Coleta de Dados Judiciais

Alexandre Costa

Coleta de Dados Judiciais

Alexandre Costa

Esse livro está à venda em http://leanpub.com/coleta

Essa versão foi publicada em 2022-02-18

Conteúdo

Coleta de dados judiciais **1**
2. Definição do universo pesquisado 3
3. Mitigando o viés de seleção 5
4. Obtendo dados previamente organizados 11
5. Obtendo dados dispersos em múltiplas páginas 21

Coleta de dados judiciais

Excerpt

1 Geração e localização de dadosO primeiro passo para uma pesquisa de dados é justamente o levantamento dos dados que serão analisados pelo pesquisador. Esse levantamento pode ser feito de duas formas: Geração de dados: quando o pesquisador precisa coletar dados que não foram ainda levantados, o que exige estratégias

—

1 Geração e localização de dados

O primeiro passo para uma pesquisa de dados é justamente o levantamento dos dados que serão analisados pelo pesquisador. Esse levantamento pode ser feito de duas formas:

1. Geração de dados: quando o pesquisador precisa coletar dados que não foram ainda levantados, o que exige estratégias por meio das quais é possível observar fatos (p.ex: por meio de observação direta, de observação participante ou de experimentos) ou coletar impressões (p.ex: por meio de pesquisas de opinião, de entrevistas, de grupos focais)
2. Localização de dados: quando o pesquisador parte de bancos de dados já coletados e organizados por procedimentos anteriores de pesquisa, limitando-se a identificar certos conjuntos de dados que podem ser utilizados em sua investigação.

Evidentemente, muitas pesquisas usam as duas formas: incorporam dados que já se encontram organizados e também produzem novos dados, o que gera um incremento nas informações disponíveis.

O foco da pesquisa de dados está na identificação de dados já disponíveis, estejam eles sistematizados ou não, e não no levan-

tamento primário de dados. Em outros momentos históricos, a disponibilidade de informações era baixa e uma parcela muito grande de qualquer pesquisa empírica era dedicada a realizar a coleta direta dos dados necessários para a aplicação das estratégias metodológicas escolhidas.

No momento atual, existe uma disponibilidade muito grande de dados que não foram devidamente tratados, de tal forma que a fronteira contemporânea da pesquisa está no desenvolvimento de nossa capacidade de coletar os dados disponíveis, organizá-los de forma adequada e classificá-los de maneiras produtivas, para eles poderem nos conduzir a conclusões originais.

O foco do curso Data Science e Direito[1] está na localização de dados, e não na geração de dados, que envolve o desenvolvimento de capacidades específicas, ligadas a cada tipo de procedimento de contato direto com a realidade. Nesse âmbito, a originalidade das pesquisas não está ligada à produção de dados primários novos, mas à produção de novas classificações, que apresentem de maneiras originais. Classificar decisões, partes, ou tipos de argumentos significa gerar novos dados, pois esse é um procedimento que enriquece nosso repertório de informações acerca do mundo.

Cabe ressaltar que informações e dados são palavras com significado muito semelhante, o que as faz ser intercambiáveis na maior parte dos contextos. As distinções conceituais realizadas entre esses vocábulos são tipicamente ligadas ao uso inglês, em que data é uma palavra usada normalmente no plural, para falar de um conjunto de informações (data set), enquanto information é uma palavra tipicamente usada no singular, para indicar que existe um conjunto de dados que, devidamente interpretados, oferecem informação acerca de um objeto.

Esse uso das palavras nos permite afirmar que a combinação de dados acerca dos exames positivos de Covid-19, combinados com dados demográficos, nos oferece informação acerca dos níveis de

[1] https://dsd.arcos.org.br/

contaminação de uma população; a combinação de dados sobre o momento de ingresso de um processo no tribunal, combinado com dados sobre a data de eventual julgamento, possibilitam termos informação sobre o tempo médio de tramitação. Porém, quando usamos as palavras no plural (informações e dados), seus significados se sobrepõem.

2. Definição do universo pesquisado

O problema de pesquisa da sua investigação deve ser suficientemente preciso para ser possível identificar acerca de que população você pretende falar. Nas pesquisas censitárias, o universo pesquisado coincide com o conjunto de objetos acerca dos quais você fará o levantamento de dados. Em pesquisas amostrais, você coletará dados sobre uma amostra, com o objetivo de fazer afirmações sobre um universo. Em estudos de caso, você falará apenas do objeto analisado, sem buscar inferências.

Como apontam Epstein e Martin, essa questão pode ser trivial em alguns casos, mas em outros ela pode ser complexa (2012). Por vezes, queremos falar do comportamento do STF, mas os dados que levantamos são acerca das decisões de certos processos julgados em um certo período (por exemplo, das ADIs ajuizadas entre 2015 e 2020). Ocorre que, com base em informações sobre esses processos, o seu universo pesquisado possivelmente será: o conjunto dessas decisões, e não o comportamento do STF.

Se quisermos falar dos comportamentos dos ministros do STF, temos de colher dados que nos permitam fazer inferências sobre esse grupo de pessoas, o que aponta para a necessidade de fazer recortes. Se levantarmos dados sobre os últimos 30 anos e buscarmos padrões de julgamento no conjunto desses dados, podemos encontrar correlações sem muito sentido.

