Enforcement.[[91]](#endnote-91) Governmental agencies are using data to track, label, and limit rights to a widening class of citizens. This is all being done under the justification of public use while an effective access to the data that would show the outcomes of these policies is either un-accessible or not allowed.

Back to judicial databases, the Supreme Court has been reluctant to broaden this justification of public information to statistical analysis of the government databases. *McClesky v. Kemp* highlights this fear of databases and analytics as Justice Powell states, “[I]f we accepted McCleskey's [racial bias claim],…we could soon be faced with similar claims as to other types of penalty [and]…claims could apply with equally logical force to statistical disparities that correlate with the race or sex of other actors in the criminal justice system.”[[92]](#endnote-92) There is a disparity between justifying the release of information with laws like Megan’s Law under the need for public knowledge and the McCleskey opinion. This seemingly double standard that individual information can be released under the justification of public knowledge but data that actually exposes what the Court is doing remains closed due to privacy concerns highlights the problem that the fear of protecting privacy is harming both privacy and transparency.

Furthermore, privacy is being eroded as private entities are recreating or extracting the data. If the courts took a more public access approach, they could assert more control by allowing for adequate controls to prevent disclosure of personal information. Organizations that spend resources to mimic judicial data would have to compete with organizations gaining access to the data by following the rules. Accepting and allowing for more transparency will allow the courts and government to create controls and find solutions help prevent loss of privacy.

IV. THE CONTROLS

The government and courts can use technological and organizational solutions allowing for transparency and mitigating privacy concerns.

# Technological Solutions

One way out of the dilemma is to allow restricted access using technology.[[93]](#endnote-93) The Global Justice XML Data Model being developed by the United States Department of Justice (DOJ) and the LegalXML project of the Organization for the Advancement of Structured Information Standards (OASIS),[[94]](#endnote-94) rely on XML to integrate the data across several governmental and judicial databases allowing interoperability and data exchange yet providing leveled access control of data.[[95]](#endnote-95) Extensible Markup Language, or "XML," transmits both data and the meaning of the data by assigning data "tags" to define both the name of a data element and the format of the data within that element.[[96]](#endnote-96) The tags can be formatted to identify the different levels of access. This would allow for a connection to a large database by a diverse number of platforms and entities while controlling the individual data at specific levels of access.[[97]](#endnote-97)

An example of the problem of access comes from the Conference of Chief Justices (CCJ) and the Conference of State Court Administrators (COSCA), which have published a set of model guidelines for granting public access to court records.[[98]](#endnote-98) The guidelines require the removal of a whole electronic document if it contains private information, even though the whole document is available at the courthouse.[[99]](#endnote-99) This creates a contradiction where the public is denied access to records it should have electronically but allowed to get private information from the courthouse that it may not be allowed to have electronically. XML reference tags could be written into PDFs or Word documents while automatically redacting individual private information based on the tags and guidelines. Furthermore, governments could utilize these reference tags to control the release of public data in large scale data-mining by allowing access based on reference tags and preventing the harmful disclosure of personal data.[[100]](#endnote-100)

# Organizational Solutions

One very intriguing solution is suggested by Peter A. Winn from the Department of Justice. Mr. Winn compares governmental databases to federal lands and the allocation of environmental resources like minerals.[[101]](#endnote-101) The governmental agency weighs the cost-benefit analysis between the companies extracting the minerals and the negative externalities, pollution or destruction, caused by the mining activity. The federal and state agencies consider all the potential costs and benefits when granting permission to a mining company to extract minerals.[[102]](#endnote-102) Specialized agencies conduct tests and propose rules that reflect a more systematic cost-benefit analysis that weighs the “overall” social welfare.[[103]](#endnote-103) “In a similar way, the problems that have been identified with the system of judicial data suggest that we look to the earlier model of environmental protection law.”[[104]](#endnote-104) Judicial data, like public lands, is a public resource. What Winn proposes is an audit and oversight management agency to address the problems of managing governmental data from the point of view of the judicial “ecosystem”.[[105]](#endnote-105) Winn advices using a public body like the Administrative Office of U.S. Courts, presently charged with running the PACER system. Entities wanting access to the data could petition this data agency for permission to mine the data. Organizations concerned with privacy would have a say in protecting the data, the way organizations like the Sierra Club protect resources on public lands.[[106]](#endnote-106) Under the guidance of federal or state agency, data mining organizations would be required to scrub data and remove all identifying information before republishing the data to the public.[[107]](#endnote-107) Contracts with these data aggregators, statistical analyzers, or data-miners would allow for adequate audit and oversight.[[108]](#endnote-108) Principals could be set in place to frame new rules and policies with mining entity by contract negotiation, by enforcement action, or by revoking the contractual license of companies engaged in harmful data mining activities.[[109]](#endnote-109)

