RICI Revista Ibero-Americana de Ciência da Informação

ISSN 1983-5213

Iniciativas científicas de arquivamento e preservação de conteúdos em mídias sociais: panorama atual

Laura Vilela Rodrigues Rezende
Universidade Federal de Goiás, Faculdade de Informação e Comunicação, Goiânia, GO, Brasil
lauravil.rr@gmail.com

Dalton Lopes Martins
Universidade Federal de Goiás, Faculdade de Informação e Comunicação, Goiânia, GO, Brasil
dmartins@gmail.com

DOI: https://doi.org/10.26512/rici.v11.n1.2018.8538

Recebido/Recibido/Received: 2017-11-16
Aceitado/Aceptado/Accepted: 2017-12-12

Resumo: O estudo realiza um mapeamento de experiências científicas que realizam coleta e arquivamento de conteúdos de mídias sociais. Apresenta inicialmente um breve referencial teórico relacionado à preservação digital e *Web 2.0*. Pesquisa de abordagem qualitativa e de cunho exploratório uma vez que buscou identificar nas experiências inovadoras questões pouco investigadas, tais como as características: tecnológicas (plataforma de mídia social utilizada, solução tecnológica utilizada para a coleta dos dados e formato de armazenamento dos dados coletados); e contextuais verificando se os metadados garantem proveniência, autenticidade e como é a gestão de direitos e tipos de acesso aos dados. Os resultados apontam que, embora o foco principal destas iniciativas não seja diretamente a preservação, esta questão está intrinsecamente contemplada. O Twitter é a plataforma principal de coleta de dados utilizada por pesquisadores. Vários são os desafios elencados, é preciso especificar como e por quem os dados coletados serão utilizados e relevante deve ser a atuação curatorial nestes projetos, visando selecionar o que de fato deve ser guardado. Por fim, as restrições das plataformas de mídias sociais e a falta de transparência e compreensão sobre suas práticas de gestão de dados atualmente dificultam a preservação deste tipo de conteúdo.

Palavras-chave: Mídia social; Preservação digital; Twitter.

ARTIGOS

Scientific initiatives of archiving and preservation of contents in social media: current overview

Abstract: The study carries out a mapping of scientific experiments that collect and archive social media content. It presents initially a brief theoretical reference related to digital preservation and Web 2.0. Qualitative research with an exploratory approach since it sought to identify in the innovative experiences questions little investigated, such as the characteristics: technological (social media platform used, technological solution used to collect the data and format of storage of the collected data); and contextual checks to see if the metadata guarantees provenance, authenticity, and how rights management and types of data access are. The results indicate that, although the focus of these initiatives is not directly on preservation, this issue is intrinsically considered. Twitter is the primary platform for data collection used by researchers. Several are the challenges listed, it is necessary to specify how and by whom the data collected will be used and relevant should be the curatorial action in these projects, to select what should in fact be saved. Finally, the restrictions of social media platforms and the lack of transparency and understanding about their data management practices currently make it difficult to preserve this type of content.

Keywords: Digital preservation; Social media; Twitter.

Iniciativas científicas de archivamiento y preservación de contenidos en medios sociales: panorama actual

Resumen: El estudio realiza un mapeo de experiencias científicas que realizan recolección y archivo de contenidos de medios sociales. Presenta inicialmente un breve referencial teórico relacionado a la preservación digital y *Web 2.0*. Investigación de abordaje cualitativo y de cuño exploratorio ya que buscó identificar en las experiencias innovadoras cuestiones poco investigadas, tales como las características: tecnológicas (plataforma de mídia social utilizada, solución tecnológica utilizada para la recolección de los datos y formato de almacenamiento de los datos recolectados); y contextuales verificando si los metadatos garantizan procedencia, autenticidad y cómo es la gestión de derechos y tipos de acceso a los datos. Los resultados apuntan que, aunque el foco principal de estas iniciativas no es directamente la preservación, esta cuestión está intrínsecamente contemplada. *Twitter* es la plataforma principal de recolección de datos utilizada por los investigadores. Varios son los desafíos enumerados, es preciso especificar cómo y por quién los datos recolectados serán utilizados y relevante debe ser la actuación curatorial en estos proyectos, con el objetivo de seleccionar lo que de hecho debe ser guardado. Por último, las restricciones de las plataformas de medios sociales y la falta de transparencia y comprensión sobre sus prácticas de gestión de datos actualmente dificultan la preservación de este tipo de contenido.

Palabras clave: Medios sociales; Preservación digital; Twitter.

1 Introdução

O cenário tecnológico mundial se apresenta repleto de recursos que tem proporcionado ao cidadão experiências de uso cada vez mais enriquecedoras, especialmente quanto à utilidade e capacidade de participação social no que é possível se fazer. Observa-se um crescente incremento na interação entre pessoas, especialmente pela criação e compartilhamento de conteúdos digitais, tendo as ferramentas colaborativas como pano de fundo, trazendo inovações, tanto no modo de produção quanto naquilo que pode ser produzido, em todas as áreas do conhecimento.

Constata-se uma proliferação de novos serviços que envolvem diversas possibilidades de interação social em torno de objetos digitais, agregando novas camadas de informações registrando essa interação que hoje se tornam fundamentais ao serem utilizadas por algoritmos que calculam a relevância social desses objetos e os organizam de forma sistemática para apresentação nas interfaces de busca. Do ponto de vista informacional, novas questões e desafios surgem relacionados com a preservação digital, considerando o aumento da complexidade das informações de relevância que hoje devem ser consideradas para que se possa caracterizar de forma completa os objetos digitais e as relações sociais ao seu entorno.

Este contexto anteriormente descrito se refere à segunda geração da Internet conhecida como *Web 2.0,* que pode ser entendida como uma nova geração de serviços e produtos, e não uma revolução tecnológica. Trata-se de uma revolução comportamental em que prevalece a participação e colaboração dos usuários no tocante ao uso das informações.

