

# Preservação do Conteúdo Web Brasileiro

Mesa-redonda apresentada no 9º Fórum da Internet do Brasil. Manaus, AM, 2/10/2019  
Salão D 16h30-18h

## Proponente e moderador

Antonio Paulo Carretta (terceiro setor)

Especialista em organização e recuperação de informação em ambientes digitais. Graduado em Biblioteconomia com mestrado em Ciência da Informação (ECA/USP) e extensão em Gestão do Conhecimento (FGV-SP). Atualmente colaborador na Fundação Bial, também atuou em centros de pesquisa e documentação na área de Comunicação.

## Palestrantes

Moisés Rockemback (comunidade científica e tecnológica)

Líder do Núcleo de Pesquisa em Arquivamento da Web e Preservação Digital (NUAWEB) na Universidade Federal do Rio Grande do Sul (UFRGS), onde é professor do Programa de Pós-Graduação em Comunicação e Informação. Doutor em Informação e Comunicação em Plataformas Digitais pela Universidade do Porto / Universidade de Aveiro e mestre em Comunicação e Informação pela UFRGS.

Daniel Gomes (governamental)

Líder do Arquivo.pt, serviço de arquivamento e preservação da Web portuguesa. Gestor da equipe de desenvolvimento web da Fundação para a Computação Científica Nacional (FCCN), em Portugal, e colabora com a Comissão Europeia como perito independente para avaliar atividades de pesquisa em TICs. Possui Doutorado em Engenharia Informática pela Universidade de Lisboa.

Aquiles Alencar Brayner (empresarial)

Pesquisador autônomo. Atuou como curador do acervo digital da British Library (Reino Unido), possui doutorado pelo King's College de Londres e mestrado em História da Arte e Estudos Latino-americanos pela Rijksuniversiteit, Groningen, e em Ciência da Informação pela City University de Londres.

Diogo Cortiz (comunidade científica e tecnológica)

Pesquisador no Centro de Estudos sobre Tecnologias Web do Núcleo de Informação e Coordenação do Ponto BR (Nic.br) e do W3C Escritório Brasil. Doutor e mestre em Tecnologias da Inteligência e Design Digital pela PUC-SP, com PhD Fellowship pela Université Paris 1 - Sorbonne. Realizou estágio pós-doutoral no laboratório de inteligência artificial e tecnologias criativas da Universidade de Salamanca. Atualmente é coordenador do Bacharelado em Design da PUC-SP.

Mariana G. Valente (terceiro setor)

Diretora do InternetLab e coordenadora do capítulo brasileiro do Creative Commons. Doutora em sociologia jurídica pela Faculdade de Direito da USP. Pesquisa direitos humanos e políticas de internet, direito autoral e acesso à cultura, ao conhecimento e à educação, gênero, direitos das mulheres e tecnologia.

### **Objetivos e resultados**

Considerando que a Web se tornou, nos anos 2000, o meio majoritário de expressão da humanidade e que o conteúdo nela publicado se renova rapidamente, esta mesa-redonda teve como objetivo defender a criação de políticas públicas para preservação do conteúdo web brasileiro. O resultado foi consensual quanto à urgência de se estabelecer modelos de arquivamento do conteúdo online.

### **Justificativa em relação à governança da Internet**

Pode-se dizer que a principal questão sociocultural da governança da Internet são justamente as políticas de conteúdo. Assim sendo, uma delas poderia vir a ser a preservação do patrimônio histórico-cultural na Web para que continue acessível em algum lugar, mesmo que offline.

