UNIVERSIDADE FEDERAL FLUMINENSE - UFF

INSTITUTO DE ARTE E COMUNICAÇÃO SOCIAL

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

MESTRADO EM CIÊNCIA DA INFORMAÇÃO

YURI MONNERAT LOTT

REGIMES E DISPOSITIVOS DE INFORMAÇÃO: ASPECTOS DA TECNOLOGIA APLICADA À VIGILÂNCIA E CONTROLE DE MASSA

NITERÓI

2017

UNIVERSIDADE FEDERAL FLUMINENSE - UFF

INSTITUTO DE ARTE E COMUNICAÇÃO SOCIAL

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

MESTRADO EM CIÊNCIA DA INFORMAÇÃO

YURI MONNERAT LOTT

REGIMES E DISPOSITIVOS DE INFORMAÇÃO: ASPECTOS DA TECNOLOGIA APLICADA À VIGILÂNCIA E CONTROLE DE MASSA

NITERÓI

2017

YURI MONNERAT LOTT

REGIMES E DISPOSITIVOS DE INFORMAÇÃO: ASPECTOS DA TECNOLOGIA APLICADA À VIGILÂNCIA E CONTROLE DE MASSA

Dissertação apresentada à Banca Examinadora do Programa de Pós-Graduação (stricto-sensu) em Ciência da Informação da Universidade Federal Fluminense como requisito parcial para obtenção do grau de mestre em Ciência da Informação.

Orientadora: Profª. Drª. Regina de Barros Cianconi

Linha de Pesquisa: 2 - Fluxos e Mediações Sócio-Técnicas da Informação

NITERÓI

2017

L884 Lott, Yuri Monnerat.
Regimes e dispositivos de informação: Aspectos da tecnologia aplicada à vigilância e controle de massa / Yuri Monnerat Lott. – Niterói: [s.n.], 2017.
116p.: il.; 30cm.

Orientadora: Prof.ª Drª. Regina de Barros Cianconi.
Dissertação (Mestrado em Ciência da Informação) –Programa de Pós-Graduação em Ciência da Informação, Universidade Federal Fluminense, 2017.
Referências: 110-119f.
1. Regimes de Informação. 2. Vigilância de Massa. 3. Big Data. 4. Algoritmos. 5. Dados Pessoais. 6. Universidade Federal Fluminense.
I. Cianconi, Regina de Barros (Orient.). II. Título.

CDD342

YURI MONNERAT LOTT

FOLHA DE APROVAÇÃO

REGIMES E DISPOSITIVOS DE INFORMAÇÃO: ASPECTOS DA TECNOLOGIA APLICADA À VIGILÂNCIA E CONTROLE DE MASSA.

Dissertação apresentada à Banca Examinadora do Programa de Pós-Graduação (stricto-sensu) em Ciência da Informação da Universidade Federal Fluminense como requisito parcial para obtenção do grau de mestre em Ciência da Informação.

Aprovado em: 26 / 04 / 2017

BANCA EXAMINADORA:

Profª. Drª. Regina de Barros Cianconi (Orientadora)
Universidade Federal Fluminense - UFF

Profª. Drª. Maria Nélida González de Gómez (Membro Interno)
Universidade Federal Fluminense - UFF

Profª. Drª. Sarita Albagli (Membro Externo)
Universidade Federal do Rio de Janeiro - UFRJ / IBICT

Profª. Drª. Elisabete Gonçalves Souza (Suplente Interno)
Universidade Federal Fluminense - UFF

Profª. Drª. Marcia Teixeira Cavalcanti (Suplente Externo)
Universidade Santa Úrsula - USU

DEDICATÓRIA

Dedico a conclusão desta dissertação à minha esposa Ana Cristina Lott, que ao meu lado dividiu as angustias e satisfações inerentes aos anos dedicados à pesquisa acadêmica, na prática, por também estar concluindo seu Mestrado, sendo exemplo de aplicação nos estudos e minha inspiração constante.

Aos meus pais Kátia e Eduardo Lott, pelo apoio incondicional aos meus projetos de vida, por todo investimento na minha educação e pelas oportunidades que serviram de base para a minha formação moral, para o meu o aprimoramento intelectual, dedicação à profissão e avanço nos assuntos acadêmicos.

AGRADECIMENTOS

Agradeço primeiramente à Profª Regina de Barros Cianconi, pelas orientações recebidas no curso do caminho percorrido, pela atenção dedicada e postura sempre pontual nas correções dos textos, pela parceria, apoio e compreensão das dificuldades, fundamentais para a conclusão deste trabalho de pesquisa.

À Profª Maria Nélida Gonzalez de Gomez, pelo incentivo para o desenvolvimento deste tema de pesquisa, pela atenção e disponibilidade nas consultas, pelo aprendizado durante as aulas de Regimes de Informação e generosidade no compartilhamento de assuntos de notável sofisticação conceitual.

À Profª Sarita Albagli, pelo acolhimento da minha participação na disciplina externa de Informação, Conhecimento e Poder, cujos temas abordados e discussões conduzidas permitiram o avanço sobre questões atuais, com referências complementares de grande pertinência e utilidade para o embasamento deste estudo.

Às Profª Elisabete Gonçalves e Profª Marcia Cavalcanti, pela disponibilidade, apoio e informações valiosas, que contribuíram para o aprofundamento de pontos específicos e diferenciais deste trabalho.

À Profª Coordenadora Ana Célia Rodrigues e ao Secretário Vitor Geraldo pelo atendimento de forma sempre prestativa e pelos trabalhos dedicados à manutenção, estruturação e organização do PPGCI-UFF, essenciais para a formação de novos pesquisadores, mestres e doutores.

À Comissão de Organização do Processo Seletivo de Mestrado 2015, ao Prof. Rodrigo Sales, pela receptividade e exímia condução dos trabalhos, ao Prof. Eduardo Murguia (*in memorian*), pelo voto de confiança.

À Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) e à Universidade Federal Fluminense (UFF) pela oportunidade da vaga e pela concessão da bolsa de estudos para a dedicação apropriada ao programa de Mestrado.

À minha família pelo incentivo constante e compreensão sobre os momentos de ausência. Aos amigos de jornada, em especial Tania Ferraro, Marta Fleming e Klaus Pereira - exemplos de determinação - pelas lições implícitas aprendidas com a superação dos desafios de uma vida dedicada aos estudos e à docência.

"(...) há estágios que, embora ocupem
insignificante fração da eternidade,
são de grande importância para os nossos objetivos,
pois neles, a entropia não aumenta e
a organização, e seu correlativo, a informação,
estão sendo criadas."

Norbert Wiener, Cibernética e Sociedade.

RESUMO

O presente trabalho consiste em uma pesquisa bibliográfica, de caráter exploratório e analítico, de abordagem multidisciplinar, que reúne aspectos socioculturais, político-legais, tecnológicos e metodológicos, para o entendimento sobre o atual regime de vigilância de massa com a aplicação das tecnologias sobre os dados pessoais. Buscou-se também identificar como a Ciência da Informação, no Brasil, vem incluindo os temas *vigilância*, *privacidade*, associados aos termos *big data* e *dados pessoais* em suas pesquisas acadêmicas de Mestrado e Doutorado. Foram abordadas questões como a geração massiva de dados a partir das novas tecnologias de informação e comunicação (NTIC), a ubiquidade da vigilância a partir da observação dos dados (*dataveillance*) e o uso de algoritmos para o processo automatizado para o reconhecimento de padrões, classificação de indivíduos (*profiling*) e distribuição seletiva de informação. A análise dos resultados demostrou que ocorre a inversão da visibilidade do poder e a transição dos objetos observados no processo de vigilância para os dados que os representam. Ao mesmo tempo, a distância ocorrida entre o observador e seus pontos de observação é mediada por novas tecnologias e algoritmos que permitem o processamento e respostas automatizadas para a gestão da informação em larga escala.

Palavras-chave: Regimes de Informação; Vigilância de Massa; Big Data; Algoritmos; Dados Pessoais.

ABSTRACT

A bibliographical research method was adopted for this work, and through an exploratory, analytical and multidisciplinary approach draws on sociocultural, political-legal, technological and methodological aspects, in order to understand the current regime of mass surveillance based on the use of personal data. The work also seeks to identify how Information Science in Brazil has begun to include the topics of *vigilance, privacy*, associated with the terms *big data* and *personal data* in Master's and PhD academic research. Issues covered include the mass generation of data through new information and communication technologies (NICT), the ubiquity of surveillance through data observation (*dataveillance*) and the use of algorithms for the automated processing of pattern recognition, classification of individuals (*profiling*) and selective distribution of information. Analysis of the results demonstrate a reversal in the visibility of power and the transition of observed objects in the surveillance process, for the data represented. Additionally, the distance between the observer and observation points is mediated by new technologies and algorithms that allow automated processing and responses for large-scale information management.

Key-words: Information Regime; Mass Sourveillance; Big Data; Algorithms; Personal Data.

LISTA DE FIGURAS

LISTA DE QUADROS

SUMÁRIO

1 INTRODUÇÃO

O atentado terrorista de 11 de setembro de 2001 às Torres Gêmeas na cidade de Nova Iorque (EUA) foi, sem dúvida, o acontecimento mais marcante deste início de séc. XXI, seja pelo estado de perplexidade em que deixou a opinião pública internacional, ou pelo impacto causado nas políticas de segurança dos Estados Unidos e dos principais países da Europa. A partir deste evento, percebeu-se um estado de alerta e insegurança geral por conta da ameaça de um inimigo público invisível e imprevisível, que resultou em uma intervenção radical dos governos sobre a privacidade no uso das tecnologias de comunicação e informação (BAUMAN, 2013). Em outubro do mesmo ano, como medida de segurança em reação ao ataque sofrido, o congresso americano instituiu o Patriot[1]Act concedendo amplos poderes às agências de segurança e inteligência para o monitoramento de movimentações financeiras e para a coleta de informações privadas de qualquer indivíduo que estivesse em território americano (GRABIANOWSKI, 2007).

Os anos que seguiram, levaram os Estados Unidos a expandir seu aparato de vigilância além dos limites legais, elevando seu poder de ubiquidade a um nível, comparável somente ao regime distópico descrito por Orwell, em sua obra de ficção "1984". Contudo, em 2013, Edward Snowden, um analista de sistemas terceirizado da Agência de Segurança Nacional Americana (NSA[2]), através dos jornais The American Post e The Guardian, publicou documentos secretos que evidenciaram os abusos do governo americano na espionagem de outros governos e na violação da privacidade de cidadãos americanos e estrangeiros. Foi possível saber da existência de um grande esquema de vigilância conhecido como Cinco Olhos (*Five Eyes*) e um sistema de monitoramento de dados - Prisma (PRISM) integrado por grandes empresas de tecnologia e compartilhado por vários agentes em rede.

Desde então, a transparência dos termos de uso e políticas de privacidade dos serviços digitais, que envolvem material de natureza pessoal e sensível, estão sendo questionados e muitos debates públicos vêm exigindo leis que garantam efetivamente a privacidade como um dos maiores direitos do indivíduo. Ao mesmo tempo, grupos conservadores, em posição oposta, apoiam o compartilhamento e a transparência total das informações pessoais em prol da segurança pública.

[1] Termo síntese em inglês para *Uniting and Strengthening America by Providing Appropriate Tools Required to Intercept and Obstruct Terrorism* (CONGRESSO AMERICANO, 2001).

[2] Sigla em inglês para National Security Agency

Ao mesmo tempo, no âmbito governamental quanto no setor privado, percebe-se que as organizações estão cada vez mais informatizadas e muitas desenvolveram seus próprios sistemas de gestão da informação. O atendimento e a prestação de serviços ao usuário, cliente ou cidadão, demandam a utilização de informações pessoais em modo digital para maior agilidade e acurácia na sua recuperação e análise. Contudo, o uso indevido das informações pessoais vem se tornando cada vez mais frequente, o que traz à tona discussões sobre os critérios de segurança dos bancos de dados e a transparência das políticas de privacidade que devem assegurar aos usuários a propriedade sobre seus dados.

Outros reflexos sociais do uso de informações pessoais podem ser destacados. Entre eles, o uso de algoritmos e inteligência artificial para o reconhecimento de padrões e categorização de indivíduos, com a consequente concessão de privilégios ou imposição de restrições, de acordo com a política de resposta estabelecida pelos gestores da informação. Esta prática, chamada por Bigo (2011) de *ban-opticon*, interfere na normatização da vida uma vez que a dependência das tecnologias da informação e dos sistemas digitais se dá em escala global. Se forem considerados os grandes centros urbanos, em muitos casos, os meios digitais representam o único viés de acesso aos serviços públicos e particulares. Deste modo, toda requisição, seja ela a solicitação de um plano de saúde ou de visto para a entrada em um país, torna-se um processo seletivo baseado em uma análise profunda, exaustiva e quase automática de todos os precedentes (de saúde, financeiro, legal e criminal) do indivíduo para o julgamento de sua requisição.

Outro ponto a ser destacado é a possibilidade (e a capacidade cada vez maior) de captação de dados a partir dos meios digitais. Para que um fato, evento ou instância da realidade possa ser computado, é necessário que este objeto seja convertido em um formato passível de ser reconhecido e processado pelos computadores. Nas palavras de Mayer-Schonberger e Cukier (2013), tudo deve ser *dataficado*, e esse processo de conversão difere substancialmente da simples digitalização uma vez que os diversos parâmetros do objeto devem estar estratificados e disponíveis para operações computacionais e análises algorítmicas.

Um dos fatores de distinção do grau de organização de uma sociedade e do seu domínio sobre as circunstâncias naturais e contingentes do seu tempo é a sua capacidade de registrar e recuperar informação. Desde a invenção da escrita, percebe-se que a humanidade vem buscando novos meios para recuperar os registros de suas atividades, em um caminho que visa, não só à preservação de suas memórias, mas também à perpetuação de seu legado

cultural. A Idade Média foi um período histórico bastante profícuo para as técnicas de tratamento e recuperação de informações. Já a Idade Moderna foi marcada pelo desejo de quantificação da natureza. De acordo com o espírito científico da época, todo objeto ou parâmetro, físico ou abstrato, poderia ser quantificado se fosse criado um instrumento de medição adequado e definido uma ordem de grandeza para ele. Partindo desse princípio, por hipótese, tudo pode ser medido e convertido em dados.

Com as facilidades tecnológicas existentes no atual cenário social e profissional, no campo da gestão e da pesquisa, a capacidade de captação de dados foi elevada a instâncias surpreendentes, chegando aos *conjuntos de dados massivos*. Em 1997, na NASA, o termo Big Data surgiu para definir, embora de forma não objetiva, a condição de uma base de dados que, pelo volume, velocidade e variedade de dados, excede as capacidades técnicas e de infraestrutura para seu armazenamento, processamento e visualização. Ao contrário dos tempos onde o registro das informações era feito basicamente em meio físico, hoje, a vasta presença de dispositivos óticos, leitores magnéticos e sensores dos mais diversos tipos, permite a conversão de objetos, fatos e eventos, do real para o digital, praticamente no momento em que eles ocorrem.

As atividades nos campos da disciplina, controle e segurança também migraram do campo analógico para o digital. Segundo Foucault (1987), as instituições disciplinares - a exemplo dos hospitais, fábricas e escolas - dependiam de uma estrutura arquitetônica para a manutenção da ordem sobre um determinado grupo de indivíduos. Nesse tempo, os instrumentos de operação eram os corredores, as baias, as catracas, os formulários de registro, os cartões ponto e o enfileiramento, visualização e contagem direta dos corpos, o que Foucault chamou de *física do poder*. Talvez o maior exemplo deste modelo seja o projeto do complexo penitenciário projetado por Jeremy Bentham, na Inglaterra do final do séc. XIX, o Panópticon. Mas, ao longo do tempo, com o advento de novos recursos para a gestão da informação, o monitoramento dos indivíduos deixou de ocorrer de forma direta e passou a ser feito de forma mediada, através dos dados que os representam. Deste modo, segundo alguns autores, a vigilância ampliou seu *spectrum* (Hookway, 2000), tornou-se distribuída (Bruno, 2013) ou líquida (Bauman, 2013) e ampliou seu poder de ubiquidade a escalas continentais, expandindo sua capacidade de visualização e rastreamento dos indivíduos ao nível das populações.

Esta pesquisa visa não somente a ressaltar a configuração atual dos regimes de poder estabelecidos pela captação, processamento e análise de conjuntos massivos de informação

(Big Data) para fins comerciais, mas também contribuir para o entendimento de um panorama mais amplo sobre as tecnologias de gestão da informação, sobretudo no âmbito pessoal, que possibilitaram e resultaram na crise atual onde se destaca a falta de privacidade e liberdade.

Considerando o atual dilema entre a necessidade de segurança e a consequente perda de privacidade com o compartilhamento de dados pessoais, no momento em que grande parte da população mundial marcha sobre a esteira globalizante e se torna cada vez mais dependente das novas tecnologias de informação e comunicação (nTIC), traz-se como ponto focal desta dissertação responder às seguintes questões: Como as novas tecnologias de gestão da informação, através dos seus métodos de captação, processamento e resposta, sobre o uso de informações pessoais, possibilitaram a instalação do atual regime de controle e vigilância de massa? Como a Ciência da Informação, no Brasil, vem incluindo os temas *vigilância*, *privacidade*, *big data* e *dados pessoais* em suas pesquisas acadêmicas de Mestrado e Doutorado, nos últimos dez anos?

Com o objetivo geral de analisar o panorama dos regimes e dos dispositivos de vigilância e controle, estabelecidos por governos e organizações privadas, e seus efeitos sobre as populações, bem como identificar se estas questões vêm sendo abordadas nas pesquisas recentes da Ciência da Informação, buscou-se, como objetivos específicos: (a) investigar as bases teóricas e conceituais a respeito dos regimes e dispositivos de informação para a vigilância e controle de massa; (b) investigar as práticas de vigilância e controle de massa, através dos processos de captação, processamento e resposta, sobre o uso de dados pessoais e seu impacto sobre a liberdade dos indivíduos e (c) verificar, nos anais do Encontro Nacional de Pesquisa em Ciência da Informação no Brasil (ENANCIB) nos últimos dez anos se as temáticas *vigilância* e *privacidade* são abordados nas pesquisa científicas deste campo do conhecimento.

Nessa pesquisa bibliográfica, de carácter exploratório e analítico, foram consultadas fontes secundárias: livros, periódicos (jornais e revistas), teses, dissertações e artigos científicos no âmbito da Ciência da Informação e outros campos, como a Sociologia e a Tecnologia da Informação, na observância de estudos que se mostraram pertinentes para o esclarecimento dos problemas existentes nas relações entre poder e sociedade, e entre as instituições e os indivíduos, nas questões de vigilância, privacidade, liberdade e segurança.

Em termos temporais, a pesquisa buscou fontes e relatos sobre os temas abordados independentemente de suas datas de publicação, utilizando referências históricas, sobretudo para a composição do seu alicerce. Contudo, deu-se ênfase para os fatos e publicações dos

últimos quinze anos por abarcar os maiores avanços tecnológicos na infraestrutura de comunicação e informação, e por estarem relacionados, diretamente ou não, aos marcos históricos do atentado terrorista de 11 de setembro às Torres Gêmeas na cidade de Nova Iorque (2001) e das revelações feitas por Edward Snowden sobre o esquema de vigilância e espionagem do Governo Americano e aliados (2013).

Para a interpretação das questões levantadas, foi utilizado o método qualitativo, de forma analítica e explicativa, ao decompor o cerne da pesquisa em seus principais componentes e apontar a relação entre eles, evidenciando seus papéis de causa e efeito.

Para identificar se a Ciência da Informação, no Brasil, vem incluindo tais temáticas em suas pesquisas acadêmicas de Mestrado e Doutorado, foram consultados os anais do Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB) dos últimos 10 anos.

A oportunidade para o desenvolvimento desta pesquisa é sinalizada no momento em que se observa, na mídia especializada e na literatura científica, a emergência de temas como Big Data, Computação nas Nuvens e Internet das Coisas, bem como o surgimento de novas disciplinas como a Data Science e das práticas de aprendizagem de máquina (*machine learning*) no campo da pesquisa.

A automação do processamento de dados em larga escala vem sendo vista como solução prioritária para os problemas de gestão e da segurança devido à capacidade de reconhecimento de padrões e à crença sobre a possibilidade de predição do comportamento. Nesse contexto, surgem a vigilância de massa por parte dos órgãos de segurança dos governos e a vigilância corporativa no setor privado. Percebe-se também a necessidade de pesquisa nesta temática para a Ciência da Informação em função da escassez de artigos publicados nas principais revistas e de trabalhos apresentados em congressos deste campo de pesquisa.

O tema em questão também se mostra relevante, considerando as mudanças socioculturais e mercadológicas em função dos avanços tecnológicos nas últimas duas décadas, para os quais ainda não existe suporte legal que garanta a estabilidade e imparcialidade de seus efeitos. O Estado e a sociedade altamente informatizados migraram suas práticas de gestão da informação para o âmbito digital, enfatizando o poder informacional e a informação como produto. Hoje, com os modos de produção e de organização pessoal altamente dependentes dos padrões estabelecidos pela indústria da tecnologia digital, antigos conceitos sobre individualidade, propriedade e privacidade tornaram-se obsoletos e por isso precisam ser discutidos e reestabelecidos. A liberdade é posta

em cheque, ao mesmo tempo em que as fronteiras que separam o público do privado tornam-se tão turvas quanto o limiar entre o real e o virtual.

Após esta Introdução, na segunda seção, é feita uma ampla abordagem sobre os regimes de vigilância. Primeiramente são apresentados os conceitos de dispositivo e regime elaborados por Foucault, e adotados por Bigo (2011), González de Gómez (2012) e Bruno (2013). De forma complementar, são trazidos elementos socioculturais, político-legais, tecnológicos e metodológicos, que caracterizam os regimes disciplinares e de controle, e as formas de poder através da materialidade da informação e dos mecanismos de mediação. Aborda-se as instituições disciplinares (Séc. XVIII e XIX) descritas por Foucault (1987) – a exemplo do Panopticon – até o surgimento do Estado Informacional (Braman, 2006b) no qual a vigilância se dá a partir dos meios digitais, de forma indireta, ubíqua e invisível, trazendo os conceitos de *panspectron* (Hookway, 2000; Braman, 2006b), vigilância líquida (Bauman, 2013) e vigilância distribuída (Bruno, 2013). Na sequência, são apresentadas as definições de meta dados, dados pessoais, dados sensíveis e citados alguns instrumentos legais para a preservação dos direitos à privacidade ou propriedade dos dados pessoais, no Brasil.

Na terceira seção é apresentado, em linhas gerais, como o processo de captação de dados pessoais se dá a partir do uso das novas tecnologias de informação e comunicação (NTICs), muitas vezes, facilitada pela permissividade e compartilhamento de informações pessoais pelos próprios usuários. São expostas as condições iniciais da vigilância com os processos de geração e captação de dados pessoais; é esclarecido como surge o Big Data através da ação de seres humanos e não humanos, e ainda, como se dá o processo de dataficação de corpos, ações e eventos.

A quarta seção apresenta alguns aspectos da tecnologia voltada para o domínio sobre a informação. Em um primeiro momento, enfatizando o aspecto físico da informação enquanto recurso, traz-se um breve histórico do desenvolvimento tecnológico voltado para o armazenamento e estruturação da informação. No segundo momento, dá-se ênfase para o aspecto semântico da informação, com a abordagem sobre os processos de tratamento e análise de dados pessoais. Apresenta-se o conceito de algoritmo e explica-se como se dá o processo de mineração de dados para a verificação da similaridade e agrupamento de objetos (*clusterização*), através do sistema vetorial.

A quinta seção aborda as possibilidades da personalização e categorização seletiva em sistemas digitais. Para isso, faz-se uma descrição do processo de captação de dados pessoais e consequente análise para o reconhecimento de padrões, criação de perfis (*profiling*) e

classificação de indivíduos, para os quais, foram trazidos dois exemplos atuais de algoritmos utilizados para a *predição de comportamento* e *distribuição seletiva da informação* (Netflix e Facebook). Por fim, fecha-se o ciclo do processo de vigilância digital abordando os usos de dados pessoais e efeitos da personalização, distribuição seletiva de informação e categorização seletiva sobre os indivíduos. São trazidos exemplos de usos de dados pessoais no campo do marketing e da segurança, entre eles o ban-opticon e outras práticas comerciais que afetam a privacidade dos usuários com o uso indevido de seus dados pessoais.

A sexta seção traz um levantamento quantitativo de como a Ciência da Informação, no Brasil, vem incluindo os temas *vigilância*, *privacidade*, *big data* e *dados pessoais* em suas pesquisas acadêmicas de Mestrado e Doutorado. Foram consultados os anais do Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB) dos últimos 10 para verificar as abordagens dos temas *vigilância* e *privacidade* junto ao uso de *Big Data*, e *dados pessoais* - termos chave que definem o escopo deste estudo.

Na última seção são feitas a análise dos resultados, considerações complementares ao conteúdo apresentado na dissertação e conclusões sobre a pesquisa.

2 REGIME DE VIGILÂNCIA

Para uma maior noção a respeito da natureza complexa e fluida, multi-articulada, abrangente e permeável do *panspectron*, ou dos regimes da vigilância líquida / distribuída que serão abordados nesta pesquisa com base em Braman (2006b), Bauman (2013), Bigo (2011) e Bruno (2013), cabe inicialmente discorrer sobre os conceitos de *dispositivo* e *regime* elaborados por Michael Foucault (1990), utilizados por muitos autores e contextualizados por Frohmann (1995) e González de Gómez (1999) no âmbito da Ciência da Informação.

Segundo Foucault (1990, p.244), o *dispositivo* é o sistema de relações existentes em um composto heterogêneo formado por "discursos, instituições, organizações arquitetônicas, decisões regulamentares, leis, medidas administrativas, enunciados científicos, proposições filosóficas, morais, filantrópicas. Em suma, o dito e o não dito". A relação entre esses componentes constitui um *apparatus* dinâmico onde a importância e a ênfase de ação oscila entre seus agentes de acordo com a resposta necessária a ser dada em cada contexto e situação (BIGO, 2006). Um dispositivo é algo de natureza funcional e operacional, embora inserido em uma estratégia de Poder; portanto, não pode ser definido *a priori* por sua intenção ou direção, e por isso, também não possui neutralidade, uma vez que sua existência é definida e justificada pelos efeitos que ele é capaz de causar (GONZÁLEZ DE GÓMEZ, 1999).

Expandindo a noção de dispositivo, chega-se ao conceito de *regime*. Segundo Foucault, um regime de verdade designa

> os tipos de discursos que ela acolhe e faz funcionar como verdadeiros; os mecanismos e as instâncias que permitem distinguir os enunciados verdadeiros ou falsos, a maneira pela qual se sanciona uns e outros; as técnicas e os procedimentos que são valorizados para a obtenção da verdade; o estatuto daqueles que se encarregam de dizer o que funciona como verdadeiro (FOUCAULT, 1994, p. 112 apud BRUNO, 2013).

A partir da fala original do autor, entende-se que um regime de verdade é constituído por forças que estabelecem um padrão de verdade ou o *status quo* de uma sociedade em determinada época. Esta condição é instituída por agentes políticos, institucionais ou individuais que através de seus discursos conseguem mobilizar a opinião de diversos públicos, em diversas instâncias, para legitimação de uma ideia. Os dispositivos descritos acima ajudam a estabelecer o conjunto de regras e o ordenamento social necessário para a manutenção da verdade proposta pelo regime (BRUNO, 2013; GONZÁLEZ DE GÓMEZ, 1999, 2012).

Frohmann (1995) ressalta a importância da informação como forma de mediação/integração das relações sociais exercidas pelo poder e estabelece como uma das questões fundamentais para os estudos das Políticas de Informação a dominação sobre a informação, conquistada e mantida, por grupos específicos. O autor define também que o regime de informação se configura em um sistema ou rede através dos quais a informação flui por determinados canais - a partir de certos produtores, via estruturas organizacionais, para usuários específicos.

González de Gómez, a partir de Frohmann, contextualiza o conceito de regime no âmbito da informação densificando a sintagma *regime de informação*, que seria

> (...) o modo informacional dominante em uma formação social, o qual define quem são os sujeitos, as organizações, as regras e as autoridades informacionais e quais os meios e os recursos preferenciais de informação, os padrões de excelência e os modelos de sua organização, interação e distribuição, enquanto vigentes em certo tempo, lugar ou circunstância (GONZÁLEZ DE GÓMEZ, 2012, p.43).

A autora enfatiza ainda que um regime de informação é constituído a partir de variáveis culturais, políticas e econômicas, inerentes a um campo social, em determinado tempo, e remetem à relação de poder através da informação que perfaz todas as instâncias do dispositivo. O poder atua autorizando, legitimando e potencializando certos *locus* anunciatórios, instituições e representações políticas que, através de seus discursos, alteram a percepção, a opinião e os resultados de comportamento de uma certa população. A formação do Regime é, portanto, "o processo pelo qual novas formas políticas emergem fora do campo da política" (BRAMAN, 2004 apud GONZÁLEZ DE GÓMEZ, 2012, p.52).

Para que seja compreendido como o estado atual de informatização da gestão, seja no setor público ou privado, viabiliza intervenções que aviltam o direito à privacidade, e como as práticas de gestão da informação em larga escala tornaram possíveis a vigilância de massa, é

necessário um breve retorno às origens dos sistemas de vigilância e controle, com a apresentação do, talvez, maior símbolo de opressão da liberdade e da privacidade, o projeto de complexo penitenciário idealizado por Jeremy Bentham: o Panopticon.

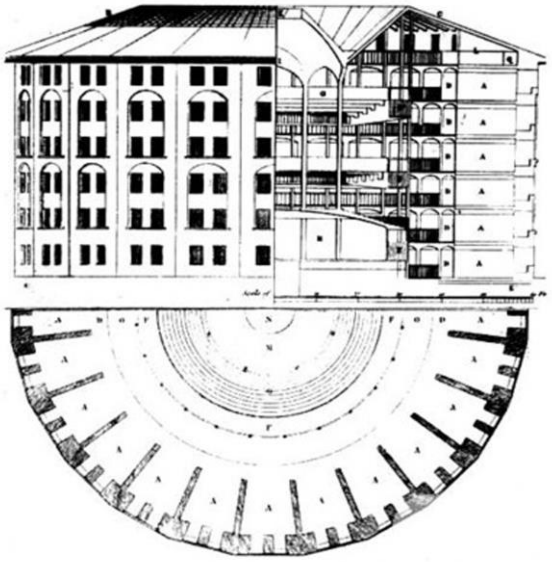
2.1 O PANOPTICON

Jeremy Bentham nasceu em Londres no ano de 1748, e foi considerado um jovem prodígio; aos doze anos ingressou na Queen´s College (Oxford) onde em 1764, com apenas dezesseis anos, tornou-se o aluno mais novo a se graduar naquela escola (SCHOFIELD, 2009). No ano seguinte, impelido pelos passos de seu pai – um notável advogado – Bentham ingressou no Lincoln´s Inn para o estudo do Direito. Logo nos primeiros contatos com a corte Inglesa, Bentham se decepcionou com a forma pela qual o sistema judiciário era interpretado e passou a buscar a reforma das leis, principalmente as do Código Penal.