As decisões do STF se dão dentro do contexto da atuação dos

outros poderes, e mudanças no contexto podem alterar o sentido de uma prática, ainda que ela permaneça aparentemente inalterada: uma pequena taxa de concessão de liminares em ADI pode ser indício de uma influência pequena da corte, mas também pode ser indício de uma forte autolimitação; um alto grau de decisões de prejudicialidade em ADI pode indicar uma influência pequena dos requerentes (que não conseguem pressionar pela inclusão dos seus processos na pauta), mas também pode indicar um poder político especialmente grande (que consegue revogar politicamente as leis impugnadas antes que o tribunal tenha chance de as julgar).

Quando estabelecemos com clareza qual é o nosso problema de pesquisa, podemos definir com relativa precisão quais são os dados que tentaremos levantar, visto que esses dados devem ser suficientes para fazer afirmações sólidas sobre o universo analisado. Porém, é comum que desejemos falar de universos maiores do que o nosso conjunto de dados permite.

Se queremos falar do comportamento da STF, não podemos levantar dados apenas sobre o controle concentrado. Se queremos falar do controle concentrado, mesmo que as ADIs sejam o maior conjunto das ações, pode ser necessário tratar conjuntamente das ADIs e das ADPFs. Outras vezes, acreditamos que podemos falar de um certo universo sem precisar adequadamente os seus limites. Se queremos falar das decisões favoráveis aos autores, talvez devamos considerar que as decisões de prejudicialidade são favoráveis (o que aumenta o universo). Se queremos falar de decisões de procedência em ADI, talvez as decisões que concedem liminares sejam tão satisfativas quanto as decisões finais.

O mais comum é que a análise dos dados nos mostre que o recorte que adotamos inicialmente não seja o mais adequado, seja porque os dados não nos permitem falar exatamente daquele recorte (mas de outros), seja porque o recorte foi muito impreciso (o que dificulta a identificação dos dados relevantes). Assim, devemos estar prontos a alterar o nosso problema de pesquisa para ele estar bem acoplado com os dados disponíveis, o que pode exigir vários ciclos de redi-

mensionamento recíproco do universo (afinando os conceitos para ser possível levantar dados adequados) e de redimensionamento dos dados a serem levantados (para que eles sejam capazes de trazer evidências sobre o universo definido).

3. Mitigando o viés de seleção

O viés de seleção ocorre quando as nossas conclusões são influenciadas pelos critérios que usamos para selecionar (ou excluir) os objetos que incluímos na nossa análise. No campo jurídico, trata-se de um viés extremamente comum e que compromete muitas das conclusões dos trabalhos acadêmicos em direito.

Um exemplo típico de sua atuação ocorre quando decidimos identificar quais são os entendimentos predominantes no STF com relação a um determinado tema e, para isso, decidimos estudar os casos emblemáticos, ou os casos de maior repercussão. É compreensível que uma análise qualitativa da argumentação exija pesquisas com um número pequeno de casos, pois não é possível fazer uma análise exaustiva de dezenas de decisões complexas no tempo de um doutorado, quanto mais de um mestrado.

Quanto maior o tempo consumido em cada análise (e algumas abordagens exigem um trabalho imenso, especialmente no caso de processos longos e decisões complexas), menor o número de elementos que poderão ingressar na nossa amostra. Porém, é muito comum que os juristas não digam que estão fazendo um estudo de caso (que não possibilita conclusões gerais), mas que estão analisando um conjunto de processos que são de alguma forma representativos do conjunto dos casos ou da jurisprudência dominante.

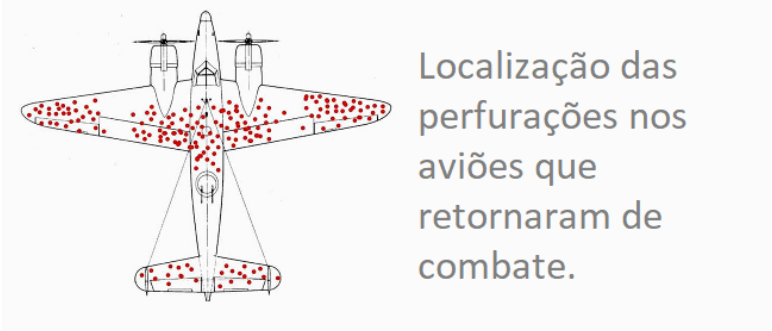
É certo que os trabalhos jurídicos ligados à dogmática não falam em amostras, pois não se trata de utilizar as técnicas estatísticas de inferir padrões de uma população a partir de uma amostra cuidadosamente definida. Porém, as abordagens jurídicas normalmente

estudam um grupo de processos (ou mesmo um processo específico) com o objetivo de inferir conclusões que ultrapassam os limites do caso concreto: estudamos casos para falar do entendimento da corte, da jurisprudência, da interpretação correta.

