Entre o tecnológico e a complexidade: aplicações da web semântica nos sistemas de informação de arquivos

Between technology and complexity: semantic web applications in archive information systems

Thais Helen do Nascimento Santos

Universidade do Porto / Universidade Federal da Paraíba
thaisnascimento.inf@gmail.com

Resumo

O presente estudo tem como objetivo explorar a web semântica nos seus desdobramentos epistemológicos dentre o paradigma tecnológico investigado por Manuel Castells e o paradigma da complexidade de Edgar Morin. A partir de tal reflexão levantamos a ocorrência das práticas de utilização dessa plataforma nos principais sistemas arquivísticos de Portugal e do Brasil. Para tanto, delineamos como metodologias de trabalho, *a priori*, a revisão do estado da arte do tema. *A posteriori*, a exploração de campo nos acervos arquivísticos ora mencionados. Como resultados, apontamos as vantagens das aplicações semânticas em sistemas e/ou portais de arquivos na interação com o usuário a partir do maior índice de precisão na recuperação, assim como as vantagens do acesso e uso da informação atinente à sua necessidade de busca.

Palavras-chave: Web semântica, paradigma tecnológico, paradigma da complexidade, recuperação da informação em sistemas de arquivos.

Abstract

The current study aims at exploring semantic web in its epistemological unfolding between the technological paradigm investigated by Manuel Castells and Edgar Morin´s paradigm of complexity. From such reflection, the occurrence of use practices of this platform in the main archival systems of Portugal and Brazil was surveyed. Therefore, a priori, the theme state-of-art review and, a posteriori, the field exploration in the mentioned archives were analyzed as work methodologies. As results, we pointed out the advantages of semantic applications in systems and/or archive sites in the interaction with the user from the highest accuracy rate in recovery as well as the advantages of information access and use related to their search need.

Keywords: *Semantic web, technological paradigm, paradigm of complexity, information recovery in archive systems.*

1. Situando a web semântica nas implicações paradigmáticas

O artigo que ora se apresenta, partilha de dois objetivos: por um lado, desdobrar as aplicações semânticas nos sistemas de informação de arquivos nos moldes dos novos paradigmas de informação e comunicação; e por outro explorar a sua operacionalização nos acervos arquivísticos. Concomitantemente, em termos metodológicos, recorremos à revisão de literatura no campo da Ciência da Informação e Comunicação (CIC), da Arquivística, assim como da Sociológica e Filosofia, com Castells (1999), Morin (2001) e Lévy (2009). No segundo momento, nos voltamos a exploração de acervos arquivísticos que fazem uso dessa ferramenta em seus sistemas.

Desde os aspectos epistemológicos até os operacionais, a informação em sua diversidade fenomenológica em face da sua finalidade comunicacional, é o ponto de partida das pesquisas em CIC. Com o advento da Internet e as suas revoluções no modo de produção, organização, preservação, acesso, mediação, uso e apropriação da informação, houve a necessidade de evolução e adaptação dos profissionais da informação (arquivistas, bibliotecários, museólogos) na utilização de ferramentas e recursos para atender as demandas sociais da 'era da informação'.

A 'era da informação' é fortemente marcada por dois aspectos: evolução tecnológica e a mediação da informação. Na convergência tecnológica para a mediação da informação urge as plataformas digitais, caracterizada por Silva (2012b, p. 7) como uma "[...] *base tecnológica concebida e usada humana e socialmente para que se produza, armazene, recupere, dissemine, comunique e transforme o fluxo informacional*". Ora, as plataformas digitais congregam à informação e à comunicação em suas características mais peculiares de base tecnológica: o não esgotamento do fluxo, instrumento de mediação info-comunicacional, constituído por *hardware* e *software*, onde sua atualização pode ocorrer em modo *online* ou *offline*.

Nesse ínterim, as acepções de materialidade da informação foram modificadas. A identificação de um suporte físico não é mais necessária aos documentos que se encontram

em formato digital. Os selos e carimbos que conferem a autenticidade e fidedignidade do conteúdo informacional estão sob os 'cuidados' da criptografia. Isto é, no novo cenário ciberespacial, o suporte documental transgride ao espaço significante potencial (LÉVY, 2009), sendo constituído por bits e bytes. A informação se desprende da materialidade e não deixa de ser um aspecto fenomenológico eminentemente humano, cultural e social.

O uso social da Internet por meio de suas diversas plataformas digitais, leva à efeito a maior apropriação midiática à comunicação e à interação. São os e-mails, *chats*, redes sociais e outras redes, como o *Messenger*, *Skype* etc. Nesse entendimento, aquiescemos com Silva (2012a, p. 50) quando em uso das palavras de Jeanneret (2007), afirma que os medias informatizados:

> *"[...] são dispositivos constituídos por aparelhos de tratamento de informação, no sentido matemático do termo, e tendo, por efeito social, fazer circular mensagens e, através disso, tornar possível troca de informações, de interpretações, de produções de conhecimento e de saberes na sociedade".*

Assim, nos novos parâmetros sociais e culturais que emergem da estrutura 'cibercultural' (Lévy, 1999) ou até mesmo 'galaxial' (Castells, 2003), a Internet e a sua gama de recursos tecnológicos devem ser utilizados e apropriados para comunicação e construção de conhecimento e saber, em virtude da necessidade de "[...] *gerenciamento global das competências nos estabelecimentos de ensino, empresa, bolsas de emprego, coletividades locais e associações*" (Lévy, 2009, p. 177).

Outrossim, as unidades de informação (arquivos, bibliotecas e museus) como umas das fontes para a construção de conhecimento e saber fazem uso de técnicas mais aprimoradas para viabilizar ao seu usuário interagir virtualmente. Nos arquivos, a web semântica vem sendo utilizada, gradativamente, para a otimização do diálogo e cooperação entre homem e máquina e o eficaz acesso à informação nos sistemas de gerenciamento arquivístico, fomentando a construção do conhecimento.