### **Metodologia e formas de participação desenvolvidas durante o workshop**

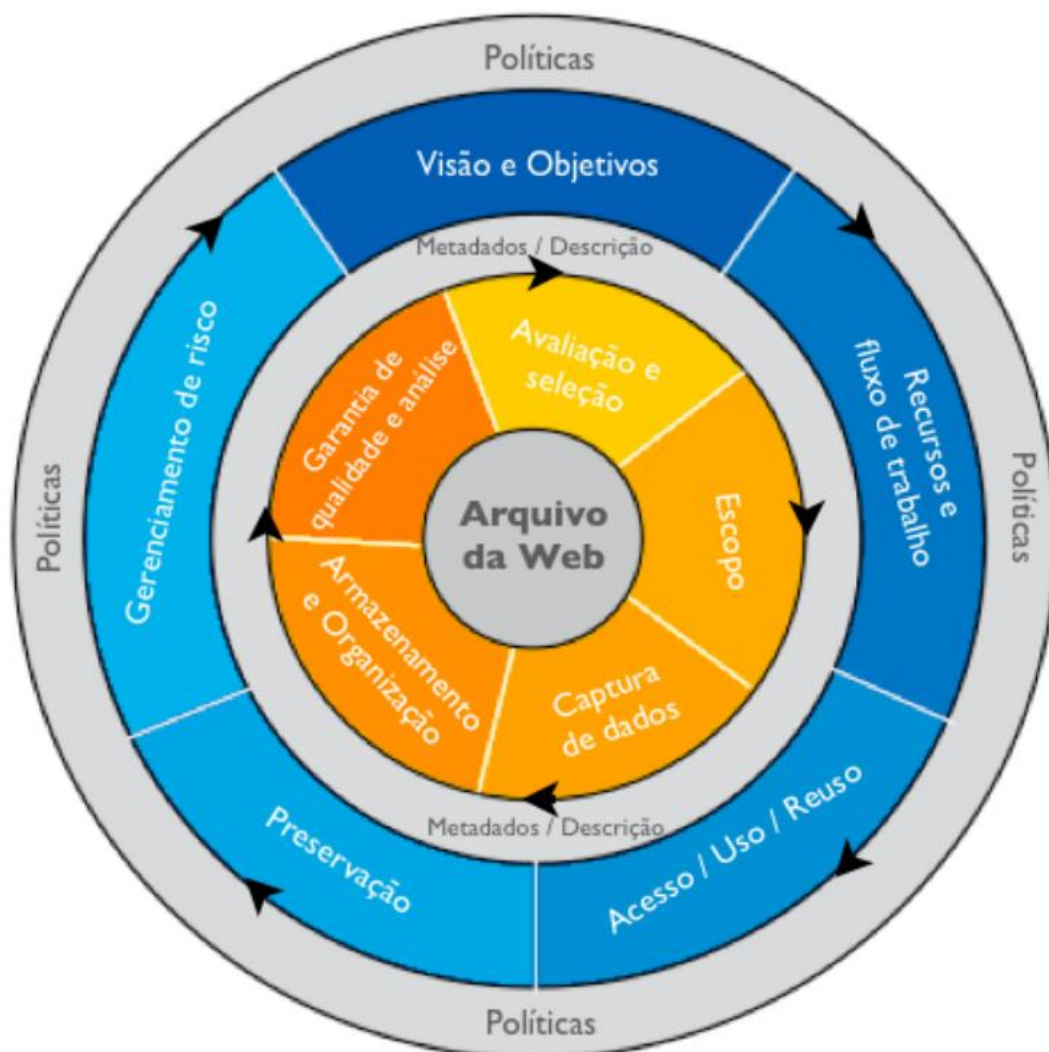
Mesa-redonda constituída por cinco palestrantes, sendo que dois deles participaram por videoconferência e vídeo pré-gravado. Os participantes da mesa receberam previamente perguntas formuladas pelo mediador, na tentativa de que suas falas fossem complementares e incrementais entre si. Por fim, a audiência foi convidada a formular perguntas à mesa.

### **Síntese dos debates**

**Moisés Rockemback** participou remotamente de Portugal, onde realizou pós-doutoramento em preservação digital e da Web. Com a intenção de introduzir o tema, sua abordagem partiu dos motivos para preservar a Web, expôs os problemas a serem enfrentados e traçou um panorama das iniciativas de arquivamento da Web, internacionais e no Brasil.

