



LUANA FARIAS SALES

INTEGRAÇÃO SEMÂNTICA DE PUBLICAÇÕES CIENTÍFICAS E
DADOS DE PESQUISA: PROPOSTA DE MODELO DE PUBLICAÇÃO
AMPLIADA PARA A ÁREA DE CIÊNCIAS NUCLEARES

Tese de doutorado
Julho de 2014



UNIVERSIDADE FEDERAL DO RIO DE JANEIRO
INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO

LUANA FARIAS SALES

INTEGRAÇÃO SEMÂNTICA DE PUBLICAÇÕES CIENTÍFICAS E DADOS DE
PESQUISA: PROPOSTA DE MODELO DE PUBLICAÇÃO AMPLIADA PARA A ÁREA
DE CIÊNCIAS NUCLEARES

RIO DE JANEIRO
2014

LUANA FARIAS SALES

**INTEGRAÇÃO SEMÂNTICA DE PUBLICAÇÕES CIENTÍFICAS E DADOS DE
PESQUISA: PROPOSTA DE MODELO DE PUBLICAÇÃO AMPLIADA PARA A ÁREA
DE CIÊNCIAS NUCLEARES**

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Informação, Convênio entre o Instituto Brasileiro de Informação em Ciência e Tecnologia e a Universidade Federal do Rio de Janeiro/ Escola de Comunicação, como requisito parcial à obtenção do título de Doutora em Ciência da Informação.

Orientadores: Profa. Dra. Rosali Fernandez de Souza
Prof. Dr. Luís Fernando Sayão

Rio de Janeiro
2014

S123	Sales, Luana Farias Integração semântica de publicações científicas e dados de pesquisa: proposta de modelo de Publicação Ampliada para a área de Ciências Nucleares / Luana Farias Sales. -- 2014. 265 f. Inclui apêndice Tese (Doutorado em Ciência da Informação) – Universidade Federal do Rio de Janeiro, Instituto Brasileiro de Informação em Ciência e Tecnologia, Programa de Pós-Graduação em Ciência da Informação, Rio de Janeiro, 2014. Orientadores: Rosali Fernandez de Souza e Luís Fernando Sayão 1. Publicação Ampliada. 2. Dados de pesquisa. 3. Relações Conceituais. I. Souza, Rosali Fernandez (Orient.). II. Sayão, Luís Fernando (Coorient.) III. Universidade Federal do Rio de Janeiro. IV. Instituto Brasileiro de Informação em Ciência e Tecnologia. IV. Título.
------	--

LUANA FARIAS SALES

**INTEGRAÇÃO SEMÂNTICA DE PUBLICAÇÕES CIENTÍFICAS E DADOS DE
PESQUISA: PROPOSTA DE MODELO DE PUBLICAÇÃO AMPLIADA PARA A ÁREA
DE CIÊNCIAS NUCLEARES**

Tese de Doutorado apresentada ao Programa de Pós-Graduação em Ciência da Informação, Convênio entre o Instituto Brasileiro de Informação em Ciência e Tecnologia e Universidade Federal do Rio de Janeiro/ Escola de Comunicação, como requisito parcial à obtenção do título de Doutora em Ciência da Informação.

Rio de Janeiro, 23 de Julho de 2014.

Profa. Dra. Rosali Fernandez de Souza – IBICT/ UFRJ
Orientadora

Prof. Dr. Luís Fernando Sayão – CNEN
Coorientador

Profa. Dra. Lena Vânia Ribeiro Pinheiro – IBICT/ UFRJ
Membro interno

Prof. Dr. Jorge Calmon Biolchini – IBICT/ UFRJ
Membro interno

Prof. Dr. Paulo Augusto Berquó de Sampaio – CNEN-IEN
Membro externo

Profa. Dra. Vera Lúcia Doyle Louzada Dodebey - UNIRIO
Membro externo

Profa. Dr. Eloísa Príncipe – IBICT/ UFRJ
Suplente Membro interno

Prof. Dr. Júlio Cesar Suíta – CNEN-IEN
Membro externo

Para Sayão, meu amigo, meu orientador, parceiro e maior incentivador, com carinho,
respeito e admiração.

Para Anna Lú, como incentivo.

Para Joana, com todo o meu amor.

AGRADECIMENTOS

Aos meus queridos orientadores Rosali e Sayão, por terem sido antes de tudo parceiros nessa caminhada, pelo incentivo, pelas orientações tão pertinentes, por terem me dado liberdade de criação dentro dos limites acadêmicos.

Aos meus mestres primeiros, professores da UFF, que me iniciaram na vida acadêmica, despertando em mim o amor pela Ciência da Informação, em especial, Prof^a. Maria Luisa e Prof. Marcondes.

Aos meus mestres mais recentes, professores do IBICT e UFRJ, que tiveram a responsabilidade de manter aceso o amor pela Ciência da Informação, em especial Prof^a. Lena Vânia e Prof. Jorge Biolchini.

Aos meus colegas de turma, pessoas tão especiais, com quem pude dialogar e trocar conhecimento.

À professora Maria Luiza Machado Campos, do Núcleo de Computação Eletrônica, sempre abrindo caminho para possíveis diálogos entre a Ciência da Informação e a Ciência da Computação.

Aos meus colegas de trabalho do IEN que dividiram comigo o entusiasmo de desenvolver esta pesquisa, em especial meus diretores, Paulo Berquó de Sampaio, Patrícia Wieland e Fábio Staude, e ainda, à equipe da Coordenação de Ensino e da Biblioteca, aos líderes de áreas temáticas que contribuíram fornecendo um pouco de seu tempo para responder algumas questões e demais colegas pesquisadores pelas dicas sempre pertinentes.

Aos meus colegas da área temática de Gestão do Conhecimento Nuclear, com quem pude dialogar e compartilhar aprendizado durante o período de elaboração desta tese.

À equipe do CIN e da Rede de Bibliotecas da CNEN pela agilidade e presteza no apoio bibliográfico, em especial à Teodora, incentivadora e revisora desta tese. Tenho muito orgulho de trabalhar com vocês!!

Às minhas amigas mães, que dividem comigo a angústia de ser mãe, estudante e profissional ao mesmo tempo.

Às minhas amigas bibliotecárias, que me inspiram e me apoiam fornecendo informações relevantes, sejam estas bibliográficas ou sobre a prática profissional.

À minha querida amiga Dilza Motta, sempre tão presente em todos os momentos da minha vida, com quem tenho a honra de partilhar todo meu aprendizado, dividir as vitórias profissionais e chorar as angústias acadêmicas e da vida pessoal. Obrigada pelo incentivo!!

À querida Hagar, pela amizade e sugestões preciosas.

À equipe de Biblioteca da Faculdade de Farmácia da UFF, onde me escondi tantas vezes para fazer essa tese fluir.

Aos meus pais Dalva e Ulisses (in memoriam) que me ensinaram o que a vida acadêmica não ensina.

À minha irmã Anna Lú, que resolveu seguir os meus passos e cursar Biblioteconomia, me enchendo de orgulho.

Às minhas primas-irmãs, que dividem comigo, desde pequena, a alegria de viver em família, em especial, a Anna Maura, pelas várias vezes que dividiu o papel de mãe da Joana comigo para que eu pudesse fazer essa tese acontecer.

Ao pai da Joana, Jorge Rogério, que sendo um super pai também foi um pouco mãe, colaborando para que eu tivesse alguns minutos extras para o estudo. Obrigada também pela paciência!

À Joana, minha jóia preciosa, que tornou a minha vida mais linda e mais rosa.

A Deus, porque para o que a ciência não prova deve haver uma explicação.

“(...)É preciso ter uma
visão capaz de situar o conjunto. É necessário dizer que não é a quantidade de
informações, nem a sofisticação em Matemática que podem dar sozinhas um conhecimento
pertinente, mas sim a capacidade de colocar o conhecimento no contexto.”

Autor: Edgar Morin

Fonte: Os sete saberes necessários à educação do futuro

SALES, Luana Farias. *Integração semântica de publicações científicas e dados de pesquisa: proposta de modelo de publicação ampliada para a área de ciências nucleares*. Rio de Janeiro, 2014. Tese (Doutorado em Ciência da Informação) -- Escola de Comunicação, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2014.

RESUMO

Esta pesquisa se desenrola sob as condições de um paradigma científico emergente, conhecido como e-Science ou 4º Paradigma Científico. Essa nova forma de fazer ciência se distingue pelo uso intensivo de redes de computadores, repositórios digitais distribuídos e pela geração extraordinária de dados de pesquisa, que é uma consequência do uso determinante de tecnologias de informação e de simulação por *software* do avanço da instrumentação científica. O ambiente informacional que se configura como consequência dessas transformações impacta sensivelmente os padrões de comunicação científica, principalmente no que tange às pesquisas cooperativas, ao compartilhamento e reuso de recursos informacionais e às formas de comunicar e de disseminar os resultados de pesquisa. Para contextualizar o seu campo de estudo, a tese contribui para delinear conceitos novos e renovados para a Ciência da Informação como e-Science, curadoria de dados de pesquisa, objetos digitais complexos, repositório de dados, CRIS (Current Research Information System) e outras infraestruturas essenciais para a gestão da pesquisa e das novas concepções de publicações acadêmicas e científicas. A pesquisa esta alicerçada sobre duas premissas: a primeira levanta a necessidade de um modelo de publicação científica que possa expressar e refletir o novo padrão de geração de conhecimento científico rico em dados, sendo capaz de integrar esses dados às publicações; a segunda preconiza que isso possa ser realizado segundo as possibilidades tecnológicas e os padrões decorrentes da Web Semântica. Essas duas premissas corporificam a formulação da

hipótese colocada pela presente tese: uma publicação científica pode ser enriquecida e estar mais próxima às novas formas de geração de conhecimento da ciência contemporânea se estiver configurada segundo um modelo que vincule por meio de relações semânticas os dados e conjuntos de dados de pesquisa à publicação convencional. O método utilizado foi o dedutivo, partindo de conceitos gerais da Ciência da Informação aplicados à especificidade das Ciências Nucleares. Isto foi realizado de duas formas: por meio de levantamento bibliográfico, para fins de análise e interpretação qualitativa dos conceitos gerais, e por meio da abordagem de análise de domínio que permitiu analisar empiricamente a área de aplicação específica. Como resultado final obteve-se uma proposta de diretrizes para uma política nacional de curadoria digital e um modelo de publicação científica para a área de Ciências Nucleares, em que os dados são ligados às publicações acadêmicas por meio de relações semânticas sistematizadas em taxonomia construída para esta finalidade. Os modelos gráficos foram utilizados como ferramenta para representar e sintetizar os conceitos resultantes. Como conclusão constata-se: alterações no ciclo da comunicação científica, a possibilidade de construção de um novo modelo de publicação científica como padrão relevante para a prática de uma ciência mais aberta e mais colaborativa, e a viabilidade de incorporação dos princípios e das teorias da Biblioteconomia e da Ciência da Informação para a organização do conhecimento técnico-científico no mundo da *eScience*.

Palavras-Chave: Publicação Ampliada; Curadoria Digital; Dados de Pesquisa; eScience; Relação Conceitual

SALES, Luana Farias. *Integração semântica de publicações científicas e dados de pesquisa: proposta de modelo de publicação ampliada para a área de ciências nucleares*. Rio de Janeiro, 2014. Tese (Doutorado em Ciência da Informação) -- Escola de Comunicação, Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2014.

Abstract

This research takes place under the conditions of an arising scientific paradigm, known as e-Science or 4th Scientific Paradigm. This new way of doing science is characterized by intensive use of computer networks, distributed digital repositories and by extraordinary generation of research data, which is a consequence of the heavy use of information and simulation technologies and advancing of scientific instrumentation. The information environment that is established as a result of these transformations significantly impacts the patterns of scientific communication, especially regarding to cooperative research, the sharing and reuse of information resources and ways to communicate and to disseminate research results. In order to create a context for their field of study, the thesis contributes to delineate new and renewed concepts for Information Science such as e-Science, curation of research data, complex digital objects, data repository, CRIS (Current Research Information System Model) and others key infrastructures for the management of research and also of new conceptions of academic and scientific publications. The research is based on two assumptions: first raises the need for a model of scientific publication that would reflect the new standard for generating scientific knowledge characterized by data richness, and being able to integrate these data to publications; the second highlights that this can be performed according to the technological possibilities and standards arising from the Semantic Web. These two assumptions embody the formulation of the hypothesis raised by this thesis: a scientific publication can be enriched and be closer to new ways of generating knowledge, which characterizes contemporary science, if it is configured according to a

model that links through semantic relations the research data and datasets to conventional publication. The method adopted was the deductive one, starting from general concepts of Information Science applied to the specificity of Nuclear Sciences. This was accomplished in two ways: through a literature review, for purposes of analysis and qualitative interpretation of the general concepts; and through the domain analysis approach that allowed empirically analyze the particular application area. As a final result was obtained a proposal of guidelines for a national policy for digital curation, and a model of scientific publication to the Nuclear Sciences area, in which the research data are linked to the academic publications by means of semantic relations systematized into taxonomy built for this purpose. Graphic models are used as a tool to represent and synthesize the resulting concepts. As a conclusion it is observed: changes in the scholarly communication cycle, the possibility of building a new scientific model as relevant standard to the practice of a more open and more collaborative science, and feasibility of incorporating the principles and theories of librarianship and Information Science for the organization of technical and scientific knowledge in the world of eScience.

Key-words: Enhanced Publication; Digital Curation; Data Research; eScience; Semantic Relations

LISTA DE FIGURAS

Figura 1: Modelo do ciclo de vida da curadoria digital	58
Figura 2: Modelo de publicação ampliada	79
Figura 3: Diagrama entidade-relacionamento para entidades básicas e propriedades.....	80
Figura 4: Interface de artigo da <i>the Cell</i> como modelo de publicação ampliada	86
Figura 5: Interface da aba “data” de uma publicação ampliada da <i>Cell</i>	86
Figura 6: Interface do demonstrador do DRIVER-II.....	88
Figura 7: Estrutura de um documento METS.....	106
Figura 8: Codificação segundo o METS de metadados descritivos	109
Figura 9: Propriedades básicas das relações.....	113
Figura 10: Classificação dos tipos de relações	115
Figura 11: Relacionamentos primários do FRBR.....	117
Figura 12: Relacionamento de reponsabilidade do FRBR	119
Figura 13: Relacionamento de assunto do FRBR.....	120
Figura 14: Modelo Básico do OAI-ORE	130
Figura 15: Publicação ampliada representada pelo modelo entidade-relacionamento, pelo modelou OAI-ORE e expresso em RDF/XML	137
Figura 16: Mapa de Avaliação das áreas temáticas	152
Figura 17: Proposta de Modelo de Curadoria Digital para o país	170
Figura 18: Recursos informacionais isolados <i>versus</i> recursos ligados semanticamente ...	174
Figura 19: Interfaces da Plataforma Carpe diEM	176
Figura 20: Intensidade da Interseção entre as áreas temáticas	181
Figura 21: Intensidade da interseção entre a área temática Realidade Virtual e demais áreas	183
Figura 22: Classes de relacionamentos	196
Figura 23: Agregação formada pelas entidades e links semânticos.....	200
Figura 24: Interface de criação e edição de publicações ampliadas	209
Figura 25: Interface do fluxo de registro de nova relação.....	210
Figura 26: Formas de apresentação do modelo de publicação ampliada.....	213
Figura 27: Exemplo de publicação ampliada e respectivas relações semânticas	214
Figura 28: Modelo de publicação ampliada	221
Figura 29: Fluxo de construção de publicação ampliada para o IEN	222

LISTA DE QUADROS

Quadro 1: Síntese dos tipos de dados de pesquisa.....	53
Quadro 2: Síntese das novas iniciativas de publicações existentes	77
Quadro 3: Critérios de avaliação das áreas temáticas do IEN.....	151
Quadro 4: Assuntos cobertos pelas áreas temáticas do IEN	179
Quadro 5: Definição dos diversos tipos de dados.....	185
Quadro 6: Tipos de dados gerados pelas áreas temáticas do IEN	187
Quadro 7: Resultado do levantamento de <i>e-prints</i> do IEN por área temática	190
Quadro 8: Classes de links semânticos	198
Quadro 9: Elementos da agregação formada por entidades e links semânticos	201
Quadro 10: Proposta de classificação dos tipos de relações para publicações ampliadas	202
Quadro 11: Extrato da Taxonomia de Relações Científicas.....	204
Quadro 12: Elementos que compõem o exemplo da publicação ampliada da pesquisa-tese	217

SUMÁRIO

1 INTRODUÇÃO	17
1.1 JUSTIFICATIVA	19
1.1.1 <i>Por que desenvolver um modelo de Publicações Ampliadas para a área de Ciências Nucleares?</i>	22
1.1.2 <i>Para que ter relações bem definidas?</i>	26
1.1.3 <i>Por que propor uma política de curadoria digital para o país?</i>	27
1.2 PROBLEMA E QUESTÕES DE PESQUISA	28
1.3 PREMISSAS E HIPÓTESE	32
1.4 OBJETIVOS	35
2 CONTEXTO TEÓRICO	36
2.1 MODELO GRÁFICO APOIADO POR BASE TEÓRICA CONCEITUAL	36
2.2 OS PARADIGMAS CIENTÍFICOS E A eSCIENCE	40
3 DADOS DE PESQUISA: O QUE SÃO E QUAIS SEUS TIPOS	49
3.1 CURADORIA DIGITAL DE DADOS DE PESQUISA	54
3.2 REPOSITÓRIOS DIGITAIS DE DADOS DE PESQUISA	60
3.3 OUTRAS INFRAESTRUTURAS DE TRATAMENTO DE DADOS DE PESQUISA	66
3.4 NOVOS MODELOS DE PUBLICAÇÃO CIENTÍFICA	71
4 PUBLICAÇÕES AMPLIADAS: O QUE SÃO E COMO SÃO FORMADAS	78
4.1 INICIATIVAS CONCRETAS EXISTENTES	84
4.2 ARQUITETURA DE PUBLICAÇÃO AMPLIADA	89
4.3 OBJETOS DIGITAIS: PRIMEIRO COMPONENTE ESSENCIAL DA PUBLICAÇÃO AMPLIADA	96
4.4 METADADOS PARA GESTÃO DE OBJETOS DIGITAIS	100
4.5 RELAÇÕES CONCEITUAIS: OUTRO COMPONENTE ESSENCIAL PARA FORMAÇÃO DE PUBLICAÇÃO AMPLIADA	110
4.6 AGREGANDO ELEMENTOS NA COMPOSIÇÃO DE PUBLICAÇÃO AMPLIADA: A NORMA OAI-ORE	128
4.7 PUBLICAÇÃO AMPLIADA E WEB SEMÂNTICA	130
5 MATERIAL E MÉTODO	139
5.1 MATERIAL	139
5.2 CLASSIFICAÇÃO METODOLÓGICA E ETAPAS SEGUIDAS	140
5.2.1 <i>Classificação metodológica da pesquisa</i>	141
5.2.2 <i>Etapas Metodológicas</i>	141
6 CAMPO EMPÍRICO: O DOMÍNIO DA ENERGIA NUCLEAR	144
6.1 ANÁLISE DE DOMÍNIO SOB A ABORDAGEM DA ESTRUTURA DO SISTEMA DE INFORMAÇÃO E COMUNICAÇÃO CIENTÍFICA NO DOMÍNIO DAS CIÊNCIAS NUCLEARES: O CASO DO INIS	145
6.2 ANÁLISE DE DOMÍNIO SOB A ABORDAGEM DA ESTRUTURA DAS INSTITUIÇÕES CIENTÍFICAS: A CNEN E O IEN	147
6.2.1 <i>A Comissão Nacional de Energia Nuclear – CNEN</i>	147
6.2.2 <i>O Instituto de Engenharia Nuclear – IEN</i>	149
6.2.2.1 <i>Áreas temáticas de pesquisa do IEN</i>	150
6.2.2.2 <i>A Plataforma Carpe diEN</i>	153
7 RESULTADOS E DISCUSSÕES	157

7.1 DIRETRIZES PARA UMA POLÍTICA DE CURADORIA DIGITAL DE DADOS DE PESQUISA PARA O PAÍS.....	158
7.2 CONSIDERAÇÕES ACERCA DO MODELO DE PUBLICAÇÃO AMPLIADA PROPOSTO.....	171
7.2.1 <i>Mapeamento das atividades de pesquisa do IEN</i>	177
7.2.1.1 Mapeamento dos temas abordados nas pesquisas desenvolvidas no IEN	178
7.2.1.2 Mapeamento de dados de pesquisa gerados no IEN	184
7.2.1.3 Mapeamento de e-prints gerados no IEN	189
7.2.1.4 Atores: organização e pessoa	191
7.2.1.5 Entidades e relações	192
7.3 PROPOSTA DE INTERFACES PARA PUBLICAÇÃO AMPLIADA.....	205
7.3.1 <i>Interface de Autoria/Edição de publicações ampliadas</i>	205
7.3.2 <i>Fluxo de autoria de publicação ampliada</i>	207
7.3.3 <i>Interface de Apresentação de publicações ampliadas</i>	211
7.3.4 <i>Serviços integrados ao modelo de publicações ampliadas</i>	218
7.4 MODELO DE PUBLICAÇÃO AMPLIADA PARA O IEN	220
8 CONCLUSÃO E PERSPECTIVAS FUTURAS	226
APENDICE A - PROPOSTA DE CLASSIFICAÇÃO DE RELAÇÕES PARA PUBLICAÇÃO AMPLIADA	246
APÊNDICE B – FORMULÁRIO PARA LEVANTAMENTO DE DADOS E DOCUMENTOS PRODUZIDOS NO IEN	248
APENDICE C – RELAÇÕES CONCEITUAIS APLICADAS AO IEN	250
APÊNDICE D - PROPOSTA DE TAXONOMIA DE RELAÇÕES PARA USO EM PUBLICAÇÕES AMPLIADAS	252

1 INTRODUÇÃO

As motivações que levaram ao desenvolvimento deste trabalho iniciaram-se a partir da atuação profissional em uma instituição de pesquisa. Conviver com pesquisadores no dia a dia, acompanhar seus anseios e angústias durante o desenvolvimento de suas pesquisas e suas expectativas em relação aos resultados, que muitas vezes se traduzem em publicação, citação, financiamento para novas pesquisas entre outros, levam a diversas reflexões sob o fazer do profissional da informação.

Para o profissional da informação, conviver em um ambiente de pesquisa é uma experiência riquíssima, pois, neste ambiente, a informação é um recurso valioso que precisa ser preservado, recuperado e disseminado corretamente. Neste ambiente, as ações dos profissionais da informação são importantes, no decorrer da pesquisa, para auxiliar na descoberta de novos recursos informacionais e, no final da pesquisa, para auxiliar no armazenamento persistente e na disseminação dos resultados da pesquisa. No entanto, assim como em qualquer campo de atuação, este profissional precisa estar totalmente ligado às mudanças, principalmente oriundas dos avanços tecnológicos que permeiam esta era e que exercem forte impacto na prática científica.

Assim, o primeiro sinal apreendido e que levou ao desenvolvimento deste trabalho foi a diminuição do número de usuários presenciais nas bibliotecas técnico-científicas. Esse tipo específico de biblioteca, que tem como usuário principal, durante os últimos anos, os pesquisadores de uma instituição de pesquisa, vem sendo menos frequentada e o motivo não está na ausência de consumo de informação, pois ao contrário, pesquisadores utilizam cada vez mais informação, porém o fazem diretamente de seus próprios computadores.

Ferramentas como bases de dados *on line*, bibliotecas digitais, Portal de Periódicos CAPES¹, Google Scholar², entre outros, têm oferecido ao pesquisador a possibilidade de acessar de forma mais ágil a informação necessária. Além disso, a própria tecnologia permite armazenar, em seus próprios computadores ou *gadgets*, as informações geradas no decorrer de suas atividades de pesquisa, além de outras informações, ditas bibliográficas, e que até pouco tempo atrás só era possível obter diretamente na biblioteca.

¹ CAPES - Coordenação de Aperfeiçoamento de Pessoal de Nível Superior. Disponível em: <<http://www.periodicos.capes.gov.br>>. Acesso em: 14 maio 2014.

² Disponível em: <<http://www.scholar.google.com.br>>. Acesso em: 14 maio 2014.

Mas, se os pesquisadores não vão mais às bibliotecas, é correto dizer que eles não precisam mais do profissional da informação? A resposta é não. Ao contrário, pesquisadores estão cada vez mais perdidos em dilúvios de dados e informação e precisam de auxílio para gerenciar o conhecimento disponível.

As mudanças pelas quais a atual biblioteca científica se depara são decorrentes de mudanças ocorridas no próprio fazer científico.

De acordo com Hey, Tansley e Tolle (2009) há aproximadamente mil anos havia apenas a ciência empírica, cuja forma de execução estava pautada na descrição de fenômenos naturais. Em seguida, veio a ciência teórica, que por meio do uso de modelos e generalizações, formulava leis e equações de grande abrangência.

Com o passar do tempo, a dificuldade de resolução dos modelos teóricos e empíricos foi crescendo e, como consequência, nas últimas décadas, o modo de fazer ciência passou a ser o da simulação via computadores. Este terceiro modelo, muito comum nos dias de hoje, promove um crescimento exponencial dos dados científicos, fazendo surgir no meio da ciência um quarto paradigma, que vem sendo chamado de *eScience*.

O quarto paradigma objetiva a exploração de dados e a unificação das teorias, experimentos e simulações utilizando metodologias e ferramentas voltadas para o gerenciamento e reuso desses dados. Portanto, pode-se afirmar que a quantidade cada vez maior de dados gerados, e a possibilidade de reutilizá-los de forma interdisciplinar e inovadora, têm alterado a forma como a ciência evolui.

A definição clássica apresentada por Borko (1968), preconiza que a Ciência da Informação, enquanto disciplina, está interessada diretamente no “corpo de conhecimentos relativos à origem, coleção, organização, armazenagem, recuperação, interpretação, transmissão, transformação, e uso da informação”, investigando “as propriedades e o comportamento da informação, as forças que governam seu fluxo, e os meios de processá-la para otimizar sua acessibilidade e seu uso.”.

Investigar este novo fenômeno de geração de dados e informações no âmbito da ciência bem como propor novas formas de coletar, organizar, armazenar, recuperar estes dados e ainda atividades como interpretação, transmissão, transformação e uso da informação - que atualmente vem sendo denominada como reuso de informação - é papel da Ciência da Informação. Desta forma, percebe-se que a Ciência da Informação enquanto disciplina e por meio de seus estudos relacionados à organização do

conhecimento e recuperação da informação, mais uma vez, se destaca, no que diz respeito ao desenvolvimento da Ciência.

As necessidades dos pesquisadores originadas a partir das novas práticas do fazer científico revelam a necessidade de criação de serviços originais de informação que contemplem as novas técnicas científicas. Nesta conjuntura, a proposta desta pesquisa vem atender essa necessidade trazendo para o âmbito da Ciência da Informação discussões dos conceitos “Repositórios”, “Objetos digitais Complexos” “Curadoria digital de dados de pesquisa” reuso de dados de pesquisa e, principalmente, “novos modelos de publicações”.

Na literatura brasileira da área de Ciência da Informação, pouco se discute sobre os temas acima citados, com exceção para o tema “Repositórios” que tem sido foco de interesse por diversas comunidades e instituições. O repositório que este trabalho pretende tratar, no entanto, é de um tipo diferente, pois é desenvolvido para armazenar, além das publicações científicas, os dados de pesquisa relacionados a essas publicações.

O tratamento fornecido a esses dados de pesquisas, no ambiente de um repositório com a finalidade de gestão, preservação e futuro reuso é o que se pretende através das técnicas de curadoria digital.

Por serem os dados de pesquisas objetos digitais informacionais distintos em termos de complexidade, diversidade e de significados em relação aos objetos digitais informacionais tradicionais, como por exemplo, um artigo de periódico apresentado em formato PDF, é necessário, para elucidar esta pesquisa, uma análise aprofundada dos conceitos que envolvem a questão dos objetos digitais informacionais. Estes conceitos são *eScience*, curadoria digital de dados de pesquisa, relacionamentos semânticos e o próprio objeto desta pesquisa, isto é: a publicação ampliada

1.1 JUSTIFICATIVA

No panorama da comunicação científica, a publicação pode ser considerada uma forte moeda para a ciência. É ela o critério principal para o estabelecimento da prioridade de uma descoberta, tornando o status da publicação um fator importante na resolução de litígios prioritários ou de reivindicações de propriedade intelectual. Prestígio acadêmico e decisões de promoção são baseados em grande parte na publicação em periódicos ou de coletâneas revisados por pares (KIRCZ, 2001). No entanto, há sinais significativos de que este principal veículo de disseminação do conhecimento gerado pelos pesquisadores e

acadêmicos esgotou a sua capacidade de refletir com fidedignidade a complexidade, a diversidade e a sofisticação da atividade científica contemporânea.

Por um lado, a comunidade científica deseja veículos tecnologicamente mais avançados, que proporcionem velocidade ao ciclo de comunicação científica, interatividade, acesso aberto, além de compartilhamento e reuso de dados e conjunto de dados. Por outro lado, as transformações comportamentais e sociais, decorrentes dos aparatos tecnológicos que permeiam e dinamizam as atividades de pesquisa, ainda estão centradas no armazenamento e na disseminação de recursos informacionais individuais, ou seja, nas publicações que sintetizam os resultados das pesquisas, mas não revelam os dados que a originaram. Nas bibliotecas de pesquisa, por exemplo, constata-se que oferecem ao usuário basicamente um artigo ou uma monografia como resultado de uma busca. Outro fato é que “muitos editores acadêmicos não aceitam outro produto de projetos de e-pesquisa, tais como base de dados, gravação de vídeos e serviços *Web*” (VERHAAR, 2008, p.9).

De acordo com Kircz (2002, p.28) “O maior avanço na comunicação científica é que agora somos capazes de usar um único veículo para todas as expressões possíveis do conhecimento científico”. É preciso então se valer de todos os artifícios trazidos pelo advento da tecnologia para fazer com que a comunicação científica siga para além de um documento simples. O novo padrão de produção de conhecimento científico, baseado na geração intensiva de conjunto de dados, demanda tipos inéditos de publicações que consigam integrar dados de toda natureza e publicações tradicionais em formatos digitais, criando um novo gênero de publicação *Web*. Atualmente, isto é possível, pois ao “traduzir o conhecimento em código binário, criamos um mono-meio que nos permite integrar todos os tipos de representações” (KIRCZ, 2002). Isso significa dizer que não importa se o documento é um artigo, um dado ou outro documento qualquer, se ele estiver em meio digital, eles são todos, na sua essência, códigos binários.

Com a chamada *eScience*, ou quarto paradigma científico, fica patente que a adição de outros recursos ao texto, como imagens, sons e interatividade, agora se torna fundamental. Assim como “as impressões de alta qualidade permitiu um avanço em herbários e atlas anatômico, a introdução de sons e simulações permite-nos apresentar informações relevantes para o leitor de uma forma muito mais realista.” (KIRCZ, 2002).

O fato de cada vez mais os dados de pesquisas estarem armazenados em repositórios digitais confiáveis³ e gerenciados sob os princípios da curadoria digital, sendo preservados e mantendo a sua capacidade de reuso, pode ser favorável à criação de uma nova infraestrutura de comunicação científica em que os conteúdos dos repositórios institucionais, ou seja, publicações científicas, sejam ligados aos conteúdos dos repositórios de dados. Kircz afirma que

embora a linguagem continue a ser um mecanismo de transferência essencial para troca de conhecimentos, a comunicação não linguística vai recuperar um pouco da proeminência que se perdeu quando a linguagem escrita possibilitou a comunicação científica surgir independente de espaço e do tempo. (KIRCZ, 2002, p.28)

A comunicação não-linguística a que Kircz se refere são os demais recursos audiovisuais e interativos - que algumas vezes aparecem na forma de dados de pesquisa, outras como Fontes relacionadas para compreensão da temática da pesquisa – que esta nova infraestrutura de disseminação da informação científica pode oferecer como suporte para o desenvolvimento de pesquisas inovadoras.

Para Seringhaus e Gerstein (2007), o foco da atenção está mudando da publicação simples para uma estrutura de publicação maior em que um *link* deve apoiar e refletir a relação entre a publicação e todos os objetos relevantes como dados, recursos *Web*, comentários etc. Este padrão está se tornando mais e mais complexo e indo para além de uma estrutura linear ou sequencial como é o modelo de publicação tradicional. Atualmente vários estudos se concentraram na possibilidade de se entrelaçar produtos de e-pesquisa que se encontram distribuídos, gerando novas modalidades de publicações científicas. De acordo com Pinheiro

a informação científica e tecnológica é parte fundamental da infraestrutura de C&T. Portanto, abordar a comunicação científica significa não somente enfocar padrões de comunicação entre pares, mas também englobar tanto a informação à qual recorrem para as suas pesquisas, quanto aquela que produzem e transmitem por diferentes canais de comunicação e tipos de documentos. (PINHEIRO, 2003, p.62)

Assim, fica claro a importância de se ter ligado às publicações, dados de todos os tipos que, no contexto atual, se configuram como objetos digitais. Desta maneira, a

³ Repositórios Digitais confiáveis – repositórios que têm como “missão oferecer à sua comunidade-alvo acesso confiável e de longo prazo aos recursos digitais por ela gerenciados, agora e no futuro” (RESEARCH LIBRARY GROUP / ONLINE COMPUTER LIBRARY CENTER - RLG/OCLC, 2002, p.5).

proposta de um novo modelo de publicação se justifica pela necessidade de integrar por meio de relações semânticas esta ampla diversidade de objetos digitais acadêmicos, possibilitando o registro de toda a memória digital da pesquisa desenvolvida, bem como uma recuperação integrada de toda a informação a ela relacionada. Este modelo está sendo chamado nesta pesquisa de Publicação Ampliada.

1.1.1 Por que desenvolver um modelo de Publicações Ampliadas para a área de Ciências Nucleares?

A área de Ciências Nucleares é uma das líderes em geração intensiva de dados. O projeto LHC (sigla para *Large Hadron Collider* ou Grande Colisor de Hádrons) do CERN - European Organization for Nuclear Research, é o principal exemplo de atividade desenvolvida na área de ciências nucleares com geração intensiva de dados. De acordo com o *Science & Technology Facilities Council* - STFC⁴, este projeto possui três focos: o desenvolvimento do colisor em si, os detectores e o *grid*. Enquanto os detectores espalhados pelo colisor identificam e medem o momento e a carga das partículas, o *grid* processa esses dados detectados, registra e os compartilha entre a comunidade internacional interessada.

Uma simples consulta à tabela de assuntos⁵ coberta pelas publicações da Agência Internacional de Energia Atômica – organismo mais representativo internacionalmente nas áreas de Ciências Nucleares – fornece um panorama esclarecedor da amplitude da área nuclear, de sua característica multi e interdisciplinar e de sua abrangência em termos científico, tecnológico, político e social dos problemas e desafios enfrentados pelos pesquisadores e tecnólogos nucleares.

CIÊNCIAS DA VIDA

Medicina nuclear; Biologia das radiações; radiofármacos; física médica; dosimetria

ALIMENTOS E AGRICULTURA

⁴ Disponível em: <<http://www.lhc.ac.uk/>>. Acesso em: 26 maio 2013.

⁵ Disponível em: <http://www-pub.iaea.org/books/IAEABooks/Subject_Areas> Acesso em: 26 maio 2013.

Irradiação de alimentos; controle de pragas; melhoramento genético de plantas; biotecnologia de plantas; fertilização e irrigação de solos; agroquímicos; produção e saúde animal.

MEDIDAS NUCLEARES, TÉCNICAS E INSTRUMENTAÇÃO

Física; química; técnicas de dosimetria; técnicas analíticas nucleares; aplicações de reatores de pesquisa e aceleradores; dados nucleares.

CIÊNCIAS DA TERRA

Geologia, mineração e mineração de urânio; hidrologia.

APLICAÇÕES INDUSTRIAIS

Processos de radiação; radiometria; traçadores.

SEGURANÇA NUCLEAR E RADIOLÓGICA

Mineração e beneficiamento do urânio; fabricação e armazenamento de combustível; usinas nucleares; Fontes de radiação e aceleradores; transporte de material radioativo; reatores de fusão; reatores de pesquisa; repositórios de rejeitos; proteção radiológica; resposta a acidentes; gestão de rejeitos radioativos; análise de segurança; gestão de qualidade; aspectos legais e governamentais.

NUCLEAR POWER

Planejamento e economia da energia nuclear; operação da geração de energia nuclear; tecnologia de reatores; garantia de qualidade; capacitação e treinamento de pessoal.

CICLO DO COMBUSTÍVEL E GESTÃO DE REJEITOS

Processamento do minério do urânio; fabricação e desempenho do combustível nuclear; gestão do combustível gasto; gestão de rejeitos radioativos.

FÍSICA DE PLASMA E FUSÃO

SALVAGUARDA

ASPECTOS LEGAIS

MEIO AMBIENTE

PROTEÇÃO FÍSICA E MATERIAL RADIOATIVO.

A Tabela de assunto da base de dados INIS ainda é mais abrangente, incluindo áreas do conhecimento como Fontes convencionais de energia, por exemplo, carvão, gás, petróleo, xisto, etc.; e ainda Fontes alternativas, tais como energia solar, eólica, hidrogênio, biomassa, marés, geotérmica etc. O planejamento, política e economia da energia é um item importante, especialmente quando tratam de análises e estudos comparativos entre a energia nuclear e as demais formas de geração de energia.

A base de dados INIS dedica ainda espaço relevante às disciplinas básicas como Física, Química e Engenharia que dão sustentação aos processos nucleares, o mesmo acontece com as áreas de computação e métodos matemáticos que são imprescindíveis nos cálculos, simulações e na geração de imagens.

Os desafios de se obter novos materiais é um item crítico para as tecnologias nucleares e tem um reflexo claro na base INIS. Há ainda espaço reservado para as novas áreas de estudo como nanotecnologia e nanociências. Por outro lado, a gestão do conhecimento nuclear ganha status de um item importante dentro dos temas de pesquisa da área nuclear, ampliando ainda mais o escopo da área.

Por fim é necessário reafirmar que as aplicações da energia nuclear na medicina, na produção de energia, na agricultura e pecuária, no meio ambiente, na produção e conservação de alimentos e no combate a pragas, bem como os aspectos de segurança são os itens de relevância crítica e de forte impacto na sociedade e que tem um rebatimento claro nas pesquisas desenvolvidas na CNEN e, particularmente, no IEN.

Nesta direção uma breve análise das linhas de pesquisa do Instituto de Engenharia Nuclear –IEN, um dos institutos de pesquisa da Comissão Nacional de Energia Nuclear do Brasil - CNEN, lócus desta pesquisa - mostrou uma grande geração de dados científicos de todos os tipos, mas principalmente de dados experimentais e dados gerados por simulação. Estes dados permanecem armazenados nos computadores pessoais dos pesquisadores, porém, sem nenhum tratamento específico tem o seu potencial de uso e reuso sensivelmente diminuído; além do mais, expostos à fragilidade intrínseca do meio digital e da adolescência tecnológica são freqüentemente perdidos para sempre.

A criação de uma infraestrutura que possibilite a preservação, o tratamento, a recuperação e o acesso compartilhado a esses dados vem sendo uma preocupação do IEN e, entre as necessidades identificadas, a contextualização dos dados a partir da ligação dos mesmos com os resultados das pesquisas publicados é uma necessidade que urge e merece atenção especial.

Neste sentido, a fim de tornar mais fácil para os pesquisadores da área de Ciências Nucleares encontrarem publicações e objetos relacionados (como por exemplo, os dados primários, originais ou não-processados) que possam ser reusados para novas pesquisas, um novo modelo de publicação deve ser desenvolvido, de forma que a ligação entre publicações e dados seja feita por meio de relações conceituais consistentes, o que acarretará em recuperação precisa e acesso compartilhado a um conjunto de informações sobre a mesma temática.

É importante mencionar também que apesar da existência de iniciativas que visam reunir a memória técnico-científica nuclear em âmbito local (como o Projeto Memória CNEN)⁶ e em âmbito internacional (base de dados bibliográficos do INIS)⁷, nenhuma delas promove a ligação semântica entre a produção científica e dados de pesquisa o que dificulta a visualiação da pesquisa de forma global. Integrar os dados brutos, os dados processados, os documentos correlatos aos resultados publicados em *e-prints* é uma forma de registrar a memória da pesquisa desenvolvida. Assim, um modelo de publicação ampliada pode ser considerado tanto uma novidade para a área de Ciências Nucleares, bem como um campo de estudo original para a Ciência da Informação que volta a sua atenção para o tratamento de dados, colocando-a par e passo com os desafios que envolvem a

⁶Disponível em: <<http://memoria.cnen.gov.br>> Acesso em: 14 maio 2014.

⁷Disponível em: <<http://www.iaea.org/inis/>> Acesso em: 14 maio 2014.

publicação científica e se aproximando da orientação por dados preconizada pela web semântica.

1.1.2 Para que ter relações bem definidas?

A ideia que está subjacente às publicações ampliadas é criar elos entre o conteúdo dos repositórios digitais, ou seja, as publicações científicas digitais, com os dados digitais que dão sustentação a essas publicações. (VERHAAR, 2008). Desta forma, a publicação ampliada pode ser considerada como uma instância de objeto digital complexo que combina vários recursos heterogêneos relacionados.

Em estudo anterior sobre relações conceituais em ontologias (SALES, 2006) verificou-se a importância de um sistema de informação possuir relações semanticamente bem estabelecidas. Em qualquer sistema de recuperação de informação, as relações podem aparecer em meio às linguagens documentárias que representam suas informações tematicamente, por exemplo, tesouros, vocabulários controlados e/ou ontologias ou ainda na base da programação do próprio sistema.

Em uma página *Web*, por exemplo, a navegação pode ser considerada mais amigável se seus *hiperlinks* forem criados por meio de relações semânticas. Em banco de dados relacionais, as relações conceituais bem estabelecidas também podem melhorar a precisão das informações recuperadas.

No âmbito da *Web* semântica, linguagens de representação como o RDF (*Resource Description Framework*) ou o OWL (*Ontology Web Language*) são instâncias de modelos em que as relações são insumo essencial para prover semântica aos dados. No caso do RDF, essas relações se manifestam sob a forma de “*property*” – propriedades que se afirmam sobre um dado e as ligam ao recurso *Web*. Mais recentemente a ideia do *linked open data* vem reafirmar o RDF enquanto padrão e, conseqüentemente, a importância das relações conceituais bem estabelecidas.

Sob outra perspectiva, mas ainda relacionada à importância das relações semânticas, Biolchini (2001) evoca a importância das associações na formação de redes de conhecimento e no estabelecimento de modelos:

uma vez que se destinam a auxiliar os processos de raciocínio e de tomada de decisão dos profissionais, os sistemas de informação devem ser desenvolvidos visando ampliar as potencialidades, não somente de busca e recuperação de unidades informacionais, como também de organização e apresentação de unidades semanticamente relacionadas, ampliando a recuperação e evocação de unidades de sentido e de conjuntos informacionais relevantes e significativos. Este desenvolvimento da organização de relações de significação pode ser conduzido através da construção de redes semânticas entre os elementos constitutivos das bases de conhecimento, compreendidos pelos termos que se encontram presentes nas mesmas bem como pelos componentes de suas respectivas definições, de modo a se estabelecer nós e grafos de associação entre tais unidades constituintes. (BIOLCHINI, 2001, p.11)

Quando o problema é representar o conhecimento, as ontologias constituem como um importante instrumento para essa finalidade. No entanto, no próprio escopo do desenvolvimento de ontologias, a questão das relações ainda é problemática. (SALES, 2006).

Uma constatação importante relatada nos estudos de Dahlberg (1978a) sobre as relações é que estas podem variar de acordo com o contexto para o qual a linguagem é construída, sendo assim faz-se necessária uma investigação que concentre seus esforços na área de aplicação desta pesquisa: a área de Ciências Nucleares.

O pressuposto desta tese é que a publicação ampliada será mais rica semanticamente, se a criação de elos entre a pesquisa e seus dados for feita por meio de relações conceituais explícitas. Como a publicação ampliada é um artefato tecnológico que visa reunir objetos digitais de diversas naturezas, porém de uma mesma temática, a recuperação dessa agregação conceitual de forma integrada será mais precisa se as relações revelarem sua semântica adequadamente.

1.1.3 *Por que propor uma política de curadoria digital para o país?*

Uma questão que também se coloca nesta tese é a necessidade de curadoria dos dados de pesquisa que deverão estar integrados na formação de uma publicação ampliada. Esta necessidade, porém extrapola os limites de um domínio, como é o caso das Ciências Nucleares, e requer diretrizes que possam ser seguidas em âmbito nacional.

Essa necessidade se justifica por que o modelo de publicação proposto por esta tese é, em uma primeira instância, um objeto digital complexo, podendo ser formado por meio de relacionamentos entre recursos distribuídos em diversos repositórios digitais, saindo

assim do alcance de uma única instituição. Uma publicação ampliada, enquanto um objeto digital complexo, pode se configurar como a combinação de um *e-print* depositado em um repositório institucional, dados de pesquisa depositados em um repositório de dados e ainda outros documentos depositados em qualquer outra base de dados de informação técnico-científica.

Neste sentido, propor diretrizes para uma política de curadoria digital para o país significa estabelecer em esfera nacional, normas e padrões comuns que propiciem a cooperação entre as instituições de pesquisa e a interoperabilidade entre esses sistemas que abrigam os recursos digitais de pesquisa (*e-prints*, dados de pesquisa, etc). Dito de outra maneira, esses recursos digitais precisam estar tratados adequadamente para que possam servir como elementos para serem agregados na formação de uma publicação ampliada.

Apesar de essa pesquisa estar voltada para aplicação em um domínio específico, a proposição do modelo deverá considerar explicitamente as estruturas envolvidas na formulação de políticas de C&T no país e aspectos geopolíticos, como legislação, financiamento, pesquisa, educação, capacitação, infraestruturas tecnológicas e de gerenciamento de C&T etc. A infraestrutura concebida como contextualização do modelo proposto serve também para prover uma contribuição para a formação de um ambiente de curadoria para os países. Assim, enquanto uma política nacional não é estabelecida, as diretrizes propostas podem servir como elementos norteadores para as instituições locais que pretendem dar algum tipo de tratamento aos dados de pesquisa, sejam para fins de preservação, de gestão, de disseminação via novos modelos de publicação ou qualquer outra finalidade.

De uma forma mais genérica, a presente pesquisa se justifica principalmente pelas alterações ocorridas na produção científica, caracterizada essencialmente pela chegada de um novo paradigma, chamado *eScience*, suas conseqüentes alterações no ciclo da comunicação científica eletrônica e a necessidade de novas formas de disseminação do conhecimento.

1.2 PROBLEMA E QUESTÕES DE PESQUISA

Pesquisadores e acadêmicos sempre procuraram incorporar os desenvolvimentos tecnológicos na intermediação e intensificação das suas relações de compartilhamento de

conhecimentos. Os pressupostos de inovação e modernidade da ciência, aliados à própria natureza social inerente à construção do saber científico, colocam as tecnologias de informação e comunicação (TICs) numa condição determinante para o estabelecimento de novas modalidades de socialização entre pesquisadores.

Iniciando com extensões e simulacros virtuais de formas tradicionais de comunicação como correio eletrônico, telefone e reuniões presenciais, as TICs, nos dias atuais, ultrapassam os limites da comunicação e recriam novas formas de fazer ciência, como são os laboratórios⁸ e os experimentos levados a cabo de forma distribuída em escala universal, baseados em redes de computadores.

O que se observa de forma concreta é que os pesquisadores têm colocado luz sobre problemas que estavam despercebidos e que, entretanto, têm desdobramentos significativos na estrutura clássica da produção científica.

Se a forma de fazer ciência mudou, fica patente que a forma como lidar com esses registros de conhecimento precisa acompanhar essas mudanças, sendo necessário pensar novas formas de registro, tratamento e recuperação e disseminação do conhecimento produzido nesses contextos mais dinâmicos.