Segundo Souza & Alvarenga (2004, p. 134) a web semântica diz respeito:

> *"[...] em sua essência, [...] a criação e implantação de padrões (standards) tecnológicos para permitir [que programas de dispositivos por meio da infraestrutura de dados da Internet], que não somente facilite as trocas de informações entre os agentes pessoais, mas principalmente estabeleça uma língua franca para o compartilhamento mais significativo de dados entre dispositivos e sistemas de informação de uma maneira geral".*

Como base em um novo paradigma de informação e comunicação em ambientes digitais (Santos, 2013), o qual aspira a melhor operacionalização e organização dos documentos na primazia de fenômeno ontológico e humano e da necessidade no rápido acesso à informação, a web semântica se apresenta como um mecanismo de sistematização síncrona dos dados para a interação entre a informação desejada e o sujeito que a usa.

Em análise dos problemas de recuperação da informação em ambientes digitais na *web*, a W3C (World Wide Web Consortium) iniciou um projeto de aprimoramento da *web* a partir da web semântica. A proposta se funda na melhoria do diálogo entre a máquina e o homem. Nos dados e metadados que configuram um documento *web*, a linguagem de marcação (que adiciona informações singulares aos documentos a fim de distingui-los) adotada passa do HTML (Hyper Text Markup Language) considerada rígida, para o XML (eXtensible Markup Language) mais flexível que permite a relação entre os dados (W3SCHOOLS, 2013).

Além da maior flexibilização com o XML, a estrutura da web semântica faz uso do padrão RDF (Barners-Lee, Hendler & Lassila, 2001), que permite a descrição de dados e metadados por meio de triplas que envolvem recurso – propriedade – valor. Para tanto, o RDF faz emprega a infraestrutura do XML, que atribui sintaxe entre descrição e propriedade dos documentos e permite aplicações inteligentes e automatizadas sobre as informações. A reunião dessas duas ferramentas com as ontologias origina o desenvolvimento do substrato para a aplicação da web semântica com a OIL (Ontology Interference Layer), que será um ponto explorado mais adiante (Moreira, Alvarenga & Oliveira, 2004).

Como extensão da *web* atual, a web semântica não se caracteriza como uma simplista ligação entre páginas da *web*. Seu objetivo principal é descrever as relações entre objetos informacionais: que A é uma parte de B e Y é um membro de Z e as suas propriedades (tamanho, peso, idade, dentre outros). Desse modo, a informação é dada com significado bem definido, permitindo melhor interação entre os computadores e as pessoas, uma vez que as máquinas estarão dotadas de ferramentas inteligentes, capazes de raciocinar, deduzindo o conteúdo dos documentos, facilitando assim a melhoria da recuperação da informação relevante (Souza & Alvarenga, 2004; Pickler, 2007; Breitman, 2005).

No prisma epistemológico, pactuamos com Silva (2012a) quando afirma que os estudos de informação partem de duas formas de abordagem: humanístico-social e tecnológica. Na digressão sob a web semântica, é possível identificá-la na abordagem e paradigma tecnológico por sua condução teórica e prática.

De ordem capitalista, o paradigma tecnológico é considerado por Castells (1999, p. 107) em posse das palavras de Freeman (1988), como:

> *"[...] um agrupamento de inovações técnicas, organizacionais, administrativas inter-relacionadas cujas vantagens devem ser descobertas não apenas em uma nova gama de produtos e sistemas, mas também e sobretudo na dinâmica da estrutura dos custos relativos de todos os possíveis insumos para produção. Em cada novo paradigma, um insumo específico ou conjunto de insumos pode ser descrito como o "fator-chave" desse paradigma caracterizado pela queda dos custos relativos e pela disponibilidade universal".*

Sob a égide econômica e social, o paradigma tecnológico caracteriza a produção e uso contemporâneos da Internet e das suas plataformas digitais por indivíduos e organizações. Desde o anseio ao rápido acesso, uso e transferência de informação ao desenvolvimento de aplicações que aproveitem os efeitos do trabalho em rede para a melhoria das ações funcionais, a Internet se insere nas atividades cotidianas do indivíduo e do seu grupo social. Desse modo, a informação e a Internet se tornam as principais características da sociedade

informacional a partir de sua flexibilidade (reconfiguração social) e da crescente convergência tecnológica.

Castells (1999) pontua cinco aspectos que edificam a base material do paradigma tecnológico:

1) *Informação como matéria-prima:* todas as plataformas digitais em operação têm como fonte de trabalho à informação, e não a informação sob as plataformas digitais, como ocorreu em revoluções tecnológicas anteriores;
2) *Penetrabilidade dos efeitos das novas tecnologias:* como a informação é um fenômeno humano, logo, todos os processos humanos envolvidos são moldados pelo novo meio tecnológico;
3) *Lógica de redes:* estruturação de um conjunto de relações para interação e maior usufruto criativo nas plataformas digitais;
4) *Flexibilidade:* corresponde a possibilidade de organização e reorganização, configuração e reconfiguração dos elementos individuais e organizacionais;
5) *Convergência de tecnologias específicas para um sistema integrado:* capacidade de integração de informações em um grande sistema visando maior interação, comunicação e cooperação estratégicas.

Na crítica tecnológica, é possível explorar a web semântica no escopo das cinco caracterizações elencadas por Castells (1999):

1) a razão do fazer semântico está em atribuir sentido aos termos indexadores, logo, todo o sistema é construído e se movimenta a partir da informação como matéria-prima;

2) razão, emoção, cultura, linguagem são aspectos humanos e norteadores para o agrupamento dos termos e no fazer sentido dessa união. Sendo assim, os processos humanos também engendram o fazer tecnológico e semântico;

3) a lógica de redes apresenta-se como um dos eixos centrais da web semântica, é a confluência linear e não-linear das relações entre os termos que permitirá ao usuário

a recuperação da informação de forma eficiente e eficaz, viabilizando, assim, a construção do conhecimento;

4) sob a ênfase tecnicista, a flexibilidade de organização e reorganização linear e não-linear da informação é possível a partir dos metadados na linguagem de marcação XML e do padrão RDF e;

5) a integração sistêmica que acentua a base material do paradigma tecnológico, alcançaria o ápice da proposta da web semântica: tecer relações entre diferentes conteúdos da *web*, embutindo ferramentas 'inteligentes' nas máquinas para a atribuição de sentido aos termos de busca, potencializando o grau de relevância do conteúdo fomentando, consequentemente, a construção de conhecimento e saber pelo usuário.