Por que preservar a Web? Ao responder esta pergunta, Moisés destaca o aspecto de que a Web se tornou, na atualidade, o meio informacional e comunicacional hegemônico, por onde trafegam cada vez mais informações. Isso significa que a Web de nossos dias será memória sociocultural dos cidadãos do futuro, quando se tornará fonte de informações para pesquisas acadêmicas e para o público em geral.

Porém, as publicações que nela circulam são dinâmicas e efêmeras, particularmente as nato-digitais. A natureza descentralizada, permeada de hiperlinks, e a rápida obsolescência tecnológica impõem desafios para a preservação da Web. O ponto de partida para enfrentá-los é o planejamento, no qual são definidos o escopo do arquivamento (se centrado em um tema, site ou domínio, p.ex.), a metodologia e as plataformas para coletar, armazenar e disponibilizar o conteúdo web. Estes três verbos correspondem às etapas do processo de arquivamento web, ou de seu ciclo de vida, como se vê neste gráfico.



Fonte: Bragg, Hanna, 2013 apud Rockemback, 2018.

Os métodos empregados nos arquivos da web vêm sendo estabelecidos pelo International Internet Preservation Consortium (IIPC) por meio da troca de experiências entre as instituições membros, que, em 2019, somam 45. Destas, a maioria situa-se em países europeus, sendo que a única sul-americana, a Biblioteca Nacional do Chile, se filiou em 2013.

No Brasil, Moisés aponta a iniciativa do Núcleo de Pesquisa em Arquivamento da Web e Preservação Digital (NUAWEB) da Universidade Federal do Rio Grande do Sul (UFRGS), criado em 2017. Nele, são realizadas pesquisas interdisciplinares sobre arquivamento de conteúdo web brasileiro, tais como websites universitários e institucionais, com um estudo de caso do governo federal, e sobre as eleições presidenciais de 2018.

Em sua participação por meio de um vídeo, **Daniel Gomes** mostra o primeiro jornal impresso português, a *Gazeta de Lisboa*, criada em 1715 e extinta em 1762. Quase 300 anos depois, suas edições continuam disponíveis ao público. Já o *Diário Digital*, jornal publicado exclusivamente online entre 1999 e 2017, onde encontrá-lo?

Essa fugacidade das publicações na Web evidencia a contradição da Era Digital: abundância de informações vs. escassez de memória. É por isso que os arquivos web são tão necessários quanto às demais instituições que se dedicam à preservação da documentação analógica. Assim, Daniel justifica a criação do [Arquivo.pt](http://Arquivo.pt).

As pesquisas no Arquivo.pt podem ser feitas por palavras-chaves ou por URL. Neste último caso, o resultado da pesquisa é um calendário de arquivamento da página pesquisada. A pesquisa pela URL <estadao.com.br>, por exemplo, resultou em uma lista de mais de 2.200 versões da página que se estendem de 2008 a 2018.

Após apresentar outros exemplos de conteúdo web preservado pelo Arquivo.pt, Daniel cita exemplos do quão significativa pode ser a recuperação de informações para os mais diversos usuários. Quanto ao valor do conteúdo preservado, que é inestimável, Daniel fez uma estimativa de quanto poderia valer, em dinheiro, o patrimônio do Arquivo.pt. O resultado, a partir de dados de 2017, foi superior do PIB de Portugal no mesmo ano.

Além da preservação extensiva do domínio português realizado pelo Arquivo.pt, qualquer pessoa pode indicar um site a ser preservado por meio do formulário disponível em [arquivo.pt/sugerir](http://arquivo.pt/sugerir). Outro serviço prestado são recomendações para criação de sites preserváveis.

Sobre a colaboração com o Brasil na preservação do conteúdo web, Daniel lembrou-nos de um protocolo assinado em 2009 pelos ministros da Ciência de Portugal e do Brasil. Um dos termos de colaboração seria justamente o arquivo da Web, mas o acordo não foi adiante.

O Brasil é o terceiro país que mais utiliza o Arquivo.pt, depois de Portugal e dos Estados Unidos, e Daniel aproveita para solicitar que o arquivo seja melhor divulgado no Brasil, de forma que os pesquisadores daqui possam explorá-lo ainda mais, uma vez que compartilhamos a mesma língua.