V. CONCLUSION

There are large government databases, particularly court and litigation databases. We should advocate for increased transparency and access to these databases to understand the data and to keep government abuse in check. The fear of databases like the circular prisons can be alleviated by greater public access to the judicial databases. Amazing things are being done with amassing and analyzing of data in the private and public field. We are beginning to view our world from a more statistical point of view in a way that we have not been able to do in history. There are concerns to transparency, including privacy concerns, but they should not paralyze us into ignoring the problem. Ignoring the problem is making it worse by creating removing privacy where it might need more protection and failing to provide adequate transparency where needed. These privacy concerns can be mitigated with technological or organizational solutions. Looking at the data as a public good or public resource, we can find solutions to controlling the extraction of public data while still protecting our long cherished privacy and public access rights.

1. Joel Stein, *Data Mining: How Companies Now Know Everything About You*, TIME Magazine, Mar. 10, 2011, at 150. [↑](#endnote-ref-1)
2. *Id*. [↑](#endnote-ref-2)
3. *Id*. [↑](#endnote-ref-3)
4. *Id*. [↑](#endnote-ref-4)
5. Edward K. Cheng, *Will Quants Rule The (Legal) World?*, 107 Mich. L. Rev. 967 (2009). [↑](#endnote-ref-5)
6. *Id*. [↑](#endnote-ref-6)
7. *Id* at 975. [↑](#endnote-ref-7)
8. *Id*. [↑](#endnote-ref-8)
9. *Id*. [↑](#endnote-ref-9)
10. *Id*. [↑](#endnote-ref-10)
11. *Id* at 976. [↑](#endnote-ref-11)
12. Peter A. Winn, *Online Court Records: Balancing Judicial Accountability And Privacy In An Age Of Electronic Information*., 79 Wash. L. Rev. 307 (2004). [↑](#endnote-ref-12)
13. *Id* at 308. [↑](#endnote-ref-13)
14. *Id*. [↑](#endnote-ref-14)
15. *Id* At 315. [↑](#endnote-ref-15)
16. *Id* at 313. [↑](#endnote-ref-16)
17. Daniel J. Soloved1, *Access And Aggregation: Public Records, Privacy And The Constitution*, 86 Minn. L. Rev. 1137, 1177 (2002). [↑](#endnote-ref-17)
18. Peter W. Martin, *Online Access to Court Records - From Documents to Data, Particulars to Patterns*, (Cornell Legal Studies Research Paper No. 08-003, 2008) at 2, available at [http://ssrn.com/abstract=1107412](http://ssrn.com/abstract%3D1107412) (last visited Mar. 23, 2011) [hereinafter Online Access to Court Records]. [↑](#endnote-ref-18)
19. *Id*. [↑](#endnote-ref-19)
20. *Id*. [↑](#endnote-ref-20)
21. U.S. Trustee Program, Data Enabled Form Standard, available at: <http://www.usdoj.gov/ust/eo/bapcpa/defs/index.htm>. [↑](#endnote-ref-21)
22. *Id*. [↑](#endnote-ref-22)