Quanto às novas formas de registro, o que se observa na prática é que os modelos de documentos eletrônicos com os quais lidamos atualmente são apenas uma reprodução digital do modelo tradicional impresso. Um artigo de periódico em formato PDF (Portable Document Format) é uma cópia digital fiel de um artigo em uma revista impressa, isto é, de leitura totalmente linear, sem a possibilidade de qualquer tipo de interação por parte do leitor.

Modelos alternativos de publicações são propostos desde há muito tempo por pesquisadores da área de Documentação e Ciência da Informação. Podem ser citados como exemplo a estação de estudo imaginada por Otlet (1934) em sua “*machine to think with*”, o Memex de Vannevar Bush (1945), o Xanadu de Ted Nelson (1974), entre outros. No entanto, durante muito tempo essas ideias não puderam ser colocadas em prática, entre outras coisas, pela ausência de tecnologia de informação para tal. (SALES; SAYÃO, 2012)

⁸Colaboratório é um termo cunhado por Wulf (1989) para designar um “centro de pesquisa sem paredes, no qual os pesquisadores de um país podem realizar suas pesquisas sem considerar a localização física, interagindo com colegas, acessando instrumentação, compartilhando dados, informações e recursos computacionais, e acessando bibliotecas digitais”.

Dodebei (2002, p.13) cita como uma das grandes contribuições de Otlet a ampliação do conceito tradicional de documento para o conceito de bibliografia, atualmente nomeada base de dados. Com a ideia do princípio monográfico de decompor a obra em partes menores e organiza-las de forma temática e hierarquizada, Otlet, tal como Bush e Nelson, antecipa o conceito de hipertexto e se mostra atual no que envolve os novos modelos de documentos propostos na literatura que possuem como foco a prática da navegação, sejam aqueles que decompõem o documento em unidades menores, como aqueles que o mantem na íntegra, mas que propõem uma ampliação das publicações por meio da adição de recursos extras, como dados, metadados, referências e outros documentos correlatos. Esses novos modelos de publicação aparecem nomeados de diversas formas, a saber: publicações compostas, publicações sobrepostas, publicações semânticas, publicações ampliadas, entre outras.

Publicação composta é o nome dado para designar uma coleção de tipos diferentes de documentos. Em meio analógico, um exemplo de documento composto é um dossiê, que só tem sentido enquanto um todo formado por suas partes e estas partes podem ser de naturezas distintas. Em meio digital, as publicações compostas são encontradas com mais frequência, como são, por exemplo, as páginas *Web*, os multimídias, as bases de dados, etc. Na literatura, a publicação composta também aparece nomeada como documentos digitais compostos ou objetos digitais compostos. (CHEUNG et al., 2008)

Publicações sobrepostas são documentos criados para anotar outro documento com informação sobre a qualidade deste recurso, por exemplo, um documento de avaliação de um artigo, os *datasets* que embasam a pesquisa registrada no artigo ou até uma nova versão do mesmo artigo. (CALLAGHAN et al., 2010)

Publicações semânticas são documentos construídos em combinação com uma ontologia para permitir que o usuário acesse o conhecimento através de buscas inteligentes. (ERICKSON, 2007a).

Publicações ampliadas (VEHAAR, 2008) são tipos de publicações que devem reunir todos os tipos já citados. A publicação ampliada é uma instância de um objeto digital composto, cujo objetivo é ligar os resultados de pesquisa aos dados que o geraram, extrapolando o limite do documento físico. Essa ligação dos resultados publicados em *e-prints*⁹ com os dados faz a publicação ampliada aderir às características das publicações compostas, pois neste caso, ela se torna a reunião de vários tipos de documentos. Por outro lado, as publicações ampliadas, ainda podem ter características de documentos semânticos, pois nascem em meio ao contexto de uma nova *Web*: a *Web* semântica.

O *World Wide Web Consortium*, ou como mais conhecido, o *W3C* é um consórcio internacional que reúne especialistas, empresas, organizações governamentais ou independentes que têm por objetivo estabelecer padrões para criação e interpretação dos dados na *Web*: o que gera um forte impacto também na interoperabilidade e na recuperação dessas informações. Para o *W3C* Brasil (2001), a *Web* semântica é uma visão que é expressa pela “ideia de ter dados na *Web* definidos e ligados de uma maneira tal, que possam ser usados por máquinas, não só com o objetivo de apresentação, mas por automação, integração e reuso de dados entre aplicativos.”.

Para que essa ideia seja possível é preciso, entre outras coisas, que os dados publicados na *Web* tenham uma gestão mínima que garanta preservação e acesso. Essa gestão vem sendo realizada por meio de técnicas de curadoria digital, que envolvem o estabelecimento de políticas para guarda e preservação dos dados, bem como o uso de metadados e ontologias para tratamento e recuperação. Neste sentido, o problema que a presente pesquisa visa investigar é a identificação de uma forma semântica para ampliar a publicação científica integrando-a aos dados que a fundamentam, bem como a outros documentos que registrem a memória da pesquisa.

Pressupondo que uma publicação ampliada se constrói a partir da sistematização do conhecimento nela representado, este estudo busca responder às seguintes questões:

1) Em um ambiente de pesquisa em que os resultados gerados - dados e publicações - precisam estar interligados entre si, como é o caso das publicações ampliadas, como garantir semântica e interoperabilidade a esses novos modelos de publicação?

⁹ Este trabalho considera *e-prints* toda e qualquer publicação avaliada e disponibilizada em formato digital.

2) Relações conceituais podem ser o elo entre publicações científicas digitais, dados de pesquisa e outros documentos digitais?

1.3 PREMISSAS E HIPÓTESE

Esta pesquisa se baseia em duas premissas para a formulação de sua hipótese. A primeira pode ser sintetizada pela necessidade de um novo modelo de publicação que integre os dados gerados pela pesquisa aos resultados divulgados em *e-prints*, como exemplificado na situação a seguir:

Existe uma lista de possibilidades viabilizadas pelas publicações eletrônicas quando somadas às potencialidades das redes de computadores e suas aplicações mais avançadas, como a *Web* e a *Web* semântica. As possibilidades se ampliam quando são incorporadas às redes sociais e suas formas de socialização e intercâmbio de ideias, e ainda mais com a característica distribuída dos repositórios que reúnem essas publicações, sem que seja necessário que as informações estejam reunidas em um mesmo lugar.

Em um ambiente científico globalizado e cooperativo, a produção de conhecimento depende fortemente do compartilhamento de recursos e informações, proporcionado pelo acesso em tempo real aos experimentos. O grande colisor de hádrons - LHC - da *Organisation Européenne pour la Recherche Nucléaire* - CERN¹⁰, conforme já mencionado, é um exemplo atual da forma contemporânea de gerar conhecimento de forma compartilhada.

A avaliação por pares, um dos pilares tradicionais da validação dos novos conhecimentos, por sua vez, pode ser enriquecida em termos de velocidade, de interatividade contínua entre revisores e autores, e de fundamentação, tendo em vista que os dados em que a publicação se baseia poderão estar disponíveis em repositórios digitais apropriados, permitindo que a pesquisa seja conferida e legitimada.

A disseminação por muitos canais, com possibilidades de versões, manifestações, expressões da mesma obra, abre oportunidades extraordinárias para a participação dos leitores. A convergência das mídias que pode integrar na mesma publicação textos, vídeos, gráficos em 3D, animações, simulações, comportamentos, bem como, a interatividade

¹⁰Disponível em: <<http://public.Web.cern.ch/public/>>. Acesso em: 20 maio 2013.

proporcionada por blogs e redes sociais incorporados às publicações recriam um ambiente propício para colaboração, colocando em contato permanente leitores com autores, configurando uma rede entrópica de geração de novas idéias.

A segunda premissa em que este trabalho se embasa para formulação de sua hipótese está calcada nas possibilidades originadas pelas novas tecnologias, em especial, as decorrentes da *Web* semântica que impulsionam uma diversidade de aplicações voltadas para integração, recuperação e compartilhamento de informações de naturezas distintas.

Como já mencionado, a publicação ampliada se caracteriza por ligar uma publicação científica aos dados que geraram os resultados nela publicados. Esta ligação pode ser feita por meio de *hiperlinks*, contudo, seus rótulos devem revelar relações que expressem alguma semântica. Essa posição é defendida porque uma publicação ampliada, assim como qualquer outro artefato tecnológico alternativo que se pense atualmente, deve ser planejado para ser inserido no contexto da *Web* semântica.

Tim Berners-Lee, criador da *Web* atual, juntamente com outros interessados em melhorar o uso da *Web*, criou em 2001 o consórcio W3C e propôs a *Web* semântica como uma nova geração da *Web* atual com a finalidade de “fornecer estruturas e dar significado semântico ao conteúdo das páginas *Web*, criando um ambiente onde agentes de *software* e usuários possam trabalhar de forma cooperativa” (BERNERS-LEE; HENDLER; LASSILA, 2001).

Assim, como se pode observar, a proposta da *Web* semântica é melhorar a *Web* através da ligação semântica de dados que podem estar distribuídos e interligados por meio de diversos aplicativos. Nesse sentido, a proposta das publicações ampliadas adapta-se à visão da *Web* semântica, a partir do princípio de que publicações devem ser ligadas aos dados que dão sustentabilidade à pesquisa divulgada.

Interessante observar também, no que tange à diferença da proposta da *Web* atual para a *Web* semântica, a primeira é classificada como uma *Web* voltada para documentos enquanto a segunda é classificada como a *Web* dos dados. Assim, mais uma vez, pode ser ressaltada a adequação da proposta das publicações ampliadas.

É importante destacar ainda que a visão da *Web* semântica sugere que repositórios de dados sejam criados, mas também que vocabulários e regras sejam estabelecidos para que estes repositórios interoperem. Pode-se citar como exemplo de tecnologia para

‘linkagem’ de dados: RDF¹¹, SPARQL¹², OWL¹³, SKOS¹⁴. (W3C BRASIL¹⁵).

No que se refere ao vocabulário, a *Web* semântica vem se utilizando do conceito de ontologias para descrever propriedades e relacionamentos sobre itens, através de regras de inferências. Berners-lee, Hendler e Lassila (2001) conceitualizam ontologia como um “documento ou um arquivo que define formalmente as relações entre os termos, sendo formada por uma taxonomia e regras de inferências”. Sendo assim, não basta ligar dados, mas é preciso representá-los conceitualmente, explicitando a semântica dessas relações. Dito de outra forma, são as relações especificadas claramente que permitirão que a *Web* atual se torne a *Web* idealizada, onde, através do compartilhamento de informações entre aplicativos, usuários poderão obter como resposta a uma consulta *Web*, não apenas documentos, mas respostas precisas a questões complexas.

Ter relações explicitadas semântica e formalmente torna-se tão importante para o escopo da *Web* semântica que em quatro das cinco linhas de atuação do consórcio W3C (dados ligados, vocabulários e ontologias, inferências, consultas e aplicações verticais) as relações são mencionadas como necessárias.

Os *hiperlinks* que ligam publicações, dados e metadados são relações que devem ser reveladas como em uma ontologia para que estas publicações ampliadas possam, não apenas ser recuperadas, mas também processadas conceitualmente possibilitando inferência pela máquina e geração de conhecimento.

Assim, com base nas premissas apresentadas, este trabalho defende a hipótese de que uma publicação pode ser ampliada a partir de um modelo que ligue os dados de pesquisa ao documento final por meio de relações que expressem uma semântica mais precisa.

¹¹ RDF ou Resource Description Framework é uma linguagem para representação da informação na Web.

¹² SPARQL ou SPARQL Protocol and RDF Query Language – é uma linguagem de consulta para bancos de dados, capaz de recuperar e manipular dados armazenados em RDF.

¹³ OWL ou Ontology Web Language é uma linguagem para descrição de ontologias.

¹⁴ SKOS ou Simple Knowledge Organization System são normas e padrões recomendados pelo W3C para apoiar o uso de sistemas de organização do conhecimento, tais como vocabulários controlados, taxonomias, tesouros no âmbito da Web semântica.

¹⁵ Disponível em: <<http://www.w3c.br/Home/WebHome>>. Acesso em: 2 abr. 2014.

1.4 OBJETIVOS

OBJETIVO GERAL

Investigar elementos conceituais, teóricos e práticos das áreas de *eScience* e de curadoria digital de dados de pesquisa visando a proposição de um modelo semântico com base em relações conceituais para estruturação de publicação ampliada para a área de Ciências Nucleares.

OBJETIVOS ESPECÍFICOS

- Investigar conceitos, padrões, tecnologia e modelos de informação que envolvam curadoria digital de dados de pesquisa no ambiente da *eScience*.
- Investigar aspectos tecnológicos, políticos, sociais, educacionais e de pesquisa para a composição de um modelo de curadoria digital de dados de pesquisa.
- Analisar novos modelos de publicação científica que atendam às necessidades do quarto paradigma científico *eScience*, bem como formas de interligação de dados de pesquisa e *e-prints* para recuperação, acesso e reuso.
- Investigar e analisar tipos de relações conceituais adequadas à elaboração de um novo modelo de publicação científica para a área de Ciências Nucleares.
- Investigar e identificar elementos para compor uma publicação ampliada para a área de Ciências Nucleares.

2 CONTEXTO TEÓRICO

Este capítulo tem por objetivo apresentar o contexto no qual se desenvolve esta tese, bem como esclarecer sobre alguns conceitos que dão sustentação à pesquisa. Inicia-se a partir da explicação do que essa tese vai considerar como modelo, passando pela contextualização do tema que se dá mediante uma nova forma de fazer ciência chamada *eScience*. Esse novo fazer científico nasce a partir da união de três paradigmas: o paradigma da ciência experimental, o paradigma da ciência teórica e o paradigma da grande geração de dados. A contextualização deste capítulo é de fundamental importância para a compreensão dos demais conceitos de que trata esta tese e que são discutidos nos capítulos que seguem.

2.1 MODELO GRÁFICO APOIADO POR BASE TEÓRICA CONCEITUAL

A presente pesquisa é fruto de estudos que caminham em direção ao estabelecimento de um novo modelo de publicação científica que usufrua dos benefícios que os avanços tecnológicos trazem para o ambiente de pesquisa, como é o caso da publicação ampliada.

A discussão do que se entende por modelo se faz importante por ser este um conceito que pode ser compreendido de diversas formas, de acordo com a área e com o propósito. É possível encontrar trabalhos envolvendo discussões sobre modelos e modelagem, nas mais diversas áreas. Toda ciência se beneficia das abstrações construídas sobre a noção de modelo, da Física à Filosofia, da Medicina às Ciências Políticas.

Esse interesse se justifica pela dificuldade do homem entender a complexidade da realidade do universo que o envolve. Assim, em uma primeira instância, o ser humano elabora modelos para: compreender o mundo ou simplesmente uma questão no mundo; estabelecer padrões de comunicação entre ele e outros seres, ou ainda para representar de forma simplificada um objeto ou uma situação no mundo.

Neste sentido, um modelo pode ser compreendido como uma "representação abstrata e simplificada de um sistema real, com a qual se pode explicar ou testar o seu comportamento, em um todo ou em partes" (COUGO, 1997, p.7). Esta abstração ou simplificação é feita através de representações de fatias da realidade com a finalidade de

“descrever formalmente alguns aspectos físicos e sociais do mundo que nos rodeia para fins de compreensão e comunicação”. (MYLOUPOULOS, 1990)

Nas áreas das ciências exatas, por exemplo, os modelos são representações da realidade ou de fenômenos que, na maioria dos casos, não podem ser vistos a olho nu, mas que precisam ser compreendidos e teorizados. Nesse contexto, os modelos podem ser físicos (protótipos ou representações) ou matemáticos (representações dadas por equações). (WEGNER; GOLDIN, 1999)

As áreas tecnológicas também possuem um conceito próprio de modelo, sendo este o resultado do processo chamado modelagem. Em geral, a abordagem pode ser de três tipos: Modelagem Conceitual, Modelagem Lógica e Modelagem Física. (COUGO, 1997). Esses processos são voltados para a elaboração de *software* e podem ser compreendidos como etapas sequenciais do desenvolvimento de um sistema. Nesse contexto, a modelagem conceitual é a etapa em que se mapeia o ponto de vista do usuário, descrevendo uma dada realidade e o objetivo do sistema. A modelagem lógica é a etapa seguinte em que a descrição é representada por meio de axiomas lógicos, já a modelagem física é a implementação de acordo com o sistema de gerenciamento de banco de dados (SGBD) escolhido.

No domínio da Ciência da Informação, um modelo é definido como “uma criação destinada a representar uma realidade e alguns de seus aspectos, a fim de torná-los descritíveis e algumas vezes observáveis” (SAYÃO, 2001, p.83). Estes modelos podem ser construídos “por meio de formalismos matemáticos, fenomenológicos ou conceituais” e permitem “testar hipóteses, tirar conclusões, caminhar no sentido da generalização e da particularização, através de processos de indução e tem sempre vida provisória”.

No domínio da Ciência da Informação, os modelos são construídos para servirem como padrões de representação de informações, tornando a recuperação e comunicação mais precisas nos sistemas de recuperação de informação (SRI). Já na Ciência da Computação, a modelagem conceitual é um estágio anterior ao desenvolvimento do sistema. Nesta área, a elaboração de modelos conceituais fornece subsídios para construção de sistemas eficazes aos seus propósitos.

Considerando que a “Ciência da Informação pode ser caracterizada como uma ciência multiparadigmática” (SAYÃO, 2001, p.56) ou dito de outra forma, que “absorve ou ‘importa’ numerosos gêneros de modelos, que de ângulos distintos descrevem a sua realidade” (DODEBEI, 2002, p.24), convém destacar que esta pesquisa se apropria do

conceito de modelo conforme em Dodebei (2002, p.19) em que os modelos são “representações simplificadas e inteligíveis do mundo, permitindo vislumbrar as características essenciais de um domínio ou campo de estudo.” No caso desta investigação, o modelo tem como objetivo representar um novo instrumento de comunicação científica que visa relacionar dados e *e-prints* originados na área de Ciências Nucleares.

Essa diversidade de conceitos de modelos trabalhados no âmbito da Ciência da Informação leva à necessidade de especificar que o modelo que interessa ao escopo deste trabalho é o modelo conceitual e/ou semântico. Este modelo é construído a partir de abstrações semânticas que especificam relacionamentos entre conceitos. Apesar dos modelos não se configurarem apenas como um conjunto de relações, os princípios para sua modelagem podem ser os mesmos utilizados para o estabelecimento de relações em linguagens de representação, afinal sob outra perspectiva essas linguagens também podem ser classificadas como modelos. Burt e Kinnucan (1990) apresentam quatro tipos de abstrações como as mais usadas nos modelos semânticos, a saber: generalização, agregação, classificação e associação.

O **modelo semântico de generalização** é elaborado através do agrupamento hierárquico de entidades, no qual os objetos de mais baixo nível são ligados como subtipos dos objetos de mais alto nível. Na Ciência da Informação, esse tipo de agrupamento é especificado através das relações de gênero-espécie, que também são chamadas de lógicas ou hierárquicas. Na Ciência da Computação, esse tipo de agregação é conhecido como relação *<is-a>*, ou *<é_um>*.

O **modelo semântico de agregação** é criado através do agrupamento de partes de um objeto e revelam a composição do mesmo, assim cada entidade contribui para visualização específica do conceito maior ao qual está ligado. Na Ciência da Informação, esse tipo de agrupamento é especificado através das relações partitivas, que algumas vezes são consideradas lógicas e/ou hierárquicas, outras vezes são consideradas ônticas e não-hierárquicas. Na Ciência da Computação, esta relação é conhecida como *<part_of>* ou *<parte_de>*, mas também vem sendo chamada - por influência de outras áreas, como a Linguística, a Filosofia e a Psicologia - de relações meronímicas.

O **modelo semântico de classificação** é construído através do agrupamento de exemplos particulares de um conceito mais geral, ou em outras palavras, através da categorização das entidades. Na Ciência da Informação, este modelo é de grande importância e possui um consistente arcabouço-teórico para sua elaboração. A Teoria da

Classificação Facetada de Ranganathan (1967), por exemplo, sugere 5(cinco) facetas para categorização de assuntos conhecidas como PMEST (Personalidade, Matéria, Energia, Espaço e Tempo). Na Ciência da Computação, esse modelo começa a aparecer com mais frequência a partir do estudo das taxonomias e das ontologias. A UFO (*Unified Fundamental Ontology*) - que é uma ontologia de topo, mas também uma teoria que vem orientando aspectos de modelagem conceitual - utiliza outros tipos de categorias, como por exemplo, <Kind>, <subkind>, <role>, <phase>, etc. (GUIZZARDI, 2006)

O **modelo semântico de associação** é construído através do agrupamento de objetos segundo quaisquer outros critérios, sempre visando uma determinada facilidade. Este modelo pode ser considerado o mais complexo e mais problemático, pois existe uma infinidade de possibilidades de relacionar conceitos no mundo. Na Ciência da Informação, essas relações são chamadas de ônticas e se referem à relação do objeto com o mundo (DAHLBERG, 1978b). O princípio para estabelecimento de relações é de grande importância na elaboração de modelos e a Ciência da Informação possui uma tradição na construção de linguagens de representação, em especial, no que compreende à elaboração de conceitos e suas relações, no entanto as relações associativas ainda aparecem denominadas de outras formas como associativas ou não hierárquicas. Por outro lado, apesar da literatura da área apresentar uma variedade de possibilidades de um conceito se relacionar associativamente com outro, no âmbito da elaboração de tesauros, essas relações aparecem representadas de uma só forma, através da sigla TA (termo associado) ou TR (termo relacionado). Contudo, cabe destacar que é na Ciência da Computação que essas relações são mais requeridas, principalmente no âmbito das ontologias e da *Web* semântica, onde elas são fundamentais para permitir inferências e respostas mais precisas às questões de busca.

A presente pesquisa insere-se no escopo da modelagem semântica de associação e se valeu deste tipo de modelo para relacionar semanticamente publicação, dados de pesquisa e outros documentos, não por meio da construção de instrumento de organização do conhecimento, como são as linguagens documentárias, mas se apropriando de alguns conceitos importantes abordados nas teorias que fundamentam a construção dessas linguagens e se utilizando de representações imagéticas para sintetizar de forma clara a proposta de um novo modelo de publicação científica. Este modelo está sendo chamado nesta tese de **modelo gráfico de base conceitual** e tem por objetivo representar por meio

de imagens uma dada abstração considerando os conceitos que envolvem essa abstração e suas relações com a realidade que a envolve.

Para Dodebei (2002, p.21) “os paradigmas podem ser entendidos como supermodelos, dentro dos quais os modelos são colocados em escala reduzida”. Seguindo esse raciocínio a seção a seguir prosseguirá explicando o contexto em que o modelo proposto por esta pesquisa se desenvolve que é o da conjuntura de um novo paradigma científico chamado *eScience*.

2.2 OS PARADIGMAS CIENTÍFICOS E A eSCIENCE

No início era apenas a ciência experimental ou empírica. Esta ciência tradicional estava embasada na observação e na verificação dos fatos por meio da experiência. Segundo Rudio (2004, p.69), “na pesquisa experimental, o pesquisador manipula deliberadamente algum aspecto da realidade, dentro de condições anteriormente definidas, a fim de observar se produz certos efeitos”. Uma característica marcante da pesquisa experimental é a existência de um experimento. É a partir da execução de um experimento, que a ciência empírica estuda a “relação entre fenômenos procurando saber se um é causa do outro” (RUDIO, 2004, p.73). Em geral, de acordo com Gil (2008, p.3) as ciências empíricas “tratam de fatos e processos”, como por exemplo, a Física, a Química e a Biologia.

Na percepção dos cientistas, as teorias são mentefatos, propriedades do mapa conceitual humano e não pertencentes somente ao domínio da realidade. Na busca por novos “esclarecimentos e conhecimentos, de novos fenômenos e eventos passou-se a recorrer à reflexão e ao conhecimento acumulado, através da formulação de hipóteses e da estruturação de modelos” (ALMEIDA; TAUHATA, 1981). Sendo assim, de acordo com Sayão a abstração passou a constituir

uma ferramenta poderosa no exercício eterno de aquisição de conhecimento, uma vez que, para se compreender a imensa variedade de formas, estruturas, comportamentos e fenômenos residentes no nosso universo, é necessário selecionar aqueles de maior relevância para o problema objeto de investigação e elaborar para eles descrições adequadas (SAYÃO, 2001, p. 82),

construindo-se “esquemas abstratos da realidade, nos quais as coisas são reduzidas a seus perfis mais convenientes” os quais são chamados de modelos.

A construção de modelos para descrição e explicação de fenômenos deu origem ao segundo paradigma científico, chamado de ciência teórica ou descritiva.

Na ciência teórica executa-se pesquisa de natureza descritiva ou explanatória. Nela, “o pesquisador procura conhecer a realidade, sem nela interferir para modificá-la” (RUDIO, 2004, p. 69). O interesse do cientista está em “descobrir e observar os fenômenos, procurando descrevê-los, classificá-los e interpretá-los” (RUDIO, 2004, p.71). Para isso, formula leis, constrói modelos e teoriza a partir da descrição da natureza de um fenômeno, de sua composição e dos processos que o constituem ou que nele se realizam. A ciência teórica pode ser bem ilustrada pelas Leis de Kepler, as Leis de Newton, as equações de Maxwell, entre outras.

Esses dois primeiros paradigmas citados – experimental e teórico - são na verdade os dois motores da ciência, que impulsionam o fazer científico. Apesar das tentativas de definição os colocar em lados distintos, atualmente - principalmente devido à interdisciplinaridade que marca a construção do conhecimento científico - esses dois paradigmas são mais complementares do que opostos, sendo difícil, em muitos casos, classificar uma pesquisa unicamente como teórica ou experimental.

Com o passar do tempo, com a complexidade crescente dos modelos criados e com o desenvolvimento das tecnologias de computação, as técnicas de simulação passaram a ser a chave para a solução de problemas científicos. Gray relata que “os modelos teóricos tornaram-se muito complicados para serem resolvidos analiticamente e as pessoas começaram a simular” (apud TOLLE; TANSLEY; HEY, 2009, p. xviii). Essas simulações vêm acompanhando as pesquisas científicas desde a última metade do último milênio e têm contribuído para uma geração de dados que cresce para além de centenas de *terabytes*.

O uso de *softwares* de simulações é a característica principal do terceiro paradigma científico, que teve como desdobramento uma produção intensa de dados. Todos esses dados, capturados ou gerados foram acumulados em bancos de dados esperando para serem analisados e utilizados por programas de computador que detectassem padrões de regularidade.

Outro atributo do terceiro paradigma é o custo elevado dessas pesquisas. De acordo com Gray (apud TOLLE; TANSLEY; HEY, 2009, p.xx) “mesmo nas ciências de “poucos dados”, observam-se as pessoas coletar informação e depois gastar muito mais energia na análise da informação do que propriamente na coleta”. Neste terceiro paradigma começa a ficar clara a necessidade de se dedicar aos dados levantados e observados durante a

pesquisa, um tratamento criterioso que garanta significado à esses dados.

Para ilustrar essa dicotomia de dados brutos versus significados, Bell (2009, p.xi) relembra que Kepler (1571-1630) – assistente do astrônomo dinamarquês Tycho Brahe (1546-1601) – foi quem a partir do caderno de observações astronômicas sistemáticas de Brahe formulou as leis do movimento planetário. Este fato estabeleceu nitidamente uma divisão entre a mineração (coleta, sistematização) e a análise de dados experimentais (tratamento semântico). Por um lado, temos os dados coletados e arquivados; por outro, a criação de teorias e o reuso.

A necessidade de criação de ferramentas, bem como o estabelecimento de políticas que auxiliem os cientistas a lidar com os dados gerados, utilizando o potencial máximo desses dados e possibilitando seu reuso e compartilhamento, são fatores que levam à formulação do quarto paradigma científico.

A ideia de troca e compartilhamento de dados é bem antiga. No escopo da Ciência da Informação, pesquisadores como Otlet (1934), Bush (1945) Ted Nelson (1974), e Licklider (1965) já vislumbravam essa possibilidade. A proposta da *Web* atual como conhecemos também esteve apoiada no ideal de troca e compartilhamento de informação científica. No entanto, a *Web* tomou um rumo inesperado, tornando-se um sistema que se disseminou em grande escala, sendo usada também para outros fins. Essa disseminação ampla acarretou uma série de dificuldades, como por exemplo, a ausência de controle sobre quem publica e quem utiliza a informação nela veiculada. Essa ausência de controle, por sua vez, provocou consequências nefastas no processo de recuperação da informação, já que a publicação indiscriminada e a falta de padrões para publicar gerou imprecisões nos resultados de busca. Foi neste contexto que o criador da *Web*, Tim Berners-Lee, e outros pesquisadores, idealizaram uma nova visão para a *Web*: a *Web* semântica - cujo objetivo era possibilitar a troca de dados e informações de forma mais rápida e precisa.

No Brasil, a necessidade de se ter dados científicos tratados e disponíveis para o acesso dos pesquisadores já era sentida, em fins da década de 1980 e início de 1990. Por exemplo, no âmbito da pesquisa sobre primatas, o Museu Paraense Emílio Goeldi, desenvolveu o projeto PRIMATAM, ligado ao Núcleo de Primatologia, cujo tratamento dos dados de pesquisa resultou em um catálogo impresso, conforme informações dadas por Suely Marques-Aguiar (2011), através de contato via e-mail. Infelizmente este catálogo se perdeu ao longo do tempo.

No final da década de 90, também PEREIRA (1999), em palestra proferida no Encontro da Associação Brasileira de Editores Científicos (ABEC), já afirmava que Ciência é “um empreendimento *open-souce*”, prevendo o movimento conhecido hoje como Ciência Aberta e ainda que “ O método científico se baseia em um processo de descoberta e um processo de justificação”, preconizando assim, a necessidade de replicação e compartilhamento dos resultados científicos conforme citação abaixo:

Para os resultados científicos serem justificados, eles têm de ser replicáveis. Replicação não é possível a não ser que a Fonte seja compartilhada: a hipótese, as condições de teste, e os resultados. O processo de descoberta pode seguir muitos caminhos, e às vezes as descobertas científicas podem ocorrer em isolamento. Mas ultimamente o processo de descoberta deve ser servido pelo compartilhamento de informação: habilitando os cientistas a prosseguir a partir de onde um outro não pôde; burilando as idéias de outros e permitindo que algo novo cresça, quando de outro modo nem teria nascido. (PEREIRA, 1999, p.10)

Em contexto internacional o que se pode observar é que com o surgimento das mídias eletrônicas, a mudança nos fluxos da comunicação científica e a rápida obsolescência tecnológica de seus suportes, a necessidade de criação de métodos para tratamento e recuperação de dados científicos toma uma proporção ainda maior. Na área de genoma, por exemplo, o acesso aberto aos dados de pesquisa já é realidade há algum tempo. Desde a década de 1980. O *International Nucleotide Sequence Database Collaboration* (INSDC) mantém o *Genbank*¹⁶ – um conjunto de bases de dados que trata informações sobre sequenciamento genômico das mais diversas espécies. Atualmente, com a *Web* e todo o desenvolvimento tecnológico, esta base é alimentada pelos pesquisadores que submetem os dados e fazem "anotações" - sobre o andamento da pesquisa, sobre novas conclusões, etc. - usando como base uma ontologia chamada *Gene Ontology*. Isso revela não apenas uma preocupação em juntar os dados, mas também em ter dados tratados de forma que seu acesso e seu compartilhamento sejam feitos de forma precisa, evitando problemas na interpretação. Esse ambiente de geração, troca, compartilhamento e reuso de dados é o que se caracteriza como *eScience*.

¹⁶ <http://www.ncbi.nlm.nih.gov/sites/entrez?db=nucleotide>

Há algum tempo, no seio das comunidades científicas, existe um consenso sobre o acesso e o compartilhamento dos dados de pesquisa serem requisitos imperativos de âmbito global. (SIEBER, 1988; HEDRICK, 1988). Com o avanço da tecnologia, surge um novo problema que é a gestão de dados de pesquisa num mundo digital interligado por redes de computadores, onde há um fluxo intenso de dados sendo gerados, processados e compartilhados. A partir desse ponto, instala-se, então, um desafio importante do nosso tempo, que é ao mesmo tempo uma oportunidade significativa e essencial para se conduzir a pesquisa científica nesse século que se inicia: a *eScience*.

O conceito de *eScience*, bem como suas características, funções e desdobramentos são apresentados aqui a partir de uma revisão de literatura de autores clássicos, considerados pioneiros na introdução do tema e que estabeleceram as bases teóricas e práticas desse novo paradigma. Esses autores são: TAYLOR (2001), GRAY (2005), HEY; TREFTHEN (2003); DE ROURE (2001, 2003, 2004).

É importante mencionar que na literatura de língua inglesa o conceito de *eScience* aparece também como *e-Research*, *cyberinfrastructure*, *cyberscience*, *enhanced science*, *data-driven-science*, entre outros nomes (CESAR JUNIOR, 2011, p.7). Alguns autores ressaltam determinadas nuances entre esses conceitos, outros consideram todos como sinônimos. Para fins de levantamento, escolhemos o termo *eScience* como termo preferido nesta pesquisa, apesar de algumas vezes, o termo também aparecer com variação na grafia, sendo escrito como *e-science*. Neste trabalho adotamos *eScience*.

O termo *eScience* foi cunhado por John Taylor (2001), diretor-geral do Conselho de Pesquisa do Office of Science and Technology (OST)¹⁷ do Reino Unido, no período de 2000-2006. Taylor tinha em seu currículo experiências anteriores como chefe de laboratórios de pesquisa europeus e percebeu que muitas áreas da ciência foram se tornando cada vez mais dependentes de novas formas de colaboração multidisciplinares. Assim, esse autor definiu *eScience* como “colaboração global em áreas chave da ciência e toda próxima geração de infraestrutura que vai possibilitar esta colaboração.” Para Taylor, a *eScience* mudaria a forma de fazer ciência.

¹⁷ Depois de algumas mudanças, atualmente o OST se refere ao *Government Office for Science* do governo britânico. Disponível em: <<http://www.bis.gov.uk/go-science/>> Acesso em 19 mai. 2013.

Contudo, o grande disseminador dos ideais da *eScience* foi James Nicholas Gray¹⁸ (Jim Gray) um cientista da computação norte-americano, que recebeu o Prêmio Turing em 1998 por contribuições originais e pela liderança técnica na implementação do sistema de desenvolvimento de bancos de dados de pesquisa e processamento de transações. Gray trabalhou como pesquisador e projetista de *software* em diversas empresas, incluindo a IBM e Microsoft Research.

Ao idealizar o quarto paradigma, Jim Gray propõe uma visão da pesquisa científica para o século 21 que possibilita os pesquisadores fazerem ciência no molde dos três paradigmas anteriores, mas lidando de forma mais eficaz com a enorme quantidade de dados gerados. Assim, a chave da mudança do quarto paradigma está no uso intensivo de dados de pesquisa, bem como no seu armazenamento, processamento, análise e compartilhamento.

Para Jim Gray (2005) a nova ciência acontece quando os dados são analisados de formas novas, sendo necessário focar na exploração de dados, na análise interativa e na integração de conjuntos de dados.

Outro autor que vem contribuindo para a disseminação do conceito de *eScience* é Anthony John Grenville Hey. Tony Hey, como é conhecido, é um pesquisador e educador das áreas de Física e Ciência da Computação. Atualmente é vice-presidente de conexões da Microsoft Research, mas anteriormente exerceu o cargo de diretor do Programa de *eScience* do governo britânico. Entre as muitas contribuições de Hey para o desenvolvimento e a disseminação da *eScience* pode ser destacada particularmente a inserção do tema na comunidade de Biblioteconomia e Ciência da Informação (HEY; HEY, 2006), enfatizando sempre a importância de se ter bibliotecários e cientistas trabalhando em conjunto em prol da curadoria dos muitos dados de pesquisas gerados a partir do paradigma da *eScience*.

Para Hey e Trefethen (2005), um dos principais fatores que sustentam a *eScience* é a “iminente disponibilidade de grande quantidade de dados decorrentes das novas gerações de experimentos e pesquisas científicas”. Segundo esses autores, novos dispositivos experimentais de alto desempenho já são implantados em vários campos da ciência e isso tem levado a um verdadeiro “dilúvio” de dados científicos.

¹⁸ Gray deixou pouca literatura escrita sobre *eScience*, pois seu desaparecimento repentino em 28 de janeiro de 2007 interrompeu suas pesquisas. Muitos dos artigos seminais sobre *eScience* possuem alguma menção ou dedicatória a Jim Gray que iniciou o movimento em prol da *eScience* por meio de palestras e discussões com seus pares.

Para gerenciar essa grande quantidade de dados gerados, Hey e Hey destacam que os cientistas

vão precisar da ajuda de motores de busca especializados e poderosas ferramentas de mineração. Para criar essas ferramentas, os dados primários terão de ser anotados com metadados relevantes, registrando informações como a procedência, conteúdo e as condições que produziram os dados. (HEY; HEY, 2006, p.5)

Para Hey e Hey (2006) a criação de enormes repositórios digitais de dados científicos distribuídos levará à necessidade de serviços de gestão semelhantes às bibliotecas digitais convencionais, bem como outros serviços específicos de gerenciamento de dados. Para estes autores, a *eScience* vai colocar as bibliotecas e os repositórios como pilares centrais do desenvolvimento das próximas gerações de infraestrutura de pesquisa.

Entre os autores mais produtivos e citados, destaca-se o pesquisador David De Roure. Físico-matemático de formação, atualmente é professor de e-Research na Universidade de Oxford, diretor da E-research Centre Oxford (OeRC). De 2009 a 2013 ocupou o cargo de Diretor Estratégico Nacional para *eScience* Social. Sua contribuição está no estabelecimento da relação entre *eScience* e *Web* semântica.

Para De Roure (2003), as TICs podem apoiar e melhorar o processo científico, possibilitando que os cientistas gerem, analisem, compartilhem e discutam suas ideias e experiências de forma eficaz. A infraestrutura tecnológica a que De Roure se refere é chamada de *Grid*. A computação em grade, como também é conhecida no Brasil, é um conjunto de recursos computacionais distribuídos para atingir um objetivo comum. Apesar de geograficamente dispersa, a tecnologia de grade pode ser dedicada a uma aplicação específica ou para diversas finalidades. Tal tecnologia permite que pesquisadores espalhados por todo o mundo se unam em prol de um objetivo comum, compartilhando dados, trocando experiências e reutilizando informações. Quando se fala em compartilhamento de informações, troca de dados e comunicação entre pessoas geograficamente distantes é preciso pensar em um acordo terminológico que propicie esta troca de maneira eficaz. Foi pensando nas barreiras intercomunicacionais que De Roure e colegas (2001, 2003, 2004) propuseram uma infraestrutura chamada *Semantic Grid* ou grade semântica, conotando uma relação semelhante à que existe entre *Web* e *Web* semântica. Este grupo de pesquisadores traz então para o contexto da *eScience*, a aplicação de ferramentas da *Web* semântica – por exemplo, metadados e ontologias – melhorando a eficácia da interoperabilidade e do compartilhamento possibilitados pela tecnologia de *Grid*.

Ao utilizar o ferramental da *Web* semântica em tecnologias de *Grid* fica claro a contribuição que pode ser dada pelo Bibliotecário e o Cientista da Informação no que envolve o novo paradigma científico, pois estes profissionais têm sido pioneiros na aplicação de técnicas e ferramentas que melhoram a precisão na recuperação da informação, como por exemplo, por meio de estabelecimento de padrões de vocabulários, como são as ontologias, e outras formas de representação de conhecimento, como os metadados. Esses padrões são recursos-chave para a *eScience*.

O padrão de metadados mais difundido atualmente, isto é, o *Dublin Core Metadata*, surgiu no seio da comunidade biblioteconômica em 1995 em um *Workshop* promovido pela Online Computer Library Center - OCLC e sua aplicação vem sendo disseminada para representação de diversos tipos de objetos digitais¹⁹.

Por outro lado, as ontologias surgem na área de Inteligência Artificial, mas seu uso vem sendo potencializado a partir da utilização de bases teóricas para construção de linguagens documentárias oriundas da Ciência da Informação e Biblioteconomia, como é o caso da Teoria da Classificação Facetada (RANGANATHAN, 1967), Teoria Geral da Terminologia (WUSTER, 1981) e Teoria do Conceito (DAHLBERG, 1978b). Atualmente há interesse massivo por parte dos cientistas da informação em aperfeiçoar o desenvolvimento de ontologias consistentes, investigando aspectos relacionados às taxonomias (NOVO, 2007), à estruturação de definições (CAMPOS, 2010), bem como às relações conceituais estabelecidas entre os conceitos (SALES, 2006) que compõem as ontologias.

Não foi por acaso que Jim Gray (2005) afirmou que a “Biblioteconomia é fundamental na nossa forma de fazer ciência”; e ainda que “cada vez mais Biblioteconomia é Ciência da Computação” e vice-versa. De acordo com Soehner, Steeves e Ward (2010, p.X) “curadoria de dados, preservação, acesso e metadados são áreas da *eScience* onde as bibliotecas encontram uma afinidade natural e consequentemente dominam as respostas de pesquisa.”. Assim, percebe-se que quando se trata do papel do bibliotecário ou do profissional da informação frente à *eScience*, sua contribuição vai ainda além do desenvolvimento de ferramentas e técnicas de representação da informação, sendo também de competência desses profissionais os demais procedimentos relacionados à gestão dos dados gerados mediante este novo paradigma de geração de conhecimento científico.

.

¹⁹ Disponível em: <<http://dublincore.org/workshops/dc1/report.shtm>> Acesso em: 20 maio 2013.

Tarefas que envolvem desde a captura ou seleção dos dados de pesquisa, sua representação considerando a preservação e a recuperação desses dados, são aspectos que fazem parte do cotidiano do profissional da informação que esteja inserido neste contexto. Há um interesse implícito na proposta desta tese que é investigar novas ferramentas de comunicação científica, como é o caso das publicações. No entanto, dado esse novo paradigma científico é preciso que essa nova ferramenta acompanhe as alterações ocorridas no fazer científico e veicule de forma precisa e representativa a pesquisa preservando toda sua riqueza e complexidade. Para isso, o modelo que está sendo proposto como novo padrão de troca de informação científica considera que junto com os resultados de pesquisa sejam veiculados também os dados gerados pela pesquisa. Este modelo expressa uma agregação de objetos digitais que devem ser relacionados semanticamente, isto é, a partir de relações que tenham o seu significado explícito conceitualmente.

Assim, o capítulo que segue apresentará de forma teórica o que se entende por dados de pesquisa, bem como seus tipos, as técnicas que envolvem seu tratamento, o local onde esses dados devem ser armazenados e por fim algumas diretrizes para o estabelecimento de uma política nacional de curadoria desses dados.

3 DADOS DE PESQUISA: O QUE SÃO E QUAIS SEUS TIPOS

Cientistas de todo o mundo têm abordado a necessidade de aumentar o acesso global aos dados de pesquisa que são produzidos em quantidade cada vez maior. Isto acontece essencialmente devido à tecnologia digital que se torna cada vez mais um elemento onipresente nos processos da construção do conhecimento científico e permite também que esse conhecimento seja compartilhado e construído de forma cooperativa. Os pesquisadores, as instituições acadêmicas e as agências de fomento à pesquisa começam a entender que estes dados, se preservados e bem gerenciados, constituem uma excelente Fonte de recursos informacionais que podem ser compartilhados e reutilizados como insumo para novas pesquisas.

A Declaração de Berlim sobre o Acesso Aberto ao Conhecimento em Ciências e Humanidades, publicada em 2003, amplia o escopo do que se entende por acesso livre ao definir que as “contribuições de acesso livre incluem resultados de pesquisas científicas originais, dados não processados e metadados, Fontes originais, representações digitais de materiais pictóricos e gráficos e materiais acadêmicos multimídia” (BERLIN, 2003).

Compreendendo a importância do tema, a D-Lib Magazine²⁰ – um periódico importante no que envolve as pesquisas em bibliotecas digitais – publicou no início de 2011, um número especial sobre dados de pesquisa onde estão apresentadas questões como acesso livre, curadoria digital, aquisição e gestão, qualidade e confiabilidade e as possíveis conexões entre dados de pesquisa e as publicações acadêmicas tradicionais. Em 2014, essa mesma revista volta a publicar outro número sobre dados de pesquisa, mas dessa vez, enfatizando a criação do *Research Data Alliance*, também conhecido como RDA – uma aliança entre interessados na preservação e no tratamento de dados de pesquisa.

O National Research Council dos EUA, em seu relatório sobre direitos privados e interesse público em bases de dados técnico-científicas, define dados como “fatos, números, letras, símbolos que descrevem um objeto, uma condição, uma situação ou outro fator” (NATIONAL RESEARCH COUNCIL, 1999, p.15). Esta definição pode variar consideravelmente entre colaboradores e de acordo com a área em que são utilizados.

²⁰ Disponível em: <<http://www.dlib.org/dlib/january11/01contents.html>>. Acesso em: 20 maio 2013.

A Organização para Cooperação e Desenvolvimento Econômico (OCDE, 2007), em seu guia para acesso aos dados de pesquisas financiadas por recursos públicos, define como dados de pesquisa “registros de fatos usados como Fontes primárias na investigação científica e que geralmente são aceitos na comunidade científica como necessários para a validação dos resultados da pesquisa.”

De acordo com Borgman (2010, p.3), alguns tipos de dados têm tanto valor imediato quanto duradouro, alguns ganham valor ao longo do tempo, outros têm valor transiente, e existem outros que podem ser facilmente recriados. No levantamento realizado pode-se verificar algumas tentativas de classificação dos dados.

A National Science Foundation - NSF no seu relatório sobre Coleções de Dados Digitais de Longa Vida, publicado em 2007, classifica os dados sob vários aspectos: de acordo com sua natureza, de acordo com sua origem e de acordo com o nível de processamento ao qual foram submetidos.

Quanto à natureza, os dados de uma coleção podem ser variados, incluindo números, imagens de vídeo ou áudio, *software* e informações de versão de *software*, algoritmos, equações, animações, simulações ou modelos.

Quanto às origens, os dados podem ser observacionais, computacionais, ou experimentais. Esta distinção é fundamental para as escolhas feitas para arquivamento e preservação.

Os **dados observacionais** são aqueles gerados a partir de observações diretas, por exemplo: a temperatura do oceano em uma data específica, a atitude dos eleitores antes de uma eleição ou as fotografias de uma supernova ou ainda os dados gerados a partir de um acidente nuclear. Como são registros históricos que não podem ou não devem ser replicados, estes dados de observação são geralmente arquivados indefinidamente e precisam ser preservados com muito cuidado para que possam ser recuperados, recompostos e reusados no futuro.

Os **dados computacionais** são aqueles originados a partir da execução de um modelo de computador ou simulação. Inclui entre eles as informações sobre o modelo (incluindo uma descrição completa do *hardware*, *software*, e dados de entrada), bem sobre sua disponibilidade. De acordo com a National Science Foundation (2007), a preservação deste tipo de dado, em um repositório de longo prazo, pode não ser necessária porque os dados podem ser reproduzidos. No entanto, embora as saídas de um modelo não precisem ser preservadas, o arquivamento do próprio modelo e de um conjunto robusto de metadados - que inclua informações técnicas sobre *hardware* e *software* - pode ser essencial para que esses dados sejam reusados em novas pesquisas científicas

Os **dados experimentais** são aqueles originados a partir de experiências, tais como medições de padrões de expressão genética, taxas de reação química, ou desempenho de um motor. Para a National Science Foundation, esses dados nem sempre precisam ser armazenados indefinidamente. Os dados de experiências que não são possíveis de ser reproduzidos com as mesmas condições experimentais, ou onde algumas condições e variáveis experimentais podem não ser conhecidas, ou ainda quando os custos de reprodução do experimento forem elevados, devem ter a preservação de longo prazo garantida. Assim, considerações de custo e de reprodutibilidade são pontos chave no desenvolvimento de políticas de preservação de dados experimentais.

Borgman (2010) inclui nesta categoria de dados “quanto à sua origem”, os **dados de registros**, referindo-se àqueles dados gerados por empresas públicas ou privadas e que também são úteis para a pesquisa científica, social e humanística.

O processo experimental é outra característica de divisão dos dados. Neste caso, os dados são classificados em: intermediários e finais.