O uso maciço da *Web 2.0* apresenta grandes inovações nas novas maneiras de relacionamento, trabalho e produção da informação que, por conta do aumento no volume e escala também tem registrado complexidades jamais vistas anteriormente. O surgimento de novos atores que trabalham de forma colaborativa na produção de informação de relevância, somadas às inovações produzidas por novos algoritmos sociais que recombinam a participação dos usuários e a informação de novas maneiras, como é possível se observar em experiências como a *Wikipedia*[1], *Uber*[2], *Waze*[3], *Netflix*[4], *Amazon*[5], *Google*[6], entre outros, trazem à tona questões relacionadas à necessidade de rever e repensar muitos dos modos de gestão contemporânea da informação. Entende-se que é exatamente nessas novas aberturas de possibilidades e no espaço produzido pelo surgimento do novo que é possível propor novos olhares para os múltiplos fenômenos informacionais do século XXI.

Neste contexto, também conhecido como *web* social, surge um questionamento em relação à memória destes materiais que possuem particularidades que precisam ser consideradas quando se trata de preservação. Arquivar o material gerado no contexto da *Web 2.0* traz para a humanidade registros relevantes do momento atual e abre questões que se tornam de extrema relevância na contemporaneidade relacionadas à preservação de novos tipos de objetos digitais.

O desaparecimento da rede social Orkut, criada pelo *Google* em 2002 e com duração de 12 anos, sendo encerrada em 2014 com suas funcionalidades dinâmicas e ficando seu conteúdo disponível para *download* até meados de 2016, é considerado um caso clássico de perda de conteúdo *web* criado de maneira colaborativa. Durante os dois anos que a rede esteve disponível para consulta, os processos de extração de conteúdo não eram considerados fáceis, especialmente por conta do grande volume de dados além da problemática do que se pode considerar um dos elementos mais importantes em redes sociais: as conexões, que ficavam fortemente comprometidas ao serem reconstituídas após as tentativas de extrações, apesar de se considerar que eram ambientes preparados e que contavam com recursos técnicos profissionais para a pesquisa.

Sem dúvida, o Orkut teve um papel extremamente importante nos processos de socialização em redes digitais no Brasil, sendo um dos países mais atuantes e participativos em termos mundiais. Cabe se perguntar: o patrimônio cultural ali produzido e armazenado em

[1] https://pt.wikipedia.org/wiki/Wikip%C3%A9dia:P%C3%A1gina_principal
[2] https://www.uber.com/pt-BR/
[3] https://www.waze.com/pt-BR
[4] https://www.netflix.com/br/
[5] https://www.amazon.com/
[6] https://www.google.com.br/webhp?hl=pt-BR&tab=ww

suas bases de dados não faz parte também do patrimônio cultural brasileiro, servindo de referência como fonte de pesquisa e, sobretudo, como dinâmica social de articulação em rede e como campo de produção de capital social e cultural de uma nação? O que perde o país com seu desaparecimento? Que impactos essas perdas produzem em nossa memória? Seriam essas questões de interesse de uma política pública para a memória digital em pleno século XXI e mesmo para a pesquisa científica na área da Ciência da Informação? Estariam nossas instituições custodiais preparadas para lidar com a arquitetura de socialização de produção informacional em rede? Tais perguntas demandariam processos de pesquisa amplos e complexos, indo muito além da proposta deste artigo, mas que cabe aqui serem pontuadas por terem sido inspiradoras desta pesquisa.

O presente estudo realizou um mapeamento e análise de experiências científicas internacionais que estão hoje realizando ações de coleta e arquivamento de conteúdos gerados em mídias sociais no intuito de atender objetivos de investigações científicas. Entende-se aqui que conhecer essas experiências, avaliar suas potencialidades e pontos críticos é um passo fundamental como ponto de partida para delinear algumas respostas possíveis para os questionamentos inspiradores supracitados. Além de trazer um panorama de iniciativas científicas ao redor do mundo que desenvolvem ações de coleta e arquivamento de conteúdos de mídias sociais, apresenta-se um breve referencial teórico com conceitos relacionados à preservação digital e *Web 2.0*.

2 Preservação digital e mídias sociais: mobilização emergente

É possível afirmar, conforme Hockx-Yu (2017), que mídias sociais são aplicações baseadas na Internet que permitem aos usuários formarem redes sociais ou comunidades baseadas em algum interesse comum — com orientações ideológicas ou sociais. Tais aplicações existem de várias formas, porém seu propósito fundamental é dar suporte à interação e comunicação entre os membros da comunidade, incluindo criação e troca de conteúdos gerados pelos usuários. Estes conteúdos podem ser criados em formato texto, áudio ou vídeo. Exemplos incluem *upload* de arquivos de vídeo para o *Youtube*[7], uso de formulários web para comentar ou classificar um item em um site de compras ou editar conteúdos de artigos em um site de enciclopédia on-line como a *Wikipedia*.

O manual *The preservation of web resources* traz questões fundamentais sobre a "fluidez" que muitas vezes caracteriza o conteúdo da *Web 2.0*, afirmando que tal característica pode dificultar a demarcação do ponto em que o conteúdo postado foi concluído estando,

[7] https://www.youtube.com/

portanto, pronto para ser arquivado (ULCC; UKOLN, 2008). Em relação às permissões de uso dos conteúdos presentes na *Web 2.0*, geralmente, quantidades significativas destes recursos informacionais são geradas por vários usuários. Conteúdos gerados por usuários distintos apresentam o desafio de solicitar aos seus detentores - podendo ser o proprietário da mídia social ou do conteúdo específico – a autorização para arquivamento. Pode-se afirmar que se trata de uma tarefa árdua e que às vezes se torna praticamente impossível de ser realizada (PENNOCK, 2013, p. 13).

Visando enfatizar a importância de se preservar conteúdos das mídias sociais, a pesquisa realizada por Salaheldeen e Nelson (2012), intitulada "*Losing my revolution: How many resources shared on social media have been lost?",* buscou estimar o quanto deste conteúdo foi arquivado em arquivos públicos da *web*, estando disponível e o quanto foi perdido sem possibilidade de recuperação. A pesquisa buscou coletar dados de seis eventos ocorridos no período de junho de 2009 a março de 2012, a saber: o surto do vírus H1N1, a morte de Michael Jackson, as eleições iranianas e seus protestos, o Prêmio Nobel da Paz de Barack Obama, a revolução egípcia e a revolta na Síria. Com as análises dos dados coletados, foi possível concluir que após o primeiro ano desde a publicação, aproximadamente 11% dos recursos informacionais compartilhados serão perdidos e depois disto a perda estimada por dia será de 0,02%.