Na sequência, Daniel responde às perguntas que lhe foram feitas pelo mediador em virtude de sua participação ter se dado por vídeo pré-gravado.

*1. Qual a relação entre a governança da Internet e a preservação do conteúdo web?*

Sendo a Web a principal aplicação da Internet, a governança da Internet tem de considerar a preservação do conteúdo web.

*2. Como convencer órgãos governamentais sobre a importância de preservar conteúdo web?*

Informação é poder. A informação no século XXI é conteúdo web. Preservar conteúdo web é poder.

*3. Como foi o exemplo de Portugal? Quais os principais obstáculos e avanços no percurso do Arquivo.pt?*

Recrutar e treinar recursos humanos especializados. Imaturidade da tecnologia para o arquivamento web. A pesquisa de texto foi um avanço em escala mundial.

*4. O arquivamento da web é um investimento tecnológico caro?*

Perder a informação sai mais caro.

##### 5. *Requer uma infraestrutura complexa?*

Sim. Requer especialização de pessoas e recursos materiais. [Archive-it.org](http://archive-it.org) é uma alternativa para começar um projeto de arquivamento web.

##### 6. *Já existem parcerias de troca de experiências entre Portugal e o Brasil?*

Houve tentativas, sem prosseguimento.

Quem quiser se manter atualizado sobre o Arquivo.pt, basta inscrever-se em [arquivo.pt/inscrever](http://arquivo.pt/inscrever).

**Aquiles Brayner** dedica sua participação às políticas de arquivamento. A coleta do domínio nacional é uma delas, feita de maneira automática por um crawler. Esse tipo de arquivamento é problemático em virtude do grande volume de dados. O problema não está no armazenamento, mas no tratamento da quantidade de informação coletada.

Outra política adotada por muitas instituições é a coleta seletiva. Estipulam-se quais os sites a serem arquivados e a periodicidade das coletas. Sendo a captura das páginas realizada com interação humana e a quantidade capturada substancialmente menor, a qualidade do arquivamento seletivo é superior ao automático.

Tais políticas de arquivamento não são excludentes, e a necessidade de seleção e arquivamento de páginas web é imperativa. Para exemplificar isso, Aquiles mostrou páginas da campanha presencial de Donald Trump que, enquanto candidato, fez uma série de promessas. Depois de eleito, aquelas promessas foram retiradas do ar. Mas ainda podem ser encontradas no [Internet Archive](http://Internet Archive).

Quanto aos aspectos positivos e sensíveis das políticas de arquivamento, o primeiro a ser elencado é o controle de acesso. Aquiles cita como exemplo a Biblioteca Nacional do Chile, que faz o arquivamento do domínio .cl, mas não o disponibiliza online devido a diversas restrições legais.

O nível de arquivamento se refere à coleta do site em sua integridade ou não, à quantas páginas serão coletadas, se não todas, e quais conteúdos nelas contidos serão priorizados, sejam textos, imagens e/ou vídeos. Já a periodicidade de coleta precisa ser estabelecida de acordo com a dinâmica de alteração/atualização dos sites selecionados.

Um dos aspectos mais sensíveis diz respeito à política de reuso dos conteúdos arquivados, determinada por uma série de legislações que protegem a privacidade, os dados pessoais e os direitos autorais. Num ambiente supranacional como a Web, como estabelecer políticas de acesso e de reuso que não estejam circunscritas às fronteiras nacionais?

No que concerne à Web brasileira, Aquiles informa que há mais de 4 milhões de endereços registrados. Se considerarmos a quantidade de páginas, esse número aumenta exponencialmente. Como já foi dito antes, as páginas web são bastante efêmeras: 80% delas desaparecem entre 12 meses e dois anos. Ou seja, perdemos muita informação, apesar de termos, no Brasil, a [Lei de Depósito Legal](http://Lei de Depósito Legal) (Lei 10.994/2004), cujo art. 2 nunca se cumpriu no sentido de promover o arquivamento da Web brasileira.