Entretanto, a imersão paradigmática proposta se pauta no fazer tecnológico pelo humano, que é um sujeito social, cultural, biológico, cognoscente, físico, etc. Um humano que pensa e repensa, cria e recria, faz e refaz, assim como organiza e reorganiza *hardwares* e *softwares*. Essa reflexão incita a crítica já levantada por Castells (1999) no caráter complexo do fazer e interagir tecnológico. Não são os bits e bytes que contemplam toda a estrutura das plataformas digitais. Historicamente, foi a convergência capitalista (interesses organizacionais e/ou individuais) que demandou recursos de comunicação mais aprimorados, minimizando a incidência de ruídos e maior velocidade na troca de mensagens nos telefones, televisores, computadores, etc. Sendo assim, o humano exerce papel fundamental no ponto de vista comportamental, histórico, social, evolutivo e tecnológico.

Nesse sentido, é crucial tecer relações paradigmáticas entre a tecnologia e a complexidade. Da base paradigmática tecnológica, Castells (1999) nos conduziu as principais ponderações do campo técnico e social. Sobre a complexidade, nos munimos com o filósofo, historiador e sociólogo, considerado como um dos maiores pensadores do século XX: Edgar Morin (2001).

A dinâmica do conhecimento desencadeou a necessidade de ordenar, hierarquizar e simplificar o fazer científico através da busca pela certeza e a minimização da ambiguidade. De ordem positivista, a simplificação trouxe resultados significativos com a evidência a partir dos números. Foi assim, que Descartes permitiu avanços à ciência e à filosofia, no século

XVII. Para o século XXI, o fazer e pensar científico evoluiu consideravelmente. Todavia, a ordenação, hierarquização e simplificação ainda continuam em uso, especialmente por pesquisadores das ciências exatas e biológicas. Sobre isso, Morin (2001, p. 21) acredita na dificuldade em *"[...] enfrentar a confusão (o jogo infinito das interretroações), a solidariedade dos fenômenos entre eles, a bruma, a incerteza, a contradição".*

No berço das Ciências Sociais e Humanas, o pensamento evolutivo, da informação como um fenômeno humano, cultural e social orienta os debates e pesquisas recentes no seio da complexidade, ou seja, do *"[...] tecido em conjunto; tecido de acontecimentos, ações, interações, retroações, determinações, acasos*" (Morin, 2001, p. 20), que constituem a produção, organização, preservação, acesso, mediação e uso da informação.

Na perspectiva informacional, as aplicações da web semântica por mais que estejam estruturadas em códigos binários, inscrevem relações conceituais que permitem o raciocínio de conteúdos com vistas à construção do conhecimento, estando, dessa forma, entre a simplicidade (artefatos da tecnologia) e a complexidade. Nesse sentido, concordamos com Morin (2001, p. 149) ao afirmar que

> *"[...] a complexidade não é apenas a união da complexidade e da não-complexidade (a simplificação), a complexidade encontra-se justamente no âmago da relação entre o simples e o complexo porque uma tal relação é ao mesmo tempo antagônica e complementar".*

Diante disto, as próximas reflexões desse trabalho irão se ocupar das relações da web semântica nos sistemas de informação em arquivos através do olhar da complexidade. Fruto das novas perspectivas em plataformas digitais, essa ferramenta permite a maior interação entre o homem e a máquina apresentando resultados satisfatórios para ambas as vertentes (humanas e tecnicistas).

2. Recursos semânticos nos sistemas de informação de arquivos: olhar sob a complexidade

Entendo a informação como fenômeno humano e tecnicista, a abordagem da web semântica é dualista, se manifesta entre a tecnologia e a complexidade, conforme evidencia o título atribuído a esse trabalho. Nesse contexto, a complexidade converge tais paradigmas na natureza interdisciplinar das CIC, que contornam o âmago da web semântica entre a Ciência da Computação e a Linguística.

A abordagem interdisciplinar no paradigma da complexidade se pauta na afirmação de Saracevic (1996, p. 46) quando declara que: "*problemas complexos demandam enfoques interdisciplinares e soluções multidisciplinares*". Nesse entendimento, o autor elenca a Biblioteconomia, Ciência da Comunicação, Ciência Cognitiva (incluindo a Inteligência Artificial) e a Comunicação como áreas centrais do fluxo interdisciplinar da Ciência da Informação (CI). No contrafluxo epistêmico-operacional das Ciência da Informação e da Comunicação, ainda podem ser elencadas como áreas dialógicas: Administração, Linguística, História, Filosofia, Direito, Psicologia, dentre outros. Tais agregações são imprescindíveis para melhor compreender o fenômeno informacional no processo comunicacional minimizando as suas incertezas e reconhecendo a sua irredutibilidade (Morin, 2001).

Nesse entendimento, este tópico se orienta na análise da informação pelo registro documental em suas novas facetas tecnológicas, para alcançar o ápice das premissas da web semântica nos arquivos.

2.1. O registro documental na flexibilidade semântica

O documento está situado entre a informação e a web semântica. Dessa forma, nos reportamos a um breve retorno histórico da concepção de documento se encontra com um dos seus principais pensadores: Paul Otlet. Considerado um dos visionários da documentação até os dias atuais, Otlet publicou, em 1934, o *Traité de Documentation: le livre sur le livre* que amplia a noção de documento e envolve: documentos particulares; bibliotecas; bibliografias; arquivos documentais; arquivos administrativos; arquivos antigos; documentos sonoros (música) e audiovisuais (cinema), assim como as coleções museológicas

e a enciclopédia. Em termos gerais, a proposta de Oltlet se firma no uso de um termo genérico que contemplaria toda a dinamicidade e diversidade dos registros informacionais: *"[...] a unidade intelectual seria o pensamento e o livro um meio de produzir utilidades intelectuais"* (Ortega & Ginez de Lara, 2010, p. 01).