23. Public Access to Court Electronic Records, Frequently Asked Questions: What Am I Charged for in CM/ECF?, available at: <http://www.pacer.gov/psc/efaq.html#CMECF>, (last visited Apr. 6, 2011) [hereinafter CM/ECF Charges]. [↑](#endnote-ref-23)
24. RECAP website, *Why It Matters?,* available at <https://www.recapthelaw.org/why-it-matters/> (last visited Mar. 23, 2011). [↑](#endnote-ref-24)
25. *Id*. [↑](#endnote-ref-25)
26. *Id*. [↑](#endnote-ref-26)
27. *Online Access to Court Records* at 8-10. [↑](#endnote-ref-27)
28. *Id*. [↑](#endnote-ref-28)
29. *Id*. [↑](#endnote-ref-29)
30. *Id*. [↑](#endnote-ref-30)
31. *Id* at 18. [↑](#endnote-ref-31)
32. *Id*. [↑](#endnote-ref-32)
33. *Id*. [↑](#endnote-ref-33)
34. *Id*. [↑](#endnote-ref-34)
35. *Id*. [↑](#endnote-ref-35)
36. *Id* at 20. [↑](#endnote-ref-36)
37. *Id*. [↑](#endnote-ref-37)
38. *Id*. [↑](#endnote-ref-38)
39. *Id*. [↑](#endnote-ref-39)
40. *Id* at 21. [↑](#endnote-ref-40)
41. Colorado Integrated Criminal Justice Information System website, *Welcome Page,* available at <http://www.colorado.gov/cs/Satellite/CICJIS/CIC/1184685280209>, (last visited Mar. 23, 2011). [↑](#endnote-ref-41)
42. Colorado Integrated Criminal Justice Information System website, *Connecting to CICJIS*, available at <http://www.colorado.gov/cs/Satellite/CICJIS/CIC/1187685227971>, (last visited Apr, 6, 2011) [hereinafter Connecting to CICJIS]. [↑](#endnote-ref-42)
43. Colorado Integrated Criminal Justice Information System website, *Cooperating Agencies,* available at <http://www.colorado.gov/cs/Satellite/CICJIS/CIC/1184685280254>, (last visited Mar. 23, 2011). [↑](#endnote-ref-43)
44. Connecting to CICJIS. [↑](#endnote-ref-44)
45. Colorado State Courts – Data Access website, *Login page*, available at <https://www.jbits.courts.state.co.us/pas/pubaccess/>, (last visited Apr. 6, 2011) [↑](#endnote-ref-45)
46. Colorado Courts website, *What is COCOURTS.com*?, available at <https://www.cocourts.com/>, (last visited Apr 6, 2011). [↑](#endnote-ref-46)
47. *Id* at 1173. [↑](#endnote-ref-47)
48. Lynn M. LoPucki, *Court-System Transparency*, 94 Iowa L. Rev. 481 (2009). [↑](#endnote-ref-48)
49. *Id* at 494. [↑](#endnote-ref-49)
50. *Id*. [↑](#endnote-ref-50)
51. *Id* at 496. [↑](#endnote-ref-51)
52. *Id* at 497. [↑](#endnote-ref-52)
53. *Id*. [↑](#endnote-ref-53)
54. *Id*. [↑](#endnote-ref-54)
55. *Id*. [↑](#endnote-ref-55)
56. *Id* at 498. [↑](#endnote-ref-56)
57. *Id*. [↑](#endnote-ref-57)
58. *Id*. [↑](#endnote-ref-58)
59. *Id*. [↑](#endnote-ref-59)
60. *Id*. [↑](#endnote-ref-60)
61. *Id*. [↑](#endnote-ref-61)
62. *Id*. [↑](#endnote-ref-62)
63. *Id*. [↑](#endnote-ref-63)
64. *Id*. [↑](#endnote-ref-64)
65. Id at 499. [↑](#endnote-ref-65)
66. Id. [↑](#endnote-ref-66)
67. *Id*. The difference in the performances of the three sets of courts during that period was statistically significant at the .001 level. [↑](#endnote-ref-67)