A inspiração para o projeto arquitetônico de um complexo penitenciário surgiu mais tarde, ao final da década de 1780, quando Jeremy visitou seu irmão Samuel na Criméia - parte do Império Russo - onde trabalhava como supervisor de uma fábrica. Para facilitar a inspeção dos operários, Samuel projetou um *layout* de produção circular onde ele, do centro, podia monitorar e manter o controle sobre toda atividade fabril. Bentham percebeu que essa estrutura dava ampla vantagem de poder para o supervisor sobre os operários e adotou o princípio que ficou conhecido como *panoptismo*, a ser aplicado em instituições que devessem exercer controle sobre um grande contingente de pessoas, com o mínimo de esforço para sua administração, a exemplo de escolas, sanatórios, abrigos e, sobretudo, prisões (SEMPLE, 1993).

Em seu retorno à Inglaterra, Bentham apresentou ao governo do estado de Londres sua proposta para o novo complexo penitenciário, que ele chamou de Panopticon (Figura 2). Percebendo nas autoridades uma inclinação inicial favorável à sua ideia, Bentham seguiu em dedicação ao projeto, de forma intensa, por mais de dez anos (fazendo inclusive investimentos às suas próprias custas), contudo, o projeto não veio a se concretizar.

Figura 1 - Projeto arquitetônico do Panopticon. Desenho de Willey Reveley, 1791



Fonte: Vigiar e Punir. Foucault, 1987.

O Panopticon tornou-se o ícone do estado utilitarista de Bentham – com todas as suas ambiguidades – e, embora não tenha sido realizado, ainda serviu de base para a arquitetura de outros complexos de segurança, a exemplo da penitenciária de Millbank (Londres, Inglaterra, 1812 ~ 1902), Pentoville (Londres, Inglaterra, 1842), o Presídio Modelo (Ilha da Juventude, Cuba, 1925 ~ 1973) e Stateville (Illinois, EUA, 1925), confirme a Figura 2.

Figura 2 - Penitenciária de Stateville (Illinois, EUA, 1925)

Fonte: http://prisonphotography.org/ (Foto: Doug DuBois e Jim Goldberg)

Sob a ótica dos estudos de arquitetura de Evans (1982 *apud* SEMPLE, 1993), o Panopticon, mais que um simples prédio, é um instrumento projetado para aumentar o poder da administração sobre a contingência da população carcerária ou, como considera Foucault (1987), mais que um simples teto, um "operador terapêutico".

Foucault, em 'Vigiar e Punir', foi um dos autores que melhor analisou o projeto Panopticon de Bentham, e não somente sob o viés arquitetônico, mas principalmente em relação aos aspectos funcionais que visavam à disciplina e o controle.

> (...) na periferia uma construção em anel; no centro, uma torre; esta é vazada de largas janelas que se abrem sobre a face interior do anel; a construção periférica é dividida em celas, cada uma atravessando toda a espessura da construção; elas têm duas janelas, uma para o interior, correspondendo às janelas da torre; outra, que dá para o exterior, permite que a luz atravesse a cela de lado a lado. Basta então colocar um vigia na torre central, e em cada cela trancar um louco, um doente, um condenado, um operário ou um escolar. Pelo efeito da contraluz, pode-se perceber da torre, recortando-se exatamente sobre a claridade, as pequenas silhuetas cativas nas celas da periferia. Tantas jaulas, tantos pequenos teatros, em que cada ator está sozinho, perfeitamente individualizado e constantemente visível (FOUCAULT, 1987, p.165-166).

Foucault chama a atenção para os dispositivos inerentes às instituições disciplinares, nos séculos XVIII e XIX, com o objetivo de discriminar os elementos da massa, que consistiam em: *individualizar, identificar e classificar*. Segundo ele, esta seria uma prática de marcação binária que todos os mecanismos de poder, em nossos dias, utilizam para ter o

domínio sobre o *anormal*. Portanto, segundo o autor, "a visibilidade é uma armadilha" (FOUCAULT, 1987, p.166).

O efeito mais dominante do Panopticon, portanto, era causar no detento a tensão e a percepção do estado de vigilância constante onde o funcionamento do poder é automático. Foucault ressalta que o efeito a longo prazo de uma vigilância constante seria capaz de criar uma entidade censora, por antecipação, de qualquer ato de infração. O poder seria despersonalizado, portanto exercido independentemente do indivíduo que estivesse por trás do posto de observação, e onipresente, atuante mesmo nos períodos em que não houvesse de fato alguém na torre de vigilância. Neste sentido, Foucault diz que Bentham declarou o princípio do Poder *visível e inverificável*, e que o Panopticon seria a máquina de dissociar o "par ver-ser-visto" onde, no anel periférico, o indivíduo é visto sem nunca ver, enquanto que na torre central, vê-se tudo, sem nunca ser visto (FOUCAULT, 1987, p.167).

Outro aspecto observado por Foucault foi a leveza dos prédios e estruturas das instituições disciplinares pois, uma vez que elas privilegiavam a visibilidade dos corpos, seus ambientes eram mais devassados e compostos por uma geometria simples e econômica. Ao contrário das *casas de segurança*, não haveria grades pesadas, paredes espessas e correntes; tudo seria feito de forma a garantir a objetividade de uma *casa de certeza*. Assim, segundo Foucault (1987), o poder pôde, de forma gradual e contínua, livrar-se de seus fardos físicos com uma arquitetura que tende ao incorpóreo e, quanto mais se aproxima deste limite, mais amplo, presente e definitivo ele se torna.

A solução do Panopticon é versátil e não está limitada aos complexos penitenciários, conforme Bentham (1791c) publicitava na venda de seu projeto. O mesmo princípio de construção e suas formas de organização poderiam ser aplicados a qualquer estabelecimento onde "pessoas devem estar sob inspeção", a exemplo de fábricas, hospitais, sanatórios e escolas; ou mais que isso, "é destinado a se difundir no corpo social; (e) tem por vocação, tornar-se uma função generalizada" (FOUCAULT, 1987, p.171).

2.2 AS INSTITUIÇÕES DISCIPLINARES E A INVERSÃO DA VISIBILIDADE

Conforme Foucault (1987) descreve o modelo de dispositivo disciplinar aplicado para conter o surto da peste em cidades da Europa tinha os mesmos efeitos de uma amputação cirúrgica, para a extração de um cisto social maligno; era o do retalhamento dos espaços

geográficos, ratificação do isolamento e identificação de cada indivíduo com seu respectivo estado de saúde.

A *hierarquia* do Corpo de Inspeção permitia o avanço da vigilância através de uma rede altamente capilarizada e capaz de verificar o estado de cada rua, quarteirão ou casa; de cada órgão ou célula da massa estratificada. A cidade era imobilizada para evitar a mistura, o atrito e a agitação, e inspecionada regularmente, *de forma direta e sensorial*; os moradores se reportavam aos soldados da guarda e aos síndicos das ruas, que se reportavam aos intendentes de quarteirão, que por fim respondiam ao prefeito; da mesma forma, em sentido contrário, o prefeito verificava a qualidade do trabalho dos intendentes, que supervisionavam o trabalho dos síndicos, e assim por diante.

Toda cidade era traçada e transcrita em um sistema de registro permanente com o qual podia-se obter profundo *controle* sobre todas as atividades de inspeção e situações encontradas. A ordem

> (...) prescreve a cada um seu lugar, a cada um seu corpo, a cada um sua doença e sua morte, a cada um seu bem, por meio de um poder onipresente e onisciente que se subdivide ele mesmo de maneira irregular e ininterrupta até a determinação final do indivíduo, do que o caracteriza, do que lhe pertence, do que lhe acontece (FOUCAULT, 1987, p.163-164).

Tem-se caracterizadas aqui, portanto, duas imagens, dois estágios da história humana com seus diferentes regimes de disciplina: em um extremo, a disciplina da instituição fechada, voltada para a repressão do mal e o poder pontual, através do controle da comunicação marginal e da supressão do tempo; no outro, o mecanismo panóptico do poder visível e inverificável, de intervenção sofisticada e abrangente. De um lado, a disciplina da exceção (séc. XVII), do outro a vigilância generalizada (séc. XVIII) e, entre esses dois pontos, um espaço de transição que é preenchido pela "extensão progressiva dos mecanismos de disciplina (... e) sua multiplicação através de todo o corpo social, a formação do que se poderia chamar grosso modo a sociedade disciplinar" (FOUCAULT, 1987, p.173).

Outras mudanças podem ser notadas sobre a visibilidade do poder. No início, a chamada *sociedade do espetáculo*[3] convergia a atenção da massa popular para a figura do Rei em seu altar. Toda demonstração de poder era revestida de um aparato cenográfico e ritualístico que ampliava a visibilidade do monarca. Contudo, como todo espetáculo, os

[3] Termo utilizado por Foucault (1987) para referir aos contextos sociais em que a figura do monarca ou déspota, e toda indumentária Real, colocava-se em local de destaque no cenário do cotidiano, inevitavelmente, chamando para si a atenção da população.

holofotes direcionados para o palco cegavam a visão dos atores, ao mesmo tempo que ocultavam os espectadores da plateia. O local de poder era visível, enquanto toda a população sob a égide da coroa permanecia anônima e desconhecida.

Com o passar da era clássica para a moderna, ocorre significativa mudança dos dispositivos de visibilidade e vigilância, através dos quais o Poder revestiu-se de um anteparo de isolamento informacional e a população passou a ser cada vez mais conhecida e controlada pelas instituições do governo.

Essa questão envolve uma arquitetura – e deve-se entender aqui por *arquitetura*, não só a forma física, aparência e estrutura funcional dos prédios, mas também todos os arranjos processuais e tecnológicos de gestão da informação – cujo objetivo não era mais chamar a atenção (palácios) ou permitir uma vista privilegiada de seu entorno (fortalezas), mas ampliar e detalhar o foco sobre seu ambiente interno, tornando visíveis os que ali se encontravam, sendo um operador de transformação. A antiga sociedade do espetáculo – onde muitos olhavam poucos através da arquitetura monumental dos templos, teatros e circos – passou a ser a sociedade moderna da física panóptica, com seus dispositivos que levaram a uma "distribuição infinitesimal do poder" (FOUCAULT, 1987, p.178).

Além de permitir o controle sobre os meios de produção – cada vez mais extensos e complexos –, os mecanismos disciplinares surgiram para dar solução a outros dois problemas modernos: a explosão demográfica e a população flutuante. Assim, segundo Foucault, com o poder disciplinar é possível assentar os nômades e assegurar a "ordenação das multiplicidades humanas"; reger, subdividir, estruturar, individualizar e classificar a massa, conduzindo esta à uma finalidade mais útil.

Da mesma forma que aponta os efeitos dos novos dispositivos de poder, Foucault também reconhece a desproporcionalidade de direitos imposta pelo regime disciplinar que se instala em oposição à relação contratual justa. Segundo ele, as disciplinas têm o efeito irreversível de introduzir assimetrias e excluir reciprocidades; ao invés de um equilíbrio, os pesos *não* se distribuem de forma igualitária na relação entre as partes, onde o cidadão, o cliente ou usuário final, o estudante, o paciente, o operário ou o detento, se tornam reféns das grandes estruturas burocráticas, em sua maioria despersonalizadas e engessadas, que impõem seu domínio regimental.

Em estudo posterior, com base na obra de Foucault, Deleuze (1992) deu continuidade à crítica dos dispositivos adotados pelos meios de produção, em geral, baseados no controle dos recursos humanos. Segundo o autor, diferente das instituições disciplinares, as formas de

controle surgidas ao final do século XX não estavam restritas ao confinamento e visibilidade direta dos corpos. Elas aprimoraram o processo de utilização da informação como meio representante dos indivíduos e através dela puderam tornar o regime de vigilância ainda mais pervasivo e eficiente.

Deleuze cita a linguagem numérica, não necessariamente binária, e a utilização de recursos digitais (diferentes dos analógicos) muito superiores na sua capacidade de reconhecer padrões numéricos, identificar estados e situações, e estabelecer a distinção entre os indivíduos dentro da massa.

Um dos aspectos enfatizados pelo autor é a superação do modelo de confinamento por instrumentos de controle normativos que incentivam o indivíduo a seguir um padrão de comportamento esperado e desejado para realizar seus objetivos. As cifras são mencionadas como um recurso de identificação que permitem o acesso à informação ou a sua restrição e os tipos de máquinas (ou recursos tecnológicos) são adotados como referência para a observação da capacidade de cada regime (disciplinar ou de controle) em manter a ordem ou a normalidade desejada sobre as populações: “as máquinas simples ou dinâmicas para as sociedades de soberania, as máquinas energéticas para as de disciplina, as cibernéticas e os computadores para as sociedades de controle” (DELEUZE, 1992, p.216).

A superação do analógico pelo o digital, assim como a transição do burocrático para o informatizado também é um assunto pertinente para o entendimento sobre a instalação do regime de vigilância atual.

2.3 O ESTADO INFORMACIONAL E OS NOVOS REGIMES DE VIGILÂNCIA

O regime burocrático despontou na década de 1940 como solução para aumentar o nível de eficiência e profissionalismo da gestão pública, onde as demandas dos cidadãos deveriam ser tratadas de forma imparcial, transparente e justa. Para isso, a administração pública deveria ser regrada por processos bem definidos e declarados para o conhecimento de todos. Para cada caso, as atividades eram registradas passo a passo através de um complexo aparato documental, normalmente representado por formulários impressos, que garantiriam a rastreabilidade das informações e a possibilidade de averiguação da idoneidade de cada ato administrativo. A princípio, a burocracia seria a solução para garantir a moralidade na gestão da coisa pública, mas em poucas décadas este modelo ficou marcado pela imagem de uma

máquina engessada (com processos rigorosos que só serviam para criar pilhas de papéis), pelo excesso de pessoal empregado, pelo desperdício e também, pela falta de produtividade.

Após o esgotamento da proposta burocrática, os governos em geral começaram a atualizar seus modelos gerenciais, agora com foco na economicidade, flexibilidade e eficiência. Em sua maioria, os planos de reforma administrativa contavam com a revolução tecnológica dos meios de comunicação e da informática para fazerem a gestão da informação. Com isso, foi realizado o processo de informatização da administração pública e de todas as atividades e setores sociais. Este seria o caminho natural para atender a demanda cada vez maior dos serviços públicos, sobretudo nos grandes centros.

Portanto, o Estado informatizado, suportado pelas novas tecnologias da informação e comunicação, passou a ter o mesmo poder de difusão e capilaridade da Grande Rede que, através dos serviços *online* e das interfaces digitais, pode chegar a qualquer cidadão que tenha um computador ou *smartphone* com acesso à Internet (BRAMAN, 2006a).

Contudo, Braman (2006b) também afirma que, apesar de todo o movimento em prol da transparência e crescente publicidade dos resultados da gestão pública, a relação informacional entre o Estado e os cidadãos ainda é desproporcional. Desde o Estado burocrático de bem-estar, a necessidade de conhecimento sobre as demandas dos indivíduos e grupos para a definição de políticas públicas vem aumentando, de modo que, no Estado Informacional, a urgência por informação passou a ser ainda maior devido às preocupações com o fator ‘segurança’ e à necessidade de monitoramento das fronteiras digitais (GONZÁLEZ DE GOMEZ, 2015).

Desta forma, a *assimetria* descrita inicialmente por Foucault e sublinhada por Braman (2006b) só tende a crescer com o aumento do uso da tecnologia digital e da informatização dos processos – o Estado, portanto, tem cada vez mais conhecimento sobre o cidadão, enquanto este tem cada vez menos capacidade para conhecer em profundidade o comportamento da gigante máquina estatal. A autora cita três fatores que respaldam essa afirmação: (1) os mecanismos de vigilância são de via-única - os cidadãos são monitorados e nem sequer têm a noção de estarem sendo observados; (2) os cidadãos perderam a capacidade de escolher, determinar, filtrar ou selecionar quais de suas informações estarão disponíveis para a consulta pública; (3) as informações deliberadamente selecionadas e disponibilizadas por terceiros funcionam como subsídios de sustentação do regime de vigilância total (BRAMAN, 2006a).

Outra comparação feita por Braman é em relação às diferentes formas de vigilância exercidas pelo Estado burocrático e o Estado Informacional. No primeiro, a autora cita a presença do dispositivo panóptico de Bentham, enquanto que, no segundo, surge o modelo *panspectron* onde as informações são coletadas de forma abrangente - sobre tudo e todos -, de forma indireta e constante, onde o sujeito alvo da vigilância é conhecido através de padrões e identificado por um processo de lógica inferencial aplicado a grandes volumes de dados (*big data*) (GONZÁLEZ DE GÓMEZ, 2015).

Braman (2006a) indica que quando Hookway (2000) apresentou o conceito de *panspectron*, há mais de uma década, esta ideia era tida por muitos como algo meramente teórico e especulativo, mas hoje, já se tornou realidade - considerando a variedade de dispositivos de localização e monitoramento existentes, desde os circuitos fechados de TV (CFTV) aos chips de identificação por radiofrequência (RFID[4]). Somam-se, a estes dois exemplos, toda a rede de comunicação eletrônica e o sistema digital de transações financeiras como outros dois grandes componentes deste regime de sensoriamento e rastreamento constantes ao qual estamos submetidos. Em contraposição ao modelo moderno do Panopticon, Hookway indica que no *panspectron* não há *a priori* a distinção/identificação/discriminação do sujeito para ativar o processo de vigilância sobre os corpos ou sobre os dados que os representam; as informações são coletadas de forma abrangente, sobre todos, sem o limite de tempo, por padrão. Um indivíduo é identificado somente em caso de necessidade, quando alguma questão em particular demanda o rastreamento de valores ou padrões informacionais já existentes em um banco de dados (BRAMAN, 2006b).

Outra contribuição relevante é a de Fernanda Bruno (2013) que elenca uma extensa lista de agentes, processos e tecnologias que fazem parte deste complexo cenário ativo com a capacidade de recolher informações pessoais, quando desenvolve o conceito de *vigilância distribuída*. Entre os itens estão: *webcams* pessoais, sistemas de controle de trânsito (radares, cancelas de pedágio, sistemas de controle de estacionamentos), sistemas de geolocalização, portões eletrônicos e mecanismos de autenticação (com senhas, biometrias, reconhecimento de movimento, etc.), cartões magnéticos, sistemas *online*, mecanismos de busca e navegadores, entre outros. A autora considera que pelo menos sete fatores contribuem para a caracterização do regime de vigilância distribuída, entre eles (1) seu carácter ubíquo, e

[4] Sigla para o termo em inglês *Radio-frequency* Identification. O RFID é uma tecnologia similar ao código que barra, usada para identificar objetos, que utiliza sinais eletromagnéticos para se comunicar com uma base de leitura a uma distância de até 6 metros (TECHNOVELGY).

descentralizado (sem hierarquias); (2) a diversidade tecnológica, de dispositivos e de práticas (relacionadas acima); (3) indiscernibilidade inicial sobre o foco da vigilância (todos estão sendo monitorados por padrão); (4) função potencial ou utilização secundária de dispositivos que foram desenvolvidos para outros fins que não a vigilância; (5) função operacional e de análise com a participação de agentes humanos e não-humanos; (6) utilização das redes de entretenimento, notícias e compartilhamento (redes sociais), além das redes especializadas em segurança; e (7) participação e colaboração do meio social, de forma independente e não estruturada, por parte dos indivíduos conectados à rede (BRUNO, 2013, p.29-36).

Bauman (2013) preferiu explicar esse fenômeno com uma analogia às características da liquidez, ao que chamou de *vigilância líquida.* Segundo o autor, os princípios que edificaram a "modernidade clássica" foram definitivamente postos em cheque ao final da Segunda Guerra, uma vez que os grandes objetivos do pensamento moderno já não tinham mais argumentos plausíveis que os sustentassem diante das tantas atrocidades feitas em nome do progresso. A partir disso, uma nova forma de pensar cresceu com as gerações XYZ e refletiu-se também na configuração de um novo campo social para o qual ele prefere não mais utilizar o conceito de *sociedade*. Segundo Bauman, a *rede* seria o termo mais adequado para caracterizar esse novo contexto marcado pela obsolescência programada, pelas respostas rápidas e pelo acesso a toda ordem de conteúdo com o mínimo de esforço; por outro lado, as tecnologias móveis incentivam a superficialidade, o comportamento nômade e diferentes formas de organização, relacionamento e associação. Somam-se a esse cenário (1) a cultura do consumo onde a imagem do *estar* tornou-se melhor que a essência do *ser*; (2) a indústria do entretenimento que lucra com a exploração do realismo da vida ordinária e (3) a situação de tensão constante no meio urbano, em função do risco eminente da violência, agravada pela possibilidade de ataques terroristas (BAUMAN, 2001).

Esses são fatores que contribuem para o processo de *emancipação do indivíduo*. Segundo Bauman (2001), a sociedade está testemunhando a chegada das reivindicações da teoria crítica sobre a autonomia, a liberdade de escolha e o direito à diversidade. Percebe-se que o prefixo "auto" nunca esteve tão presente quanto nesses tempos do *auto*didatismo - recursos da Internet e as plataformas de ensino à distância (EAD); do *auto*serviço - sistemas de atendimento personalizado e individualizado que permitem o usuário atender às suas próprias necessidades; e do *auto*retrato (*selfie*) - padrão fotográfico em que o indivíduo retrata a si mesmo. Os recursos de comunicação e acesso à informação, antes compartilhados ou utilizados de forma coletiva, foram aos poucos se multiplicando e ao mesmo tempo

restringindo-se ao acesso individual. A exemplo disso tem-se o computador pessoal (PC) no lugar das televisões; os *smartphones* e a telefonia móvel em substituição do telefone "fixo" residencial; os terminais de autoatendimento nos bancos e os serviços de *home banking* em lugar das grandes agências bancárias e das filas para o atendimento pessoal; e os serviços de conteúdo de filmes e séries (YouTube e Netflix) ao invés das grandes salas de cinema, teatros e plenárias em espaço público. Gonzalez de Gómez (2012) considera que estes novos dispositivos tecnológicos permeiam todos os campos da vida social e permitem aos indivíduos realizarem seus projetos de autonomia. Os antigos agentes reguladores – os gigantes dos meios de comunicação (*broadcasters*) – que estabeleceram os padrões da indústria cultural e mantiveram por muitas décadas a produção pasteurizada de conteúdo que homogeneizava a massa de audiência com base em um mínimo denominador comum, hoje foram substituídos por incontáveis canais de produção que conseguem atender de forma equânime "os universos de referência recortados e recontextualizados em suas grades discursivas, (... e) os universos valorativos e vivenciais de seus públicos" (GONZÁLEZ DE GÓMEZ, 1999, p.28).

2.4 DADOS PESSOAIS E METADADOS

Hoje, os cidadãos dos grandes centros, dependentes das tecnologias digitais, estão constantemente conectados a diversas redes, acessando inúmeras bases de dados e envolvidos por todo tipo de informação. Circuitos fechados de câmera, cartões magnéticos, publicações nas redes sociais e telefones portáteis que guardam as informações dos usuários, marcam seus passos com a precisão de horas, minutos e segundos. O mito do Big Brother e da sociedade de controle preconizada por George Orwell nunca esteve tão presente quanto no cenário atual.

Virtualmente, cada *click* ou toque em uma tela interativa, é suficiente para criar um novo conjunto de dados que podem conter muitas informações determinantes sobre o indivíduo. Deste modo começa a vigilância, o monitoramento das ações e a capacidade de prever as escolhas de pessoas ou grupos, de forma massiva. Pouco-a-pouco, de forma natural e progressiva, as pessoas estão sendo absorvidas pela tecnologia de rede e seguem caminhando para o estado de letargia em que não conseguem mais perceber os agraves do regime de vigilância de massa no qual elas estão inseridas. Aos poucos, o usuário das mídias digitais e novas tecnologias de comunicação e informação (nTIC) segue aceitando (e colaborando) com a invasão da sua privacidade ou talvez entregando a sua liberdade em troca

de uma falsa sensação de segurança. Quem sabe ainda, transferindo seus dados pessoais para domínios públicos e oferecendo, para uma rede de audiência efêmera, suas preferências, e até mesmo os bastidores da sua vida mais íntima (PANOPTYKON FUNDATION, 2015).

Wood et al. (2006, p. 04), definem *vigilância* como "a observação de informações pessoais, de forma proposital, rotineira e sistemática com objetivos de controle, direitos, gestão, influencia ou proteção". Já Andrejevic e Gates (2014) contribuem com algumas distinções entre a vigilância e a vigilância de massa, considerando que esta última utiliza grandes bases de dados, onde o "alvo" da observação não são pessoas específicas, mas os padrões de comportamento gerais que podem ser percebidos e reconhecidos. Contudo, a partir da identificação de certos padrões classificados como nocivos à sociedade e à segurança pública, os órgãos de vigilância passam a fazer uma observação mais acurada dos dados gerados pelo indivíduo suspeito. Desta forma, a vigilância de massa trabalha com duas instâncias de material para análise: os dados ou conteúdo propriamente dito, gerados pelos usuários e os metadados gerados pelos sistemas.

Metadados, segundo o IBGE, são "dados que descrevem os dados", ou "informações úteis para identificar, localizar, compreender e gerenciar os dados". Assim, é possível fazer uma série de filtros em uma grande base de dados e selecionar melhor a amostra que se deseja tratar e observar (INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA). O prefixo 'meta' sugere definição ou descrição, por isso tendem a resumir a informação básica a respeito de um dado. Os metadados são, portanto, informações complementares relacionadas ao conteúdo principal produzido. Através deles, tornam-se muito mais fáceis as tarefas de localização, recuperação e uso de uma instancia de informação já armazenada, sobretudo se o conteúdo em questão não estiver na forma de texto (ex: vídeo, imagens, objetos, etc.) (ROUSE, 2014).

Usando a rede de microblog Twitter[5] como exemplo, tem-se mensagens de, no máximo, 140 caracteres, o que poderia não significar muito para um analista de *Big Data*, se não fossem os mais de 30 campos de metadados relacionados a cada *tweet*. Com eles, pode-se saber muito mais sobre o emissor da mensagem, por exemplo: nome do usuário, data e hora da publicação, geo-localização, IP da máquina, *hashtags* utilizadas, etc. (KIRKPATRICK, 2011).

[5] Rede social que permite aos usuários enviar e receber atualizações pessoais de outros contatos, em textos de até 140 caracteres. Os textos são conhecidos como *tweets*, e podem ser enviados por meio do *website* do serviço, por SMS, por aplicativos para celulares, entre outros meios.

Esses dados podem ser usados pelo Twitter para formular estatísticas de acesso e uso com fins comerciais ou de utilidade pública. Sabendo qual é a opinião da sua audiência e onde ela está localizada, o Twitter pode vender inserções direcionadas para qualquer público específico. Além disso, seus dados podem servir para o mapeamento de situações de emergência, mapeamento de epidemias, verificação de situações de trânsito ou ocorrências em regiões específicas. Lembrando a filosofia de trabalho dos analistas de *Big Data*, todo padrão identificado ou reconhecido pode ser investigado de forma mais acurada.

Pasquinelli traz um conceito diferente para 'metadados', com um teor relacionado a ideia de *mais valia* proposta por Marx. Na concepção do autor, 'metadados' pode ser logicamente definido como "uma mensuração da informação, a computação de sua dimensão e a sua transformação em valor" (PASQUINELLI, 2015, p.63, tradução nossa). Em um contexto econômico e social com a ascensão do Big Data, os metadados se tornam fundamentais por permitirem: 1) a mensuração do valor das relações sociais; 2) a melhoria dos processos e da inteligência artificial; e 3) o monitoramento e previsão do comportamento.

Com o advento das redes sociais, é possível conhecer a relação e o grau de ligação entre seus usuários, assim como mensurar o fluxo de informação entre eles. Os metadados também podem servir para a melhoria dos processos produtivos, principalmente daqueles ligados diretamente à gestão e análise da informação. Mas, a função dos metadados que mais tem chamado a atenção dos gestores de negócio é a capacidade de fazer amplas leituras sobre o comportamento de um grupo ou população, ressaltando conclusões de alto nível que podem servir para a previsão de comportamento futuro.

Segundo Richards e King (2014) as leis que garantem a privacidade, confidencialidade e transparência devem abranger a proteção dos metadados, pois estes oferecem um meio mais fácil e ainda mais relevante para as operações de controle e vigilância. Em algumas ocasiões estas referências podem ser tão expressivas quanto os dados pessoais e dados sensíveis na identificação de um indivíduo.

A segurança dos dados pessoais e a preservação da identidade são direitos básicos de todo cidadão, garantidos por Lei. A Constituição Nacional, em seu artigo 5°, incisos X, XI e XII, preserva a intimidade, a vida privada, a honra e a imagem das pessoas; e qualifica como inviolável o sigilo das comunicações pessoais (BRASIL, 1988). A Declaração Internacional dos Direitos do Homem, em seu artigo 12, prescreve que "ninguém será sujeito a interferências em sua vida privada, em sua família, em seu lar ou em sua correspondência, nem a ataques à sua honra e reputação" (ONU, 1948).

Contudo, empresas de tecnologia e os próprios governos passaram a usar os dados pessoais e dados sensíveis dos clientes e cidadãos como mercadoria. Hoje em dia, parece que tudo é passível de ser negociado; basta a oferta de uma quantidade razoável de dinheiro ou qualquer premissa de segurança ou força maior, quase sempre, não justificada ou esclarecida de forma apropriada (ANDREWS; LINDEMAN, 2013).

Segundo o texto do anteprojeto de Lei de proteção de dados pessoais no Brasil, "dados pessoais" são aqueles pelos quais um indivíduo pode ser identificado (PENSANDO O DIREITO). Isso inclui informações próprias e intransferíveis (como nome completo e número de identidade) ou descrições específicas que em conjunto possam localizar o indivíduo dentro de um contexto. Já os "dados sensíveis", são aqueles que podem ensejar a discriminação do seu titular por se referirem, por exemplo, à opção sexual, convicções religiosas, filosóficas ou morais, e opiniões políticas. Este anteprojeto de Lei visa proteger os direitos fundamentais de liberdade, intimidade e privacidade, garantindo os direitos do cidadão brasileiro sobre suas informações. Neste contexto, a palavra "consentimento" é fundamental. Nada pode ser feito sem a anuência do titular dos dados, lembrando que qualquer prestadora de serviços - operadora de banco de dados – está, na verdade, administrando a propriedade de terceiros e não detém direitos sobre os dados que opera.