O ponto de vista de Otlet sobre o documento orientou vários outros pesquisadores, em especial na França e na Espanha. Ortega & Ginez de Lara (2010) em recente pesquisa desenvolvida, levantaram obras de diversos pensadores da documentação que refletem o pensar otleriano do documento, desde a Segunda Guerra Mundial até a década passada. Longe de esgotar um tema relevante e profundo de base epistêmica da Ciência da Informação, a intenção desse tópico está na discussão do documento sob a influência das plataformas digitais.

No propósito de estruturar uma linearidade reflexiva, exploramos Silva (2012a, 2012b) em sua análise do impacto das Tecnologias de Informação de Comunicação (TIC) na noção de documento. Em uma digressão epistemológica, o autor apresenta a apreensão de documento adotado por Otlet que é acompanhado por outros estudiosos da área, até as configurações mais atuais do documento nas plataformas digitais.

O documento é compreendido por dois elementos: o simbólico/intelectual (a inscrição) e o material/natural (o suporte). Tendo como ponto de partida a segunda acepção, a materialidade do suporte, já abordamos no tópico anterior, as modificações ocorridas diante da produção e uso dos documentos nas plataformas digitais. Entretanto, ainda é um impasse para investigadores das CIC que o material ainda exista na potencialidade do ciberespaço. O palpável, o tangível, o quantificável ainda são 'objetos' de consumo pelos profissionais da informação e até mesmo para comunicólogos. Sobre isso, Silva (2012a, p. 36) advoga que "[...] *a materialidade do documento (seja em papel, seja eletrônico) não apaga a existência menos tangível da informação e esta subiste conformada, ou ajustada aos limites materiais e formas decorrentes da "condição documentária*"". Acrescente-se, os novos formatos do material em 'espaço significante potencial' no ciberespaço (Silva, 2012a; Lévy, 2009).

Nessas novas configurações materiais da natureza do suporte documental, os informáticos denominaram a justaposição da informação em códigos binários como "objeto digital". Sob a ênfase tecnológica, o termo foi amplamente adotado em estudos que acentuam a relação das CIC com a Ciência da Computação. Outrossim, Silva (2012b) critica a apropriação desse termo na CI, pela limitação que atribui a erudição de documento em meio digital como um objeto, ou seja, uma coisa perceptível pelos sentidos.

> *"[...] em CI, essa expressão informática além de redutora é perniciosa, porque equívoca, omitindo a presença na "coisa material" de informação humana e social. Daí que a expressão alternativa "documento eletrônico", mesmo que raie a redundância, mantenha óbvia acuidade e valor denotativo, sendo utilizável em CI...". (Silva, 2012b, p. 05).*

Silva (2012b), nessa afirmação, pontua uma disparidade conceitual que ocorre no caso brasileiro. O Dicionário Brasileiro de Terminologia Arquivística (Arquivo Nacional, 2005, p. 75) apresenta duas formas de apreensão do documento em plataforma digital: o digital e o eletrônico. O documento digital compreende o "*documento codificado em dígitos binários, acessível por meio de sistema computacional*". Em sentido um tanto análogo, o documento eletrônico é um "*gênero documental integrado por documentos em meio eletrônico ou somente acessível por equipamentos eletrônicos, como cartões perfurados, disquetes e documentos digitais*". Em outros termos, o documento digital representa a unidade de registro de informações a partir de dígitos binários, enquanto o documento eletrônico é a configuração que assume a unidade de registro de informação acessível por mídia eletrônica. Na prática, por um lado, a dissociação das nomenclaturas deve-se ao fato de que o documento digital nasce, tramita, tem acesso e uso em meio totalmente digital. Por outro lado, o documento eletrônico é o que se integra as plataformas digitais por meio da digitalização. Sua produção é física, contudo, seu trâmite, acesso e uso ocorrem sob a mesma plataforma do documento que nasceu em berço digital. No entanto, é cabível pontuar que tal dissociação teórica e prática ainda é pouco explorada.

Na transição evolutiva que se desenvolve no reflexo da complexidade, a concepção de documento eclode no seu elemento simbólico/material (a inscrição). Nesse ínterim, os

aspectos sociais e culturais configuram a inscrição em um suporte de informação com vistas à comunicação, necessidade vital humana. Sendo assim, o documento se situa entre a informação e comunicação como meio de transmissão.

> *"[...] O dinamismo da informação é, hoje, um fato indesmentível, e o fenômeno infocomunicacional obriga a que percebamos as assimetrias e a profunda complementariedade entre o ato informacional e o ato comunicacional. O elo entre os dois reside no ato documental que hoje é manifestante complexo, por força da tecnologia digital transversal a todos os setores e níveis da vida humana e social". (Silva, 2012b, p. 22).*

No intercurso da produção e comunicação da informação em plataforma digital, as caracterizações essenciais do documento não são extintas. Paralelamente, o documento também se modifica em face das transformações e evoluções paradigmáticas. Por esse entendimento, Silva (2012b) elabora a designação 'docmedia', como a unidade de articulação das funções de inscrição e de transmissão da informação em plataformas digitais. Para exemplificação, o autor se apropria das caracterizações de um *smartphone*: 1) é um documento produzido pela mente de um sujeito/humano que o materializou; 2) produz documento e transmite informação. Sendo assim, *"[...] o documento é o meio e o meio ou média é o documento*". (Silva, 2012b, p. 20).

Portanto, o conceito de 'docmedia' se integra às discussões da web semântica na sua flexibilidade interativa. O documento em plataforma digital semântica é produzido para a interação com outros documentos efetivando a transmissão e comunicação, ou seja, é um meio de comunicação da informação registrada e de outros registros que combinem relações temáticas para a produção do conhecimento.