Os **dados intermediários** são aqueles recolhidos durante investigações preliminares, já os **dados finais** são aqueles em que os pesquisadores conduzem o experimento ou recolhem dados sob uma variedade de circunstâncias, no entanto, relatam apenas os resultados que consideram mais interessantes, ou seja, são dados selecionados sob algum critério.

Os dados também podem ser classificados de acordo com seus estágios de geração em: dados brutos e dados derivados

Os **dados brutos** ou **crus**, também chamados de **dados não-processados**, são aqueles recolhidos inicialmente na sua forma bruta e sujeitos a estágios posteriores de refinamento e análise, dependendo dos objetivos da pesquisa.

Já os **dados derivados** são aqueles gerados a partir de atividades de processamento ou curadoria - podendo ser os dados crus em um segundo estágio - ou seja, depois do refinamento e análise. Embora os dados não processados possam ser a forma completa, os dados derivados podem ser mais facilmente reutilizados por outros.

Harvey (2010) apresenta outra classificação para os dados. O autor os divide em coleções chamando-os de: **coleção de pesquisa** - formada por dados locais gerados em um laboratório ou projeto; **coleção de comunidade** - formada por bases de dados internacionais geradas a partir de pesquisas integradas, como por exemplo, a MGI – Mouse Genome Informatics; e **coleção de referência** que reúne dados referenciais sobre determinados experimentos. Ex: Protein Data Bank.

Lyon (2007, p.15) cita a classificação do NERC - Natural Environment Research Council, um conselho de pesquisa britânico, que apoia as atividades de pesquisa, formação e transferência de conhecimento nas ciências ambientais, e divide os dados: em **dados canônicos** - aqueles que não mudam e **dados episódicos** – aqueles que podem mudar; e também a classificação do *International Union of Crystallography*²¹ - IUCR: **dados crus** (imagem de arquivos), **dados primários** (fatores de estrutura) e **dados derivados** (seis dimensões do modelo estrutural) refletindo o domínio da cristalografia. Harvey (2010, p. 49) inclui ainda metadados como outra categoria de tipo de dados.

De acordo com duas premissas fundamentais apresentadas por Souza (2012, p.4) no que diz respeito às classificações “os mesmos objetos e ideias podem ser organizados e representados de formas diferentes e toda classificação está relacionada a um propósito definido de construção e uso de informação”. Sendo assim, todas essas classificações refletem o propósito para o qual foram construídas e um contexto específico.

Confirmando ainda essa afirmação, Borgman (2010) ressalta que pesquisadores coletam dados para diversos fins, usando vários métodos, podendo tanto a finalidade quanto os métodos influenciar no que consideram como "dados", e em que condições estes pesquisadores estão dispostos a compartilhar seus dados com os pares.

No quadro a seguir é possível verificar uma síntese da tipologia de dados de pesquisa apresentada na literatura.

²¹ Disponível em: <<http://www.iucr.org/>>. Acesso em: 1 jun. 2014.

Quadro

1: Síntese dos tipos de dados de pesquisa

CARACTERÍSTICA DE DIVISÃO	NSF	BORGMAN	HARVEY	LYON
Quanto à natureza	Número Imagem Software			
Quanto à origem	Observacionais Computacionais Experimentais	Registro		
Quanto ao nível de processamento	Intermediário Finais			
Quanto ao estágio de geração		Brutos Derivados		Crus Primários Derivados
Quanto à formação de coleções			De pesquisa De comunidade De referência	
Quanto à mutabilidade				Canônicos Episódicos

Fonte: A autora

Compreendido o que são os dados de pesquisa e como são classificados, faz-se necessário compreender também as técnicas que envolvem seu gerenciamento e preservação. Esse conjunto de técnicas é chamado de **curadoria digital de dados de pesquisa** e serão apresentadas com mais detalhes na seção que segue.

3.1 CURADORIA DIGITAL DE DADOS DE PESQUISA

O grande volume de dados digitais científicos produzidos atualmente, bem como o anseio pela preservação, compartilhamento e reuso dos mesmos conduz à necessidade de uma gestão de informação que propicie que esses fatores sejam contemplados. A fragilidade inerente à mídia na qual são registrados - a mídia digital - e a rápida obsolescência tecnológica criam a necessidade de estabelecimentos de políticas que assegurem a salvaguarda dos dados de modo que seu reuso possa ser garantido no futuro.

Tanto a gestão quanto o estabelecimento de políticas de preservação fazem parte de um leque de atividades que são englobadas pelo conceito de curadoria digital. Por ser uma área emergente, ainda há algumas divergências sobre o seu conceito, sendo muitas vezes confundido com preservação digital, arquivo digital, entre outros. Beagrie (2008, p.4) destaca a importância de se reconhecer que estes termos ainda podem ser percebidos de forma diferente por indivíduos e disciplinas distintos, já que são estudados de maneira interdisciplinar. A seguir serão apresentadas algumas das definições encontradas na literatura e um pequeno histórico do desenvolvimento do conceito de curadoria digital.

De acordo com Beagrie (2008, p.4) o termo "curadoria digital" foi usado pela primeira vez no "Digital Curation: digital archives, libraries and eScience seminar" patrocinado pela Coligação de Preservação Digital e pelo Centro Espacial Britânico, em Londres, no dia 19 de outubro de 2001. Este seminário reuniu especialistas internacionais de diversos setores (arquivistas, bibliotecários, especialistas em informação, gerentes e gestores de dados de *eScience*) para discutir sobre a evolução da curadoria de dados e da preservação digital. (BEAGRIE; PHOTEN, 2001)

A contribuição para este diálogo bem sucedido foi a seleção cuidadosa do termo "curadoria digital" utilizado para o seminário. O novo termo foi beneficiado pelo uso já consagrado da palavra "curadoria" pelos setores de biblioteca, museu, e ciências biológicas. Em todos os três setores, o termo implica não só a preservação, mas também a manutenção de uma coleção ou banco de dados, adicionado de algum valor de conhecimento. (BEAGRIE, 2008, p.4)

Assim como nas Bibliotecas e nos Museus a coleção de objetos é acompanhada por descrição do contexto e do histórico da pesquisa realizada a partir da *expertise* da equipe, o uso do termo "curadoria digital", por analogia, foi estendido também aos objetos digitais, transferindo essas abordagens para as coleções digitais.

Dessa forma, pode-se compreender que o termo curadoria digital envolve o registro do contexto e do histórico dos objetos digitais, tendo em vista a natureza mutável desses objetos em oposição aos analógicos. Um documento digital pode ter várias versões, pode estar armazenado em lugares diferentes, pode precisar de *softwares* distintos para ser compreendido e pode ter o seu conteúdo ampliado a partir da ligação com outros documentos que embasem ou atestem a veracidade de seu conteúdo.

Segundo Beagrie (2008, p.5) o conceito de criação de coleções como parte da curadoria - envolvendo a seleção e a manutenção de um corpo de conhecimento - pode ser encontrado em muitas outras disciplinas e setores, por exemplo, em centros de dados para as ciências sociais, oceanográficas e outras ciências ambientais.

Outra área que também já utiliza o termo “curadoria” há algum tempo são as ciências biológicas, sendo que nessa área o termo é ligado à manutenção e publicação de bases de dados, tais como o *GenBank*: banco de dados digitais genômicos, de iniciativa internacional, em que o valor adicionado é derivado de anotação, ligação, gestão, validação e editoração por especialistas.

Ainda conforme Beagrie (2008), a palavra "curadoria" já havia sido adotada por John Taylor em 2001 que cunhou o termo *eScience*, ao se referir à aquisição e curadoria de grandes e valiosas coleções de dados primários, o que proporcionou um contexto político valioso para a adoção do termo.

Em 2003, o Joint Information Systems Committee - JISC incorporou o termo curadoria digital e juntamente com o *eScience core programme* criou o Digital Curation Centre - DCC²². Ainda em 2003 foi realizada a 1ª Conferência Internacional de Curadoria Digital, onde em uma sessão sobre "O que é Curadoria Digital" foi debatida a definição de curadoria de dados em termos das atribuições do DCC (KERR; REDDINGTON; WILKINSON; 2005). Assim, a definição atualmente utilizada pelo DCC, publicada em sua página *Web* informa que a curadoria digital “envolve a manutenção, a preservação e a agregação de valor a dados de pesquisa durante o seu ciclo de vida”; e que a gestão ativa sobre esses dados diminui as possibilidades de perda de valor a longo prazo, reduzindo os riscos de perda devido à obsolescência digital. Além disso, minimiza a duplicação de esforços na criação de dados de pesquisa, reforçando o valor dos dados existentes à medida que eles são disponibilizados para reuso em novas pesquisas de qualidade.

Daisy Abbott expande a ideia de curadoria digital definindo-a como

²² Disponível em: <www.jisc.ac.uk/>. Acesso em 20 maio 2013

todas as atividades envolvidas na gestão de dados, desde o planejamento da sua criação – quando os sistemas são projetados -, passando pelas boas práticas na digitalização, na seleção dos formatos e na documentação, e na garantia de estarem sempre disponíveis e adequados para serem descobertos e reusados agora e no futuro. (ABBOTT, 2008, p.1)

Assim, a curadoria digital inclui ainda a gestão de grandes conjuntos de dados para uso diário, assegurando, por exemplo, que eles possam ser pesquisados, recuperados, mantendo a capacidade de serem lidos e interpretados.

A curadoria digital, em síntese, garante a sustentabilidade dos dados para o futuro, não deixando, entretanto, de conferir valor imediato a eles. Os recursos estratégicos, metodológicos e as tecnologias envolvidas nas práticas da curadoria digital facilitam o acesso persistente a dados digitais confiáveis por meio da melhoria da qualidade desses dados, do seu contexto de pesquisa e da checagem de autenticidade. Dessa forma, a curadoria contribui para assegurar a esses dados validade como registro arquivístico, garantindo que eles possam ser usados no futuro como evidência legal. O uso de padrões comuns para tratamento de diferentes conjuntos de dados, como metadados e ontologias, no contexto da curadoria digital, cria mais oportunidades de buscas transversais e de colaboração.

Para compreender as atividades que giram em torno da curadoria digital de dados de pesquisa, o Digital Curation Center (DCC) propõe um modelo que representa as etapas do ciclo de vida dos dados de pesquisa. Modelos como este proposto pelo DCC são importantes segundo Harvey (2010, p. 37), pois “ao mapear os passos e ações necessárias para cada contexto específico que eles descrevem, permitem estratégias e ações abrangentes para que uma curadoria digital eficaz seja planejada e desenvolvida”. Para a informação digital especificamente, eles moldam a forma de estudar os processos que a envolvem. De acordo com Humphrey (2006, p.1) “esses modelos representam o curso da vida de um sistema mais amplo através de uma série de fases ou estágios relacionados sequencialmente em que a informação é produzida ou manipulada”. Neste contexto, existem diversos outros modelos que podem ser encontrados na literatura sobre informação digital, no entanto o que interessa para o escopo deste trabalho é o ciclo de vida da curadoria do Data Curation Center - DCC.

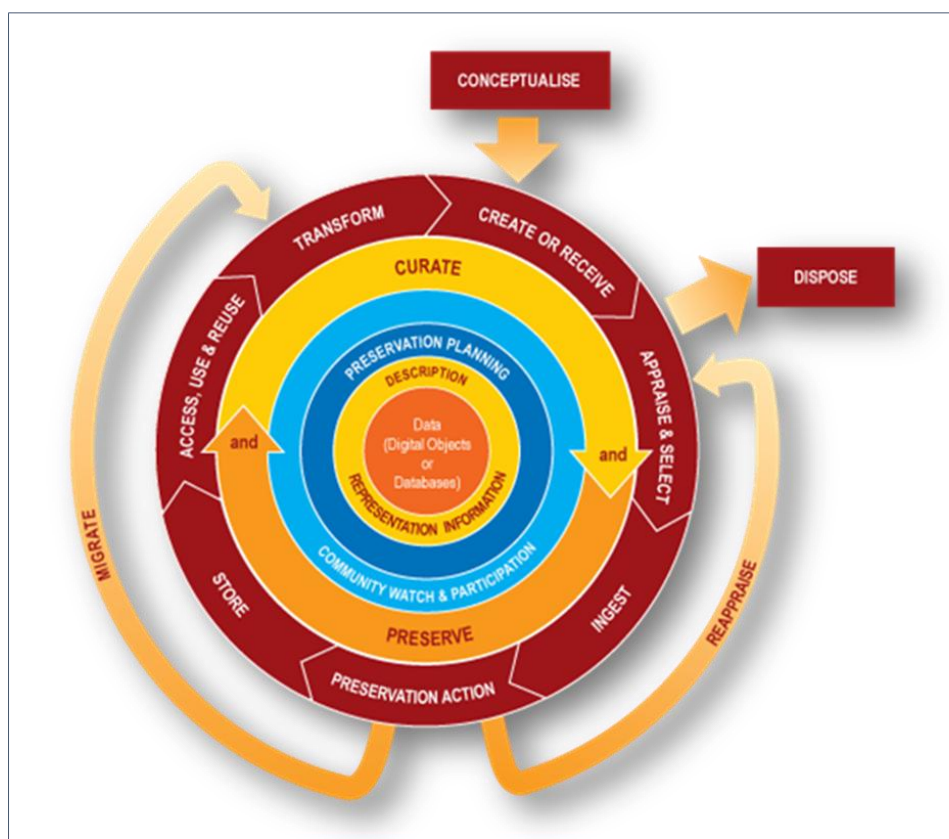
O modelo para o ciclo de vida da curadoria do DCC reflete uma visão de alto nível dos estágios necessários para o sucesso do processo de curadoria e de preservação de dados de pesquisa. Este modelo parte da concepção ou do recebimento dos dados, estando orientado para o planejamento das atividades de curadoria nas organizações ou consórcios a fim de garantir que todos os passos do ciclo sejam cumpridos. Entretanto, isto não implica que todas as organizações devam entrar no ciclo no primeiro estágio; na realidade, a operacionalização dos estágios dependerá das áreas de necessidade de cada organização.

Os elementos chave deste modelo são: dados, objetos digitais e bases de dados. No centro do ciclo de vida da curadoria está o dado digital, que é qualquer informação codificada em formato binário. A ideia de dado inclui: os objetos digitais simples e os objetos digitais complexos.

As ações que completam a totalidade do ciclo de vida incluem: a descrição e a representação da informação efetivada pela atribuição de metadados – administrativos, técnicos, estruturais e de representação – de acordo com os padrões apropriados; a definição de um plano de preservação que englobe todo o ciclo de vida da curadoria digital; a manutenção do monitoramento sobre as atividades das comunidades envolvidas, bem como participação da comunidade no desenvolvimento de padrões, de ferramentas e de *software* adequados ao problema, que possam ser compartilhados; estar continuamente alerta e promover ações administrativas e gerenciais planejadas para a curadoria e preservação por todo o ciclo de vida da curadoria.

A sequência de ações do modelo de ciclo de vida da curadoria digital proposto pelo DCC pode ser observada na Figura 1 a seguir:

Figura 1: Modelo do ciclo de vida da curadoria digital



Fonte: DCC(2008)

Dos 11 estágios deste ciclo (conceituação, criação e recebimento, avaliação e seleção, incorporação, ação de preservação, armazenamento, acesso uso e reuso, transformação, eliminação e reavaliação), os três últimos (eliminação, reavaliação e migração) podem ser aplicados ocasionalmente se necessário. Esses estágios podem ser explicados da seguinte forma:

1. Conceituação – refere-se à fase de concepção e planejamento de como o dado será criado, incluindo os métodos de captura e as opções de armazenamento.
2. Criação e recebimento – referem-se à criação do dado, incluindo o elenco de metadados necessários à sua gestão e compreensão, ou seja, metadados administrativos, descritivos, estruturais e técnicos; os metadados de preservação podem ser também incluídos neste momento.

3. Avaliação e seleção – fase em que é feita a avaliação do dado e a seleção de quais serão os objetos dos processos de curadoria e de preservação por longo prazo; manter-se aderente às práticas, às políticas pertinentes e exigências legais.
4. Captura – fase de transferência do dado para um arquivo, repositório, centro de dados ou outro custodiante apropriado.
5. Ação de preservação – fase em que se promovem ações para assegurar a preservação de longo prazo e a retenção do dado de natureza oficial; essas ações devem assegurar que o dado permaneça autêntico, confiável e capaz de ser usado enquanto mantém sua integridade; essas ações de preservação incluem: a limpeza do dado e a sua validação, a adição de metadados de preservação; adição de informação de representação e a garantia de estruturas de dados ou formatos de arquivos aceitáveis.
6. Armazenamento – fase em que o dado é armazenado de forma segura mantendo a aderência aos padrões relevantes.
7. Acesso, uso e reuso – fase em que se assegura que o dado possa ser cotidianamente acessado tanto pela sua comunidade alvo, quanto pelos demais usuários interessados no seu reuso; isto pode ser realizado na forma de informação disponível publicamente; controle de acesso robusto e procedimentos de autenticação podem ser aplicados.
8. Transformação – refere-se à fase de criação de novos dados a partir do original, por exemplo, pelo processo de migração para diferentes formatos ou pela criação de subconjuntos - realizada por meio de seleção ou formulação de consultas – derivando novos resultados que podem ser publicados.
9. Eliminação – fase que pode ocorrer ou não. Nela pode-se eliminar o dado que não foi selecionado para curadoria e preservação de longo prazo de acordo com políticas documentadas, diretrizes e exigências legais.
10. Reavaliação – fase que também pode ocorrer ou não. Nesta fase, os dados são reavaliados, podendo haver um retorno ao dado cujos procedimentos de avaliação foram falhos para nova avaliação e possível seleção.

11. Migração – fase não obrigatória em que pode ser feita a conversão dos dados para um formato diferente. Ocorre geralmente quando há necessidade de compatibilizar os dados com o ambiente de armazenamento ou para assegurar a imunidade do dado em relação à obsolescência de *hardware* e de *software*.

Um levantamento no Google utilizando o termo “*data curation*” revela que já existem diversas iniciativas relativas à curadoria de dados de pesquisa no mundo. Essas iniciativas partem de bibliotecas, de agências de fomento, das instituições de pesquisa, mas principalmente do governo nacional de cada país. Desta forma, é promissor observar que cada vez mais, dados de pesquisa estão sendo armazenados em repositórios confiáveis, onde, gerenciados sob os princípios da curadoria digital, são preservados e mantêm a sua capacidade de reuso.

Definido o que é curadoria digital de dados de pesquisa, é importante compreender que esses dados, na maioria das vezes, são armazenados em locais específicos denominados repositórios de dados. O capítulo a seguir discutirá este conceito.

3.2 REPOSITÓRIOS DIGITAIS DE DADOS DE PESQUISA

O modelo de publicação proposto por esta tese é viabilizado tecnicamente e gerencialmente pela existência de ambientes como os repositórios digitais que “reúnem de maneira organizada a produção científica de uma instituição ou área temática” (IBICT, 2012) e que também podem armazenar os dados curados gerados pela pesquisa publicada.

De acordo com o Instituto Brasileiro de Informação em Ciência e Tecnologia - IBICT (2012) os repositórios digitais (RDs) são “bases de dados *online* que [...] armazenam arquivos de diversos formatos”, resultando em uma série de benefícios tanto para os pesquisadores quanto para as instituições ou sociedades científicas, proporcionando maior visibilidade aos resultados de pesquisas e possibilitando a preservação da memória científica de sua instituição.

A literatura apresenta uma diversidade de tipos de repositórios digitais. A classificação mais comum considera o objetivo para o qual o repositório é construído, dividindo-os em **Repositórios Institucionais** – aqueles voltados para a o armazenamento da produção científica de uma determinada instituição, como por exemplo, a Plataforma

Carpe diEN²³ repositório do Instituto de Engenharia Nuclear - e **Repositórios Temáticos**, voltados para a reunião da produção técnico-científica sobre uma determinada temática extrapolando os limites institucionais, como é o caso do E-Lis²⁴ – repositório que armazena a produção científica da área de Ciência da Informação.

Apesar do termo repositório não ser novo na área tecnológica, o conceito no âmbito da comunicação científica envolve características que devem ser destacadas para diferenciá-lo de uma base de dados comum, a saber: 1) contêm documentos digitais e não apenas referências, 2) armazenam documentos digitais referentes à produção técnico-científica, 3) são planejados para que a informação ali depositada seja preservada ao longo do tempo, 4) utilizam tecnologia e padrões abertos que permitam que o sistema seja interoperável com outros repositórios 5) permitem autosubmissão, 6) estão disponíveis na *Web*. Essas características são fundamentais e permitem que o repositório cumpra a sua função de preservar e disseminar a informação técnico-científica seja de uma instituição ou de uma dada área temática.

Outras características adicionadas podem ainda transformar um repositório digital em uma biblioteca digital, por exemplo, a criação de serviços de informação com base nas informações depositadas nos repositórios.

Para esta pesquisa é especialmente interessante trabalhar com outra classificação para repositórios, isto é, a classificação que considera o tipo de material depositado. Sendo assim, o texto segue tratando os conceitos de repositórios de *e-prints* e repositórios de dados.

Os repositórios de *e-prints*, de uma forma simples, podem ser definidos como uma espécie de repositório digital que armazena *e-prints*, cujo objetivo é “melhorar o processo de comunicação científica, facilitando o acesso à literatura” (PINFIELD; JAMES, 2003, p.6). Neste caso, *e-prints* estão sendo considerados como toda versão eletrônica de publicação científica (teses, artigos, relatórios, etc) (PINFIELD; JAMES, 2003, p.1), sem fazer distinção entre *pré-print* – publicações que ainda não passaram por *peer-review* e *pos-print* – documentos já avaliados e publicados. Este tipo de repositório se enquadra tanto na categoria de repositório institucional quanto na categoria de repositório temático, pois ambos armazenam versões eletrônicas de documentos científicos. Em ambos os casos há uma vasta literatura em língua portuguesa que trata sobre estes tipos de repositórios.

²³ Disponível em: <<http://carpedien.ien.gov.br>>. Acesso em: 14 maio 2014.

²⁴ Disponível em: <<http://eprints.rclis.org/>>. Acesso em: 14 maio 2014.

No entanto, os repositórios que mais interessam ao contexto dessa pesquisa são os repositórios de dados científicos. Este tipo de repositório é mais raro no contexto brasileiro. As iniciativas existentes de repositórios desse tipo no país surgem no âmbito de consórcios internacionais como é o caso do Projeto Genoma Humano²⁵ ou como inovação proposta pelo usuário, isto é, os próprios pesquisadores, que em alguns casos isolados, sentem a necessidade de organizar e compartilhar seus dados com pares e alunos e começam a organizar banco de dados para tratar seus dados, como por exemplo, na área ambiental e de geociências.²⁶

Os repositórios de dados, em grande parte dos casos, se distinguem dos repositórios digitais em geral, pois o seu conteúdo, isto é, os dados, possuem características próprias necessitando de um tratamento apropriado. (RODRIGUES, 2010). As anotações sobre a análise de uma determinada substância química, por exemplo, vão requerer metadados de tratamento específicos, enquanto outro *dataset* pode necessitar de outro conjunto de metadados para ser recuperados.

O Relatório D-24 do Projeto Repositório Científico de Acesso Aberto de Portugal – RCAAP, de 2010, relata que as soluções tecnológicas adotadas para repositórios de dados algumas vezes são as mesmas adotadas para outros tipos de repositórios, mas que também existem soluções desenvolvidas para um caso específico podendo-se identificar dois tipos de práticas. A primeira existe em domínios que já têm uma prática estabelecida de registro e compartilhamento de dados, como é o caso do domínio genômico e da astronomia em que o paradigma vigente é o do uso de base de dados e não repositórios. A segunda surge mais recentemente impulsionada pelo movimento de acesso livre em que as agências de fomento à pesquisa tomam a iniciativa de recolherem os dados, neste caso,

Como estas acções são freqüentemente entregues aos mesmos serviços que gerem os repositórios institucionais, e os próprios conjuntos de dados estão muitas vezes ligados às publicações que neles se baseiam, o paradigma aqui é o de uso das plataformas de repositórios²⁷ (RODRIGUES, 2010, p.24)

Observa-se que os repositórios de dados se caracterizam, em termos tecnológicos, por se estruturarem, na maioria dos casos, na forma de bancos de dados factuais. Entretanto, devido às condicionantes de cada disciplina e à variedade e complexidade

²⁵ Disponível em: <<http://www.genome.gov/10001504>> Acesso em: 20 maio 2011.

²⁶ Informações obtidas com pesquisadores por meio de Fontes informais de comunicação.

²⁷ Um exemplo deste tipo de plataforma é o dSpace: sistema mais amplamente utilizado no Brasil para implementação de repositórios.

típica dos resultados de pesquisa, o panorama geral dos repositórios de dados é extremamente heterogêneo em termos de conteúdo, concepção, tratamento dos dados e de gestão. Sendo assim, na prática, a escolha da tecnologia a ser aplicada no desenvolvimento do repositório de dados digitais está relacionada à forma como se configurou o processo de curadoria dos dados, à que tipo de dado será armazenado, ao processo de pesquisa (*workflow*) que resultará nos dados e às possibilidades de reuso.

É importante destacar que apesar da diversidade dos modelos em que os repositórios de dados são baseados e se apresentam, estes se tornam uma peça importante para enfrentar o desafio do acesso permanente e da preservação de longo prazo aos dados de pesquisa pelas comunidades científicas interessadas. Esse desafio contemporâneo se sobrepõe aos ciclos de geração de conhecimento científico e abre grandes oportunidades para a pesquisa científica.

Segundo a página do SURF,²⁸ repositórios de dados de pesquisa são bancos de dados digitais que garantem o acesso a resultados de pesquisa agora e no futuro, tendo como perspectiva primordial o acesso aberto ao que é considerado também um patrimônio digital da humanidade. Tornar a informação disponível nesses repositórios assegura que o conhecimento não ficará selado em bases de dados comerciais, enfatiza a página. Dessa forma, as redes de repositórios de coleções de dados contribuem de forma decisiva para o fortalecimento e a expansão dos princípios do livre acesso.

São muitos os benefícios para as comunidades científicas da implantação de repositórios de dados padronizados e interoperáveis. Vejamos alguns desses benefícios que são mais diretamente perceptíveis:

- Amplia a visibilidade dos resultados de pesquisa, posto que, via de regra, só é formalmente disseminada a fração que está registrada nas publicações acadêmicas.
- Proporciona mecanismos de preservação de longo prazo, em termos de preservação digital, de arquivamento seguro e de curadoria digital.
- Permite que o material depositado esteja disponível *on-line* continuamente para ser consultado e citado mais frequentemente.
- É um instrumento chave para os processos de reformatação e recriação de dados proporcionados pela curadoria digital.

²⁸ Disponível em: <http://www.surf.nl/en>. Acesso em: 8 jun. 2014.

- Abre a possibilidade de criação de novos serviços de informação para pesquisadores e gestores a partir da análise dos dados arquivados e a integração de dados e publicações acadêmicas.
- Permite a criação de redes de repositórios interoperáveis
- Aumenta o grau de reuso dos dados minimizando a duplicação de esforços e otimizando os investimentos na geração de dados.

A relevância dos repositórios e bancos de dados de pesquisa se reflete no número e importância das organizações internacionais envolvidas na padronização, organização e disseminação desses recursos informacionais e, sobretudo, na inserção deles na “ciberinfraestrutura” mundial de informação para a pesquisa. Segue abaixo a relação de algumas organizações que têm um papel chave nos processos de alavancagem dos repositórios de dados e de suas coleções:

- DataCite²⁹ – organização sem fins lucrativos formada nos fins do ano de 2009 em Londres com a responsabilidade de: facilitar o acesso aos dados de pesquisa disponíveis na Internet; aumentar a aceitação de dados de pesquisa como uma contribuição legítima e citável para um registro acadêmico; e dar apoio ao arquivamento de dados de forma a permitir a verificação dos resultados de pesquisa e o reuso dos dados para futuros estudos.
- REGISTRY OF RESEARCH DATA REPOSITORIES (re3data.org)³⁰ – fundada pelo GermanResearch Foundation entre 2012 e 2014, tem como objetivo criar um registro global dos repositórios de dados de pesquisa que cubram diferentes disciplinas acadêmicas, tendo como perspectiva promover a cultura de compartilhamento, aumento do acesso e melhor visibilidade dos dados de pesquisa.
- RESEARCH DATA REPOSITORIES - Databib³¹ – define-se como uma ferramenta para apoiar as pessoas a identificarem e a localizarem repositórios *on-line* de dados de pesquisa. O Databib tem como objetivo responder às seguintes

²⁹Disponível em: <www.datacite.org>. Acesso em: 5 jun. 2014.

³⁰Disponível em: <www.re3data.org>. Acesso em: 1 jun. 2014.

³¹ Disponível em: <www.databib.org>. Acesso em: 5 jun. 2014.

indagações de pesquisadores: Quais são repositórios apropriados para um pesquisador submeter seus dados? Como o usuário acha repositórios apropriados e descobre conjunto de dados que se enquadre nas suas necessidades? Como as bibliotecas podem ajudar os usuários a localizar e integrar dados na sua pesquisa ou atividade de ensino?

- RESEARCH DATA ALLIANCE - RDA³² – tem como objetivo construir pontes técnicas e sociais que permitam o compartilhamento aberto de dados de pesquisa, baseando-se na visão de que pesquisadores e inovadores compartilham livremente dados de forma transversal sobrepondo-se a tecnologias, disciplinas e países no sentido de equacionar os grandes desafios da sociedade.

O re3data.org e o Databib pretendem até o ano de 2015 reunir seus projetos em um único serviço que será gerenciado pelo DataCite. O objetivo dessa fusão é reduzir a duplicação de esforços e melhorar os serviços oferecidos à comunidade de pesquisadores através de um único registro de repositórios de dados que incorpore o melhor dos dois projetos. Os princípios que regem essa fusão são os seguintes:

- Transparência – os metadados e as interfaces dos registros estarão acessíveis de forma aberta; os metadados estarão acessíveis sob os termos da Creative Commons³³.
- Garantia de qualidade – um *workflow* de dois estágios, que irá garantir a qualidade e a atualização dos registros, consistindo de uma revisão por um conselho editorial internacional somada a uma revisão voltada para a consistência do registro.
- Desenvolvimento de funcionalidades inovadoras – desenvolvimento cooperativo de novas funcionalidades para o registro e a integração com um “ecossistema” global de infraestruturas voltadas para as necessidades de uma ciência aberta e voltada para dados.

³² Disponível em: <<https://rd-alliance.org>>. Acesso em: 5 jun. 2014.

³³ Creative Commons é um conjunto de licenças que permite que detentores de copyright (isto é, autores de conteúdos ou detentores de direitos sobre estes) possam abdicar em favor do público de alguns dos seus direitos inerentes às suas criações, ainda que retenham outros desses direitos.

- Gestão compartilhada – o registro será gerenciado igualmente por representantes de cada um dos projetos.
- Sustentabilidade – ambos os projetos trabalharão em conjunto em torno de uma estrutura sustentável de governança e de uma infraestrutura permanente para o registro.

Os repositórios de dados de pesquisa cumprem assim a difícil missão de construir “pontes de dados” entre disciplinas e domínios altamente fragmentados que caracterizam a pesquisa global atual, criando uma metáfora de blocos de construção que podem redesenhar novas descobertas através de análises transversais dos dados produzidos por diferentes disciplinas. O compartilhamento e o trabalho cooperativo são chaves para isso, porém eles só podem se tornar possíveis tendo como base um conjunto de tecnologias e padrões que se consolidem em torno desse conceito imprescindível – tão antigo e tão inovador - de banco de dados de pesquisa.

A intensa geração de dados que caracteriza a pesquisa contemporânea impõe ao mundo das ciências, seja de que disciplina for, a necessidade imprescindível do desenvolvimento de infraestruturas que assegurem a máxima acessibilidade, estabilidade e confiabilidade dos dados cujo valor de reuso esteja patente; além do mais, é necessário que essas infraestruturas garantam não apenas que os dados possam ser intercambiados, mas também que outras informações importantes para o estabelecimento de parcerias, para a gestão das atividades de pesquisa ou para a maximização do potencial de inovação de uma pesquisa possam ser recuperadas. Nessa direção, o que se nota é que mais e mais universidades, centros de pesquisa e mesmo órgãos governamentais que coletam dados e indicadores, estão trabalhando na construção e aperfeiçoamento de infraestruturas que se configurem em ambientes tecnológicos gerenciais e confiáveis, como será descrito na subseção a seguir:

3.3 OUTRAS INFRAESTRUTURAS DE TRATAMENTO DE DADOS DE PESQUISA

A crescente complexidade das atividades de pesquisa, a imensa geração de dados e informações e a necessidade de gerenciar processos propiciaram o surgimento de

infraestruturas tecnológicas com vistas ao tratamento e à recuperação dessas informações. Essas infraestruturas vêm sendo criadas não apenas para o armazenamento de dados, mas principalmente para gerenciar os processos e as etapas das atividades de pesquisa. Os benefícios são vistos não apenas pelos pesquisadores, mas pelos gestores, pelas agências de fomento, pelas empresas, bem como pelo público em geral. Essas infraestruturas permitem a contextualização das atividades científicas, otimizam os fluxos de trabalho, tornando a produção mais transparente, além de padronizá-las e permitir sua avaliação e reavaliação para o bom andamento das pesquisas, bem como para o reuso de dados e para a viabilização de novas descobertas.

Um exemplo de infraestrutura nesses moldes é o Current Research Information System – CRIS, que consiste em um modelo de dados que descreve um conjunto de objetos de interesse para as atividades de pesquisa e uma série de ferramentas que possibilitam ao usuário (pesquisador, gestor, etc) a gestão de seus dados de pesquisa em todos os processos, incluindo alocação de recursos, avaliação de projetos, identificação de novos mercados para produtos de pesquisa, análise de tendências, entre outros serviços.

Em geral, o CRIS é construído para uma dada comunidade, como por exemplo, o United States Data Agriculture - USDACRIS³⁴, que fornece documentação e relatórios para as atividades agrícolas, ciência dos alimentos, nutrição humana e silvicultura.

No entanto, a ideia do CRIS não é nova. Há aproximadamente 40 anos diversos sistemas nos moldes do padrão CRIS vêm sendo desenvolvidos pelo mundo, muitas vezes com outros nomes, mas sempre como mecanismo de apoio à organização e à recuperação de informações relevantes para a comunidade científica.

Normalmente, o CRIS tem informações sobre os projetos, pessoas, unidades organizacionais, programas de financiamento, resultados de pesquisa (produtos, patentes e publicações), instalações e equipamentos, e eventos, ou seja, todo tipo de informação que de alguma forma pode dar apoio às atividades de Pesquisa & Desenvolvimento (P & D), seja para um financiador, para uma instituição de pesquisa, para o pesquisador, para o público ou para os meios de comunicação.

São exemplos de informações constantes nos CRIS, o currículo dos pesquisadores e suas páginas, portfólios de projetos de pesquisa, bibliografias, instituições com pesquisas correlatas, informações sobre oportunidades de inovação, informações sobre instalações e equipamentos, eventos etc.

34 Disponível em: <<http://cris.nifa.usda.gov/>>. Acesso em: 15 jan. 2014.

O sucesso dos CRIS, somado à riqueza informacional da *Web* e à proliferação de uma grande variedade de sistemas voltados para as comunidades científicas tornaram a informação para a pesquisa heterogênea e distribuída, como consequência, a busca por esse tipo de informação transformou-se numa tarefa árdua para os usuários. Dito de outra maneira, a informação agora armazenada e tratada estava distribuída em sistemas diversos fazendo com que o usuário gastasse muito tempo navegando separadamente por cada um deles.

Lopatenko (2001) mostra esse problema no seu artigo sobre recuperação da informação em CRIS. Segundo ele, normalmente pesquisadores ou gestores de informação em políticas de pesquisa não se limitam apenas à informação armazenada em um dos sistemas existentes, ao contrário, informações de pesquisa em qualquer área da ciência e tecnologia estão espalhadas por uma variedade de sistemas de informações heterogêneos e por isso há uma forte necessidade de reunir todas as informações possíveis ou, de pelo menos, o sistema apontar para onde essas informações podem ser encontradas.

Lopatenko enfatiza a importância de saber se a informação reunida na pesquisa é efetiva e completa. No entanto, segundo ele, pesquisas anteriores revelaram que a integração de dados de instituições de pesquisa não resolve o problema, especialmente se as instituições forem regidas por órgãos diferentes ou se não usufruem de benefícios diretos de participação em tais redes de informação.

Nesta direção, o autor reafirma a necessidade de encontrar uma solução para o problema de integração dos dados, que passa pela implementação de um padrão com três características essenciais: 1) fácil de implementar para qualquer participante, 2) flexível o suficiente para abraçar a diversidade, a estrutura e o significado dos dados em diferentes estados, organizações, ou áreas da ciência e 3) poderoso para fornecer serviços de recuperação de informações sofisticados. Para isso sugere o uso de ontologia e de padrões sugeridos pelo *W3C Consortium*³⁵.

Nesta direção, a Comunidade Europeia criou o European CRIS (EUROCRIS)³⁶, uma organização sem fins lucrativos voltada para o desenvolvimento de sistemas de informação de pesquisa e a interoperabilidade entre esses sistemas.

A idéia de fazer esses sistemas interoperarem é permitir que o usuário final possa acessar a informação disponibilizada em CRIS distribuídos e heterogêneos, bem como em

³⁵ Disponível em: <<http://www.w3.org/>>. Acesso em: 15 jan. 2014.

³⁶ Disponível em: <<http://www.eurocris.org/>>. Acesso em: 15 jan. 2014.

repositórios em um local único. Para isso o EUROCRIS vem adotando uma série de estratégias, como: troca de experiência entre os membros em geral; criação do DRIS (diretório de CRIS); estudo e desenvolvimento de atividades conjuntas de P & D; conferência bienal sobre CRIS; reuniões semestrais com os membros, seminário estratégico anual, *workshops*, ligações com parceiros estratégicos, desenvolvimento de estratégia e infraestrutura e o mais importante deles, o desenvolvimento do Common European Research Information Format - CERIF, um padrão recomendado aos estados-membros da comunidade europeia inicialmente com a finalidade de facilitar o intercâmbio de informações entre bases de dados de projetos de pesquisa.

Criado em 1991, o CERIF, com o passar do tempo, precisou ser revisto e assim foi também estendido a outros tipos de informações além dos projetos de pesquisa. Nessa direção, a versão CERIF2000 apresentou diretrizes para um modelo de dados CRIS mais completo e um núcleo base que permitia a troca de informações de maneira flexível possibilitando que a maioria dos CRIS existentes pudessem manter suas características próprias e ainda assim interoperar com os demais CRIS existentes na comunidade.

O CERIF2008- última versão disponível - descreve um modelo de dados formal que permite a interoperabilidade entre os sistemas de gestão da investigação, a partir de informações sobre pessoas, projetos, organizações, publicações, patentes, eventos, prêmios, equipamentos etc, um modelo de dados físico (JÖRG et al., 2009a ; JÖRG et al., 2009b) e um formato de troca de dados em XML (JÖRG et al., 2009b).

Além disso, de acordo com Ivanovic, Surla e Rackovic (2011), o modelo de dados CERIF tem uma camada semântica que permite a classificação de entidades e suas relações de acordo com algum esquema de classificação. Outras entidades do modelo de dados CERIF estão ligadas à camada semântica através da entidade <cfClass> que descreve o papel da pessoa na criação do resultado (autor da publicação, editor da publicação, presidente do conselho de eventos, gerente de projetos, etc), a classificação do resultado da pessoa (ex: monografia, revista impressa, etc), a classificação das publicação em que o resultado é publicado (ex: principal revista de importância internacional, revista nacional, etc), a classificação do evento onde o resultado é apresentado (conferência de importância internacional, conferência de importância nacional, etc) e a classificação do prêmio que é dado à pessoa (Prêmio de excelência internacional, prêmio nacional etc.).

Complementarmente, de acordo com a página mantida pelo grupo gestor, esta versão incluiu a recomendação de um tesauro multilíngue chamado Ortelius que padronizou a indexação de assunto e os códigos utilizados para as áreas de atividades econômicas e produtos e ainda uma lista controlada de valores e atributos de determinados elementos (por exemplo: o papel de uma pessoa no projeto)

Em suma, a inovação apresentada pelo CERIF está na sua estrutura de dados formais, garantindo a integridade dos dados e evitando múltiplas instâncias dos mesmos valores de atributos; no uso de relações n:n permitindo declarar o papel e a duração temporal dos projetos; na preservação das características individuais de cada sistema e em sua essência multilíngue. Interessante observar que assim como essa pesquisa, o modelo CERIF está preocupado não apenas em identificar as entidades a serem descritas, mas também as relações que elas possuem umas com as outras, o que propicia a formação de uma rede interligada de informações.

No Brasil, as iniciativas semelhantes ao CRIS são raras e o que se encontrou mais próximo foi a Plataforma Lattes³⁷, entretanto, o sistema CRIS conforme concebido na Europa considera não apenas informações sobre pessoas e instituições, como é o caso do Lattes, mas seu primeiro e principal objeto são os projetos de pesquisa e, mais recentemente, os dados não processados gerados por estes projetos, o que não se encontra em nenhuma das agências brasileiras de financiamento, que seriam as principais interessadas. O que se observa, portanto, é que no Brasil ainda não há um sistema avançado de gerenciamento, acesso e compartilhamento da produção científica nacional, como é o EUROCRIS.

Considerando como pressupostos básicos que os estoques de informações digitais são elementos fundamentais para o desenvolvimento da ciência e tecnologia, para os processos de inovação, para a educação e a cultura e para os empreendimentos governamentais e privados; fica claro que o futuro desses domínios e processos dependerá, em doses variadas, da competência das instituições responsáveis em prover acesso persistente a estes estoques, e que a capacidade de exploração, reutilização e transversalidade disciplinar desses recursos informacionais dependerá da sofisticação de tratamento e de gestão por que eles tiverem passado desde seu planejamento.

Resta então estabelecer que informação deverá ser preservada, quem é o responsável pela preservação, que infraestrutura deverá ser desenvolvida, que controles

³⁷ Disponível em: <<http://lattes.cnpq.br/>>. Acesso em: 15 jan. 2014.

sociais, éticos e legais deverão ser aplicados, e, finalmente, quem pagará por tudo isso. As decisões são urgentes, pois o acesso aos dados no futuro vai depender de como vamos equacionar todas as variáveis que se sobrepõem.

Deslocando o olhar para os dados digitais, há um consenso entre gestores de C&T, pesquisadores e profissionais das áreas de Ciência da Informação e de Tecnologia da Informação de que em coleções digitais de dados pesquisa – principalmente em vista de sua complexidade, diversidade e fragilidade intrínseca – os dados só podem ser preservados e gerenciados ao longo do tempo para acesso e reuso por meio de compromissos sustentáveis e duradouros que se entrelaçam em várias instâncias. Assim fica evidente a necessidade do estabelecimento de uma política nacional que alinhe as diversas dimensões sob a qual a curadoria digital de dados de pesquisa deve ser tratada.

É preciso enfatizar ainda que é o estabelecimento de uma política de curadoria digital de dados que abrirá a possibilidade de elaboração de um novo modelo de publicação que ligue os *e-prints* aos dados, tendo em vista que para que esse modelo se consolide de maneira consistente, os dados utilizados como elementos para compor o modelo precisam estar curados por meio de políticas, normas e padrões adotados nacionalmente.

A seção a seguir discorre sobre novos modelos de publicação com o objetivo de apresentar como a literatura vem expondo a necessidade de veículos de comunicação científica que explicitem com fidedignidade a complexidade e a riqueza das pesquisas técnico-científicas atuais.

3. 4 NOVOS MODELOS DE PUBLICAÇÃO CIENTÍFICA

Há algum tempo a comunidade científica sente a necessidade de novos modelos de publicações. Modelos que possam explicitar mais claramente a pesquisa desenvolvida, seus métodos, seu material, que revelem com nitidez a intenção do pesquisador e dos diversos projetos de pesquisa, que possam servir de base para avaliações mais confiáveis e para testes de veracidade e replicação de experimentos, ou simplesmente, estarem disponíveis para serem reutilizadas mais eficazmente para novas pesquisas. Modelos que não sejam apenas uma cópia eletrônica da página impressa, como é o documento PDF, mas que permitam ao pesquisador interagir com outros pesquisadores e também com os sistemas automatizados subjacentes às novas concepções de publicações.

No início do século XX, ainda muito aquém dos avanços tecnológicos recentes, Paul Otlet, pai da Documentação, já criticava o modelo tradicional de publicação científica e, em diversas palestras já mencionava a necessidade de um novo padrão de publicação (VAN DEN HEUVEL; RAYWARD, 2011). De acordo com estes autores, Otlet tinha plena consciência da necessidade de um novo formato para o livro que deveria ter por base uma forma colaborativa de representação do conhecimento.

Aparentemente o avanço tecnológico seria rapidamente apropriado pelas comunidades científicas possibilitando uma rápida transição para um novo modelo de publicação acadêmica, a exemplo do que vem acontecendo com os demais veículos de informação desde o surgimento da Web, entretanto, não foi isso que se observou. Desde meados da década de 90, o advento da Internet vem trazendo diversas oportunidades de melhoria dos canais de comunicação científica. As publicações *on-line* revolucionaram a pesquisa e a descoberta de informações, aumentando a amplitude e a facilidade de acesso a todo tipo de material, inclusive conjuntos de dados científicos. No entanto, quando se trata do artigo científico - enquanto principal veículo para comunicar pesquisa – as publicações apresentaram pouca melhoria, principalmente no que diz respeito à aplicação de tecnologias de *Web 2.0* e *Web* semântica, que poderiam certamente beneficiar esse veículo em termos de representação, estrutura, organização e apresentação. Essa constatação pode ser evidenciada no primeiro texto de Kircz sobre novas práticas para publicação eletrônica, onde o autor mostra que não basta juntar informações, mas é preciso reuní-las sob um contexto para que possam servir de base para a argumentação entre os pares:

um documento eletrônico não é uma versão eletrônica de um documento tradicional em papel, com enfeites, como *hiperlinks*, fotos coloridas e animações ilustrativas. Em vez disso, um documento eletrônico é um documento que compreende uma variedade de diferentes tipos de informações que são apresentadas juntas por um autor, a fim de formular um argumento científico abrangente. (KIRCZ, 2001, p.266)

Segundo Aalbersberg, Dunham e Koers (2011), nas últimas décadas, as “melhorias no artigo acadêmico têm sido principalmente em termos de entrega (eletronicamente), descoberta (pesquisa de texto completo), bem como uma série de melhorias específicas, em menor escala, como a possibilidade de fazer *upload* de dados complementares”. Entretanto, em termos de estrutura e forma, o artigo atual é, em geral, o mesmo que nas primeiras revistas científicas do século XVII. Com o intuito de estabelecer uma nova infraestrutura para disseminação e compartilhamento do conhecimento científico que novos modelos de publicação que utilizem o máximo potencial tecnológico vêm sendo propostos.

Em 2006, Hunter visualizou um “pacote de publicações científicas” que encapsula e relaciona, na forma de objetos compostos, dados brutos com os seus subprodutos, publicações e metadados contextuais de proveniência e administrativos.

A última palestra ministrada por Jim Gray em 2007 e que depois virou um dos capítulos do livro “*The fourth paradigm*” organizado por Hey, Tansley e Tolle em 2009 é dividida em dois momentos. No primeiro, Gray fala sobre a necessidade de ferramentas para ajudar os cientistas a capturar, curar, analisar e visualizar seus dados. Em um segundo momento, aborda o tema comunicação científica, apontando o problema do padrão de publicação acadêmica tradicional e destacando o desafio de se adotar um novo padrão de publicação, que faça interoperar artigos e dados brutos, além de permitir, via sistema de colaboração, anotações e comentários sobre os artigos. Este modelo foi chamado de *overlay documents* ou “documentos sobrepostos”.

De acordo com Cavalli (2009), o conceito de publicação sobreposta surge em 1996 em um artigo seminal de Paul Ginsparg – apontado como um dos fundadores do movimento *Open Access* - onde considera que as melhorias implementadas no ambiente eletrônico, como a queda das barreiras físicas e financeiras para disseminação, possibilitariam a criação de novos tipos de documentos. Para Cavalli, era possível se imaginar que os artigos fossem repartidos em pequenos fragmentos, graduados de acordo com sua importância para a pesquisa ou com algum outro critério de utilidade, e ainda poderiam conter qualquer tipo de informação sobreposta, como por exemplo, os arquivos de dados crus ou não-processados. O discurso ainda continua propondo diferentes formas de adicionar valor aos dados crus que deveriam estar depositados em repositórios digitais.