Segundo Goulart e Serafim (2015) os termos de política de privacidade, em geral, não são claros o suficiente. Normalmente são extensos e escritos com linguagem que não facilita o entendimento do usuário sobre as regras de uso. Além disso, a referência aos dados pessoais normalmente não se encontram evidentes no texto. Como exemplo da falta de assertividade e clareza nos termos de uso, segue um trecho retirado da política de privacidade do Windows 8:

> A Microsoft pode acessar ou divulgar informações sobre você, incluindo o conteúdo de suas comunicações, para: (a) cumprir a lei ou responder a solicitações legítimas ou processos judiciais; (b) proteger os direitos ou a propriedade da Microsoft ou dos nossos clientes, incluindo a aplicação dos nossos contratos ou políticas que regem o uso dos softwares, ou (c) adotar providências quando acreditarmos, de boa fé, que esse acesso ou divulgação seja necessário para proteger a segurança pessoal dos funcionários da Microsoft, dos clientes ou do público em geral (Microsoft, 2012).

O Marco Civil da Internet (Lei 12.965/14), que está em vigor desde 26 de junho de 2014, estabelece princípios, garantias, direitos e deveres para o uso da Internet no Brasil. No que diz respeito à segurança dos dados dos usuários, a Lei garante a proteção aos dados pessoais e determina que as ações das empresas de Internet sejam mais transparentes. Outro

avanço está nas garantias da liberdade de expressão e privacidade das comunicações onde se busca o meio digital como amplo espaço de debate democrático, aberto e livre, ao mesmo tempo em que os e-mails passam a receber o mesmo tratamento sigiloso dos meios de comunicação tradicionais (CULTURA DIGITAL; GOULART e SERAFIM, 2014a). Outra característica vista hoje em dia nas redes sociais da Internet é a regulação do conteúdo publicado a partir da própria comunidade de usuários, que pode ser considerado uma modalidade de *sousveillance* onde "todos vigiam todos". Os principais sistemas de interação e organização de redes sociais, contam com mecanismos de denúncias de material impróprio, *spams*[6] ou violações de direitos autorais (*copyright*).

[6] Termo em inglês que designa uma mensagem eletrônica recebida, mas não solicitada pelo usuário.

3 GERAÇÃO E CAPTAÇÃO DE DADOS PESSOAIS

Seguindo o fluxo natural da gestão da informação, primeiramente ocorrem a geração ou captação do material informacional a ser processado. Neste sentido, torna-se relevante a exposição de alguns eventos históricos e marcos tecnológicos que contribuíram para a configuração do regime atual da coleta massiva de dados pessoais.

3.1 A EXPLOSÃO DA INFORMAÇÃO: DO MEMEX AO BIG DATA

Quando Fremont Rider publicou *The Scholar and the Future of the Research Library*[7], em 1944, era estimado que as bibliotecas das universidades americanas dobrassem de tamanho a cada 16 anos, e que ainda, se essa tendência permanecesse, as mesmas chegariam em 2020 com mais 200 milhões de volumes impressos (PRESS, 2013). Este talvez tenha sido o primeiro alerta sobre os futuros problemas de armazenamento e recuperação da informação.

No ano seguinte, quando a 2ª Guerra Mundial já estava em vias de terminar, Vannevar Bush publicou o ensaio *As We May Think*[8] considerando que os cientistas, ao serem liberados das tarefas de guerra, deveriam voltar seus esforços para o desenvolvimento de soluções de gestão da informação, para criar o mais vasto arquivo de informação recuperável da História (STROTHER et al., 2012). Neste mesmo ensaio Bush usou a expressão "explosão da informação" que desde então tornou-se uma referência histórica.

Contudo, por mais surpreendentes que pudessem ser as previsões de Rider e Bush, as décadas seguintes mostrariam que a produção de informação seria maior que todas as previsões daquela época, mesmo que fosse contabilizada somente a produção de livros (excluindo outras formas de conteúdo).

Apesar de terem feito estimativas bem respaldadas, Rider e Bush não contavam com o advento dos computadores pessoais e de muitos outros dispositivos eletrônicos que serviriam para manter os usuários da tecnologia conectados em rede 24 por dia, com a possibilidade de produção, publicação e compartilhamento de todo tipo de conteúdo, em tempo real.

---

[7] Em português: "A Academia e o Futuro das Bibliotecas para Pesquisa" – tradução nossa.

[8] Ensaio de Vannevar Bush publicado em 1945 na revista *The Atlantic*, em que o autor faz concepções a respeito de uma máquina (servidor de conteúdo) cujo comportamento seria semelhante ao cérebro humano na recuperação de informações.

Em 2000, Peter Lyman e Hal Varian – dois professores da Universidade da Califórnia – realizaram um estudo para estimar a quantidade de informação produzida anualmente em escala global. A produção de conteúdo foi estimada em 1,5 exabytes[9] por ano. A pesquisa serviu ainda para afirmarem que o disco magnético (*hard disc*) viria a ter sua capacidade de registro dobrada a cada ano e que, em pouco tempo, o mesmo seria adotado internacionalmente como a tecnologia padrão para o armazenamento de dados (LYMAN; VARIAN, 2000).

Já em 2010, Kenneth Cukier publica na revista *The Economist* uma reportagem especial intitulada *Data, data everywhere*[10]. Nesta matéria, o autor revela fatos impressionantes sobre a geração de dados a partir de instrumentos de segurança ou voltados para a Ciência, a exemplo do telescópio do Sloan Digital Sky Survey (SDSS) que, nas suas primeiras semanas de operação, registrou mais dados que toda a história anterior da Astronomia. Cukier (2010) ainda considera que

> [...] o mundo comporta uma quantidade inimaginável de informação digital que ainda está expandindo com vasta rapidez (...). Este efeito está sendo percebido em todas as áreas: dos negócios à ciência, dos governos às artes. Para este fenômeno os cientistas e os engenheiros da computação criaram a expressão: **Big Data** (CUKIER, 2010 - tradução nossa, grifo nosso).

Outro caso citado foi o do Walmart, gigante do varejo que operava mais de 1 milhão de transações de clientes por hora com mais de 2,5 petabytes[11] de dados – o equivalente a 167 vezes a quantidade de conteúdo em livros da Biblioteca do Congresso Americano (CUKIER, 2010). Hoje a Biblioteca do Congresso Americano possui mais de 37 milhões de livros e outros materiais impressos (LIBRARY OF CONGRESS). Segundo o *site* Compare Business Products[12], este acervo é considerado o 10ª maior do mundo em termos de quantidade de informação armazenada, atrás de outras como as das empresas Amazon, YouTube e AT&T.

A primeira vez que o termo *Big Data* apareceu em uma publicação foi em 1997 quando os engenheiros da NASA Michael Cox e David Ellsworth relataram a impossibilidade de aproveitamento de um conjunto de dados, por este exceder à capacidade de memória e processamento dos computadores da época (COX e ELLSWORTH, 1997). Desde então, esta

[9] Um exabyte equivale a um bilhão de gigabytes, ou 1000^7 bytes.
[10] Em português: "Dados em todo lugar" – tradução nossa.
[11] Um petabyte equivale a 1000^5 bytes.
[10] Na lista apresentada pelo *site* não foi mencionado o *data center* da NASA e não foi informado o método de cálculo utilizado para estimar a capacidade de armazenamento das bases de dados citadas no ranking, contudo, a lista é válida como uma noção aproximada da realidade.

questão vem se tornando cada vez mais frequente. Empresas e governos vêm enfrentando o dilema da automação e do aumento da conectividade às redes, sem conseguirem efetivamente administrar a sobrecarga de dados gerada a partir dos usuários, clientes, parceiros, máquinas e todo tipo de dispositivo digital. Dumbill (2014) considera que um *Big Data* surge a partir dos 3 Vs: quando o *volume*, a *velocidade* e a *variedade* dos dados excede a capacidade de memória e processamento dos computadores, tornando seu aproveitamento algo extremamente caro e complexo.

Boyd e Crawford, no artigo Critical Questions for Big Data[13] publicado no Information, Comunication and Society Journal, definem *Big Data* como

> um fenômeno cultural, tecnológico e acadêmico baseado na interação de três fatores: (1) Tecnologia: maximização da precisão dos algoritmos e do poder de computação para reunir, analisar, relacionar e comparar grandes conjuntos de dados; (2) Análise: processamento de grandes conjuntos de dados para identificar padrões para atender às necessidades de ordem econômica, social, técnica e legal; e (3) Mitologia: a ampla crença de que grandes conjuntos de dados possibilitam uma forma mais avançada de inteligência e conhecimento que podem gerar *insights* até então impossíveis de se alcançar, de forma objetiva e confiável (BOYD; CRAWFORD, 2012, p. 02, tradução nossa).

A grande questão que move as empresas e os governos para o *Big Data* é a (nova) possibilidade de encontrar relações entre pontos distantes de um sistema complexo e sem necessariamente entender suas causas, podendo fazer previsões para o futuro. Os *Big Data* - esses conjuntos de dados massivos que advêm de situações reais - são extremamente complexos e impossíveis de serem interpretados pela mente humana, fato que trouxe novos e poderosos métodos de análise a partir de algoritmos e uso de inteligência artificial por meio da aprendizagem de máquina (*machine learning*). Os dados gerados a respeito da população das grandes cidades, por exemplo, têm demandado investimento em novas tecnologias que permitam seu processamento e aproveitamento para a melhoria de serviços públicos como a erradicação de epidemias, mapeamento de ações criminosas e organização do trânsito, entre outros benefícios (ADREJEVIC; GATES, 2014).

Para os propósitos desta pesquisa, portanto, a noção de *Big Data* abrange: (1) a questão dos massivos conjuntos de dados sem precedentes na História, (2) a nova capacidade técnica de processamento e análise desenvolvida para o aproveitamento desses dados em

[13] Em português: "Perguntas fundamentais para o Big Data" – tradução nossa.

tempo real e (3) a mudança de cultura das empresas e governos a partir da possibilidade de previsão do comportamento de sistemas complexos.

3.2 A GERAÇÃO DE DADOS NO CAMPO SOCIAL ATRAVÉS DE SERES HUMANOS E NÃO HUMANOS

Considerando principalmente o processo de informatização para a automação de sistemas no campo governamental e privado, percebe-se um crescimento substancial na geração de dados digitais a partir da década de 2000. Alguns fatores como a popularização dos computadores pessoais e a ampliação da área de cobertura dos serviços de acesso à Internet podem ser citados como causa para o surgimento da *web* social ou Web 2.0.

De acordo com a explanação de Tim O´Reilly, a Web 2.0 é um conceito que define o funcionamento da Internet a partir do "estouro da bolha" no ano 2000, quando passou a ser uma plataforma e contar com a colaboração efetiva dos usuários na publicação, revisão e classificação de conteúdo. Nesta segunda fase, a *web* passou a ser de fato interativa, dando mais poder de publicação para os usuários e permitindo maior interação dos usuários entre si (o que significa o embrião das redes sociais como Twitter, Instagram, YouTube e Facebook, entre tantas outras).

Além dos serviços disponíveis nos *sites*, surgiram os *blogs*[14], os *wikis*[15], as *tags*[16] e também diversos aplicativos (como o pioneiro Napster, por exemplo) que permitiram a organização de múltiplas redes de compartilhamento de arquivos (O´REILLY, 2005; GOVERNOR, NICKULL, HINCHCLIFFE, 2009). Deste modo, a tecnologia não estava conectando mais um usuário da Rede apenas a um servidor de conteúdo, mas também a outros usuários (P2P)[17]. A noção de como a Web funciona através deste novo padrão colaborativo é de grande importância para entender como a tecnologia e a infraestrutura da Rede - hoje com o amplo acesso sem fio (Wi-Fi) - serviu como ferramenta de organização, mudando radicalmente a forma como as pessoas trabalham e gerenciam suas informações.

[14] Páginas pessoais, de acesso público, que refletem a opinião ou transmitem informações sobre o autor.

[15] *Websites* que servem como ferramentas de colaboração onde uma comunidade de usuários pode trabalhar em autoria conjunta.

[16] Palavras-chave que identificam (indexam) um conteúdo digital (texto, áudio ou vídeo) facilitando sua recuperação pelos mecanismos de busca.

[17] Acrônimo do termo em inglês *Peer to Peer*, que se refere a uma rede formada apenas por computadores clientes, que se conectam entre si sem a dependência de um servidor central (TECHTERMS).

A partir disto, cada vez mais, as pessoas estão informatizando suas ações e adotando novas ferramentas para ampliar suas capacidades de controlar as informações que elas criam e têm que lidar todos os dias. Os dados estão em toda parte: nas relações dos cidadãos com o Governo, na interação dos usuários com as suas contas de serviços, nos negócios e até mesmo atrelados aos corpos dos indivíduos. Em suma, como disse O´Reilly, "enquanto ainda não afundamos em um oceano de dados, estamos percebendo que quase tudo pode (ou deve) ser instrumentado" (O´Reilly, 2011, p. 06).

O fenômeno do *Big Data* pode ser considerado a nova "explosão" da informação. Até a popularização da Internet, a produção de conteúdo estava nas mãos dos grandes grupos de comunicação e das tradicionais editoras de mídia impressa e televisão. Após o processo de popularização dos computadores pessoais, a publicação de conteúdo passou a não ser mais uma exclusividade das grandes empresas (*broadcasters*). Em julho de 2014, o Nielsen Group divulgou o resultado de uma pesquisa[18] informando que no Brasil há mais de 120 milhões de pessoas com acesso à Internet (NIELSEN GROUP, 2014).

A Web 2.0 também diz respeito ao novo *locus* de armazenamento e processamento de dados, a "Nuvem" - servidores externos, que podem ser acessados de qualquer ponto da rede internacional de computadores. Um exemplo deste recurso são as ferramentas *online* do Google Docs – softwares que funcionam a partir do navegador, sem estarem instalados na máquina do usuário e que salvam seus dados em um servidor da Rede, não utilizando o disco rígido da máquina local (GRIFFITH, 2015).

Mas a formação do *Big Data* não se dá apenas através dos computadores. Seguindo a lei de Moore[19], existe uma relação inversamente proporcional entre a redução de tamanho dos processadores e o aumento das suas capacidades de processamento, a ponto de permitirem todo tipo de tecnologia portátil e acessórios funcionais (*wearable technology*) (THE GUARDIAN, 2015). A exemplo disso, estão os *smartphones*, *smart watches*, sensores óticos, GPS, controladores de temperatura e até mesmo marca-passos (REUTERS, 2009), entre outros utilitários para lazer, negócios, esporte, saúde e segurança, capazes de aferir as condições do ambiente e fazer o *upload* dos dados, em tempo real. Conforme O´Reilly,

[18] A pesquisa considerou pessoas de 2 a 15 anos de idade com acesso à internet por meio de computadores em domicílios e considerou pessoas de 16 anos ou mais com acesso à internet por meio de computadores em qualquer ambiente (domicílios, trabalho, *lan houses*, escolas, igrejas e outros).

[19] Lei proferida por Gordon E. Moore, na qual o número de transistores dos chips teria um aumento de 100%, pelo mesmo custo, a cada período de 18 meses.

> [...] as pessoas estão passando cada vez mais tempo *online*, e deixando um rastro de dados por onde quer que elas passem. Os aplicativos móveis deixam um rastro de dados ainda mais rico, uma vez que eles são dotados de geolocalização ou comportam vídeos e áudios. Tudo isso pode ser analisado (O´Reilly, 2011, p. 07 - tradução nossa).

Conforme Licklider (1960), através de diversos dispositivos eletrônicos (muitos deles acoplados ao próprio corpo) é possível considerar que os indivíduos já vivem em simbiose com ferramentas que ampliam as suas capacidades de gerenciamento da informação[20]. Ações, movimentos, escolhas e interferências feitas no campo real (tangível) de fato passaram a ser convertidos, em tempo real, para o campo digital, o que nos traz um novo conceito de representação da realidade. Assim, está sendo criado um "*backup*" do mundo real, lotando bancos de dados inteiros com parâmetros que representam os diversos contextos sociais e criam os perfis como cidadãos, profissionais, pesquisadores, consumidores, etc. (WOLF, 2015).

A geração massiva de dados também não está restrita à ação humana ou aos serviços que demandam a transferência e processamento de dados pessoais.

Hoje, o conceito de Internet das Coisas (IoT – *Internet of Things*) diz respeito às máquinas ou equipamentos - desde carros, eletrodomésticos e tecnologias de segurança - que podem monitorar suas próprias atividades ou as do ambiente à sua volta, e guardar ou transmitir dados através da rede, com ou sem fio (STROUD, 2015; GOULART e SERAFIM, 2014b).

Segundo Burrus (2014), a Internet das Coisas é muito maior do que a maioria das pessoas pode imaginar, pois as possibilidades de interação entre as máquinas e o ambiente, e das máquinas entre si, são praticamente infinitas. Este padrão já está sendo testado por muitos adotantes primários da tecnologia no campo da saúde, da segurança, da eficiência industrial e energética e, em breve, estará presente na maioria das casas (*smart houses*) e sistemas integrados das cidades inteligentes (*smart cities*).

[20] Licklider (1960) desenvolveu estudos para compreender a relação homem-máquina em seu processo de simbiose para aumentar a capacidade humana no processamento de informações e síntese dos múltiplos aspectos de uma situação complexa para a tomada de decisão.

4 TECNOLOGIAS PARA O ARMAZENAMENTO, TRATAMENTO E ANÁLISE DE DADOS

Para o maior entendimento sobre o processo de vigilância de massa que ocorre nos dias atuais, devido à plataforma tecnológica instalada, onde a observação direta dos corpos foi superada pela análise dos dados, de forma remota, invisível e em larga escala (*dataveillance*), é necessário, primeiramente, mostrar como a tecnologia, os processos e métodos voltados para a gestão da informação evoluíram ao longo do tempo e como a informação, enquanto recurso, foi progressivamente sendo armazenada e estruturada.

Uma das primeiras definições de Gestão de Recursos da Informação (GRI) encontradas na literatura é a formulada por Horton (1979, p. 99, tradução nossa): "um modelo de trabalho que visa o gerenciamento dos recursos de dados, de forma ordenada e sistemática". Ao longo do tempo, este conceito recebeu a contribuição de outros autores e teve seu escopo e abrangência atualizados de acordo com as necessidades organizacionais e as tecnologias disponíveis em cada período.

Davenport (2002, p. 27) contribui para o entendimento da evolução das práticas de gestão da informação trazendo uma abordagem com foco na estruturação da informação a partir dos meios de fixação, registro ou armazenamento. Através da história da base, mídia, ou suporte para o registro, é possível perceber as vias tecnológicas que auxiliaram o processo de estruturação e conversão da informação, do estado físico para o digital, trazendo novas possibilidades para sua gestão em larga escala (chegando-se a atual vigilância e controle de massa).

Abaixo são listadas as quatro etapas definidas pelo autor em ordem cronológica, das quais, em função do foco temático deste trabalho, serão destacadas apenas as três primeiras. É importante mencionar também que, apesar de haver evolução nos métodos de gestão e das formas de estruturação da informação, as abordagens ou tipos citados não são necessariamente sequenciais e podem coexistir em um mesmo período e contexto.

1. Informação não-estruturada;
2. Informação estruturada em papel;
3. Informação estruturada em computadores;
4. Capital intelectual (conhecimento).

A *informação não-estruturada* é a mais antiga das abordagens e permanece, ainda hoje, em um estado bem próximo do que era nos primórdios da escrita. Mesmo com tantos avanços ocorridos no campo da tecnologia da informação, esta modalidade é, sem dúvida, a que ainda oferece os maiores problemas para o controle, mensuração e valoração da informação.

A *informação estruturada em papel* é a forma mais comum de abordagem – considerando toda história da gestão da informação –, e abrange registros, documentos, livros e todo tipo de material impresso. Um ponto patente desta vertente é a facilidade para o controle físico da informação, o que torna possível, e mais evidente, seu ciclo de vida e seus fluxos, desde a criação (ou reprodução) até sua guarda e disseminação.

A *informação estruturada em computadores* tornou-se, a partir da década de 1960, o grande foco dos profissionais de TI e, ao mesmo tempo, a esperança de solução para os problemas da *overdose* de informação. Viu-se então a possibilidade de otimizar os recursos de informação evitando o desperdício de papel e material de impressão, fazendo o uso de dados compartilhados. Para tal, foram necessários a desvinculação dos dados das estações de trabalho e o desenvolvimento de redes e servidores, possibilitando o múltiplo acesso a um documento, sem a necessidade de redundância.

Marchand (1985 apud SAVIC, 1992) apresenta sua visão de como a Gestão da Informação (GI) teria evoluído ao longo dos anos, através de quatro estágios. O trabalho deste autor se mostra relevante por citar os principais recursos tecnológicos disponíveis em cada período histórico.

O primeiro deles seria o *controle físico da informação*, de origem no início do Séc. XX e com ênfase na gestão de arquivos físicos, documentos impressos, arquivos, relatórios e correspondências. As principais ferramentas utilizadas nesse tempo eram o papel, os gabinetes de arquivo, a máquina de escrever, o telefone e o microfilme. O *design* de escritório também tinha grande relevância para o ganho de eficiência sobre o controle físico da informação.

O segundo estágio seria a *gestão da tecnologia de automação*, que se estendeu da década de 1960 até meados de 1970. A tecnologia da época contava com os computadores de segunda e terceira geração, as máquinas de reprografia, os processadores de texto e os primeiros equipamentos de transmissão de voz.

Ainda de acordo com Marchand (1985 apud SAVIC, 1992), o terceiro estágio foi o da *gestão dos recursos de informação* que perdurou de meados da década de 1970 até o final dos anos 1980. O amadurecimento na visão da gestão da informação trouxe à questão o

envolvimento de executivos de maior ascensão dentro das organizações, cujas funções incluíam o planejamento estratégico e a integração da gestão dos recursos tecnológicos para a informação. Nesta época já eram utilizados computadores pessoais (PC) e as estações multifuncionais de trabalho.

Por fim, como quarta etapa de evolução do conceito e prática da gestão da informação, tem-se o que Karl Wiig chamou de *gestão do conhecimento*[21], com base em sistemas de suporte à decisão e de inteligência corporativa (CIANCONI, 2003). Segundo Marchand (1985 apud SAVIC, 1992), a característica mais evidente deste período é a crescente dependência das organizações, em todos seus níveis, da tecnologia da informação.

Em 1986, Marchand, em parceria com Horton, atualizou seu modelo inicial, redefinindo a nomenclatura de algumas etapas e acrescentando a *Gestão estratégica da informação*. As etapas elencadas a seguir – embora suas nomenclaturas possam enfatizar suas aplicações no mercado corporativo – são adotadas neste trabalho como a síntese do desenvolvimento da gestão da informação, de forma geral, ao longo das últimas sete décadas (MARCHAND; HORTON, 1986 apud SAVIC, 1992, p. 134, tradução nossa):

1) Gestão de documentação impressa;
2) Gestão da tecnologia de automação;
3) Gestão dos recursos de informação da organização;
4) Inteligência e análise da competitividade em negócios;
5) Gestão estratégica da informação.

Savic (1992) lança um olhar sobre as práticas de GRI, que tiveram suas origens a partir da confluência de três fatores: a explosão da informação, a proliferação do papel e o uso extensivo das tecnologias de tratamento da informação.

Um fenômeno ocorrido após a segunda guerra mundial foi a *explosão da informação* ou, como chamado também por alguns autores, "poluição da informação". A expressão se dá pela grande quantidade de informação gerada na época e pelo surgimento de novos tipos de mídia para o armazenamento de dados, publicação de conteúdo e transferência de informação a exemplo de livros, jornais, relatórios, CD-ROMs e correspondências diversas. Dentre as tecnologias de comunicação havia o rádio, a telefonia, a televisão, a rede científica e militar de computadores e os serviços de informação e satélites. Contudo, o que a princípio deveria ser uma solução para as necessidades de informação dos indivíduos, empresas e governos, se

[21] O conceito de gestão do conhecimento, em especial com Nonaka e Takeuchi (1997), passa a se referir ao processo de produção de conhecimento (tácito e explícito), distinguindo-se da gestão da informação que se refere ao conhecimento registrado / explícito.

converteria em um novo problema, pois toda informação produzida não poderia ser recuperada se antes não houvesse o seu devido tratamento.

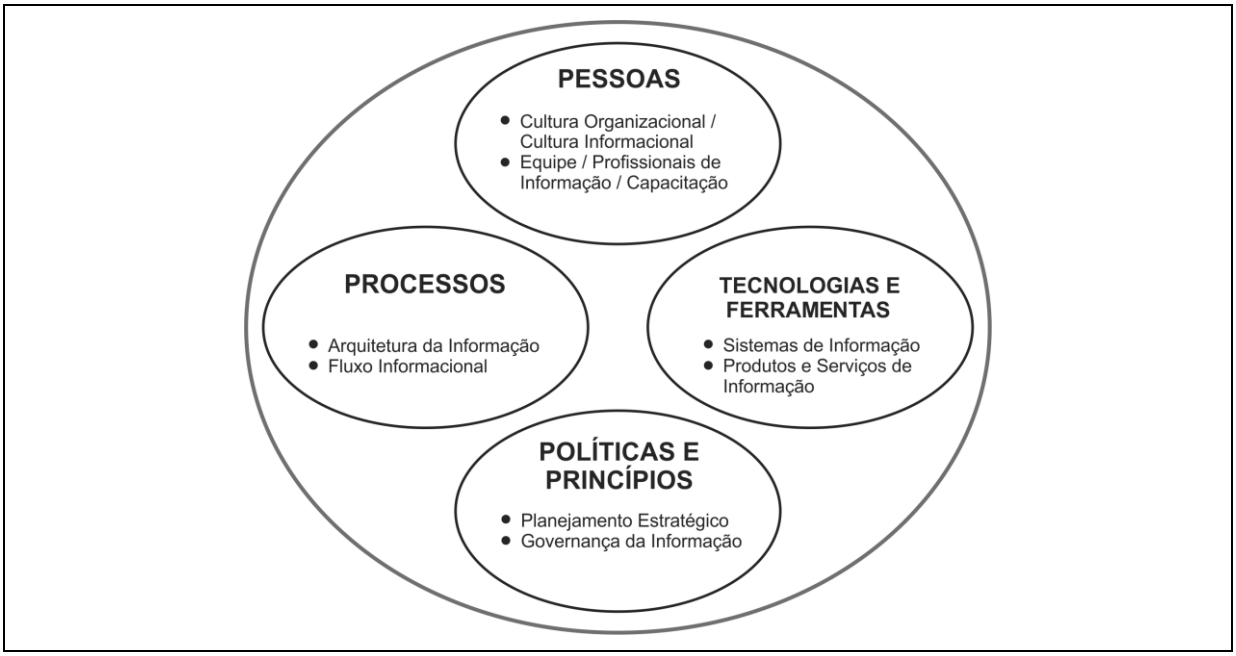
Em segundo lugar, e ainda como consequência direta da explosão da informação, houve a proliferação do papel. Para uma rápida noção sobre o tamanho da dependência desta matéria prima, podem ser citadas três referências estatísticas sobre o ano de 1990, quando a Unesco estimou que o consumo *per capta* de papel nos Estados Unidos era de 83,2 Kg (por ano) e a American Paper Institute calculou que a produção de papel, nos Estados Unidos, dobrou de 1980 a 1990, chegando a duzentos e quarenta milhões de toneladas por ano. Sobre esta mesma década, Barber (1990 apud *op. cit.*) apontou que noventa e cinco por cento dos registros da época ainda era feito em papel. Tudo isso levou os gestores da época à conclusão de que algo deveria ser mudado para garantir a sustentabilidade nas atividades das empresas privadas e dos governos.

O terceiro fator que contribuiu para o desenvolvimento da GRI foi o *uso extensivo das tecnologias de tratamento da informação*. De fato, as tecnologias da informação trouxeram mudanças radicais e irreversíveis para a sociedade pós-industrial, em todos seus aspectos. A automatização de processos, a digitalização de dados e o aumento da capacidade de processamento dos computadores mudaram radicalmente as formas de gestão nos governos, nos negócios e na agricultura, e ao mesmo tempo, as formas de comunicação, relacionamento e consumo de cultura na sociedade.

Com o passar dos anos, o computador pessoal tornou-se mais acessível e foi adotado por grande parcela da sociedade, levando para milhares de famílias e indivíduos os benefícios da organização de arquivos e armazenamento seguro de dados. A popularização dos computadores pessoais e, mais tarde, dos *smartphones* foi um dos principais fatores que permitiram o monitoramento de dados e informações pessoais em larga escala.

Com base nas considerações dos autores supracitados, Martins e Cianconi (2013) elaboraram uma figura que explicita os conjuntos de recursos contidos no modelo Gestão dos Recursos da Informação, em sua visão integrativa:

Figura 3 - Recursos incluídos no modelo de gestão integrada da informação



Fonte: MARTINS; CIANCONI, 2013, pg. 4.

A partir da figura acima, é possível perceber a associação dos vários recursos abarcados pela gestão da informação, em diferentes níveis. Embora seus componentes não possam ser dissociados, para fins desta pesquisa, foi enfatizado o aspecto tecnológico das mediações e práticas de (1) captação, (2) tratamento, processamento e análise, e (3) respostas, em função da vigilância e controle, referenciando os demais aspectos socioculturais (pessoas), legais (políticas e princípios) e funcionais (processos), sempre que pertinente para a contextualização e embasamento das questões abordadas.

Além da tecnologia e dos métodos que visam o controle da informação sob seus aspectos físicos, evoluíram também os princípios de Tratamento Temático, e as técnicas de indexação para a organização e gestão da informação em seu aspecto semântico.

4.1 TRATAMENTO TEMÁTICO DA INFORMAÇÃO

Conforme afirmado anteriormente, o excesso de informação pode ser tão prejudicial quanto a falta dela. Essa comparação se dá pela premissa de que toda informação, para ser útil, deve receber tratamento adequado e ser armazenada segundo critérios lógicos e consistentes que garantam a sua futura recuperação.

“Indexação e resumos tem suas origens em um ponto do tempo em que alguém percebeu que os registros escritos deveriam ser organizados” – assim Cleveland, D. e Cleveland, A. (2013, p. 33, tradução nossa) começam contando a história dessas atividades fundamentais para o Tratamento Temático da Informação (TTI) consideradas por muitos autores tarefas de fundamental importância para Ciência da Informação (BORKO, 1977; GUIMARÃES, 2008).

Desde a invenção da escrita, percebe-se que a humanidade vem buscando novos meios para recuperar os registros de suas atividades, em um caminho que visa, não só à preservação de suas memórias, mas também à perpetuação de seu legado cultural. A Idade Média, neste sentido, foi um período histórico bastante profícuo diante das técnicas de indexação e recuperação da informação. Com o crescimento do estudo erudito, que culminou na invenção da imprensa e na publicação de um volume considerável de livros, surgiram muitos recursos permitiram maior controle no uso da informação, tornando-a também mais acessível, dentre eles o índice de final de livro, a lista alfabética de autores e o periódico de resumos (BEARE, 2007 apud *op.cit.*, p.36).