2.2. Web semântica e arquivos: uma proposta para recuperação e acesso da informação

As plataformas digitais na *web* concentram grande parcela da produção documental, porém, sem nenhuma estratégia de indexação dos documentos para a sua recuperação. Assim, concordamos mais uma vez com Souza & Alvarenga (2004, p. 133) quando asseveram que:

> *"Embora tenha sido projetada para possibilitar o fácil acesso, intercâmbio e a recuperação de informações, a web foi implantada de forma descentralizada e quase anárquica; cresceu de maneira exponencial e caótica e se apresenta hoje como um imenso repositório de documentos que deixa muito a desejar quando precisamos recuperar aquilo de que temos necessidade."*

Destarte, as atividades arquivísticas estão ocupando mais espaço no âmbito digital. As organizações em sua função administrativa, histórica ou social fazem uso gradativo da *web* para agilizar e facilitar as tomadas de decisões diárias. Não só isso, os usuários esperam e necessitam de sistemas cada vez mais fáceis de manipulação para a interação e recuperação da informação. Nesse contexto, as atividades desenvolvidas com documentos nos suportes físicos são adaptadas para execução nas plataformas digitais viabilizando o atendimento das novas demandas.

A indexação é um *"[...] processo pelo qual documento ou informações são representados por termos, palavras-chave ou descritores, propiciando a recuperação da informação.*" (Arquivo Nacional, 2005, p. 107), e se constitui como uma das principais etapas de gerenciamento arquivístico, responsável pelo arquivamento e posterior acesso aos documentos. Em consonância com o vocabulário controlado, a indexação vai permitir *"[...] organizar e recuperar documentos – e informações – com consistência, gerando, consequentemente, confiança no sistema*", (Smit & Kobashi, 2003, p. 14). Integrando a indexação com recursos tecnológicos, as atividades de recuperação da informação em sistemas de arquivos ganham mais recursos sofisticados na otimização dos serviços de acesso e uso da informação com a implantação da web semântica.

Ambicionando aprimorar a interação entre o homem e a máquina, especialmente na recuperação da informação na *web*, Barners-Lee, Hendler & Lassila (2001, p. 01) definiram como planos da World Wide Web Consortion (W3C) a implementação da web semântica: "*The Semantic Web is not a separate Web but an extension of the current one, in which information is given well-defined meaning, better enabling computers and people to work in cooperation*".

A *web* atual permite que os computadores apresentem a informação, ficando a cargo do humano a interpretação dos dados. Um exemplo comum e prático é a recuperação da informação através dos buscadores. A *web*, na amplitude documental que concentra, se ocupa em apresentar o maior índice de revocação dos documentos, cabendo ao sujeito que necessita da informação precisar os que são atinentes a sua busca.

Nesses moldes, os aprimoramentos propostos pela web semântica são empregados em programas e dispositivos especializados para que possam interagir entre si. Um projeto que envolverá toda a rede de computadores, visa embutir inteligência nas máquinas, maximizando a eficiência na troca de informações das ações e atividades diárias.

A linguagem franca se sobressai como um dos pontos nevrálgicos das aplicações semânticas. Sua base linguística se pauta no comprometimento ontológico, ou seja, na escolha que leva a selecionar um determinado termo ao invés de outro, por meio de uma padronização terminológica.

A ontologia, comumente abordada sob a óptica filosófica, do grego 'ontos' (ser) e 'logos' (palavra), *"[...] é o estudo da existência de todos os tipos de entidades, abstratas ou concretas, que constituem o mundo*". (Lima-Marques, 2006, p. 17). Seus princípios surgem entre os séculos XVII e XVIII com Aristóteles na perspectiva de estabelecer dez caracterizações básicas para classificar o tudo.

Na CI, as concepções ontológicas aparecem em face dos mecanismos teóricos e práticos dos profissionais da informação. Como uma instância primária do pensamento humano, que enseja a representação primária do conhecimento na percepção, identificação e representação do documento, a ontologia auxilia nas atividades de representação e indexação da informação (Santos & Souza, 2013).

No que tange ao desígnio computacional, a ontologia (situada nos estudos de Inteligência Artificial) é amplamente utilizada nos recursos que envolvem "*[...] representação, reuso, compartilhamento, aquisição e integração de conhecimento, processamento de linguagem*

natural e tradução automática, comunicação de informações entre sistemas, agentes, empresas ou pessoas, recuperação da informação e especificação e software." (Moreira, Alvarenga & Oliveira, 2004, p. 01).

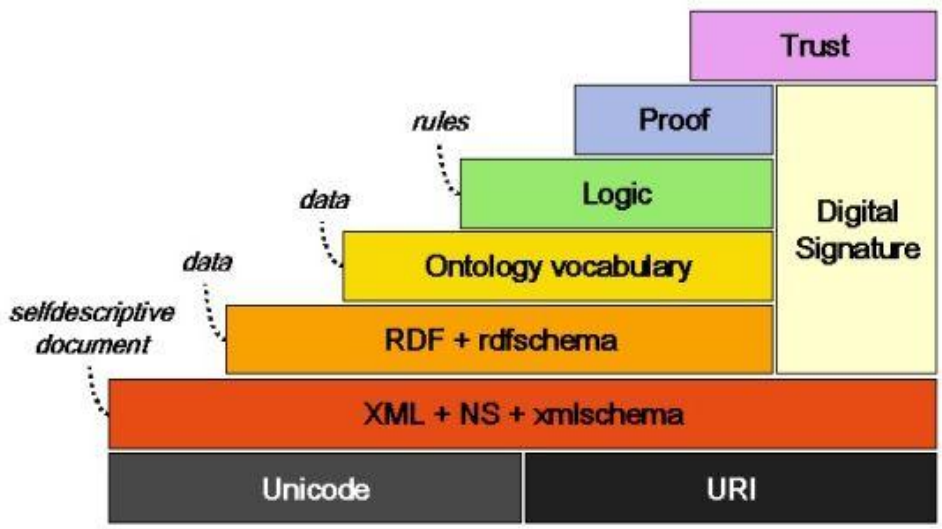
Acrescendo à linguagem franca por meio das ontologias, o sistema da web semântica requer outros recursos computacionais flexíveis, como linguagem de marcação e padrão de metadados. Na preocupação em assegurar a apresentação e a navegação em rede, o HTML (HyterText Markup Language) foi projetado com limitações semânticas, possibilitando apenas a editoração gráfica de fontes, cores e outras configurações básicas, implicando rigidez na legibilidade por máquinas e humanos. Diante dessa problemática, a W3C projetou o XML (eXtensible Markup Language), uma linguagem de marcação para atender necessidades especiais na descrição de dados. De acordo com Freitas (2013, p. 36), o XML "[...] *permite a representação de outras linguagens de forma padronizada*". Sua estrutura mais flexível comporta a separação do conteúdo de formatação, criação de *tags* sem limitação, interligação com bancos de dados distintos, dentre outros.