Quando estudamos algumas decisões do STF e pretendemos extrair desse trabalho conclusões sobre a jurisprudência do STF sobre o princípio da igualdade ou sobre o princípio da proporcionalidade, nós incluímos certas decisões como objetos de estudo porque a consideramos exemplares ou paradigmáticas. Isso faz com que tratemos esses casos como se eles fossem uma amostra, ou seja, um conjunto representativo da população acerca da qual se pretende fazer afirmações. Ocorre, porém, que é muito comum que esses critérios de relevância sejam bastante influenciados pelas nossas percepções individuais (os processos que nós consideramos mais relevantes) ou coletivas (os processos mais conhecidos, mais citados, mais célebres).

O viés de seleção não é apenas a interferência de nossas preferências pessoais, mas também pode ser a interferência de certas percepções coletivas, de certos processos institucionais por meio dos quais alguns elementos se tornam visíveis e outros permanecem ocultos.

A Wikipedia indica que um caso clássico desse viés ocorreu na segunda guerra, quando os militares fizeram um levantamento dos locais em que havia furos de balas nos aviões que voltavam de combate, e chegaram ao seguinte padrão:



Com base nesses dados, onde você sugeriria que os aviões deveriam ter a blindagem reforçada? W. Allen Wallis narra que os militares se inclinaram inicialmente a proteger as partes que tinham sido mais alvejadas, mas que esse diagnóstico foi contestado por Abraham Wald, um dos membros do Grupo de Pesquisa Estatística na Columbia University, que sugeriu justamente o oposto:

> "The military was inclined to provide protection for those parts that on returning planes showed the most hits. Wald assumed, on good evidence, that hits in combat were uniformly distributed over the planes. It follows that hits on the more vulnerable parts were less likely to be found on returning planes than hits on the less vulnerable parts, since planes receiving hits on the more vulnerable parts were less likely to return to provide data. From these premises, he devised methods for estimating vulnerability of various parts." (Casselman[2])

Embora a Wikipedia e vários outros sites repitam essa história de uma forma romanceada e o trabalho estatístico de Wald seja bem mais técnico e complexo do que sugerir o reforço das partes não atingidas nos aviões que retornaram (Wald, 1943[3]), trata-se de um

[2] http://www.ams.org/publicoutreach/feature-column/fc-2016-06
[3] https://apps.dtic.mil/dtic/tr/fulltext/u2/a091073.pdf

caso evidente em que uma análise pouco refletida poderia levar a resultados desastrosos, de modo que ela se tornou uma das bases para o que se veio a chamar de viés de sobrevivência (survivorship bias).

Para evitar o viés de seleção, a estratégia básica é usar amostras aleatórias, dentro do universo que se deseja descrever. Não podemos analisar o universo dos aviões em combate a partir de uma amostra dos aviões que retornaram. Não podemos analisar o universo das ADIs a partir do mapeamento daquelas que foram julgadas, pois os processos que levam ao julgamento de uma ação não são aleatórios. Não podemos descrever a atuação do STF a partir dos processos mais visíveis (na mídia, nas citações processuais, nos artigos científicos) porque essa visibilidade decorre de certos critérios de seleção.

Além disso, devemos ter em mente um exemplo que é dado por Marcos Silva:

> Imagine que queremos medir quanto um treinamento sobre **Data Science** aumenta a produtividade dos funcionários de uma empresa. Nossa primeira abordagem é mandar um e-mail para toda empresa dizendo que há 50 vagas e os primeiros a se cadastrarem farão o curso. Fazemos o treinamento e medimos a produtividade antes e depois do treinamento. Podemos dizer que essa diferença observada é culpa do treinamento? Infelizmente não, pois provavelmente as primeiras pessoas a se cadastrarem no programa não são aleatórias, provavelmente são as mais atentas, mais rápidas e até mesmo mais esforçadas para fazer um curso extra e ao medir o resultado do experimento teríamos que escolher aleatoriamente seus participantes. (Silva[4])

[4] https://medium.com/data-hackers/10-coisas-que-voc%C3%AA-precisa-saber-sobre-vi%C3%A9s-e-causalidade-af4e7ac644c8

Esse tipo de viés pode comprometer pesquisas baseadas na resposta voluntária a questionários, mas também pode comprometer pesquisas que avaliem o comportamento judicial. Várias pesquisas sobre o controle concentrado de constitucionalidade analisam conjuntos restritos de dados, especialmente de decisões. Mas certos conjuntos de decisões podem ser como os aviões que voltam de combate ou como as pessoas mais dispostas a aprender coisas novas: analisá-las pode nos dizer algo sobre a amostra, mas pode ser enganoso generalizar as conclusões.

Pode ser que certas inovações institucionais (como os julgamentos em lista ou a ampliação dos julgamentos virtuais) tenham impactos grandes no comportamento do STF. Mas também pode ocorrer que essas mudanças representem uma via que facilita a resolução rápida de alguns processos (como ADIs que esperavam há anos com liminar concedida, ou ADIs sobre assuntos repetitivos) que não seria estratégico julgar no plenário presencial, que admite um número muito restrito de processos. Com isso, é possível que analisar os processos julgados pelo STF em 2020 gere um conjunto de dados distorcido por questões não aleatórias, que pode distorcer as nossas percepções.