Em período mais recente – para evitar redundâncias na descrição dos adventos tecnológicos inerentes a cada período históricos – no século XX, após a Segunda Guerra Mundial, o computador, com todas as capacidades e funções de processamento utilizados até então para os cálculos logísticos e precisão cibernética da artilharia, foi considerado por Cleveland, D. e Cleveland, A. (2013) o mais importante advento tecnológico aplicado à indexação. Os autores destacam a contribuição de um pioneiro da Ciência da Informação e notável inventor da IBM – International Business Machine, Hans Peter Luhn pelo desenvolvimento de um dos primeiros algoritmos para o reconhecimento de padrões. A tecnologia de Luhn, conhecida como KWIC (*key word in context*), era capaz de aumentar o reconhecimento de palavras em seus contextos, reduzindo sensivelmente os erros de ambiguidade. Este trabalho seria o precursor de muitos outros no campo da indexação automática, reconhecimento de padrões e de interpretações com caráter de inteligência artificial que seguem até hoje.

Com referência em Lancaster (2004, p.1) e Mai (2005, p. 599), pode-se dizer que o principal propósito da indexação e resumos é a *representação do objeto* informacional original, de forma sintética e objetiva, que possa ser inserida em uma base de dados para possibilitar sua posterior localização e recuperação. Este processo é composto por duas etapas principais: Análise conceitual e Tradução.

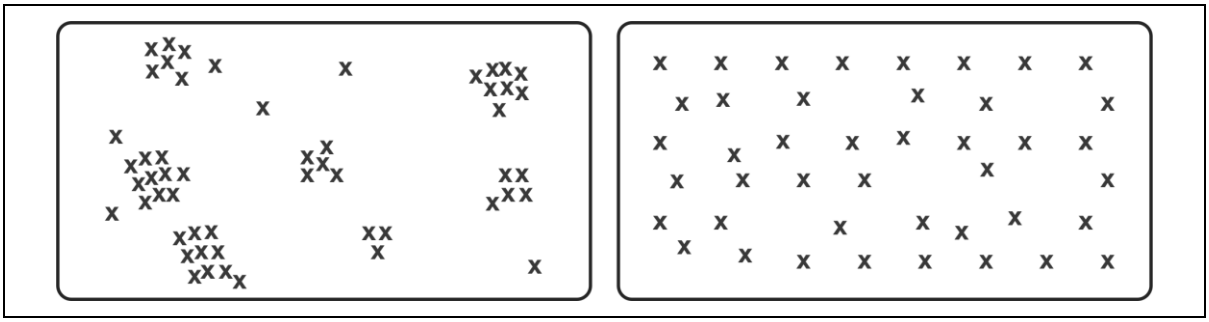
A *análise conceitual* trata, em primeira instância, de identificar o assunto principal do documento. Neste ponto, Mai (2005) ressalta as diferenças entre a indexação *orientada ao documento*, na qual o indexador se baseia somente nos atributos do documento para definir seu assunto independentemente de qualquer contexto de uso, e a indexação *orientada ao usuário*, na qual o indexador deve ter em mente os interesses e o vocabulário do usuário para fazer a indexação do assunto. Em outras palavras, não existe uma única forma correta de indexar um documento - o mesmo objeto pode ser classificado de maneiras diferentes e igualmente adequadas para propósitos específicos, dependendo da política de indexação e do interesse dos seus usuários (LANCASTER, 2004, p.9).

Já a *tradução*, significa como representar o conteúdo de um documento da forma mais fiel e objetiva possível, utilizando a linguagem de um sistema de indexação para fazer o referenciamento do objeto em uma lista de assuntos ordenados. Esta vinculação do documento a um índice ou base de dados vai possibilitar a posterior localização do seu conteúdo dentro do acervo. Lancaster (2004) complementa que a tradução pode se diferenciar quanto à extração de termos, fazendo o uso de palavras e expressões contidas no documento ou complementar suas referências (ou pontos de entrada) com outras palavras consoantes ao vocabulário controlado para adequar sua classificação.

Neste ponto, destaca-se a importância do vocabulário controlado como instrumento auxiliar para a desambiguação de termos similares, e orientação do indexador na escolha do termo mais apropriado para a representação de uma ideia, garantindo uma categorização mais consistente (LANCASTER, 2004, p. 19). Borko (1977, p. 362) afirma que o vocabulário controlado é o instrumento responsável por fazer a agregação dos documentos similares em *clusters*, submetendo os mesmo em uma disposição em núcleos de assuntos consolidados no espaço informacional. Neste caso, a requisição de um item específico vai direcionar a busca para itens similares, prevenindo a revocação excessiva e garantindo maior grau de precisão.

Uma representação visual da comparação entre o espaço informacional organizado em *clusters* e outro onde os itens se dispersam de forma homogênea, pode ser visualizada na imagem abaixo:

Figura 4 - Agrupamento em clusters x Dispersão homogênea



Fonte: Reproduzido pelo autor com base em Borko (1977, p. 362).

Todos os pontos tratados até aqui, em relação à indexação utilizada em acervos bibliotecários e arquivísticos tem uma profunda relação com os métodos utilizados na gestão de banco de dados contendo informações pessoais. Conforme dito por Lancaster (2004), o resultado de uma indexação é a representação de um objeto. O mesmo é considerado para o tratamento de dados pessoais e dos atributos que caracterizam o indivíduo, o que Lyon (2007) com base em Haggerty e Ericcson (2000) chama de duplo informacional (*data-double*). Ou seja, a representação do usuário, em suas características pessoais, preferências ou ações, através de parâmetros objetivos, que permite a vigilância dos dados (*dataveillance*) e a ação à distância.

Após a captação de dados pessoais – já descrita anteriormente, através da exposição voluntária dos usuários, ou por vias menos transparentes – a intenção de todo analista de dados é aumentar seu entendimento sobre a população a ser estudada, e isso se faz com o estabelecimento de objetivos e a eleição de critérios para categorização dos indivíduos e agrupamento dos mesmos em perfis (*profiling*).

Mas, uma vez que o *Big Data* é constituído de dados em múltiplos formatos, captados muitas vezes a esmo e de forma não estruturada, para chegar à análise dos dados em última instância, primeiramente é necessário a utilização de tecnologias e métodos para o tratamento dos dados brutos.

4.2 MINERAÇÃO DE DADOS (*DATA MINING*)

Quando um conjunto de dados se encontra na forma de texto não estruturado, este precisa receber tratamento adequado para que todo material seja “garimpado” e depurado. Isso ocorre quando os dados brutos extraídos de um fato ou evento, principalmente em grande

escala como é o caso do *Big Data*, representam a realidade com todas suas imperfeições. Em outras palavras, a massa de dados coletada pode conter, inicialmente, dados incompletos, inconsistentes, imprecisos, redundantes ou com "ruídos", que precisam ser corrigidos e trabalhados para oferecerem maior utilidade (FACELI et al., 2011).

O processo de mineração de dados está inserido no campo temático da Recuperação da Informação (RI) e visa, sempre, em última instância a extração de informações e identificação de padrões – processo este chamado de *descoberta de conhecimento* (*knowledge discovery*). Alguns autores optam por distinguir os dois processos. Alguns consideram a mineração de dados um conceito mais abrangente que inclui a fase de pré-processamento, responsável pela preparação (limpeza, integração, transformação e redução) do material que será, posteriormente, processado e analisado; outros assumem que a mineração de dados propriamente dita ocorre somente a partir do momento em que o conjunto de dados a ser estudado já se encontra devidamente "depurado" e armazenado de forma estruturada, no formato atributo-valor, para a aplicação das rotinas computacionais (FACELI et al., 2011; WEISS et al., 2005).

Em Mineração de Textos, rotinas de processamento normalmente são aplicadas para reduzir a complexidade do material coletado e aumentar a evidência do seu conteúdo essencial em relação à massa de dados irrelevantes. Neste caso, através de uma abordagem estatística e semântica, algoritmos de preparação são utilizados para (1) eliminar palavras e termos comuns que não agregam valor semântico ao texto (como artigos e preposições), (2) identificar toda palavra ou expressão e associá-las a um termo raiz, cadastrando sua posição de origem e frequência no texto e (3) atribuir valores (pesos) para a relevância de cada palavra ou expressão, de acordo com o seu aparecimento em locais de maior destaque ou importância no texto (MORAIS; AMBRÓSIO, 2007; SOARES DA SILVA, 2016).

Esses recursos são amplamente utilizados na mineração de textos na Web (*webmining*) para a extração de informação relevante, cadastramento ou indexação das páginas em bases de dados para facilitar buscas. Além disso, os algoritmos de indexação da maioria das ferramentas de busca verificam, entre outros fatores, o grau de semelhança e os *links* entre as páginas de diferentes *sites* para fazer uma ponderação da relevância de seus conteúdos visando apresentar melhores resultados de busca para os usuários. A verificação do grau de semelhança entre os documentos é feita através de uma análise de conteúdo ou comparação automática de seus atributos, que resulta na aproximação desses objetos com a formação dos *clusters* - conjuntos temáticos específicos. Estes mesmos princípios de verificação da relação entre documentos, agrupamento e ponderação da relevância, são utilizados tanto na

Bibliometria quanto na elaboração de perfis de usuários (*profiling*) – pois ambos decorrem dos mesmos princípios de consistência matemática.

Bibliometria é a análise estatística de livros, artigos e outras publicações (OXFORD). Sua função é "analisar, quantificar e mensurar fenômenos de comunicação, com a construção de representações precisas do seu padrão de comportamento e interação, para fins de conhecimento, avaliação e administração" (DE BELLIS, 2009, p.3, tradução nossa). Originalmente, o trabalho da bibliometria estava limitado à classificação de artigos e publicações científicas, considerando autores, instituições, campos de pesquisa, etc. para a elaboração de índices de produtividade. Na sequência, outras técnicas, envolvendo um número maior de variáveis ou atributos, foram adotadas, resultando em instrumentos como o índice de citações e a co-citação de análises, que contribuíram para uma representação mais qualificada do desenvolvimento das pesquisas em seus campos temáticos (DE BELLIS, 2009).

Para exemplificar, como se dá o processo de agrupamento (clusterização), parametrização e recuperação de objetos (sejam livros ou pessoas), expõe-se aqui, com base em Soares da Silva (2016), de forma didática, um caso hipotético de indexação para uma coleção com quatro documentos (D1, D2, D3 e D4), listados abaixo com seus conteúdos na ocorrência dos termos A, B e C.

Documento	**Conteúdo**
D1	AAB
D2	ABC
D3	BBC
D4	ACC

Portanto, o documento 1 (D1) possui o conteúdo "AAB", com 2 ocorrências do termo A e 1 ocorrência do termo B, e assim por diante.

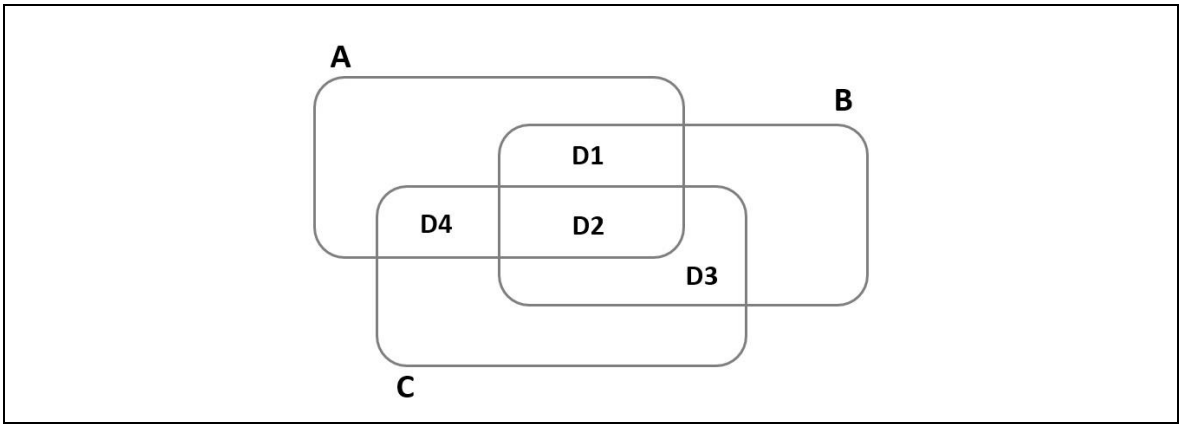
Fazendo a inversão da relação entre objetos e atributos, dispondo os documentos em função dos termos, têm-se a seguinte estruturação dos dados:

Termo	**Localização e número de ocorrências (Dn,freqt)**
A	(1,2)(2,1)(4,1)
B	(1,1)(2,1)(3,2)
C	(2,1)(3,1)(4,2)

Na primeira linha, constam a localização (documentos de origem) e a frequência (número de ocorrências) do termo A. O par ordenado (1,2) denota que o termo em questão é encontrado no documento 1, com 2 ocorrências. Seguindo esta lógica, pode-se aferir que o termo A está presente nos documentos 1, 2 e 4, com o total de 4 ocorrências em toda coleção (2 + 1 + 1).

Este índice invertido é estruturado para facilitar a localização dos termos (ou atributos) pois é através deles que os usuários definem suas necessidades de informação, em suas sentenças de busca. Os termos são dispostos em ordem alfabética na primeira coluna e relacionados aos seus documentos de origem (localização) com seus graus de ocorrência. Assim, o sistema obtém os primeiros graus de pertencimento e agrupamento dos documentos em relação aos termos (SOARES DA SILVA, 2016). Uma representação desta afirmação, na forma de conjuntos, pode ser vista da imagem abaixo.

Figura 5 - Representação da relação entre documentos em função dos termos



Fonte: Elaborado pelo autor com base em Soares da Silva (2016).

Supondo que um usuário queira consultar a base de dados descrita acima e fazer uma pesquisa (P), é possível comparar a requisição feita aos documentos disponíveis na coleção para perceber qual objeto seria o mais pertinente. A tabela abaixo permite a comparação entre os termos incluídos na pesquisa (P) e as ocorrências dos termos nos documentos, para verificar o grau de semelhança entre P e D, ou o nível de pertinência de cada D em relação a P.

Exemplo 1: P x Dn - Pesquisa (P) por “A + B” em relação aos documentos D1...D4.

	A	B	C	Result.
D1	2	1	0	3
D2	1	1	1	2
D3	0	2	1	2
D4	1	0	2	1
P	1	1	0	

O resultado expresso na última coluna foi calculado com um modelo de covalência linear extremamente simples, fazendo o somatório da multiplicação do peso de cada termo na pesquisa pelo grau de ocorrência dos mesmos nos documentos. Sendo assim, o resultado obtido por D1 em relação a P é proveniente da expressão [(1x2) + (1x1)] = 3. No exemplo em questão, os termos A e B encontram-se na sentença de pesquisa com igual peso (1). Os números que aparecem na última linha não representam apenas o grau de ocorrência dos termos na sentença de pesquisa, uma vez que não faz sentido reforçar o grau de necessidade sobre um termo replicando o mesmo no campo de busca.

Com base na situação acima, tem-se D1 com a maior pontuação, sendo, portanto, o documento mais pertinente para a requisição feita em P.

D2 e D3 encontram-se empatados. Ambos possuem 2 pontos, mas observando suas composições percebe-se que D2 obteve seu êxito com base em uma equalização de seus pontos entre os termos A e B, enquanto D3 teve sua atribuição de pertinência em função de uma concentração de suas propriedades no termo B. O modelo de cálculo (ou algoritmo) utilizado para este exemplo é extremamente simples e, por isso, também incapaz de resolver problemas mais complexos. Pergunta-se, portanto: haveria alguma forma de desambiguar este empate? Que critérios adicionais poderiam ser adotados para definir, entre a generalização ou a especialização, o que é mais pertinente?

Se fosse possível a consideração de outras características do usuário como localização geográfica, resultado de buscas anteriores ou histórico de navegação, o algoritmo do mecanismo de busca em questão poderia atribuir pesos diferentes para cada termo requisitado ou outros critérios de desempate. No segundo exemplo, faz-se um cálculo ponderado do grau de pertinência dos documentos em relação aos termos requisitados, com diferentes pesos ou intensificador de prioridade.

Exemplo 2: Pesquisa (P) por “A + B” com diferentes pesos para A e B.

	A	B	C	Result.
D1	2	1	0	7
D2	1	1	1	4
D3	0	2	1	2
D4	1	0	2	3
P	3	1	0	

Nesta nova situação, a diferença de peso, ou de nível de prioridade, atribuída para o termo A na sentença de pesquisa P, foi suficiente para desempatar o resultado da busca, tendo como documentos mais pertinentes: D1, D2, D4 e D3. Nesta ordem, D4 se mostrou mais pertinente que D3, ao contrário do resultado obtido na condição anterior.

Outro ponto de intermediação de grande utilidade é o grau de *Relevância* dos documentos – algo um tanto distinto do critério *Pertinência* já citado. A pertinência diz sobre a adequação de um documento à uma necessidade informacional, mas a relevância refere-se à importância ou grau de contribuição de um documento para a coleção da qual faz parte, em função de um termo específico. Neste caso, pode-se levar em consideração (1) o grau de destaque de um termo em relação ao conjunto de termos da coleção e (2) o percentual de participação de cada termo na sua composição de um documento.

Termos que aparecem em muitos documentos não contribuem para aumentar a relevância ou distinção de um documento qualquer que os contenham. Segundo Soares da Silva (2016), para mensurar o grau de novidade de um termo em relação à coleção, usa-se o logaritmo do número total de documentos da coleção sobre o número de documentos que contém o termo em questão. Esta grandeza é chamada de *Inverse Document Frequency* (IDF). Quanto mais raro é um termo, mais alto é o seu IDF.

$$IDFt = \log (Nd / ndt)$$

Onde: IDFt = *Inverse Document Frequency* em função do termo t; Nd = número total de documentos da coleção; ndt = número total de documentos da coleção que contêm o termo t.

A representatividade de um termo na composição de um documento, ou *Term Frequency* (TF), é uma grandeza que pode ser calculada através de um raciocínio bastante próximo à regra de porcentagem (MORAIS; AMBRÓSIO, 2007; SOARES DA SILVA, 2016).

TF(t,d) = freq(t,d) / maxfreq(T,d)

Onde: TF(t,d) = *Term Frequency* do termo t no documento d; freq(t,d) = frequência (ou número de ocorrências) do termo t no documento d; maxfreq(T,d) = máxima frequência (ou maior número de ocorrência) de um termo T no documento d.

Por fim, determina-se a relevância de um documento d na coleção, em função (1) do grau de destaque de seus termos em relação ao conjunto de termos da coleção e (2) do percentual de participação de cada termo na sua composição, fazendo a multiplicação das duas grandezas apresentadas, IDF e TF:

Peso(t,d) = IDF x TF

Onde: Peso(t,d) = peso do termo t em relação ao documento d.

Exemplo 3: Pesquisa (P) por “A + B” utilizando os pesos de cada termo em relação aos documentos.

No caso do D1, tem-se:

Peso de A em relação a D1:

Peso(A,D1) = [log (Nd / ndA)] x [freq(A,D1) / maxfreq(T,D1)];

Peso(A,D1) = [log (4 / 3)] x [2 / 2];

Peso(A,D1) = [0,28] x [1];

Peso(A,D1) = 0,28

Peso de B em relação a D1:

Peso(B,D1) = [log (Nd / ndB)] x [freq(B,D1) / maxfreq(T,D1)];

Peso(B,D1) = [log (4 / 3)] x [1 / 2];

Peso(B,D1) = [0,28] x [0,5];

Peso(B,D1) = 0,14

Peso de C em relação a D1:

Peso(C,D1) = [log (Nd / ndC)] x [freq(C,D1) / maxfreq(T,D1)];

Peso(C,D1) = [log (4 / 3)] x [0 / 2];

Peso(C,D1) = [0,28] x [0];

Peso(C,D1) = 0

OBS: Para fins de simplificação, neste exemplo, foi demonstrado apenas o cálculo da composição de D1 em função dos termos da coleção (A,B,C), onde somente A e B contribuíram para a representatividade de D1, uma vez que o valor do termo C é zero por não constar no documento em questão.

Após os cálculos, abaixo seguem os pesos de relevância cada termo para os documentos que os contêm, em relação aos demais termos e documentos da coleção.

	A	B	C	Result.
D1	0,28	0,14	0	0,42
...	...	...	...	...
P	1	1	0	

Outro recurso amplamente utilizado para verificar o grau de semelhança ou disparidade entre objetos (documentos, produtos, etc) ou calcular a relevância de um documento em relação a uma sentença de busca, é o *modelo vetorial*. Com este modelo de representação, é estabelecido um campo vetorial cujo número de eixos é o mesmo número de termos existentes em uma coleção e cada documento é representado por um vetor-resultante da soma dos vetores gerados por cada um de seus termos em seus respectivos eixos (MORAIS; AMBRÓSIO, 2007; SOARES DA SILVA, 2016).

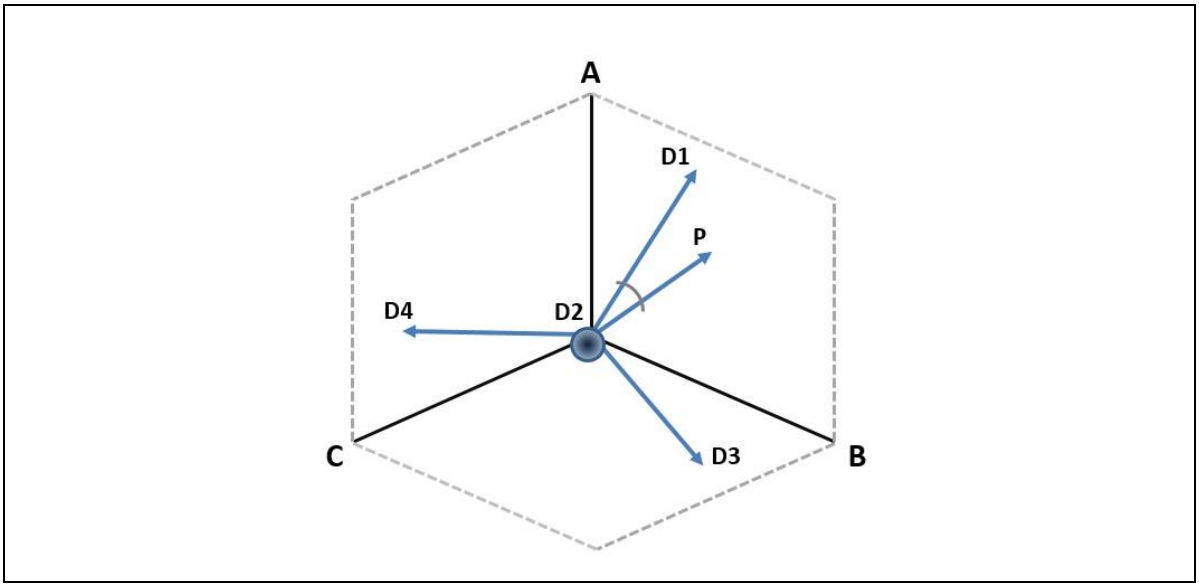
No caso do exemplo 3 desenvolvido acima, tal coleção de documentos e termos deve ser representada, no modelo vetorial, através de um campo vetorial com três eixos (A,B,C) e, consequentemente, três planos (AxB, AxC, BxC). Os documentos D1, D2, D3 e D4 são representados por vetores que resultam da soma dos seus vetores-componentes gerados por cada um dos termos que os compõem. Para D1 tem-se, portanto:

D1: (pesoA; pesoB; pesoC);

D1: (0,28; 0,14; 0)

De forma simplificada, os outros documentos (D2, D3, D4) e a Pesquisa (P) foram estimados e representados na forma de vetor no campo vetorial.

Figura 6 - Representação dos documentos e pesquisa no campo vetorial



Fonte: Elaborado pelo autor com base em Soares da Silva (2016).

Observando a imagem acima, percebe-se que o vetor que representa a pesquisa P encontra-se no plano AxB uma vez que o termo C não foi incluído na sentença de busca. O documento mais semelhante ou pertinente à necessidade de busca é aquele de maior proximidade (ou menor coseno) em relação ao vetor da pesquisa.

Faz-se importante esclarecer que o campo vetorial exemplificado contém apenas 3 eixos (A,B,C) uma vez que estes são os termos encontrados nos documentos da coleção. Um caso real pode conter centenas ou milhares de termos e eixos no campo vetorial e, uma situação dessas, exige uma capacidade de abstração elevada para o seu entendimento e seria impossível de ser representada graficamente. O modelo de cálculo (algoritmo) utilizado nos exemplos acima, também foi apresentado de maneira bastante simplificada, para fins didáticos. Em casos gerais, os algoritmos utilizados para calcular resultados de busca, ou que fazem a sugestão de produtos em função do perfil de um usuário, são surpreendentemente mais sofisticados e, por isso, capazes de resolver problemas intensamente mais complexos. Os algoritmos são o principal fator de diferenciação dos serviços de recuperação de informação e análise com a extração de informações de valor de um banco de dados, a exemplo do reconhecimento de padrões. Alguns levam em consideração centenas ou até milhares de parâmetros que são utilizados na identificação de contextos (da busca, dos documentos e perfil do usuário), interpretações semânticas e desambiguações, para maior eficiência e

eficácia de seus resultados.

Algumas fórmulas vistas nos exemplos trazidos aqui derivam dos campos da probabilidade e da estatística. Isso afirma que os princípios e métodos utilizados na indexação e recuperação da informação em acervos bibliográficos ou arquivísticos, ou nos estudos biliométricos, são os mesmos utilizados para gerenciar os dados de uma população de indivíduos (DE BELLIS, 2009; SOARES DA SILVA, 2016).

4.3 ALGORITMOS

Algoritmos formam a base da programação de computadores. São módulos de código que definem uma sequência de tarefas a ser seguida em uma ordem específica e podem conter parâmetros de verificação de determinadas situações para consequentes tomadas de decisão (CRUZ, 1997; WITTEN; FRANK; HALL, 2011).

Todo algoritmo tem um objetivo, ou visam a resolução de um problema – aqueles destinados à mineração de dados, por exemplo, podem ser orientados para transformar ou extrair informações – mas sua especificidade pode ser também seu fator de limitação.

As rotinas de processamento (ou linhas de código para a orientação de um agente autômato) podem ser restringidas pela sua capacidade de resolver tarefas específicas. Isso é o que difere antigas rotinas de programação, praticamente estáticas e limitadas à execução de uma única tarefa, para os algoritmos atuais utilizados nas práticas de *Machine Learning*. Ao desenvolverem um algoritmo, os programadores buscam sua maior versatilidade para que possa ser aplicado a um mesmo problema com diversas configurações ou a diversos problemas de estruturas similares.

Cruz (1997) traz um simples exemplo de como um algoritmo pode ser aplicado para a resolução de um problema. Assim é dada uma situação inicial em que cinco rãs estão posicionadas em uma fileira de seis caixas, conforme a figura abaixo:

	Rã 1	Rã 2	Rã 3	Rã 4	Rã 5

Pretende-se, através de uma rotina lógica, inverter a ordem das rãs, mantendo a primeira caixa vazia ao final, seguindo as premissas de que 1) duas rãs não podem ocupar a mesma caixa, 2) uma rã pode passar diretamente para uma caixa vazia à sua direita ou

esquerda ou 3) pular uma (e apenas uma) caixa ocupada para chegar a uma caixa vazia.

O problema em questão é factível e, aparentemente, não oferece grande complexidade para um ser humano adulto (ou até mesmo uma criança), mas o fato de encontrar um caminho para a solução do problema ainda não significa que um algoritmo foi desenvolvido. Ao descrever passo-a-passo o movimento que cada uma das rãs deve fazer para chegar-se o resultado final desejado, representa apenas a definição de uma rotina estática de operação e não um algoritmo. Para essa confirmação, cabem aqui duas perguntas: 1) Se for alterada a ordem inicial das rãs, a mesma rotina de operação seria útil para obter o resultado final desejado? 2) O passo-a-passo, descrito, de forma específica e rígida, é a forma mais eficiente de obter resultado final? (Qual seria o menor número de passos possível para a realização do mesmo objetivo?).

Deste modo, percebe-se claramente a diferença entre uma rotina estática (e limitada) de operação e um algoritmo - que recebe algumas condições de atuação (premissas ou regras), a ser aplicado sobre um conjunto de dados, para uma operação específica, e que possui certo grau de versatilidade para resolver problemas similares, ou, ainda, mudar seu comportamento de acordo com algumas atribuições dinâmicas que ofereçam ganho de performance.

Esta condição não estática ou até mesmo de autorregularão para ganho de performance, é a condição de todo algoritmo aplicado às práticas de *Machine Learning*.

Assim como visto em Witten, Frank e Hall (2011), aqui não se visa trazer uma discussão filosófica sobre o que é aprendizagem e se, de fato, uma máquina é capaz de aprender. Independentemente de ter ou não faculdades cognitivas reais, o termo é aplicado no campo da inteligência artificial para designar algoritmos com capacidade de 1) complementar sua programação de acordo com informações coletadas em suas operações e 2) identificar padrões.

Segundo Berry e Linoff (2004), no campo da mineração de dados, um algoritmo pode ser utilizado para solucionar problemas de ordem intelectual, econômico ou de negócios, e realizar as seguintes funções, operações ou resultados:

- Estimativas
- Classificação
- Predição
- Agrupamento por afinidade ou regras de associação
- *Clusterização*
- Descrição e criação de perfis

Estimativas referem-se a cálculos ponderados para a aproximação de um valor que não se pode obter com exatidão. Neste caso, a precisão não é o mais importante, mas sim, ter uma noção do valor de determinada característica do objeto de estudo. Como referência é possível citar a estimativa da propensão de um cliente a comprar determinado produto, feita com base em seu perfil de consumo, preferências pessoais, estilo de vida, etc. A estimativa de um valor é, portanto, um passo que precede a tarefa de classificação (BERRY; LINOFF, 2004).

Classificação é uma das mais frequentes tarefas em mineração de dados e uma constante no processo de organização entre os seres humanos – uma necessidade básica para a compreensão da realidade, com grande ênfase no desenvolvimento das ciências e no espírito modernista – ao mesmo passo da categorização e da ordenação. Classificação consiste em analisar as características de um novo objeto e associá-las a parâmetros pré-estabelecidos. No caso da prática de CRM (*customer relationship management*) de uma carteira de clientes, a classificação é aplicada a uma característica (específica) de um cliente, relacionando a mesma a uma escala de valores com classes definidas, como por exemplo, risco de inadimplência entre os níveis baixo, médio e alto.

Predição está associada às duas tarefas anteriores e relacionada à noção antecipada de valores de determinada característica de um objeto ou conjunto de características, individuais ou de um grupo, que denotem um comportamento. A predição depende da medição periódica dos parâmetros a serem analisados ou da leitura mais extensa de uma série histórica que permita a ligação dos pontos na direção de um futuro determinado. Assim também são necessários estudos de casos prévios que tenham aferido um resultado específico, que possa ser previsto com base nos primeiros passos de um padrão de comportamento conhecido.

Agrupamento por afinidade ou regras de associação são estabelecidas a partir de padrões de comportamento e leis de causa e efeito. Ainda tendo como parâmetro a observação do comportamento do consumidor, em muitos casos, não é difícil verificar a relação entre dois itens da cesta de compras. Se dois produtos estão associados - como massa de macarrão e molho de tomate, ou leite e cereais -, a escolha de um item pré-dispõe a seleção do outro, por padrão. Portanto, estes dois itens podem ser associados, e esta noção é a base de muitos sistemas de recomendação.