Também em 2012, a União das Nações Unidas para a Educação, Ciência e a Cultura (UNESCO) publicou a Declaração de Vancouver intitulada "A memória do mundo na era digital: digitalização e preservação" (UNESCO, 2012). Este relevante documento sinalizou para a emergência do tema de preservação digital trazendo questões ligadas aos direitos humanos dos cidadãos. Ao abordar o tema da preservação digital, a UNESCO considera para a elaboração deste documento alguns fatores impactantes nesta temática, dentre os quais figuram: o risco de desaparecimento da herança cultural das nações; a necessidade de conservação do patrimônio universal representados pelos produtos do conhecimento humano, sobretudo o patrimônio documental; a crescente produção, distribuição e uso dos recursos informacionais em formato eletrônico constituindo novo tipo de patrimônio, o digital; o reconhecimento de que o acesso a este vasto contingente de informação proporcionará maiores oportunidades a toda humanidade, possibilitando a criação, comunicação, assim como a troca de experiências; a urgente necessidade de se promover a preservação do patrimônio digital ao redor do mundo em benefício das gerações atuais e futuras; os documentos nativos digitais e digitalizados devem ter a mesma importância na definição de padrões e modelos que garantam acesso contínuo.

No Brasil, o Conselho Nacional de Arquivos (CONARQ) baseado nas recomendações da UNESCO, veio através de carta aberta convocar “os setores públicos e privados, envolvidos com a produção e proteção especial dos documentos em formato digital, a envidarem esforços para garantir sua preservação e acesso contínuo [...]” especialmente por considerar que esta produção digital representa um patrimônio fortemente ameaçado de extinção, assim como “de falta de confiabilidade, e que sua preservação em benefício das gerações atuais e futuras é uma preocupação urgente no mundo inteiro” (CONARQ, 2016, p. 1). Devido à grande relevância e em função do alcance que se obteve com as propostas sobre o tema da preservação digital iniciadas pela UNESCO, inúmeras instituições e pesquisadores se posicionaram a favor da necessidade de proporcionar as condições necessárias e seguir as recomendações oferecidas pela literatura para de fato estabelecer ações e práticas voltadas para efetivar a preservação digital.

Conforme observado em vários trabalhos sobre a guarda da informação em formato digital, nota-se a insuficiência em assegurar a longevidade de acesso e preservação destes materiais ao se promover somente sua guarda em soluções tecnológicas e aplicações de *software*. Tal fato se deve à constatação certa de que os novos recursos de *software e hardware* se tornam obsoletos em um curto intervalo de tempo estando ainda suscetíveis a danos e acidentes que podem encurtar sua perspectiva de funcionamento em relação ao tempo de uso estimado. Trazendo a problemática da preservação de objetos digitais para o contexto das mídias sociais, cabe retomar inicialmente a questões iniciais fundamentais. O modelo referencial *Open Archival Information System* (OAIS) – ISO 14721:2012, ou Sistema Aberto para Arquivamento de Informação (SAAI), que segundo Sayão (2010, p. 13), trata-se de "uma infraestrutura conceitual que descreve o ambiente, as interfaces externas, os componentes funcionais e os objetos de informação, associados com um sistema responsável pela preservação de longo prazo de materiais digitais".

Embora o modelo referencial já tenha sido criticado por não conseguir abarcar aspectos práticos do arquivamento de conteúdos *web* de maneira geral, um de seus destaques é a definição dos usuários primários da informação a ser preservada, ou a “comunidade designada”. Tal abordagem vem garantindo o planejamento e desenvolvimento de vários casos de sucesso e pode ser considerado o ponto central ao se considerar ações de preservação digital. Complementando, O *Digital Preservation Handbook* da Digital Preservation Coalition (DPC, 2017) argumenta que as melhores práticas em preservação encorajam as organizações a considerar o "início" da preservação digital como sendo o ponto de criação do objeto digital para que possa ocorrer o efetivo entendimento necessário sobre como o objeto será usado e por quem pode ser capturado.

Thomson (2016) afirma que, por causa do conhecimento especializado, as comunidades designadas ou usuários primários articulados com seus requisitos podem ser considerados algo essencial e decisivo para muitos outros componentes de uma estratégia de preservação digital. Assim, definir e monitorar a comunidade designada ajuda a definir os requisitos de metadados uma vez que são boas fontes de informações contextuais necessárias para tornar os dados inteligíveis. Da mesma forma, o desenvolvimento de uma abordagem para preservar as mídias sociais será fortalecido pela compreensão de quem e como os dados preservados serão utilizados. Novos formatos de dados, como os conteúdos das mídias sociais têm uma necessidade ainda maior de uma comunidade designada claramente definida se comparado com outros formatos de dados mais estabelecidos. No atual cenário marcado pela ausência de melhores práticas e padrões estabelecidos para se preservar este tipo de conteúdo, os requisitos dos usuários ajudarão a determinar os formatos, metadados e modos de acesso garantindo que o conteúdo preservado possa ser acessado na íntegra, no futuro.

Depois de percorridas as devidas ressalvas em relação aos requisitos de armazenamento de conteúdo em formato digital, em especial nas mídias sociais, sobretudo em função dos obstáculos a serem transpostos para que se possa garantir a preservação destes documentos, vale destacar aqui as restrições impostas por plataformas de mídias sociais e os dilemas éticos em se reutilizar as interações sociais dos indivíduos sem que estes saibam.