Se houvesse um número muito grande de decisões (de dezenas de milhares de decisões singulares, por exemplo), a adoção de uma amostra randômica poderia ser uma solução adequada. Porém, o fato de contarmos com poucas centenas de processos julgados a cada ano, não temos um conjunto de dados suficientemente amplo para que essas estratégias se apliquem de forma adequada, ainda mais quando as nossas questões enfocam subconjuntos muito restritos:

1. Será que processos ajuizados por partidos de oposição têm menos chance de ter uma decisão de procedência?
2. Nos processos em que se argumenta com base no princípio da igualdade, que tipo de critério de discriminação é considerado proporcional?

No discurso jurídico dogmático, um dos grandes desafios é observar um pequeno número de decisões (às vezes uma decisão apenas) e extrair delas (ou dela) um fundamento determinante, uma ratio decidendi que sirva como precedente, um paradigma adequado para julgamentos futuros. A dogmática busca estabelecer padrões estáveis a partir de um conjunto muito restrito e heterogêneo de elementos: textos legais, interpretações, decisões, etc.

Não existe propriamente uma tentativa de evitar o viés de seleção, mas o de justificar discursivamente o caráter paradigmático de determinadas teses jurídicas, que possibilita extrair das fontes do direito certas interpretações que devem guiar a atuação futura dos juristas. Nas abordagens quantitativas, um caso concreto nunca pode ser representativo. O que pode ser representativo é uma amostra randômica, de um tamanho adequado, que permita minimizar as margens de erro.

No campo jurídico, porém, o debate dogmático muitas vezes é voltado a definir se se uma decisão deve ser considerada um precedente adequado, sobre se um acórdão fixa ou não uma tese, sobre se um provimento alcança ou não um caso concreto. Tratamos de poucos elementos e tentamos extrair o máximo de orientação dessas balizas discursivas. Tentamos justificar discursivamente o caráter representativo desses elementos, o que nos coloca em choque com as abordagens científicas, que avaliam populações de elementos a partir de investigações voltadas a limitar, tanto quanto possível, os vieses de seleção e de confirmação.

No campo da pesquisa empírica em direito, há estratégias para mitigar esses vieses. Uma saída, viabilizada pela pesquisa em bancos de dados, é realizar pesquisas censitárias, que limitam o viés de seleção por meio do próprio abandono da amostragem. Quando isso não viável, pode ser possível ampliar as amostras, tornando-as mais variadas e, com isso, mitigando os vieses. De uma forma ou de outra, é preciso ter em mente que toda abordagem lida com um número finito de observações e que qualquer recorte (e qualquer classificação) interfere nos resultados. Essa é uma dificuldade que

pode ser mitigada, mas que não pode ser totalmente superada nas pesquisas em direito.

4. Obtendo dados previamente organizados

Uma das formas mais eficientes de trabalhar com informações é buscar bancos de dados previamente organizados por outros pesquisadores. Para além da economia de tempo, bancos de dados que já foram utilizados em outras pesquisas costumam ser mais maduros e podem propiciar diálogos com pesquisas anteriores.

Existem na internet uma série de dados acessíveis, que podem ser utilizados como base para que o pesquisador realize as suas análise. Portanto, é necessário que os pesquisadores conheçam os dados disponíveis para analisar se é possível enfrentar com base neles os problemas de pesquisa escolhidos.

Quando o pesquisador tem sorte, ele encontra os dados devidamente organizados em tabelas sistematizadas, completas e bem categorizadas. Mas esse grau de sorte é raro. O mais comum é que os dados existam de modo disperso em uma multiplicidade de páginas, que são acessíveis individualmente, mas que não são sistematizadas em bancos de dados abrangentes.

4.1 Bancos de dados judiciais

No judiciário, a maioria dos dados é disponível dessa forma porque o que se utiliza são bancos centrados em fornecer um serviço de acompanhamento processual. É preciso oferecer a cada advogado informações precisas sobre o estado atual dos processos e sobre os seus andamentos, o que é feito de forma eficaz por vários sistemas. Porém, esses sistemas são focados nos processos individuais, e não em populações de processos, o que não apenas faz com que os

processos sejam publicados de forma individualizada, mas também faz com que a arquitetura do sistema seja feita de modo a otimizar a observação processo a processo.

Se o foco dos sistemas fosse o de produzir mapas de populações de processos (em vez de fotografias individuais), seria necessário desenvolver sistemas de classificação mais robustos, que definissem os atributos de um modo mais preciso. No caso dos sistemas processuais, pode ser suficiente apresentar o inteiro teor das decisões, que serão lidos por pessoas envolvidas no processo e por ela interpretadas. Cada classificação envolve uma interpretação, e cabe a cada parte (e a seus advogados) interpretarem o sentido e o alcance das decisões judiciais.

Cada nova classificação gera a possibilidade de erros e de ambiguidades que podem gerar insegurança processual. Toda classificação feita pelos tribunais (p.ex.: processos repetitivos e individuais, processos simples e complexos, controle material ou formal, etc.) pode ser contestada e pode ter implicações na velocidade e no resultado dos julgamentos. Por isso mesmo, é de se esperar que os tribunais utilizem categorias muito descritivas (e pouco avaliativas) e que sejam muito literais com relação às disposições dos julgadores.