Clusterização é "a tarefa de segmentar uma população heterogênea em um certo número de subgrupos homogêneos" (BERRY; LINOFF, 2004, p. 11, tradução nossa). A diferença entre a classificação a *clusterização* é que este último não se baseia em categorias preestabelecidas ou na associação de objetos a modelos pré-definidos - os objetos são agrupados por padrões de semelhança, definidos por um usuário ao aplicar filtros sobre uma

base de dados, ou de forma automática, por um algoritmo de clusterização. Essa tarefa normalmente é predecessora de outras como a criação de perfis.

A criação de perfis é uma tarefa de grande importância, sobretudo quando se trata de mineração de dados de uma população de indivíduos. Cada objeto (neste caso, pessoas) possui várias características, desde traços pessoais e demográficos, até preferências e padrões de comportamento, que podem denotar desde uma propensão à compra de um produto ou ao risco de um atentado terrorista. Pelas características pessoais, que podem ser consideradas em um número indefinido de itens, a gestão de um banco de dados pode definir perfis segundo seus objetivos e políticas para a classificação dos indivíduos. No caso do CRM, exemplificado em parágrafos anteriores, a criação de perfis é baseada em um modelo previamente estabelecido, e esta tarefa precede o reconhecimento de padrões, assim como a predição de comportamento para a ação antecipada na manutenção de uma situação desejada ou para a transformação de cenários indesejados. Em outros casos, como para fatores de segurança, este item é particularmente importante, e serve de base para a gestão da vigilância sobre os indivíduos apontados pelo sistema como próximos de um padrão de risco já conhecido.

Todas as tarefas descritas acima estão relacionadas ao objetivo de obter, de forma detalhada e precisa, maior gestão sobre um grupo massivo de usuários, uma vez que, ao contrário de outros objetos como documentos, produtos ou ativos tangíveis, os indivíduos têm a característica diferencial do poder de ação – e isto pode ser usado tanto na decisão de compra de um produto, quanto para fraudes financeiras ou crimes de ordem física e massiva. Portanto, nos próximos tópicos deste trabalho, será dada ênfase ao uso das técnicas descritas para a gestão de banco de dados de usuários para fins mercadológicos ou de segurança. Em ambos os casos, a gestão sobre os indivíduos se dá pelo monitoramento de seus dados (*dataveillance*) e com ações remotas através do próprio sistema. As consequências para o usuário podem ser desde a visualização de um anúncio ou a recomendação adequada de um produto que atenda suas preferências e necessidades, até a avaliação (positiva ou negativa) de um crédito bancário ou da obtenção de visto para uma viagem internacional.

O que será mostrado a seguir são as formas e os efeitos da personalização ou distribuição seletiva da informação para o usuário em função do seu perfil e das políticas e objetivos do sistema de gestão da informação.

5 EFEITOS DA PERSONALIZAÇÃO E CATEGORIZAÇÃO SELETIVA

Na era da economia da informação e da informação como produto, a vigilância do consumo ou a coleta sistemática de dados pessoais e transacionais, tornaram-se práticas amplamente aplicadas nos processos de marketing, no atendimento aos clientes e na geração de novos produtos. Para a gestão de bases de clientes (ou de uma população) em larga escala, foram desenvolvidas tecnologias que automatizam o processo de análise de dados, com múltiplos níveis de verificação de casos - seguindo modelos de árvores de decisão -, para a oferta de respostas precisas e mais adequadas a cada demanda (PRIDMORE, 2006).

Dentre as técnicas utilizadas para a segmentação de uma base de usuários (clientes ou cidadãos) está a criação de perfis, com uso de algoritmos de classificação e *clusterização*, para a personalização de produtos e serviços, ou repostas automatizadas para manter ou fazer a contenção de situações diversas.

5.1 CRIAÇÃO DE PERFIS (*PROFILING*)

Para o estudo e entendimento maior sobre a diversidade de uma população - em suas necessidades, desejos e tendências de comportamento -, primeiramente, é indispensável a obtenção dos dados pessoais de cada indivíduo, de forma sistemática e seguindo um plano de categorização.

Segundo Pridmore (2006), os tipos de dados a serem captados dependem dos objetivos de negócio ou propósitos de cada organização e da sua capacidade em obter dados em sua relação com o usuário. Segundo o autor, no caso de uma relação comercial, os dados dos usuários a serem investigados podem se dividir em quatro categorias: geográficos, demográficos, psicográficos e comportamentais.

Os dados geográficos são utilizados para caracterizar uma região, trazendo informações sobre a densidade populacional, concentração de mercado e outros fatores socioeconômicos. Esta determinação é feita através de parâmetros que acusam a localização do usuário como código postal, prefixo telefônico (DDD), endereço de e-mail (em casos corporativos), endereços de IP, etc.

Os dados demográficos incluem nome, idade, gênero, estado civil, renda mensal, formação educacional, etnia e ocupação. Estes dados são qualificadores dos dados geográficos pois relacionam aspectos sociais e econômicos às regiões geográficas.

Os dados psicográficos são os que descrevem os indivíduos em termos de personalidade, classes, valores, estilo pessoal ou momento de vida.

Os dados comportamentais (muito utilizado no estudo das motivações e modelo de decisão de compra) aferem determinadas ações, principalmente em relação ao consumo como frequência e horários de compra, nível de lealdade à marca, preferências e nível de responsividade a estímulos de marketing, entre outros.

Todos os tipos de dados acima citados podem ser coletados de várias maneiras. No caso da relação do usuário com o Estado, as informações do indivíduo são geradas no momento do seu nascimento e posteriormente ampliadas ou acumuladas no momento de cadastro em outros bancos de dados (Cadastro de Pessoa Física, Título de Eleitor, Carteira de Habilitação, etc).

Na relação comercial - para fins de garantia de soldo e cumprimento de contratos -, é exigido o cadastramento de informações básicas do usuário, comprovadas por documentos pessoais e referências de localização. A necessidade de conhecimento da instituição ou organização, sobre o usuário, pode se estender ao logo do tempo em função de uma consequência da relação de atendimento, em que os novos benefícios exigem em contrapartida mais dados pessoais.

Por outro lado, de maneira implícita e com o mínimo de resistência, a simples utilização do serviço ou produto digital, a exemplo dos cartões de crédito ou planos de telefonia móvel - cada vez mais sofisticados em termos de conexão e transferências de informação -, já predispõe naturalmente o rastreamento de dados que denotam o comportamento do consumidor. Esses dados são utilizados para analisar a frequência e intensidade de consumo que irá classificar o usuário em certos perfis ou categorias pré-estabelecidas.

Pridmore (2006) faz ainda uma observação relevante - sobretudo quando consideramos o valor da informação no cenário do mercado atual de alta competitividade e mudanças rápidas -, que toda a base de dados interna (desenvolvida por meios internos da organização) podem ser complementados por bases externas, provindas de setores governamentais (dados públicos) ou de empresas especializadas em comercialização de base de dados pessoais – os *Data Brokers*, ou corretores de dados.

Um dos maiores, e mais conhecidos, *data brokers* do mundo é a Experian[22]. Algumas das especialidades dessa empresa são a análise de risco de crédito e detecção de fraudes financeiras. Outro grande *player* é o Instituto Nielsen[23], especializado em segmentação de mercado e análise de dados demográficos. Em sua página de serviços, a empresa oferece soluções customizadas para o entendimento de uma região, sobre vários aspectos, em vários níveis de granularidade, capaz de discriminar um bairro, uma rua ou até mesmo um quarteirão. Já a Aristotle[24]é especializada em dados eleitorais, faz a segmentação de perfis de eleitores por regiões com base em dados demográficos e psicográficos e presta serviços de assessoria para campanhas políticas. Em sua página de apresentação, a empresa afirma que todos os presidentes americanos, desde Ronald Reagan (1983), utilizaram seus serviços especializados.

> O maior valor em coletar dados pessoais de usuários está na sobreposição de várias bases de dados complementares para criar informações úteis. (...) A conexão feita entre essas bases são o resultado das técnicas de mineração de dados com objetivos de formar *clusters* e identificar padrões e relações particulares em um conjunto de dados (PRIDMORE, 2006, p. 29, tradução nossa).

Outra prática usual hoje em dia é a formação de cooperativas de dados, em que várias empresas, geralmente parceiros de negócio ou rede de fornecedores, unem-se de forma complementar para extrair maior valor dos dados. Em muitos termos de política de privacidade são mencionadas parcerias com terceiros que podem ter acesso aos dados pessoais dos usuários para fins de processamento e análise.

O Facebook, por exemplo, desenvolveu uma rede de parceiros oferecendo um serviço de cadastramento e *login* facilitado para o usuário através da sua conta na rede social. Em contrapartida, uma vez o usuário tendo feito o cadastro, mesmo não estando *logado*, o Facebook consegue rastrear todas suas atividades no site do parceiro, monitorando cada ação desde a leitura de artigos com palavras-chave, seções ou páginas visitadas, textos copiados, páginas impressas, *downloads* e transações feitas. Tudo isso para obter conhecimento avançado sobre as preferências e padrões de comportamento dos usuários com objetivos de oferecer conteúdo pertinente (seja este, publicações de outros usuários ou anúncios publicitários) (PEREZ, 2007).

---

[22] http://www.experian.com/corporate/areas-of-expertise.html
[23] http://www.nielsen.com/us/en/solutions.html
[24] http://aristotle.com/about/

Ainda sobre a captação de dados, Cufolgu (2014) informa que o cadastro de um usuário pode ser composto por dados estáticos e dinâmicos, coletados através de processos explícitos ou implícitos. Por dados estáticos entende-se os de origem demográfica, que nunca ou raramente são alterados, enquanto os dados dinâmicos são aqueles provenientes da ação do usuário - hoje em dia, facilmente rastreados pelo uso de tecnologias e sistemas digitais. Quanto ao método de captura, ocorre o processo explícito quando os dados são claramente solicitados pelo requerente e deliberadamente fornecidos pelo usuário, ao contrário do processo implícito onde o usuário não percebe a coleta de dados resultantes direto de suas ações.

De fato, usuários esperam que informações pessoais tenham que ser fornecidas para efetuarem alguma transação e normalmente eles recebem algum tipo de compensação por isso, como descontos ou algum privilégio que confira seu grau de fidelidade. Por outro lado, constata-se que os consumidores em geral não estão preocupados com os efeitos colaterais da vigilância do consumo em seu dia-a-dia (CUFOGLU, 2014).

Ao fazer adesão a um serviço totalmente digital como o Netflix, o usuário preenche uma ficha de cadastro inicial, com seus dados pessoais – nome completo, endereço, R.G., CPF, número do cartão de crédito, etc. (dados estáticos – processo explícito) e, ao começar a usar o sistema, navega por seções de interesse, assiste alguns filmes ou séries, em determinadas horas do dia, em dias específicos da semana, por tempo determinado, e com isso, alimenta o sistema com ricas informações sobre suas preferências pessoais e padrão de comportamento, de forma dinâmica, automática e inadvertida (dados dinâmicos – processo implícito).

Quadro 1 – Comparação entre tipos de perfis de usuários

Tipo	Descrição	Técnicas Usadas	Vantagens	Desvantagens
Explícito	Usuário cria manualmente o cadastro	Questionários e avaliações	Informação coletada de alta qualidade	Grande esforço exigido do usuário para atualização do cadastro
Implícito	Sistema cria o cadastro do usuário de acordo com histórico de uso e interação	Algoritmos de *Machine Learning*	Mínimo esforço exigido do usuário para a atualização do cadastro	Demanda inicial de grande quantidade de interação entre usuário e conteúdo para a criação de um cadastro acurado
Híbrido	Combinação de dados explícitos e implícitos	Ambos	Equalização das falhas com a adoção dos pontos fortes de ambos os tipos	N/A

Fonte: Cufoglu, 2014, p. 3, tradução nossa.

Cofuglu (2014) contribui ainda com uma explanação sobre os métodos para a criação de perfis **com base no conteúdo** ou através de um **processo colaborativo**.

O método com base no conteúdo se respalda na premissa da consistência de respostas, onde cada indivíduo tende a tomar as mesmas decisões, dada as mesmas circunstâncias. Neste caso, as respostas dos usuários são presumidas pelo seu histórico de comportamento.

Já o método colaborativo assume que dois usuários classificados no mesmo perfil ou categoria possuem comportamentos similares. Neste caso, a resposta de um usuário, para uma determinada demanda, pode ser tomada como "resposta mais provável" para o outro, com certo grau de confiança. Usando mais uma vez o Netflix como exemplo, usuários com mesmo perfil de uso tendem a fazer avaliações similares sobre filmes e séries. Neste caso, a média das atribuições para um filme ou série, dentro de um grupo, pode ser considerada uma boa referência de avaliação para qualquer outro membro do grupo. Os dados dinâmicos são escalonáveis, assim, a cada membro do grupo que assiste e avalia um filme ou série, o valor desta atribuição para o grupo é atualizado, confirmando ou refutando uma tendência inicial (GOMEZ-URIBE; HUNT, 2015).

Abaixo, apresenta-se um quadro comparativo dos métodos para criação de perfis:

Quadro 2 – Comparação entre métodos para a criação de perfis

Método	Descrição	Vantagens	Desvantagens
Baseado em conteúdo	Seleção ou filtro de conteúdo de uma linha de dados – registro histórico	Análise objetiva de uma base de dados grande e/ou complexa, de material digital (ex: multimídia) sem a interferência direta do usuário	1. Dependência do conteúdo estático 2. Dificuldades para apresentar novas opções/alternativas de conteúdo em função do efeito "visão em túnel"
Colaborativo	Seleção ou filtro baseado em similaridades de perfis e comportamento	1. Independência do conteúdo estático 2. Maior acurácia na preservação do padrão de recomendação em função de escolhas esporádicas e aleatórias	1. Carência: baixa capacidade de predição na recomendação de um novo item inserido na base de dados em função da não ocorrência de avaliações suficientes 2. Alta influência ou determinação das primeiras avaliações, que podem determinar o caminho de aceitação do novo item
Híbrido	Combinação dos dois tipos de seleção ou filtro	Equalização das falhas com a adoção dos pontos fortes de ambos os métodos	Equalização das falhas com a adoção dos pontos fortes de ambos os métodos

Fonte: Adaptado de Cufoglu, 2014, p. 4, tradução nossa.

A mineração de dados e da formação de perfis de usuários refletem diretamente na capacidade das organizações adaptarem suas abordagens de marketing e oferecerem produtos

mais adaptados às necessidades dos clientes. A origem desta prática está na customização de produtos a partir de linhas de montagem mais flexíveis e dos sistemas de configuração de pedidos, para a composição do produto mais adequado às necessidades do cliente (seja pelo critério de performance e/ou de custo final) (PRIDMORE, 2006).

Mas, em termos de adaptação, existe um passo além da customização (ou da simples configuração de um produto): a personalização de ambientes, centrais de atendimento e configurações de serviços, de acordo com o perfil do usuário, definidos por suas ações e escolhas, ou com base em suas preferências e padrão de comportamento. Essas práticas, na maioria dos casos, não são ostensivas e passam desapercebidas aos olhos do usuário comum, no entanto, oferecem ganho de produtividade, usabilidade e maiores benefícios na recomendação de itens ou seleção de conteúdo.

5.2 PERSONALIZAÇÃO

Segundo Cufoglu (2014, p. 1, tradução nossa), "serviços personalizados tem como objetivo relacionar os requisitos, preferências e necessidades dos usuários a configuração de uma oferta ou solução exclusiva". Complementando essa definição, Blom (2000, apud CUFOGLU, 2014, p. 2, tradução nossa) diz que personalização "é um processo de mudança de funcionalidade, conteúdo informacional ou distinção de um sistema para aumento de relevância pessoal para um indivíduo".

A relação entre anúncios e usuários feita pela Google, a partir de palavras-chave inseridas em um campo de busca, já se estabeleceu como um padrão nas práticas da publicidade online, assim como o sistema de recomendação de produtos da Amazon, onde o usuário recebe sugestões de produtos similares com base em sua navegação, visualização e seleção. Contudo, dois outros exemplos de negócio em meio digital - Netflix e Facebook - podem ser explorados aqui para maior noção de como as informações pessoais dos usuários são coletadas e expandidas, registradas e analisadas, classificadas em perfis e respondidas com base em sistemas de personalização.

5.2.1 Netflix e seu algoritmo de recomendação

Para alguns sistemas online, o algoritmo de recomendação pode ser o maior diferencial para a excelência na satisfação do usuário e essencial para a escalabilidade na gestão da base de dados. Carlos Gomez Uribe, vice-presidente de inovação em produto da Netflix, diz que o segredo do sucesso de seu sistema de recomendação está na sua adaptação e aprendizado constantes sobre as preferências e estilo dos usuários (GOMEZ-URIBE; HUNT, 2015).

O negócio da Netflix, no início da sua trajetória, era o envio de DVDs por correios, e seu maior problema era refinar o sistema de avaliação dos itens do seu acervo buscando prever quantas estrelas cada usuário daria para cada filme, série ou documentário. Desde o início, os gestores do sistema estavam cientes sobre as diferenças de gostos e nuances de comportamento dos usuários e, portanto, da necessidade de segmentação de cada perfil para uma abordagem mais adequada. Deste modo, não seria viável classificar os itens em termos gerais - fazendo uma média simples das avaliações dos usuários que apertaram o botão de *play*. Ao mesmo tempo, com a transição do seu modelo de negócio para o ambiente exclusivamente digital, o crescimento exponencial do seu acervo tornou impossível a apresentação de todos os itens existentes para o acesso. Logo, o problema da Netflix se tornou: desenvolver um sistema escalonável que pudesse "aprender" com a interação dos usuários, reconhecer padrões de comportamento e fazer recomendações assertivas. Apesar das mudanças a serem feitas, um ponto do modelo original de gestão do acervo que foi mantido pela empresa é a participação de profissionais especializados em tratamento temático da informação para fazer a classificação e indexação dos novos itens para o acervo, tal qual é feito em uma biblioteca ou arquivo. Contudo, depois de serem inseridos no sistema, cada item adquire uma classificação dinâmica a partir da interação dos usuários.

Assim, em 2006, a Netflix promoveu um concurso global para o desenvolvimento de seu novo algoritmo de recomendação, oferecendo o prêmio de um milhão de dólares para o vencedor. Muitos trabalhos foram submetidos e testados ao longo de três anos, até que, em setembro de 2009, o time Bellkor´s Programatic Chaos foi proclamado vencedor com um produto que obteve 10,06% de vantagem em performance sobre o Cinematch - algoritmo da base original (VAN BUSKIRK, 2009).

Este caso ganhou popularidade quando foram divulgados os diversos aspectos do comportamento humano considerados pelo algoritmo vencedor para seu alto índice de acertos na predição da avaliação de itens e recomendações mais assertivas. Entre os mais de cem pontos adotados estão os critérios de frequência: recenticidade e momento, das avaliações que, respectivamente, dizem respeito ao tempo de relevância de cada avaliação feita pelo usuário - com ênfase para as mais recentes - e a ponderação das avaliações feitas nos diversos dias da semana, onde foi percebido que as melhores atribuições ocorriam entre a sexta-feira e o domingo, em função do estado de ânimo do usuário.

Joe Sill – um dos desenvolvedores da equipe vencedora – informou que a experiência lhe proporcionou um grande aprendizado pela busca de diferentes modelos de algoritmos, com diferentes níveis de complexidade e granularidade, em função dos aspectos comportamentais e emocionais que definem o usuário (VAN BUSKIRK, 2009). Com este novo modelo as recomendações e avaliações (estrelas) atribuídas a cada item passaram a ser relativas a cada perfil de usuário. Em outras palavras, usuários com perfis altamente distintos passaram a receber *ratings* e recomendações diferentes para um determinado filme, uma vez que o sistema passou a considerar, entre muitos fatores, a avaliação que o próprio usuário fez para itens similares e a avaliação que outros membros do mesmo grupo (perfil) de usuários fizeram para o filme em questão.

Além dessa, outras melhorias foram implementadas no sistema de distribuição dos vídeos por *streaming* e mudanças foram feitas na interface do programa para facilitar a interação do usuário e suas escolhas, tornando a plataforma cada vez mais personalizável.

Com sua base de operações instalada em ambiente 100% digital, a Netflix passou a ser capaz de registrar, de forma ampla e absoluta, todas as ações de seus usuários. Isto representa um processo diferente da Dataficação (comentado anteriormente) - neste caso, não ocorre a conversão, mas a geração direta dos dados dentro da plataforma, de forma estruturada, a partir das ações (escolhas, preferências e comportamento de uso) dos usuários, possibilitando, além da criação de perfis, uma gama indescritível de cruzamentos e métricas.

Gomez-Uribe e Hill (2015) enfatizam a importância dos esforços aplicados na melhoria dos algoritmos e sobre os experimentos realizados com grupos de teste. Na Netflix, todas as propostas de melhoria do sistema são amplamente testadas antes de serem efetivadas. Em termos de Big Data, isso quer dizer que: qualquer alteração na quantidade de variáveis consideradas pelos algoritmos ou nos pesos dessas variáveis pode gerar mudanças significativas nos resultados do sistema, para ganho ou perda de performance. Um exemplo

dado sobre este assunto foi o processo de submissão e desenvolvimento das propostas para o concurso promovido entre 2006 e 2009. Em primeira instância, os algoritmos propostos pelas equipes tiveram performance nitidamente superior em relação ao Cinematch, mas, em segundo momento essa vantagem começou a cair. Mayer-Schonberger e Cukier (2013) comentam que a performance de um algoritmo está intimamente ligada à quantidade de dados a ser processada e, embora não haja uma conclusão definitiva sobre isso, tudo indica que exista uma ordem de grandeza, ou faixa de tamanho de um banco de dados, em que cada algoritmo consiga operar obtendo melhores índices de performance.

Para evitar perdas inesperadas em sua capacidade de acertos, a Netflix testa suas variações com grupos isolados (ou "ilhas") de usuários. Isto evidencia a sofisticada arquitetura de sistema e a grande capacidade que a empresa dispõe para a gestão dos dados, e da performance de cada variação do algoritmo, para cada amostragem de usuários, com suas características de perfil e tamanhos específicos.

Por outro lado, quantidade não é o único objetivo dos engenheiros de produto e tecnologia da Netflix. A abordagem mais adequada, para o maior índice de acertos, depende diretamente da cultura de cada grupo – suas preferências e modo de fazer escolhas. Neste sentido, Gomez-Uribe e Hill (2015) dizem que um fator chave na melhoria contínua do sistema da Netflix está nos sofisticados métodos de teste A/B, desempenhados com muitos grupos de usuários. O aprendizado advindo desse tipo de teste tem influência direta sobre o algoritmo de recomendação propriamente dito – aquele responsável por apresentar as opções de filmes na tela para os usuários. Patel (2017) diz que testes A/B são experimentos controlados, através dos quais busca-se verificar a melhor configuração do produto em relação às preferências, comportamento de uso e modos de escolha do usuário. O método consiste em definir duas versões de um produto (com variação em apenas uma de suas características) e apresentar cada uma delas para duas amostras de usuários selecionadas de um mesmo grupo ou perfil. Ao final do teste, observa-se qual das duas versões do produto obteve maior índice de sucesso em sua adequação às necessidades do usuário.

Usando a interface do Netflix como exemplo, percebe-se que uma simples variação da posição, cor ou tamanho, do ícone de busca pode resultar em diferentes taxas de acesso a este recurso. No intuito de conhecer a melhor forma para disponibilizar esse recurso, múltiplos testes são realizados – apresentando diferentes versões do objeto, com duas variações apenas para cada característica – verificando as diferentes taxas de acesso, cliques ou conversão. No caso da apresentação dos filmes, outras variações podem ocorrer desde o texto da sinopse até

a sua imagem representativa, que pode ser a própria capa do filme, uma cena específica, ou o destaque para um vilão, um mocinho, um galã, uma donzela, o ator principal ou um coadjuvante (dependendo da popularidade de cada um) – o que for mais conveniente para atrair mais público para assistir ao filme.

5.2.2 Facebook e seu *feed* de notícias

Em termos de como um algoritmo pode fazer a distribuição seletiva da informação, não existe exemplo maior que o Facebook. Esta famosa plataforma de rede social digital é conhecida também pelas críticas que recebe em relação a sua pervasividade sobre os dados pessoais dos usuários e capacidade suspeita de influenciar ou controlar o estado de ânimos das pessoas.

Em seu início, em 2004, o Facebook era apenas uma ferramenta para construção de perfis que auxiliava a interação entre pessoas, mas não tardou para que seus gestores compreendessem o potencial que seu produto tinha para criar engajamento do público e as possibilidades em obter cada vez mais informações úteis sobre seus usuários (PHILLIPS, 2007).

Aos poucos, seu o modelo de negócio foi mudando (ou se definindo) de um catálogo de perfis para ser uma plataforma de gestão de conteúdo e distribuição seletiva da informação. Hoje, a empresa fatura mais de sete bilhões de dólares por trimestre com a veiculação de anúncios publicitários (BBC, 2016) e o que faz o Facebook ser tão rentável não é somente sua capacidade de alcance, com milhões de usuários inscritos, mas seu poder de segmentação e seleção na hora de apresentar um anúncio, garantindo que a mensagem do anunciante chegue até seu público alvo com o mínimo de dispersão.

Para tornar seu ambiente cada vez mais atrativo e manter a frequência de acessos dos usuários, o Facebook tem como principal ferramenta seu algoritmo de seleção de notícias, cujo objetivo é apresentar o conteúdo mais relevante para cada usuário, garantindo o máximo do seu engajamento (CONSTINE, 2016). Para seu sucesso, este algoritmo precisa considerar todos os fatores envolvidos no processo de participação e escolha dos usuários, o que certamente envolve modelos complexos de como as pessoas percebem a relevância dos conteúdos apresentados e fazem as suas escolhas.

Um dos parâmetros utilizados para medir a popularidade das publicações são as "curtidas", os comentários, os compartilhamentos e o tempo de leitura (inferido pelo tempo de permanência de um *post* no centro da tela). Além disso, o Facebook faz, constantemente, consultas ao seu público, através de *surveys* presenciais ou através da própria plataforma, para aferir o índice de relevância dos conteúdos apresentados. O serviço incentiva constantemente os usuários a darem *feedbacks* e fazerem ajustes em suas linhas de notícias para auxiliar o algoritmo a identificar os padrões do conteúdo mais relevante para cada um. Entre as várias opções de configuração do *feed* de notícias está a de ocultar uma publicação, clicar em "ver menos publicações como esta" ou simplesmente excluir uma pessoa da lista de contatos. Por outro lado, o Facebook busca aumentar sua assertividade estimando o nível de afinidade entre os usuários a partir da frequência e nível de interação entre eles, reconfigurando automaticamente e constantemente seus algoritmos de seleção e de ordenação de *posts*.

Figura 7 – Caixa de controle para personalização do feed de notícias do Facebook

Fonte: Facebook.com.

Segundo Constine (2016), estima-se que o algoritmo do Facebook considera mais de cem variáveis e, embora ninguém saiba explicar exatamente como ele opera, é possível estimar, em termos gerais, quais são seus principais fatores: C – criador dos posts (*creator)*, P – performance da publicação (*post*), T – tipo da publicação (*type*) e R – recenticidade ou nível de atualidade da publicação (*recency*).

Quadro 3 – Principais fatores do algoritmo de seleção de notícias do Facebook

Visibilidade das publicações = C x P x T x R

Fonte: Elaborado pelo autor com base em Constine, 2016.

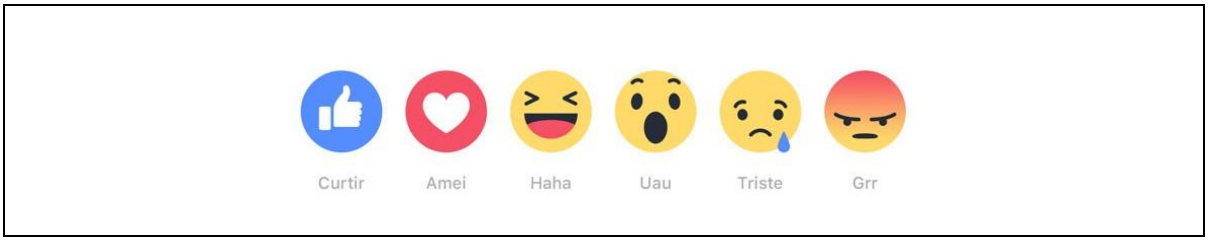
Interpretando a formula exposta no quadro acima, primeiramente tem-se o item C relativo à força do vínculo ou grau de importância do autor do post para o usuário leitor. O Facebook afere o nível de interação entre dois usuários a partir das "curtidas" e comentários, ou interações através da janela de *chat*. O mesmo pode ser dito para os graus de parentescos declarados entre os participantes, as marcações de "melhor amigo", os vínculos profissionais ou contatos pessoais frequentes confirmados pelo app do Facebook instalado no celular (e sua capacidade extraordinária de monitoramento da localização dos usuários via GPS) (OREMUS, 2016).

Em segundo, tem-se P para a popularidade ou performance das publicações. Embora o algoritmo do Facebook já possua alta capacidade para o reconhecimento de imagens ou um dicionário de palavras-chave frequentemente relacionadas a publicações importantes, a melhor maneira de aferir a relevância de uma publicação é pelo engajamento que ela gera na rede de usuários. De acordo com os critérios de relevância do Facebook, se uma publicação é inicialmente apresentada apenas para um círculo restrito de amizades, dependendo do nível de aceitação do seu conteúdo, esta pode estender seu tempo de exposição e alcançar outras instancias de amigos cadastrados.

O terceiro fator T é referente ao tipo do conteúdo. No caso do Facebook, as postagens se diferem primeiramente em 1) conteúdo pessoal (de autoria do usuário) ou 2) compartilhamento de terceiros (normalmente notícias de outros sites ou blogs). Depois tem-se a distinção da mídia utilizada: texto, foto, ilustração, vídeo ou outros recursos de publicação do Facebook. Por fim, o teor do conteúdo – neste quesito, além da identificação de algumas palavras-chave ou reconhecimento da linha editorial da fonte original do conteúdo compartilhado, o Facebook possui um artifício bastante eficaz para aferir o sentimento do usuário ao ler uma postagem, que são os botões de reação alternativos do "curtir" – amei, haha, uau, triste e grr (raiva) – ou os *emoticons* que informam o estado de ânimo do autor do *post* em relação ao seu conteúdo publicado (OREMUS, 2016, CONSTINE, 2016).

Por último está o fator R de recenticidade (atualidade) ou proximidade da publicação, que considera de maior relevância os *posts* mais próximos em termos de tempo. Este item também releva a distância física do usuário para conteúdos publicados com referência geográfica próxima ao do seu local de acesso ou por lugares já visitados e registrados.

Figura 8 – Botões de reação do Facebook



Fonte: Facebook.com.