Pennock (2013, p .13, tradução nossa) destaca o dilema do contexto do que deve ser coletado e preservado no *Twitter*[8]:

> O arquivamento do *Twitter*, não deve considerar somente os *tweets* de forma isolada uma vez que são conversas estabelecidas. Sendo assim, pensar em se arquivar uma simples página do *Twitter,* significa considerar que somente um lado da conversa está sendo guardado. Como estabelecer limites para se definir com coerência uma coleção do *Twitter*? Seria necessário capturar todas as respostas de perfil de usuário, além deste perfil na íntegra? Seria necessário arquivar os perfis dos respondentes visando garantir informações suficientes sobre o contexto de um conteúdo específico? Baseando-se na relevância dos links para o *Twitter*, todos os links que foram compartilhados por um perfil também deveriam ser arquivados? Como o arquivo poderia garantir consistência temporal entre os links e o conteúdo presentes nestes sites, dada a curta duração de um link no *Twitter*? Definir os limites de uma rede social envolve questões bem mais complexas do que aparentemente parece ser algo simples.

O UK Data Forum ou Fórum de dados do Reino Unido de 2013 (UK DATA FORUM, p. 16a) apresenta o plano estratégico de cinco anos para guiar o desenvolvimento e utilização de dados para pesquisa social e econômica. O plano destaca que no contexto das mídias sociais,

[8] https://twitter.com/

milhões de interações humanas que ocorrem diariamente são registradas, criando enormes recursos de dados que em potencial auxiliam no entendimento de padrões de comportamentos sociais. Analisar as mídias sociais representa uma oportunidade de investir em pesquisas sociais de larga escala que seriam dificultadas utilizando-se outras ferramentas de coleta de dados, tais como as entrevistas ou questionários.

Apesar de que vários destes questionamentos ainda não tenham sido solucionados, faz-se necessário concentrar esforços para a captura, armazenamento e preservação de conteúdos relevantes presentes nas mídias sociais antes que estes sejam danificados ou perdidos totalmente. Diante do desafio de se definir estruturas, padrões e ferramentas tecnológicas que sirvam como referência para a coleta e arquivamento de conteúdos de mídias sociais, várias são as tentativas que têm se proposto a discutir e implementar tais ações. O presente estudo buscou identificar iniciativas científicas relevantes que têm arquivado conteúdos de mídias sociais ao nível mundial. Algumas das experiências listadas a seguir são referências na área e foram extraídas do relatório: *Digital Preservation Coalition Technology Watch Report,* de fevereiro de 2016, o qual apresentou como tema central a preservação de conteúdos de mídias sociais. Outras iniciativas foram localizadas após buscas realizadas em bases de dados bibliográficas e na *Web* utilizando-se os termos: "*web archiving", "research data", "digital preservation + social media", "social repository".*

3 Iniciativas científicas de arquivamento de conteúdos em mídias sociais

As iniciativas científicas que têm realizado este tipo de arquivamento, em geral são vinculadas a algum centro de investigação ou laboratório de análises de dados de pesquisas. Trata-se de um estudo de abordagem qualitativa de cunho exploratório uma vez que se buscou investigar e comparar iniciativas científicas inovadoras que coletam e arquivam conteúdos de mídias sociais. O presente estudo não pretende se tornar exaustivo no tocante à quantidade de experiências relatadas, mas ilustrativo uma vez que, além de apresentar experiências relevantes de arquivamento de conteúdos de mídias sociais em pesquisas científicas expõe também ferramentas de coleta e arquivamento consideradas pioneiras que poderão servir de inspiração de uso para futuros estudos.

Seguindo a definição de preservação digital de Ferreira (2006, p. 20) que consiste na capacidade de garantir que a informação digital permanece acessível e com qualidades de autenticidade suficientes para que possa ser interpretada no futuro recorrendo a uma plataforma tecnológica diferente da utilizada no momento da sua criação, buscou-se identificar nas experiências as seguintes características:

- Tecnológicas

- Identificar qual a plataforma de mídia social a ser preservada;
- Identificar qual a solução tecnológica utilizada para a coleta dos dados das mídias sociais e se foram utilizados serviços de terceiros;
- Identificar qual o formato de armazenamento dos dados coletados.

- Contextuais
  Identificar se os metadados coletados garantem:
  - Proveniência: os metadados de preservação devem registrar informações sobre a história do objeto desde sua origem, traçando a sua cadeia de custódia e de propriedade;
  - Autenticidade: os metadados de preservação devem incluir informações suficientes para validar que o objeto é de fato o que diz ser e que não sofreu alterações – intencionais ou não – não documentadas;
  - Gestão de direitos e tipos de acesso (público ou restrito à uma comunidade específica) conforme listados pelos autores Lavoie e Gartner (2005, p. 5, tradução nossa).

De maneira geral, as iniciativas científicas que têm coletado dados de mídias sociais estão utilizando estas coleções em pesquisas no campo das Ciências Sociais. Adiciona-se ao caráter de arquivamento o dinamismo do contexto investigativo das pesquisas científicas que acaba por determinar o escopo das coleções conforme os objetivos das investigações.

3.1 Social Data Science Lab: The COSMOS Platform of Cardiff University – UK

O Laboratório de Dados Sociais Científicos da Cardiff[9] University no Reino Unido mantém e distribui a plataforma COSMOS que visa facilitar análises de grande quantidade de dados extraídos de mídias sociais (o *Twitter,* em especial) de forma grátis e acessível para uso sem fins lucrativos. Trata-se de uma das mais relevantes iniciativas relacionadas com análises de dados de mídias sociais e desenvolvimento de novas metodologias computacionais de utilização destes dados (incluindo tipos de dados, metadados e formatos) em pesquisas no campo das Ciências Sociais. As pesquisas realizadas tentam responder a um vasto leque de questões relacionadas com criminalidade, tensões, riscos e bem-estar nas comunidades (*online* e *off-line*). Por meio da API nativa do *Twitter*, os dados brutos e seus metadados são extraídos para uma base de dados local gerada pela plataforma COSMOS. Estes dados originais em formado JSON são indexados e podem ser consultados por meio de uma camada intermediária de software que inclui atributos adicionais extraídos dos metadados originais dos dados brutos. Obedecendo a política de coleta, uso e desenvolvimento da plataforma de mídia social, a plataforma COSMOS segue com algumas restrições: No máximo 1% dos dados do *Twitter* podem ser coletados diariamente; os dados coletados não podem ser armazenados em

[9] http://socialdatalab.net/

dispositivos na nuvem e nem compartilhados com outras instituições (THOMSON, 2016, p. 28, tradução nossa).