Uma base de dados pode classificar todas as extinções processuais a partir de uma classe geral de extinções processuais, mas um sistema de andamentos precisa dar aos advogados a informação de que o processo foi julgado prejudicado, ou que foi extinto sem julgamento de mérito ou que houve um indeferimento da inicial. Assim, a falta de um sistema complexo de classificações não é uma falha dos sistemas judiciais atuais, mas é uma decorrência previsível do fato de que esses sistemas são voltados a um acompanhamento processual.

Não existe um sistema classificatório objetivamente correto. Existem múltiplos sistemas classificatórios e cada um implica a adoção de certos critérios, que serão mais ou menos vagos e que terão utilidade para resolver determinados problemas. A academia pode

produzir múltiplos sistemas de classificação, alguns compatíveis entre si, outros concorrentes, e este é o seu papel: multiplicar as interpretações, enriquecer o conhecimento.

Esse não é o papel dos tribunais, cuja função é decidir e ser transparente com relação a suas decisões. Essa transparência exige que as informações tenham grande fidelidade com as decisões, e todo conjunto de decisões será composto por unidades que usam categorias diversas, que usam repertórios conceituais incompatíveis, que não formam um sistema claro.

4.2 Acesso por APIs

Para finalidade de pesquisa, o ótimo seria que os órgãos fornecessem uma Application Programming Interface (API), ou seja, uma interface que permitisse aos programadores construir programas que utilizassem os dados contidos nos sistemas do tribunal.

O que os tribunais disponibilizam tipicamente são páginas com informações de cada processo e também certos mecanismos de busca, mas eles não oferecem um acesso direto aos dados, para que os pesquisadores possam criar ferramentas que busquem informações dentro dos próprios bancos de dados. O que nós podemos fazer é acessar as páginas que o Tribunal disponibiliza para consulta e retirar de cada uma delas as informações desejadas, o que é uma forma adequada de interação para seres humanos, mas não para máquinas.

Uma API é uma interface (ou seja, um programa que dá acesso aos dados) otimizada para o uso de máquinas, de programas que pedirão os dados à API. Uma API voltada à pesquisa permitiria a consulta de vários processos ao mesmo tempo (em vez de um por vez) e permitiria a consulta de populações de processos, voltando como resposta uma base de dados consolidados, em vez de uma página individual para cada processo.

O TRT da 3a Região disponibilizou uma API para consultas sobre contratos, cujos termos estão definidos na página:

> https://portal.trt3.jus.br/internet/transparencia/dados-api/contratos[5]

Essa página indica o endereço no qual podem ser feitas consultas que retornam uma lista dos contratos celebrados pelo Tribunal ao longo de um determinado mês. O TRT3 esclarece que a pesquisa deve ser feita pelo endereço seguinte, indicando o mês e o ano que se pretende buscar:

> https://transparencia.trt3.jus.br/odata4/transparencia.1/transparencia/contrat

Além disso, a página indica quais são as informações oferecidas como resposta a essa solicitação, que consistirá em uma lista que indicará cada contrato e certos atributos de cada um deles (data de assinatura, minuta, valor, etc.), o que permite que o pesquisador elabore um programa capaz de incorporar à sua base de dados as informações que julgar relevante.

Assim, ter uma API não significa apenas dar acesso aos dados, mas criar um sistema específico para mediar a forma como outros programas (inclusive os que você vai aprender a fazer) poderiam solicitar dados ao Tribunal. E esse trabalho de criar uma API de informações processuais não foi assumido por tribunais que têm sistemas voltados a oferecer informações processuais e pesquisas de jurisprudência.

4.3 Lei de Acesso à Informação

A Lei de Acesso à Informação (Lei 12.527/2011) garante aos cidadãos o acesso a todas as informações disponíveis e impõe aos órgãos

[5] https://portal.trt3.jus.br/internet/transparencia/dados-api/contratos

públicos a obrigação de criar sistemas adequados de consulta e divulgação de dados, mas o fato é que os órgãos públicos (ao menos ainda) não desenvolveram sistemas adequados a prestar as informações disponíveis de forma adequada às pesquisas acadêmicas, que se interessam normalmente por populações de processos e não por situações individuais.

Tabelas que consolidam as informações são raras e as instituições judiciais não têm o dever de produzi-las, visto que eles foram expressamente desonerados dessa função pela Resolução 215/2015 do CNJ:

> Art. 12. Não serão atendidos pedidos de acesso à informação: [...]
> III – que exijam trabalhos adicionais de análise, interpretação ou consolidação de dados e informações, serviço de produção ou tratamento de dados que não seja de competência do órgão ou entidade;

Como o que se exige dos tribunais é a informação individual dos processos, a consolidação de informações na forma de uma tabela agregadora é vista normalmente como um trabalho extra, que não pode ser exigido. Apesar de não ser obrigatório, alguns tribunais ofereceram esse serviço de modo bastante amplo ao longo da década passada, especialmente o próprio STF. Porém, nos últimos anos, o Tribunal tem rejeitado pedidos de informações consolidadas feitas por pesquisadores, sob o argumento de que as informações já estavam disponíveis, tanto individualmente, quanto de forma agregada na página de Estatística do STF, que foi elaborada justamente para esse fim.