Algumas controvérsias envolvem a reputação do Facebook e sua capacidade em determinar o que os usuários vão receber em suas linhas de notícia. Uma delas é um experimento realizado em 2014 por três pesquisadores - um membro da equipe de Data Science do Facebook e dois membros dos departamentos de Comunicação e Ciência da Informação da Universidade de Cornell -, em que foram manipuladas a linha de notícias de 700 mil usuários, para estudar o "contágio da emoção das redes sociais".

Mais uma vez, depara-se com uma política de privacidade de termos vagos que dão margem a um grande leque de interpretações sobre o que de fato será feito com os dados pessoais e conteúdo publicado pelos usuários. Alguns questionamentos foram levantados pela comunidade acadêmica e pelas áreas da Psicologia e Saúde sobre os preceitos da ética envolvida nessa pesquisa, e também sobre as fontes financiadoras do projeto pois, caso tenha tido participação de verba pública, o experimento deveria seguir o Código Americano de Regulamentos Federais, onde se define que experimentos, envolvendo pessoas, "que possam oferecer algum tipo de desconforto aos participantes" deve ter o consentimento de participação declarado pelos mesmos (WALDMAN, 2014; KRAMERA; GUILLORYB; HANCOCKB, 2014).

Outro caso, mais recente, foi a utilização do Facebook como a principal ferramenta de marketing social na campanha de eleição de Donald Trump, em 2016. Algumas suspeitas, ainda infundadas, insinuam a participação deliberada da citada rede social na manipulação das linhas de notícia dos usuários, e da falta de controle sobre a implantação de falsos boatos que circularam pela rede, e teriam favorecido o candidato republicano na corrida eleitoral. Através

da campanha feita no Facebook, a comissão eleitoral de Trump conseguiu levantar um fundo de 250 milhões de dólares, em doações.

Gary Coby – Diretor de Comunicação da Comissão Nacional Republicana –, declarou que a grande vantagem oferecida pelo Facebook enquanto plataforma de comunicação foi o condicionamento de seus usuários em clicar e se engajar nas campanhas. Através da rede social é possível ter o *feed-back* do público alvo, em tempo real (LAPOWSKY, 2016) e medir a eficiência de cada ponto de ação. Outra vantagem da ferramenta explorada foi a sua capacidade de segmentação, direcionamento de anúncios e a alta capacidade de gerenciamento de testes A/B. Conforme declarado por Gary Coby (LAPOWSKY, 2016), "quanto mais testes você fizer, mais oportunidades terá de encontrar a melhor configuração (para os anúncios)". A campanha de Trump no Facebook veiculava por dia uma média de 40.000 tipos de anúncios diferentes – muitos dos quais eram variações de teste A/B – chegando ao auge de 175.000 variações de anúncios no dia do terceiro debate presidencial.

Mais um aspecto relevante a ser ressaltado sobre o Facebook é o efeito colateral de sua alta capacidade de adaptação e personalização do conteúdo: a criação das bolhas ideológicas (ou câmaras de eco) dentro da plataforma. O sistema circular de confirmação sobre o que o usuário quer ver na sua linha de notícias reforça uma tendência que se intensifica progressivamente até o ponto de não se obter informações diferentes do habitual ou desejado, ou qualquer opinião contrária àquela que o usuário já está inclinado. Segundo Pariser (2011), o ambiente altamente seletivo das redes sociais e dos mecanismos de busca pode ser um agravante para a polarização de opiniões que prejudica o debate democrático. A seletividade dos assuntos e dos *locus* discursivos gera uma distorção na percepção da realidade e um processo de alienação nos usuários, que passam a viver em uma bolha ideológica.

Para demonstrar como a realidade pode ser apresentada de diferentes maneiras para os usuários do Facebook, em função de suas crenças pessoais, inclinações políticas e assuntos de interesse, o The Wall Street Journal criou uma página que compara duas linhas de notícia – uma azul, representando o partido Republicano e configurada para receber notícias de alinhamento conservador, e outra vermelha, representando o partido Democrata e configurada para apresentar notícias de tendência liberal. Na página do serviço[25], o usuário pode escolher entre um dos principais temas envolvidos nos debates eleitorais e comparar as diferentes frentes de opinião (KEEGAN, 2016).

[25] http://graphics.wsj.com/blue-feed-red-feed/

Com a dependência cada vez maior de sistemas para a gestão da informação e o uso cada vez mais imersivo de serviços digitais, muitos problemas ocorrem ainda pela falta de legislação específica que assegure os direitos dos usuários sobre a propriedade e segurança de seus dados pessoais.

As tentativas de predição de comportamento e práticas de personalização parecem não ter limites e as consequências do uso indiscriminado de informações sensíveis sobre o usuário podem trazer consequências sociais bastante graves.

A seção seguinte é dedicada à apresentação de situações e casos que contribuem para o entendimento prático dos efeitos nocivos da personalização e categorização seletiva.

5.3 EFEITOS NOCIVOS DA PERSONALIZAÇÃO E CATEGORIZAÇÃO SELETIVA

A seletividade de informações e determinação de respostas automáticas podem ter efeitos recursivos bastante nocivos. Além das ilhas ideológicas, apresentadas acima - que distorcem a representação da realidade apresentada para o usuário -, está o ban-opticon, proposto por Bigo (2006) e algumas práticas comerciais que afetam a privacidade dos usuários com o uso indevido de seus dados pessoais.

5.3.1 O Efeito Ban-opticon

Grande parte da rotina diária, pessoal e profissional dos indivíduos, depende do trânsito no meio digital e do acesso a sistemas, banco de dados e contas virtuais. De acordo com os padrões atuais do mercado onde “tempo é dinheiro” e das múltiplas facilidades que dão suporte ao estilo de vida moderno, o uso das tecnologias de acesso à rede e gestão digital da informação tornou-se praticamente inevitável.

Assim como a abrangência do domínio da vida contemporânea, onde não há mais divisão entre o real e o virtual, as *fronteiras e limites de acesso* também se tornaram físicos e digitais. Didier Bigo (2006) faz uma excelente exposição dos problemas enfrentados pela União Europeia para organizar os fluxos de imigração entre as fronteiras dos países. Muitos mecanismos são implementados para este objetivo, desde os mais evidentes como cancelas, roletas, muros e portões, até os baseados puramente na informação, a exemplo dos vistos,

conferência da autenticidade de documentos, entrevistas e análise de históricos. Esta talvez seja a parte mais fácil da tarefa de isolar grupos ou garantir que certos padrões de indivíduos não entrem em determinados espaços, mas conforme alertado por Bauman (2001) e Bigo (2006) ainda existem fronteiras virtuais e internas, mais difíceis de serem definidas e gerenciadas. Ainda, conforme González de Gómez (2015), Sandra Braman ressalta a dificuldade de ordenamento do Estado Informacional em função das múltiplas esferas tecnológicas e jurídicas que transpassam o seu território. Neste sentido, Bigo pergunta: "Como será possível discernir as fronteiras internas e distinguir o 'nocivo' dos outros quando estão todos dentro de um mesmo país?" (BIGO, 2011, p.55).

Visto que a observação direta sobre os corpos não é mais suficiente para a manutenção da normalidade, o caminho para a resolução das questões vistas acima está na utilização das múltiplas bases de dados – enquanto repositórios de massivos conjuntos de dados que representam a realidade – e na criação de mecanismos para a identificação e controle de padrões, da forma mais discreta, automática e não intrusiva possível.

A todo instante, quaisquer usuários dos serviços digitais estão passíveis de serem identificados e perfilados. Basta a digitação de uma senha, a inserção de um cartão magnético ou a simples passagem de um crachá por uma roleta eletrônica para o indivíduo ser reconhecido; a partir disto, o sistema (seja ele qual for) apresenta a nós apenas as opções cabíveis ao nosso *status*, enquanto, todas as outras opções não apresentadas aos nossos olhos, representam a margem de restrição - a fronteira invisível que bloqueia nosso perfil ao acesso de um espaço limitado. Bigo (2006, p.34) criou a expressão *ban-opticon*, combinando o termo "*ban*" de Jean Luc Nancy, e reconfigurado por Giorgio Agaben; e o termo "*opticon*" usado por Foucault, para caracterizar o regime de exceção e exclusão possibilitados hoje em dia pelas tecnologias de controle da informação. Este termo também traz a concepção do isolamento e da rejeição, da repulsa e do banimento. A arquitetura das instituições disciplinares utilizadas como instrumentos do poder, descritas por Foucault (1987), foram substituídas por dispositivos mais sutis e fluidos, capazes de estratificar a massa e criar perfis, da mesma forma que conseguem restringir e segregar, com alto grau de eficiência, evitando ao máximo o constrangimento ou a suspeita de se estar sendo excluído.

O autor faz um alerta para o efeito "bola de neve" causado por estes dispositivos que reforçam a situação do indivíduo classificado em determinada categoria, dificultando a transição do mesmo entre as fronteiras de classes, *status* ou perfis sociais. A situação financeira de um indivíduo por exemplo, pode ser a "chave" para o seu acesso ou restrição a

um sistema de crédito, da mesma forma que a idade e o histórico de saúde são determinantes para a definição do risco que uma empresa terá em aceitar um cliente como associado de um plano de saúde ou de um seguro de vida.

5.3.2 Dados Pessoais: Práticas de Uso e Abusos

Desde o momento em que se tornou possível guardar informações sobre o comportamento do consumidor no momento das suas decisões de compra, as empresas em geral, e principalmente os *sites* de *e-commerce*, começaram a explorar novas formas de aumentar suas vendas. No início, essas práticas restringiam-se à oferta de produtos em promoção ou similares aos anteriormente buscados pelos clientes, mas hoje está tomando proporções que só a ficção científica[26] era capaz de conceber. Muitos internautas, pouco informados sobre as práticas de distribuição seletiva da informação a partir do reconhecimento de seus padrões de navegação e escolhas, acabam intrigados quando ofertas sobre determinado assunto do seu interesse, inexplicavelmente, começam a aparecer em todo lugar da *web* ou até mesmo em sua caixa de e-mails. Isso ocorre porque, através da navegação do usuário, cada *click* dado, cada assunto acessado ou palavra pesquisada, são armazenados para a formação de grupos de cadastros, que são usados para aumentar a efetividade de futuras promoções de vendas (HAYS, 2004).

Em alguns casos, a utilização de dados pessoais, por parte dos fornecedores de serviços, visa a entrega de benefícios e facilidades, e não interfere ou prejudica em nada a situação dos usuários, clientes ou consumidores. Mas há também os casos em que o uso dos dados advindos do comportamento dos usuários pode ser feito de forma abusiva, considerando a invasão de privacidade, a interferência no direito à identidade e escolha, ou até mesmo a divulgação ou comercialização de dados sigilosos particulares.

Um caso sobre o aproveitamento legítimo de dados massivos advindos do comportamento do consumidor é o do Walmart em que a diretora executiva de informação, Linda M. Dillman, e sua equipe, perceberam em relatórios de períodos anteriores que um determinado produto havia alcançado picos de vendas (sete vezes maiores que o normal) em momentos que precediam a chegada de um furacão. A partir deste aprendizado, a rede de

[26] Em uma cena do filme *Minority Report*, o personagem principal protagonizado por Tom Cruise recebe ofertas personalizadas a partir da sua identificação feita pela retina. Para assistir a cena, basta digitar "*minority report shopping scene*" no campo de busca do YouTube.

supermercados passou a reforçar os estoques daquele produto nas unidades que se encontravam nas rotas dos tornados, garantindo o máximo das vendas em função da alta demanda e atendendo melhor seus clientes (HAYS, 2004).

Mas, a análise do comportamento do consumidor e o reconhecimento de perfis de compras nem sempre são isentos de intrusão no campo da vida privada. Em caso mais recente, a empresa Target, através do seu gerente de *business intelligence*, o estatístico Andrew Pole, fez um estudo sobre os hábitos de consumo de uma série de jovens mulheres, identificando o padrão de compras daquelas que estavam grávidas. Com isso, a Target começou uma campanha silenciosa e efetiva abordando todas as clientes identificadas como "grávidas" com ofertas específicas para cada uma das fases das suas gestações (DUHIGG, 2012).

Para citar um caso genuinamente brasileiro sobre interferência no direito à identidade e falta de transparência no uso de dados pessoais, em fevereiro de 2010 a Serasa Experian anunciou o lançamento de um produto inédito no mercado nacional: o MOSAIC. Segundo eles, a "melhor radiografia da sociedade brasileira" ou "o maior e mais completo estudo que cruza dados cadastrais da Serasa Experian, do Censo do IBGE e da Pesquisa Nacional de Amostra Domiciliar (PNAD)". No mesmo anúncio o presidente da Serasa Experian enfatiza que "essa eficiente segmentação conduz ao aumento das vendas e à fidelização dos consumidores", pois com a posse das informações privilegiadas sobre os hábitos, necessidades e desejos dos indivíduos, as empresas podem planejar de maneira mais eficaz a segmentação da oferta de seus produtos e serviços (SERASA, 2010). Se antes a Serasa Experian era um serviço de proteção de crédito, agora passou a ser declaradamente um "serviço de informações para apoio na tomada de decisões das empresas". As informações pessoais dos cidadãos brasileiros, confiadas ao Governo e que seriam destinadas para a melhoria das políticas públicas, passam a ser vendidas para empresas privadas.

Em 2013 houve mais uma polêmica, de âmbito global, envolvendo o acesso e uso indevido de informações, que veio à tona quando Edward Sownden, profissional da computação, entregou para os jornais The Guardian e Washington Post documentos sigilosos da Agência de Segurança Nacional Americana (NSA). Estes documentos denunciaram abusos de poder de autoridades americanas pela invasão da privacidade de cidadãos americanos e estrangeiros. Além disso, foi divulgado que o governo americano, através das empresas de tecnologia da Internet, tinha amplo acesso a informações sigilosas de empresas e de outros governos. O fato colocou em cheque as relações internacionais dos Estados Unidos com

alguns países e levou ao descrédito a política de privacidade das empresas envolvidas - entre elas, Microsoft, Yahoo, Google, Facebook e Apple (ELLIOTT; RUPAR, 2013).

O material divulgado por Snowden apresentava em detalhes como funciona o programa de coleta de dados da NSA chamado PRISM - sem dúvida, o projeto mais audacioso da História em termos de processamento de dados massivos (*Big Data*) e controle da informação (THE WASHINGTON POST, 2013). Isso só foi possível após o ataque terrorista de 11 de setembro de 2001, quando o congresso americano aprovou novas medidas de segurança, ampliando o poder dos órgãos de defesa para a vigilância de massa (ELLIOTT; RUPAR, 2013). Desde então a vigilância do Estado Americano foi intensificada sob o argumento de que a segurança contra o terrorismo é mais importante que a privacidade dos cidadãos.

Em um grande evento de tecnologia digital e novas mídias, o diretor de tecnologia da CIA, sr Ira Hunt, fez uma declaração que expressou bem o apetite das agências de inteligência americana por informação em suas campanhas de vigilância em massa:

> o valor de um elemento de informação só é conhecido quando o conectamos a "algo mais" que aparece em um ponto futuro da nossa trajetória. Uma vez que não conseguimos conectar aqueles pontos que não possuímos, somos impulsionados para uma política de, fundamentalmente, tentar coletar tudo e manter a guarda desses dados para sempre (Sledge, 2013 - tradução nossa).

A partir dos documentos vazados, é possível ler também, nas próprias falas da NSA e do GCHQ[27] a vigência de uma "era de ouro" da vigilância eletrônica, na qual, somente o programa TEMPORA da GCHD é capaz de gravar 39 bilhões de "eventos" por dia e o programa da NSA chamado DISHFIRE pode coletar uma média de 190 milhões de mensagens de texto (NYST, 2015).

E para deixar os teóricos da conspiração ainda mais alertas em relação às práticas abusivas das empresas de tecnologia sobre a invasão de privacidade de seus clientes, em fevereiro de 2015 a companhia Samsung divulgou em seu *site* um termo aditivo à política de privacidade de sua Smart TV - que possui funções avançadas de interação como o usuário por comando de voz ou gestos - dizendo que a empresa pode "coletar, usar, armazenar e compartilha com terceiros, informações reconhecidas pela Smart TV" e que, portanto, os usuários deveriam estar atentos sobre suas conversas pessoais e assuntos sensíveis em caso de

[27] Sigla em inglês para General Communication Head Quarter - organização britânica responsável por abastecer o Governo e as Forças Armadas britânicas com informações de Inteligência.

estarem próximos da TV. Talvez a *realidade* devesse agora dizer que qualquer semelhança com a *ficção* é mera coincidência[28] (LOMAS, 2015).

Práticas como estas estão se tornando cada vez mais frequentes o que faz despertar a preocupação e reação de alguns segmentos da opinião pública e, uma vez que a situação entre governo, empresas e cidadãos/consumidores se torna instável, pelo atrito de interesses, direitos e deveres, surgem novos atores para tentar equilibrar essa relação – fundações, comissões de ética, ONGs, coletivos de ativistas, movimentos da sociedade civil organizada, campanhas na Internet, entre outros.

Além dos casos citados anteriormente, existem aqueles que o uso indevido de informações ou material de cunho pessoal se dá, não entre empresa e cliente, ou governo e cidadãos, mas na relação interpessoal dos usuários da rede.

Goulart e Serafim (2015) comentam o caso de uma dupla sertaneja que lançou o videoclipe "Vou Jogar na Internet". O trabalho artístico, considerado polêmico e politicamente incorreto, fala de um homem que filmou momentos de intimidade com sua namorada, sem o consentimento dela, e após o término do namoro, chantageava a parceira ameaçando publicar os vídeos na Rede. Minutos depois do lançamento deste videoclipe na Internet, os usuários da rede YouTube reagiram com veemência, repudiando o vídeo com inúmeros comentários que exigiam a sua retirada do ar. Com esta censura, feita pela própria comunidade de usuários, os administradores do *site* ficaram cientes do conteúdo impróprio e o vídeo foi banido.

O senador Romário (PSB-RJ) - autor do projeto de lei que tipifica o *revenge porn*[29] como crime - comentou o episódio. "Isso não é brincadeira. As consequências para as vítimas são gravíssimas. A integridade física, moral e psicológica das vítimas são abaladas depois de terem a vida íntima exposta desta forma", declarou. A proposta apresentada pelo senador prevê pena de até três anos de detenção, além de indenização à vítima (STREIT, 2015).

Em 2012, ocorreu no Brasil o caso Carolina Dieckmann[30], que resultou na lei, de mesmo nome, que classifica criminalmente delitos informáticos. Este caso não foi classificado como *revenge porn*, mas teve consequências igualmente nocivas para a moral da vítima o que

[28] A matéria da TechCrunch denunciando as falhas de privacidade da Smart TV da Samsung faz uma menção direta à obra 1984 de George Orwell, 1984, que cunhou o termo "Big Brother".

[29] Termo em inglês que significa "vingança pornográfica" através do ato de publicar material, contendo sexo explícito ou nudez, de uma pessoa sem o consentimento da mesma.

[30] A atriz teve seu computador invadido e seus arquivos pessoais subtraídos, incluindo fotos íntimas, que foram publicadas na Internet e rapidamente se espalharam pelas redes sociais.

colocou em pauta no cenário nacional um sério questionamento: até que ponto a privacidade digital está segura? (OLIVEIRA JR, 2012).

6 A TEMÁTICA DO PRESENTE ESTUDO NA PRODUÇÃO DA CIÊNCIA DA INFORMAÇÃO NO BRASIL

Para identificar se a Ciência da Informação, no Brasil, vem incluindo os temas *vigilância, privacidade, big data* e *dados pessoais* em suas pesquisas acadêmicas de Mestrado e Doutorado, foram consultados os Anais do Encontro Nacional de Pesquisa em Ciência da Informação (ENANCIB) dos últimos dez anos, por critério de recenticidade e pela possibilidade de serem referentes às questões envolvendo o uso das Novas Tecnologias de Informação e Comunicação (NTIC), com o aparecimento do termo *vigilância* e *privacidade* junto ao uso de *big bata* e *dados pessoais* - termos chave que definem o escopo deste estudo.

A pesquisa foi feita através de consulta a três bases de dados: 1) o repositório Benancib[31], as pesquisas apresentadas e publicadas no Encontros Nacionais de Pesquisa e Pós-Graduação em Ciência da Informação – ENANCIB, de 1994 (sua primeira edição) até 2014; 2) Anais do XVI ENANCIB; 3) Anais do XVII ENANCIB[32].

O repositório Benancib é uma coleção do projeto Questões em Rede, um sistema online que possui uma base de dados com informações sobre todos os trabalhos apresentados e publicados no ENANCIB de 1994 a 2014. Este sistema permite a busca dos artigos através de 1) navegação por data de publicação, autor, título ou assuntos; através de 2) busca simples por palavras-chave; ou por 3) busca avançada, com filtro por: autor, título, palavra-chave, resumo, idioma, ano do evento, cidade do evento, edição do evento, número do GT, nome do GT, assuntos e referências. O resultado da navegação ou busca é apresentado através de uma lista dos itens encontrados, representados por suas formas de citação, que podem ser listados por título ou data de publicação, em ordem ascendente ou descendente. Ao clicar em um dos itens apresentados, o sistema abre uma página com o título, autor, resumo, palavras-chave, *link* para *download* do artigo em PDF e *link* para informações completas sobre o trabalho.

[31] http://repositorios.questoesemrede.uff.br/repositorios/handle/123456789/2

[32] As consultas foram feitas de forma separada pois ao Anais do ENANCIB XVI e XVII ainda não constam indexados no Benancib, até o momento de realização desta pesquisa, contudo, os critérios utilizados foram os mesmos para possibilitar a consolidação dos dados.

6.1 PESQUISA NO REPOSITÓRIO BENANCIB

Na **navegação exploratória**, inicialmente feita no repositório Benancib, através do diretório de assuntos, foram identificados: dois trabalhos cadastrados no assunto vigilância, nenhum trabalho cadastrado no assunto privacidade, um trabalho cadastrado no assunto big data e nenhum trabalho cadastrado no assunto dados pessoais.

Quadro 4 – Pesquisa por navegação em diretório de assuntos do Benancib

Repositório: Benancib	
Método utilizado: Navegação por diretório de assuntos	
Assunto	Itens encontrados
vigilância	2
privacidade	0
big data	1
dados pessoais	0

Fonte: Elaborado pelo autor.

O resultado acima mostrou-se insuficiente para conclusões acerca da representatividade dos termos na produção da Ciência da Informação, de modo que foram buscados outros métodos de busca na consulta dos repositórios.

Em **busca simples** por cada um dos termos (assuntos), foi verificado um resultado mais expressivo pois, cada artigo pode apresentar o termo buscado, desde seu título (de maior importância) até suas referências. Portanto, em segundo momento, na análise dos resultados foi feita a distinção entre os trabalhos que apresentaram os termos buscados em seu título, resumo, palavras-chave e corpo de texto, daqueles que simplesmente apresentaram os termos nas referências. Em um total de 2721 itens do repositório, 28 continham a palavra *vigilância*, 97 continham a palavra *privacidade*, 1150 continham *big data* e 2445 continham *dados pessoais*.

Quadro 5 - Busca simples pelos termos *vigilância*, *privacidade*, *big data* e *dados pessoais* no Benancib

Repositório: Benancib (2721 itens no total)	
Método utilizado: Busca simples pelos termos	
Assunto	**Itens encontrados**

vigilância	28
privacidade	97
big data	1150
dados pessoais	2445

Fonte: Elaborado pelo autor.

Observando o resultado obtido, foi definido que os dois termos com menor grau de ocorrência (vigilância e privacidade) seriam os de maior relevância para o refinamento da busca a fim de apontar os trabalhos com maior proximidade temática em relação à presente dissertação. Portanto, a partir dos 28 artigos contendo *vigilância* e dos 59 contendo *privacidade*, foi verificado o grau de importância de cada termo para cada artigo.

6.1.1 Busca simples por *vigilância* no Benancib

A partir do resultado inicial da busca por *vigilância*, 17 artigos foram descartados: 1 por conter o termo apenas no resumo, 1 por conter o termo apenas no corpo do texto, 13 artigos por conterem o termo somente nas referências e, curiosamente, 2 por não conterem referências sobre o assunto embora o termo estivesse presente no resumo e no corpo de texto.

Dos 11 textos restantes, 3 apresentavam o termo *vigilância* no contexto da vigilância sanitária, 2 no contexto da vigilância em saúde, 1 no contexto de monitoramento do ambiente externo de marketing, 1 no contexto da observância das práticas de responsabilidade social de empresas e governos, e apenas 4 no contexto da vigilância conforme adotado neste trabalho: "a observação de informações pessoais, de forma proposital, rotineira e sistemática com objetivos de controle, direitos, gestão, influencia ou proteção" (WOOD et al., 2006, p. 04).

Quadro 6 - Contextualização do termo *vigilância* em textos selecionados no Benancib

Amostra: 11 trabalhos com a ocorrência do termo *vigilância* nas referências e em pelo menos um dos campos de indexação (título, resumo, palavras-chave ou corpo do texto).		
Método utilizado: análise de conteúdo		
Assunto	**Contexto**	**No. de trabalhos encontrados**
Vigilância	Vigilância Sanitária	3
	Vigilância em Saúde	2

	Marketing – monitoramento de ambiente externo	1
	Observatório sobre Responsabilidade Social	1
	Observação de informações pessoais, de forma proposital, rotineira e sistemática com objetivos de controle, direitos, gestão, influencia ou proteção.	4

Fonte: Elaborado pelo autor.

Considerando apenas os 4 textos em que o contexto de *vigilância* está compatível com a definição adotada nesta dissertação, fez-se um quadro para a comparação dos campos em que o termo ocorre, a fim de identificar os mais relevantes.

Quadro 7 - Seleção de textos com o termo *vigilância* no Benancib

Amostra: 4 trabalhos com a ocorrência do termo *vigilância* nas referências e em pelo menos um dos campos título, resumo, palavras-chave ou corpo do texto, compatível com o contexto ou definição de vigilância adotados nesta dissertação.							
Método utilizado: análise de conteúdo							
Item	Campo de ocorrência do termo *vigilância*					Ano	GT
	Título	P.Chave	Resumo	Texto	Refs.		
1	1	0	1	1	1	2011	11
2	1	1	1	1	1	2014	3
3	1	1	1	1	1	2014	5
4	0	0	0	1	1	2014	5
Legenda: P.Chave: Campo de palavras-chave; Refs.: Campo de referências; 1: Ocorrência do termo no referido campo; 0: Não ocorrência do termo no referido campo.							
Itens: Textos de maior relevância contendo o termo vigilância no contexto desta dissertação							
1	CAVALCANTE, Ricardo Bezerra; PINHEIRO, Marta Macedo Kerr. Sistema de informação da atenção básica: relações de poder, centralização e vigilância. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 12., 2011, Brasília, Anais... Brasília: UnB, 2011.						
2	BEZERRA, Arthur Coelho. "Culturas de vigilância", "regimes de visibilidade": novos caminhos para a pesquisa em ciência da informação. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 15., 2014, Belo Horizonte. Anais... Belo Horizonte: UFMG, 2014.						
3	BEZERRA, Arthur Coelho; PIMENTA, Ricardo Medeiros; ORMAY, Larissa Santiago. Vigilância, vigilância inversa e democracia: do panoptismo ao midiativismo. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 15., 2014, Belo Horizonte. Anais... Belo Horizonte: UFMG, 2014.						
4	ANTONIUTTI, Cleide Luciane; ALBAGLI, Sarita. Uso do big data em campanhas políticas eleitorais. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 15., 2014, Belo Horizonte. Anais... Belo Horizonte: UFMG, 2014.						

Fonte: Elaborado pelo autor.

Com este resultado, percebeu-se que o ano de 2014 foi o mais profícuo em termos de publicação de trabalhos com o tema em questão, contendo 3/4 das publicações e o GT 5 (Política, Economia e Informação) foi o que mais contribuiu, com 50% das publicações válidas, de acordo com os critérios de seleção estabelecidos neste trabalho.

6.1.2 Busca simples por *privacidade* no Benancib

Em **busca simples** por *privacidade*, obteve-se, inicialmente, o resultado de 97 itens com a ocorrência do termo. A partir deste montante inicial, foi adotado o critério de seleção em que 1) o termo *privacidade* deveria ocorrer no corpo do texto e em mais um campo de indexação do artigo - como título, resumo, palavras-chave ou referências; 2) a ocorrência *privacidade* não deveria ser apenas uma menção ou citação simples, mas ter também sua discussão ampliada no âmbito do direito, de respaldo e preservação da intimidade a partir de dados pessoais. Definido os parâmetros de seleção, foram destacados 5 trabalhos.

Quadro 8 - Seleção de trabalhos com o termo *privacidade* no Benancib

Amostra: 5 trabalhos com a ocorrência do termo *privacidade* no corpo do texto e em mais um campo de indexação do artigo e sendo discutido no âmbito do direito, de respaldo e preservação da intimidade a partir de dados pessoais.							
Método utilizado: análise de conteúdo							
Item	Campo de ocorrência do termo *privacidade*					Ano	GT
	Título	P.Chave	Resumo	Texto	Refs.		
1	0	1	0	1	0	2007	4
2	0	1	0	1	0	2011	5
3	0	1	0	1	1	2014	3
4	0	1	0	1	0	2014	5
5	0	0	0	1	1	2014	5
Legenda: P.Chave: Campo de palavras-chave; Refs.: Campo de referências; 1: Ocorrência do termo no referido campo; 0: Não ocorrência do termo no referido campo.							
Itens: Textos de maior relevância contendo o termo *privacidade* no contexto desta dissertação							
1	ISONI, Miguel Maurício; VIDOTTI, Silvana Aparecida Borsetti Gregorio. Percepções de segurança e ameaças em ambientes de tecnologia da informação. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 8., 2007, Salvador. Anais... Salvador: UFBA, 2007.						
2	MARQUES, Rodrigo Moreno; PINHEIRO, Marta Macedo Kerr. Assimetria de informação na Lei Geral de Telecomunicações: uma análise dialética. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 11., 2010, Rio de Janeiro. Anais... Rio de Janeiro: IBICT, 2010						

3	BEZERRA, Arthur Coelho. “Culturas de vigilância”, “regimes de visibilidade”: novos caminhos para a pesquisa em ciência da informação. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 15., 2014, Belo Horizonte. Anais... Belo Horizonte: UFMG, 2014.
4	BEZERRA, Arthur Coelho; PIMENTA, Ricardo Medeiros; ORMAY, Larissa Santiago. Vigilância, vigilância inversa e democracia: do panoptismo ao midiativismo. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 15., 2014, Belo Horizonte. Anais... Belo Horizonte: UFMG, 2014.
5	ANTONIUTTI, Cleide Luciane; ALBAGLI, Sarita. Uso do big data em campanhas políticas eleitorais. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 15., 2014, Belo Horizonte. Anais... Belo Horizonte: UFMG, 2014.

Fonte: Elaborado pelo autor.

Com este resultado, percebeu-se que o ano de 2014 foi o mais profícuo em termos de publicação de trabalhos com o tema em questão, contendo 3/5 das publicações e o GT 5 (Política, Economia e Informação) foi o que mais contribuiu, com 60% das publicações válidas, de acordo com os critérios de seleção estabelecidos neste trabalho.