A iniciativa tem se mostrado inovadora em se tratando de avanços no tratamento e análises de dados de mídias sociais. Várias parcerias têm sido feitas entre o Laboratório de Dados Sociais Científicos da Cardiff University e outros órgãos internacionais, incluindo as empresas proprietárias das plataformas de mídias sociais, visando incrementar as estratégias de desenvolvimento, armazenamento, coleta e análises destes dados. Tais iniciativas demonstram que este projeto precede ações inovadoras futuras de arquivamento e preservação de dados de mídias sociais, especialmente em se tratando de estruturas de bases de dados e enriquecimento de dados com atributos visando manter os contextos específicos em que foram criados.

3.2 Social Repository of Ireland

O Repositório Social da Irlanda[10] foi lançado em junho de 2015. O projeto foi criado e é implementado por um consórcio de cinco instituições científicas que juntas definiram políticas, diretrizes e capacitações referentes ao repositório. São elas: Royal Irish Academy[11] (RIA), instituição líder, National University of Ireland[12] - Maynooth (NUIM), Dublin Institute of Technology[13] (DIT), National University of Ireland – Galway (NUIG) e National College of Art and Design[14] (NCAD). Trata-se de um projeto que visa explorar os desafios de se arquivar dados de mídias sociais relevantes sobre a Irlanda preservando-os em um repositório digital confiável, o *Digital Repository of Ireland*[15] considerando aspectos legais e éticos. O projeto coleta dados do *Twitter*, por meio de sua API utilizando palavras chave e *hashtags*, especificamente sobre figuras públicas irlandesas, localizações geográficas e instituições relevantes para o país. Os algoritmos de coleta utilizados também fazem categorizações que permitem aos usuários buscarem conteúdos por temas específicos: esporte, política, eventos relevantes, dentre outros, utilizando metadados fornecidos pelo *Twitter*. Novamente, obedecendo a política de uso da plataforma de mídia social, o projeto pode coletar no máximo 1% dos dados do *Twitter* diariamente, não pode armazenar estes dados coletados em dispositivos na nuvem e nem os compartilhar com outras instituições. Exemplos de coleções que foram criadas são: dados de *tweets* contendo a opinião de cidadãos Irlandeses sobre

[10] https://www.insight-centre.org/content/social-repository-ireland
[11] https://www.ria.ie/
[12] http://www.nui.ie/
[13] http://www.dit.ie/
[14] http://www.ncad.ie/
[15] http://www.dri.ie/

casamento de pessoas do mesmo sexo cujo *referendum* constitucional ocorrido em maio de 2015 aprovou tal união (THOMSON, 2016, p. 29, tradução nossa). O projeto também coletou em 2016 dados de *tweets* sobre o "Brexit", termo utilizado para caracterizar a saída do Reino Unido da União Europeia (BBC BRASIL, 2016).

Uma vez que a equipe envolvida nesta iniciativa reconhece a relevância das mídias sociais como registro cultural e histórico de uma nação, um importante legado deste projeto para a União Europeia foi conduzir a elaboração de uma Carta Magna para Projetos de Dados. Esta iniciativa começou em 2014 e ainda está em andamento. Vários eventos têm marcado a trajetória da elaboração desta carta e as nações que fazem parte do bloco entendem que para que a sociedade possa desfrutar dos benefícios do *Big Data*[16] (conjuntos de dados complexos em que são aplicadas soluções tecnológicas de extração, curadoria, visualização e análises) é preciso registrar as preocupações públicas e criar uma plataforma justa em que todos possam confiar (O'SULLIVAN, 2016).

3.3 GESIS Leibniz Institute for the Social Sciences – GERMANY

Desde 1986, o GESIS – Leibniz-Institute for the Social Sciences[17] é a maior instituição independente de Ciências Sociais da Alemanha. Hoje conta com mais de 300 colaboradores e está localizado em duas cidades: Mannheim e Cologne. GESIS dispõe de serviços e infraestrutura relevantes para gestão de dados de pesquisas nacionais e internacionais. A instituição implementou um projeto piloto de arquivamento de dados de mídias sociais provenientes do *Twitter* e *Facebook* durante as eleições parlamentares de 2013 na Alemanha. Vários desafios surgiram ao longo do projeto, dentre eles a ausência de padrões e documentação sobre o assunto e a questão da volatilidade dos dados. No caso do *Twitter*, depois de extraídos por meio da API nativa, os dados em formato JSON ou XML são disponibilizados somente para pesquisadores autorizados, garantindo assim as questões relacionadas à política de acesso e uso do *Twitter e Facebook*. Uma vez que se trata de uma iniciativa de armazenamento de dados coletados das mídias sociais, a equipe do projeto não desenvolve API, deixando esta tarefa a cargo dos pesquisadores após terem acesso aos dados. Vale citar que o projeto também enfrentou dificuldades relacionadas com a autenticidade dos dados, não sendo possível a replicação exata da experiência de uso da mídia social. A coleta e arquivamento de dados do *Facebook* ocorreu em pareceria com o Laboratório de Computação em Ciências Sociais da Escola de Negócios de Copenhagen (CSSL) utilizando a ferramenta

[16] http://mike2.openmethodology.org/wiki/Big_Data_Definition
[17] https://www.gesis.org/en/institute/

SODATO (*Social Data Analytics Tool*)[18]. São coletados dados do *Facebook* em formato texto e imagens (THOMSON, 2016, p. 30).