O padrão RDF (*Resource Description Framework*) é outro recurso que foi apropriado pela web semântica. O RDF é um padrão que identifica e descreve partes de documentos ou dados por um esquema de triplas de recurso-propriedade-valor (Souza & Alvarenga, 2004), potencializando o acesso aos padrões de metadados publicados na *web*. Freitas (2013, p. 38) complementa que a principal vantagem do RDF como linguagem de descrição de recursos sobre o XML, "[...] *reside na liberdade de ignorar as imposições da estrutura do documento, referindo-se apenas a dados sobre o conteúdo*", adicionando, assim, mais semântica ao documento sem a necessidade de alterar a sua estrutura.

A soma desses recursos (XML + RDF + Ontologias) resulta no OIL (Ontology Interference Layer) que alicerça o substrato da web semântica. O conjunto das vantagens de cada recurso acima descrito, engendra funções semânticas baseadas em lógica de descrições, modelagens de orientação a objetos e duas variações de sintaxe: uma baseada no XML e outra no RDF (Moreira, Alvarenga & Oliveira, 2004).

Nessas premissas, os recursos básicos da web semântica são articulados nas camadas de tecnologia da *web* e padrões. Essas camadas concebem a arquitetura básica da web semântica proposta pela W3C, conforme observamos na figura abaixo.

Figura 1 – ***As camadas da web semântica***



Fonte: Koivunem; Millher (2002, p. 34).

De acordo com Koivunem & Miller (2002), o *Unicode* e a cama URI comprovam a utilização de caracteres internacionais e fornecem meios para identificação dos objetos semânticos. A camada XML (com *namespace* e *Schema*) integram as definições da web semântica com outros padrões baseados em XML. A camada RDF e *RDFSchema* é responsável por tecer afirmações sobre objetos URI e definir vocabulários (esta é a camada onde é possível inserir recursos de *links*). A ontologia acompanha a evolução dos vocábulos, pois é a responsável por definir as relações entre os conceitos. A camada de assinatura digital tem como função detectar as alterações nos documentos. Por fim, as camadas superiores *Logic*, *Proof* e *Trust* ainda não possuem formas bem definidas. Entretanto, o objetivo de cada uma, respectivamente, são: escrita de regras; executa as regras e avaliar em conjunto com a camada de confiança para aplicações e dados.

A partir base geral de organização e operacionalização da web semântica, o próximo tópico é destinado a conhecer projetos semânticos que estão sendo desenvolvidos em acervos arquivísticos. Para abarcar contextos locais distintos, selecionamos explorar o caso do PAEIS da Diretoria-Geral de Arquivos em Portugal e do Portal Semântico do CPDOC no Brasil.

3. Usos e aplicações semânticas na recuperação da informação em arquivos

Nas tessituras teórico-conceituais apontadas, a web semântica se manifesta como uma ferramenta para o futuro. Contudo, a ausência de formas sólidas de sistematização da arquitetura semântica não deve direcionar ao entendimento da inexistência de recursos semânticos em portais, *websites*, sistemas de recuperação da informação, etc. É pertinente destacar, que a W3C projeta a web semântica para a *web* na sua totalidade, não impedindo que sistemas de menores dimensões utilizem as ferramentas para melhorar a recuperação da informação aos seus usuários.

Nessa perspectiva, para ilustrar e elucidar as vantagens das aplicações semânticas, selecionamos dois projetos arquivísticos para breve apresentação. Na tentativa de explorar contextos diferentes optamos pela experiência de aplicações e projetos em um acervo arquivístico português e um brasileiro, respectivamente.

Em uma reunião realizada no ano de 2006, o governo português apontou à Diretoria-Geral de Arquivos (DGARQ) a necessidade de estabelecer regras comuns para identificação e organização de seus documentos. Tal apontamento, levou os diretores gerais a elaboração de ações estratégicas com vistas à comunicação integrada com a gestão da documentação no âmbito da Administração Central do Estado (ACE).

A integração dessas ações originou o plano de 'Interoperabilidade Semântica', que, conforme Lourenço, Henriques & Penteado (2011, p. 235) objetiva "[...] *uma estrutura hierárquica representativa das funções daquele universo [...] que deverá servir de referência para uma posterior integração nos planos de classificação dos diferentes serviços públicos e nos respectivos sistemas de gestão de documentos de arquivo*". Em outros termos, o projeto

tem como ênfase o aprimoramento das atividades de classificação, representação, indexação e atribuição de uma linguagem franca para a comunicação entre os órgãos. Assim sendo, as caracterizações principais do projeto dialogam com as diretrizes da web semântica. Porém, a adoção do termo 'interoperabilidade' parte de uma denominação compartilhada entre a administração pública portuguesa, em especial nas ações de modernização administrativa.

Dois projetos comportam a primeira fase de ações: o Metainformação para a Interoperabilidade (MIP) e a Macroestrutura Funcional (MEF). O MIP consiste em elementos de metainformação que objetivam facilitar a interoperabilidade na gestão, no acesso no e uso das informações. Em consonância, o MEF se dirige a representação conceitual de funções para a harmonização de valores dos elementos obrigatórios de metainformação estabelecidos no MIP a partir de um código de classificação. A reunião desses projetos comporta o sistema geral PAEIS (Programa Administração Eletrônica e Interoperabilidade Semântica) destinado a produzir e implantar recursos semânticos interoperacionais para a Administração Pública e outras entidades estatais (DGARQ, 2013).