Conforme Cavalli (2009), a definição mais aceita pela comunidade para publicação sobreposta é dada por Van De Sompel et al. (2004). Para estes autores uma publicação sobreposta pode ser definida como: “uma publicação que não publica nenhum artigo original, mas seleciona artigos existentes em outros lugares, acrescenta certo valor e publica os resultados como um serviço para os usuários de sua base”.

Neste modelo a informação está descentralizada, isto é, ela pode estar em diversos repositórios, mas o criador da publicação sobreposta (que pode ser um editor, um bibliotecário, o especialista em curadoria etc) as reúne e acrescenta algum valor, que pode ser expresso por meio de metadados ou alguma relação com outro documento. Na visão de Ginsparg (1997), este tipo de publicação deve ser feito em uma interface que indique um conjunto de leituras essenciais de um dado assunto em um dado período, auxiliando o leitor a avaliar a informação dada.

Outro modelo de publicação encontrado na literatura é a publicação composta, ou objetos científicos compostos, do inglês, “*scientific compound objects*”. Este modelo, proposto por Cheung e colaboradores (2008) se define por encapsular vários conjuntos de dados (*datasets*) e recursos gerados ou utilizados durante a realização de um experimento científico, ou descoberta do processo, dentro de um único objeto composto para publicação e intercâmbio. Para estes autores, o potencial dos objetos compostos só é atingido se as informações estruturais de seus componentes forem compreensíveis tanto pelo homem quanto pela máquina.

Outro modelo chamado publicações semânticas visa “combinar documentos e ontologias permitindo que usuários acessem o conhecimento de várias maneiras” (ERICKSON, 2007a). Ao adicionar anotações padronizadas por meio de ontologias aos documentos digitais, estes documentos se tornam mais semânticos sendo possível melhorar a precisão da informação recuperada, processar conhecimento de forma automatizada e ainda gerar novos serviços.

Outra definição de publicações semânticas é a de Shotton que afirma ser esta

qualquer coisa que melhore o significado de um artigo de periódico publicado, facilite a sua descoberta automatizada, permite a sua ligação com os artigos semanticamente relacionados, fornece acesso a dados dentro do artigo em forma de recurso, ou facilite a integração de dados entre os artigos. (SHOTTON, 2001, p.86)

Para Shotton, a semântica do documento se dá a partir da contextualização da publicação. Essa contextualização pode ser feita de diversas maneiras, inclusive a partir da disponibilização dos dados relacionados ao conteúdo do documento.

Nesic (2010), em contraposição à definição de Erickson (2007a), afirma que este conceito deveria denotar não apenas documentos anotados com ontologia e sim uma nova categoria de documentos que pudesse contribuir plenamente com o ambiente idealizado pela Web semântica. Para Nesic, um documento semântico deve ser construído com base em quatro princípios:

1. O conteúdo do documento deve ser completamente consultável, com elementos endereçáveis e de granularidade diferentes;
2. Todo o documento deve ser identificado exclusivamente com URIs (*Uniform Resource Identifiers*);
3. Todo o documento deve ser anotado com conjuntos substanciais de metadados;
4. O conhecimento humano expresso no documento deve ser também representado de uma forma que possa ser processado por máquinas (isto é, os agentes de *software*);

Nesic (2010) recomenda a partir desses princípios uma nova definição para os documentos semânticos, a saber: “recurso composto unicamente identificado e semanticamente anotado”. Para ele, o recurso composto é formado por unidades menores que também devem ser identificados e anotados semanticamente. Assim, o autor propõe um modelo semântico de documento caracterizado por ter conteúdo (dados) e conhecimento representados nele de forma compreensível para humanos e agentes de software. Cada documento pode possuir diferentes tipos de relacionamentos com outros documentos ou com outros recursos (páginas *Web*, instituições, dados, etc) também devidamente identificados por suas URIs. Para definir a estrutura lógica do documento são utilizadas relações hierárquicas entre seus componentes.

Assim como o modelo de Nesic (2009), existem outras iniciativas de criação de publicação semântica, cada uma seguindo em uma direção, mas todas com o mesmo objetivo de melhorar a capacidade das publicações digitais de transmitirem conhecimento e significado. Marcondes (2011), por exemplo, propõe um modelo de publicação em que a semântica é explicitada no próprio texto do documento, isto é, nas conclusões que são fornecidas pelos autores e representadas em formato legível por máquina permitindo recuperação semântica e descoberta de novos conhecimentos. Outro exemplo é o modelo proposto por Erickson (2007b) em que a semântica é adicionada ao documento a partir de anotações feitas com base em uma ontologia OWL, assim, os conceitos existentes no documento são ligados à ontologia permitindo ao usuário alternar entre a navegação do documento e a ontologia.

Shotton (2009) também apresenta um modelo semântico de publicação em que o enriquecimento do conteúdo da publicação se dá por meio de acesso às bases de dados inclusas no documento. O modelo compreende marcação semântica de termos textuais, com *links* para outros recursos de informação relevantes, números interativos, lista de referência reordenável, resumo do documento que contém síntese do estudo, uma nuvem de *tags* e análise de citação. O modelo inclui ainda dois novos tipos de enriquecimento semântico: o primeiro, dicas para permitir "citações no contexto" e o segundo, uma taxonomia que reúne termos semanticamente relacionados. Além disso, publicam planilhas de dados e figuras para *download*, com as devidas informações de procedência, e demonstram vários tipos de fusão de dados (*mashups*), com os resultados de outros artigos de pesquisa e com o Google Maps.

Enquanto os modelos semânticos de Marcondes e Erickson acima citados se preocupam com a padronização para a melhoria da recuperação, o modelo de Shotton (2009), apesar de toda a sua sofisticação, permanece na extração de significado via contexto.

A definição de Shotton (2009) é relevante, pois caminha na mesma direção do modelo que esta tese propõe, ou seja: expandindo o conceito de publicação ampliada para o conceito de publicação ampliada semântica, onde os dados devem ser ligados às publicações com base em alguma semântica. Neste caso, a ideia é que o modelo se utilize da contextualização proporcionada pela adição de recursos como os sugeridos por Shotton (marcação semântica, *link* para dados de pesquisa, resumo com sínteses etc), mas que também se utilize da padronização descritiva e terminológica promovida pelo uso de metadados e pelo controle do vocabulário e suas relações.

Percebe-se então a preocupação da maioria das iniciativas com questões referentes à precisão da informação recuperada ou com o conhecimento processado. Além disso, tanto o uso de tecnologias semânticas quanto o uso de ontologias integram as novas propostas de publicação ao mundo da *Web Semântica* onde interoperabilidade e resposta precisa a questões de busca são desejáveis para potencializar o uso da *Web*.

O Quadro 2 a seguir apresenta uma síntese das características marcantes em cada um dos modelos de publicação.

Quadro 2: Síntese das novas iniciativas de publicações existentes

CARACTERÍSTICAS	PUBLICAÇÃO SOBREPOSTA	PUBLICAÇÃO COMPOSTA	PUBLICAÇÃO SEMÂNTICA
Liga dados	X	X	X
Utiliza metadados	X	X	X
Divide o documento em fragmentos menores (granularidade)	X		X
Reúne de documentos já existentes	X		
Informação descentralizada (distribuída por diversos repositórios)	X		
Utiliza Ontologias			X
Permite anotação			X
Gera novos serviços			X
A publicação dá significado aos dados (contextualiza)			X
Componentes possuem URI		X	X

Fonte: A autora

Como se pode perceber no quadro 2 acima, cada novo modelo de publicação proposto possui características relevantes. No entanto, nenhum dos modelos consegue reunir todas as propriedades importantes para a efetivação de um modelo que usufrua o potencial máximo das tecnologias vigentes. O modelo semântico que reúne características propostas por diversos autores (SHOTTON, 2009; NESIC, 2010; ERICKSON, 2007b) foi o modelo que mais acumula possibilidades para enriquecimento da publicação. Assim, pode-se compreender que ainda há uma lacuna a ser preenchida por um modelo de publicação que consiga integrar todas essas características.

O capítulo a seguir apresenta o objeto central dessa pesquisa: as publicações ampliadas e a sua formação, que é, de certa forma, a reunião de todos os tipos de publicação acima descritos.

4 PUBLICAÇÕES AMPLIADAS: O QUE SÃO E COMO SÃO FORMADAS

Em 2008, o Digital Repository Infrastructure Vision for European Research II - DRIVER-II38 – uma associação europeia de esforços voltados para a construção de uma infraestrutura de integração de repositórios desenvolvidos pela Comunidade Europeia - publicou um relatório sobre modelo de objetos em que definiu o termo *enhanced publications* ou publicação ampliada - termo que vem sendo adotado no país - da seguinte forma:

objetos digitais compostos que combinam *e-Prints* com um ou mais recursos de dados, um ou mais registros de metadados, ou qualquer combinação destes”. De acordo com o relatório de autoria de Peter Verhaar, *e-Prints* são entendidos como “um recurso textual como trabalho acadêmico original, que se destina a ser lido por seres humanos, que apresenta algumas reivindicações acadêmicas e que geralmente contém uma interpretação ou uma análise de determinados dados primários. (VEHAAR, 2008, p.11)

Conforme o relatório, o surgimento das publicações ampliadas foi impulsionado pela

constatação de que a publicação tradicional é limitada na sua capacidade para incorporar os resultados de todo o processo de descoberta científica, especialmente quando grandes conjuntos de dados foram gerados, o texto acadêmico normalmente pode apresentar apenas os dados da pesquisa de uma forma condensada. (VERHAAR, 2008, p.11).

Com o reconhecimento da importância do arquivamento, gestão e curadoria dos dados para posterior reuso e também para validação e replicação das pesquisas, vem crescendo o número de repositórios confiáveis que armazenam dados científicos.

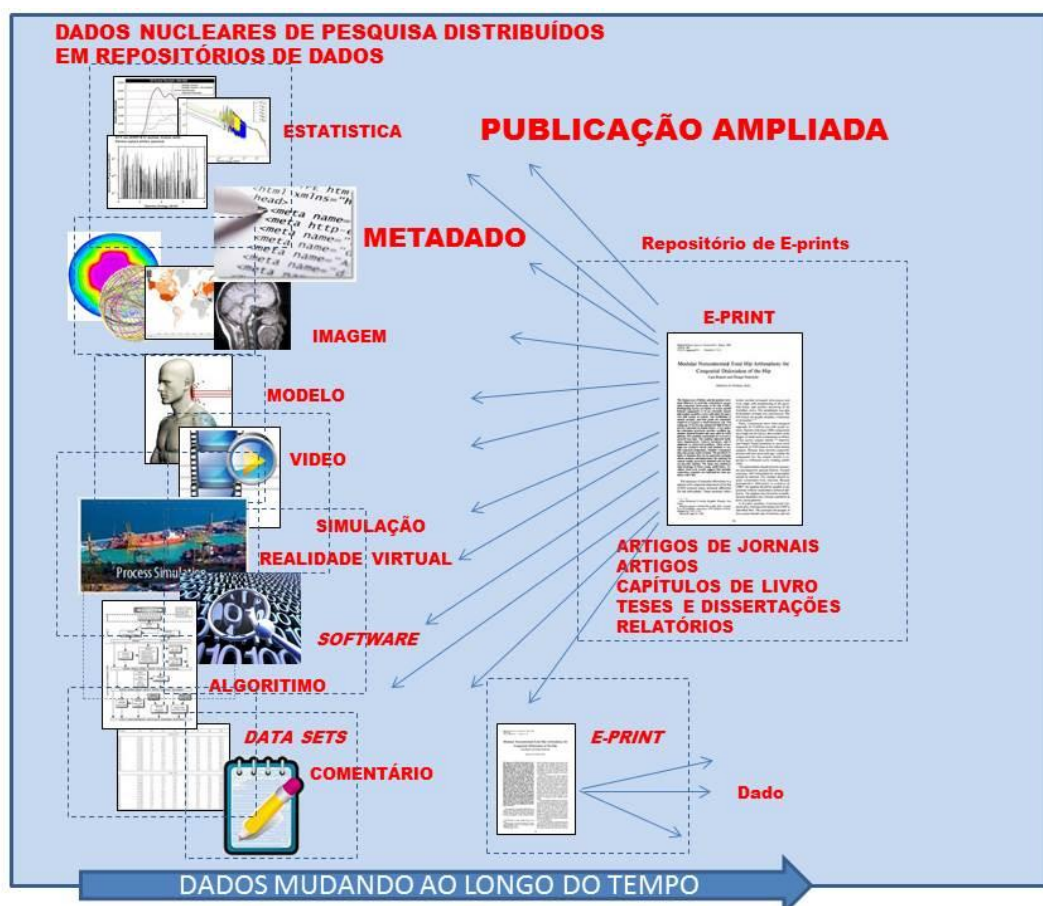
Contudo, o relatório constata como uma deficiência da infraestrutura atual da comunicação acadêmica o fato desses dados geralmente não estarem ligados às publicações científicas em que são discutidos. Assim, as publicações ampliadas foram criadas com o objetivo de criar pontes entre o conteúdo de repositórios institucionais e os repositórios de dados.

Mas como é formada uma publicação ampliada? “Uma publicação pode ser ampliada a partir da adição de um ou mais recursos a um *e-Print*” (VERHAAR, 2008, p.7). Estes recursos podem ser aqueles produzidos ou consultados durante a criação do texto e que, geralmente apoiam, justificam, ilustram ou esclarecem as afirmações científicas que são apresentadas em uma publicação. Woutersen-Windhouver e Brandsma (2009)

³⁸ Disponível em: <<http://www.driver-community.eu/>> Acesso em: 18 jun. 2014.

afirmam que uma “publicação é ampliada com dados de pesquisa, materiais extras, dados de pós-publicação, registros de base de dados (ex. *protein data bank*) e tem uma estrutura baseada em objeto com *links* explícitos entre objetos”. Assim, um objeto pode ser parte de um artigo, um *dataset*, uma imagem, um filme, um comentário, um módulo ou um *link* para informação em uma base de dados. A Figura 2 ilustra um modelo de publicação ampliada em que um *e-print* é ligado a diversos tipos de dados de pesquisa e os dados – armazenados em repositório específicos - são versionados e variam ao longo do tempo.

Figura 2: Modelo de publicação ampliada

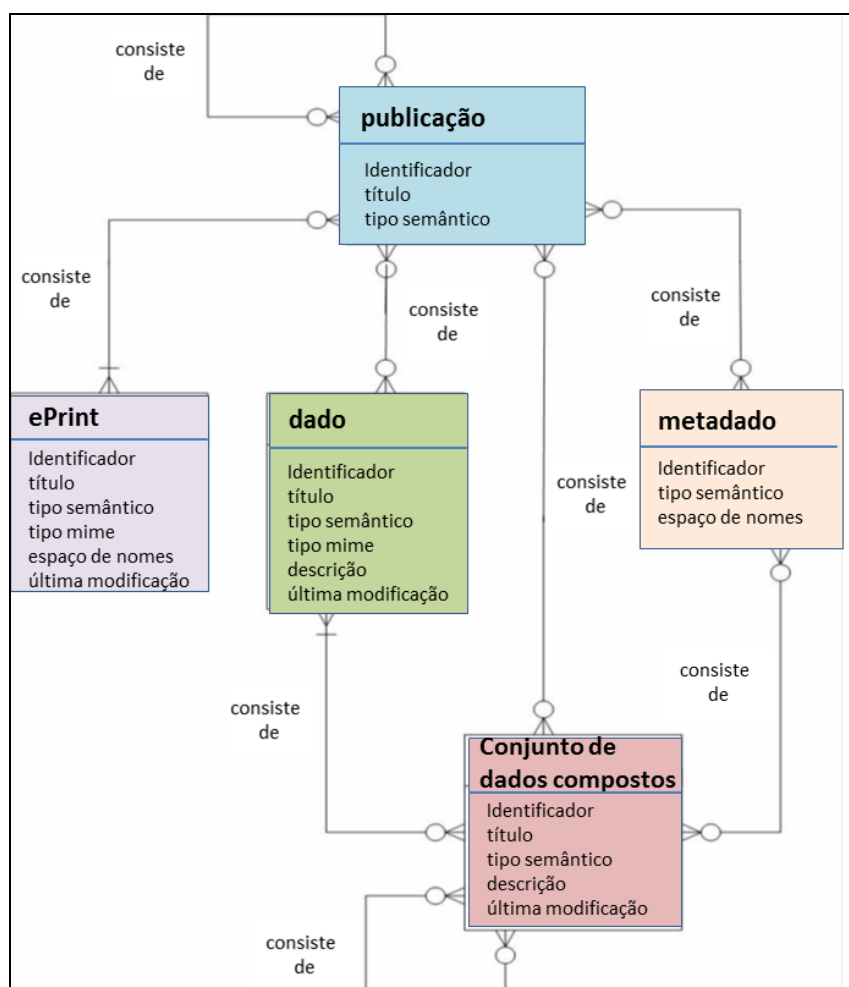


Fonte: A autora

Em geral, um manuscrito acadêmico é armazenado em um repositório institucional ou temático junto com outros componentes de outros repositórios que são adicionados a essa publicação como parte do fluxo de trabalho de projetos de pesquisa científica. Um exemplo simples de uma publicação ampliada pode se consistir num *e-print* combinado meramente com seu registro de metadados. Em outra definição de dados fornecida pela National Science Foundation (NSF) que abrange os métodos científicos a partir do qual os dados são atualmente obtidos, o dado pode ser definido como: "quaisquer entidades de

dados complexos obtidos a partir de observações, simulações de experiências, modelos e conjuntos de ordem superior, juntamente com a documentação associada necessária para descrever e interpretar estes dados" (NSF, 2007, p.2, grifo nosso). Ao incluir a “documentação associada” em sua definição, a NSF assume que metadados também são parte dos dados de pesquisa e que devem ser gerenciados. Como ilustração, a Figura 3 de Vehaar (2008) mostra o metadado também considerado como um tipo de dado.

Figura 3: Diagrama entidade-relacionamento para entidades básicas e propriedades



Fonte: VEHAAR, 2008 (tradução nossa)

No relatório de Vehaar (2008) foi identificada ainda uma série de dez requisitos e recomendações que caracterizam uma publicação ampliada, a saber:

1. Deve ser sempre possível especificar as partes componentes de uma publicação ampliada.

2. A publicação ampliada e seus componentes devem ser disponibilizados como recurso *Web* identificados por URI. O mesmo vale para seus componentes.
3. Deve ser possível agregar objetos digitais compostos à publicação ampliada.
4. Deve ser possível acompanhar as diferentes versões das publicações ampliadas como um todo e das suas partes constituintes.
5. Deve ser possível registrar as propriedades básicas da publicação e dos outros recursos que estão adicionados a ela.
6. Deve ser possível registrar a autoria da publicação ampliada e dos itens que a compõem.
7. Deve ser possível assegurar a preservação de longo prazo das publicações ampliadas.
8. Deve ser possível registrar as relações entre os recursos *Web* que fazem parte da publicação ampliada.
9. Instituições que oferecem acesso a publicações ampliadas devem assegurar que elas possam ser recuperadas.
10. Instituições que oferecem acesso a publicações ampliadas devem assegurar que estas estejam disponíveis como documentos baseados na norma OAI-ORE, que será vista com mais detalhes na seção 4.4.

Estes requisitos são apresentados visualmente no diagrama entidade-relação (figura 3). Este diagrama mostra que as publicações ampliadas consistem de cinco tipos de entidades: *e-prints*, dados, metadados, conjuntos de dados compostos (*datasets*) e outras publicações ampliadas.

Em outra publicação do DRIVER-II elaborada em conjunto com o SURF-Foundation, em 2009, intitulada “Enhanced Publications: Linking Publications and Research Data in Digital Repositories” é descrita, entre outros itens, a estrutura de uma publicação ampliada, que é formada de elementos fornecidos pelo autor, pelos avaliadores e pelo editor da publicação. (VERNOOY-GERRITSEN, 2009b)

De acordo com Lagoze e colaboradores (2008), os elementos fornecidos pelo autor são: **tipo de semântica**, por exemplo, artigo, simulação, vídeo, conjunto de dados, *software*; **tipo de mídia**, por exemplo, texto, imagem, áudio, vídeo, misto; **formato de mídia**, por exemplo, PDF, XML, MP3; **endereço da rede**, pois os componentes podem ter diferentes endereços; e **acessibilidade por diferentes repositórios**. Além desses, a publicação pode conter também os anexos e apêndices, porém estes não são obrigatórios.

Os elementos fornecidos pelos avaliadores não são mencionados como obrigatórios, mas a partir dos exemplos fica claro que estes elementos devem ser: **notas on line**, **comentários dos colegas**, **discussão interativa** e **classificação do artigo**.

Os elementos fornecidos pelos editores também não são mencionados como obrigatórios, mas a literatura deixa transparecer que são todas as facilidades que o editor deve fornecer para melhorar a precisão da recuperação da informação, aumentando a visibilidade do artigo e o interesse do leitor. Como exemplo, pode-se citar: o uso de **linguagens de marcação** para auxiliar a busca e a disseminação, a indexação ou “**tagueamento**” do artigo, o **uso de normas** e padrões para publicação, o **uso de vocabulários controlados** ou *link* para ontologias e ainda **interface amigável**, o que permite que o usuário identifique facilmente outros recursos.

O relatório do SURF-Foundation também descreve algumas propostas de modelos para publicações ampliadas e afirma que em todos eles há uma tentativa de adicionar estrutura pelo uso de metadados nos mais diferentes níveis. Estes modelos são a base da proposta da publicação ampliada. Os modelos são: modelo de Kircz, modelo de Hunter, modelo de Marcondes, Modelo de Van de Sompel e Modelo de Seringhaus e Gerstein.

O modelo apresentado por Kircz (2002) é uma forma modular de publicação, assim um artigo acadêmico é dividido em módulos vinculados por *links* entre si. Um módulo pode ser uma parte do artigo, por exemplo, resumo, informação bibliográfica ou pode ser um conjunto de dados ou de informações registradas em outras mídias. Um módulo é definido como uma representação de uma unidade conceitual de informação, unicamente identificada e autocontida cujo objetivo é comunicar a informação que ele porta (HARMSZE, 2000, p.39). Cada módulo possui seu conjunto de metadados, o que permite que cada parte da unidade de informação possa ser citada separadamente, reutilizada para diferentes propósitos ou ainda recombinação, formando novos documentos complexos.

O modelo de Hunter (2006) tem por base os conceitos, assim como o de Marcondes (2009) - já citado neste trabalho como modelo de publicação semântica. Já o modelo de

Lagoze e colaboradores (2008) é baseado em classes e instâncias.

Outra proposta de infraestrutura para publicações ampliadas encontrada na literatura é a de Seringhaus e Gerstein (2007), que sugere que as publicações ampliadas devem: 1) capturar uma gama de dados em formato digital e facilitar seu depósito em banco de dados juntamente com a publicação do manuscrito; 2) indexar todos os artigos de periódicos em texto completo, associando palavras-chave e identificadores com os registros do banco de dados e *links* para livros, *sites* de laboratório e comentário de alto nível; 3) fornecer vários níveis de revisão por pares, comentários da comunidade e anotação; 4) produzir artigos totalmente legíveis por máquina, fornecendo marcação inteligente e resumos digitais estruturados; 5) ser uma rede de informações acessíveis através de um único portal.

Percebe-se que todos esses modelos tornam-se mais estruturados quando revelam sua semântica mais explicitamente. A estruturação semântica não é apenas importante para a leitura e compreensão humana, mas também é útil para a máquina, pois à medida que torna a informação legível, possibilita a criação de vários tipos de processamento, como por exemplo, a mineração de dados.

É esperado também que “imagens estáticas e em movimento, sons, simulações e em breve também informação tátil possam ser trocadas, experimentadas, e, portanto, analisadas e interpretadas, por diferentes pessoas em diferentes ambientes culturais e épocas” (KIRCZ, 2002). Todos estes componentes do documento eletrônico devem estar aderentes com o corpo de normas e padrões pertinentes que garantam sua qualidade e integridade. A seção a seguir apresenta algumas iniciativas já existentes e tem por finalidade ilustrar esse novo modelo de publicação acadêmica que vem sendo chamado de “Publicação Ampliada”.

4.1 INICIATIVAS CONCRETAS EXISTENTES

Em 2009, a editora acadêmica Elsevier³⁹ lançou o projeto “Artigo do futuro”, visando repensar o artigo científico na era eletrônica. A ideia do projeto é oferecer uma nova plataforma para comunicar ciência no mundo digital atual. O conceito de “artigo do futuro” foi desenvolvido em estreita colaboração com a comunidade científica, envolvendo *feedback* de centenas de pesquisadores. A contribuição oriunda do projeto foi encaminhada em três direções: **apresentação** - oferecendo uma navegação *online* mais próxima da experiência da leitura; **conteúdo** – entregando junto ao artigo outras informações relevantes como dados científicos, código de computador e arquivos multimídia; e **contexto** – ligando o artigo a outros recursos científicos de confiança no contexto do artigo.

Assim como a Elsevier, diversas outras iniciativas isoladas podem ser encontradas na *Web*. Conforme apresentado por Shotton (2009) a *SourceOECD*⁴⁰ uma biblioteca *on-line* de bancos de dados estatísticos, livros e periódicos da Organização para Cooperação e Desenvolvimento Econômico (OCDE)⁴¹ disponibiliza não apenas seus quadros estatísticos, mas também um guia de exportação que cria e transfere os dados para uma planilha em formato MS excel.

Outro exemplo é o periódico “The New England Journal of Medicine – NEJM”⁴² que possui um site altamente interativo, onde apresenta não apenas novos artigos, mas também uma variedade de publicações de áudio e vídeo, junto com um desafio de imagem em que estimula os leitores a fazerem o diagnóstico médico correto a partir das imagens exibidas.

Essas iniciativas isoladas certamente são frutos do movimento científico em prol dos novos modelos de publicações e, conforme já definido, aparecem na literatura denominadas de diversas formas como “publicações sobrepostas”, “publicações compostas”, “publicações semânticas” entre outras.

Ambinder (2012), em sua dissertação de mestrado, faz um levantamento exaustivo das iniciativas existentes em termos de novos modelos de publicação e as classifica em: **iniciativas de uso de linguagem XML** para marcação e publicação de artigos científicos na *Web*; **uso de ontologias** em publicações científicas e **sistemas inovadores de**

³⁹ Disponível em: <<http://www.articleofthefuture.com/>>. Acesso em: 14 maio 2014.

⁴⁰ Disponível em: <<http://www.sourceoecd.org/>>. Acesso em: 27 maio 2013.

⁴¹ Disponível em: <<http://www.oecd.org/>>. Acesso em: 18 maio 2013.

⁴² Disponível em: <<http://content.nejm.org/>>. Acesso em 27 maio 2013.

publicações científicas eletrônicas.

Outro artigo que faz uma análise detalhada das iniciativas de publicações ampliadas existentes é “*Rich Internet Publication: show what you tell*” de Breure, Voorbij e Hoogerwerf (2011) onde os autores relatam recursos que utilizam ações de integração, visualização e exploração de informação acadêmica não-linear.

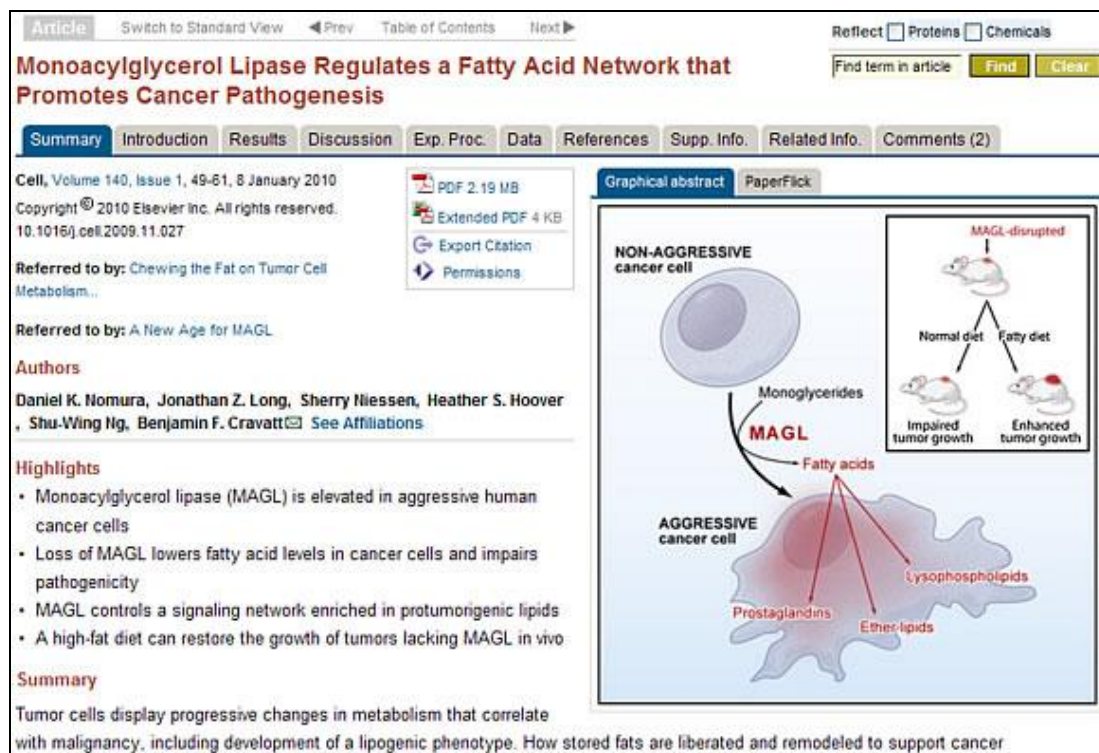
Neste trabalho serão destacadas apenas duas experiências especiais no que diz respeito às publicações ampliadas, por serem as que mais se assemelham ao modelo que esta tese vem propor: a da Revista *The Cell*⁴³ e o demonstrador do Driver-II, utilizando uma experiência de *enhanced publication* sobre jornalistas na Holanda pelo *Data Archiving and Networked Services*⁴⁴ - DANS.

A nova interface da revista *The Cell* possui como característica mais importante a estratificação do artigo de forma explícita e a integração interna das informações exibidas. Soma-se a isto um resumo textual e outro gráfico, além de um vídeo em que os autores apresentam suas descobertas. No topo possui uma aba que corresponde à divisão tradicional do artigo científico (sumário, introdução, resultados, discussão, referências, etc) (Figura 4) e ainda uma aba ‘*data*’ que fornece gráficos e outros recursos utilizados como dados para a pesquisa. Interessante é que estes recursos estão ligados ao corpo do texto, de modo que cada quadro pode ser visto em sincronia com seu próprio contexto. (Figura 5). Possui ainda uma aba ‘*Comments*’ que abre um *blog* com comentários do autor, reações dos pares, entre outros.

⁴³ Disponível em: <<http://www.cell.com/>>. Acesso em 27 maio 2013.

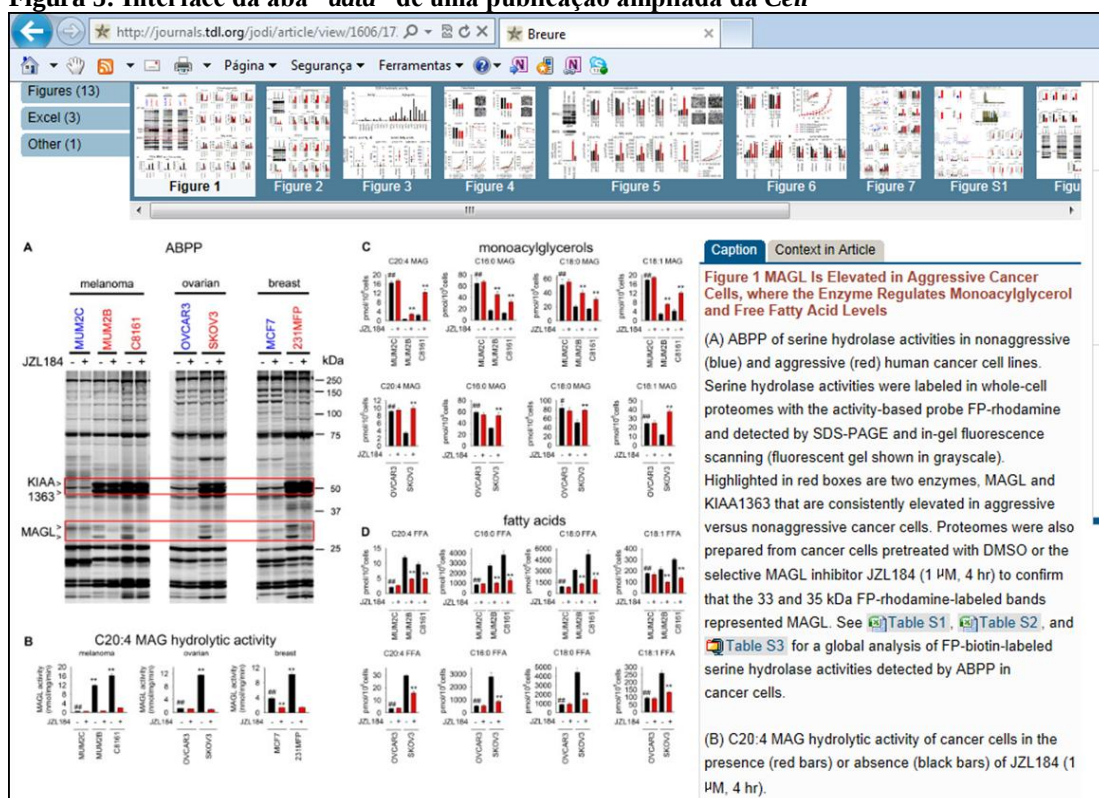
⁴⁴ Disponível em: <<http://www.dans.knaw.nl/en>>. Acesso em: 27 maio 2013.

Figura 4: Interface de artigo da *the Cell* como modelo de publicação ampliada



Fonte: Breure; Voorbij; Hoogerwerf (2011)

Figura 5: Interface da aba "data" de uma publicação ampliada da *Cell*



Fonte: Breure, Voorbij e Hoogerwerf (2011)

Outra Iniciativa é a do DRIVER-II que oferece uma visão abrangente dos diversos componentes de uma publicação ampliada. O texto é disponibilizado em formato PDF e, conjuntos de dados, dependendo do assunto, podem ser exibidos em um formato ou consultados através de uma interface de banco de dados (Figura 6).

O DRIVER II tem o objetivo de estabelecer uma infraestrutura de repositórios digitais coesa para apoiar o desenvolvimento da pesquisa na Europa, oferecendo serviços com funcionalidades sofisticadas para os pesquisadores e o público em geral. As informações a serem fornecidas incluem as três categorias: dados da pesquisa, materiais extras e dados pós-publicação (como comentários e ranqueamento). Além disso, integra os metadados ligando os objetos uns aos outros. O demonstrador do DRIVER-II oferece também a possibilidade de reuso de objetos digitais compostos, pois utiliza o padrão OAI-ORE que permite a definição e identificação de agregações, possibilitando a combinação de recursos distribuídos com vários tipos de mídia (textos, imagens e vídeo). A Figura 6, mais adiante, mostra uma publicação ampliada exibida via interface do demonstrador do DRIVER-II que consiste dos dados da pesquisa de um livro. O livro em si está disponível como um arquivo completo e uma coleção de capítulos separados. Dados SPSS⁴⁵ são acessíveis através de 'Source'

⁴⁵ Dados SPSS são dados obtidos por meio do *software* da IBM-SPSS, usado para apoiar a tomada de decisão através de aplicativos de *data mining*, *text mining* e estatística.

Figura 6: Interface do demonstrador do DRIVER-II

The interface shows a top navigation bar with tabs: Content, Relations, Comments, Source, and DRIVER 2 - Demonstrator. A 'Parents: none' button is on the right. The main content area is divided into two panels.

Panel 1: Journalists in The Netherlands

EP: Journalists in The Netherlands

An enhanced publication about journalists in The Netherlands. It is the result of a reserach to journalists and current social changes such as multiculturalism, the use of internet and the integration of news and entertainment. This enhanced publication consists of tha data from the survey and the book. The book itself is available as a complete file and a collection of separate chapters.

Author	M.J.P. Deuze
Persistent ID	urn:x-driver:378001
Publisher	Data Archiving and Networked Services
Created	2004-10-08
Modified	2008-10-17
Rights	OpenAccess

Panel 2: Journalists in The Netherlands

This book provides a profile of journalists in The Netherlands, looking at their professional role in contemporary multicultural, digital and commercial society. The analysis futhermore compares Dutch, Australian, German, British and American journalists, offering a comprehensive international overview of this cornerstond profession in todays' democracy.

Author	M.J.P. Deuze
Persistent ID	urn:x-driver:378004
Type	doctoralThesis
Publisher	University of Amsterdam
Created	2002-03-13
Modified	2002-03-13
Rights	RestrictedAccess

Panel 3: Resources

Surveydata 'Journalism in the Nether...

Author	M.J.P. Deuze
Persistent ID	urn:x-driver:378002
Type	dataset
Publisher	Data Archiving and Networked Services
Modified	2008-09-11
Rights	OpenAccess

Open

Fonte: Breure; Voorbij; Hoogerwerf (2011)

Esses modelos de publicação ampliada existentes agregam publicação e dados de pesquisa, mas sem expressar a semântica da relação existente entre os seus componentes. A proposta desta tese é que essas relações sejam explícitas e expressem o porquê de um objeto está ligado a outro. Isso é importante para a visualização do todo e suas partes e também para uma recuperação de informação precisa. A seção a seguir apresentará como uma publicação ampliada deve ser estruturada, isto é, sua arquitetura, detalhando posteriormente cada um de seus componentes

4.2 ARQUITETURA DE PUBLICAÇÃO AMPLIADA

Uma publicação ampliada, como já mencionado, é uma instância de um objeto digital composto, formado por vários elementos e suas relações. Estes elementos são objetos digitais simples que podem ser chamados de unidades atômicas, pois se constituem como menor unidade de um objeto digital composto. Entretanto, dada a complexidade e o dinamismo temporal inerente a estes objetos informacionais, outros requisitos devem ser considerados para a configuração de um registro completo de uma publicação ampliada, tais como selo de tempo (*timestamp*), identificação persistente, dispositivos de resolução de *Uniform Resource Identifier* - URI em endereços, marca de revisão por pares que garantam melhores níveis de confiabilidade. Esses requisitos são, na verdade, metadados usados para descrever minuciosamente cada elemento de um agregado que forma uma publicação ampliada.

Reforçando as características cruciais de um documento acadêmico, o relatório do SURF (VERNOOY-GERRITSEN, 2009a) recomenda que na composição de uma publicação ampliada, cada função do sistema de comunicação científica – ou seja, registro, certificação, disseminação, arquivamento e recompensa – deve ser reimplementada segundo os novos contextos em que se apresentam.

O **registro** é a principal função do sistema de comunicação científica. É por meio dos registros que as descobertas, invenções e inovações são tornadas públicas e reivindicadas como originais e como garantia de autoria, entretanto, esses registros devem ser expandidos para que possam representar mais fielmente a complexidade das pesquisas contemporâneas. Altman e King (2007) propõem que uma publicação ampliada, por conter em sua estrutura dados e outros componentes, além dos três componentes clássicos de citação de qualquer publicação tradicional - autor (s), título e data de publicação - deve possuir uma forma de registro especial, incluindo outros três metadados: selo de tempo, *Universal Numeric Fingerprint* (UNF) e um identificador único. Estes metadados adicionais são necessários para identificar um determinado objeto de forma inequívoca, garantir uma localização estável, fácil recuperação e verificação do estudo.

O *timestamp*, ou selo de tempo, é uma sequência de caracteres de informação codificada que identifica a ocorrência de certo evento por data e hora do dia, considerando as frações de segundo. O *timestamp* de uma publicação é registrado pela editora com a data em que o manuscrito é recebido. Para o caso de uma publicação ampliada o selo de tempo

deve ser assinalado para a publicação como um todo e também para cada unidade atômica que a compõe.

O UNF, ou impressão digital numérica universal, garante que um objeto digital não foi modificado de maneira significativa e evidencia se objetos complexos são iguais ou diferentes, posto que qualquer mudança intencional ou acidental sofrida pelos objetos digitais muda a UNF resultante. Sua representação se dá na forma de uma sequência curta de números e caracteres, com comprimento fixo e que resume todo o conteúdo do conjunto de dados de forma que uma alteração em qualquer parte dos dados produza um UNF completamente diferente. Para as publicações ampliadas, o UNF é de extrema importância, pois, por ser uma agregação de dados que podem ser corrigidos, atualizados e mantidos por mecanismos de curadoria distintos, a verificação se dois objetos são o mesmo ou não pode ser uma tarefa complexa. O UNF é obtido pela tradução dos dados, em primeira instância, em uma forma canônica com graus fixos de precisão numérica; aplicação de uma função hash criptográfica resultando em uma cadeia curta de caracteres, independente do meio de armazenamento e do formato do objeto, assim o mesmo valor pode ser mantido mesmo quando o conjunto de dados é transferido entre programas, sistemas ou plataformas.

O Unique Global Identifier (UGI), por sua vez, é um nome curto ou uma cadeia de caracteres única que identifica e nomeia o conjunto de dados de forma permanente, independentemente da sua localização e de tecnologias. As publicações ampliadas precisam de um UGI para a publicação como um todo e também para cada objeto que a compõe. A infraestrutura dos esquemas de identificação persistente e globalmente única deve estar associada a serviços de resolução, cuja função é traduzir o nome ou identificador persistente de um objeto em endereço corrente que localize o recurso, por exemplo, em Uniform Resource Locator (URL) que são reconhecidos pelos navegadores. Entretanto, o identificador global único não deve apenas resolver o objeto em si, mas deve apontar para uma página contendo os metadados descritivos e estruturais do objeto. Esta página com os metadados descritivos deve conter um *link* para o acesso ao objeto real, bem como uma descrição textual do objeto, a citação, a documentação completa e qualquer outra informação pertinente.

Outra função essencial do sistema de publicação é a **certificação** cuja função é estabelecer a validade da solicitação de um registro acadêmico que, na maioria dos casos, se concretiza por meio do processo de revisão por pares conduzida pelas editoras científicas. No que tange as publicações ampliadas é preciso certificar também os objetos

suplementares. De acordo com o relatório do SURF (VERNOOY-GERRITSEN, 2009a, p.49), em algumas áreas há um esforço para construir sistemas de avaliação por pares em torno dos dados, de modo que estes possam ser julgados formalmente em termos de coerência, *design*, consistência e confiabilidade de acesso.

Além do registro e da certificação, outra função do sistema de publicação científica é conscientizar as comunidades interessadas sobre novas afirmações e conclusões, ou seja, **disseminar** informação sobre as novas descobertas garantindo aos autores visibilidade, autoridade e prestígio. As publicações ampliadas auxiliam nessa disseminação à medida que seus complementos estão distribuídos e também indicam a publicação de origem. Assim, pesquisadores podem ter acesso à informação, via *site* da editora, via repositório institucional temático ou de dado ou através do *blog* de um pesquisador que funcione como outro canal de informação, mas que também seja um componente da publicação ampliada. Para que isso seja possível o SURF recomenda que:

1. A publicação e os objetos relacionados devem ambos ter um único identificador para garantir que possam ser referenciados;
2. Os *links* entre a publicação e os materiais suplementares devem ser configuradas em ambos os lados, a partir da publicação dos materiais complementares e vice-versa;
3. Os objetos devem ser depositados em local confiável e publicado em regime de acesso aberto, o que resultará em mais *downloads* e muito mais citações .

É importante mencionar ainda que para a publicação ampliada cumprir a sua função de dar ciência sobre as pesquisas desenvolvidas é preciso que os materiais complementares sejam armazenados em locais adequados, como repositórios temáticos, nacionais e internacionais e que utilizem identificadores persistentes para preservar e disseminar esses materiais suplementares. Lynch (2007) declara que revistas são menos adequadas, pois nem toda revista aceita materiais suplementares; nem sempre as revistas têm uma política clara sobre a preservação de dados ou as ferramentas para trabalhar com os dados; algumas revistas impõem restrições sobre a quantidade de dados que aceitarão; o acesso aos dados é somente para assinantes; e por fim, nem sempre é claro até que ponto os materiais

suplementares são parte da revisão por pares.

Um exemplo de local alternativo para armazenar conjuntos de dados é o *Dataverse NetWork*⁴⁶ - DVN, uma plataforma criada pelo Harvard Institute of Quantitative Social Science para pesquisadores armazenarem e gerenciarem seus dados de pesquisa, com descrições textuais de seus estudos e metadados e liberá-los ou não, conforme desejarem. Assim como esta, outras plataformas utilizam tecnologia de computação em grade (*grid*) para controle e compartilhamento de dados. Essa tecnologia é interessante, pois muitas aplicações científicas e de engenharia geram e requerem acesso à grande quantidade de dados distribuídos.

Somando às funções já citadas, a questão do **arquivamento** também deve ser considerada na composição de uma publicação ampliada. Esta função visa preservar a produção científica ao longo do tempo. As bibliotecas vêm desempenhando o papel de guardiãs da produção bibliográfica dos pesquisadores, mas no caso das publicações ampliadas, que envolve também o arquivamento de materiais suplementares, como dados de pesquisa, quem será o responsável por esse arquivamento seguro? Editoras? Bibliotecas? Instituições de pesquisa? O relatório do SURF (VERNOOY-GERRITSEN, 2009a) considera “que as universidades e instituições de pesquisa devam assumir a responsabilidade de arquivamento de seus próprios materiais acadêmicos digitais nos quais se têm investido”, assegurando que os resultados de pesquisas estarão acessíveis no futuro, independente das mudanças nos métodos de armazenamento e distribuição de dados e acesso à rede, independente dos formatos do *software*, dos bancos de dados, dos sistemas operacionais ou do *hardware*.

Por último, é preciso considerar também na composição de uma publicação ampliada a função de **recompensa** de uma publicação científica. A recompensa para um pesquisador é a própria citação em uma revista de alto impacto, que gera, por exemplo, visibilidade e novas citações. No caso da publicação ampliada, apesar de ser vista como um todo relacionado tematicamente, o relatório do SURF sugere que publicações e conjuntos de dados sejam encarados como dois objetos de comunicação científica separados, com referências próprias, a fim de que os materiais suplementares tornem-se visíveis. (VERNOOY-GERRITSEN, 2009a). Pesquisadores devem ser recompensados também pela citação de suas imagens, ferramentas, comentários, bancos de dados, etc, mas

⁴⁶ Disponível em: <<http://thedata.org/>>. Acesso em: 30 mar. 2014.

para isso é preciso que esses objetos possuam uma identificação única e que sua citação seja clara e ainda que novas políticas de avaliação sejam definidas.

Mais uma vez, para a formação de publicações ampliadas, relações bem estabelecidas se tornam essenciais para compor, descrever, gerenciar, preservar, prover serviços, possibilitar recuperação precisa da publicação ou de seus componentes, tornar o documento interoperável em todos os níveis, entre outras funções.

A evolução da curadoria de dados digitais cria grandes oportunidades no mundo da publicação acadêmica. Já não é mais possível condensar toda a pesquisa científica no formato tradicional de artigo, por isso faz-se necessária a incorporação de outros elementos que apoiem a revisão por pares. Essa nova identidade da publicação científica requer a contextualização desses novos elementos por meio de novos relacionamentos. A partir disto uma questão que merece atenção é: que tipos de relações devem ser usados na ligação desses dados à publicação?

O relatório do DRIVER-II (VERHAAR, 2008) apresenta, como um dos requisitos das publicações ampliadas, a existência de relações entre os recursos *Web* que fazem parte dessa publicação. Nesse sentido, o relatório oferece uma classificação dos tipos de relações que podem ocorrer entre os componentes de uma publicação ampliada, a saber: relações de inclusão, relações sequenciais, relações de versionamento, relações de linhagem, relação de manifestação e relação de citação.