Ao mesmo tempo em que nos ressentimos de informações do passado que não chegaram até nós, somos negligentes em relação ao arquivamento de páginas web. Assim, no futuro seremos conhecidos como a Era do Erro 404, e a sociedade que está por vir sofrerá o impacto desse vácuo informacional. Na opinião de Aquiles, esse quadro não se alterará se deixarmos a guarda do conteúdo web apenas sob a responsabilidade dos governos. Todos nós somos responsáveis pelo legado histórico-cultural a ser deixado para as próximas gerações.

**Diogo Cortiz** parte da página de Erro 404 para tratar dos trabalhos desenvolvidos pelo W3C com o intuito de evitá-la. Um dos problemas para o arquivamento de páginas web é a mudança de comportamento no uso da tecnologia. Quando se fala em arquivar páginas, trata-se de arquivar URIs (identificadores de recursos), os quais não são necessariamente adotados pelas redes sociais para onde migraram os usuários da Web 2.0.

Diante do desafio de preservar os URIs, o W3C aprovou, em 2017, a recomendação [\*Data on the Web Best Practices\*](#), cujo item 8.11 trata da preservação de dados. Entre as boas práticas relacionadas, a de nº 27 se refere aos identificadores de recursos, os quais, quando não encontrados, retornam a página 404. A ideia é passar a usá-la apenas quando ocorre erro.

Suponhamos que um blog tenha sido retirado do ar. Se apenas o conteúdo for removido, a tentativa de acessá-lo resultará na página 404. O W3C recomenda que, neste caso, o servidor seja configurado para retornar um outro código de resposta: o 410, que corresponde a 'gone'(retirado). Essa medida será importante para os arquivadores da Web, pois, quando os crawlers se depararem com o código 410, saberão que aquela página não deve ser arquivada.

Outro código a ser usado em substituição ao 404 é o 303. Este avisará o usuário que a página solicitada não está disponível, mas se encontra arquivada no Internet Archive ou no Arquivo.pt, por exemplo. Ou seja, se a página desaparecida estiver arquivada em algum lugar, o código 303 avisará onde ela está.

O URI corresponde à representação mais recente do recurso digital por ele identificado. Para superar esse estado de "eterno presente" dos URIs, surgiu, em 2009, o [Projeto Memento](#), liderado pelo Laboratório Nacional Los Alamos, nos Estados Unidos. Seu objetivo é prover a arquitetura Web de uma dimensão temporal, de forma que o usuário poderá selecionar entre diferentes versões de uma página ao longo do tempo. Mas isso só será possível se as versões da página estiverem arquivadas em algum lugar e se os arquivos web tiverem o protocolo Memento.

**Mariana Valente** ressalta que pensar em direitos não significa pensar, logo de saída, nas restrições impostas pela Lei de Direitos Autorais. Antes de tudo, precisamos pensar no direito de acesso à cultura. Pensar em memória, em patrimônio é pensar em direitos previstos na Constituição Federal, os quais são tão caros à cidadania quanto o direito à propriedade intelectual.

Portanto, quando se pensa em preservação de conteúdo web, os direitos autorais devem ser balanceados em relação ao direito conferido ao patrimônio histórico-cultural. O arquivo Arquivo.pt, por exemplo, dá acesso ao conteúdo preservado, mas alerta seus usuários sobre a necessidade de respeitar a legislação de direitos autorais vigente.

Os direitos autorais têm limitações e exceções que favorecem outros direitos, como o direito à informação. Permitem usos de materiais que não são ou deixaram de ser protegidos, além de possibilitar o uso de obras protegidas em situações específicas. A citação é uma delas, propiciando o exercício da crítica, estudo e pesquisa.

Uma das limitações aos direitos autorais que é essencial à proteção do patrimônio artístico, histórico e cultural é a assim chamada cópia para preservação. Está presente na legislação de muitos países, mas não na brasileira, de maneira explícita. Nossa [Lei de Direitos Autorais](#) (Lei 9610/98) é a que tem menos previsões de limitações e exceções se comparada a de outros países. É evidente que uma análise sistemática do ordenamento jurídico, ponderando direitos tão valiosos quanto os autorais, autoriza as instituições brasileiras a fazerem cópias de preservação.