Ocorre, contudo, que essas páginas consolidadas não têm todas as informações disponíveis nos sistemas e não as oferecem do modo mais adequado para as pesquisas, sendo que essa mudança de postura administrativa foi um dos motivos pelos quais integrantes do Grupo de Pesquisa em Política e Direito se dedicaram a desenvolver

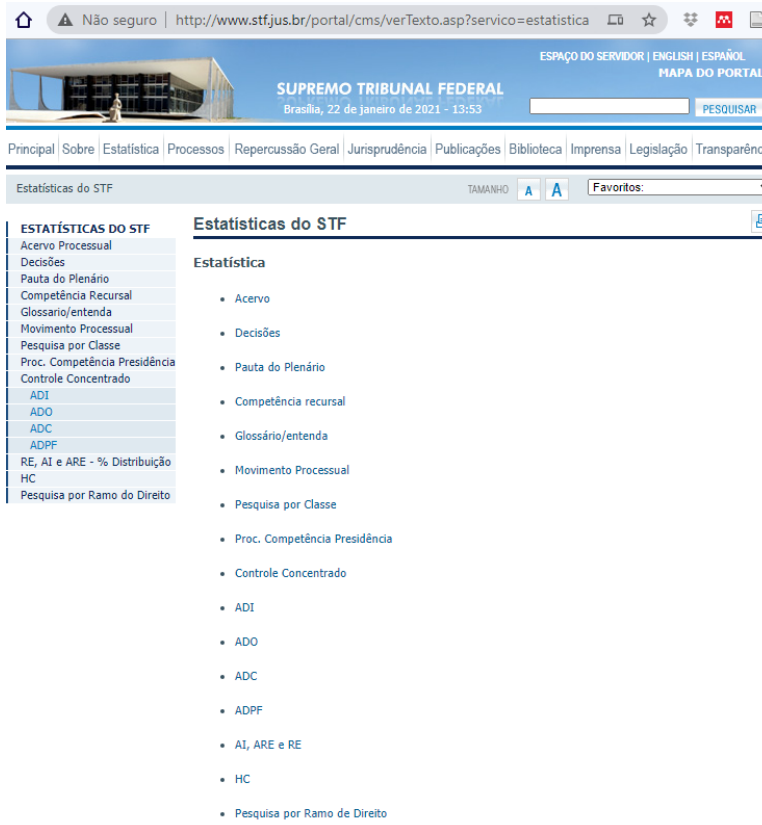
as habilidades necessárias para obter os dados diretamente nas páginas do Tribunal, por meio de programas de extração de dados.

4.4 Página de Estatística do STF

No lugar de oferecer as informações solicitadas pela Central do Cidadão, o STF investiu na formulação de uma página de informações mais complexa, que fosse capaz de oferecer ao público dados de forma organizada. O desenvolvimento dessa página aumentou a transparência, pois passou a divulgar não apenas dados fragmentados, mas também consolidações dos dados, que são úteis para a sociedade em geral.

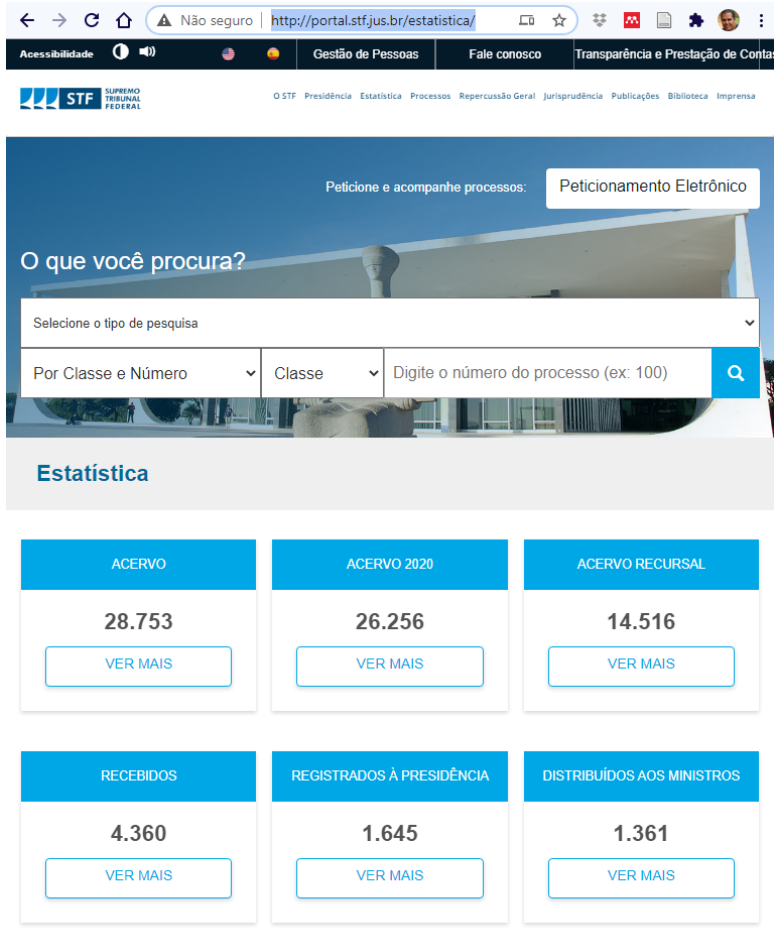
A versão anterior da página de estatística[6] ainda está disponível no portal do STF, e ela oferece algumas páginas com gráficos e os dados subjacentes em arquivos do tipo .mhtml, que mescla arquivos .html com outros elementos gráficos. Esses arquivos não são abertos de forma integrada com o navegador, mas são baixados e depois tendem a ser abertos pelo Microsoft Internet Explorer em uma janela separada. Dentro desses arquivos existem links para algumas tabelas do Excel (arquivos .xlsx) com os dados subjacentes.