As **relações de inclusão** são aquelas que expressam que uma unidade está fisicamente ou logicamente dentro de outra unidade. Esta relação é a mais comum, pois ocorre sempre que unidades são agrupadas dentro de uma unidade maior, como por exemplo, o capítulo de uma tese.

As **relações sequenciais** são aquelas utilizadas para registrar a ordem em que os recursos devem ser consultados, é o caso, por exemplo, de uma tese em que os capítulos sejam arquivados como partes separadas. O objetivo desse tipo de relação é estabelecer um caminho de leitura dentro de um documento.

As **relações de versionamento** são aquelas que revelam a existência de diversas versões de um mesmo recurso.

As **relações de linhagem** são aquelas que fornecem informações sobre a ordem em que os dados de pesquisa são produzidos. Esta linhagem se refere às etapas da cadeia de processamento em que dados científicos são gerados e permite que os pares possam traçar as várias fases do fazer científico.

As **relações de manifestação** são aquelas que revelam os diferentes formatos técnicos de um recurso, por exemplo, uma tese pode estar disponível em formato PDF e/ou HTML. Além disso, essa mesma tese pode conter imagens em TIFF ou JPG, isto significa que uma publicação ampliada pode ser o agrupamento de diversos formatos de arquivo e isso precisa ser expresso também por meio de relações.

As **relações de citação** são aquelas que ligam os recursos a outros recursos que são citados por este recurso por meio de referências. Essas relações geralmente apontam para recursos externos à publicação, mas que podem estar armazenadas num mesmo repositório.

Outra classificação apresentada pelo relatório do DRIVER II é quanto à existência de relações inversas. Neste contexto, as relações podem ser unidirecionais e bidirecionais. As relações unidirecionais são aquelas em que o recurso A tem uma relação com B, mas o inverso não pode ser assumido automaticamente; enquanto a relação bidirecional expressa que se A se relaciona com B, B também se relaciona com A, por exemplo, a relação *<hasPart>* deve ter uma relação inversa expressa por *<isPartOf>*. Apesar de parecer redundante e nem sempre ser possível criá-las - já que os componentes de uma publicação ampliada, na maioria dos casos estão distribuídos por repositórios distintos - o uso da relação bidirecional é importante, pois permite que cada recurso carregue em si informações explícitas sobre o que o envolve.

A norma OAI-ORE (Open Archives Initiative Object Reuse and Exchange), que será descrita na seção 4.4, define um vocabulário e ainda alguns relacionamentos para objetos complexos que podem ser usados para construção de uma publicação ampliada. De acordo com a norma, os tipos de relações que existem entre as entidades ou entre uma entidade e um valor literal são: *<aggregates>*, *<isAggregatedBy>*, *<describe>*, *<isDescribedBy>*, *<lineage>*, *<proxyFor>*, *<proxyIn>* e *<similarTo>*.

A relação Agrega, representado por *<ore:aggregates>*, expressa que um objeto do recurso é um membro do conjunto de recursos agregados e tem como inversa a relação *<ore:isAggregatedBy>* que expressa que um conjunto de recursos possui determinados agregados.

A relação Descreve, representada por *<ore:describe>*, expressa que um dispositivo chamado “mapa de recurso” descreve o objeto (uma agregação) e tem como inversa a relação *<ore:isDescribedBy>*, significando que o objeto é descrito pelo mapa de recurso.

A relação de Linhagem é utilizada quando há necessidade de expressar a cadeia de proveniência de onde vieram os recursos que formam a agregação. Desta forma, *<ore:lineage>* é uma relação entre dois objetos substitutos ou representantes de um recurso, com linhagens distintas do mesmo dado, sendo que ambos devem ser *Proxy* do mesmo recurso.

A relação *<ore:proxyFor>* é usada para ligar o Proxy ao recurso agregado, ou seja, ao componente da agregação, enquanto que a relação *<ore:ProxyIn>* liga o Proxy à agregação, isto é, ao recurso total.

Já a relação Similar é representada por *<ore:similarTo>* e liga uma agregação a outra expressão ou manifestação desta agregação, por exemplo, o conjunto de slides apresentados em um congresso e os anais deste mesmo congresso.

De acordo com o projeto *Digital Fedora Object Relationship* do Fedora Commons (2008), os objetos digitais se relacionam de muitas maneiras, podendo ser uma parte de outro objeto, membro de uma coleção, uma derivação ou descrição de outro objeto, ou ainda podem ser substitutos ou proxies de outros objetos. Esses relacionamentos se dão por meio de um conjunto de relações genéricas comuns e úteis na criação de redes de objetos digitais, a saber: *<isConstituentOf>*, *<hasConstituent>*, *<hasMember>*, *<isMemberOf>*, *<isSubsetOf>*, *<hasSubset>*, *<hasCollectionMember>*, *<isDerivationOf>*, *<hasDerivation>*, *<isDependentOf>*, *<hasDependent>*, *<isDescriptionOf>*, *<hasDescription>*, *<isMetadataFor>*, *<hasMetadata>*, *<isAnnotationOf>*, *<hasAnnotation>*, *<hasEquivalent>*. Essas relações podem ser refinadas ou estendidas de acordo com a necessidade de cada comunidade de usuários.

No Fedora, as relações são estabelecidas de um objeto para outro objeto, seguindo o seguinte padrão de afirmação: o primeiro objeto Fedora é considerado o ‘sujeito’ da afirmação; o segundo é a própria ‘relação’ e o terceiro é uma ‘propriedade do sujeito’. No Fedora, o objeto alvo é o objeto relacionado. Um exemplo de sentença na língua inglesa é: *<MyCatVideo> <is a member of collection> <GreatCatVideos>*.

Essas relações objeto-a-objeto são armazenadas no Fedora como metadados de objetos digitais dentro de um fluxo especial de dados. Este fluxo é reconhecido por um identificador chamado "RELS-EXT" (que significa relacionamento externo) que é usado

exclusivamente para afirmar relações entre objetos digitais.

Percebe-se que as relações em um objeto digital complexo se dão em tríades, que são declarações ou afirmativas que podem ser expressas em RDF e assim deve ser feito para modelar as relações de uma publicação ampliada.

Percebe-se também que apesar dos trabalhos citados apresentarem a importância das relações e inclusive citarem alguns exemplos, mesmo nos trabalhos mais detalhados, como o da ontologia de Relacionamentos do Fedora, não existe uma sistemática para estabelecimento dessas relações.

As seções que seguem visam detalhar cada um dos elementos que compõem a publicação ampliada, isto é, os objetos digitais, os metadados e as relações e, em seguida apresentará a norma utilizada para integrar esses elementos e como integrar esse novo modelo ao contexto da Web Semântica.

4.3 OBJETOS DIGITAIS: PRIMEIRO COMPONENTE ESSENCIAL DA PUBLICAÇÃO AMPLIADA

O conceito de publicação ampliada conforme proposto por esta tese é viabilizado pela TICs que propicia que publicações científicas e dados de pesquisa, na maioria dos casos, nasçam em formato digital. Essas publicações e dados nascidos em formato digital são em sua essência objetos digitais e por isso pode-se afirmar que as publicações ampliadas são formadas necessariamente por estes objetos. Assim sendo, torna-se de fundamental importância para a construção de uma publicação ampliada a compreensão do que vem a ser objetos digitais.

Objetos digitais⁴⁷ são definidos pelo modelo de referência do Open Archival Information System (OAIS)⁴⁸ “como qualquer item formado por uma cadeia de bits” (CCSDS, 2002). Esses objetos digitais são originados de duas formas: a partir da conversão de seus originais analógicos para um formato digital através de processos de digitalização ou obtidos diretamente de algum dispositivo digital, por exemplo, uma

⁴⁷ A literatura também apresenta outros termos para designar o mesmo referente, como por exemplo, documentos digitais, artefatos digitais, porém objeto digital é o mais usado.

⁴⁸ Reference model for an Open Archival Information System - OAIS é um documento desenvolvido pelo Consultative Committee for Space Data Systems - CCSDA que estabelece um sistema de arquivamento de informações físicas ou digitais para fins de preservação e disponibilização para uma dada comunidade.

máquina fotográfica, ou diretamente do computador, quando são criados desta forma, sem nenhum equivalente analógico.

Para que um objeto digital possa ser compreendido por seres humanos é preciso muito mais que uma simples cadeia de bits. É necessária uma série de atributos que o tornam significativos tanto para os sistemas de computadores quanto para o homem.

De acordo com a National Information Standards Organization (NISO, 2004, p.12), objetos digitais “são conceitualmente equivalentes a itens que podem ser encontrados no acervo de bibliotecas, na coleção de museus, e em fundos arquivísticos”. Isso significa dizer que, assim como um documento bibliográfico - que para ser gerenciado no ambiente de um sistema de biblioteca precisa ser identificado e descrito através de suas características intrínsecas - um objeto digital deve conter informações contextuais e estruturais que permitam sua identificação, recuperação, preservação e, sobretudo, que seu conteúdo e estrutura possam ser corretamente interpretados e experimentados pelo público-alvo. Essa descrição é feita por meio de informação de representação e, segundo definição do modelo de referência do Open Archival Information System - OAIS, é esta informação de representação (ou seus metadados) que dão ao objeto digital, enquanto cadeia de bits, um status de objeto digital informacional. (CCSDS, 2002)

Kallinikos, Aaltonen e Marton (2010) apresentam quatro atributos do objeto digital: 1) os objetos digitais são editáveis, isto é, são passíveis de serem modificados continuamente e de muitas formas em contraste com os objetos físicos que não mudam; 2) os objetos digitais são interativos, ou seja, podem ter incorporados às suas funções diversos arranjos ou serviços; 3) objetos digitais podem ser acessados e modificados por meio de outros objetos digitais, por exemplo, uma foto que é editada por um aplicativo de edição de fotos; 4) objetos digitais são distribuídos, isto significa dizer que raramente esses dados estão dentro de uma única Fonte ou instituição.

De acordo com o *Data Curation Center*⁴⁹, os objetos digitais informacionais podem ser classificados quanto à sua composição, em: objetos digitais simples e objetos digitais complexos. Os primeiros são aqueles expressos por um único arquivo ou formato, como por exemplo, um documento pdf ou uma imagem jpg. Os segundos são documentos que incluem agregação de formatos de arquivos, assim como diversos tipos de metadados, por exemplo, um banco de dados ou uma *homepage* - que possuem metadados de preservação, de direitos, além dos metadados estruturais.

⁴⁹ Disponível em: <<http://www.dcc.ac.uk/resources/curation-lifecycle-model>>. Acesso em: 17 abr. 2013.

Quanto maior a complexidade do objeto digital informacional, maior é o número de metadados que ele necessita para ser bem representado e consequentemente recuperado, acessado, interpretado, interoperável e processável.

O modelo de dados subjacente ao dicionário PREMIS (OCLC/RLG, 2005) – sigla para *Preservation Metadata: Implementation Strategies* - nos oferece uma conceituação importante para a compreensão das camadas estruturais de um objeto digital complexo. No domínio do PREMIS, um Objeto (ou Objeto Digital) é uma unidade discreta de informação em formato digital. Por outro lado, o conjunto de conteúdos que é considerado como uma unidade intelectual para propósito de gestão e descrição é chamada de Entidade Intelectual. Um livro em particular, um mapa, uma fotografia ou uma base de dados, são exemplos de entidades intelectuais. Este conceito corresponde à ideia de objeto informacional como tratado nesta pesquisa.

A entidade Objeto definida pelo PREMIS tem três subtipos: arquivo, sequência de bits e representação. Um **arquivo** é uma sequência ordenada e identificada de bits reconhecida por um sistema operacional; é similar à ideia comum de arquivo computacional; um arquivo pode ser lido, escrito, copiado e tem nome e formato. Uma **sequência de bits** – em inglês, *bitstream* – é um conjunto de bits embutido num arquivo. Um exemplo de um arquivo com sequências de bits embutidas é um arquivo TIFF contendo duas imagens.

Para o PREMIS, **representação** são as diversas formas de apresentação de uma entidade intelectual que possa ser utilizada por um ser humano⁵⁰. (OCLC/RLG, 2005). Assim, definem representação como um conjunto de arquivos, incluindo metadados estruturais necessários para a apresentação (ou renderização, para usar o neologismo já comum derivado do inglês *rederization*) de uma entidade intelectual. O próprio PREMIS (OCLC/RLG, 2005, p.7) enumera alguns exemplos: um artigo de periódico pode estar completo num único arquivo PDF, constituindo uma representação; outro artigo pode ser representado por uma imagem TIFF para cada uma das suas 12 páginas e outro exemplo ainda é um arquivo XML contendo os metadados estruturais que mostram a ordem das páginas, sendo a representação de outra representação.

⁵⁰ O PREMIS preferiu o termo “representação” para evitar o termo “manifestação” usado pelo FRBR, entretanto há uma correspondência próxima entre os conceitos que os termos expressam (OCLC/RLG, 2005, p.8).

Um repositório digital – por exemplo, uma biblioteca digital ou um repositório voltado para a preservação - pode utilizar mais de uma representação para a mesma entidade intelectual. O essencial é que para uma entidade intelectual ser apresentada e usada por seus usuários, todos os arquivos que constituem pelo menos uma versão desta entidade, devem ser identificados, armazenados e mantidos de forma que possam ser reunidos e apresentados para um usuário em algum ponto do espaço e do tempo, como é esperado que seja uma publicação ampliada, objeto desta tese que será aprofundado mais adiante.

Compreendido que, no contexto desse estudo, dados de pesquisa se configuram como objetos digitais e que esses dados precisam ser gerenciados para que pesquisadores possam recuperá-los, acessá-los e reusá-los é importante destacar que mesmo objetos digitais simples possuem uma quantidade razoável de metadados, gerados automaticamente ou adicionados manualmente, o que revela que a representação via metadados é o caminho para a gestão do dado digital de pesquisa e um componente importante da publicação ampliada.

Para comprovar essa última afirmação, em 1994, antes da explosão do paradigma da *eScience*, Hachen e colegas já mencionavam que “um importante aspecto da gestão de dados científicos é a gestão de metadados”. Expandindo a definição mais difundida de metadados, i.e, dados sobre dados, esses autores formulam uma boa explicação do que vem a ser metadados para os cientistas.

metadados é a informação requerida para identificar dados de interesse baseada no conteúdo, validade, Fontes, pré-processamento ou outras propriedades selecionadas. Nas bases de dados científicas, onde muitos tipos de dados são armazenados, os metadados associados devem ser preservados e acessíveis para que os dados possam ser processados com significado no futuro. (HACHEN et al., 1994, p.1)

Os autores afirmam também que “metadado é um conceito relativo” isso significa que os metadados podem variar de acordo com o objeto a ser representado e também de acordo com a finalidade e/ou interesse de quem utilizará o dado. A questão que se coloca neste momento é “como gerenciar esses dados?”, é o que será abordado na seção a seguir:

4.4 METADADOS PARA GESTÃO DE OBJETOS DIGITAIS

A Informação de Representação é um dispositivo para reconstituição do significado da informação que tem por objetivo converter *bits* em conteúdos mais expressivos aos sentidos, ou seja, em texto, em imagem, em tabela, etc.

Se por um tempo foi suficiente o uso de metadados descritivos para gerir objetos digitais e torná-los recuperáveis, intercambiáveis etc, hoje, a realidade mostra que somente esses padrões não são mais suficientes, pois são inúmeras as questões que devem ser consideradas nos processos de gestão de acervos digitais: “Como identifico para sempre um recurso digital? Qual é o formato de arquivo? Qual é o *software* que criou o objeto? E como o arquivo é aberto? Qual a versão desses programas? Que tipo de *hardware* é preciso? Que equipamentos são necessários? Preciso de folha de estilo? De esquemas? Qual é a relação do objeto com outros objetos digitais? Qual a proveniência do objeto? Quais são os direitos de propriedade intelectual associados ao objeto? Quais os mecanismos que garantem a integridade e autenticidade do objeto? Qual foi a cadeia de custódia do objeto?”

A resposta para essas questões precisam ser registradas na representação do objeto para assegurar preservação e confiabilidade das informações nele contidas. Assim, para responder questões referentes ao ambiente tecnológico, contexto, referência, direitos, proveniência, integridade, autenticidade etc, uma variedade de metadados vem sendo criada em contextos específicos.

No âmbito da Biblioteconomia, o uso de metadados sempre esteve associado à representação descritiva de documentos bibliográficos visando sua recuperação futura. No caso dos objetos digitais a utilização de metadados incorpora também outras perspectivas, pois precisa garantir, além da recuperação, a preservação das informações neles contidas. No artigo “Uma outra face dos metadados”, Sayão apresenta este fato, afirmando, entre outras coisas, que os “desafios do mundo digital, foram redesenhando a ideia puramente descritiva de metadados, criando expansões para o seu conceito com o intuito de abrigar novos propósitos e funções” (SAYÃO, 2010, p.4.).

Sob esta perspectiva, a NISO (National Information Standard Organization) propõe uma definição para metadados que engloba essas novas funções: “Metadado é a informação estruturada que descreve, explica, localiza, ou possibilita que um recurso informacional seja fácil de recuperar, usar ou gerenciar”. (NISO, 2004, p.1)

Os metadados são utilizados não apenas para descrever o conteúdo do documento, mas também para atribuir identificação persistente, recompor a estrutura do documento, explicitar as relações com outros objetos, com versões, manifestações/expressões, evidenciar declarações de direitos, por exemplo: acesso e *copyright*; informar as dependências técnicas, instruir a preservação, registrar proveniência, isto é, a história do objeto, garantir a integridade e autenticidade do objeto, ou seja, se foram alterados de forma não documentada.

Para compreender melhor essas diversas funções, ou seja, gerenciar, preservar, recuperar, acessar e manter a capacidade de ser interpretado são necessários vários tipos de metadados, que Kenney et al. (2001) classificam os metadados em: metadados descritivos ou intelectuais, metadados estruturais e metadados administrativos.

Os **metadados descritivos** são aqueles usados para a descrição do conteúdo dos recursos informacionais, permitindo a indexação e a catalogação, como por exemplo, o MARC, Dublin Core, os *metatags*, entre outros. A descrição pressupõe também a identificação persistente que deve ser registrada por esquemas padronizados como o DOI - *Digital Object Identifier*⁵¹ ou o *Handle System*⁵².

Os **metadados estruturais** são aqueles que fornecem informações sobre a estrutura de armazenamento das Fontes de dados, isto é, são aqueles que dão forma ao documento, por meio de informações para juntar as partes, dar sequência e estrutura. Este tipo de metadado documenta o relacionamento entre objetos: artigos, itens, volumes, páginas e capítulos de livros. Esses metadados, essenciais para os objetos compostos, estão geralmente codificados em documentos XML.

Os **metadados administrativos** são aqueles que controlam o acesso a cada um dos recursos informacionais identificados registrando informações referentes à obtenção de informação, direitos de reprodução, critérios de seleção para digitalização e dependências técnicas.

⁵¹ Disponível em: <<http://www.doi.org/index.html>>. Acesso em 20 maio 2013.

⁵² Disponível em: <<http://www.handle.net/>>. Acesso em 20 maio 2013.

Além disso, a literatura evidencia também a existência dos metadados de preservação, que estão espalhados pelas categorias anteriores (descritivos, estruturais e administrativos). Pode-se citar como exemplo o modelo de metadados do PREMIS cuja base conceitual é o Open Archival Information System - OAIS (CCSDS, 2002). O modelo PREMIS foi criado a partir de um consenso extraído das experiências acumuladas de muitas e variadas instituições – museus, bibliotecas, arquivos, governo e iniciativa privada – e da *expertise* dos principais profissionais provenientes da Austrália, Nova Zelândia, Estados Unidos, Grã-Bretanha, Holanda e Alemanha. Assim, foi formado um grupo de trabalho cuja experiência convergiu para uma fundamentação prática da preservação digital, para o intercâmbio de informações de preservação e para a interoperabilidade entre repositórios. De acordo com Sayão (2010), essa experiência pode ser revelada por meio dos dois objetivos do PREMIS:

- 1) Definir um conjunto essencial de elementos de metadados de preservação que seja implementável e de larga aplicação, devendo esse núcleo ser apoiado por um dicionário de dados, desenvolvido para oferecer diretrizes e recomendações para o preenchimento e para a gestão dos elementos de metadados.
- 2) Identificar e avaliar estratégias alternativas para codificar, armazenar, gerenciar e intercambiar metadados de preservação, especialmente os essenciais, no contexto de um sistema de repositório digital.

Os primeiros resultados do trabalho do Grupo foram sumarizados no relatório *“Implementing preservation repositories for digital materials: current practice and emerging trends in the cultural heritage”* (OCLC/RLG, 2004), cuja principal contribuição foi a elaboração de um guia que definiu um conjunto de metadados necessários para apoiar a preservação digital de longo prazo. Nesse documento, foi dada atenção especial aos metadados que descrevessem a proveniência de um objeto e seus relacionamentos com outros objetos pertencentes a um mesmo repositório.

Este relatório gerou o Dicionário de Dados PREMIS, que se concentra em um núcleo específico de metadados próprios para preservação, deixando de lado outros tipos de metadados, como por exemplo, os descritivos. Esse núcleo é chamado de metadados essenciais.

Além disso, outra contribuição importante do grupo foi a criação de um conjunto de esquemas XML para apoiar o uso do Dicionário de Dados por instituições que gerenciam e intercambiam metadados de preservação e que estejam em conformidade com a proposta do PREMIS (OCLC/RLG, 2005).

Uma questão interessante é que o Grupo de Trabalho PREMIS estabeleceu que o glossário não consideraria elementos de metadados e sim unidades semânticas. Conforme Sayão

Essa diferença é sutil, porém importante: uma unidade semântica é uma peça de informação ou de conhecimento, enquanto um elemento de metadados é uma forma definida de representar essa informação em um registro de metadados, em um esquema ou numa base de dados. (SAYÃO, 2010, p.24).

Assim como num instrumento de Organização do Conhecimento - como tesauro, taxonomia, ontologia etc, em que uma propriedade pode ser parte de um conceito ou um novo conceito ligado a outro mais abrangente - cada unidade semântica do PREMIS pode ser vista como uma propriedade de uma entidade. Por exemplo, o tamanho em *bytes* é uma unidade semântica que é uma propriedade de uma entidade objeto. Além disso, unidades semânticas têm valores: para um objeto específico o valor do tamanho pode ser "843200004", por exemplo.

As unidades semânticas do Dicionário de Dados PREMIS se referem a objetos, eventos, agentes e direitos. A quinta entidade do modelo, isto é, a entidade intelectual, é considerada fora do escopo, pois é bem servida por metadados descritivos. Essas entidades têm papéis associados à preservação digital e são definidas da seguinte forma (OCLC/RLG, 2005):

Entidade intelectual – conjunto coerente de conteúdos que é reconhecido como uma unidade, por exemplo, livros, artigos, bases de dados;

Objeto – unidade discreta de informação em forma digital, constituindo o que realmente é armazenado e gerenciado pelo repositório, por exemplo, um arquivo PDF. As unidades semânticas para objetos podem ser especificadas em três níveis: cadeia de bits (*bitstream*), arquivos (*files*) e o conjunto de arquivos que completam a apresentação de uma Entidade Intelectual, ou seja, a representação (*representation*);

Evento –ações que envolvem ou afetam os objetos no repositório, por exemplo, uma ação de migração de mídias ou formatos;

Agente – pessoa, organização ou *software* que desempenha papéis associado com um Evento ou declarações de Direitos;

Direitos – direitos e permissões vinculadas ao objeto, por exemplo, permissão para cópia.

O modelo inclui ainda, para cada entrada de unidade semântica, um lugar para notas sobre como criar ou utilizar a unidade semântica. Em alguns casos, o grupo acrescentou informações adicionais, tais como a razão para a definição de uma unidade semântica, ou questões que surgiram nas deliberações do grupo, como, por exemplo, o fato de que cada componente semântico deve herdar a aplicabilidade da unidade semântica que o contém.

Cada entrada no Dicionário de Dados oferece os seguintes atributos de uma unidade semântica:

- **Nome da unidade semântica:** nomes são concebidos para serem descritivos e únicos dentro do Dicionário de Dados. São esses nomes que auxiliarão a troca de metadados à interoperabilidade entre os repositórios.
- **Componentes semânticos:** cada componente tem sua própria entrada no final do Dicionário de Dados. A unidade semântica que tem componentes semânticos não tem qualquer valor próprio. Apenas as unidades semânticas no nível mais baixo têm valores.
- **Definição:** registra o significado da unidade semântica.
- **Justificativa:** registra o porquê de a unidade semântica ser necessária, se isto não é autoevidente a partir da definição.
- **Restrição de dados:** registra como o valor da unidade semântica deve ser codificado.
- **Categoria de Objeto:** registra se a unidade se aplica a uma representação, arquivo ou objeto *bitstream*.
- **Aplicabilidade:** registra se um objeto é aplicável a determinado escopo ou categoria
- **Exemplos:** registra exemplos de valores da unidade semântica. Pretende ser ilustrativo.

- **Repetibilidade:** uma unidade semântica pode ser designada como "Repetitivo" e pode assumir vários valores.
- **Obrigação:** um valor para a unidade semântica pode ser obrigatório (se aplicável) ou opcional.
- **Criação / Manutenção notas:** registra notas sobre como os valores para a unidade semântica podem ser obtidos e/ou atualizados.
- **Notas de Uso:** informações sobre o uso pretendido para a unidade semântica, ou esclarecimento da definição.

A diferença do PREMIS para os esquemas de metadados mais conhecidos é que o esquema PREMIS não especifica apenas como os metadados devem ser representados em um sistema, mas também define o que o sistema precisa entender e o que ele deve ser capaz de exportar para outros sistemas (CAPLAN, 2009). Isso é muito interessante, pois permite o diálogo com outros esquemas de metadados.

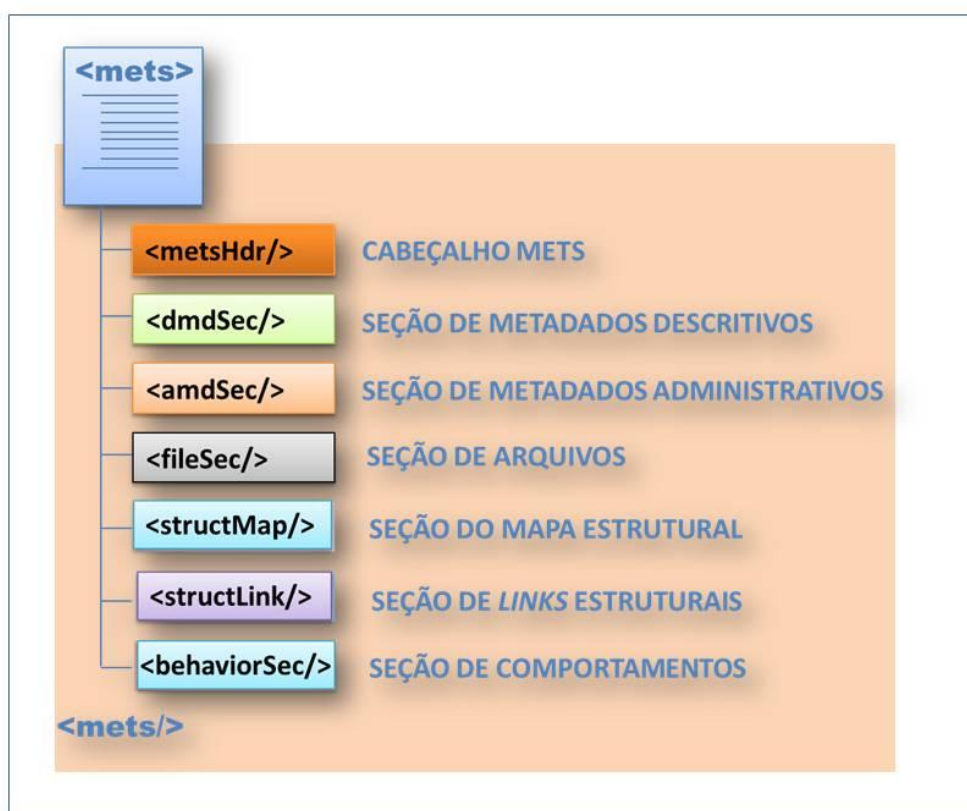
O Dicionário de dados PREMIS tem objetivos muito semelhantes aos instrumentos de organização do conhecimento mais conhecidos, isto é, visa melhorar a recuperação e a troca de informações através da padronização semântica de seus conceitos e relações. Além disto, sua forma de esquematizar as relações entre as unidades semânticas revela a necessidade de estruturas que sistematizem a relação existente entre os grupos de metadados dos esquemas existentes. Dito de outra forma, não basta criar os metadados, mas é preciso também criar uma estrutura consistente que evidencie as relações entre eles.

Fica claro que um objeto digital inserido num ambiente de um sistema de informação para ser gerenciado, recuperado, preservado e ter sua estrutura recomposta, precisa de um número significativo de metadados. Este fato coloca uma questão crítica: como esse conjunto de metadados pode estar organizado e vinculado ao objeto correspondente?

A solução passa pela elaboração de “invólucros” conceituais que permitam o registro, ordenamento e a codificação de todos os metadados de um objeto digital, especialmente os dos chamados objetos digitais complexos. Nessa direção, Digital Library Federation - DLF desenvolveu a norma conhecida como METS – sigla para Metadata Encoding Transmission Standard.

O METS é um esquema XML que oferece um mecanismo flexível para codificar todos os tipos de metadados associados a um objeto digital – descritivos, administrativos, estruturais - e para exprimir as ligações complexas entre esses metadados no ambiente de um repositório. Por conseguinte, o METS estabelece um padrão útil para a gestão de objetos digitais no âmbito de um repositório e o intercâmbio deles entre repositórios (ou entre repositórios e seus usuários); além do mais, oferece a possibilidade de associar um objeto digital com comportamentos ou serviços. O METS considera que os objetos que compõem um objeto complexo devam ser modelados como estruturas de árvore, assim como um livro com subcapítulos. Cada nó desta árvore deve ser associado a um conjunto de metadados expressos em XML, conforme ilustrado na Figura 7.

Figura 7: Estrutura de um documento METS



Fonte: A autora com base em METS

O esquema METS se organiza em sete seções principais:

1) Cabeçalho METS - contém metadados que descrevem o documento METS em si, incluindo informação como o criador, editor, etc.

2) Metadados Descritivos - pode apontar para outros metadados externos ao documento METS (ex. um registo MARC num OPAC ou um registo EAD mantido num servidor Web), ou conter metadados descritivos embutidos, ou ambos.

3) Metadados Administrativos - oferecem informações sobre como os arquivos foram criados e armazenados, sobre direitos de propriedade intelectual, informações sobre o objeto original a partir do qual o objeto digital foi derivado, e informação sobre a proveniência dos arquivos que compõem o objeto digital (i.e., relações de arquivos originais/derivados, e informação de migração/transformação). Assim como os metadados descritivos, os metadados administrativos podem ser tanto externos ao documento METS, ou codificados internamente.

4) Seção de Arquivos - lista todos os arquivos que contêm as versões eletrônicas do objeto digital. Elementos <file> podem ser agrupados em elementos <fileGrp>, para permitir a subdivisão de arquivos por versão do objeto.

5) Mapa Estrutural - é o coração do documento METS. Ele esboça uma estrutura hierárquica para o objeto da biblioteca digital e liga os elementos dessa estrutura aos arquivos com conteúdos e metadados referentes a cada elemento, como numa taxonomia.

6) Ligações Estruturais - permite aos criadores METS registrar a existência de hiperlinks entre nós na hierarquia esboçada no Mapa Estrutural. Esta seção tem um valor particular na utilização do METS para descrever sites.

7) Comportamento - pode ser usada para associar comportamentos executáveis (serviços) com o conteúdo no objeto METS. Cada comportamento numa seção tem um elemento de definição de interface que representa uma definição abstrata do conjunto de comportamentos representado por uma seção de serviço particular.

Entre as seções do METS, mais uma vez, o que chama a atenção é a necessidade de

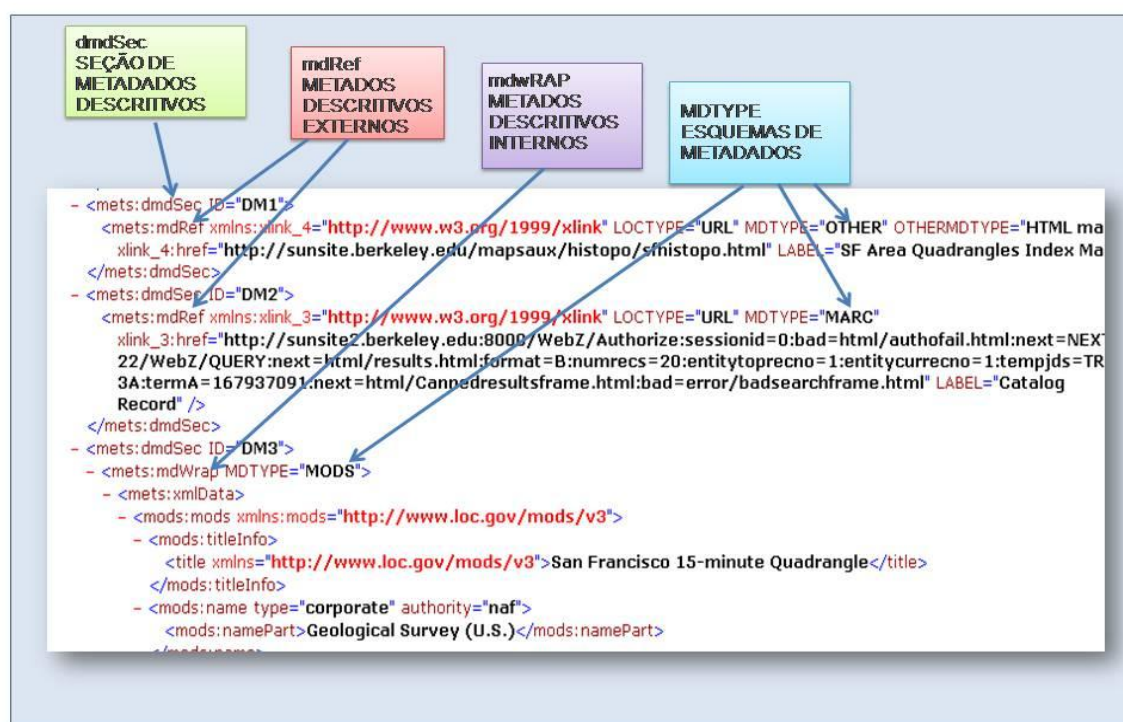
uma estrutura sistemática para organização desse amplo conjunto de metadados, que aparentemente é solucionado pelas seções Mapa Estrutural e Ligações Estruturais.

A seção do mapa estrutural do esquema METS define uma estrutura hierárquica que pode ser apresentada aos usuários do objeto da biblioteca digital para lhes permitir navegar nele. O elemento <structMap> codifica essa hierarquia como série de elementos <div> encaixados. Cada <div> contém informação em atributos que especifica que tipo de divisão é, e também pode conter múltiplos apontadores METS (<mptr>) e elementos apontadores de arquivos (<fptr>) para identificar o conteúdo correspondente a esse <div>. Apontadores METS especificam outros documentos METS como contendo a informação relevante para o <div> que os contém. Isto pode ser útil quando se codifica grandes coleções de material (ex. todos os números de uma revista científica) para manter o tamanho de cada arquivo METS relativamente pequeno.

Já a Seção de ligações estruturais do formato METS é a mais simples de todas as principais seções METS, em termos de forma, contendo apenas um único elemento, <smLink> (embora esse elemento possa ser repetido). Esta seção visa registrar a existência de *hyperlinks* entre itens dentro do mapa estrutural, geralmente elementos <div>. Esta é uma funcionalidade útil caso se pretenda utilizar o METS para arquivar sites, e se pretenda manter um registro da estrutura do hipertexto dos sites separadamente dos arquivos HTML do site em si. Enquanto o mapa estrutural sistematiza os vínculos entre os objetos de forma hierárquica, a seção de ligações se utiliza de outros tipos de relações para sistematizar vínculos não hierárquicos entre os objetos.

A Figura 8 a seguir apresenta um fragmento de documento descrito em METS, mostrando como estão codificados os metadados descritivos internos e as referências aos metadados externos.

Figura 8: Codificação segundo o METS de metadados descritivos



Fonte: A autora

Uma das aplicações mais importantes do METS é a utilização como a unidade de informação flui num repositório destinado ao arquivamento confiável definido pelo modelo conceitual ISO/OAIS. O conteúdo a ser preservado juntamente com o conjunto de metadados, que garantem a preservação da estrutura e da semântica do objeto, são estruturados em pacotes METS – pacotes de submissão, pacotes de arquivamento e pacotes de disseminação –, formando unidades informacionais que são preservadas, recuperadas, disseminadas e intercambiadas.

Outro modelo importante de metadados é o *Open Archival Information System* – OAIS, que serviu de referencial teórico para a criação do PREMIS. O OAIS define os tipos de metadados necessários para a preservação como parte de uma taxonomia de classes de objetos de informação (CCSDS, 2002). Essa taxonomia detalha os níveis de informação de descrição e preservação de cada objeto digital e serve para explicitar de forma lógica as informações que serão utilizadas para preservação e recuperação. Esse modelo também ressalta a necessidade de uma estrutura sistemática para organização desses metadados, no entanto, como ele é apenas um modelo conceitual, a execução prática ficou para os modelos empíricos oriundos dele, como foi visto anteriormente no PREMIS.

Fica compreendido então que uma publicação ampliada é formada por objetos digitais, metadados e ainda por relações conceituais que conectam esses objetos formando um agregado de informações. Assim, a seção a seguir apresentará esse último elemento também essencial para formação deste novo modelo de publicação acadêmica.

4.5 RELAÇÕES CONCEITUAIS: OUTRO COMPONENTE ESSENCIAL PARA FORMAÇÃO DE PUBLICAÇÃO AMPLIADA

Mey (1995, p.6) afirma que "a riqueza da catalogação repousa nos relacionamentos entre os itens estabelecidos de forma a criar alternativas de escolha para os usuários". Catarino e Souza explicam essa afirmação, no contexto da Web, de uma forma bem didática dizendo que

se um usuário busca determinada obra, ele terá a oportunidade de optar por diferentes versões ou suportes, ou seja, as diferentes manifestações do mesmo item, como livro, dvd, CD-ROM, documento *online*, bem como em diferentes edições, traduções, línguas, e isso certamente ampliará o universo de pesquisa e de escolha desse usuário. Os relacionamentos entre os itens também lhe permitem localizar itens por ele desconhecidos, uma vez que a catalogação propicia a reunião de todos os itens sobre um determinado assunto, aumentando e melhorando a busca. (CATARINO; SOUZA, 2012, p.85)

Para a formação de publicações ampliadas, relações conceituais bem estabelecidas se tornam essenciais para compor, descrever, gerenciar, preservar, prover serviços, possibilitar recuperação precisa da publicação ou de um de seus componentes, tornar o documento interoperável em todos os níveis, entre outras funções.

Relações ou Relacionamentos, como também são chamados na literatura, podem ser definidos amplamente, como “uma associação entre duas ou mais entidades ou entre duas ou mais classes de entidades” (GREEN, 2001, p.3). No domínio da organização do conhecimento, os relacionamentos são mecanismos usados para indicar a existência de conceitos que guardem alguma semelhança contextual, propiciando uma recuperação mais precisa. Por um lado, as relações permitem que o usuário tome ciência da existência no sistema de outros documentos interessantes ao seu propósito. Por outro lado, as relações permitem sistemas mais eficazes, garantindo consistência na “adoção de algoritmos baseados em regras de associação” (KIETZ et al., 2000).

Sheth, Arpinar e Kashyap (2003) afirmam que os relacionamentos são fundamentais para a proposta da *Web* semântica, “pois associam os significados às palavras, aos termos e às entidades”, e ainda podem ser considerados “a ‘chave’ para novas percepções e ideias”. Além disso, alegam que “a descoberta do conhecimento é antes de tudo a descoberta de novos relacionamentos”, evidenciando assim, a importância das relações em instrumentos como ontologias.

De acordo com Green (2001, p.3), para explicitar um relacionamento, deve-se primeiramente designar todas as partes vinculadas ao relacionamento e, em seguida, especificar a natureza do relacionamento. Além disso, os participantes desta relação devem ser revelados ou rotulados de modo a associar a cada participante o papel que o mesmo desempenha. Soma-se a isso a especificidade da natureza de uma relação que deve estar acompanhada pelo próprio conjunto de propriedades a serem consideradas por ela. Neste sentido, a relação também se torna um conceito.

Um relacionamento, porém, possui muitas propriedades a serem consideradas no seu estabelecimento, a saber: aridade, simetria, cardinalidade, transitividade e semântica. Cada uma dessas propriedades gera uma cadeia de tipos de relações.

Primeiramente, de acordo com sua natureza, as relações podem ser classificadas em abstratas ou concretas. As **relações abstratas** expressam associação entre entidades de classes (ex: Pessoa <born in> Lugar) e as **relações concretas** expressam associações entre entidades específicas (ex: Michelangelo <born in> Caprese). As relações abstratas são usadas na modelagem de dados e as relações concretas como instâncias.

As relações também podem ser classificadas de acordo com o número de participantes envolvidos pelo relacionamento, isto é, de acordo com sua aridade. Se uma relação envolve dois elementos, ela é chamada de **relação binária**, se envolve três, ela é chamada de **relação ternária**. Uma relação que envolve muitos participantes é chamada de **relação n-ária**. Ao modelar qualquer contexto pode-se verificar que, em geral, todas as relações são n-árias, pois todos os elementos podem ter algum tipo de associação. Uma entidade está associada a uma segunda, que está associada a uma terceira, que está associada a uma quarta e assim sucessivamente. De acordo com Green

embora a conversão das relações n-árias em conjuntos de relações binárias possa ser visto como confuso conceitualmente, a regularidade das relações binárias, muitas vezes a torna uma opção atraente. Esta atratividade também é um perigo, uma vez que cria a tentação de restringir relações com relacionamentos binários, sem dar o passo necessário de criar uma classe. (GREEN, 2001, p.4)

As relações são classificadas também de acordo com sua simetria, em **relação assimétrica** ou **relação simétrica**. As relações são **assimétricas** quando a relação de A com B, não é a mesma relação de B com A, por exemplo: João <loves> Maria, mas não é necessariamente verdadeiro que Maria <loves> João. No entanto, uma relação é **simétrica** quando esta correspondência é verdadeira, por exemplo, Ana <is a cousin of> Maria. Se Ana é prima de Maria, Maria também é prima de Ana.

Outra forma de classificar as relações é quanto à sua **cardinalidade**. Esta propriedade refere-se ao número de membros participantes de cada classe de entidade e a possibilidade de ser associado através da relação e vice-versa. No caso das relações binárias, as relações podem ser *one-to-one* (um-para-um), *one-to-many* (um-para-muitos) e *many-to-many* (muitos-para-muitos). Um exemplo de **relação um-para-um** é a relação ‘casamento’ em uma sociedade monogâmica. (Um) marido <is married to> (uma) esposa e vice-versa. Já em uma relação de **um-para-muitos**, pode-se citar como exemplo a relação de paternidade, por exemplo, (um) homem <is a biological father> de filhos (muitos), assim, um homem pode ter muitos filhos e muitos filhos podem ter um mesmo pai. E em uma relação de **muitos-para-muitos** pode se citar a relação ‘pais’, Person <is parent of> person, significando que muitas pessoas podem ser pais de muitas outras pessoas e muitas pessoas podem ter pais.

As relações ainda podem ser classificadas de acordo com a sua **transitividade**. Uma relação é **transitiva** se ela transmite sua relação para a entidade associada subsequentemente, sendo uma relação determinada por seus aspectos lógicos, podendo ser explicada da seguinte maneira: se A <está ligada de certa maneira para> B e B <está relacionada na mesma maneira> a C, também é verdade que A <está relacionada na mesma maneira> com C. A herança hierárquica é um exemplo de relação transitiva, assim, uma tabela de classificação <is_a> linguagem documentária, a CDD <is_a> tabela de classificação, então a CDD <is_a> linguagem documentária.

Outra propriedade identificada por Green (2001, p.6) está relacionada ao plano da língua e ao posicionamento de uma relação dentro de uma estrutura classificatória. Assim Green explica que os **relacionamentos fechados** correspondem aqueles expressos linguisticamente como paradigmáticos, por exemplo, hyponímia, meronímia, sinonímia e homonímia (hierárquicos e de equivalência em um tesouro) e os **relacionamentos abertos** correspondem aqueles expressos linguisticamente por meio de relações sintagmáticas (associativos). Como pode se perceber, cada propriedade identificada por Green promove uma classificação para os tipos de relações. Essas propriedades e as relações geradas podem ser visualizadas na Figura 9 a seguir:

Figura 9: Propriedades básicas das relações

Quanto à sua posição estrutural	• Relação Fechada • Relação Aberta
Quanto à sua natureza	• Relação Abstrata • Relação Concreta
Quanto à sua aridade	• Relação binária • Relação Ternária • Relação n-ária
Quanto à sua simetria	• Relação Simétrica • Relação Assimétrica
Quanto à sua cardinalidade	• Relação um para um • Relação um para muitos • Relação muitos para muitos
Quanto à sua transitividade	• Relação Transitiva • Relação Não-transitiva

Fonte: A autora baseado em Green(2001)

Green (2001, p.7) continua sua investigação sobre relacionamentos a partir da identificação dos participantes do relacionamento, apontando as principais entidades que este campo envolve - por exemplo, pessoas (autor, tradutor, editor, etc); unidades bibliográficas (unidades intelectuais, como o texto e unidades físicas, como livros, periódicos, páginas *Web*, etc); assuntos, conceitos, palavras, conhecimento; usuários e suas necessidades. Em seguida apresenta uma lista não-exaustiva com 12 exemplos de relações.

- 1) Pessoa <*produces*> bibliographic unit
- 2) Bibliographic unit <*is part of*> Bibliographic unit
- 3) Bibliographic unit <*describes*> Bibliographic unit
- 4) Bibliographic unit <*cites*> Bibliographic unit
- 5) Bibliographic unit <*links to*> Bibliographic unit
- 6) Text <*has*> Subject
- 7) Concept <*in*>Text
- 8) Word <*conveys*> Concept
- 9) Knowledge <*conveyed by*> texto
- 10) User<*has*> Knowledge
- 11) User<*has*> Information need
- 12) Knowledge <*addresses*> information need

Green (2001) propõe então uma nova classificação para os relacionamentos na Organização do Conhecimento, dividindo-os em quatro áreas: 1) **Relacionamentos entre unidades de registro do conhecimento ou bibliográficos** que são baseados na descrição bibliográfica destas unidades, como as relações bibliográficas apresentadas pelo Functional Requirements for Bibliographic Records - FRBR, por exemplo; 2) **Relacionamentos intratextuais e intertextuais**, incluindo relacionamentos baseado na estrutura do texto, relacionamentos de citação e *links* de hipertexto. 3) **Relacionamentos de assunto ou conceituais**, como apresentados em tesouros ou outros instrumentos de classificação; 4) **Relacionamentos de relevância**, isto é, relacionamento entre o usuário e sua necessidade de informação.

As relações que interessam ao contexto das publicações ampliadas são as três primeiras: Relacionamento entre unidades de Registro do conhecimento, Relacionamentos intratextuais e intertextuais e Relacionamento de assunto. Neste trabalho, optou-se por não trabalhar os Relacionamentos de relevância, porque o conceito de relevância envolve uma série de questões que extrapolam o objetivo desta tese. Para exemplificar esta complexidade, não seria possível discutir esse conceito sem considerar estudo de usuário ou outros assuntos correlatos.

Para fins de síntese e melhor compreensão, a seguir tem-se uma representação sistemática da classificação sugerida por Green, na qual a autora incorpora as relações sugeridas pelo FRBR (Figura 10). Após essa representação, seguirá um detalhamento dos relacionamentos considerados relevantes para o modelo proposto por esta tese.