68. 86 Minn. L. Rev. 1137, 1140. [↑](#endnote-ref-68)
69. *Id* at 1186. [↑](#endnote-ref-69)
70. *Id*. [↑](#endnote-ref-70)
71. *Id*. [↑](#endnote-ref-71)
72. American Bankruptcy Law Journal, *Thinking About The Private Matters In Public Documents: Bankruptcy Privacy In An Electronic Age, 75 Am. Bankr. L.J. 437, 441 (2001)*. [↑](#endnote-ref-72)
73. *Id*. [↑](#endnote-ref-73)
74. Harry Surden, *Structural Rights in Privacy*, 60 SMU L. Rev. 1605, 1618 (2007). [↑](#endnote-ref-74)
75. 86 Minn. L. Rev. 1137, 1150. [↑](#endnote-ref-75)
76. Peter A. Winn, *Online Court Records: Balancing Judicial Accountability and Privacy in an Age of Electronic Information*, 79 Wash. L. Rev. 307 (2004). [↑](#endnote-ref-76)
77. *Id*. [↑](#endnote-ref-77)
78. *Id*. [↑](#endnote-ref-78)
79. *Id*. [↑](#endnote-ref-79)
80. *Id*. [↑](#endnote-ref-80)
81. *Id* at 1183. [↑](#endnote-ref-81)
82. *Id*. [↑](#endnote-ref-82)
83. *Id*. [↑](#endnote-ref-83)
84. 86 Minn. L. Rev. 1137, 1183. [↑](#endnote-ref-84)
85. *Id*. [↑](#endnote-ref-85)
86. Michelle Alexander, *The New Jim Crow: Mass Incarceration in the Age of Colorblindness* 239, n. 2 (1st ed. 2010). [↑](#endnote-ref-86)
87. *Id* at 92. [↑](#endnote-ref-87)
88. *Id* at 14. [↑](#endnote-ref-88)
89. *Id* at 55. [↑](#endnote-ref-89)
90. Gregory M. Silverman, *Rise of the Machines: Justice Information Systems and the Question of Public Access To Court Records Over The Internet*, 79 Wash. L. Rev. 175, 181 (2004). [↑](#endnote-ref-90)
91. U.S. Immigration and Customs Enforcement website, *Secure Communities*, available at: <http://www.ice.gov/secure_communities/> last visited Apr. 21, 2011. [↑](#endnote-ref-91)
92. *Mccleskey V. Kemp*, 481 U.S. 279 (1987). [↑](#endnote-ref-92)
93. Gregory M. Silverman, *Rise Of The Machines: Justice Information Systems And The Question Of Public Access To Court Records Over The Internet*, 79 Wash. L. Rev. 175, 186 (2004). [↑](#endnote-ref-93)
94. *Id*. at 199. [↑](#endnote-ref-94)
95. *Id*. [↑](#endnote-ref-95)
96. Global Justice XML (Archive), *Extensible Markup Language (XML) and Its Role in Supporting the Global Justice XML Data Model*, available at <http://www.it.ojp.gov/documents/What_is_XML_article.pdf>. [↑](#endnote-ref-96)
97. *Id*. [↑](#endnote-ref-97)
98. 79 Wash. L. Rev. 175, 206. [↑](#endnote-ref-98)
99. *Id*. [↑](#endnote-ref-99)
100. *Id*. [↑](#endnote-ref-100)
101. Peter A. Winn, *Judicial Information Management In An Electronic Age: Old Standards, New Challenges*, 3 Fed. Cts. L. Rev. 135, 174 (2009). [↑](#endnote-ref-101)
102. *Id* at 170. [↑](#endnote-ref-102)
103. *Id* at 170. [↑](#endnote-ref-103)
104. *Id.* [↑](#endnote-ref-104)
105. *Id*. [↑](#endnote-ref-105)
106. *Id* at 173. [↑](#endnote-ref-106)
107. *Id* at 174. [↑](#endnote-ref-107)
108. *Id*. [↑](#endnote-ref-108)
109. *Id* at 173. [↑](#endnote-ref-109)