Enquanto ações de longo prazo, o projeto semântico organizado pela Diretoria-Geral de Arquivos, ainda está em vias de execução. Como prevê a adesão voluntária das entidades e alterações de culturas organizacionais, as atividades demandam tempo e envolvimento de diferentes equipes técnicas.

Nos trilhos práticos, o Brasil dispõe do CPDOC (Centro de Pesquisa e Documentação de História Contemporânea do Brasil), vinculado a Escola de Ciências Sociais e História da Fundação Getúlio Vargas (FGV), que foi criado em 1973, para "[...] *abrigar conjuntos documentais relevantes para a história recente do país, desenvolver pesquisas históricas e promover cursos de graduação e pós-graduação*" (Souza, Higuchi & Silva, 2013, p. 153). O acervo acondiciona documentos pessoais de personalidades públicas, registros de programa de história oral, dicionário biográfico brasileiro, dentre outros.

Em diversos suportes, os documentos passaram/passam por processo de digitalização desde o ano de 2008. São documentos textuais, fitas cassete, fitas rolo, fotografias, entrevistas

orais, etc, que relatam a história republicana do Brasil. Diante da riqueza histórica, social, cultural e informacional em face do comprometimento com o tratamento arquivístico da sua documentação, o CPDOC é considerado como um dos mais importantes acervos do país.

Assim como o caso português, foi a identificação da necessidade de melhorias na recuperação da informação que motivou uma ação interdisciplinar com a Escola de Matemática Aplicada (EMAp) da mesma instituição. A equipe interdisciplinar realizou um diagnóstico dos documentos e dos sistemas de tratamento adotados, apontado a necessidade de melhor descrição dos dados através de uma padronização descritiva, assim como de interação entre os sistemas existentes.

Nesse cenário, tem início o projeto para implantação do Portal Semântico, que nas palavras de Souza, Higuchi & Silva (2013, p. 154-5)

> *"[...] prevê a migração de todo o acervo atual para uma base de dados comum em formato RDF triplestore, e a unificação dos padrões de descrição entre todos os fundos e sistemas, o que envolve a criação de ontologias de descrição e de domínio. Como objetivo, pretende-se oferecer uma interface única para buscas temáticas transversais e integradas, utilizando-se conceitos e categorias de conceitos relativos ao domínio da História Contemporânea Brasileira – como pessoas, acontecimentos e locais – através de todos os sistemas/acervos atuais".*

Iniciado em 2010, a implantação do Portal Semântico desenvolveu outros cinco projetos exordiais, dentre os quais estão: projeto de reconhecimento de faces e personagens, projeto de alinhamento de som e texto, projeto de mineração de textos, projeto de "wikificação" do DHBB e o projeto de criação de ontologias a partir dos descritores de sistemas. A interligação desses projetos compreende um sistema único com tecnologias abertas e preconizadas pelo projeto web semântica da W3C.

O Portal Semântico traz como vantagens ao acervo do CPDOC: acesso unificado aos acervos dos sistemas, buscas transversais entre os sistemas, interligação entre os acervos através de

linguagem franca, padrões únicos de descrição adotados mundialmente (viabilizando a interação/interoperabilidade com sistemas externos), dentre outros.

4. Considerações finais

O percurso traçado por este artigo nos permitiu explorar a web semântica diante do paradigma tecnológico e do paradigma da complexidade, no seio dialógico interdisciplinar com a Arquivística, a Linguística e a Ciência da Computação. Nesse sentido, contornamos como primeiro objetivo do estudo, desdobrar as aplicações semânticas nos sistemas de informação de arquivo sob os paradigmas de informação e comunicação, nas perspectivas sociológicas, filosóficas e epistemológicas apresentadas por Castells (1999) e Morin (2001), respectivamente. Dentre as configurações do paradigma tecnológico, assim como da complexidade, consideramos que a web semântica é um fenômeno tecnológico, em todos os seus aspectos. Contudo, é inerente a elementos subjetivos, em especial, por dois fatores: 1) produção técnica realizada por humanos, em contextos socioculturais específicos; 2) surge das necessidades humanas cognitivas de maior interação e comunicação com as plataformas digitais.

No âmago das CIC, a web semântica se apresenta como uma ferramenta presente nos sistemas utilizados pelas unidades de informação. Eficaz para interação e comunicação entre diversos sistemas, o segundo objetivo deste artigo se concentrou em explorar a sua operacionalização em acervos arquivísticos. Em termos didáticos, um tópico foi destinado a essa exposição que transitou entre a realidade portuguesa e brasileira. O PAEIS (Programa Administração Eletrônica da Interoperabilidade Semântica) que está em desenvolvimento pela Diretoria-Geral de Arquivos de Portugal, se encontra em vias de articulação com as entidades administrativas participantes e visa à integração das atividades de classificação, representação, indexação (atribuição de linguagem franca) para a comunicação entre as entidades da Administração Central do Estado (ACE). Já em pleno funcionamento, a ilustração brasileira que merece destaque é o Portal Semântico do CPDOC (Centro de Pesquisa e Documentação Histórica Contemporânea) da Fundação Getúlio Vargas, que oferece uma interface única para buscas temáticas transversais e integradas no domínio da História Contemporânea Brasileira através de todos os seus conjuntos documentais.

Enquanto uma plataforma digital, a web semântica oferta recursos de interação entre o meio e o receptor da informação, aprimorando o processo comunicacional na melhor recepção e uso da informação para produção do conhecimento. Essa vantagem se destaca diante da expectativa e necessidade dos usuários dos sistemas de informação de arquivos, em melhores e avançados mecanismos de busca, recuperação por precisão e não por revocação de documentos, e acesso às informações cruzadas com outros documentos e/ou 'docmedias' capazes de agregar maiores resultados à sua busca e acesso, gerando, consequentemente, um usuário satisfeito.