Figura 10: Classificação dos tipos de relações



Fonte: A autora baseado em Green (2001)

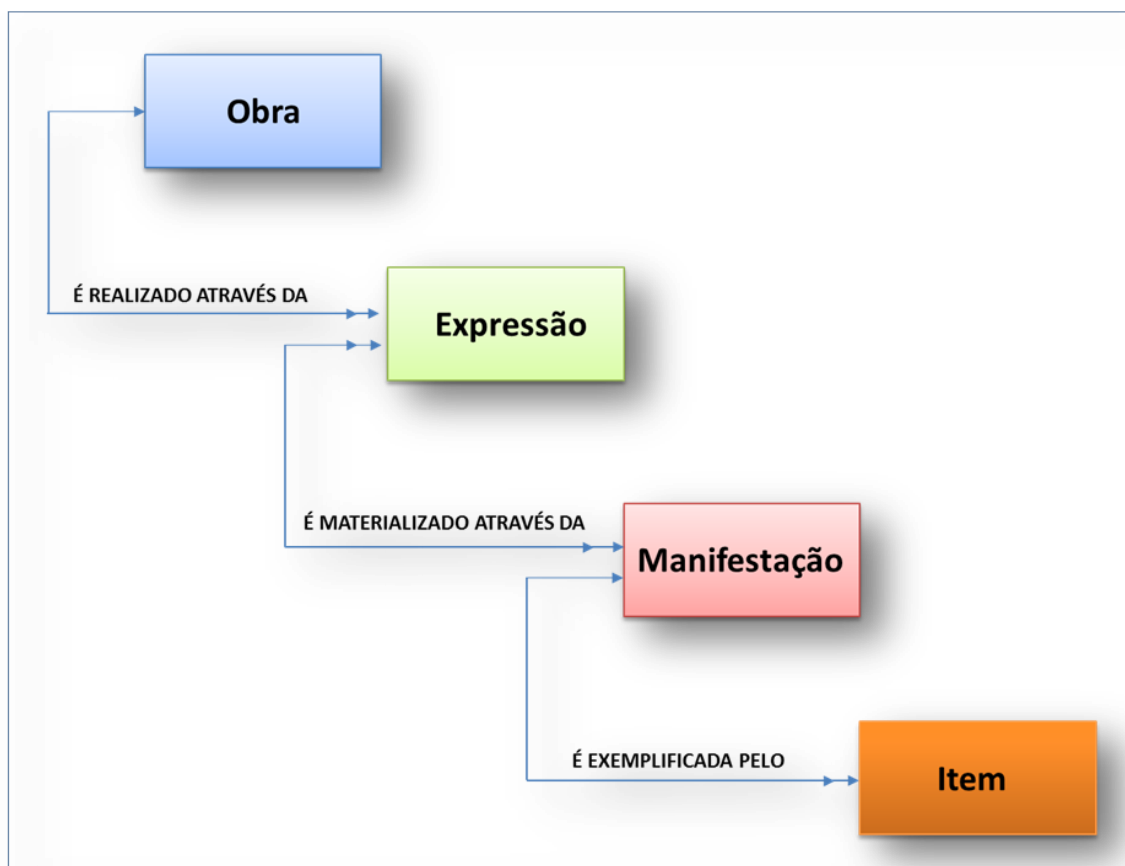
1) Relacionamentos Bibliográficos

Relacionamentos bibliográficos podem ser amplamente definidos como relações que envolvem a catalogação descritiva de unidades bibliográficas, considerando as unidades físicas materiais de um lado e as unidades intelectuais de outro. (GREEN, 2001). Estes relacionamentos podem ser considerados no nível abstrato e no nível concreto.

No nível abstrato, duas questões devem ser pensadas primeiramente: Que classes de entidades bibliográficas necessitam ser reconhecidas e como essas classes gerais se relacionam umas com as outras? O modelo FRBR (IFLA, 2009) indica quatro entidades bibliográficas, sendo duas de natureza intelectual (obras e expressões) e duas de natureza material (manifestações e itens). Entre estas entidades, a International Federation of Library Associations and Institutions - IFLA (2009) aponta a existência dos seguintes tipos de relacionamentos que podem ser usados na modelagem e na construção de sistemas bibliográficos, a saber.

Relacionamento primário - são aqueles que estão implícitos entre as entidades bibliográficas (obra, expressão, manifestação e item), conforme mostrados na Figura 11 a seguir.

Figura 11: Relacionamentos primários do FRBR



Fonte: IFLA (2009) tradução nossa

A IFLA, no documento sobre o FRBR, publicado em 2009, explica estas relações da seguinte forma:

As relações descritas no diagrama indicam que a obra pode ser realizada através de uma ou mais do que uma expressão (a seta dupla na linha que liga a obra, portanto, é a expressão). Uma expressão, por outro lado, é materializada através de uma e apenas uma obra (a seta única no sentido inverso da expressão é a linha que liga a obra). Uma expressão pode ser materializada em uma ou mais do que uma manifestação, de igual modo uma manifestação pode materializar uma ou mais do que uma expressão. Uma manifestação, por sua vez, pode ser exemplificada por um ou por mais de um item, mas um item pode exemplificar uma e apenas uma manifestação. (IFLA, 2009, p.13, grifo nosso).

Relacionamento de conteúdo - são aqueles relacionamentos que podem ser visualizados como um *continuum* a partir de uma obra original, incluindo as **relações de equivalência** - aquelas que se dão entre cópias exatas da mesma manifestação de uma obra ou entre um item original e reproduções do mesmo, desde que o conteúdo intelectual e autoria sejam preservadas. Ex: cópias, microfichas, etc; **relações derivadas** - que se dão entre uma obra bibliográfica e a modificação baseada no trabalho. Ex: edições, adaptações, mudança de gênero, paráfrases, traduções etc; e **relações descritivas (ou referencial)** - que se dão entre uma entidade bibliográfica e a descrição, a avaliação ou a revisão desta entidade, tais como entre uma obra e um livro de revisão que o descreva ou ainda edições anotadas ou comentadas, críticas etc.

Relacionamento parte-todo – são aqueles relacionamentos que se dão entre uma entidade bibliográfica e uma parte componente desta entidade, por exemplo, um volume de uma série ou o capítulo de um livro. Esta classe, inclui também as **relações sequenciais ou cronológicas** - que se dão entre uma entidade bibliográfica que continua ou precede a outra, como por exemplo, títulos de uma série, partes de uma monografia ou entre os vários volumes de um periódico – e ainda as **relações de acompanhamento** - que ligam entidades bibliográficas e os materiais que o acompanham. Nas relações de acompanhamento, há casos em que uma entidade é predominante e a outra é subordinada a esta, como no caso de um texto e seus suplementos. Outro caso de relação de acompanhamento também se refere a uma entidade bibliográfica que prover acesso à outra, como no caso dos índices ou dos catálogos de biblioteca. Ou ainda há casos de entidades de status equivalente, mas sem arranjo cronológico específico, por exemplo, as partes de um *kit*.

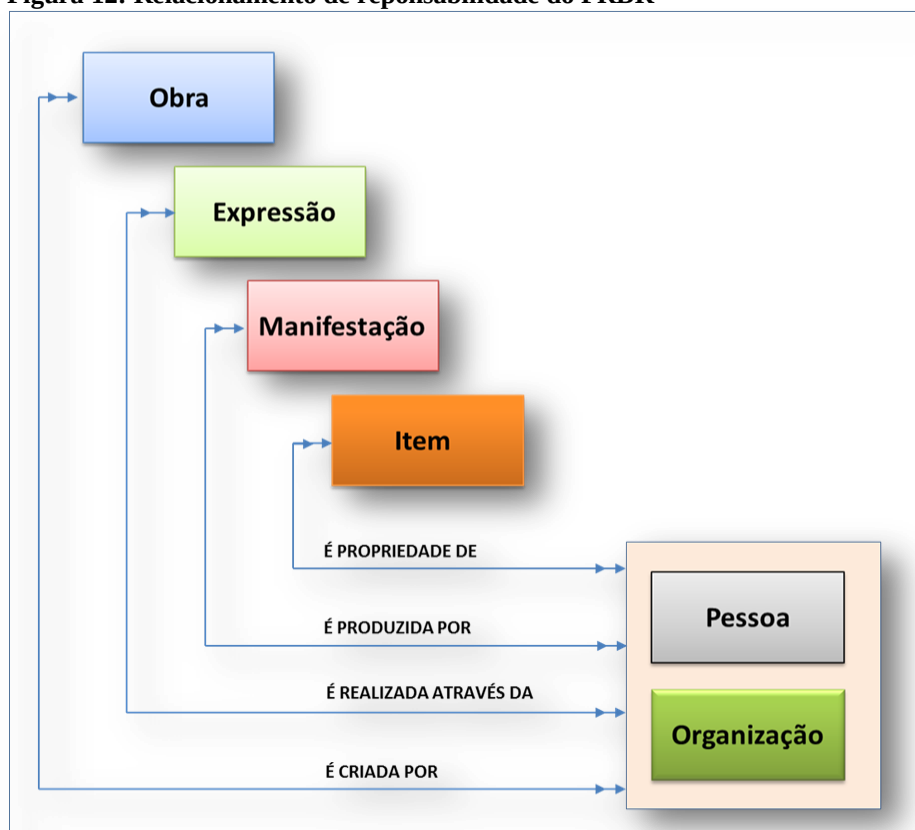
Relacionamento de característica compartilhada - são aqueles relacionamentos que se dão entre entidades bibliográficas que não estão relacionadas de nenhuma forma, mas coincidentemente, possuem o mesmo autor, o mesmo título, o mesmo assunto ou outra característica usada como ponto de acesso em um catálogo.

Relacionamento de responsabilidade – Esse tipo de relacionamento é considerado pelo FRBR em um segundo grupo de entidades que representam os responsáveis pelo conteúdo intelectual ou artístico, a produção física e disseminação, ou a custódia das entidades do primeiro grupo. As entidades do segundo grupo incluem pessoa física e pessoa jurídica (uma organização ou grupo de indivíduos e/ou organizações). A figura 12 a seguir mostra as relações de "responsabilidade" que existem entre as entidades do segundo

grupo e as entidades do primeiro grupo. A IFLA explica essas relações da seguinte forma:

A Figura indica que a obra pode ser criada por uma ou mais pessoa e/ou por uma ou mais Instituições. Por outro lado, uma pessoa ou uma Instituição pode criar uma ou mais obras. Uma expressão pode ser realizada através de uma ou mais de uma pessoa física e/ou jurídica, e uma pessoa física ou jurídica pode realizar uma ou mais de uma expressão. A manifestação pode ser produzida por uma ou mais de uma pessoa física ou jurídica, a pessoa física ou jurídica pode produzir uma ou mais de uma manifestação. Um item pode ser propriedade de um ou mais de uma pessoa e/ou entidade coletiva, uma pessoa física ou jurídica pode ser proprietário de um ou mais de um item. (IFLA, 2009, p.14, grifo nosso)

Figura 12: Relacionamento de responsabilidade do FRBR

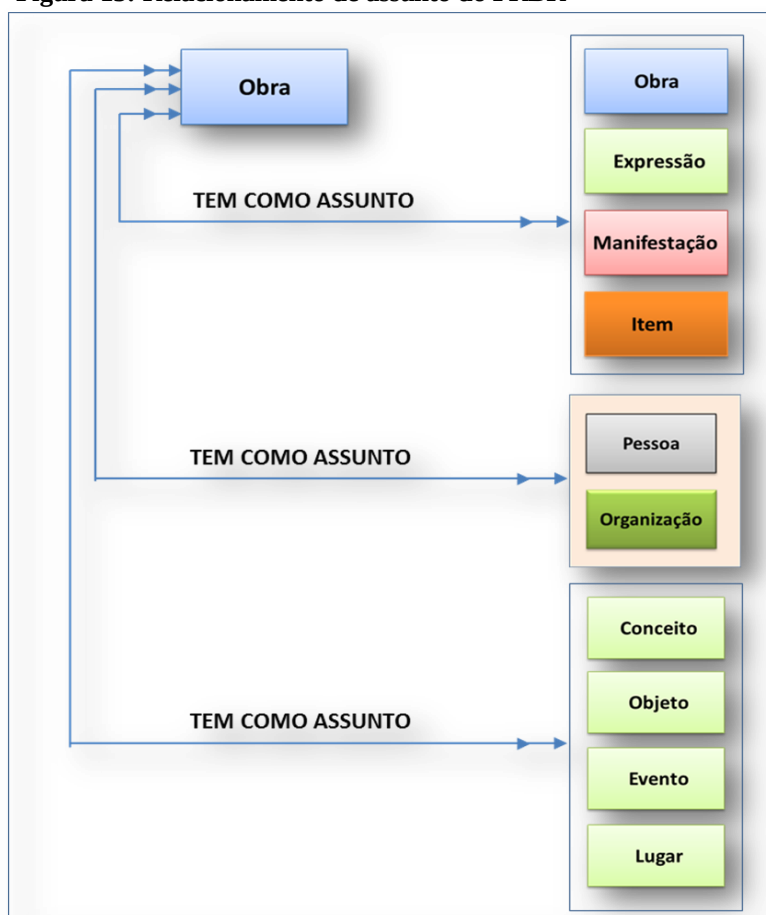


Fonte: IFLA (2009) tradução nossa

Relacionamento de assunto – Este tipo de relação também é considerado pelo FRBR em um grupo à parte, isto é, no terceiro grupo. Este terceiro grupo reúne entidades ligadas à identificação do assunto na obra, como o conceito (que para o FRBR, é uma noção abstrata ou idéia), objeto (uma coisa material), evento (uma ação ou ocorrência) e lugar (a localização). A explicação para as relações entre essas entidades é mostrada na Figura 13 a seguir e explicada pela IFLA da seguinte maneira:

A Figura indica que uma obra pode ter como assunto um ou mais do que um conceito, objeto, evento e/ou local. Por outro lado, um conceito, objeto, evento e/ou o local pode ser assunto de uma ou mais de uma obra. A Figura mostra também as relações de assunto entre a obra e as entidades do primeiro e segundo grupo. A Figura indica que uma obra pode ter como assunto uma ou mais que uma obra, expressão, manifestação, item pessoa, e/ ou Organização. (IFLA, 2009, p.17, grifo nosso).

Figura 13: Relacionamento de assunto do FRBR



Fonte: IFLA (2009) tradução nossa

É importante destacar que as relações bibliográficas, apresentadas pelo FRBR, são especialmente importantes para as publicações ampliadas, que são tipos de documentos que reúnem informações de natureza bibliográfica.

2) Relacionamentos Textuais

Como se pode perceber, o advento das tecnologias vem permitindo que as relações entre os documentos sejam estabelecidas não apenas entre documentos - como há muito tempo era feito, a partir das pistas dos catálogos bibliográficos e também das citações feitas no texto – mas também dentro do próprio texto, a partir, por exemplo, de uma escrita hipertextual. Os relacionamentos que têm por objetivo indicar outros documentos são chamados de intertextuais, já os relacionamentos que se dão dentro do próprio documento são chamados de intratextuais.

Os relacionamentos bibliográficos já citados são tipos de relacionamentos intertextuais, pois indicam, manifestações, expressões, que são outros documentos que guardam algum tipo de relação com a obra principal.

Green (2001, p.8) inclui entre os relacionamentos intertextuais, os *links*, mas estes podem ter as duas naturezas. Existem os *links* intratextuais que, assim como as citações, relacionam documentos que estão fora do texto, e existem *links* intertextuais que remetem a outras partes do próprio texto. De uma forma ou de outra, Green defende que estas relações sejam estabelecidas de forma semântica a partir da técnica de *links* tipados.

Esta técnica se define por estabelecer um *link* para outro documento revelando o caráter semântico desse *link*, isto é, ao invés de simplesmente apontar para o documento, o *link* especifica também que tipo de relação este documento tem com o outro, ou seja, se é um índice do outro documento, uma citação, uma revisão, etc. A técnica dos *links* tipados não é recente. Estudos apontam que sistemas anteriores à Internet já tinham como característica comum o uso de *links* tipados, no entanto havia uma dificuldade em aplicá-los. Atualmente, a versão 4 do HTML, padrão sugerido pelo W3C possui recursos para *links* tipados utilizando o <rel> (relação para a frente) e o <ver> (relação inversa) para atributos. Esses recursos podem ser aplicados para ligar o documento total <link> tag , e para ligar parte do documento, usando <a>tag. Por exemplo, a tag <link rel="contents" href="top.html"> especifica que o documento "top.html" é um sumário, outro exemplo é tag <link rel = "next" href = "cap3.html"> especifica que "cap3.html" é o próximo documento em seqüência lógica, após o documento que o leitor está lendo.

O HTML 4⁵³ estabelece como padrão os seguintes *links* tipados: *alternate*, *stylesheet*, *start*, *next*, *prevcontents*, *index*, *glossary*, *copyright*, *chapter section*, *subsection*,

⁵³ Disponível em: <<http://www.w3.org/TR/html4/types.html>>. Acesso em 13 maio 2013.

apêndix, help, bookmark.

Green (2001) não menciona, mas o W3C já avançou bastante em abordagens semânticas para estabelecimento de relações hipertextuais. Cabe aqui acrescentar que outra técnica criada pelo consórcio foi o *link* semântico, em que, com o uso de XML, se descreve, não apenas o tipo de semântica como no *link* tipado, mas também outras propriedades sobre o recurso apontado, por exemplo, em XML, um título de livro descrito como `<BookTitle isbn="0812511816"/>`. Dentro dos moldes da *Web* semântica, um *link* semântico faz parte de uma rede semântica chamada *Semantic Link Network* - SLN (ZHUGE; SUN; ZHANG, 2010), isto é, um conjunto de nós e *links* semânticos com regras de raciocínio associado a ele. Os nós semânticos podem ser recursos, classes de recursos ou até *links* semânticos, com lógica associada. As regras de raciocínio podem gerar novos *links* de forma automática e a semântica da SLN pode mudar, se essas regras forem alteradas.

3) Relacionamentos de assuntos

Esse tipo de relacionamento é de grande importância porque a recuperação por assunto é o tipo de recuperação mais usada. No entanto, este tipo de recuperação, na maioria das vezes se torna complexa, por inúmeros problemas ocorridos durante a indexação, ocasionados, por exemplo, pela polissemia linguística usada nos textos, pela falta de coerência intra e interindexador, má indexação, entre outros. Alguns desses problemas são solucionados pelo uso de recursos oferecidos por uma linguagem documentária, como um tesauro, por exemplo. O Relacionamento de assunto é um desses recursos e pelo qual podem ser indicadas temáticas associadas que muitas vezes se encontram dispersadas nos sistemas de informação. Com a especialização do conhecimento, é difícil encontrar um documento que trate de apenas um assunto, pois as temáticas estão cada vez mais inter, multi e transdisciplinares.

A riqueza das linguagens documentárias - independente da unidade de representação utilizada, se palavra, cabeçalho de assunto, notação ou termo - está na existência de relações. São essas relações que permitem ao usuário escolher entre o assunto preferido e não preferido. Green (2001, p.11) assegura que “na tentativa de estabelecer relacionamentos entre assuntos, devem ser consideradas as relações entre conceitos”, por isso essas relações podem ser chamadas também de conceituais.

Nestes instrumentos, conforme Neelamegham e Maitra (1978), as relações podem ser divididas em: hierárquicas e não-hierárquicas. Estes autores incluem entre as relações hierárquicas, não só as associativas, mas também as de equivalência. As relações hierárquicas surgem da ligação entre conceitos subordinados e superordenados, como é o caso, da Relação de Gênero-espécie e Parte-todo. Assim, todas as relações que não possuem esse tipo de ligação são consideradas não-hierárquicas.

Outros autores apresentam também outro tipo de classificação, em que as relações são divididas em três subgrupos: hierárquica, equivalência e associativa (CLARKE, 2001).

Para Clarke (2001), as relações hierárquicas são aquelas em que um termo de escopo mais amplo ambrange outro de menor alcance e podem ser divididas em: genéricas, partitivas e instanciais.

As relações **hierárquicas genéricas** são aquelas que revelam um relacionamento entre um gênero e uma espécie, dito de outra forma, são relações em que, em uma hierarquia, o termo subordinado se liga ao superordenado por ser um “tipo de”.

As **relações hierárquicas partitivas** são aquelas que revelam um relacionamento entre um todo e uma parte, isto é, são relações em que o termo subordinado se liga ao superordenado por ser uma “parte de”. Conceitualmente, existem muitas maneiras de um termo ser parte de outro. No escopo da Lógica, existe uma teoria voltada só para o estudo dos tipos de partes, chamada Mereologia Clássica. (ESCHENBACH; HEYDRICH, 1995).

As **relações hierárquicas instanciais** são definidas por Clarke (2001, p.44) como relações em que o termo superordenado é representado por um único membro e os subordinados são nomes próprios. Essa relação forma uma classe. Na prática de construção de tesouros, pode-se observar esse tipo de relação que também é representada em listas de identificadores fora da estrutura do tesouro. Clarke menciona que essas relações podem ser encontradas no *Thesaurus of ERIC Descriptors* e no *CAB Thesaurus*, no entanto não menciona a origem teórica deste tipo de relação. Pode-se acreditar que esse tipo de relação tenha origem em estudos vindo da área de Orientação a Objeto, onde conforme Campos (2001, p.104) “a relação entre uma classe de objeto e sua instância é chamada de instanciação”.

Por outro lado, existem as relações não-hierárquicas, também chamadas de sintagmáticas, que são relações que se dão em uma linha horizontal, isto é, entre elementos de um mesmo nível. Entre elas pode-se citar as relações de equivalência e as relações associativas

As **relações de equivalência** têm por finalidade controlar a forma do termo utilizado no vocabulário, em geral, essa variação é dada por sinônimos ou quase-sinônimos.

As **relações associativas**, em geral, englobam qualquer outro tipo de relação que não esteja expressa em nenhuma das categorias anteriormente citadas. São relações problemáticas, pois na maioria das linguagens documentárias não há uma regra para seu estabelecimento, o que as tornam relações subjetivas.

Outra classificação interessante para as relações conceituais é sugerida pela Teoria Geral da Terminologia de Eugene Wüster (1981), no qual este teórico coloca como centro de suas discussões as relações entre os termos, dividindo-as, em uma primeira instância, em lógicas e ônticas. As **relações lógicas** são relações de abstração, ou seja, relações que repousam no fato de que dois conceitos têm, pelo menos, uma característica comum, como as relações de Gênero-especie. Já as **relações ônticas** são aquelas que possuem proximidade no tempo e no espaço, isto é, todas as outras relações que não são identificadas por meio de abstração. Wüster inclui entre as relações ônticas, as relações de contato (coordenação – como as de parte-todo e encadeamento – como as relações de antecessão e sucessão) e relações de causalidade (relações entre gerações e entre estágios). Como se pode perceber, a área de Terminologia possui uma visão própria sobre as relações entre conceitos. Esta classificação das relações em lógicas e ônticas pode ser vista também em outras literaturas da área, como em Felber (1984), por exemplo. O interessante desta classificação é a separação das relações partitivas das hierárquicas e a junção das mesmas com as associativas.

De fato, é muito peculiar estudar as partitivas separadamente, pois elas possuem características totalmente próprias e são bem mais complexas que as relações genérico-específicas, já que existem diversas formas de uma parte se relacionar com o todo, fazendo com que mereçam uma atenção especial. O estudo das relações partitivas também recebe uma atenção especial por parte dos filósofos, os quais estabeleceram uma linha de estudo chamada Mereologia Clássica que já vem sendo aplicada em diversos domínios, inclusive na modelagem de sistemas embasando a construção de axiomas lógicos e melhorando a precisão das informações recuperadas.

A forma de ver as relações parece estar relacionada ao objetivo com o qual são empregadas em cada área. Na Ciência da Informação, essas relações são estudadas para serem aplicadas na elaboração de instrumentos de indexação, sendo assim, a reunião das

relações genérico-específicas e partitivas em uma hierarquia se torna facilitador para a visualização da estrutura informacional de uma área. Na área de Terminologia, as relações são utilizadas para compor terminologias técnicas, sendo necessário que as relações sejam bem especificadas para que um termo seja aplicado de forma adequada. Na área de Ciência da Computação, as relações são utilizadas para melhorar a lógica dos sistemas de recuperação, sendo assim, essas relações precisam ser ainda mais específicas, pois para a máquina é fundamental evitar as redundâncias. Observa-se então, na Ciência da Computação, uma ênfase diferente, no estudo das relações. Enquanto a Ciência da Informação e a Terminologia se preocupam em nomear as categorias de relações, a Ciência da Computação se preocupa em especificar o tipo de relação existente entre os objetos no mundo.

Assim, é importante mencionar ainda, aqui, outra classificação dada às relações, que utilizou como critério a área empregada. Ao sistematizar as relações conceituais em ontologias, Sales (2006, p. 83) identificou que estas relações tinham configurações diferentes na Ciência da Informação e na Ciência da Computação. Desta forma, foi definido como **relações categoriais** “aquelas relações que revelam duplas de categorias, ex: coisa-processo, material-produto etc” e como **relações formais** “aquelas relações que revelam o tipo de relação existente entre as duplas de categorias, ex: *caused_by*, *occurs_in* etc.” Nesse mesmo trabalho, Sales (2006, p.103) identificou que para as ontologias era necessário que houvesse os dois tipos de relações: as *categoriais* e as *formais*, sugerindo o uso de uma tríade de relações que envolvesse esses dois tipos. Essa tríade foi chamada em trabalhos posteriores de “modelo triádico de relações” (SALES, 2008, 2012).

Como se pode perceber, existem inúmeras tentativas de classificação das relações associativas ou, conforme chama o FRBR, relações não-hierárquicas. No entanto, mais importante que a tipologia de relações é o estabelecimento de um método que abdique da subjetividade e forneça às linguagens de representação maneiras eficientes de criar as relações.

É bom destacar que a necessidade de estabelecimento de um novo modelo, que vá para além do FRBR, se explica pelo fato do modelo ter sido desenvolvido para relações que estão dentro de um catálogo bibliográfico. Para o contexto da publicação ampliada a expansão dos estudos que envolvem os relacionamentos se faz necessária principalmente porque a relação que a publicação ampliada faz com outros documentos, dados, pessoa, organizações, etc é externa e segue para além do catálogo bibliográfico, ainda que também possa haver relações com objetos de dentro de um mesmo sistema. Isto significa dizer que a publicação ampliada é integradora de diversos sistemas e justifica o porquê do qualitativo “ampliada”, ou seja, uma publicação é ampliada à medida que expande o limite do documento para além do sistema (repositório ou revista) no qual está inserida.

Quanto à forma como as relações devem aparecer no contexto das publicações ampliadas, a literatura mostra que para as publicações ampliadas, as relações também precisam ser como nas ontologias, isto é, triádicas, considerando as categorias e a forma como essas categorias se associam (PARINOV, 2012). Contudo, mais do que a forma é necessário que se pense como essas relações podem ser estabelecidas.

Neste sentido, em 1987, Motta, com base na Teoria do Conceito de Dahlbergh, considerando a subjetividade existente na criação de relações associativas no âmbito das linguagens documentárias, estabelece o **método relacional** como nova abordagem para construção de tesauros na expectativa de atender a carência de bases metodológicas que orientassem a formação de relações consistentes. Segundo a autora, este método “consiste na análise das definições dos conceitos que integrarão o sistema, de forma a identificar suas características, e na estruturação desses conceitos, tendo em vista as relações entre eles.” (MOTTA, 1987, p.39). Esse método é interessante, pois a análise a partir das definições evita erros conceituais, auxilia na desambiguação de termos equivalentes e ainda serve de suporte para identificar as relações e padronizá-las.

Sales (2008) também sugere um método para estabelecimento de relações, no entanto, não para aplicação em instrumentos como tesauros, mas para aplicação em ontologias ou em modelagem conceitual (SALES; SAYÃO; MOTTA, 2012). Este método, doravante chamado **método relacional-categorial**, se valeu de abordagens advindas da Teoria do Conceito (DAHLBERGH, 1978a) e da Teoria da Classificação Facetada (RANGANATHAN, 1967) para “estabelecer relações a partir da combinação das categorias existentes no domínio mapeado”.

Green (2001), considerando que a característica mais importante de uma relação é a sua semântica, sugere quatro formas de explicitar as relações.

A primeira opção renuncia a qualquer tentativa de identificar a natureza ou a semântica das relações, mas conta com a enumeração dos tipos de participantes e da compreensão da relação subjacente, como por exemplo, os TR (termo relacionado) dos tesouros.

A segunda opção se refere a um tipo de relação entre as classes que se mantém por convenção, por exemplo, as relações em que as categorias são previamente estabelecidas.

A terceira opção simplesmente nomeia o tipo de relação, por exemplo, um capítulo *<is parte of>* unidade bibliográfica. Esta opção considera que o usuário esteja familiarizado com este tipo de relacionamento e possa a partir de sua própria experiência pessoal estender a compreensão semântica com base no rótulo da linguagem natural.

A quarta opção é a mais explícita e usa uma linguagem formal para estabelecer a semântica do tipo de relacionamento. Tal opção vai ser encontrada em sistemas de representação do conhecimento que incorporam tipos de relacionamentos. Geralmente, os símbolos da linguagem formal são transformados com base na experiência pessoal em linguagem natural.

Cabe registrar, portanto, que “especificar a natureza operacional do relacionamento que existe entre as entidades ou entre as classes de entidades é identificar a semântica das relações (GREEN, 2001, p.5, grifo nosso), o que vem reafirmar a necessidade do modelo de relações considerar as entidades que o envolve como elementos essenciais para a especificação da semântica do relacionamento, isto é, como uma tríade; <categoria1> relacionamento <categoria2>.

Finalizando essa seção deve-se destacar ainda que, de acordo com Green (2001), a razão real para focar as relações vem da combinação de duas consequências oriundas da explosão informacional: o registro da informação em inúmeras formas e o aumento da gestão de oferta de Fontes de informação por meios automáticos. Para a autora, “a expressão e a manipulação das relações é talvez a maior esperança para adicionar maior qualidade nos sistemas de recuperação” (GREEN, 2001, p.14). De fato, ao considerar as publicações ampliadas como item que agrega informações sobre uma mesma temática em diferentes expressões ou manifestações, são as relações que, em meio automatizado, permitem que esta agregação seja feita de forma semântica, possibilitando a recuperação conjunta de todas essas expressões ou manifestações. A presente investigação segue nessa

direção, propondo que os objetos digitais sejam distribuídos em categorias que se relacionam formalmente.

Tendo apresentado os componentes de uma publicação ampliada, uma questão que surge é como agregar esses elementos tornando-os uma unidade de informação? Esta pergunta é respondida nas subseções a seguir.

4.6 AGREGANDO ELEMENTOS NA COMPOSIÇÃO DE PUBLICAÇÃO AMPLIADA: A NORMA OAI-ORE

Uma publicação ampliada por ser um agregado de componentes é classificada como uma instância de um objeto digital complexo. Para que um objeto digital seja recuperado, assim como qualquer outro tipo de informação, ele precisa ser descrito preferencialmente por meio de normas e padrões que tornem a recuperação mais precisa. No caso de um objeto digital complexo, a norma OAI-ORE é a que se ajusta a esses propósitos, permitindo descrever as agregações de objetos de forma padronizada.

Para tornar objetos digitais complexos intercambiáveis - com propriedades, outros objetos e relações - é preciso que esses elementos sejam descritos de forma padronizada no todo e em suas partes. Para isso, o Open Archive Initiative - OAI define uma norma para descrição e intercâmbio de agregação de recursos *Web* chamada de Object Reuse and Exchange (OAI-ORE). Conforme explicitado na página *Web* do OAI-ORE⁵⁴,

esta agregação, algumas vezes chamada de objetos digitais compostos, pode combinar recursos distribuídos com tipos múltiplos de mídia, incluindo texto, imagens, dado e vídeo. O objetivo da norma é expor o conteúdo rico dessa agregação para aplicações que suportem sistemas de autoria, depósito, intercâmbio, visualização, reuso e preservação. (OAI-ORE, tradução nossa)

A norma equaciona o problema básico que é a ausência de forma padronizada para descrever os elementos constituintes do objeto digital composto e os limites de uma agregação. (LAGOZE; VAN DE SOMPEL, 2007).

Apesar de não ter sido criada especificamente para tratamento de publicações ampliadas, a norma OAI-ORE é perfeitamente adequada para este fim já que uma publicação ampliada é em sua essência um objeto digital complexo, formado pela agregação de diversos recursos digitais, que podem estar distribuídos por vários locais.

⁵⁴ Disponível em: <www.openarchives.org/ore>. Acesso em: 29 maio 2013.

O ponto de partida para o entendimento do OAI-ORE é a compreensão de que cada unidade de um objeto complexo deve ser descrita separadamente com metadados que registrem sua estrutura, linhagem e identificação persistente.

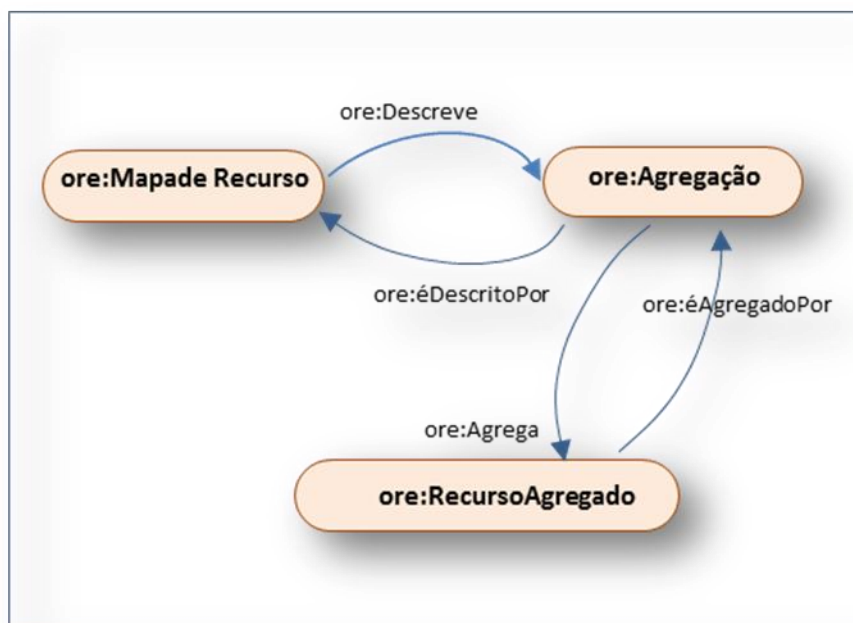
Diferente da outra norma do *Open Archive Initiative*, a OAI-PMH (Protocol for Metadata Harvesting), a OAI-ORE é considerada mais rica por requerer dos objetos digitais uma camada de interoperabilidade que vai além dos metadados puramente descritivos, considerando também na representação de cada elemento os metadados de preservação e administrativos.

A *World Wide Web* é construída sobre a noção de unidades atômicas de informação, também chamada de "recursos", que são identificados por meio de URIs - Uniform Resource Identifiers, que dão a essas unidades um caráter único permitindo que ela seja encontrada. Além dessas unidades atômicas, os recursos, muitas vezes são agregações de unidades formando um objeto complexo. Como já mencionado, a OAI-ORE define os padrões para descrição dessas agregações de recursos na *Web*.

As especificações do OAI-ORE são baseadas no modelo ORE. O modelo ORE apresenta um Mapa de recursos que permite a associação de uma entidade com agregações de recursos, fazendo afirmações sobre a estrutura e a semântica de cada recurso agregado, bem como da agregação como um todo. O formato de serialização primário para mapear o recurso é o perfil da unidade atômica expressa em RDF - Resource Descriptive Framework – uma linguagem para representação da informação na *Web* recomendada pelo consórcio W3C, para descrever por meio da sintaxe XML três elementos: recurso, propriedade e valor. Por ser expresso em RDF, o mapa de recursos também pode ser serializado por qualquer outro formato que permita serialização em RDF.

Este mapa de recursos de agregação é então uma forma de representar um conjunto de recursos associados por meio de relações. Este conjunto representado chama-se recurso agregado e deve ter um URI, garantindo que exista um único recurso de agregação para todo o mapa. A Figura 14 a seguir ilustra o modelo básico da norma OAI-ORE.

Figura 14: Modelo Básico do OAI-ORE



Fonte: VERNOOY-GERRITSEN, 2009^a

É interessante observar que o foco do OAI-ORE está mais no recurso e menos nos repositórios, podendo ser adotado para encapsular recursos distribuídos por diversos repositórios, que em geral é o que acontece com os dados de pesquisa e as publicações. Nesse sentido, a norma OAI-ORE é indicada para composição de novos modelos de publicação em que objetos digitais de diferentes naturezas, por exemplo, publicações e dados de pesquisa, se unem formando um novo objeto. A norma OAI-ORE continuará sendo discutida na seção a seguir, porém dentro do contexto da *Web Semântica*, enquadrando a publicação ampliada dentro da visão da *Web* em que a semântica dos dados é de fundamental importância. Este enquadramento vem ratificar a importância da investigação aqui realizada.

4.7 PUBLICAÇÃO AMPLIADA E WEB SEMÂNTICA

Como já mencionado, publicações ampliadas consistem na combinação de dados e informações heterogêneas, podendo vir de disciplinas distintas, em formatos distintos e também em idiomas diferentes. A visão da *Web semântica* é focada no compartilhamento e

reuso de dados, diferente da Web tradicional que é focada no compartilhamento de documentos apenas. Isto faz com que a ideia das publicações ampliadas seja totalmente apropriada aos propósitos da Web semântica.

O W3C, sigla em inglês para World Wide Web Consortium, responsável pela disseminação das ideias da Web semântica e também pela criação de normas e padrões que possibilitem a concretização desta nova visão da Web, recomenda que os dados sejam acessados usando uma arquitetura Web que utiliza URIs para cada recurso e relaciona os dados uns aos outros, assim como documentos ou partes de documentos. Para as publicações ampliadas, a integração dessas soluções sugeridas pelo W3C é especialmente importante, pois possibilita a criação de uma lógica de conexão entre os termos, o que permite uma integração semântica entre as partes da publicação e também entre os dados.

De acordo com o W3C, as tecnologias da Web semântica podem ser usadas para diversas aplicações, por exemplo: para **integração de dados**, interligando dados que estão em vários locais e formatos distintos; para **classificação e descoberta de recursos**, melhorando a capacidade dos motores de busca em domínios específicos, isto é, fazendo com que os resultados das buscas sejam mais relevantes; para **catalogação**, descrevendo o conteúdo e suas relações disponíveis em um determinado site, página ou biblioteca digital; para **agentes de software inteligentes**, facilitando o compartilhamento de conhecimentos, para **classificação de conteúdos**, descrevendo coleções de páginas que representam documentos sobre uma mesma temática, para descrever **os direitos** associados às páginas da Web, registrando a propriedade intelectual (por exemplo, o *creative commons*), entre outros. Todos esses recursos permitirão que a Web forneça respostas mais inteligentes às solicitações de seus usuários.

Para que isso seja possível, o W3C (2001) afirma que o mais importante é ser “capaz de definir e descrever as relações entre os dados (ou seja, os recursos) na Web”. A Web atual conecta páginas, mas a Web semântica quer conectar dados, independente de onde estejam. Esta conexão precisa ser realizada por meio de relações bem definidas, pois são elas que facilitarão o intercâmbio automático dos dados.

Com relação à interoperabilidade das publicações ampliadas, as informações de cada uma de suas partes precisam ser suficientemente detalhadas e precisas. Para isso, é necessária uma representação apropriada tanto no nível sintático quanto no nível semântico. Para a representação no nível sintático, o W3C sugere o uso de soluções como URI, XML, *XMLschema* e RDF. Já para o nível semântico, a sugestão é o uso de ontologias.

Como já mencionado, o URI serve para dar uma identidade única aos recursos *Web*. O XML ou *eXtensible Markup Language* é uma linguagem de marcação que define um conjunto de regras para a codificação de recursos em um formato legível por homens e por máquina, possibilitando o compartilhamento desses recursos via *Web*. Esta linguagem é a que fornece mais detalhes sobre o recurso, pois incorpora metadados em sua estrutura. Já o *XMLschema*, é uma linguagem baseada em XML que serve para validar as regras utilizadas na estruturação e conteúdos de documentos XML, como tipo, tamanho, regras de preenchimento, entre outros.

O *Resource Description Framework* - RDF, que é um dos alicerces da *Web* semântica, fornece uma definição formal para o intercâmbio de dados, com base em três componentes: **sujeito ou recurso**, que pode ser qualquer objeto que possua um URI, por exemplo, uma página *Web*, um documento XML, ou um conjunto de dados; **predicado ou propriedade**, que é a parte que identifica uma característica do recurso, por exemplo, o criador da página *Web*, e **indicação ou objeto**, que é a parte que identifica o valor de um recurso (BRIETMAN, 2005).

O predicado ou propriedade expressa uma relação do recurso com o objeto, por exemplo, o site <www.w3.org/standards/semanticWeb> criado pelo W3C, onde o endereço <www.w3.org/standards/semanticWeb> é o recurso, a relação é “*created by*” e a indicação é “W3C”. O RDF permite então a integração dos dados com um pouco mais de semântica.

Como se pode verificar, a *Web* semântica oferece ferramentas que permitem que dados heterogêneos sejam compartilhados e reutilizados. Como as publicações ampliadas consistem de dados heterogêneos, tipos de informações de diferentes disciplinas, em línguas diferentes que precisam ser integrados semanticamente, as soluções oferecidas pela *Web* semântica podem e devem ser utilizadas na construção desse tipo de publicação. O problema de lidar com informações heterogêneas é sempre o fato dos mesmos elementos poderem expressar significados diferentes para casos e tipos diferentes. A integração semântica é um processo que utiliza a representação conceitual dos dados e suas relações

para eliminar possíveis ambiguidades. Essa representação conceitual é feita por meio de ontologias – que são

instrumentos de representação do conhecimento definidos em uma linguagem formal e processável por máquina que possibilitam a descrição dos aspectos semânticos dos conteúdos informacionais, explicitando seus relacionamentos de modo detalhado a partir de restrições lógicas que possam ser processadas de forma automatizada, possibilitando inclusive relacionamentos baseados na Lógica de Segunda Ordem. (RAMALHO, 2006, p.59)

No contexto da Web Semântica, ontologias são estruturas informacionais que permitem que o significado dos recursos *Web* seja padronizado garantindo a interoperabilidade entre sistemas e podendo ser utilizados na integração de metadados por meio de um esquema global. O W3C recomenda que para a integração de dados heterogêneos, as ontologias sejam expressas em OWL - *Ontology Web Language*: uma linguagem baseada em XML própria para instanciar ontologias por meio de descrições de classes, suas respectivas propriedades e relacionamentos.

Em uma publicação ampliada, o uso de ontologias para integrar documentos digitais apoia não apenas o acesso à informação de forma integrada e precisa como também o reuso unificado destas informações.

De acordo com Tello (2002), as ontologias são instrumentos formados por: **conceitos** - que são idéias básicas sobre o que se pretende formalizar, **classes e subclasses** - que podem estar organizadas em uma taxonomia, **relações** - que devem representar os tipos de interação entre as classes de um domínio, **funções** – que casos especiais de relações que não possuem inversa, mas expressam alguma característica única para seus elementos, **axiomas** - que são teoremas ou regras de inferência para determinar a verdade das sentenças de forma lógica e **instâncias** - que são utilizadas para representar elementos específicos, ou seja, os próprios dados. Assim, as publicações ampliadas podem se beneficiar do uso de ontologias à medida que se valerem dos componentes dessa ontologia para ligar as publicações aos diversos tipos de dados que a geraram, permitindo uma contextualização lógica e semântica.

Como já mencionado, registrar as propriedades básicas da publicação e dos outros recursos que estão adicionados a ela é um dos dez requisitos apresentados por Vehaar (2008, p.17) no relatório do DRIVER-II para compor uma publicação ampliada. Essas propriedades são registradas por meio de metadados que quando padronizados e controlados permitem que a publicação seja interoperável.

A interoperabilidade é atualmente a chave principal para a disseminação da informação científica. Quando sistemas interoperam, a fronteira da disseminação é expandida e tornam a informação acessível por diversas plataformas *Web*. Na composição da publicação ampliada, a propriedade de ser interoperável é uma característica ainda mais importante, pois na maioria das vezes é formada por informações que estão distribuídas em diversos sistemas. Desta forma, faz-se necessário que as informações que a compõem utilizem um núcleo de metadados padrão que as tornem compatíveis.

Vehaar (2008, p.27) descreve alguns modelos que visam padronizar a terminologia utilizada na composição de publicações ampliadas, a saber: o modelo abstrato *Dublin Core Metadata Initiative*, DCMI⁵⁵; o Modelo de Referência Conceitual CIDOC/CRM⁵⁶; o Vocabulário ORE⁵⁷ - *Object Reuse and Exchange*; e a ontologia de relacionamentos do FEDORA⁵⁸ (*Flexible Extensible Digital Object Repository Architecture*). Estes modelos serão analisados a seguir.

O modelo abstrato DCMI especifica os componentes e os construtos usados no esquema de metadados Dublin Core, descrevendo uma estrutura de informação chamada *Description set DC* e especifica como essa estrutura deve ser interpretada. O modelo abstrato DCMI baseia-se no RDF e fornece um modelo de informação que é independente de qualquer sintaxe de codificação, permitindo uma melhor compreensão dos tipos de descrições codificados.

Outro modelo citado no documento de Vehaar (2008) é o CRM - *Conceptual Model Reference* do Comitê Internacional de Documentação - CIDOC-*International Committee for Documentation* do Conselho Internacional de Museus - ICOM - *International Council of Museums* - que fornece definições e uma estrutura formal para descrever os conceitos e relações implícitas e explícitas utilizadas na documentação do patrimônio cultural. Este modelo fornece um quadro semântico comum e extensível para que qualquer informação sobre herança cultural possa ser mapeada, auxiliando assim na formulação de requisitos para sistemas de informação e servindo como um guia de boas práticas para modelagem conceitual neste domínio. Apesar de ter sido construída para a área do patrimônio cultural, a CRM do CIDOC é perfeitamente aplicável a outros campos, pois é um vocabulário de alto nível, definindo conceitos genéricos, como classe, subclasse, superclasse, nota de

⁵⁵ Disponível em: <dublincore.org>. Acesso em: 15 maio 2014.

⁵⁶ Disponível em: <www.cidoc-crm.org/>. Acesso em: 15 maio 2014.

⁵⁷ Disponível em: <www.openarchives.org/ore/1.0/vocabulary>. Acesso em: 15 maio 2014.

⁵⁸ Disponível em: <fedoraproject.org/pt>. Acesso em: 15 maio 2014.

escopo, propriedade, instância, intensão, extensão, domínio, herança, herança múltipla, etc. Além disso, quantifica os tipos de propriedades, se um-para-um, um-para-muitos, muitos-para-um, muitos-para-muitos, entre outros.

Por outro lado, o vocabulário fornecido pelo OAI-ORE visa descrever itens de interesse e expressar as relações entre eles. É necessário assinalar que como este vocabulário usa o princípio de reuso de vocabulários existentes para termos que não são específicos e fundamentais para o modelo ORE, ele só cobre os aspectos não abrangidos por outros vocabulários, como por exemplo, os do DCMI ou do RDF. Os termos do vocabulário estão enquadrados no espaço de nomes (*namespace*) ORE e são usados para construir mapas de recurso ORE, que descrevem as agregações de recursos, como visto anteriormente na seção 4.4.

O ORE define uma classe semântica como um conjunto de recursos descritos com ORE. Isto ajuda os aplicativos a compreenderem o que a agregação contém e representa, por exemplo, uma agregação de artigos de periódico composta de um periódico, um número do periódico, um volume do periódico, um número especial, uma lista de citação, etc. As principais entidades de interesse para o OAI-ORE são: *Agregação*, *Recurso Agregado*, *Proxy* e *Mapa de Recurso*.

Uma **agregação** é um conjunto de recursos disponíveis agrupados de tal modo que o conjunto pode ser tratado como um único recurso. Pode-se dizer que uma publicação ampliada é uma instância de agregação. Um **recurso agregado** é um recurso que está incluso em uma agregação específica, ou seja, é uma unidade membro da classe de recurso, ou um componente da publicação ampliada. Um **proxy** é um representante de um recurso agregado incluso numa agregação específica, desta forma pode-se citar o recurso agregado ao invés de citar a agregação, como por exemplo, um capítulo de um livro que pode ser citado ao invés de citar o livro no todo. O Proxy pode ser o DOI de um documento. O conceito **mapa de recurso** se refere a grafos RDF que descrevem a agregação para um formato legível por máquina.