[6] http://www.stf.jus.br/portal/cms/verTexto.asp?servico=estatistica



Em 2020, o STF lançou uma nova página de Estatística[7] , com um formato mais contemporâneo e amigável, no qual os gráficos estão disponíveis em .html, ou seja, em arquivos que são abertos de forma integrada com o navegador. São gráficos mais cuidados e interativos, compatíveis com os dashboards que tem sido utilizados atualmente e com o formato típico do PowerBI da Microsoft. Todavia, ainda é uma página em desenvolvimento, e vários dos dados só estão disponíveis na versão anterior.

[7] http://portal.stf.jus.br/estatistica/



Essa apresentação é mais cuidada e mais complexa do que a do modelo anterior, com 3 avanços importantes:

1. Ao final, existe um artigo em PDF com uma documentação da base de dados, esclarecendo o sentido das variáveis de modo a viabilizar uma melhor interpretação dos dados;
2. Os dados subjacentes aos gráficos podem ser baixados em versão .xlsx (arquivo de Excel), mas também em versão .csv, que é uma

forma de arquivo compatível com qualquer leitor de dados, o que viabiliza a sua manipulação direta por meio de Python;
3. Os gráficos são interativos, podendo ser modificados a partir da inserção de filtros, de forma a possibilitar uma exploração razoável dos dados na própria página.

4.5 Dimensionar as pesquisas aos bancos de dados disponíveis

Como os dados disponíveis de forma organizada (os bancos de dados) não são tão numerosos nem tão completos como todos gostariam, uma das estratégias de pesquisa mais eficientes é modelar o projeto de tal forma que o problema de pesquisa possa ser respondido a partir das informações disponíveis.

Uma das formas de realizar essa compatibilização é o pesquisador começar o processo por meio da análise cuidadosa dos bancos de dados, procurando neles padrões que justifiquem uma investigação mais detida. Esse é um tipo de trabalho que tende a ser mais rápido (visto que trabalha com dados já compilados) e pode chegar a conclusões interessantes.

A maioria das perguntas importantes depende da construção de bases de dados novas ou do desenvolvimento das bases disponíveis. Porém, como estamos em um tempo no qual as bases jurídicas existentes não foram suficientemente mapeadas e exploradas, consideramos que trabalhar com os dados existentes ainda é uma opção capaz de gerar pesquisas interessantes.

O artigo Controle de constitucionalidade no Brasil: eficácia das políticas de concentração e seletividade[8], publicado em 2016 na Revista DireitoGV (Qualis A1), seguiu essa estratégia. Douglas Zaidan encontrou no site do STF uma tabela com informações interessantes sobre o número de processos ajuizados e julgados ao longo dos anos

[8] https://www.scielo.br/scielo.php?pid=S1808-24322016000100155&script=sci_abstract&tlng=pt

e, com Alexandre Costa e Felipe Farias, exploraram esses dados até que localizassem um padrão que merecia ser investigado: em certo momento, houve um aumento acelerado no número de processos decididos, sendo que não houve um incremento proporcional no número de acórdãos publicados.

Quantitativamente, tínhamos identificado uma correlação: em certo momento, o número de processos julgados era equivalente ao de acórdãos publicados, e aos poucos essa situação foi mudando até que o número de processos julgados era praticamente 10x maior do que o número de acórdãos. Como interpretar essa combinação de fatos? A hipótese a que chegamos é que houve um incremento no número de decisões monocráticas, cujo alto índice de incidência é identificado em várias pesquisas contemporâneas.

Não tentamos demonstrar quantitativamente essa correlação (pois não levantamos os processos para avaliar o que ocorreu neles), mas estudamos o contexto político e normativo e identificamos que houve mudanças legislativas que abriram espaço para uma atuação monocrática mais intensa e consideramos que essa explicação era compatível com as observações posteriores. Assim, consideramos que os resultados dessa análise qualitativa nos ofereceram uma narrativa que tornava compreensível as transformações evidenciadas nos dados que obtivemos em uma tabela já organizada pelo STF.

Esse é um artigo que foi escrito de forma relativamente rápida e combinou análise documental com pesquisa de dados. Trata-se de um trabalho de pesquisa que não partiu de uma intuição para a busca de dados, mas que partiu da análise dos dados para gerar uma hipótese explicativa, que foi explorada por meio de uma pesquisa qualitativa (basicamente documental). Ele é mencionado aqui como exemplo de que essa exploração dos dados pode gerar pesquisas interessantes e que podem ser publicadas em revistas jurídicas de ponta.