A defesa do domínio público deve ser constante no que concerne às bibliotecas, arquivos e museus. Não como um vazio de direitos, mas como possibilidade dos cidadãos fazerem usos livres dos bens culturais. Seja para o seu arquivamento, para torná-los acessíveis online, como faz a Biblioteca Nacional Digital, quanto para o seu reuso criativo por outros autores. O Internet Archive é uma instituição exemplar no que diz respeito ao acesso online de conteúdos digitais em domínio público, além do arquivamento da web em âmbito global.

Outro recurso criado para resguardar o direito ao acesso e de usufruto de bens culturais são as licenças [Creative Commons](#), por meio das quais os autores permitem que suas obras sejam usadas de acordo com o estabelecido em cada uma das licenças, sem necessidade de prévia autorização e sem ônus.

O direito à privacidade há que ser respeitado no âmbito do arquivamento web, considerando que sites podem conter dados pessoais, ou mesmo informações criadas em uma situação em que a pessoa tinha uma expectativa de privacidade. Considerando que se as pessoas fornecerem informações ou se expressarem em ambientes cuja preservação de longo prazo não estava de antemão prevista, é lhes facultado o direito à remoção de seus dados e ao esquecimento, como assim prevê a [Lei Geral de Proteção de Dados](#) (Lei 13.709/2018).

A Lei de Depósito Legal, por sua vez, não trata do arquivamento e disponibilização de conteúdo web especificamente, mas já existem legislações em outros países que autorizam o arquivamento em associação com o direito autoral, cópias para preservação e com o direito à privacidade. Essa é uma questão a ser discutida com urgência no Brasil.

### Consensos

O conteúdo web brasileiro é bastante fluído e há que ser preservado como memória histórico-cultural de nossa sociedade. Se nenhuma medida for tomada nesse sentido, nossa época sofrerá, no futuro, de escassez de informações, apesar de terem sido tão abundantes em nossos dias.

### Dissensos

Se a Lei de Depósito Legal confere ou não mandato à Biblioteca Nacional para coletar, armazenar e disponibilizar online o conteúdo da web brasileira, sem prévia autorização dos detentores dos direitos autorais.

### Pontos a aprofundar

A política pública de preservação da web brasileira deveria se pautar pelo arquivamento do domínio nacional, como no caso Português, ainda que coexistam projetos de arquivamento descentralizados? Quais os prós e contras do arquivamento extensivo (domínio nacional) e intensivo (seletivo)?

### Perguntas

[Luan], não-identificado

Eu tenho uma pergunta para o Diogo. Como você trabalha no Nic.br, eu gostaria de perguntar o seguinte. Neste ano, o governo federal fez uma mudança bem brusca nos sites governamentais, e sumiu muita coisa de vários deles, vários documentos importantes. Eu gostaria de saber se vocês estão em contato com o governo federal quanto a essa readequação do conteúdo do portal deles.

Só um adendo ao que a Mariana disse sobre a questão de privacidade: na semana passada, houve uma decisão do Tribunal de Justiça da União Europeia sobre desindexação e indexação de conteúdos sensíveis. O tribunal deixou em aberto para que as plataformas equilibrem os direitos de privacidade e acesso à informação pública.

Carlos Afonso, membro do MAG do IGF

Sou um dos facilitadores de um fórum de melhores práticas que existe já há algum tempo no IGF sobre preservação e criação de conteúdo local. A ênfase desse fórum de melhores práticas neste ano é justamente a preservação de conteúdo levando em conta alguns aspectos que não foram mencionados aqui, e incluindo os que foram mencionados, obviamente. Um deles é a questão dos patrimônios culturais que são representados ou reproduzidos digitalmente e que podem estar em risco de mutilação ou desaparecimento devido à situações políticas. Nós tivemos aqui no Brasil dois exemplos recentes, o caso da Fiocruz, em que a publicação de um relatório sobre uso de drogas no Brasil foi censurada pelo governo e desapareceu da internet. Há o risco do site da Comissão Nacional da Verdade ser mutilado ou ser apagado. O Instituto de Segurança Pública do governo Witzel, no estado do Rio de Janeiro, também está ameaçado, porque contém dados de violência da polícia militar.