3.4 DOCNOW – *Documenting the now* – UNITED STATES

O projeto americano DocNow é um exemplo de ação colaborativa entre pesquisadores e profissionais da informação que se utilizam de dados das mídias sociais para afirmar a importância da preservação destes conteúdos em escala nacional. Trata-se de um esforço colaborativo entre a Universidade de Maryland[19], a Universidade da Califórnia[20] em Riverside e a Universidade de Washington[21] em St. Louis contando com o financiamento da Fundação Andrew W. Mellon[22]. O projeto surgiu inspirado nas reações postadas no *Twitter* após o tiroteio policial de Michael Brown em Ferguson, Missouri, em 2014 (Wortham, 2016). Um dos arquivistas que lideram a equipe, Berguis Jules, na ocasião deste tiroteio, após o ocorrido se viu como milhões de americanos: conectado no *Twitter* em busca de notícias, reações e comentários sobre o fato. Nos dias seguintes, a *hashtag #IfTheyGunnedMeDown* que significa "e se eles me matarem?" desafiaram a narrativa liderada pela mídia convencional dando início a uma discussão relevante sobre estereótipos raciais e a brutalidade policial. A partir daí o arquivista começou a coletar dados do *Twitter* sobre o tiroteio e se mostrou extremamente indignado com a maneira como as mídias sociais mudam a forma de pensar dos cidadãos sobre fatos e acontecimentos da história.

Neste contexto, o DocNow tem um forte compromisso em priorizar práticas éticas ao trabalhar com conteúdo de mídias sociais, especificamente o *Twitter*, até o presente momento, em termos de coleta e preservação a longo prazo. Este compromisso se estende à noção do *Twitter* de honrar a intenção do usuário e seus direitos como criadores de conteúdo. O projeto tem como objetivo permitir que pesquisadores, estudiosos, profissionais da informação, entre outros, coletem e preservem os conteúdos das mídias sociais referentes aos eventos historicamente significativos utilizando-se de uma ferramenta amigável. No *site* do projeto, é possível ter acesso a várias ferramentas desenvolvidas com a filosofia de *software* livre disponíveis para uso. São elas:

[18] http://cssl.cbs.dk/software/sodato/
[19] https://www.umd.edu/
[20] http://www.ucla.edu/
[21] https://www.washington.edu/
[22] https://mellon.org/

- **Hydrator:** permite substituir os conjuntos de dados de identificadores de *tweets* que alguém já possua por outros mais completos contendo todos os metadados;
- **Tweet Catalog:** Catálogo de conjuntos de dados que estão disponíveis publicamente na web. Permite baixar conjuntos de dados de identificadores de *tweets* em formato JSON;
- **TWARC:** Ferramenta em linha de comando que permite baixar dados do *Twitter* em formato JSON;
- **DiffEngine:** Ferramenta que permite acompanhar as mudanças nos artigos de notícias por meio de seus *feeds* RSS.

4 Análise comparativa das iniciativas científicas de arquivamento de conteúdos de mídias sociais

O quadro 1 traz um comparativo das experiências científicas de captura e arquivamento dos dados de mídias sociais e apresenta uma característica relevante que é o foco na utilização dos dados para análises investigativas, com um propósito específico de cada problemática elaborada. Isto sugere que o foco principal destas iniciativas não seja diretamente a preservação para garantir acesso ao longo prazo, embora esta questão também seja intrinsecamente contemplada.

Quadro 1 – Análise comparativa de iniciativas científicas para preservação de mídias sociais

INSTITUIÇÃO/ PROJETO DE PRESERVAÇÃO DE MÍDIA SOCIAL	CARACTERÍSTICAS TECNOLÓGICAS			CARACTERÍSTICAS CONTEXTUAIS E DE ACESSO AOS DADOS	
	MÍDIA SOCIAL	**API e soluções tecnológicas adicionais**	**ARMAZ. DOS DADOS**	**METADADOS CONTEMPLADOS: proveniência, autenticidade e gestão de direitos**	**ACESSO AOS DADOS COLETADOS**

The COSMOS Platform of Cardiff University – UK	*Twitter*	• API do *Twitter*
		• COSMOS *Plataform*
			JSON	• Proveniência (dados brutos originais coletados e mantidos);
				• Autenticidade (não é garantida):
				  • todos os metadados originais visíveis;
				  • interface de indexação e busca dos dados;
				• Gestão de direitos: conforme políticas e termo de acesso, uso e desenvolvimento da mídia social.
					• Acesso local e remoto restrito a pesquisadores vinculados à instituição;
					• Dados coletados de usuários comuns obedecendo às restrições impostas pelo *Twitter*.

Digital Repository of Ireland / Social Repository of Ireland	*Twitter*	• API do *Twitter*	**JSON**	• Proveniência (dados brutos originais coletados e mantidos);
				• Autenticidade (não é garantida):
				  • todos os metadados originais visíveis;
				  • interface de indexação e busca dos dados gerando bases de dados específicas;
				• Gestão de direitos: conforme políticas e termo de acesso, uso e desenvolvimento da mídia social.
					• Acesso local e remoto restrito a pesquisadores científicos e jornalistas autenticados com *login* institucional;
					• Dados coletados de usuários comuns obedecendo às restrições impostas pelo *Twitter*.

GESIS Leibniz Institute for the Social Sciences (GERMANY)	*Twitter/ Facebook*	• API do *Twitter*;
		• SODATO (*Social Data Analytics Tool*)
			Twitter: **JSON** *Facebook*: **Texto e Imagens**	• Proveniência (não é garantida): Dados são provenientes de coleta realizada por projetos de pesquisa.
				• Autenticidade (não é garantida):
				  • metadados originais podem ter sido excluídos pela equipe do projeto que realizou a coleta;
				  • interface de indexação e busca dos dados conforme coleta realizada e base de dados específicas;
				• Gestão de direitos: conforme políticas e termo de acesso, uso e desenvolvimento da mídia social.
					• Acesso local e remoto restrito a pesquisadores autenticados com *login* institucional;
					• Dados coletados de usuários comuns obedecendo às restrições impostas pelo *Twitter*.

Docnow
 	
(USA)
	Twitter	• API do *Twitter;*
		• *Hydrator;*
		• *Tweet Catalog;*
		• *Twarc;*
		• *Diff Engine.*
			JSON	• Proveniência (dados brutos originais coletados e mantidos);
				• Autenticidade (não é garantida):
				  • todos os metadados originais visíveis;
				  • interface de indexação e busca dos dados gerando base de dados específicas;
				• Gestão de direitos: conforme políticas e termo de acesso, uso e desenvolvimento da mídia.
					• Acesso local e remoto sem restrições;
					• Dados coletados de usuários comuns obedecendo às restrições impostas pelo *Twitter*.