6.2 PESQUISA NOS ANAIS DO ENANCIB XVI E XVII

Os trabalhos do XVI ENANCIB[33] podem ser visualizados no site do evento. Na sessão ANAIS, é possível buscar os artigos através de navegação por sobrenome do autor em ordem alfabética ou por GT. Na sessão PESQUISA, é possível fazer uma busca simples, ou uma busca avançada parametrizada por categorias (autor, título, texto completo e documentos suplementares), por data de publicação ou por termos indexados. O resultado da navegação ou busca é apresentado através de uma lista de itens, representados por seus títulos, nome(s) do(s) autor(es) e *link* para download do artigo em PDF, por critério de ordenação não reconhecido. Ao clicar em um dos itens apresentados, o sistema abre uma página com o título, autor, resumo e *link* para download do artigo em PDF.

Para os trabalhos do XVII ENANCIB[34], na sessão Anais do ENANCIB 2016, é possível fazer uma navegação por sobrenome do autor em ordem alfabética ou por GT, ou uma busca simples limitada por título do trabalho ou nome do autor. Na barra lateral direita é possível fazer buscas simples ou avançadas parametrizadas por autor, título, resumo, termos indexados ou texto completo. O resultado da navegação ou busca é apresentado através de uma lista de itens, representados por seus títulos, nome(s) do(s) autor(es) e *link* para download do artigo em PDF, cujo critério de ordenação não foi identificado. Ao clicar em um dos itens

[33] http://www.ufpb.br/evento/lti/ocs/index.php/enancib2015/
[34] http://www.ufpb.br/evento/lti/ocs/index.php/enancib2016/

apresentados, o sistema abre uma página com o título, autor, resumo e *link* para download do artigo em PDF.

6.2.1 Busca simples por *vigilância* nos Anais do ENANCIB XVI e XVII

Em **buscas simples,** pelo termo *vigilância*, feita nos Anais do ENANCIB XVI e XVII, foi obtido o seguinte resultado:

Quadro 9 - Resultado do termo *vigilância* nos Anais do ENANCIB XVI e XVII

Termo buscado:	**Anais**	**Itens encontrados**
Vigilância	ENANCIB XVI	1
	ENANCIB XVII	2

Fonte: Elaborado pelo autor.

Este resultado de busca foi ratificado com a aplicação dos mesmos critérios utilizados para a seleção de itens na busca feita no Benancib.

Quadro 10 - Seleção de trabalhos com o termo *vigilância* nos ENANCIB XVI e XVII

Amostra: 3 textos com a ocorrência do termo *vigilância* nas referências e em pelo menos um dos campos título, resumo, palavras-chave ou corpo do texto, compatível com o contexto ou definição de vigilância adotados nesta dissertação.							
Método utilizado: análise de conteúdo							
Item	Campo de ocorrência do termo *vigilância*					Ano	GT
	Título	P.Chave	Resumo	Texto	Refs.		
1	1	1	1	1	0	2015	3
2	0	1	1	1	0	2016	2
3	1	1	1	1	1	2016	4
Legenda: P.Chave: Campo de palavras-chave; Refs.: Campo de referências; 1: Ocorrência do termo no referido campo; 0: Não ocorrência do termo no referido campo.							
Itens: Textos de maior relevância contendo o termo *vigilância* no contexto desta dissertação							
1	BEZERRA, Arthur Coelho. Vigilância e Filtragem de Conteúdo das Redes Digitais - desafios para a competência crítica em informação. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 16., 2015, João Pessoa. Anais... João Pessoa: UFPB 2015.						
2	MOURA, Maria Apareceida. DECIFRA-ME OU DEVORO-TE: CONTEXTO, SIMILARIDADE SEMÂNTICA E TERMINOLOGIA ESPECIALIZADAEM SERVIÇOS DE INTELIGÊNCIA NO BRASIL. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 17., 2016, Salvador. Anais... Salvador: UFBA 2016.						
3	PÉREZ, Lisandra Guerrero Pérez; NASSIF, Mônica Erichsen. FATORES DE INFLUÊNCIA NA AVALIAÇÃO DOS						

	OBSERVATÓRIOS SOCIAIS DO BRASIL ENTENDIDOS COMO SISTEMAS DE VIGILÂNCIA INFORMACIONAL. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 17., 2016, Salvador. Anais... Salvador: UFBA 2016.

Fonte: Elaborado pelo autor.

6.2.2 Busca simples por *privacidade* nos Anais do ENANCIB XVI e XVII

Em **buscas simples,** pelo termo *privacidade*, feita nos Anais do ENANCIB XVI e XVII, foram encontrados, respectivamente, 2 e 1 itens. Contudo, com a aplicação dos mesmos critérios utilizados para a seleção de itens na busca feita no BENANCIB, 2 itens foram descartados por apresentarem somente a ocorrência do termo no corpo de texto do artigo. Tendo feita essa correção, apresenta-se abaixo o resultado obtido.

Quadro 11 - Resultado do termo *privacidade* nos Anais do ENANCIB XVI e XVII

Termo buscado:	**Anais**	**Itens encontrados**
Privacidade	ENANCIB XVI	1
	ENANCIB XVII	0

Fonte: Elaborado pelo autor.

Quadro 12 - Seleção de trabalhos com o termo *privacidade* nos ENANCIB XVI e XVII

Amostra: 1 trabalho com a ocorrência do termo *privacidade* no corpo do texto e em mais um campo de indexação do artigo e sendo discutido no âmbito do direito, de respaldo e preservação da intimidade a partir de dados pessoais.							
Método utilizado: análise de conteúdo							
Item	Campo de ocorrência do termo *privacidade*					Ano	GT
	Título	P.Chave	Resumo	Texto	Refs.		
1	0	1	1	1	1	2015	8
Legenda: P.Chave: Campo de palavras-chave; Refs.: Campo de referências; 1: Ocorrência do termo no referido campo; 0: Não ocorrência do termo no referido campo.							
Itens: Textos de maior relevância contendo o termo *privacidade* no contexto desta dissertação							
1	AFFONSO, Elaine Parra Affonso; SANT'ANA, Ricardo César Gonçalves. Anonimização de Metadados de Imagem Digital por maio do Modelo K-Anonimato. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 16., 2015, João Pessoa. Anais... João Pessoa: UFPB 2015.						

Fonte: Elaborado pelo autor.

Consolidando os resultados obtidos com a consulta das três bases de dados, tem-se os seguintes resultados para cada termo:

Quadro 13 – Publicações com termo *vigilância* no ENANCIB, no contexto abordado nesta dissertação, nos últimos dez anos

	Grupo de Trabalho (GT)											
	1	2	3	4	5	6	7	8	9	10	11	**Total**
2007												
2008												
2009												
2010												
2011											1	**1**
2012												
2013												
2014			1		2							**3**
2015			1									**1**
2016		1		1								**2**
Total		**1**	**2**	**1**	**2**						**1**	**7**

Fonte: Elaborado pelo autor.

No total, 7 publicações abordaram o tema *vigilância* com o mesmo conceito utilizado neste trabalho, e tendo o termo em questão presente nas referências e em pelo menos um outro campo de indexação.

No ano de 2014 foram publicados 3 artigos (que equivalem a 42% desta amostra), seguido do ano 2016 com duas publicações, e dos anos 2011 e 2015, com uma publicação cada.

Os GTs que mais contribuíram para a abordagem deste tema foram os GTs 3 e 5, com duas publicações cada, seguidos dos GTs 2, 4 e 11, com uma publicação cada.

Quadro 14 – Publicações com termo *privacidade* no ENANCIB, no contexto abordado nesta dissertação, nos últimos dez anos

	Grupo de Trabalho (GT)											
	1	2	3	4	5	6	7	8	9	10	11	**Total**
2007				1								**1**
2008												
2009												
2010												
2011					1							**1**
2012												
2013												
2014			1		2							**3**
2015								1				**1**
2016												
Total			**1**	**1**	**3**			**1**				**6**

Fonte: Elaborado pelo autor.

No total, 6 publicações abordaram o tema *privacidade* com o mesmo conceito utilizado neste trabalho, tendo o termo em questão presente no corpo do texto e em mais um campo de indexação e sendo discutido no âmbito do direito, de respaldo e preservação da intimidade a partir de dados pessoais.

No ano de 2014 foram publicados 3 artigos (que equivalem a 50% desta amostra), seguido dos anos 2007, 2011 e 2015, com uma publicação cada.

Os GT que mais contribuiu para a abordagem deste tema foi o GT 5, com três publicações, seguido dos GTs 3, 4 e 8, com uma publicação cada.

Vale a pena citar que dentre os trabalhos selecionados com o termo *vigilância* e aqueles com o termo *privacidade*, 3 artigos são comuns aos dois grupos.

6.3 OCORRÊNCIA DOS TERMOS *BIG DATA* E *DADOS PESSOAIS*

Para verificar a relação dos artigos selecionados com os termos *big data* e *dados pessoais*, foi feita uma busca simples na amostra dos grupos "Vigilância" (com 7 itens) e "Privacidade" (com 6 itens). Os resultados obtidos são apresentados abaixo:

Quadro 15 – Ocorrência dos termos *big data* e *dados pessoais* na amostra

Amostra		Ocorrência dos Termos	
Termo	Total de itens	Big Data	Dados Pessoais
Vigilância	7	1	0
Privacidade	6	2	1

Fonte: Elaborado pelo autor.

Foi verificado que, na amostra do termo *vigilância*, houve apenas uma ocorrência do termo *big data* e nenhuma ocorrência do termo *dados pessoais*. Enquanto que na amostra do termo *privacidade*, houve duas ocorrências do termo *big data* e uma ocorrência do termo *dados pessoais*.

7 ANÁLISE DOS RESULTADOS

O estudo realizado possibilita destacar inicialmente algumas observações feitas por Foucault (1987) que impressionam por sua acurácia e objetividade quase proféticas sobre a natureza do regime de vigilância, que é possível notar nos dias atuais. A partir da reprodução indireta da fala do autor, são feitos alguns comentários para apontar como a teoria tomou forma e hoje vem sendo observada, em termos práticos, na realidade.

Foucault chama a atenção para os dispositivos criados para descriminar os elementos da massa através da individualização, identificação e classificação. Esta seria uma prática de marcação binária que todos os mecanismos de poder, em nossos dias, utilizam para ter o domínio sobre o anormal.

Segundo Bauman (2001), a sociedade deu lugar a um campo social composto por agrupamentos de indivíduos. As políticas do governo, o cenário de mercado altamente competitivo e as NTIC causaram o processo de emancipação do indivíduo. Os sistemas de segurança e de marketing tornaram-se capazes de agregar, processar e analisar conjuntos massivos de dados pessoais, e criar categorias de objetos e demandas a serem observados e atendidos de acordo com uma política preestabelecida. A observação não é mais feita diretamente sobre os corpos, mas através dos dados que os representam em uma base digital (*dataveillance*). A identidade e o comportamento dos indivíduos são definidos e reconhecidos por padrões binários que determinam ações do sistema a esses indivíduos, seja no meio virtual ou real.

Outro aspecto observado por Foucault foi a leveza dos prédios e estruturas das instituições disciplinares compostos por uma geometria simples e econômica. Ao contrário das "casas de segurança", tudo seria feito de forma a garantir a objetividade de uma "casa de certeza". Assim, de forma gradual e contínua, o poder assumiu uma arquitetura que tende ao incorpóreo, destinada a se difundir no corpo social e se tornar uma função generalizada.

As estruturas do poder não são mais ostensivas, alicerçadas por concreto ou facilmente identificadas na paisagem urbana. Suas bases estão distribuídas em uma rede geograficamente dispersa de múltiplos agentes. O poder está difundido e generalizado em cada sistema, processo ou artefato tecnológico capaz de gerar dados sobre os indivíduos que os utilizam. A vigilância é abrangente e constante (tudo sobre todos), e cada ação pode ser objetivamente registrada, rastreada e analisada.

Foucault ainda comenta que o projeto de Bentham visava à sustentabilidade e manutenção de seus propósitos iniciais, deste modo, as instituições deveriam seguir regras de

transparência e se reportar à sociedade - a maior supervisora das instituições disciplinares ou o "grande comitê do tribunal do mundo".

Neste ponto, infelizmente o que existe hoje é uma distopia. Apesar das características do Estado Informacional (BRAMAN, 2013) e dos movimentos de transparência com a apresentação irrestrita dos resultados da administração pública, ainda existe uma discrepante assimetria informacional entre o governo e os cidadãos. A situação mais grave talvez esteja na relação entre os usuários e os produtores de tecnologia, onde as políticas de privacidade, os termos de uso e as leis de proteção de dados pessoais são insuficientes para garantir o direito à privacidade e à propriedade. As tecnologias estão fechadas como "caixas pretas" que ocultam os reais processos de compartilhamento, venda e uso dos dados pessoais e da propriedade intelectual de seus usuários.

A antiga sociedade do espetáculo – onde muitos olhavam poucos através da arquitetura monumental dos templos, teatros e circos – passou a ser a sociedade moderna da física panóptica, com seus dispositivos que levaram a uma "distribuição infinitesimal do poder". Ao mesmo tempo que os dispositivos disciplinares crescem em número, também aumentam seu poder de penetração através de uma estrutura distribuída, ramificada e interligada.

Neste ponto, Foucault parece descrever o mecanismo da grande rede, onde os grandes produtores de conteúdo, que antes determinavam o padrão cultural a ser consumidos pelas massas, hoje passam a perder força diante do surgimento de outras mídias. Os canais de comunicação se multiplicaram e as nTIC permitiram múltiplos benefícios aos usuários, desde o acesso individual ao conteúdo, a formação de comunidades de interesse, a interação, até a produção de conteúdo independente. Contudo, o conceito de vigilância distribuída nos faz lembrar que a rede também é o ambiente onde todos podem vigiar todos. O excessivo compartilhamento de informações pessoais motivados pela recompensa inestimável da atenção de uma audiência nunca está isento do crivo de um sistema colaborativo de vigilância compartilhada, onde todos observam todos em prol da moral e dos bons costumes. Os dispositivos disciplinares assumiram inúmeras formas e hoje podem ser representados por uma extensa lista de objetos insuspeitos, agentes, processos e tecnologias que fazem parte deste complexo cenário ativo com a capacidade de recolher informações pessoais, do qual todo nós somos usuários, participantes, vítimas e colaboradores.

Com base nas reflexões apresentadas, foi elaborado um quadro-síntese comparativo das diversas modalidades de regime de disciplina, controle e vigilância, e os mecanismos e mediações que definem a materialidade do objeto observado e da informação gerenciada. A

seguir o Quadro 16: Tipos de regime x objetos de informação, mediações e visibilidade do poder.

Quadro 16 – Tipos de regime x objetos de informação, mediações e visibilidade do poder

Tipo de Regime	**Localização**	**Objetos de Informação**	**Mediações**	**Características do Poder**
Regime de exceção	Final do século XVII e século XVIII. Cidades europeias assoladas pelas epidemias, como a peste e a lepra.	Os corpos dos indivíduos (informação como objeto), registros feitos pelos agentes de inspeção.	Observação direta, sistema de inspeção sensorial, sistema de registro auxiliar.	Direto e hierarquizado, visível, verificável e determinado.
Regime disciplinar	Final do século XVIII e século XIX. Instituições disciplinares: hospitais, escolas, fábricas, sanatórios, penitenciárias.	Os corpos dos indivíduos, produção (informação como objeto), registros feitos pelas autoridades disciplinares.	Observação direta, posto de observação, documentos auxiliares (prontuários médicos, pautas de presença de classe, folhas de ponto), testes escritos, instrumentos de contagem, sistema de registro auxiliar, etc.	Direto e hierarquizado, visível, verificável/inverificável e determinado.
Regime de vigilância total: vigilância líquida, vigilância distribuída, *panspectron*	Século XX e início do século XXI. Sociedade da informação, Estado informacional, ambiente virtual, redes sociais e sistemas *online.*	Informação com *input* direto em meio digital, informações advindas de dispositivos sensoriais, dados que representam o sujeito em múltiplas bases de dados.	Sistemas de processamento e análise de dados, dispositivos para o reconhecimento de padrões (óticos, de som, sequências binárias, etc.), algoritmos, bancos de dados, *data centers*, *datawarehouse*, *big data.*	Indireto, invisível, inverificável e indeterminado.

Fonte: Elaborado pelo autor.

Conforme pode ser verificado no Quadro 1, a informação perde materialidade à medida que o agente observador se distancia do objeto observado; ao mesmo tempo, múltiplos recursos surgem como mediadores e operadores da informação. O resultado desse processo é o aumento da abrangência e da ubiquidade do poder, com a diminuição da sua visibilidade e determinação. A capilaridade do poder de vigilância e sua abrangência exigem o desdobramento hierárquico, mas as tecnologias informacionais vêm aumentar o poder informacional em sua capacidade de gestão, dispensando o emprego de grandes contingentes de pessoas/funcionários para o mesmo efeito. Os artefatos tecnológicos de vigilância e

segurança são ferramentas que amplificam o poder dos gestores, dando-lhes amplas vantagens de monitoramento e controle sobre os grupos e indivíduos a serem observados.

Os grandes produtores de conteúdo, que antes determinavam o padrão cultural a ser consumido pelas massas, hoje passam a perder força diante do surgimento de outras mídias. Os canais de comunicação se multiplicaram e as NTIC permitiram múltiplos benefícios aos usuários, desde o acesso individual ao conteúdo, à formação de comunidades de interesse, a interação, até a produção de conteúdo independente.

Contudo, o conceito de vigilância distribuída nos faz lembrar que a rede também é o ambiente onde todos podem vigiar todos. O excessivo compartilhamento de informações pessoais, motivado pela recompensa da atenção de uma audiência, nunca está isento do crivo de um sistema colaborativo de vigilância compartilhada, onde todos observam todos em prol da moral e dos bons costumes. Os dispositivos disciplinares assumiram inúmeras novas formas, e hoje podem ser representados por uma extensa lista de objetos insuspeitos, agentes, processos e tecnologias que fazem parte desse complexo cenário ativo, com a capacidade de recolher informações pessoais, do qual todos nós somos, voluntária ou involuntariamente, usuários, participantes, colaboradores, e/ou vítimas.

Um dos aspectos que caracterizam o presente estudo é a ênfase sobre o ambiente informacional no qual as questões são analisadas. Fez-se aqui uma breve referência histórica ao termo cunhado por Vannevar Bush ao final da década de 1945, propondo, com óbvia menor importância, mas que não poderia deixar de ser citada, a nova ou "segunda explosão da informação".

A partir do volume, velocidade e variedade dos dados que excedem a capacidade tecnológica padrão-comum instalada, a dimensão do Big Data traz um novo paradigma de pesquisa a ser considerado, menos pautado em conclusões definitivas ou explicações sobre as leis de causa e efeito dos fenômenos, mas que busca e se realiza com a simples identificação de padrões. O tempo exigido para as respostas também foi redefinido, em muitos casos, tornando-se quase nulo, ou simultâneo às próprias ações geradoras dos dados.

Além dos aspectos físicos, foi observado que o termo Big Data traz também a conotação de um movimento coletivo impulsionado pelas empresas de tecnologia, e a crença de que o domínio sobre os conjuntos massivos de dados permite evidenciar padrões, fazer predições e trazer *insights* até então impossíveis de alcançar.

O processo de captação de dados teve central importância na presente exposição e foi amplamente abordado, citando 1) a web 2.0 com suas características de participação e emancipação do usuário final que passou de mero leitor/espectador para gerador de conteúdo,

2) a popularização das redes sociais e aplicativos de mobilização coletiva que fizeram da exposição pessoal um potencial a ser buscado em troca da audiência, que incentiva a participação do indivíduo com o compartilhamento de seus dados pessoais, 3) a Internet das Coisas e o advento de aparatos tecnológicos dotados de sensores capazes de aferir as condições do seu ambiente e 4) os acessórios pessoais tecnológicos (*wearables technologies*), e outros dispositivos, que permitem a Dataficação ou conversão direta dos movimentos e ações em dados.

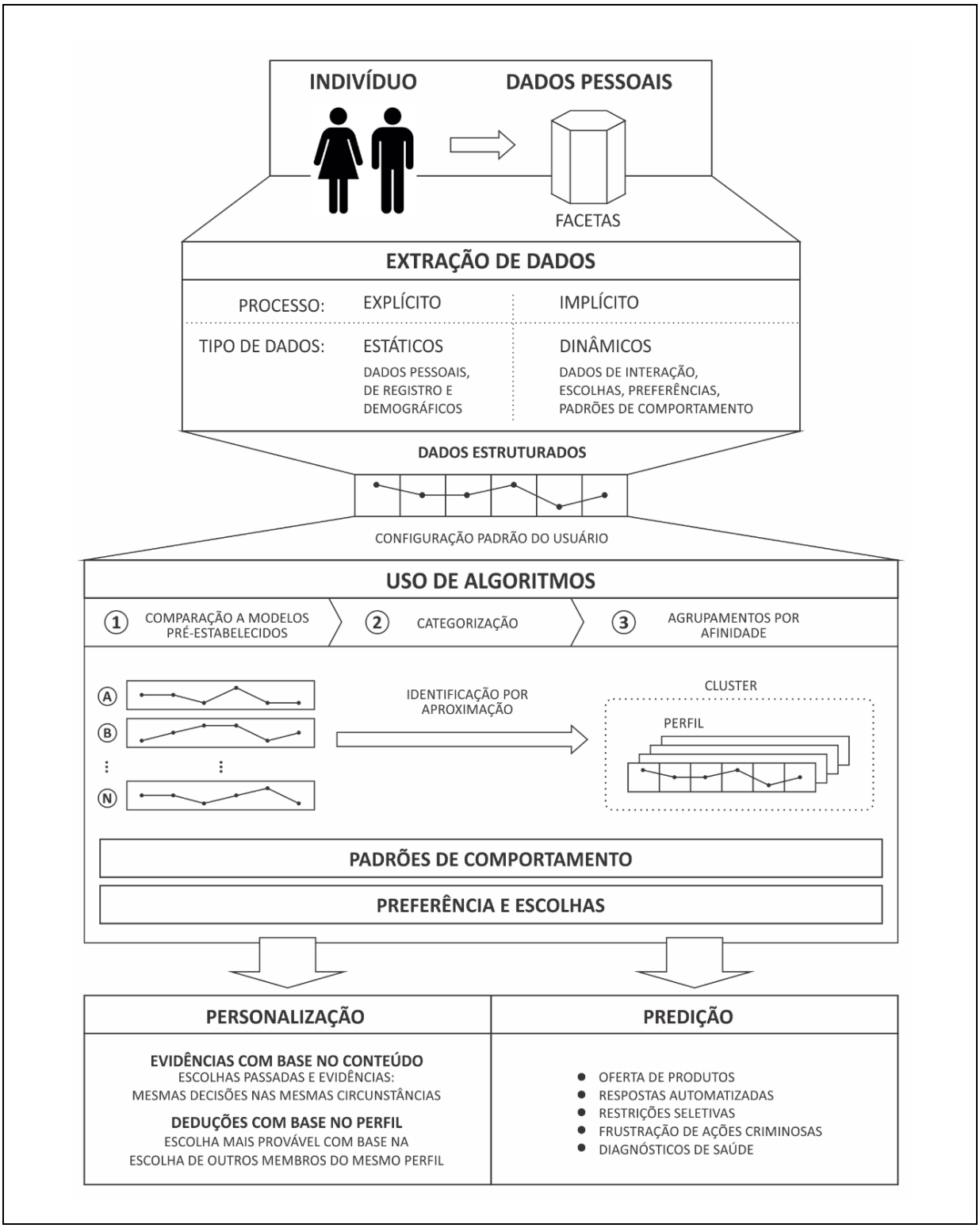
A capacidade de registrar dados em larga escala permite uma simulação cada vez mais fiel da realidade física em uma instância informacional que pode ser processada, analisada, distorcida ou reconfigurada. Foi percebido que o amplo uso de sistemas de informação e a grande dependência dos aparatos tecnológicos, seja no âmbito profissional ou pessoal, dissolveram por completo as fronteiras entre o real e o virtual, fazendo com que o indivíduo permaneça conectado à grande rede, perfazendo-se em várias nuances ou instâncias de convívio da realidade, podendo ser influenciado diretamente pelas respostas feitas sobre a análise dos seus dados.

No contexto contemporâneo da organização social e informacional, foi enfatizado que os dispositivos disciplinares deixaram de intervir diretamente sobre os corpos dos indivíduos, antes, feitas a partir do que Foucault chamou de *Física do Poder* - estruturas formatadoras e delimitadoras do comportamento, ordenadoras da massa de indivíduos, restringidoras dos atritos, do contágio, dos rompantes e das insurreições, compressora dos corpos, definidoras do tempo e do espaço - e passaram a atuar sobre o duplo informacional, ou os dados que representam os indivíduos. A vigilância dos dados (*dataveillance*) ocorre, portanto, de forma difusa, alheia à atenção do indivíduo e independente dos meios físicos, pois não exigem proximidade ao objeto observado ou a necessidade de verificação sensorial para atuar ou provocar seus efeitos. O indivíduo, na pós-modernidade, também se tornou fragmentado e disperso, com suas várias instancias informacionais, mais ou menos representativas de si, registradas em múltiplos bancos de dados, sujeitas à vigilância e à influência das respostas automáticas que os sistemas inferem sobre elas.

Foram apresentadas as diferenças sobre os meios explícitos e implícitos de captação de dados estáticos e dinâmicos. Este processo, com base em plataformas de inteligência artificial e no uso de algoritmos de mineração de dados, permite a criação de modelos, o reconhecimento de padrões, a classificação, a formação de *clusters*, a predição e a intervenção sobre os indivíduos a partir de seus dados.

Com base em Berry e Linoff (2004), Pridmore (2006) e Cufolgu (2014), foi elaborado o infográfico (Figura 9) que sintetiza, em termos visuais, o processo de extração de dados pessoais com objetivo de personalização e predição.

Figura 9 - Processo de Extração de Dados Pessoais para Personalização e Predição



Fonte: Elaborado pelo autor.

Em primeira instância, tem-se a conversão ou tradução do indivíduo em dados, representados no infográfico pelos ícones de um homem e uma mulher seguidos de um prisma de seis lados. O prisma representa a simplificação ou objetivação do indivíduo convertido em uma tabela de dados com suas as dimensões ou facetas. O número de faces é relativo à quantidade de variáveis, características ou pontos de medição, considerados para o indivíduo, dependendo da organização de cada banco de dados ou dos propósitos da extração. Neste caso, portanto, cadastros mais extensos ou um conjunto de dados mais complexos seriam representados por um prisma com maior número de faces.

A extração de dados pode se dar de forma explícita ou implícita, dependendo da transparência do processo para o usuário, que pode ou não ter noção que seus dados estão sendo gravados a partir da interação com o sistema.

Após a extração dos dados pessoais, de forma explícita ou não, o indivíduo é simplificado e convertido em um conjunto de dados estruturados, representado na figura pelo prisma aberto, com uma barra de seis *slots* (lacunas), onde, em cada faceta, uma característica do indivíduo é quantificada e definida.

Com a estruturação dos dados, as características do indivíduo (sequenciadas e quantificadas), passam a ser parâmetros que podem ser processados e interpretados por algoritmos que fazem 1) o reconhecimento de padrões, 2) classificação e a 3) aproximação por afinidade.

No primeiro passo estão os modelos - definições, arbitrárias ou não, feitas pelos gestores do banco de dados -, onde alguns padrões característicos são estabelecidos como parâmetros de comparação para a classificação dos indivíduos.

Os *clusters* são agrupamentos de objetos, feitos por critérios de aproximação, dentro do espaço informacional, enquanto que o *perfil* é a nomenclatura que define o padrão de um grupo de indivíduos. Alguns perfis podem ser observados no sistema Mosaic da SERASA Experian[35], e trazidos aqui como exemplo: aspirantes sociais, periferia jovem, empreendedores e comerciantes.

Considerando principalmente os dados demográficos, as preferências e escolha, e o padrão de comportamento do indivíduo, é possível a personalização de produtos, a configuração de interfaces ou de ambientes virtuais, e as respostas automáticas com a distribuição seletiva de informação. A predição é um ponto importante a se destacar no processo de Marketing, da Segurança e da Saúde, para a antecipação de situações permitindo

[35] https://www.serasaexperian.com.br/mosaic/segmentacao.html

a preparação de respostas para a variação de mercado, a frustração de ações criminosas ou o diagnóstico precoce de doenças.

Outro aspecto a ser ressaltado sobre este estudo, foi a sua intenção de expor e exemplificar os métodos de identificação de padrões, classificação e formação de *clusters* a partir da tecnologia de mineração de dados e da representação vetorial, demonstrando que os métodos utilizados na indexação e organização de documentos são os mesmos utilizados para a gestão de uma base de dados pessoais. Ou seja, o indivíduo é visto nesse contexto como um objeto de múltiplas facetas, dimensões ou características que podem ser estratificadas, estruturadas, quantificadas, processadas e analisadas.

A partir da interação do indivíduo com sistemas operacionais, informacionais ou virtuais, foi ressaltado o processo de repostas individualizadas, distribuição seletiva da informação ou personalização em função das informações pessoais, preferências e comportamento do usuário. Um aspecto fundamental desse processo está no uso de algoritmos de aprendizado de máquina (*machine learning*) para o reconhecimento dos padrões e atualização constante do sistema de ponderação e cálculo das respostas, em função da interação do usuário. Para melhor entendimento, foram trazidos exemplos de dois dos maiores sistemas de recomendação e seleção de conteúdo da atualidade (Netflix e Facebook) apresentando seus recursos de personalização, critérios de seleção e como essas propriedades podem ser utilizadas para influenciar as escolhas do usuário, seja na compra de um produto, escolha de um filme ou na decisão do voto.

Com a descrição dos recursos de interação e de publicação disponibilizados por essas redes sociais, é possível notar a existência de métodos para a captação (ou dataficação) do sentimento, estado de ânimo ou reação emocional dos usuários. Neste sentido, percebe-se que, aos poucos, algo tão subjetivo e intangível quanto o sentimento do ser humano, está se tornando objeto quantificável, na forma de dados estruturados que podem ser relacionados a outros fatores para maior conhecimento sobre os padrões de comportamento, escolhas e preferências dos usuários.

Foi ressaltado que, todo processo de personalização pode ser tornar excessivo pela ênfase recursiva e gerar bolhas ideológicas, câmaras de eco ou distorções na percepção da realidade a partir de uma exposição altamente seletiva e parcial dos fatos.

Outro fenômeno semelhante, embora mais nocivo, está na capacidade de gestão dos dados pessoais, de forma massiva, para a seleção, discriminação ou exclusão de indivíduos que, em sua maioria, feitas de forma automática por algoritmos de análise de cadastro, podem

agravar a situação de desigualdade social ao reforçar um padrão situacional desfavorável para um indivíduo (*ban-opticon*).