Além destas entidades, o ORE define também as relações entre elas. Estas relações podem se referir a outro objeto, mas este objeto também pode ser um valor literal em vez de outro recurso. Alguns são conceitos abstratos, por exemplo, a declaração de direitos que pode ser incorporada como um *string* ou uma referência a um recurso externo.

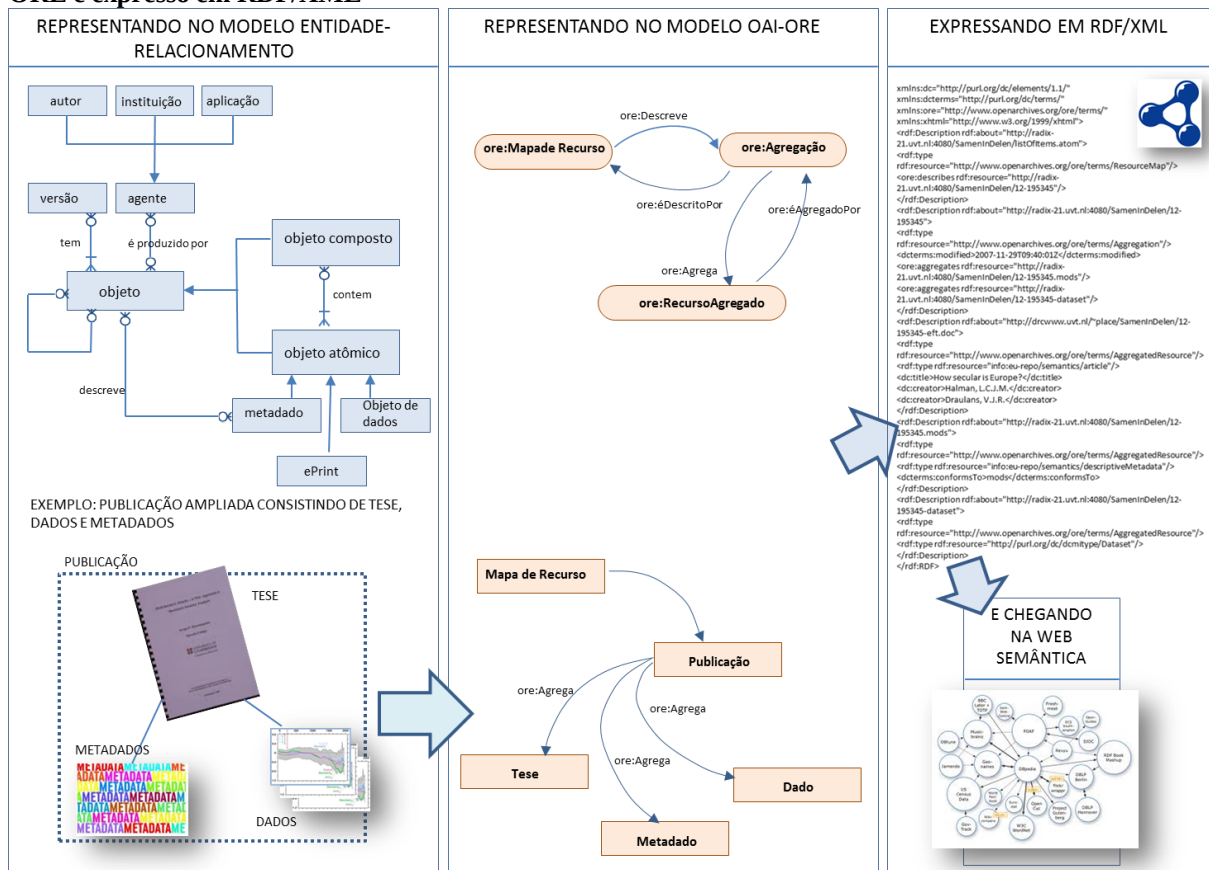
Por último, Vehaar (2008) cita a ontologia de relações do Fedora, que é um *software* de código aberto originalmente desenvolvido por pesquisadores da *Cornell*

University como uma arquitetura para armazenamento, gestão e acesso de conteúdos na forma de objetos digitais complexos. O Fedora define um conjunto de abstrações para expressão de objetos digitais, seus relacionamentos e comportamentos (serviços). O sistema é amplamente utilizado para construção de repositórios, pois utiliza o protocolo OAI-PMH e recursos que garantem a preservação digital, como alinhamento ao OAIS. Além disso, o Fedora fornece suporte RDF o que permite que o sistema seja integrado com a tecnologia de triplas semânticas. O sistema de repositório Fedora definiu um conjunto padrão de relações comum na ontologia de relacionamento Fedora.

É importante mencionar que, para o Fedora, as relações são essenciais, pois permitem acesso e gestão de metadados dentro do repositório. Por exemplo, organizam objetos em coleções de apoio à gestão, apoiam o *OAI harvesting* e a navegação pelo usuário, são usadas como relações bibliográficas entre objetos, tais como definidas pelo *Functional Requirements for Bibliographic Records - FRBR*, definem semântica entre os recursos, modelam uma rede de recursos com base em informações contextuais (por exemplo, *links* de citação ou anotações colaborativas), codificam hierarquias naturais entre objetos, fazem ligações cruzadas para *harvesting* de objetos (mostrando, por exemplo, que um determinado documento de uma coleção também pode ser parte de outra coleção).

Para concluir esta seção, a Figura 15 ilustra como uma instância de agregação - composta por uma publicação ampliada formada por um *e-print*, dados e metadados - pode ser representada pelo modelo entidade-relacionamento. A Figura 15 ilustra também como a agregação pode ser descrita usando o vocabulário OAI-ORE e ainda como ela pode ser expressa em RDF/XML, aproximando-se do padrão tratado pela *Web* semântica.

Figura 15: Publicação ampliada representada pelo modelo entidade-relacionamento, pelo modelo OAI-ORE e expresso em RDF/XML



Fonte: A autora, com base em VERNOOY-GERRITSEN, 2009a

Embora iniciativas de publicações ampliadas e semânticas já existam, a literatura mostra que a maioria dessas abordagens possui limitações, incluindo:

- As relações entre os conjuntos de dados e publicações, em geral, são *one-to-one*.
- Os *hyperlinks* são relativamente fixos e envolvem pouca semântica ou nenhum apoio para informações de procedência.
- Falta de suporte para acesso multi-nível a dados ou informações.
- Os sistemas de publicação existente parecem apoiar o acesso livre somente de publicações.
- Falta de sistemas baseados em regras ou um modelo diferente para apresentação.

- A falta de flexibilidade ou extensibilidade - que permitam o cientista escolher interativamente um conjunto de recursos gerados a partir de uma experiência ou de investigação.

O capítulo a seguir descreverá os materiais e métodos utilizados para o desenvolvimento desta pesquisa, bem como as etapas metodológicas.

5 MATERIAL E MÉTODO

Em seu livro “Como escrever uma tese”, Humberto Eco (1977, p.5) afirma que fazer uma tese é “uma experiência de trabalho metódico, quer dizer, construir um objeto que como princípio também possa servir aos outros”. Para que o objeto de estudo de um pesquisador seja compreensível pelos outros se faz necessário o comprometimento compartilhado com a forma de execução da pesquisa. Este capítulo foi desenvolvido com o objetivo de relatar a forma como se encaminhou esta pesquisa. O embasamento teórico para tais escolhas metodológicas foi levantado na literatura sobre métodos e técnicas de pesquisa social, em especial o livro de Gil (2008).

5.1 MATERIAL

O material utilizado para compor essa pesquisa foi essencialmente de dois tipos: material bibliográfico e dados de pesquisa.

O material de cunho bibliográfico abrangeu duas grandes áreas: Ciência da Informação e Ciências Nucleares, porém com duas finalidades distintas, a saber:

Primeiramente, o material bibliográfico levantado no âmbito da Ciência da Informação serviu para dar suporte teórico ao desenvolvimento deste estudo. Todo o encaminhamento desta pesquisa foi embasado em conceitos e estudos anteriores sobre os assuntos em questão.

Em um segundo momento, outro material bibliográfico foi levantado no âmbito da área de Ciências Nucleares com outra finalidade: verificar a inovação do tema em sua área de aplicação.

O levantamento bibliográfico da área de Ciência da Informação foi realizado na base de dados Library, Information Science and Technology⁵⁹ (LISTA)- por ser uma base da área de Ciência da Informação que está disponível no Portal de Periódicos CAPES⁶⁰, com acesso aos resumos e ao texto completo. Além disso, foram considerados também periódicos sobre temas específicos da tese, como por exemplo, o “*The International*

⁵⁹ Disponível em:

<<http://Web.ebscohost.com.ez106.periodicos.capes.gov.br/ehost/search/basic?sid=90ea5b73-67b3-4aa9-9537-95d75e0aef67%40sessionmgr113&vid=1&hid=123>>. Acesso em: 26 dez. 2012.

⁶⁰ Disponível em: <<http://www.periodicos.capes.gov.br/>>. Acesso em: 26 dez. 2012.

Journal of Digital Curation”.

A partir do material levantado foi escolhida uma amostragem por acessibilidade ou conveniência do autor. De acordo com Gil (2008, p.94), este tipo de amostragem é bastante usado em estudos exploratórios ou qualitativos, como é o caso desta pesquisa. Assim, a escolha dos artigos que fizeram parte da revisão bibliográfica, que se encontra relatada ao longo desta tese, foi feita buscando sempre utilizar Fontes de responsabilidade dos principais autores de cada temática.

O levantamento de material bibliográfico na área de Ciências Nucleares foi realizado com a finalidade de verificar a existência ou não de modelo semelhante ao proposto por esta tese. Para tal, foi realizado o levantamento em uma base internacional da área Nuclear, a base de dados bibliográficos do International Nuclear Information System (INIS) –, da Agência Internacional de Energia Atômica (AIEA/ONU).

Os dados de pesquisa trabalhados foram levantados por meio de entrevistas (APÊNDICE B) realizadas com os líderes das áreas temáticas de pesquisa do Instituto de Engenharia Nuclear - IEN.

5.2 CLASSIFICAÇÃO METODOLÓGICA E ETAPAS SEGUIDAS

As classificações não são mutuamente exclusivas, pois uma pesquisa pode assumir simultaneamente diversas características. A literatura na área de Ciências Sociais apresenta diversas abordagens classificatórias para a pesquisa científica. Demo (1987, p.13), por exemplo, as classifica em: pesquisa teórica - dedicada ao estudo de teorias e ao desenvolvimento de modelos; pesquisa metodológica – voltada para o modo de fazer ciência; pesquisa empírica – dedicada à face experimental e observável dos fenômenos; e pesquisa prática - ou pesquisa-ação – que visa a interferência em uma dada realidade social. De acordo com este autor, esta pesquisa seria classificada como uma pesquisa teórica.

No entanto, a pesquisa realizada por esta tese utilizou como referencial teórico as diversas abordagens mencionadas por Gil (2008), em seu livro intitulado “Métodos e técnicas de pesquisa social”. Seguindo a linha desse autor, esta tese pode ser classificada: quanto à sua finalidade, quanto ao método utilizado, quanto ao delineamento da pesquisa e quanto à abordagem a ser utilizada para a análise dos dados.

5.2.1 *Classificação metodológica da pesquisa*

Quanto à finalidade, essa pesquisa pode ser classificada como uma pesquisa pura e aplicada. Pura, pois se pretende que ela contribua para construção de um modelo teórico e, aplicada, por ser aplicada em uma área específica do conhecimento: a área de Ciências Nucleares.

Quanto aos níveis de pesquisa ela é considerada uma pesquisa exploratória, pois tem como finalidade principal “desenvolver, esclarecer e modificar ideias para formulação de problemas e hipóteses pesquisáveis para estudos posteriores” (GIL, 2008, p.27), ou seja, nesse contexto, esta pesquisa visa investigar conceitos novos no âmbito da Ciência da Informação – como “objetos digitais complexos”, “publicações ampliadas”, “curadoria digital”, “dados de pesquisa”, “relações semânticas” entre outros. Busca também esclarecer e modificar ideias, identificando problemas e hipóteses que poderão ser o ponto de partida para novos estudos e o estabelecimento de uma possível linha de pesquisa.

Quanto ao método utilizado, esta pesquisa utiliza o método dedutivo, pois “parte do geral e, a seguir, desce ao particular” (GIL, 2008, p. 9). Com base em princípios reconhecidos como verdadeiros, segue de uma abordagem genérica sobre os conceitos trabalhados e faz uma aplicação na área de Ciências Nucleares.

Quanto ao delineamento, a presente pesquisa é considerada de cunho bibliográfico, pois seus conceitos foram levantados em literatura já publicada, como livros e artigos de periódicos.

Quanto à abordagem, a pesquisa aqui relatada é classificada como qualitativa, pois todo o material obtido - desde o bibliográfico até os dados coletados por entrevistas - foi analisado, interpretado e descrito textualmente.

5.2.2 *Etapas Metodológicas*

Visando o alinhamento com os objetivos expostos na seção 1.4, esta investigação seguiu as seguintes etapas metodológicas.

- a. Levantamento e análise de material bibliográfico sobre novos modelos de publicação científica adequados ao contexto da *eScience*.

Consistiu em levantamento realizado no site Google Acadêmico e indicações de especialistas no assunto.

- b. Levantamento, análise de material bibliográfico sobre padrões, tecnologias e modelos de informação que envolvem curadoria digital de dados de pesquisa.

Consistiu em levantamento realizado na base LISTA, disponível no Portal de Periódicos CAPES, no Google Acadêmico e indicações de especialistas no assunto.

- c. Levantamento, identificação e sistematização dos elementos importantes para a proposição de um modelo de curadoria digital.

➤ Análise de três relatórios considerados fundamentais no endereçamento de questões de curadoria de dados de pesquisa: os relatórios da National Science Foundation (NSF), do Digital Data Curation (DCC) e da Organização para a Cooperação e Desenvolvimento Econômico (OCDE).

- d. Caracterização do campo empírico aplicado à área de Ciências Nucleares, que culminou em duas subetapas, a saber:

➤ Mapeamento conceitual da área de Ciências Nucleares

- Realizada por meio de análise de domínio com a abordagem de análise das estruturas institucionais e comunicação científica da área de Ciências Nucleares.

➤ Mapeamento das atividades de pesquisa do IEN

- Realizada também por meio de análise de domínio, mas sob a abordagem de indexação e estudos terminológicos.

Esta subetapa consistiu na análise dos assuntos presentes nos cadernos de áreas temáticas, por meio de indexação com metodologia do INIS – principal sistema de informação da área de Ciências Nucleares - e em entrevista com os líderes das áreas temáticas, para identificar os tipos de dados gerados no ambiente de pesquisa no IEN. (APÊNDICE B)

e) Identificação dos tipos de *e-print* e dados de pesquisa utilizados na área de Ciências Nucleares.

- Esta etapa consistiu nas seguintes subetapas:
 - Análise do Material bibliográfico digital produzido pelos pesquisadores e depositados na Biblioteca do IEN
 - Análise dos dados obtidos via entrevista com os líderes de pesquisa
 - Identificação das entidades que formarão o modelo de publicação ampliada e suas relações

f) Levantamento de normas e padrões a serem utilizados em interface de publicação ampliada para a área de Ciências Nucleares.

- Consistiu em levantamento no Google e indicações de especialistas no assunto.

g) Levantamento de material bibliográfico sobre relações conceituais no âmbito da Ciência da Informação que possam ser úteis à construção de publicações ampliadas.

- Consistiu em levantamento realizado na base LISTA, disponível no Portal de Periódicos CAPES, no Google Acadêmico e indicações de especialistas no assunto.

O capítulo a seguir visa apresentar o campo empírico escolhido, ou seja, a área de Ciências Nucleares, sob a fundamentação das abordagens de análise de domínio de Hjørland (2002) será descrito o principal modelo de comunicação científica da área: O INIS e ainda a estrutura organizacional de uma das principais instituições de pesquisa na área de Ciências Nucleares do país: O Instituto de Engenharia Nuclear da Comissão Nacional de Energia Nuclear.

6 CAMPO EMPÍRICO: O DOMÍNIO DA ENERGIA NUCLEAR

A caracterização da área de Ciências Nucleares é fundamental para a compreensão do problema de pesquisa de que trata esta tese, pois conforme Souza

as questões de ordem prática apresentam-se como forte componente de viabilização de uso dos instrumentos de organização do conhecimento e representação da informação nos ambientes definidos de atuação e aplicação. Esses ambientes revelam a necessidade de criação de “artifícios” de ajustes entre as necessidades de informação e as práticas de tratamento da informação. (SOUZA, 2012, p.6)

As publicações ampliadas podem ser consideradas como esse “artifício” que ajusta as necessidades de informação dos usuários da área de ciências nucleares e as práticas atuais de tratamento da informação, como por exemplo, a curadoria digital de dados de pesquisa. Sendo assim, este capítulo vem apresentar a área de Ciências Nucleares, enquanto campo empírico para o desenvolvimento de relações conceituais em publicações ampliadas.

Hjorland (2002), em seu artigo *“Domain Analysis in Information Science: eleven approaches – tradicional as innovative”* reúne onze abordagens interessantes que podem ser utilizadas para análise de um domínio. Essas abordagens estão sendo adotadas com frequência por cientistas da informação ao tratarem de uma área que não é de sua especialidade. As abordagens são:

- 1) Produção de Guias de Literatura ou Portais Especializados
- 2) Construção de classificações especializadas e tesouros
- 3) Indexação e recuperação nas especialidades
- 4) Estudos empíricos de usuários
- 5) Estudos bibliométricos
- 6) Estudos históricos
- 7) Estudos documentais e de gênero
- 8) Estudos epistemológicos
- 9) Estudos terminológicos, linguagens especializadas, bases de dados semânticos e estudos de discurso
- 10) Estruturas e instituições na comunicação científica
- 11) Cognição científica, conhecimento especializado e inteligência artificial

Como o autor afirma que essas abordagens não são exaustivas nem mutuamente exclusivas, neste trabalho foi decidido trabalhar com três das onze abordagens. Uma delas é usada neste capítulo para descrever o campo empírico, a saber: estrutura e instituições na comunicação científica. As outras duas foram utilizadas no capítulo da descrição dos resultados, sendo elas: indexação nas especialidades e estudos terminológicos.

6.1 ANÁLISE DE DOMÍNIO SOB A ABORDAGEM DA ESTRUTURA DO SISTEMA DE INFORMAÇÃO E COMUNICAÇÃO CIENTÍFICA NO DOMÍNIO DAS CIÊNCIAS NUCLEARES: O CASO DO INIS

A área de Ciências Nucleares tem uma longa tradição na organização, tratamento e na disseminação da informação, que remonta a antigos sistemas de informação, como o construído em torno da base de dados ENDS - European Nuclear Documentation System, ainda na década de 1960. O International Nuclear Information System - INIS, criado sob a responsabilidade da Agência Internacional de Energia Atômica - AIEA, órgão das Nações Unidas, em 1970, deu prosseguimento à política de valorização da informação nuclear, como insumo estratégico para o desenvolvimento das aplicações pacíficas da energia nuclear.

O princípio operacional do INIS estava baseado no controle da literatura técnico - científica da área nuclear dentro das fronteiras de cada país pelos centros nacionais (no Brasil, por exemplo, o CIN/CNEN – Centro de Informações Nucleares, da Comissão Nacional de Energia Nuclear). O tratamento descentralizado exigiu o desenvolvimento de uma infraestrutura que incluía regras, *softwares*, padrões, treinamento e fóruns internacionais para o estabelecimento de diretrizes e políticas de coleta e tratamento técnico – catalogação e indexação – e intercâmbio.

Assim como para a área de Documentação e Ciência da Informação, o sonho de se ter dados tratados e interligados aos seus documentos é antigo (OTLET, 1934; BUSH, 1945; NELSON, 1974, entre outros). Para a área de Ciências Nucleares este desejo também vem de longa data. Muito antes do advento das atuais Tecnologias de Informação e Comunicação (TIC's), a comunidade internacional nuclear já se preocupava com a coleta, tratamento e disseminação de dados nucleares.

Em relatório datado de 1967, o pesquisador emérito da CNEN, Bernhard Gross, cuja participação na formação do INIS foi de fundamental importância, apresenta a idéia que embasou a criação dos sistemas de informação da área nuclear na época.

O atual sistema informativo de muitas nações em desenvolvimento é falho devido à dispersão, insuficiência ou mesmo à falta de aparelhamento e amparo às bibliotecas científicas. As nações que já estão em condições de acompanhar o desenvolvimento técnico mundial e de aproveitar o que ele transmite deverão começar também com a modernização e subsequente automação do seu sistema de informações técnico-científicas, sob pena de se verem, gradativamente, privadas das suas Fontes usuais de informação. (GROSS, 1967.p.4)

Tomando como base os pressupostos de automação, a formação descentralizada de bases de dados e cooperação, aliados às diretrizes propostas pelo Manual do United Nations International Scientific Information - UNISIST, a AIEA propõe um sistema de informação de abrangência mundial que levasse em conta as necessidades dos países desenvolvidos e em desenvolvimento. Nas palavras do Prof. Gross:

A Agência Internacional de Energia Atômica, reconhecendo esta situação, propõe-se a organizar e realizar um projeto bastante ambicioso que vem ao encontro das exigências dos países em desenvolvimento e, ao mesmo tempo, contém elementos que interessam também às nações mais avançadas, assegurando, assim, cooperação ativa das mesmas. Este projeto prevê um Sistema Internacional de Informações Nucleares (INIS) inteiramente automatizado com centro de memória e armazenamento em Viena, mas operando com uma rede descentralizada de coleção de dados. (GROSS, 1967, p.5)

O sistema proposto considerava a interdisciplinaridade e a multidisciplinaridade das Ciências Nucleares: “A cobertura do sistema será a mais ampla possível, incluindo tudo que, direta ou indiretamente, está relacionado com a ciência nuclear e atômica” (GROSS, 1967, p.5).

A automação como princípio, contribui decisivamente para o progresso dos sistemas de informação e para os processos de formação de bases de dados legíveis por computador. Para corresponder às necessidades do centro mecanizado, estas informações deverão ser encaminhadas em forma computada, de modo a serem 'lidas' diretamente pela máquina do centro internacional. (GROSS, 1967, p.5)

Entre os objetivos da IAEA estavam não somente o tratamento e disseminação da informação bibliográfica, mas também de dados numéricos. Assim, para reunião da informação bibliográfica a IAEA criou o INIS e para o registro das referências a medidas, cálculos e avaliações de seções de choque, entre outras constantes neutrônicas microscópicas criou o CINDA (Computer Index of Nêutron Data). (FERRAZ, 1975).

Atualmente, a área Nuclear pode ser considerada líder em geração intensiva de dados. O conjunto de dados gerados no domínio da Física de Alta Energia tem sido historicamente o maior conjunto de dados produzidos entre as comunidades científicas. Esta comunidade foi também a primeira a adotar a tecnologia de computação em *grid*, que é a principal tecnologia para compartilhamento de dados entre usuários conectados através de uma rede de computadores. Todos esses dados são frutos da criação de diversas plataformas de computação para geração intensiva de dados.

Ainda que, no domínio nuclear, iniciativas de gerenciamento desses dados existam em uma ou outra comunidade, pouco se fala sobre curadoria desses dados, no sentido de preservá-los para uso e reuso futuro. Apesar da existência de outras iniciativas que visam reunir a memória técnico-científica nuclear, nenhuma delas busca a ligação semântica entre a produção científica e seus dados. Neste sentido, um modelo semântico para construção de publicação ampliada pode ser considerado uma novidade na área.

6.2 ANÁLISE DE DOMÍNIO SOB A ABORDAGEM DA ESTRUTURA DAS INSTITUIÇÕES CIENTÍFICAS: A CNEN E O IEN

Seguindo uma das abordagens indicadas por Hjørland (2002), isto é, a da análise da estrutura de instituições, bem como aspectos que envolvem a comunicação científica do domínio a ser mapeado, esta subseção apresenta uma breve descrição do que é a Comissão Nacional de Energia Nuclear - CNEN e o Instituto que serve de campo empírico para a realização dessa pesquisa: o Instituto de Engenharia Nuclear - IEN.

6.2.1 A Comissão Nacional de Energia Nuclear – CNEN

A CNEN⁶¹ é uma autarquia federal brasileira vinculada ao Ministério da Ciência, Tecnologia e Inovação - MCTI. Criada em 1956 pelo Decreto n. 40.110 e vinculada ao MCTI desde 1999. A CNEN tem como principais funções regular o uso da energia nuclear no Brasil e desenvolver pesquisas sobre aplicação de técnicas nucleares. Seu objetivo é levar, com segurança, os benefícios da energia nuclear a um número cada vez maior de brasileiros.

Atualmente, a atuação da CNEN se divide em duas grandes áreas: Radioproteção e Segurança Nuclear e Pesquisa e Desenvolvimento de Tecnologias Nucleares.

A área de Radioproteção e Segurança Nuclear visa a segurança dos trabalhadores que lidam com radiações ionizantes, da população em geral e do meio ambiente. Com esse objetivo, atua no licenciamento de instalações nucleares e radioativas; na fiscalização de atividades relacionadas à extração e à manipulação de matérias-primas e minerais de interesse para a área nuclear; no estabelecimento de normas e regulamentos; na fiscalização das condições de proteção radiológica de trabalhadores nas instalações nucleares e radiativas; no atendimento a solicitações de auxílio, denúncias e emergências envolvendo Fontes de radiações ionizantes; no desenvolvimento de estudos e na prestação de serviços em metrologia das radiações ionizantes. O transporte, o tratamento e o armazenamento de rejeitos radioativos são regulamentados por normas técnicas e procedimentos de controle. O controle do material nuclear existente no País também é de responsabilidade da CNEN e tem por objetivo garantir seu uso somente para fins pacíficos.

A área de Pesquisa e Desenvolvimento investe no emprego da tecnologia nuclear em medicina, agricultura, indústria e meio ambiente. Além da produção de radioisótopos e radiofármacos, amplamente utilizados em medicina nuclear, as atividades abrangem os processos e tecnologias em radiodiagnóstico e radioterapia; Fontes industriais de radiação; tecnologia de reatores; estudos sobre neutrônica, operação e manutenção de reatores; desenvolvimento de novos materiais; instrumentação e controle; tecnologia de esterilização e preservação de alimentos por meio da irradiação; ensaios citogenéticos; pesquisas de vacinas por meio da irradiação de venenos; ensaios mecânicos não destrutivos; reagentes; processos de caracterização de bacias hidrológicas e de efluentes líquidos e gasosos; e processos para análise ambiental, apenas para citar algumas.

Como serviços oferecidos pela CNEN destacam-se: a produção de radioisótopos e radiofármacos para a área médica; a formação especializada para o setor nuclear; o

⁶¹ Disponível em: <<http://www.cnen.gov.br>> Acesso em: 29 maio 2014

licenciamento, inspeção e controle de instalações e atividades com materiais nucleares e radioativos; a metrologia das radiações ionizantes; o controle de radioproteção e dosimetria; o recolhimento e armazenamento de rejeitos radioativos; o atendimento às emergências radiológicas e nucleares; e o fornecimento de informações técnico-científicas.

A CNEN é sediada no Rio de Janeiro e atualmente (2014) conta com 14 unidades distribuídas por nove estados brasileiros. Entre essas unidades estão os institutos de pesquisas que buscam avanços em aplicações de técnicas nucleares e de áreas correlatas dando suporte à linha de atuação de Pesquisa e Desenvolvimento. O Instituto de Engenharia Nuclear - IEN é um desses institutos e também está localizado no Rio de Janeiro, mais precisamente no campus da Universidade Federal do Rio de Janeiro – UFRJ, na Ilha do Fundão.

6.2.2 O Instituto de Engenharia Nuclear – IEN

O Instituto de Engenharia Nuclear - IEN⁶², desde 1962 - ano de sua fundação – vem contribuindo, através de suas pesquisas, para o domínio nacional de tecnologias da área nuclear e correlatas. A missão do IEN é “contribuir para o bem-estar da sociedade e seu desenvolvimento sustentável por meio de inovações tecnológicas e formação de recursos humanos para os setores nuclear e correlatos”. Desde 2005 a visão deste Instituto é voltada para "ser um centro de excelência com substanciais contribuições em conhecimento, inovação e transferência de tecnologia para a sociedade". A partir de suas atividades de pesquisa, o IEN gera produtos e serviços como patentes, publicações, licenciamento de tecnologias, fornecimento de radiofármacos, ensaios e análises de materiais, recolhimento de rejeitos radioativos, consultorias e formação de recursos humanos.

Para cumprir sua missão de formação de recursos humanos e, alinhado com sua visão de transferir tecnologia para a sociedade, o IEN conta também com um Programa de Pós-Graduação em Ciência e Tecnologia Nucleares - PPGIEN, que oferece o curso de Mestrado Acadêmico em Engenharia de Reatores Nucleares, preparando graduados em engenharia e ciências exatas para a indústria de geração nucleoeletrica, para a pesquisa e desenvolvimento da tecnologia de reatores e para as atividades de licenciamento de reatores realizadas pela CNEN.

⁶² Disponível em: <<http://www.iem.gov.br>> Acesso em 29 maio 2014.

6.2.2.1 Áreas temáticas de pesquisa do IEN

Em 2012, após se configurar a nova direção do Instituto de Engenharia Nuclear surgiu a necessidade de se organizar de maneira mais sistemática as pesquisas desenvolvidas na instituição, para que o apoio às pesquisas desenvolvidas fosse fornecido de maneira justa às áreas que realmente eram produtivas e/ou que tivessem perspectiva de, a partir deste apoio, se tornarem em alguma espécie de inovação científica e tecnológica para o país.

Assim, de acordo com o Comunicado DIEN⁶³ "Cadernos de Áreas Temáticas de P&D do IEN/CNEN", em 19 de fevereiro de 2013 foi lançada no Instituto uma chamada convocando pesquisadores, tecnologistas e analistas em C & T a identificarem parceiros, a se organizarem em áreas temáticas de atuação e a submeterem suas propostas onde esclarecessem sobre: identificação da proposta, qualificação do principal problema a ser abordado, objetivos e metas a serem alcançados, metodologia a ser empregada, principais contribuições científicas ou tecnológicas da proposta (Impacto científico, Impacto Tecnológico) orçamento detalhado, cronograma físico-financeiro, identificação dos demais participantes do projeto, grau de interesse e comprometimento de empresas com o escopo da proposta, quando fosse o caso, indicação de colaborações ou parcerias já estabelecidas com outros centros de pesquisa na área, disponibilidade efetiva de infraestrutura e de apoio técnico para o desenvolvimento do projeto e estimativa dos recursos financeiros de outras Fontes que serão aportados pelos eventuais agentes públicos e parceiros privados.

O objetivo desta chamada foi

dar tratamento institucional à iniciativa e à criatividade dos profissionais do IEN, incorporando-as à gestão do Instituto e, agregando ao planejamento do IEN a visão de futuro daqueles que lidam na fronteira da geração do conhecimento, seja pesquisando, ensinando ou pensando na inovação tecnológica. (IEN, 2013, p.1)

Uma comissão formada pela Direção do IEN, Chefes de Divisão, Coordenadores e Representantes do IEN nas Câmaras Técnicas da Divisão de Pesquisa e Desenvolvimento (DPD) da CNEN foi designada para avaliar os 19 projetos submetidos, os quais foram julgados a partir de critérios pré-estabelecidos, conforme podem ser visualizados no Quadro 3 a seguir:

⁶³ Comunicados enviados pela diretoria do IEN (DIEN) aos servidores por correio eletrônico.

Quadro 3: Critérios de avaliação das áreas temáticas do IEN

CRITÉRIOS TÉCNICOS	Avaliação
Impacto científico	
Impacto Tecnológico	
Impacto Econômico	
Impacto Social	
Impacto Ambiental	
Clientes do Projeto (efetivos e potenciais)	
Qualificação e Adequação da Equipe Técnica	
Infraestrutura disponível	
CRITÉRIOS ESTRATÉGICOS	
Viabilidade financeira	
Geração de receita própria	
Visibilidade institucional	
Estabelecimento de parcerias	
Aquisição de novas competências técnico científicas no IEN	
Institucionalização do conhecimento	
Aderência às demandas do setor nuclear	

Fonte: SAMPAIO (2013)

Estes critérios foram computados considerando a média e o desvio padrão dos conceitos atribuídos pelos avaliadores, onde: Indicadores Técnico (IT) e Estratégico (IE) são calculados da seguinte forma:

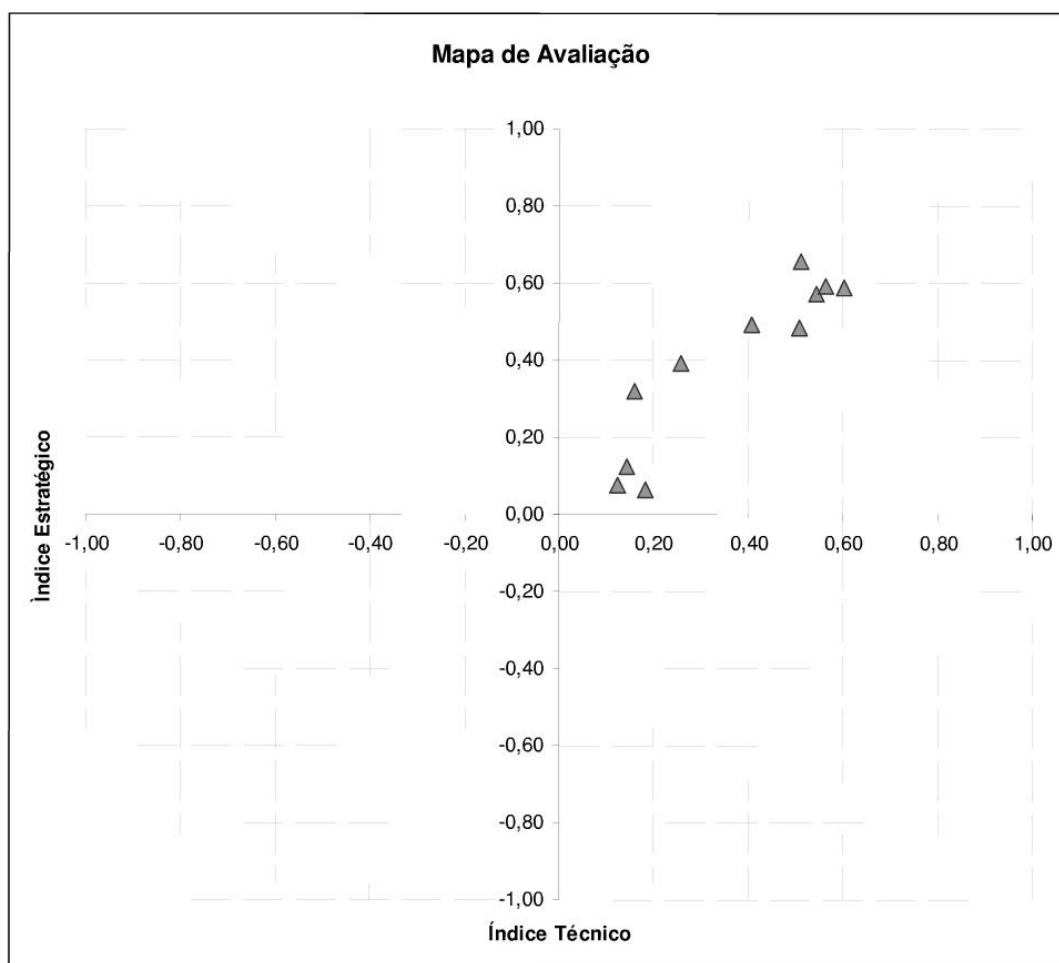
$$IT = (m\acute{e}dia\ t\acute{e}cnica - 3) / (desvpad\ t\acute{e}cnico + 2)$$

$$IE = (m\acute{e}dia\ estrat\acute{e}gica - 3) / (desvpad\ estrat\acute{e}gico + 2)$$

$$-1 < IT < 1 \text{ e } -1 < IE < 1$$

Assim, foram aprovadas automaticamente sete (7) propostas que obtiveram simultaneamente IT e IE maiores que 0,2, conforme reveladas no mapa apresentado na Figura 16 a seguir:

Figura 16: Mapa de Avaliação das áreas temáticas



Fonte: SAMPAIO, 2013

Além das sete aprovadas automaticamente, uma oitava foi admitida após reajustes com a direção. As áreas temáticas aprovadas foram:

- 1) Engenharia e Tecnologia de Reatores Nucleares
- 2) Química Nuclear e Radioquímica
- 3) Desenvolvimento de Instrumentação Nuclear
- 4) Desenvolvimento de Tecnologia para Sistemas Complexos
- 5) Realidade Virtual Aplicada à Área Nuclear

- 6) Aplicação de Técnicas Nucleares na Indústria, Saúde e Meio-Ambiente
- 7) Gestão do Conhecimento Nuclear: Preservação, Disseminação e Compartilhamento do Conhecimento gerado no IEN
- 8) Desenvolvimento e caracterização de materiais funcionais e estruturais para o setor nuclear

As áreas temáticas aprovadas vêm recebendo do IEN o apoio institucional para realização de suas pesquisas. De acordo com a Diretoria do Instituto de Engenharia Nuclear – DIEN, isto se dá “através da liberação de recursos humanos e de infraestrutura do Instituto, respeitadas as condições de pleno atendimento das demandas corporativas da CNEN” (IEN, 2013). Além disso, “as áreas temáticas selecionadas recebem também estímulo do Instituto para buscarem recursos financeiros, seja do próprio orçamento da CNEN e/ou de órgãos externos de fomento à pesquisa, desenvolvimento e inovação.” (IEN, 2013)

Como contrapartida, o IEN requer das áreas temáticas apoiadas o “compromisso de registro institucional de toda a produção técnico-científica gerada e da prestação de contas técnica anual, através de relatório de progresso e de apresentação em seminário institucional.”. Interessante ainda é que a própria Instituição assume que

o registro da produção técnico-científica do Instituto e o compromisso de prestação de contas técnica farão parte da Gestão do Capital Intelectual do IEN, na busca da identificação de formas e meios para projetar a competência técnica existente no Instituto em benefício da sociedade brasileira. (IEN, 2013, p.1).

Para tal registro foi desenvolvido o Repositório Institucional do IEN, ao qual está sendo chamado “Plataforma Carpe diEN”, com a finalidade de armazenamento, preservação e disseminação do conhecimento e ainda de ser uma ferramenta de apoio à gestão do capital intelectual do Instituto.

6.2.2.2 A Plataforma Carpe diEN

O Instituto de Engenharia Nuclear possui uma alta produção técnico-científica entre teses, relatórios, artigos, patentes etc. Essa grande produção de conhecimento gera também uma infinidade de dados que sem um tratamento adequado, muitas vezes se perdem dentro dos computadores pessoais e/ou dos *pen drives* dos pesquisadores.

Em âmbito mais geral, nos últimos anos as simulações computacionais intensas e o processamento de grandes conjuntos de dados tornaram-se uma necessidade na pesquisa de Engenharia Nuclear. Particularmente alguns grandes projetos de pesquisa em Engenharia Nuclear, como por exemplo o do Reator Termonuclear Experimental Internacional - ITER, patrocinado pela Agência Internacional de Energia Atômica – IAEA, que requer uma cooperação internacional entre vários países. Hoje em dia, pesquisadores utilizam todos os tipos de recursos existentes em diferentes lugares, desenvolvem *software*, recolhem conjuntos de dados e realizam experimentos de forma colaborativa. Assim, pode-se afirmar que as metodologias oriundas da *eScience* vêm tendo um papel cada vez mais importante na pesquisa em Engenharia Nuclear.

O IEN tem considerado a importância de ter reunido junto à sua produção técnico-científica os dados gerados em decorrência de suas atividades de ensino e pesquisa, pois acredita que estas informações reunidas podem servir, entre outras coisas, como ferramentas que inspirem novas pesquisas e apoiem a tomada de decisão.

O primeiro movimento em direção à reunião dessas informações foi a criação do Repositório Institucional, denominado **Plataforma Carpe dIEN**, onde pudesse ser depositada a memória técnico-científica do Instituto abrangendo tanto as publicações quanto os dados originados a partir dessas pesquisas registradas nas publicações. O objetivo da plataforma é auxiliar os pesquisadores na gestão e preservação de seus dados de pesquisa, encorajando-os a explorarem e a adotarem novas formas de comunicação científica por meio da utilização de ambientes digitais e de novas práticas de colaboração.

A Plataforma CarpedIEN pode ser vista como uma estratégia para dar ordenação e visibilidade à informação científica produzida pela instituição e é definida como um “repositório voltado para o arquivamento, gestão, preservação e disseminação de dados e informações em formatos digitais gerados em decorrência das atividades de ensino e pesquisa do IEN” (SALES, 2013). A definição de sua política, bem como o modelo de metadados construído para a plataforma considerou em sua essência a preservação e o uso de padrões de tratamento para representação dos documentos técnico-científicos, bem como dos dados de pesquisa. Assim sendo, pode ser considerada uma primeira iniciativa rumo à curadoria digital de dados de pesquisa na Instituição.

Entre os objetivos da plataforma pode-se citar: inserir o IEN nos fluxos nacionais/internacionais promovidos pela interoperabilidade dos repositórios; dar maior visibilidade à produção acadêmica do Instituto (*e-prints* mais visíveis, disseminação rápida e mais citados); aumentar a oferta de serviços de informação mais qualificados para tecnólogos, pesquisadores e alunos; criar um ambiente de interação e troca de idéias entre o corpo de pesquisadores; organizar e aumentar o nível de disponibilidade, acesso e transparência das informações geradas pela instituição; preservar a memória técnico-científica digital; gerar indicadores de produção acadêmica; servir de apoio para tomada de decisão administrativa; mapear o conhecimento produzido; servir de instrumento de apoio para a gestão de conhecimento.

Assim como a maioria dos repositórios institucionais brasileiros, a Plataforma Carpe diEN foi desenvolvida a partir da configuração do *dSpace*, um software livre e de Fonte aberta, desenvolvido pelo Massachusetts Institute of Technology – MIT com a Hewlett-Packard e adotado mundialmente. A escolha desse software se deu, entre outros motivos, por essa ampla utilização, pelo incentivo por parte do IBICT de adoção do mesmo em âmbito nacional⁶⁴ e principalmente por ser um sistema que utiliza o *Open Archives Initiative Protocol for Metadata Harvesting* - OAI-PMH: um protocolo desenvolvido pela Open Archives Initiative, que define um mecanismo para coleta de registros de metadados em repositórios, permitindo que a Carpe diEN seja interoperável com outros repositórios existentes no Brasil e no mundo, o que propicia a disseminação da produção técnico-científica em âmbito global.

Além disso, o *dSpace* permite o depósito dos mais diversos tipos de objetos digitais, o que propiciou que a Carpe diEN fosse configurada também para receber os dados gerados pelas pesquisas desenvolvidas no IEN. Essa configuração se deu a partir da combinação e esquemas de metadados como o Dublin Core – já adotado amplamente para representação de documentos bibliográficos, porém, em seu formato qualificado para atender também os dados de pesquisa – e o Data Cite – esquema utilizado especificamente para representar dados de pesquisa.

⁶⁴ Disponível em: <<http://www.ibict.br/pesquisa-desenvolvimento-tecnologico-e-inovacao/Sistema-para-Construcao-de-Repositorios-Institucionais-Digitais>>. Acesso em: 29 maio 2014.

Contudo, criar um repositório sem desenvolver técnicas e metodologia para tratamento e preservação desses dados seria um fato inútil. Por esta razão, pesquisas vêm sendo realizadas em direção à busca de princípios teóricos e metodológicos que faça desta plataforma um repositório confiável, provedor de serviços de informação sendo incluído na agenda de pesquisa do Instituto como uma ferramenta capaz de gerir o conhecimento produzido.

Essas técnicas envolvem a curadoria digital das publicações e dos dados de pesquisa, tornando a Plataforma Carpe diem um ambiente propício para a criação de publicações ampliadas e um campo empírico favorável para a realização desta tese.

Na Plataforma Carpe diem são armazenados documentos como: produção técnico-científica, dados de pesquisa, material didático, documentos administrativos, documentos históricos, apresentações, entre outros objetos digitais. No entanto, estes documentos são tratados ainda de maneira isolada, isto é, dentro de coleções distintas, sem que documentos de uma mesma temática interajam entre si.

A existência de um novo modelo de documento, como a publicação ampliada, que possa interligar semanticamente os resultados de pesquisas veiculados nas publicações científicas aos dados que fornecem sustentação a essas publicações, bem como a outros documentos é uma idéia que agrada aos pesquisadores, pois é uma forma rápida e precisa de se obter a maior quantidade de informação possível sobre um mesmo tema. Desta forma, ao buscar uma informação sobre “Sistema Kipros para produção do radioisótopo iodo-123”, o usuário pode recuperar e visualizar de forma integrada uma variedade de documentos, dados, metadados, anotações e outros recursos de naturezas distintas, porém, referentes a esta mesma temática.

Além disso, ligar as publicações aos dados de pesquisa é uma forma de registrar a memória da pesquisa desenvolvida bem como dar confiabilidade às investigações, na medida em que os dados replicados podem permitir chegar ao mesmo resultado novamente.

Um levantamento bibliográfico inicial sobre o tema foi realizado na base INIS da IAEA, que é referência para a área Nuclear. No entanto, sobre o tema “publicações ampliadas” nada foi encontrado, o que revela a originalidade desta pesquisa na área de Ciências Nucleares.

O capítulo a seguir apresenta os resultados obtidos nesta pesquisa.

7 RESULTADOS E DISCUSSÕES

Este capítulo tem por objetivo apresentar e discutir os resultados da pesquisa proposta por esta tese. A investigação tem como foco principal a especificação de um novo modelo de publicação acadêmica para a área de Ciências Nucleares que é caracterizado por vincular semanticamente os resultados de pesquisa publicados em um *e-print* aos dados que o originaram.

Entretanto, é preciso observar que a elaboração do modelo demandou uma investigação que extrapolou as fronteiras mais técnicas sobre como os dados e *e-prints* podem ser ligados constituindo as publicações ampliadas. A exigência de estender o estudo decorre do fato de que a gestão de dados de pesquisa – que são elementos essenciais na composição desse novo conceito de publicação - requer ambientes tecnologicamente complexos e políticas e ações de âmbito nacional. Dessa forma, como contextualização ao estudo, tornou-se necessário analisar e alinhar os diversos elementos que compõem o cenário multidimensional que consolida um ambiente de e-pesquisa através das lentes de uma ótica mais abrangente.

Portanto, o que era para ser restrito a um contexto específico, isto é, o da área de Ciências Nucleares, se mostrou relevante como uma contribuição para o estabelecimento de uma política mais genérica de curadoria digital de dados de pesquisa para o país.

Esta contextualização, por si só, acabou por se constituir em uma parte importante da pesquisa. Para uma compreensão mais direta, ela foi expressa também na forma de uma representação gráfica que articula instâncias que devem ser consideradas na formulação de uma política nacional de curadoria digital de dados de pesquisa. Dessa forma, o capítulo inicia-se descrevendo o resultado dessa análise e prossegue apresentando algumas considerações acerca de questões que influenciaram na escolha das entidades e na proposição das etapas de construção do modelo. Nas seções seguintes são descritos os elementos que irão compor o modelo, ou seja, as entidades e as relações; e em seguida são descritas as etapas do processo de construção do modelo e sua explicitação por meio de interfaces; por fim, é apresentado o modelo proposto em forma gráfica.

7.1 DIRETRIZES PARA UMA POLÍTICA DE CURADORIA DIGITAL DE DADOS DE PESQUISA PARA O PAÍS

Considerando diversas ações envolvidas na articulação de um ambiente de e-pesquisa, a análise realizada – à guisa de contextualização - foi estruturada em nove instâncias que refletem aspectos que se articulam para compor uma infraestrutura de dimensão nacional⁶⁵ voltada para a curadoria de dados de pesquisa. As instâncias consideradas são as seguintes: política, infraestrutura organizacional, desenvolvimento de coleções de dados, pesquisa, infraestrutura tecnológica e de padronização, formação de recursos humanos, sustentabilidade econômica, implicações sociais, legais e éticas e disponibilização de serviços. A Figura 17 representa de forma breve essas instâncias e as interações que se dão entre elas.