Quadro 1 – Comparativo da captura e arquivamento dos dados de mídias sociais

Fonte: os autores (2017).

Um ponto de destaque é o *Twitter* sendo a plataforma principal de coleta de dados utilizada por pesquisadores. Isto sinaliza que os *tweets* se apresentam como conteúdos com menos complexidade, se comparados a outros tipos de mídias sociais, como por exemplo os *posts* do *Facebook*. As iniciativas científicas destacadas utilizam a API da plataforma de mídia social para a captura dos dados e desenvolveram sua própria aplicação que realiza cruzamentos de dados e análises conforme o propósito das investigações.

Em relação aos metadados identificados nas experiências, os dados brutos originais são coletados e mantidos, isto garante sua proveniência, com exceção do GESIS do Leibniz Institute for the Social Sciences, que deixa a cargo dos pesquisadores dos projetos definições sobre as coletas e manutenções dos dados. A autenticidade não é garantida nestes projetos visto que geralmente não ocorre a imitação do ambiente original simulando a experiência dos usuários, os dados geram bases de dados específicas que alimentam os cruzamentos dos dados desejados. Em relação aos direitos de acesso, uso e desenvolvimento de *software*, os projetos seguem as regulamentações das plataformas de mídias sociais. Para os casos científicos os dados são coletados seguindo-se critérios pré-estabelecidos e são provenientes dos perfis de usuários de maneira geral, isto gera questões referentes à privacidade que devem ser consideradas conforme especificações de normativas das plataformas e legislações locais.

5 Conclusão

O uso de conteúdos das mídias sociais é algo recente e que tem ganhado relevância por parte de instituições científicas, especialmente em se tratando das ciências sociais. Vários são os desafios elencados relacionados com a prática do arquivamento e preservação deste tipo de conteúdo. Dentre eles, é possível destacar as mudanças contínuas nas plataformas de mídias sociais, especialmente relacionadas ao surgimento de novas aplicações e desaparecimento de outras, conferindo volatilidade a este tipo de conteúdo. A questão do uso colaborativo também é algo que dificulta o planejamento do arquivamento deste tipo de dados, uma vez que é preciso se considerar questões de permissões, direitos de acesso e os vínculos estabelecidos durante as trocas de informações. No tocante à seleção do material, é preciso se considerar os aspectos que demarcam o contexto em que as postagens ocorreram, tais como a temporalidade e temática dos conteúdos.

A transparência nas ações de coleta e arquivamento de conteúdos de mídias sociais deve ser um fator relevante, não somente por questões éticas, mas também visando conferir confiança aos usuários destas mídias. É preciso especificar como e por quem os dados coletados serão utilizados, principalmente em se tratando de pesquisas científicas.

Independente da finalidade de se preservar os conteúdos de mídias sociais é prudente afirmar que esta nova prática social de comunicação colaborativa tem contribuído para a avalanche informacional incessante, a qual desafia os mecanismos criados até então para coleta e arquivamento desta massa documental. Relevante deve ser a atuação curatorial nestes projetos, visando selecionar e então coletar o que de fato deve ser guardado.

A curadoria de dados de pesquisa pode ser considerada atualmente um assunto prioritariamente estratégico para vários países. Iniciativas munidas com soluções de infraestrutura tecnológica e material bibliográfico estão disponíveis visando favorecer a implementação de ações de gestão de dados investigativos. Comunidades internacionais lideradas por Austrália, Reino Unido, Estados Unidos e a União Europeia como um todo têm se voltado para promover iniciativas de gestão de dados de pesquisa. O Australian National Data Service *(ANDS)*[23] possui como principal objetivo gerenciar os dados australianos de pesquisa provenientes de instituições científicas, culturais, agências governamentais, dentre outras, tornando-os visíveis e acessíveis visando sua reutilização por meio de um portal de descobertas. O Reino Unido por meio de suas agências de fomento à pesquisa (Research Councils UK*)*[24] tem exigido dos pesquisadores a elaboração e implementação de planos de dados de pesquisa nas convocatórias e execuções de projetos investigativos financiados. Em 2013 o governo americano publicou um memorando[25] visando incrementar o acesso aos dados de pesquisas científicas financiadas por agências federais de fomento investindo valores consideráveis em uma política de acesso aberto às publicações e dados brutos dos projetos. No início de 2017, 99% das agências federais de fomento já tinham disponibilizado seus dados de pesquisas financiadas (SHEEHAN, 2017). A União Europeia também foi pioneira nas questões de compartilhamento no campo científico com seu plano piloto de acesso aberto aos dados de pesquisa por meio do projeto *Horizon 2020*[26]. Já é consenso que estes dados não possuem somente valor científico, mas também impacto econômico no tocante aos índices de inovação, conforme sinaliza Oettinger (2015) ao afirmar que a Comissão Europeia decide trabalhar por uma economia baseada em dados, ou seja, ciência aberta para uma economia de dados e do conhecimento.

É possível afirmar que a reutilização de dados de pesquisa por parte de pesquisadores, docentes, discentes e comunidade em geral traz ganhos inestimáveis, além de economia em

[23] Disponível em: <http://www.ands.org.au/>. Acesso em: 9 nov. 2017.
[24] Disponível em: <http://www.rcuk.ac.uk/>. Acesso em: 9 nov. 2017.
[25] Disponível em: <https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/ostp_public_access_memo_2013.pdf>. Acesso em: 9 nov. 2017.
[26] Disponível em: <https://ec.europa.eu/programmes/horizon2020/>. Acesso em: 9 nov. 2017.

relação a levantamentos e dados previamente coletados (ERWAY; RINEHART, 2016). Neste contexto de incentivo à reutilização, os dados coletados das mídias sociais se mostram como uma massa documental relevante e que pode ser utilizada em vários projetos científicos com temáticas e objetivos distintos sem que isto os prejudique ou diminua em termos de inovação e descobertas. Por fim, as restrições das plataformas de mídias sociais e a falta de transparência e compreensão sobre suas práticas de gestão de dados atualmente dificultam e em casos extremos até mesmo impedem a preservação deste tipo de conteúdo. No entanto, as soluções criativas apresentadas e o crescente movimento de comunidades de práticas que trocam informações sobre suas experiências parecem trazer importantes progressos na salvaguarda deste relevante legado cultural para atuais e futuros pesquisadores, profissionais da informação e comunidade de maneira geral.