Assim sendo, a grande preocupação desse fórum de melhores práticas neste ano, especialmente, diz respeito a essas situações de preservação de conteúdo e também à preservação de patrimônio histórico-cultural, a herança cultural dos países, que são, por exemplo, os idiomas e culturas originários que possam estar reproduzidas em forma digital e que podem desaparecer. Na minha opinião, é muito importante nós pensarmos em como podem ser preservados esses conteúdos em risco, devido a situações políticas ou conjunturas históricas.

Eu só queria mesmo ressaltar esse ponto: existe todo um trabalho dentro do IGF, já que este é o pré-IGF brasileiro, de desenvolver as melhores práticas de preservação do patrimônio cultural reproduzido ou representado em forma digital. Não é só uma questão de direito autoral, etc. Em alguns casos há risco de perda de informação fundamental para a sociedade.

Alberto César, jornalista e fotógrafo

Em relação ao que foi dito na primeira pergunta, suponhamos que, em um caso hipotético de sites oficiais de órgãos públicos – sejam federais, estaduais ou municipais –, em que a retirada de conteúdo, ou a sua modificação no sentido de dificultar o acesso pelo usuário, poderia se configurar improbidade administrativa? Ou se a forma como isso se deu, sem consulta, é que qualificaria esse crime?

## Respostas

**Diogo Cortiz** responde à primeira pergunta: o Nic.br não tem nenhum processo de arquivamento, então, não acompanhamos essas mudanças feitas nos sites governamentais, até porque o Nic.br não é um órgão do governo. Além disso, o Nic.br nunca foi consultado sobre esse assunto, seja por órgãos governamentais ou por instituições que se ocupam dessa temática. A mudança foi uma decisão do governo, e o Nic.br não tem, ainda, um mecanismo que pudesse preservar esse conteúdo.

**Mariana Valente** responde à última pergunta: se não estou enganada, seria difícil isso configurar um crime de improbidade pela ausência de dever de disponibilização. Existem deveres de disponibilização de conteúdos online, mas relativos a contextos específicos, como por exemplo, a Lei de Acesso a Informação, que obriga órgãos do Estado a divulgar informações específicas, relativas à transparência. E a lei prevê sanções caso não seja respeitada. Mas não me parece que a retirada de um conteúdo online configure uma violação ou um crime em particular.



Aproveito para comentar o ponto que o [Luan] trouxe para o debate, quando falou das decisões na Europa sobre indexação e desindexação. Essa questão também se coloca sobre o arquivamento. Quando em um arquivo, você pode procurar por palavra-chave, que pode ser o nome de uma pessoa, e o resultado da busca traz informações do passado dessa pessoa, que já não estão mais disponíveis, me parece que fica colocada a questão do direito ao esquecimento. Há que se pensar em que medida e com que meios os projetos de arquivamento web dariam conta disso, levando-se em consideração que são atividades sem fins lucrativos. Isso terá que ser analisado quando pensarmos em uma regulação para o futuro.

**Mesa redonda disponível no canal do Nic.br no YouTube:**

<[https://www.youtube.com/watch?v=ZanQR9-WR\\_E](https://www.youtube.com/watch?v=ZanQR9-WR_E)>

### **Relatora**

Etoile Shaw (empresarial)

Formada em jornalismo (Puccamp), com pós-graduação em história da arte (Faap). Atuou em jornais, revistas e editoras, com destaque para: Editora da Unicamp, Museu de Arte Contemporânea da Universidade de São Paulo (MAC/USP), Banco de Dados da Folha de S.Paulo e Editora Ática. Criou a empresa Quati Produções Culturais, por meio da qual tem atuado no mercado editorial.