Além disso, segundo Souza & Alvarenga (2004), a aplicação de recursos semânticos em sistemas de informação apresenta vantagens significativas aos projetos de novos e melhorados motores de busca, a construção de novas interfaces com o usuário, indexação automática de documentos, construção automática de tesauros e vocabulários controlados, gestão do conhecimento organizacional e gestão da informação estratégica e da inteligência competitiva, dentre outros.

Dentre a análise epistemológica e operacional estabelecidas nas tessituras conceituais traçadas, outros direcionamentos investigativos podem ser apontados no intuito de orientar estudos posteriores aos interessados na temática, tais como: comportamento informacional do usuário no sistema ou portal semântico; satisfação nos processos comunicacionais (acesso, interação e uso) desses sistemas ou portais; até as ações tecnicistas de representação e organização semântica da informação.

5. Referências Bibliográficas

Arquivo Nacional (Brasil). (2005). *Dicionário Brasileiro de Terminologia Arquivística.* Rio de Janeito: Arquivo Nacional.

Berners-Lee, T., Hendler, J. & Lassila, O. (2001). *The Semantic Web*. Scientific American. Disponível em: http://www-sop.inria.fr/acacia/cours/essi2006/Scientific%20American_%20Feature%20Article_%20The%20Semantic%20Web_%20May%202001.pdf (acedido em 16 de outubro de 2013).

Breitman, K. (2005). *Web semântica: a Internet do futuro.* Rio de Janeiro: LTC.

Castells, M. (1999). *A sociedade em rede – a era da informação: economia sociedade e cultura.* São Paulo: Paz e Terra.

Castells, M. (2003). *A galáxia Internet: reflexões sobre a Internet, negócios e sociedade*. Rio de Janeiro: Jorge Zahar.

DGARQ (Diretoria-Geral de Arquivos). (2013). *Programa Administração Eletrônica e Interoperabilidade Semântica: perguntas frequentes.* Disponível em: http://dgarq.gov.pt/files/2013/04/FAQ-sobre-PAEIS_MEF-e-MIP_2013-03-06.pdf (acedido a 17 de outubro de 2013).

Freitas, F. L. G. (sem data). *Ontologias e a web semântica*. Disponível em: http://www.inf.ufsc.br/~gauthier/EGC6006/material/Aula%203/Ontologia_Web_semantica%20Freitas.pdf (acedido a 13 de outubro de 2013).

Koivunem, M. & Miller, E. (2001). *W3C Semantic Web Activity*. In Semantic Web Kick-Off in Finland: Vision, Technologies, Research, and Applications, Finlândia.

Lévy, P. (2009). *Cibercultura.* São Paulo: Editora 34.

Lima-Marques, M. (2006). *Ontologias: da filosofia à representação do conhecimento*. Brasília: Thesaurus.

Lourenço, A., Henriques, C. & Penteado, P. (2011). *Nos modelos e instrumentos da gestão da informação arquivística na Administração Pública: a MacroEstrutura Funcional (MEF)*. In 8º Congresso Nacional da Administração Pública: desafios e soluções, Lisboa.

Moreira, A., Alvarenga, L. & Oliveira, A. De P. (2004). O nível do conhecimento e os instrumentos de representação: tesauros e ontologias. *DataGramaZero,* 6(5), p. 01.

Morin, E. (2001). *Introdução ao pensamento complexo*. Lisboa: Instituto Piaget.

Ortega, C. D. & Ginez de Lara, M. L. (2010). A noção de documento: de Otlet aos días de hoje. *DataGramaZero*, 2(11), p. 01.

Pickler, M. E. V. (2007). Web semântica: ontologias como ferramentas de representação do conhecimento. *Perspectivas em Ciência da Informação*, 1(12), PP. 65-83.

Santos, H. (sem data). *Desafios paradigmáticos e globais no campo da Ciência da Informação e Comunicação: em busca de novas fundamentações.* Disponível em: http://storage.campus.ua.sapo.pt/files/5316e06dbce8a4349e65724639c8b34a/Texto_Helena_Santos.pdf (acedido em 28 de setembro de 2013).

Santos, T. H. do N. & Souza, A. S. C. de. (2013). Registros de conhecimento em face da descrição documental: um estudo multidisciplinar. In. E. C. dos Santos & F. F. De Souza (Org.). *Seminário de Saberes Arquivísticos (SESA): práticas de leitura e escrita na universidade* (pp. 97-112). Curitiba: Appris.

Saracevic, T. (1996). Ciência da Informação: origem, evolução e relações. *Perspectivas em Ciência da Informação*, 1(1), pp. 41-62.

Silva, A. M. da. (2012a). O impacto do uso generalizado das TIC no conceito de documento: ensaio analítico e crítico (I). *Prisma.com*, 16. Disponível em:

http://revistas.ua.pt/index.php/prismacom/article/viewFile/1319/pdf (acedido a 27 de setembro de 2013).

Silva, A. M. da. (2012b). O impacto do uso generalizado das TIC no conceito de documento: ensaio analítico e crítico (II). *Prisma.com*, 18. Disponível em: http://revistas.ua.pt/index.php/prismacom/article/viewFile/2229/pdf (acedido a 27 de setembro de 2013).

Smit, J. W. & Kobashi, N. Y. (2003). *Como elaborar vocabulário controlado para aplicação em arquivos*. São Paulo: Arquivo do Estado/Imprensa Oficial do Estado de São Paulo.

Souza, R. R. & Alvarenga, L. (2004). A web semântica e as suas contribuições para a Ciência da Informação. *Ciência da Informação*, 1(33), p. 132-141.

Souza, R. R., Higuchi, S. & Silva, D. L. da. (sem data). *Desenvolvimento de Ontologias para o Portal Semântico do CPDOC.* Disponível em: http://ceur-ws.org/Vol-776/ontobras-most2011_paper17.pdf (acedido a 17 de outubro de 2013).

W3C Brasil. (sem data). Sobre o W3C. Disponível em: http://www.w3c.br/Sobre (acedido a 16 de outubro de 2013).

W3SCHOOLS. (sem data). *A web semântica.* Disponível em: http://www.w3schools.com/web/web_semantic.asp (acedido a 10 de outubro de 2013).