Referências

BBC BRASIL. *O que é Brexit e como pode afetar o Reino Unido e a União Europeia?* 2016. Disponível em: http://www.bbc.com/portuguese/internacional-36555376 Acesso em: 7 maio 2017.

CONSELHO NACIONAL DE ARQUIVOS (CONARQ). *Coletânea da Legislação Arquivística Brasileira e Correlata.* Rio de Janeiro: Arquivo Nacional, 2017. Disponível em: http://www.conarq.arquivonacional.gov.br/index.php/coletanea-da-legislacao-arquivistica-e-correlata Acesso em: 10 ago. 2017.

DIGITAL PRESERVATION COALITION (DPC). *Creating digital materials.* Glasgow, UK, 2017. Disponível em: http://dpconline.org/handbook/organisational-activities/creating-digital-materials Acesso em: 31 out. 2017.

ERWAY, R.; RINEHART, A. *If you build it, will they fund? making research data management sustainable*. Dublin, OH: Online Computer Library Center, 2016. Disponível em: http://www.oclc.org/content/dam/research/publications/2016/oclcresearch-making-research-data-management-sustainable-2016.pdf Acesso em: 1 ago. 2017.

HOCKX-YU, Helen. Archiving social media in the context of non-print legal deposit. In: IFLA WORLD LIBRARY AND INFORMATION CONGRESS, 80., 2014, Lyon, França. *Proceedings…* Lyon, França: IFLA, 2014. Disponível em: http://library.ifla.org/999/1/107-hockxyu-en.pdf Acesso em: 10 ago. 2017.

LAVOIE, Brian; GARTNER, Richard. *Technology watch report: preservation metadata*. Dublin, OH: OCLC, 2005. Disponível em: http://www.dpconline.org/docman/technology-watch-reports/88-preservation-metadata/file Acesso em: 10 ago. 2017.

OETTINGER, Günther. *Open science for a knowledge and data-driven economy.* Brussels: European Community, 2015. Disponível em: https://ec.europa.eu/digital-single-market/en/news/open-science-knowledge-and-data-driven-economy Acesso em: 11 ago. 2017.

ORGANIZAÇÃO DAS NAÇÕES UNIDAS PARA A EDUCAÇÃO, A CIÊNCIA E A CULTURA (UNESCO). *A memória do mundo na era digital: digitalização e preservação*. Brasília, DF, 2012. Disponível em: http://www.unesco.org/new/fileadmin/MULTIMEDIA/HQ/CI/CI/images/mow/unesco_ubc_vancouver_declaration_pt.pdf Acesso em: 10 ago. 2017.

O'SULLIVAN, B. (Org.). The Magna Carta for Data Project. In: EUROPEAN COMMISSION'S ANNUAL MEETING OF THE JOINT RESEARCH CENTRE COMMUNITY OF PRACTICE ON BIG DATA, 1., 2016, Ispra, Italy. *Proceedings…* Ispra, Italy: INSIGHT (CENTRE FOR DATA ANALYTICS), 2016. Disponível em: https://www.insight-centre.org/content/magna-carta-data-project Acesso em: 7 maio 2017.

PENNOCK, Maureen. *Web archiving.* Glasgow, UK: Digital Preservation Coalition, 2013. 50 p. ISSN 2048 7916. Disponível em: www.dpconline.org/component/docs/doc_download/826-webarchivingpreviewmarch2013 Acesso em: 10 ago. 2017.

SALAHELDEEN, Hany M.; NELSON, Michael L. *Losing my revolution: how many resources shared on social media have been lost?* Ithaca, NY: Cornell University Library, 2012. Disponível em: https://arxiv.org/pdf/1209.3026.pdf Acesso em: 10 ago. 2017.

SAYÃO, Luís Fernando. Uma outra face dos metadados: informações para a gestão da preservação digital. *Encontros Bibli: Revista Eletrônica de Ciência da Informação*, Florianópolis, v. 15, n. 30, p.1-31, 2010. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/12528 Acesso em: 17 mar. 2017.

SHEEHAN, Jerry. *Making Federal Research Results available to all.* [Washington]: The White House, 2017. Disponível em: https://obamawhitehouse.archives.gov/blog/2017/01/09/making-federal-research-results-available-all Acesso em: 10 ago. 2017.

THOMSON, Sara D. *Preserving Social Media*. London: Digital Preservation Coalition, 2016. 47 p. ISSN 2048-7916. Disponível em: http://dx.doi.org/10.7207/twr16-01 Acesso em: 10 ago. 2017.

UK DATA FORUM. *UK strategy for data resources for social and economic research: a five-year plan to inform and guide the development and utilization of data and related resources for social and economic research.* [London], 2013. Disponível em: http://www.esrc.ac.uk/files/research/uk-strategy-for-data-resources-for-social-and-economic-research/ Acesso em: 10 ago. 2017.

UNIVERSITY OF LONDON. COMPUTER CENTRE (ULCC); UKOLN. *Preservation of web resources handbook*. [London], 2008. Disponível em: http://www.jisc.ac.uk/publications/programmerelated/2008/powrhandbook.aspx Acesso em: 10 ago. 2017.

WORTHAM, Jenna. How an archive of the Internet could change history. *The New York Times Magazine,* New York, p. MM18, June 21, 2016. Disponível em: https://www.nytimes.com/2016/06/26/magazine/how-an-archive-of-the-internet-could-change-history.html?_r=0 Acesso em: 30 out. 2017.

Recebido/Recibido/Received: 2017-11-16
Aceitado/Aceptado/Accepted: 2017-12-12