5. Obtendo dados dispersos em múltiplas páginas

Vários dos dados relevantes para os pesquisadores em direito estão dispersos em páginas com informações ligadas a cada processo particular ou a cada decisão. Vários tribunais organizam seus dados como uma combinação de um banco de processos (com informações variadas acerca de cada um dos processos individualmente considerados) e um banco de decisões (com uma série de informações ligadas a cada decisão individualmente considerada).

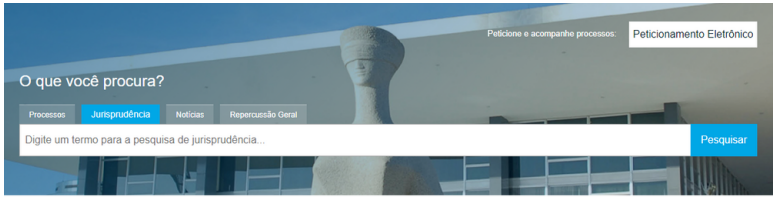
Com isso, o conteúdo da decisão costuma "habitar" um banco ligado ao serviço de jurisprudência, com informações e metadados (classificações) inseridos pelas seções de documentação e jurisprudência.

Já as informações processuais costumam "habitar" um outro local: um banco ligado à área que lida com a tramitação dos processos, sendo alimentado por informações das Secretarias e dos Gabinetes.

Essa estrutura nem sempre tem esse formato. No caso do STJ, por exemplo, os metadados dos processos e das decisões estão no mesmo banco de dados relacional (DB2), sendo possível fazer a pesquisa, extração e cruzamento dos dados alimentados com relativa facilidade, já que são tabelados. Porém, apenas técnicos da informática do Tribunal conseguem fazer essas pesquisas, sob demanda, pois o acesso direto ao banco de dados não é disponível nos sistemas da rede interna do STJ e da internet. Assim, tanto o público interno como o público externo precisa utilizar em suas pesquisas as ferramentas disponibilizadas na página, que também tem uma ferramenta de consulta processual diversa da pesquisa de jurisprudência.

No caso do STF, as buscas no banco de jurisprudência podem ser feitas na página Pesquisa de Jurisprudência[9].

[9] http://portal.stf.jus.br/jurisprudencia/



Essa pesquisa normalmente é feita por "termos", que são buscados dentro dos campos dos bancos de decisão, pois o objetivo é buscar decisões que tratem das questões de interesse do pesquisador. Isso faz com que o pesquisador possa pesquisar tanto por processos (como ADI 2994) ou por assuntos (como "criação de municípios"), e o resultado dessa pesquisa é uma lista de decisões.

Na nova versão da página de pesquisa, uma funcionalidade interessante para os pesquisadores é que é possível exportar os dados na forma de tabela .csv, com alguns dos atributos da decisão: Processo em que ocorre, Relator, Data de publicação, Data de julgamento, Órgão julgador e Ementa.

A consulta processual é feita em uma aba diferente da mesma estrutura que comporta a pesquisa de jurisprudência, sob a pergunta geral "O que você procura?".

Essa pesquisa feita por processo aponta para uma lista de processos (quando se indica somente o número) ou para a página de um processo específico, quando os termos da pesquisa conseguem individualizar um processo.

Nessa página estão presentes as informações acerca do processo e, inclusive, há um botão de "Jurisprudência", que realiza uma busca na pesquisa de jurisprudência com dados referentes ao processo contido na página. É nessa página que encontramos informações sobre as partes e os andamentos, inclusive informações específicas sobre os andamentos que noticiam a tomada de decisões.

A pesquisa no banco de processos e no banco de decisões costuma ser independente, como no caso do STF, e não há uma garantia total de compatibilidade entre as informações contidas em ambos. Por um lado, são lançados "andamentos sobre decisões" que podem ter categorias diferentes das usadas no banco de jurisprudência. Por outro, os metadados inseridos pela jurisprudência nem sempre seguem classificações idênticas aos dados dos sistemas de acompanhamento processual, tipicamente mais literais com relação às palavras usadas pelos ministros.

A união dos dados desses dois sistemas nem sempre é simples. Um mesmo processo pode ter várias decisões e é desafiador criar um modelo de dados que permita uma análise adequada do modo como as decisões se acoplam. Uma mesma decisão pode resolver vários processos e servir como precedentes a centenas de outras decisões de vários tipos.

Essa multiplicação de complexidades dificulta a construção de um

modelo de dados unificado e toda tentativa de unificação aponta para bancos de dados que são mais complexos do que as tabelas "bidimensionais" com as quais os pesquisadores de ciências sociais costumam trabalhar.

Por tudo isso, um dos desafios das pesquisas de dados é construir um modelo de dados adequado aos objetivos da pesquisa. Quais serão as unidades de análise? Como os processos serão ligados com as decisões? Essas são questões intimamente ligadas à metodologia e ao referencial teórico da pesquisa.

No que toca à coleta desses dados dispersos, é possível buscar cada página de forma individual, copiar os dados e lançar em uma tabela. Porém, esse tipo de atividade repetitiva pode ser substituída de forma eficaz por programas de computador que realizem essas operações de formas reiteradas, buscando as informações de cada página e agrupando-as em um banco de dados consolidados.