A respeito da investigação feita sobre como a *vigilância* e a *privacidade* vêm sendo abordadas nas pesquisas da Ciência da Informação, além dos artigos citados na seção 10, pode-se destacar também o estudo de Bembem, Santana e Da Costa Santos (2015) que traz uma interessante análise sobre a ocorrência do termo *privacy* nas publicações do Journal of the Association for Information Science and Technology (JASIST) nos anos 2013 e 2014, e também o número 2, volume 12 da LIINC em Revista, que traz treze artigos exclusivamente sobre questões relacionadas ao tema da "Privacidade e Vigilância nos Meios Digitais".

Uma observação final é que, no decorrer do estudo, foi percebido que as políticas de uso dos dados pessoais declaradas pelas empresas de tecnologias ou de serviços não definem claramente o uso que será feito dos dados coletados dos usuários. Em sua maioria, fazem ressalvas com termos extremamente evasivos que dão margem a um amplo espectro de interpretações e sugerem a prática de compartilhamento dos dados com parceiros comerciais ou tecnológicos, o que contribui para agravar os problemas.

8 CONCLUSÃO

Entende-se que o presente estudo atingiu seus objetivos, por ter apresentado substancial conteúdo sobre as questões propostas.

Com base em fontes bibliográficas impressas e diversas referências atuais disponíveis no meio *online,* a temática da vigilância foi abordada de forma ampla, contudo, tendo foco nas mediações tecnológicas para a captação e análise de dados pessoais.

Inicialmente foi abordado o processo histórico da tecnologia, ressaltando os meios de armazenamento e nos métodos de gestão, que permitiram o domínio progressivo sobre a contingência da *overdose* de informação.

Em seguida, com base em Foucault, foi feita uma exposição (com posterior análise) sobre a inversão da visibilidade do poder e a transição das mediações e dos aspectos materiais da informação, desde o cerne do período moderno até os dias atuais, onde os corpos passaram a ser vigiados através de seus duplos informacionais (*data-double*).

A terceira parte do estudo enfatizou o processo metodológico de captação, tratamento e classificação de dados pessoais, com o uso de algoritmos sofisticados capazes de fazer o reconhecimento de padrões e predições de comportamento. Foram apresentados em linhas gerais o funcionamento dos algoritmos de recomendação e seleção de notícias do Netflix e Facebook, e também alguns "efeitos colaterais" da distribuição seletiva da informação, com a ocorrência de bolhas ideológicas, reforço de condições socialmente desfavoráveis (*ban-opticon*) e distorções na percepção da realidade.

Com o propósito de identificar como os temas *vigilância* e *privacidade*, associados aos termos *big data* e dados pessoais vêm sendo abordados nas pesquisas da Ciência da Informação, no Brasil, foi feita uma pesquisa nos Anais do ENANCIB dos últimos dez anos, verificando que, apesar da abrangência do tema e do grande reflexo das questões envolvidas no cotidiano dos ambientes informatizados, a produção de artigos com este foco ainda é incipiente na pós-graduação em Ciência da Informação no Brasil.

Acredita-se que esta pesquisa, por sua ênfase na tecnologia aplicada à vigilância de massa, de modo abrangente e atual, ofereça contribuições significativas tanto no campo teórico quanto prático. Através de uma abordagem multidisciplinar, o estudo reuniu aspectos socioculturais (pessoas), legais (políticas e princípios) e funcionais (processos), apresentando-os de forma sequencial e didática, de modo a permitir ampla visão de conjunto sobre os regimes de informação existentes, voltados para a vigilância e controle dos indivíduos, a partir de seus dados pessoais.

O tema em questão (vigilância), para ser compreendido de forma satisfatória, exige certa amplitude no campo de observação. Tal abrangência do assunto reforça, por um lado, o aspecto multidisciplinar da Ciência da Informação, mas, por outro (questões de tempo e ferramental metodológico) exige um recorte teórico. Neste sentido, para o estudo da vigilância de dados pessoais no campo da segurança pública - o que inclui os serviços de inteligência para a contenção de ações criminosas e terroristas – o escopo ficou limitado ao conteúdo publicado em periódicos e nos meios de comunicação devido à natureza sigilosa das suas práticas.

Outros assuntos relevantes para área da Ciência da Informação foram observados durante o curso da pesquisa, mas por razões de delimitação do seu escopo e abrangência, não puderam ser aprofundados. Faz-se aqui, portanto, a menção a esses tópicos como indicação para projetos futuros, entre eles: a carência de legislação específica para a proteção e garantias da propriedade de dados pessoais; a ênfase sobre a ética na captação e uso dos dados pessoais; e a prática da associação (*assemblage*) de empresas no ramo da tecnologia para a exploração do uso dos dados pessoais.

Espera-se que o presente estudo possa contribuir para o avanço da Ciência da Informação enquanto área de pesquisa, por trazer conteúdo atual e questões emergentes que afetam o cotidiano dos indivíduos que necessitam fazer uso da tecnologia da informação, a partir da entrega do conteúdo e esclarecimento das questões que se propôs a investigar, servindo assim como fonte de informação, referência ou, ainda, incentivo para outros pesquisadores desenvolverem trabalhos sobre o mesmo tema.

9 REFERÊNCIAS

ADREJEVIC, M.; GATES, K. Big data surveillance: introduction. **Surveillance & Society**, v. 12, n. 2, p. 185-196, 2014.

AFFONSO, E. P. A.; SANT'ANA, R. C. G. Anonimização de Metadados de Imagem Digital por maio do Modelo K-Anonimato. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 16., 2015, João Pessoa. **Anais**... João Pessoa: UFPB 2015.

ANDREWS, W.; LINDEMAN, T. The black budget: funding the intelligence program. **The Washington Post**, 29 ago. 2013. National. Online. Disponível em: <http://www.washingtonpost.com/wp-srv/special/national/black-budget/>. Acesso em: 13 jun. 2015.

ANTONIUTTI, C. L.; ALBAGLI, S. Uso do big data em campanhas políticas eleitorais. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 15., 2014, Belo Horizonte. **Anais**... Belo Horizonte: UFMG, 2014.

BAUMAN, Z. **Modernidade líquida**. Tradução Plínio Dentzien. Rio de Janeiro: Zahar, 2001.

_____. **Vigilância líquida**: diálogos com David Lyon. Tradução Carlos Alberto Medeiros. Rio de Janeiro: Zahar, 2013.

BBC. Quanto dinheiro o Facebook ganha com você (e como isso acontece). **G1**, 10 nov. 2016. Tecnologia e Games. Online. Disponível em: <http://g1.globo.com/tecnologia/noticia/2016/11/quanto-dinheiro-o-facebook-ganha-com-voce-e-como-isso-acontece.html>. Acesso em: 11 fev. 2017.

BENTHAM, J. **Panopticon**: postscript; Part I - containing further particulars and alterations relative to the plan of construction originally proposed, principally adapted to the purpose of a panopticon penitentiary-house. London: T. Payne, 1791.

_____. **Panopticon**: postscript; Part II - containing a plan of management for a panopticon penitentiary-house. London: T. Payne, 1791.

_____. **Panopticon**: or the inspection-house. London: T. Payne, 1791.

BERRY, M. J. A., LINOFF, G. S. **Data mining techniques**: for marketing, sales and customer relationship management. Indianápolis: Wiley Publishing Inc, 2004.

BEZERRA, A. C. “Culturas de vigilância”, “regimes de visibilidade”: novos caminhos para a pesquisa em ciência da informação. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 15., 2014, Belo Horizonte. **Anais**... Belo Horizonte: UFMG, 2014.

_____. Vigilância e Filtragem de Conteúdo das Redes Digitais - desafios para a competência crítica em informação. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 16., 2015, João Pessoa. **Anais**... João Pessoa: UFPB 2015.

BEZERRA, A. C.; PIMENTA, R. M.; ORMAY, Larissa Santiago. Vigilância, vigilância inversa e democracia: do panoptismo ao midiativismo. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 15., 2014, Belo Horizonte. **Anais**... Belo Horizonte: UFMG, 2014.

BIGO, D. Introduction-globalized (in)security: the field and the ban-opticon. In: BIGO, D. **Illiberal practices of liberal regimes**: the (in)security games. L'Harmattan, 2006.

_____. Security, exception, ban and surveillance. In: LYON, D. (Org.). **Theorizing surveillance**: the panopticon and beyond. New York: Routledge, 2011. p. 46-68.

BOOCH, G. The human and ethical aspects of big data. **IEEE Computer Society**, jan./fev. 2014. Online. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=6750430>. Acesso em: 11 jun. 2015.

BORKO, H. Toward a theory of indexing. **Information Processing & Management**, v. 13, p. 355-365, 1977.

BOYD, D.; CRAWFORD, K. Critical questions for Big Data. **Information, Communication & Society**. v. 15, n. 5, p. 662-679, 2014.

BRAMAN, S. Information policy and power in the information state. In: BRAMAN, S. **Change of state**. Massachusetts: MIT Press, 2006a. p. 313-328.

_____. Tactical memory: the politics of openness in the construction of memory. **Fisrt Monday**, v. 11, n. 7, 2006b. Disponível em: <http://firstmonday.org/ojs/index.php/fm/article/view/1363>. Acesso em: 01 dez. 2015.

BRASIL. **Constituição da República Federativa do Brasil**. Brasília, Secretaria Especial de Informática, 1988. Senado Federal. Disponível em: <http://www.senado.gov.br/legislacao/const/con1988/con1988_05.10.1988/con1988.pdf>. Acesso em: 23 jun. 2015.

BRUNO, F. **Máquinas de ver, modos de ser**: vigilância, tecnologia e subjetividade. Porto Alegre: Sulina, 2013.

BURRUS, D. The internet of things is far bigger than anyone realizes. **Wired**, nov. 2014. Online. Disponível em:< http://www.wired.com/2014/11/the-internet-of-things-bigger/>. Acesso em: 05 jun. 2015.

CAVALCANTE, R. B; PINHEIRO, M. M. K. Sistema de informação da atenção básica: relações de poder, centralização e vigilância. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 12., 2011, Brasília, **Anais**... Brasília: UnB, 2011.

CIANCONI, R, B. **Gestão do conhecimento**: visões de indivíduos e organizações no Brasil. Tese de doutorado. Instituto Brasileiro de Informação em Ciência e Tecnologia - IBICT, Universidade Federal do Rio de Janeiro - UFRJ, Rio de Janeiro, 2003.

CLEVELAND, Donald B.; CLEVELAND, Ana D. **Introduction to indexing and abstracting.** 4ª Ed. Santa Barbara: ABC-Clio, 2013.

COMPARE BUSINESS PRODUCTS. Top 10 Largest Databases in the World. **Compare Business Products**, 17 mar. 2010. Online. Disponível em: <http://www.comparebusinessproducts.com/fyi/10-largest-databases-in-the-world>. Acesso em: 12 jun. 2015.

CONSTINE, J. How Facebook news feed works. **Techcrunch**, 6 sep. 2016. Online. Disponível em <https://techcrunch.com/2016/09/06/ultimate-guide-to-the-news-feed>. Acesso em: 21 jan. 2017.

COX, M.; ELLSWORTH, D. Application-controlled demand paging for out-of-core visualization. **NASA**, 1997. Disponível em: <http://www.nas.nasa.gov/assets/pdf/techreports/1997/nas-97-010.pdf>. Acesso em: 10 jun. 2015.

CRUZ, A. J. O. **Algoritmos**. Apostila. 1997. Disponível em: <http://equipe.nce.ufrj.br/adriano/c/apostila/algoritmos.htm>. Acesso em: 8 jan. 2017.

CUFOGLU, A. User profiling: a short review. **International Journal of Computer Applications**, v. 108, n. 3, dec. 2014.

CUKIER, K. Data, data everywhere. **The Economist**, 25 fev. 2010. Special report: managing information. Online. Disponível em: <http://www.economist.com/node/15557443>. Acesso em: 15 jun. 2015.

CULTURA DIGITAL. Marco civil da internet entra em vigor. **Cultura Digital**, 23 jun. 2014. Online. Disponível em:< http://culturadigital.br/marcocivil/>. Acesso em: 18 jun. 2015.

DAVENPORT, T. **Ecologia da informação**: por que só a tecnologia não basta para o sucesso na era da informação. Bernadette Siqueira Abrão. São Paulo: Futura, 2002.

DE BELLIS, N. **Bibliometrics and citation analysis**: from the science citation index to cybermetrics. Lanham: Scarecrow Press, 2009.

DELEUZE, Gilles. **Conversações**: 1972 – 1990. Rio de Janeiro: Ed. 34, 1992.

DUHIGG, C. How companies learn your secrets. **New York Times**, 19 fev. 2012. Magazine. Online. Disponível em: <http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html>. Acesso em: 18 jun. 2015.

DUMBILL, E. What is big data?: an introduction to the big data landscape. **O'Reilly**, 2012. Online. Disponível em: < https://beta.oreilly.com/ideas/what-is-big-data>. Acesso em: 10 jun. 2015.

ELLIOTT, K.; RUPAR, T. Six months of revelations on NSA. **The Washington Post**, 23 dez. 2013. National. Online. Disponível em:< http://www.washingtonpost.com/wp-srv/special/national/nsa-timeline/>. Acesso em: 10 jun. 2015.

FACELI, K. et al. **Inteligência artificial**: uma abordagem de aprendizado de máquina. Rio de Janeiro: Editora LTC, 2011.

FOUCAULT, M. **Vigiar e punir**: nascimento da prisão. 20.ed. Tradução Raquel Ramalhete. Petrópolis: Editora Vozes, 1987.

FOUCAULT, M. **Microfísica do poder**. Organização, Introdução e Revisão Técnica de Roberto Machado. Rio de Janeiro: Graal, 1990.

FROHMANN, B. Taking information policy beyond infomation Science: Applying the actor network theory. In: H. A. Olson, & D. B. Wards (Eds.) **Proceedings of the 23rd Annual Information Science**, 7 – 10 June 1995, Edmonton, Alberta.

GOMEZ-URIBE, C. A.; HUNT, N. The Netflix recommender system: algortithms, business value and innovation. **ACM Transactions on Management Information Systems**, v. 6, n. 4, dec. 2015.

GONZÁLEZ DE GÓMEZ, M. N. O caráter seletivo das ações de informação. **Informare**: Cadernos do Programa de Pós-Graduação em Ciência da Informação, v. 5, n. 2, p. 7-31, 1999.

_____. Regime de informação: construção de um conceito**. Informação e Sociedade**: Estudos, João Pessoa, v. 22, n. 3, p. 43-60, set./dez. 2012.

_____. Políticas e regimes de informação. In: GARCIA, J. C. R, TARGINO, M. G. (Org.). **Desvendando facetas da gestão e políticas de informação**. 2. ed. João Pessoa: Editora da UFPB, 2015. p. 321-351.

GOULART, G.; SERAFIM, V. Especial marco civil da internet. **Segurança Legal**, 11 abr. 2014. Episódio 47. Podcast. Online. Disponível em: <http://www.segurancalegal.com/2014/04/episodio-47-especial-marco-civil-da.html>. Acesso em: 02 jun. 2015.

_____. Internet das Coisas. **Segurança Legal**, 10 out. 2014. Episódio 60. Podcast. Online. Disponível em: <http://www.segurancalegal.com/2014/10/episodio-60-internet-das-coisas.html>. Acesso em: 02 jun. 2015.

_____. Ética e Tecnologia. **Segurança Legal**, 10 abr. 2015. Episódio 73. Podcast. Online. Disponível em: <http://www.segurancalegal.com/2015/04/episodio-73-etica-e-tecnologia.html>. Acesso em: 02 jun. 2015.

GOVERNOR, J.; NICKULL, D.; HINCHCLIFFE, D. Web 2.0 Architectures. **O'Reilly Media**, 2009. Dissecting Web 2.0 Examples. Online. Disponível em: <http://archive.oreilly.com/pub/a/web2/excerpts/web2-architectures/chapter-3.html>. Acesso em: 10 jun. 2015.

GRABIANOWSKI, E. How the patriot act works. **How Stuff Works**, 2007. Culture. Online. Disponível em: <http://people.howstuffworks.com/patriot-act.htm>. Acesso em: 01 mar. 2016.

GRIFFITH, E. What is cloud computing?. **PC Magazine**, 17 abr. 2015. Review. Online. Disponível em: <http://www.pcmag.com/article2/0,2817,2372163,00.asp>. Acesso em: 13 jun. 2015.

GUIMARÃES, J. A. C. A dimensão teórica do tratamento temático da informação e suas interlocuções com o universo científico da International Society for Knowledge Organization (ISKO) **Revista Ibero-americana de Ciência da Informação (RICI)**, v.1 n.1, p.77-99, jan./jun. 2008.

HAGGERTY, K. D.; ERICSON, R. V. The surveillant assemblage. **The British Journal of Sociology**, v. 51, p. 605-622, 2000.

HAYS, C. L. What Wal-Mart knows about customers' habits. **New York Times**, 14 nov. 2004. Business. Online. Disponível em: <http://www.nytimes.com/2004/11/14/business/yourmoney/14wal.html>. Acesso em: 18 jun. 2015.

HOOKWAY, B. **Pandemonium**: The rise of predatory locales in the postwar world. Princeton: Princeton Architectural Press, 2000.

HORTON, F. W. **How to harness information resources**: a systems approach. Cleveland: Association for Systems Management, 1974.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA - IBGE. **Metadados.Online**: IBGE. Disponível em: <http://www.metadados.ibge.gov.br/>. Acesso em: 21 jun. 2015.

ISONI, Miguel Maurício; VIDOTTI, Silvana Aparecida Borsetti Gregorio. Percepções de segurança e ameaças em ambientes de tecnologia da informação. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 8., 2007, Salvador. **Anais**... Salvador: UFBA, 2007.

KARVALICS, L. Z. Information society dimentions. **Academy.edu**, 2009. Disponível em: <https://www.academia.edu/652113/Information_Society_Dimensions>. Acesso em: 07 ago. 2015.

KEEGAN, J. Blue feed, red feed: see liberal Facebook and conservative Facebook, side by side. **The Wall Street Journal**, 2016. Online. Disponível em <http://graphics.wsj.com/blue-feed-red-feed/>. Acesso em: 25 fev. 2017.

KIRKPATRICK, M. What a tweet can tell you. **Read Write**, 16 nov. 2011. Web. Online. Disponível em: <http://readwrite.com/2011/11/17/what_a_tweet_can_tell_you>. Acesso em: 23 jun. 2015.

KRAMERA, A. D. I.; GUILLORYB, J. E.; HANCOCKB, J. T. Experimental evidence of massive-scale emotional contagion through social networks. **Proceedings of the National Academy of Science - PNAS**, v. 111, n. 24, p. 8788-8790, jun. 2014.

LANCASTER, F. W. **Indexação e Resumos:** teoria e prática. Trad. Antonio Agenor Briquet de Lemos. 2ª Ed. Brasília, DF: Briquet de Lemos, 2004.

LAPOWSKY, I. Here's how Facebook actually won Trump the presidency. **Wired**, 2016. Business. Online. Disponível em <https://www.wired.com/2016/11/facebook-won-trump-election-not-just-fake-news/>. Acesso em: 12 fev. 2017.

LIBRARY OF CONGRESS. Facinating facts. **Library of Congress**. About the Library. Online. Disponível em: <https://www.loc.gov/about/fascinating-facts/>. Acesso em: 10 jun. 2015.

LICKLIDER, J. C. R. Man-computer symbiosis. Reimpresso e publicado online com a permissão de IRE Transactions on Human Factors in Electronics, v. HFE-1, p. 4-11, mar. 1960.

LOMAS, N. Samsung edits orwellian clause out of TV privacy policy. **Tech Crunch**, 10 fev. 2015. Popular Posts. Disponível em: <http://techcrunch.com/2015/02/10/smarttv-privacy/>. Acesso em: 25 jun. 2015.

LYMAN, P.; VARIAN, H. R. How much information. **University of California at Berkley**, 2010. Online. Disponível em: <http://www.sims.berkeley.edu/research/projects/how-much-info/>. Acesso em: 16 jun. 2015.

LYON, D. **Surveillance studies**: an overview. Cambridge: Polity Press, 2007.

MAI, Jens-Erik. Analysis in indexing: document and domain centered approaches. **Information Processing and Management**, v. 41, n. 3, p. 599-611, 2005.

MARQUES, Rodrigo Moreno; PINHEIRO, Marta Macedo Kerr. Assimetria de informação na Lei Geral de Telecomunicações: uma análise dialética. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 11., 2010, Rio de Janeiro. **Anais**... Rio de Janeiro: IBICT, 2010.

MARTINS, S. C.; CIANCONI, R. B. Gestão da informação: estudo comparativo de modelos sob a perspectiva integrativa dos recursos de informação. In: Encontro Nacional de Pesquisa em Ciência da Informação, 14, 2013, Florianópolis. **Anais**... Rio de Janeiro: IBICT, 2013. Disponível em: <http:// http://enancib.sites.ufsc.br/index.php/enancib2013/XIVenancib/paper/view/362/>. Acesso em: 10 jun. 2016.

MICROSOFT. Política de privacidade do Windows 8 e do Windows Server 2012. **Microsoft**, 2012. Online. Disponível em: <http://windows.microsoft.com/pt-BR/windows-8/windows-8-privacy-statement#T1=statement>. Acesso em: 25 jun. 2015.

MORAIS, E. A. M.; AMBRÓSIO, A. P. L. Mineração de textos. **Relatório Técnico**, 2007. Instituto de Informática da Universidade de Goiás - UFG, Online. Disponível em: <http://www.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_005-07.pdf>. Acesso em: 27 nov. 2016.

MOURA, Maria Apareceida. Decifra-me ou Devoro-te: Contexto, Similaridade Semântica e Terminologia Especializada em Serviços de Inteligência no Brasil. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 17., 2016, Salvador. **Anais**... Salvador: UFBA 2016.

NIELSEN GROUP. Número de pessoas com acesso à internet no Brasil supera 120 milhões. **Nielsen Group**, 30 jul. 2014. Online. Disponível em: <http://www.nielsen.com/br/pt/press-room/2014/Numero-de-pessoas-com-acesso-a-internet-no-Brasil-supera-120-milhoes.html>. Acesso em: 15 jun. 2015.

NONAKA, I.; TAKEUCHI, H. **Criação de conhecimento na empresa**: como as empresas japonesas geram a dinâmica da inovação. Rio de Janeiro: Campus, 1997.

NYST, C. It's time for our governments to stop eavesdropping and start listening. **Privacy International**, 28 jun. 2015. Online. Disponível em: <https://www.privacyinternational.org/?q=node/94>. Acesso em: 11 jun. 2015.

OLIVEIRA JR, E. Q. A nova lei Carolina Dieckmann. **JusBrasil**, 2012. Online. Disponível em: <http://eudesquintino.jusbrasil.com.br/artigos/121823244/a-nova-lei-carolina-dieckmann>. Acesso em: 30 jun. 2015.

ONU. Declaração universal dos direitos humanos - adotada e proclamada pela resolução 217 A (III) da Assembléia Geral das Nações Unidas em 10 de dezembro de 1948. **UNESCO**, 1998. Disponível em: <http://unesdoc.unesco.org/images/0013/001394/139423por.pdf>. Acesso em: 13 jun. 2015.

O´REILLY, T. **Big data now**. Sebastopol: O'Reilly Media, 2011.

_____. What is Web 2.0: design patterns and business models for the next generation of software. **O'Reilly Media**, 30 set. 2005. Online. Disponível em: <http://www.oreilly.com/pub/a/web2/archive/what-is-web-20.html>. Acesso em: 10 jun. 2015.

OREMUS, W. Who controls your Facebook feed. **Slate**, 03 jan. 2016. Cover story. Online, Disponível em: <http://www.slate.com/articles/technology/cover_story/2016/01/how_facebook_s_news_feed_algorithm_works.html>. Acesso em: 12 fev. 2017.

OXFORD. **English dictionaries**. bibliometrics. Online. Disponível em: < https://en.oxforddictionaries.com/definition/bibliometrics >. Acesso em: 02 jun. 2015.

PANOPTYKON FUNDATION. Home. **Panoptykon Fundation**, 2015. Online. Disponível em: <en.panoptykon.org>. Acesso em: 12 jun. 2015.

PARISER, E. **The filter bubble**: how the new personalized web is changing what we read and how we think. New York: Penguin Press, 2011.

PASQUINELLI, M. Italian Operaismo and Information Machine. **Theory, Culture & Society**, v.32, n.3, p.49-68, 2015.

PATEL, N. Como fazer teste AB rapidamente (e aumentar a taxa de conversão). **Neil Patel**, 2017. Blog. Online. Disponível em:<http://neilpatel.com/br/blog/como-fazer-teste-ab-rapidamente-e-aumentar-a-taxa-de-conversao/>. Acesso em: 07 fev. 2017.

PENSANDO O DIREITO. Anteprojeto de lei para a proteção de dados pessoais. **Pensando o Direito**. Secretaria de Assuntos Legislativos do Ministério da Justiça do Brasil - SAL/MJ. Debates, Proteção de Dados Pessoais, Textos em debate. Online. Disponível em: <http://pensando.mj.gov.br/dadospessoais/texto-em-debate/anteprojeto-de-lei-para-a-protecao-de-dados-pessoais/>. Acesso em: 23 jun. 2015.

PEREZ, J. C. Facebook's beacon more intrusive than previously thought. **PC World**, 2007. Security. Online. Disponível em: <http://www.pcworld.com/article/140182/article.html>. Acesso em: 15 jun. 2015.

PÉREZ, L. G.; NASSIF, M. E. Fatores de Influência na Avaliação dos Observatórios Sociais do Brasil Entendidos como Sistemas de Vigilância Informacional. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 17., 2016, Salvador. **Anais**... Salvador: UFBA 2016.

PHILLIPS, S. A brief history of Facebook. **The Guardian**, 25 jul. 2007. Tech. Online. Disponível em: <https://www.theguardian.com/technology/2007/jul/25/media.newmedia>. Acesso em: 11 fev. 2017.

PRESS, G. A very short history of big data. **Forbes**, 09 may 2013. Tech. Online. Disponível em: <http://www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/>. Acesso em: 13 jun. 2015.

PRIDMORE, J. Expert report: consumption and profiling. **Surveillance & Society**, p. 27-35, sep. 2006. Disponível em: <https://www.personuvernd.is/media/frettir/surveillance_society_full_report_final.pdf>. Acesso em: 08 nov. 2016.

REUTERS, T. New York woman receives wireless pacemaker. **PC Magazine**, 10 aug. 2009. News & Analisys. Online. Disponível em: <http://www.pcmag.com/article2/0,2817,2351371,00.asp>. Acesso em: 14 jun. 2015.

RICHARDS, N. M.; KING, J. H. Big data ethics. **Wake Forest Law Review**, 19 may 2014. Disponível em: <http://ssrn.com/abstract=2384174>. Acesso em: 10 jun. 2015.

ROUSE, M. Metadata. **What Is**, jul. 2014. Online. Disponível em: <http://whatis.techtarget.com/definition/metadata>. Acesso em: 08 jun. 2016.

SAVIĆ, D. Evolution of information resource management. **Journal of Librarianship and Information Science**, v. 24, n. 3, sep. 1992.

SCHOFIELD, P. **Bentham**: a guide for the perplexed. London: Continuum, 2009.

SEMPLE, J. **Bentham's prison**: a study of the panopticon penitentiary. Oxford: Oxford University Press, 1993.

SERASA EXPERIAN. Serasa Experian lança o Mosaic: a melhor radiografia da sociedade brasileira. **Serasa Experian**, 03 fev. 2010. Serasa Consumidor. Online. Disponível em: <http://www.serasaconsumidor.com.br/serasa-experian-lança-o-mosaic-a-melhor-radiografia-da-sociedade-brasileira-11/>. Acesso em: 20 jun. 2015.

SLEDGE, M. CIA's Gus Hunt on Big Data: we ‘try to collect everything and hang on to it forever’. **Huffington Post**, 20 mar. 2013. Tech. Online. Disponível em: <http://www.huffingtonpost.com/2013/03/20/cia-gus-hunt-big-data_n_2917842.html>. Acesso em: 10 jun. 2015.

SOARES DA SILVA, A. **Processamento intensivo de texto com MapReduce**. In: Big data summer school - NCE/UFRJ, 23 fev. 2016. 137 slides. Material apresentado no curso. Microsoft Power Point.

STREIT, M. Dupla sertaneja cria polêmica com a música “Vou jogar na internet”. **Portal Fórum**, 08 abr. 2015. Blog. Online. Disponível em: <http://www.revistaforum.com.br/blog/2015/04/dupla-sertaneja-cria-polemica-com-a-musica-vou-jogar-na-internet/>. Acesso em: 28 jun. 2015.

STROTHER, J. B.; ULJIN, J. M.; FAZAL, Z. Information overload: an international challenge for professional engineers and technical communicators. **IEEE Press**, 12 out. 2012. Online. Disponível em: <http://onlinelibrary.wiley.com/book/10.1002/9781118360491>. Acesso em: 13 jun. 2015.

STROUD, F. Internet of things. **Webopedia**, 2015. Disponível em: <http://www.webopedia.com/TERM/I/internet_of_things.html>. Acesso em: 18 jun. 2015.

THE GUARDIAN. Wearable technology: the latest news and comment on wearable technology. **The Guardian**, jun. 2015. Tech. Online. Disponível em: <http://www.theguardian.com/technology/wearable-technology>. Acesso em: 12 jun. 2015.

THE WASHINGTON POST. NSA slides explain the PRISM data-collection program. **The Washington Post**, 06 jul. 2013. Politics. Online. Disponível em: <http://www.washingtonpost.com/wp-srv/special/politics/prism-collection-documents/>. Acesso em: 12 jun. 2015.

VAN BUSKIRK, E. How the Netflix prize was won**. Wired**, 2009. Business. Online.

Disponível em: <https://www.wired.com/2009/09/how-the-netflix-prize-was-won/>. Acesso em: 08 fev. 2016.

WALDMAN, K. Facebook's unethical experiment. **Slate**, 2014. Science. Online. Disponível em <http://www.slate.com/articles/health_and_science/science/2014/06/facebook_unethical _experiment_it_made_news_feeds_happier_or_sadder_to_manipulate.html>. Acesso em: 12 fev. 2017.

WEISS, S. et al. **Text mining**: predictive methods for analyzing unstructured information. Springer, 2005.

WITTEN, I.; FRANK, E.; HALL, M. **Data mining**: practical machine learning tools and techniques. 3.ed. Burlington: Elsevier, 2011.

WOOD, D. M. et al. A report on the surveillance society: report for the UK information commissioner's office. **Surveillance Studies Network**, 02 nov. 2006. Disponível em: <http://news.bbc.co.uk/2/shared/bsp/hi/pdfs/02_11_06_surveillance.pdf>. Acesso em: 23 jun. 2015.

WOLF, G. The data driven life. **The New York Times**, 02 may 2010. Magazine. Online. Disponível em: <http://www.nytimes.com/2010/05/02/magazine/02self-measurement-t.html>. Acesso em: 13 jun. 2015.