➤ **Instância Política**

Nos últimos anos, agências de financiamento de pesquisas de vários países e de alguns organismos internacionais vêm introduzindo a exigência de que a gestão de dados de pesquisa e um plano de compartilhamento de dados façam, obrigatoriamente, parte da solicitação de auxílio para os projetos de pesquisa. Ações dessa natureza traduzem o reconhecimento, por parte dos formuladores de políticas de C&T, de que a preservação de dados de pesquisa traz benefícios perceptíveis para a sociedade. A partir dessa constatação é necessário, portanto, o estabelecimento de linhas de ações que assegurem a organização e a governança apropriadas para a atividade de preservação desses estoques informacionais. Além do mais, é imprescindível garantir um fluxo contínuo de recursos destinados à sobrevivência por longo prazo das atividades de curadoria digital.

O que se observa é que a lacuna provocada pela inexistência de políticas coerentes, acessíveis e transparentes de arquivamento e acesso a dados de pesquisa revela-se como barreira para a pesquisa interdisciplinar e para a gestão efetiva de coleções de dados. Por outro lado, um esforço significativo está sendo dirigido, em escala mundial, no desenvolvimento de políticas e diretrizes que ordenem a gestão de dados de pesquisa. Estas iniciativas são levadas a cabo por um amplo espectro de instituições: pelos órgãos

⁶⁵ Este modelo de curadoria digital de dados de pesquisa foi publicado em artigo no n.1, v. 6 da Revista Tendências da Pesquisa Brasileira em Ciência da Informação (TPBCI), da Associação Brasileira de Pesquisa e Pós Graduação em Ciência da Informação - ANCIB, intitulado “Dados de pesquisa: contribuição para um modelo de curadoria digital para o país”. (SAYÃO; SALES, 2013)

nacionais ligados à gestão de C&T, pelas agências de fomento à pesquisa, pelas instituições de pesquisa individualmente, como universidades e centros de pesquisa, e por outros parceiros-chave da comunidade internacional, como os organismos de padronização e organizações ligados ao movimento de livre acesso.

Mas um modelo neutro é um desafio inalcançável no contexto atual. Segundo o relatório do Digital Data Curation de autoria de Ruusalepp (2008), a ausência de um modelo universal voltado para o compartilhamento de dados de pesquisa é um desdobramento direto dos diferentes modelos de financiamento praticados pelos países individualmente. Os fóruns responsáveis pela formulação das políticas de gestão de dados devem atentar para um fato determinante destacado pelo relatório: Por causa das diferenças na gestão, práticas e usos de coleções de dados em diferentes domínios da pesquisa, as políticas nacionais devem permanecer num patamar suficientemente geral para poderem ser efetivamente úteis na prática. (RUUSALEPP, 2008)

A gestão para o acesso e reuso de coleções de dados de pesquisa, portanto, demanda uma infraestrutura de muitas faces, com muitos atores e costurada por compromissos políticos e financeiros duradouros. Subjacente a essa estrutura é necessário o desenvolvimento de um conjunto amplo de ações políticas de abrangência nacional, que estejam, porém, em harmonia com as políticas praticadas pelas principais iniciativas internacionais - incluindo o princípio de livre acesso aos dados de pesquisa e de transparência pública, e que considere as prioridades, as idiossincrasias e as políticas das comunidades científicas e acadêmicas.

Assim, a camada política de um modelo de gestão e compartilhamento de dados de pesquisa para o país deve incluir:

- Fóruns para definição de políticas que tenham a participação de: Gestores de C&T, agências financiadoras de pesquisa (CNPq, CAPES, FAP's), geradores de dados de pesquisa (universidades, centros e institutos de pesquisa), organizações com tradição na área de preservação digital, como o Arquivo Nacional, órgãos responsáveis por aumentar os conteúdos de valor na Internet, como o Comitê Gestor da Internet, sociedades científicas, etc.
- Diretrizes e recomendações sobre padrões e tecnologias para a criação e implantação de rede de repositórios digitais de dados de pesquisa que sejam federados e interoperáveis.

- Linhas de financiamento de pesquisa em áreas de interesse como: preservação e curadoria digital, repositórios digitais, visualização de dados, ambientes colaborativos, metadados etc.
- Documentos estabelecendo diretrizes e estratégias para o desenvolvimento de uma ciberinfraestrutura nacional voltada para o arquivamento, acesso e reuso de dados de pesquisa.
- Exigências para depósito, gestão e disseminação de dados de pesquisa de projetos financiados com fundos públicos.
- Enquadramento da gestão de dados de pesquisa como elemento essencial na formulação de políticas de ciência, tecnologia e inovação.

➤ **Instância organizacional**

Por muitos séculos as bibliotecas e outras instituições de patrimônio intelectual armazenaram continuamente informações para uso corrente e futuro. Este fato moldou a forma como estas instituições foram organizadas e gerenciadas. Hoje, como afirma Pérez-González (2010, p.2), se consolidou uma transformação qualitativa e irreversível. “A criação digital, as novas formas de comunicação em rede e os modelos de consumo da informação digital implica que autores, editores e instituições de pesquisa tenham que enfrentar novas estratégias, políticas e de infraestrutura, que permitam novas formas de gestão”.

Esse desafio é mais contundente quando se pensa em dados de pesquisa, caracterizados pela sua condição heterogênea, dinâmica e distribuída.

A trajetória de desenvolvimento da pesquisa científica, nas condições que hoje se apresenta, faz crer que as instituições acadêmicas precisarão de algum nível de curadoria de dados de pesquisa. Entretanto, é irreal se pensar que cada instituição individualmente poderá estabelecer capacidade local e própria de curadoria digital. Erway e Lavoie (2012) sustentam que a necessidade por especialização em cada área do conhecimento e a necessidade de um largo espectro de conhecimentos técnico e práticas em curadoria, aliadas aos riscos que devem ser assumidos e ao atingimento de uma economia de escala tornam insensata a opção de replicar uma vasta gama de serviços de curadoria, infraestrutura, *expertise*, em cada instituição de pesquisa.

Por outro lado, a diversidade de empreendimentos científicos sugere que uma pluralidade de modelos institucionais e de abordagens de gestão de dados específicos são mais efetivos em atender às necessidades dos usuários (OCDE, 2007), assegurar a qualidade dos dados e a interação entre usuários; entretanto, é necessário observar que a especialização em disciplinas pode levar a uma indesejável compartimentalização que anula um dos benefícios esperados com a curadoria digital que é encorajar a pesquisa interdisciplinar e a interpretação de dados em diversos contextos.

A abordagem nacional adotada por alguns países pode ser viável, dependendo da escala adotada. Por exemplo, a implantação de uma rede interoperável de repositórios de dados de pesquisa pode ajudar na descoberta de coleções relevantes de dados para reuso que podem facilitar a pesquisa multidisciplinar (ERWAY; LAVOIE, 2012). Esta abordagem pode ser aliada, primariamente, a ações colaborativas baseadas na criação de grupos de especialistas em assuntos, que recorrem à *expertise* de um *pool* de especialistas em vários aspectos técnicos de curadoria de dados. O trabalho colaborativo entre especialistas em assunto e em curadoria digital pode assistir a uma coletividade ampla de pesquisadores depositantes de grandes áreas de conhecimento, como Astronomia, Ecologia, Ciências Sociais, Saúde Pública etc., em âmbito nacional.

➤ **Instância de desenvolvimento de coleções de dados de pesquisa**

As bibliotecas de pesquisa e os repositórios digitais têm como um desafio do nosso tempo a tarefa monumental de coletar uma quantidade extraordinária de dados digitais gerados pela pesquisa contemporânea. Entretanto, o chamado “dilúvio de dados” que caracteriza a Big Science, aliado à complexidade e ao alto custo dos processos de curadoria e de preservação de dados, vão exigir que as organizações de pesquisa estabeleçam prioridades sobre o que eles vão finalmente coletar, mesmo diante das dificuldades teóricas e práticas de se operacionalizar conceitos tais como “avaliação de informação”, “valor da informação” e “necessidade de informação”.

Palmer, Weber e Cragin (2011, p.1) enfatizam que a definição dos critérios de seleção de dados de pesquisa “é, num certo sentido, o que os desenvolvedores de coleções nas bibliotecas de pesquisa e nos arquivos sempre fizeram”. Mais explicitamente: julgar que Fontes de informação têm valor suficiente para as suas comunidades-alvo para que se

justifiquem os investimentos em formação de coleção, arquivamento, curadoria e preservação.

O potencial informacional crescente dos dados digitais distribuídos em rede de computadores transforma a visão que caracterizava dados de pesquisa, ainda registrados em mídia impressa, como simples subproduto dos processos de pesquisa. Nesse contexto, os dados só eram considerados na sua configuração final e, via de regra, eram descartados quando os projetos eram concluídos. A tecnologia digital interfere intensamente nas bases dessa ótica de avaliação: muitos tipos de dados científicos devem ser vistos hoje como componentes fundamentais da infraestrutura de sistemas modernos de pesquisa, cujo valor é expandido pelo acesso amplo, pelo seu potencial de reuso e, dessa forma, podem ter um longo ciclo de vida. “O valor do dado aumenta com o seu uso”, enfatiza Uhler (2010).

Sob este ponto de vista, se destaca como maior desafio, quando do estabelecimento de políticas de desenvolvimento de coleções de dados de pesquisa, a definição de métricas e de modelos de avaliação que determinem - ou, de certa forma, predigam - as possibilidades de reuso de um particular conjunto de dados (PALMER; WEBER; CRAGIN, 2011), embora considerando as incertezas decorrentes desta qualificação.

Por outro lado, as ações para aquisição e retenção de dados de pesquisa - dependendo da área de conhecimento, natureza, formato, complexidade desses recursos, para citar algumas características - vão demandar estratégias de formação de coleções, infraestruturas tecnológicas e gerenciais e investimentos em curadoria digital em escalas bastante distintas. Além do mais, os dados necessários para dar apoio a pesquisas mais ativas, em termos da intensidade de uso e de geração de dados, como por exemplo, em Astronomia, exigem coberturas mais seletivas e estratégicas, serviços de preservação e acesso, e, sobretudo, garantia de qualidade e de integridade.

Profissionais das áreas de Biblioteconomia e Ciência da Informação, cujos critérios e princípios de desenvolvimento de coleções são orientados pela avaliação de necessidades de comunidades de usuários, podem efetivamente adaptar suas práticas para a formação de coleções para repositórios de dados.

Fica patente, portanto, a necessidade do desenvolvimento e implantação de modelos teóricos e práticos de avaliação e de desenvolvimento de coleções de dados de pesquisa que venham ao encontro dos objetivos globais de formação de uma rede transversal, robusta, funcional e interoperável, que apoie os desafios da pesquisa científica contemporânea (PALMER; WEBER; CRAGIN, 2011).

Porém, no desenvolvimento de coleções de dados de pesquisa, outros problemas se interpõem. Um dos mais relevantes é assegurar que os dados possam manter a sua capacidade de apresentação e transmissão de informação e conhecimento ao longo do tempo e do espaço.

Disponibilizar os dados na Internet é apenas uma das etapas de um ciclo complexo, e que isoladamente não garante que os dados possam ser acessados, reusados, e, sobretudo, tenham seus significados e estruturas recompostos agora e no futuro. Tendo em vista que os *bits* não falam por si próprios e não impressionam nossos sentidos, para que eles possam manter a sua capacidade de serem interpretados em domínios distintos, transversalmente, é necessário que eles estejam suficientemente organizados e documentados. Dessa forma, torna-se imprescindível que informações contextuais – semânticas e estruturais – acompanhem os dados digitais de forma que eles estejam autodescritos. Isto é efetivado por meio de modelos conceituais de informação, expressos na prática por esquemas de metadados que documentam, por exemplo, os elementos semânticos, as partes dos objetos e suas relações, as dependências técnicas, a proveniência, a identificação persistente, as restrições e direitos associados aos dados, as possíveis intervenções sofridas e seus efeitos. Ou seja, os metadados devem registrar idealmente tudo que deve ser de interesse do usuário, incluindo modelos de dados, equipamentos especiais, especificação da instrumentação, linhagem dos dados e muito mais.

Os metadados cumprem um papel de ponte para o futuro nas estratégias de preservação; além do mais ajudam na presunção de integridade e autenticidade dos dados digitais de pesquisa. A qualidade e precisão dos esquemas de metadados adotados e o rigor da sua aplicação são de crucial importância na garantia de que as coleções de dados possam ser acessadas e os dados usados e reutilizados interdisciplinarmente pelo tempo que for necessário.

➤ **Instância de Pesquisa**

A inserção dos conhecimentos de curadoria na agenda de pesquisa de áreas de conhecimento como Ciência da Informação e Ciência da Computação torna-se essencial para a geração de um corpo consolidado de conhecimento que possa ser debatido em todas as áreas que lidam com intensidade com informações e dados digitais. A fragmentação da pesquisa em curadoria digital, que caracteriza a área de estudo no país, se dá pela

necessidade que alguns domínios de conhecimento, como Medicina e Ecologia, têm em gerir seus dados e extrair significado e viabilizar o reuso. Porém, permanece a necessidade de pesquisas coordenadas e de se criar linhas de investigação interdisciplinares, incentivadas por programas de fomento com perspectiva integradora, que possam gerar conhecimentos teóricos e práticos comuns e também específicos.

Essas ações de pesquisa, nos seus desdobramentos práticos, podem criar as bases para a produção de materiais de referência para a gestão de dados de pesquisa, como manuais, *guidelines*, cursos, normas e padrões, que têm, finalmente, importância crítica para as instâncias tecnológicas, de padronização e de formação de recursos humanos.

Alguns tópicos de uma possível agenda de pesquisa mostram a diversidade e interdisciplinaridade do problema:

- Dispositivos tecnológicos de visualização e compartilhamento de dados de pesquisa; mineração de dados;
- Modelos e técnicas para processamento inteligente e de descoberta de dados por meio de taxonomias e ontologias; integração com os padrões da *Web* semântica e do *linked data*;
- Concepção de novos tipos de publicação acadêmica que considerem vinculações semânticas entre dados e *e-prints*; impactos dessas publicações na comunicação científica;
- Metodologias de gestão de coleções de dados de pesquisa;
- Modelos de custo na implantação de sistemas de curadoria de dados de pesquisa;
- Interoperabilidade e integração de repositórios de dados de pesquisa;
- Impactos éticos e legais, propriedade intelectual, acesso aberto a dados de pesquisa *versus* privacidade.
- Desenvolvimento de esquemas de metadados voltados para a curadoria de dados de pesquisa.
- Avaliação de *datasets* e critérios de qualidade

➤ **Instância de infraestrutura tecnológica e de padronização**

O armazenamento seguro, a recuperação e o acesso a coleções de dados de pesquisa, além da exploração desses recursos por meio de serviços de informação e de aplicações computacionais – como, por exemplo, mineração e visualização de dados –, exigem um conjunto de tecnologias e de padrões apropriados provenientes, em maior escala, da Tecnologia de Informação e da Ciência da Informação. De igual importância são as normas e padrões que permeiam as ações de preservação e de curadoria digital e os vários níveis de interoperabilidade entre repositórios de dados e informações de pesquisa. Normas, padrões e protocolos, além de *hardware*, *software* e infraestrutura de rede se tornam essenciais na composição de ambientes de alta tecnologia conhecidos como “ciberinfraestrutura”, que tem como objetivo mais geral a integração de serviços e recursos distribuídos para arquivamento, acesso e visualização.

Compreende-se por ciberinfraestrutura, como nos esclarece Pérez-González (2010, p. 3), “uma nova forma de cultura científica que se sustenta em uma robusta infraestrutura tecnológica de alto nível”. Os dispositivos oferecidos por essa infraestrutura dão apoio a mecanismos inéditos de colaboração, baseados no acesso a uma quantidade extraordinária de dados, recursos informacionais interpretados e reutilizados por potentes ferramentas de observação, visualização e simulação. Uma ciberinfraestrutura “é um meio que permite acesso e circulação de conhecimento distribuído, em que colaboram e se comunicam diferentes comunidades e disciplinas, rompendo fronteiras culturais, geográficas e temporais”, complementa Pérez-González (2010).

- Em torno desta questão cabem algumas ações práticas na direção da formulação de uma política de gestão de dados de pesquisa:
- Definição de um elenco de normas, padrões e protocolos de especificações abertas, de aceitação internacional.
- Estabelecimento de ambientes de ciberinfraestrutura de abrangência nacional;
- Integração das ações já em andamento por instituições brasileiras vocacionadas para o problema.

➤ **Instância de formação de recursos humanos**

“A Sustentabilidade humana é crítica para assegurar continuidade e consistência ao longo do tempo de serviços de curadoria de dados de pesquisa”, afirmam Mayernik e seus colaboradores (2012, p.12). Isto nos indica que estruturas educacionais e de recompensa apropriadas são componentes necessários para a promoção das práticas de acesso e compartilhamento de dados. Essas considerações se aplicam a quem financia, produz, gerencia e usa dados de pesquisa (OCDE, 2007).

O problema de coletar, organizar, indexar, arquivar e disseminar grandes coleções de dados – embora não seja um problema novo – é amplificado de forma extraordinária no ambiente da *eScience*. Curadores de dados provenientes das bibliotecas especializadas, dos arquivos e de setores da tecnologia da informação são capazes de gerir, inserir nos sistemas e preservar coleções de dados de pesquisa. Entretanto, os especialistas em assunto é que serão capazes de realizar as análises necessárias à reinterpretação e reuso dessas coleções. Isso significa que é necessário compor equipes de curadoria que articulem dinamicamente *expertises* de natureza distinta.

A necessidade de profissionais de informação multidisciplinares, que conjuguem conhecimento de áreas científicas e de engenharias, com conhecimento de biblioteconomia, ciência da informação e informática, delineia uma nova classe profissional, chamada por alguns autores de “profissional de *eScience*” (STANTON, 2011) ou ainda “cientista de dados”, cuja missão é resolver problemas de gestão de informação em larga escala para pesquisadores com o uso de ferramentas inovadoras.

Considerando a extrema variação dos dados, os ambientes mais efetivos de gestão de curadoria são aqueles que permitem uma troca dinâmica de *expertise*, práticas e conhecimentos entre membros da equipe. “O compartilhamento de *expertise* desempenha um papel central nas operações em curso e no desenvolvimento de qualquer solução em curadoria de dados” (MAYERNIK et al., 2012, p. 12). Nessa direção, profissionais sofisticados de gestão de dados permitem que pesquisadores pratiquem uma ciência melhor, e ainda tornam possível que os profissionais de tecnologia da informação criem infraestruturas mais confiáveis, mais produtivas e mais eficazes, criando uma ponte entre os vários domínios. A capacidade de traduzir as necessidades de informação do cientista em ferramentas da ciberinfraestrutura torna-se uma função essencial no fluxo gerido por este novo profissional de informação (STANTON, 2011, p. 91).

Como não há capacitação formal nessa área, os profissionais de gestão de dados terão que construir seus conhecimentos, ao longo do tempo, no trabalho cotidiano de curadoria e de articulação com áreas finalísticas. Dessa forma, se tornarão capazes de oferecer treinamento para novos usuários e novos profissionais de curadoria. Entretanto, é necessário estabelecer meios para acumulação, sistematização e disseminação desses novos conhecimentos, e também uma articulação direta com as instâncias preocupadas com a pesquisa na área de curadoria e preservação de dados. Parece bastante natural que as demandas por profissionais da *eScience* recebam acolhidas dos cursos tradicionalmente vocacionados para tal como Biblioteconomia, Arquivologia e Ciência da Informação, ressaltando-se que a Ciência da Computação já deu passos importantes nessa área.

➤ **Instância de Sustentabilidade econômica**

Considerações sobre a persistência do acesso aos dados de pesquisa, na sua condição de elemento chave nas infraestruturas nacionais e internacionais de pesquisa – principalmente em relação aos dados financiados por recursos públicos - não podem ser avaliadas como extensões ou algo acessório nos projetos e programas de pesquisa. A facilitação do acesso, a gestão e a preservação desses dados requerem planejamentos orçamentários específicos e suporte financeiro apropriado. Essa constatação tem origem na própria natureza da curadoria digital que é um processo que se desenrola indefinidamente no tempo e no espaço; isto implica que o fluxo de fundos para a curadoria deve se compatibilizar com o ritmo dessa continuidade, o que parece óbvio, mas que na prática é frequentemente negligenciado.

Dessa forma, além da possível diversidade de arranjos dos vários atores envolvidos na pesquisa científica, persiste como condição crítica para um futuro de longo prazo para os dados de pesquisa o reconhecimento de que a alocação contínua de recursos é um passo fundamental para os processos de curadoria. “Na ausência desse reconhecimento, o objetivo de manter por longo prazo o acesso a dados de pesquisa de qualidade não será alcançado” (ERWAY; LAVOIE, 2012 p.3).

Entretanto, assegurar a sustentabilidade econômica de conjunto de dados de pesquisa – e os serviços gerados a partir deles – ultrapassa a mera alocação de recursos. Na opinião de Erway e Lavoie (2012), o processo envolve a utilização eficiente destes recursos e a alavancagem de parcerias e colaboração no sentido de se alcançar uma

economia de escala. Isto pode significar na prática que o estabelecimento de arranjos institucionais abrangentes e organicamente comprometidos seja essencial na sustentabilidade das coleções de dados de pesquisa de valor contínuo.

É necessário enfatizar ainda que modelos de custo sustentáveis para serviços de curadoria ou mesmo de preservação digital não são ainda bem entendidos, e não há na literatura da área formas e metodologias padronizadas para a condução dos processos de curadoria. Em termos mundiais, diferentes organizações adotam diferentes modelos financeiros.

Em termos práticos e mais imediatos, constata-se que o sucesso da implementação e operação de qualquer serviço de curadoria de dados digitais de pesquisa vai exigir uma análise minuciosa de todos os custos conhecidos e esperados para o futuro imediato, combinados com estratégias que assegurem a cobertura desses custos de forma contínua.

➤ **Instância social, legal e ética**

Há um consenso nítido de que entre as principais barreiras sociais, éticas e legais interpostas entre as comunidades interessadas e o pleno acesso aos dados de pesquisa, está um quadro deficiente de proteção ao direito de propriedade intelectual, a dificuldade de documentar os dados para reuso e os problemas associados com a proteção da confidencialidade e privacidade. Há ainda uma tensão latente e não resolvida entre confidencialidade e abertura dos dados.

A legislação de cada país e os acordos internacionais, particularmente em áreas como direitos de propriedade intelectual e proteção da privacidade, afetam diretamente o acesso aos dados de pesquisa e as práticas de compartilhamento, e devem ser profundamente considerados no projeto dos acordos de acesso de dados (OECD, 2007).

No ambiente acadêmico tipicamente não se reconhece completamente os direitos de propriedade intelectual relativos à produção e compartilhamento de dados. Faltam mecanismos de atribuição de crédito e de recompensa, de tal forma que o pesquisador que oferece abertamente seus dados para seus pares possa ser citado e reconhecido como autor em qualquer situação e publicação que faça uso dos dados gerados por suas pesquisas. Na

direção dessa demanda, o Data Cite⁶⁶ estabelece formas padronizadas de citação de dados e coleções de dados.

➤ **Instância de serviços**

O acesso às coleções de dados de pesquisa, na forma de serviços convencionais e inovadores, dirigidos a segmentos variados de usuários, devem fazer parte das políticas de gestão de dados na qualidade de objetivo essencial. Além das facilidades tradicionais – como busca avançada, disseminação seletiva e *browsing* – os dados devem estar preparados para serem capturados por aplicações computacionais que proporcionem novas análises, estatísticas, indicadores e sirvam também de *input* para, por exemplo, sistemas de apoio à decisão e sistemas educacionais. É necessário ainda que as ciberinfraestruturas possam oferecer diferentes modalidades de interoperabilidade, como via OAI-PMH, OAI-ORE e *Linked Data*. As interfaces para apresentação dos dados – preferencialmente via portais *Web* - cumprem um papel importante na otimização do acesso, uso e reuso dos dados. Nessa direção, as representações baseadas em tecnologias semânticas, taxonomias e ontologias tornam-se metodologias relevantes na descoberta de recursos.

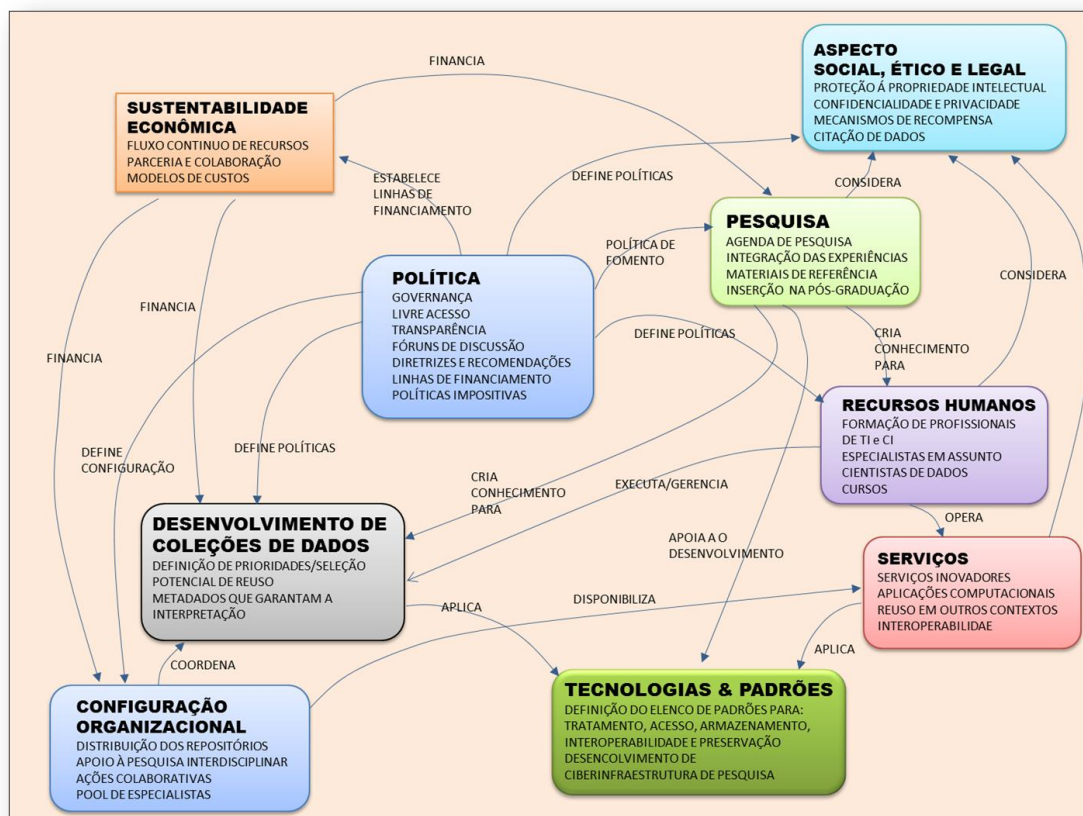
O oferecimento de serviços baseados em coleções de dados de pesquisa amplia o escopo de atuação das bibliotecas de pesquisa e as recoloca no centro dos acontecimentos. Entretanto, esse novo papel impõe grandes desafios no delineamento de novos fluxos de trabalho e na implantação de infraestruturas tecnológica e gerencial para essas bibliotecas. Além do mais, um monitoramento regular se torna essencial, posto que novos conceitos de dispositivos informacionais para acesso e distribuição de informações de pesquisa estão permanentemente surgindo. Um dos mais importantes atualmente, como já visto na seção 2.8, é o CRIS – sigla para Current Research Information System. Um CRIS consiste basicamente num modelo de dados descrevendo objetos de interesse para as atividades de pesquisa e um conjunto de ferramentas para a gestão de dados. O objetivo do sistema é assistir o usuário em todos os processos de pesquisa, incluindo alocação de recursos, avaliação de projetos, identificação de novos mercados para produtos de pesquisa, análise de tendências e muito mais.

Essas camadas propostas para a elaboração do modelo de curadoria digital estão relacionadas, como podem ser verificadas na Figura 17 a seguir. Esta visualização

⁶⁶ Disponível em: <www.datacite.org>. Acesso em: 5 set. 2013

integrada permite compreender a importância de abranger todos os aspectos na elaboração de um modelo para o país.

Figura 17: Proposta de Modelo de Curadoria Digital para o país



Fonte: A autora

A proposição de uma política de curadoria digital para o país se justifica dentro deste trabalho pelo fato de que para a criação de uma publicação ampliada é necessário que haja a curadoria de dados digitais de pesquisa não apenas dentro de uma Instituição, mas em outras instituições também. A publicação ampliada, por ser um objeto digital complexo, pode ser construída a partir do relacionamento com dados disponibilizados em outros repositórios institucionais ou temático. No entanto, para que isso seja possível é preciso que esses repositórios ou que a curadoria dos dados digitais de pesquisa utilizem em seus processos de construção ou submissão normas e padrões comuns, tornando todas essas informações interoperáveis. Isso vai além do alcance institucional sendo necessário, portanto, uma intervenção nacional.

Se por um lado a política nacional ainda é apenas uma proposta, por outro, as instituições, bem como grupos de pesquisa no país, começam a se mobilizar para apoiar a

construção ou adaptação de seus repositórios e a discutir os parâmetros para disponibilização, acesso e disseminação desses dados. É neste sentido que a proposta desta tese se enquadra, isto é, no sentido de propor uma forma de preservar e disseminar esses dados dentro do seu contexto, ou seja, integrado às publicações que disseminam seus resultados. Assim, as seções a seguir vêm explicitar o resultado principal desta pesquisa que é a proposição de um modelo de publicação ampliada que integre semanticamente publicações e dados de pesquisa.

7.2 CONSIDERAÇÕES ACERCA DO MODELO DE PUBLICAÇÃO AMPLIADA PROPOSTO

O processo de construção do saber científico estabelece um espectro amplo de relações entre os produtos de pesquisa gerados nos diversos fluxos de produção científica. Por exemplo, dados de pesquisa e os diversos tipos de publicações acadêmicas e científicas decorrentes da atividade científica guardam entre si significações informacionais ricas. Além disso, as relações ocorrem também entre os diversos atores envolvidos no ciclo de geração de conhecimento, como, por exemplo, pesquisadores, gestores e instituições científicas e de fomento à pesquisa; e ainda entre os objetos de pesquisa e esses mesmos atores.

A concepção dos sistemas de informação para a pesquisa, especialmente os moldados pelo padrão CRIS – Current Research Informations System – exploram o conceito de relações como um recurso relevante para todo o ciclo de gestão da pesquisa. O Relatório CERIF deixa clara essa importância quando define informação para a pesquisa como “a informação sobre entidades de pesquisa, tais como pessoas, projetos, organizações, publicações, patentes, produtos, fomento, etc e o **relacionamento** (grifo nosso) entre eles” (EUROCRIS, 2012a, p.2). As relações relevantes e flexíveis, a possibilidade de intercâmbio em formatos padronizados XML e a especificação de uma matriz semântica para essas relações constituem instrumentos poderosos para a formulação de sistemas de informação escaláveis e orientados à qualidade.

Há, portanto, uma profícua interlocução entre todas as entidades existentes no mundo da pesquisa, o que torna esses diálogos recursos importantes de informação e objetos de pesquisa de diversas áreas da Ciência da Informação, especialmente da

Comunicação Científica. Entretanto, para que se tornem objetos de informação e que possam fazer parte das infraestruturas de informação voltadas para a pesquisa, esses recursos precisam ser formalizados, têm que serem passíveis de processamento por computador e, finalmente, serem oferecidos na forma de serviços pelos sistemas de informação existentes ou pelos sistemas mais inovadores. Isto significa dizer que os sistemas de informação precisam permitir que esses novos recursos sejam estruturados, armazenados, preservados, intercambiados, disseminados e reusados.

Algumas das relações que são objeto do presente estudo são claras, bem conhecidas e tem os seus significados imediatamente reconhecidos dentro de um domínio específico. “Autoria” e “orientação” são exemplos de relações explícitas cujos significados que transmitem são claramente conhecidos no meio acadêmico e científico. Porém, diversas relações possíveis de se estabelecerem entre os objetos de pesquisa e entre atores do universo de pesquisa, ricas de significados, não estão explicitadas, padronizadas e, conseqüentemente, não são facilmente identificadas, ou não têm os seus contornos semânticos bem delineados.

Essas relações não formalizadas existem de forma subjetiva no universo abstrato dos pesquisadores, mas não se revelam facilmente e não expressam informação e conhecimento. Do ponto de vista prático, não podem ser registradas, automatizadas e incorporadas aos sistemas de informações como OPACS e às redes de repositórios digitais. Como desdobramento direto, essas relações não reveladas não podem ser caracterizadas como recurso informacional e não podem ser reutilizadas pela comunidade de pesquisadores. De uma maneira abreviada, um dos problemas em que o presente estudo se debruça é a formalização e a padronização semântica das relações que ocorrem no universo de pesquisa da área de Ciências Nucleares, representado pelo IEN.

No mundo digital, onde está ancorada, atualmente, a maioria das publicações científicas, a oferta de acesso *on-line* a artigos é acompanhada normalmente de *hyperlinks* para referencias, citações e para dados suplementares – por exemplo, uma página *Web* de uma instituição mencionada. Entretanto, a conexão com redes sociais, como *blogs*, a relação com outros materiais, como multimídia e dados de pesquisa e com contextos semânticos, por exemplo, via estruturas em XML, não é comumente concretizada nos dias de hoje.

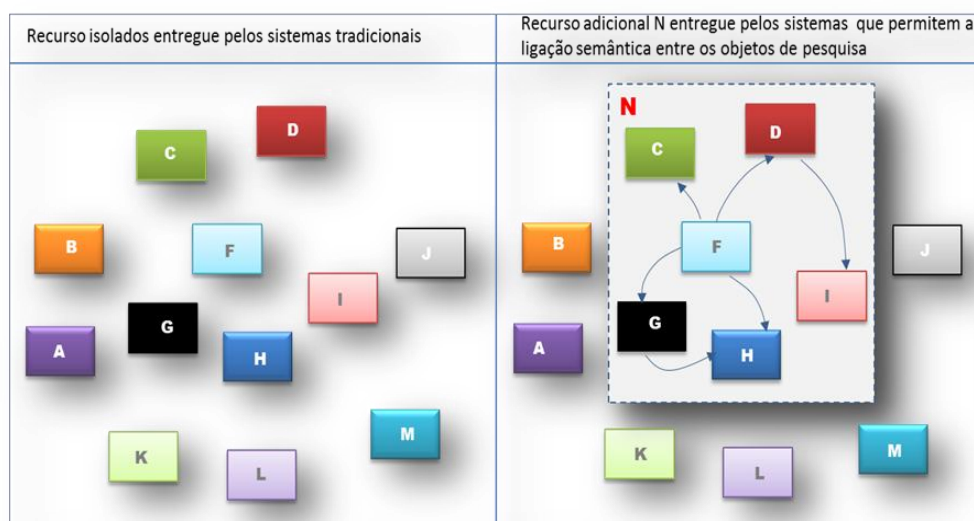
Publicações e objetos relacionados são processados separadamente como objetos individuais, dessa forma, a conexão entre eles não é fácil de ser identificada. Como as

relações entre objetos de pesquisa não são oferecidas pelos sistemas de informação atuais, fica difícil de constatar se objetos relacionados existem e estão disponíveis, argumentam Woutersen-Windhouver e Brandsma (2009, p.21) no Relatório do SURF Foundation.

Enfatizando o que já foi discutido nos capítulos anteriores, não obstante a apropriação de tecnologias avançadas pelos sistemas de informação para a pesquisa, eles continuam entregando para o usuário como resultado final uma publicação digital- que é um fac-símile da publicação impressa tradicional – isolada e independente dos outros objetos digitais ao qual está vinculada objetiva ou subjetivamente. Por exemplo, uma publicação em formato PDF. Essa concepção ignora e não consegue expressar a diversidade e a riqueza de conteúdo dos produtos de pesquisa – por exemplo, dados e conjunto de dados, programas de computadores, anotações, vídeos – ou seja, não conseguem dar conta da complexidade das relações que se estabelecem entre as entidades intervenientes no processo de geração de conhecimento científico.

Um modelo de sistema de informação que considere as vinculações semânticas entre as diversas entidades é “fortemente demandado pela comunidade, desde que, tipicamente, os cientistas não produzem unicamente textos (artigos, livros, etc), mas também relacionamentos científicos entre objetos de pesquisa”, enfatiza Parinov (2013, p.17), que vai além e argumenta que as técnicas existentes para visualizar tais relacionamentos – por exemplo, os mecanismos de citação – ainda não estão suficientemente avançados para os desafios da moderna informação científica e tecnológica (PARINOV, 2013, p.17).

O modelo proposto por esta tese argumenta a favor de trazer os padrões de vinculações semânticas praticadas no ambiente CRIS para as publicações científicas e acadêmicas que, dessa forma, podem ser integradas semanticamente com outros objetos de pesquisa e podem explicitar também os papéis de cada ator envolvido. A Figura 18 mostra de uma maneira simples que a vinculação semântica entre elementos do universo de pesquisa é um mecanismo que permite conferir aos elementos, antes isolados, novas interpretações e transmissão de novos conhecimentos, concretizando um novo recurso de informação.

Figura 18: Recursos informacionais isolados versus recursos ligados semanticamente

Fonte: A autora

A visualização das múltiplas relações semânticas entre os objetos de pesquisa permite novas inferências e aumenta o potencial de reuso desses recursos; além do mais, destaca as características essenciais dos objetos de pesquisa e atores envolvidos nas atividades de um determinado domínio de pesquisa, definindo contornos para o estabelecimento de padrões de agregações para esse domínio específico, como, por exemplo, para a área de Engenharia Nuclear, campo empírico do presente estudo.

É importante que essas relações possam ser expressas nas publicações ampliadas, estendendo a capacidade de transmissão de informação e conhecimento das publicações tradicionais. Para tal é necessário um grau mínimo de padronização que possa contribuir para a construção de publicações ampliadas bem formadas, que possam ser identificadas, recuperadas, intercambiadas, armazenadas e preservadas.

Além do mais, é desejável que os sistemas de informação possam oferecer ferramentas simples para autoria de publicações ampliadas e para registros de relações padronizadas, que permitam que o pesquisador possa trabalhar com os diversos objetos de pesquisa delineando novos constructos, e que tenha a autoria dessa nova publicação e das relações reconhecidas.

Essas relações podem estar pré-definidas e disponíveis na forma de taxonomias exibidas por interfaces *Web* para que pesquisadores e gestores de sistemas de informação possam mais facilmente compor agregações com semânticas bem definidas. Isso possibilita que os cientistas possam trabalhar com os objetos de pesquisa como um jogo de blocos de

montar, onde é possível criar publicações ampliadas, expressando opiniões, estabelecendo novas formulações de hipóteses, novos objetos intelectuais capazes de serem compartilhados, delineando novas ideias de pesquisa e de problemas reais.

Por fim, é importante que essas agregações semânticas possam ser entregues e apresentadas ao pesquisador na forma de interfaces avançadas, interativas e navegáveis que explorem a complexidade e a riqueza informacional do recurso e permita o intercâmbio via protocolos padronizados com outros sistemas.

Um dos conceitos essenciais na definição do modelo proposto é a noção de *link semântico*, cuja aplicação em sistemas de informação permite que pesquisadores possam expressar seus conhecimentos, hipóteses e opiniões a cerca da relação entre qualquer par de objetos de informação de pesquisa armazenados em um repositório digital local ou em repositórios digitais geograficamente distribuídos.

A implementação por meio de sistemas automatizados independentes ou por camadas de aplicações sobre repositórios digitais estabelece um ambiente de colaboração, de reuso dos objetos de informação de pesquisa e de criação de novas formulações de publicações.

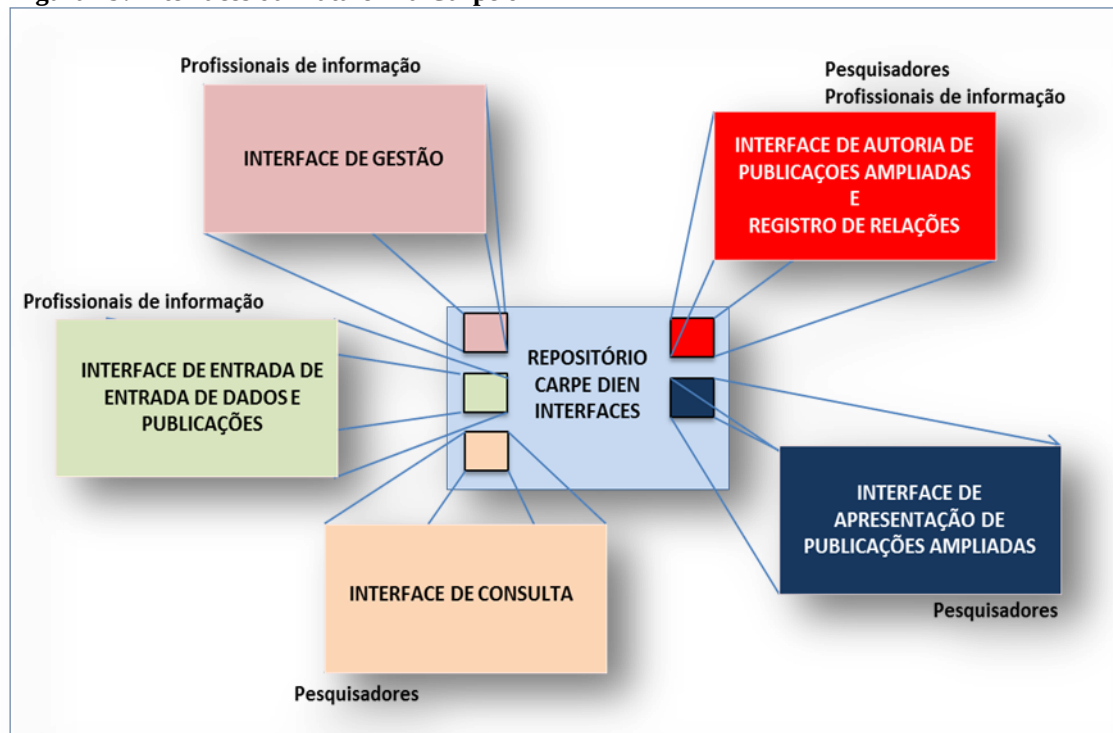
Para formalizar as relações que podem existir entre os objetos de pesquisa – e os colocarem de forma mais adequada ao processamento automatizado – são utilizadas como padrão as ontologias científicas e vocabulários semânticos já existentes, ampliados com elementos específicos para a área nuclear. Isto significa dizer que relações pré-definidas – dispostas em estruturas organizadas de vocabulários semânticos – estarão disponíveis como parte dos serviços do sistema de informação e creditadas como recurso informacional. Além do mais, novas relações que representem conexões não previstas poderão ser dinamicamente registradas pelos pesquisadores.

Nessa direção, as interfaces de interação com o sistema de informação – instanciada, no caso do presente trabalho pelo repositório de informações e dados de pesquisa, Plataforma Carpe diem – contarão com mais duas opções, além das tradicionais interfaces de consulta, de gestão e de entrada de dados: interface de autoria de publicação ampliada, que inclui dispositivo de registros de relações, e interface de visualização de publicação ampliada.

Esses serviços adicionais, consubstanciados pelas interfaces, são especificados e prototipados como parte dos resultados da presente pesquisa. A Figura 19 ilustra de forma simplificada a arquitetura desse repositório “ampliado” com essas novas interfaces, que

dessa forma se aproxima de ambiente virtual de pesquisa definido pelo padrão CRIS – Current Research Information System.

Figura 19: Interfaces da Plataforma Carpe diEN



Fonte: A autora

O ambiente de informação definido pelo CRIS mantém uma visão global dos resultados de pesquisa gerados no âmbito de uma instituição acadêmica, que são registrados de forma padronizada. Dessa forma, os sistemas de informação internos podem disponibilizar visões possíveis das atividades de pesquisa do IEN. Estes sistemas oferecem, entre outras coisas, detalhes sobre: autores, unidades da organização, *links* profissionais de autores de artigos, ofertas de apoio financeiro, projetos, parceiros, resultados de pesquisa.

Na direção da formalização do campo empírico e da adequação de seus requisitos à formulação de publicações ampliadas apropriadas à atividade de pesquisa do IEN, e à especificação de um ambiente informatizado que possam gerir essas publicações, foi necessário o levantamento de alguns dados que caracterizassem as áreas de atuação do IEN, seus fluxos, interseções e, sobretudo, os dados gerados por essas atividades.

No cerne desse levantamento está o mapeamento das atividades de pesquisa do IEN e o potencial de reuso dos seus dados de pesquisa nos diversos contextos do Instituto; a

possibilidade de uma gestão mais dinâmica e a de curadoria digital desses dados, tendo em vista que a geração de publicações ampliadas os tornam mais visíveis, perenes e acessíveis. O que se constata é que de outra forma os dados estariam dispersos e sujeitos à fragilidade intrínseca das mídias digitais e à obsolescência tecnológica. Ligar os dados de pesquisa às publicações é uma forma de registrar o andamento de atividades de pesquisa do IEN, manter a memória científica institucional viva e ainda promover gestão do conhecimento produzido na instituição.

O que pode ser observado é que o armazenamento e a catalogação dos dados de pesquisa em repositórios, como a Plataforma Carpe diEN, formalizam os dados como recurso informacional relevante, tendo em vista que o tratamento técnico os contextualizam permitindo que os mesmos passem do status de objetos digitais para status de objetos de informação, podendo ser utilizados, disseminados e reutilizados.

7.2.1 Mapeamento das atividades de pesquisa do IEN

A área de Ciências Nucleares é formada por um amálgama de disciplinas técnicas e científicas que se combinam inter e multidisciplinarmente formando um domínio amplo de conhecimentos científicos, tecnologias, práticas e padrões. As contribuições que são incorporadas cotidianamente pelos pesquisadores nucleares ao patrimônio de saberes científicos da humanidade - não obstante as controvérsias em torno da percepção de risco da energia nuclear – se desdobram em inúmeras aplicações que têm impactos em todos os segmentos importantes da sociedade, como por exemplo, nos setores de produção de energia, saúde, agricultura, meio-ambiente, indústria, entre muitos outros.

A riqueza, a diversidade e a amplitude da área nuclear estão refletidas também nas atividades de pesquisa do IEN.

A Engenharia Nuclear, foco principal do Instituto, reforça e autentica a pluralidade de possíveis abordagens. Como evidência desse fato, verifica-se que o Instituto atua em diversas frentes de pesquisa com diferentes objetivos, metodologias e propósitos. Essa parece ser, portanto, por sua natural diversidade, uma área propícia para o reuso de dados e informações de pesquisa e aplicação de conceitos de publicações mais avançadas.

A caracterização das atividades de pesquisa do IEN - que inclui o mapeamento dos temas de pesquisa e a identificação dos produtos de pesquisa gerados por essas atividades - foi delineada tendo como base duas estratégias principais:

- Questionário aplicado por meio de entrevista dirigido aos líderes das áreas temáticas;
- Análise dos assuntos presentes nos cadernos de áreas temáticas. Isto foi realizado por meio da metodologia de indexação usada para tratamento de documentos técnico-científicos que são incorporados na base de dados do International Nuclear Information System – INIS, da Agência Internacional de Energia Atômica - AIEA, principal base de dados da área de Ciências Nucleares. Para isso foi utilizado como instrumento terminológico o tesouro do INIS e *expertise* de um indexador especialista na área nuclear.

Somam-se às estratégias formais acima o conhecimento da autora sobre os fluxos de pesquisa do Instituto e a experiência adquirida na atuação por quase dez anos como bibliotecária, líder do grupo de pesquisa em gestão do conhecimento nuclear e do projeto do repositório de dados Carpe dIEN.

A principal motivação desses levantamentos foi reunir conhecimento sobre o IEN, seus fluxos e seus produtos de pesquisa com vistas à modelagem das relações semânticas a serem incorporadas às publicações ampliadas voltadas para o Instituto. Mais formalmente, a intenção desse mapeamento foi levantar os tipos de produção bibliográfica produzidos pelo Instituto; levantar os dados de pesquisa gerados pelas atividades de pesquisa do Instituto; estimar a potencialidade de reuso dos dados de pesquisa por diferentes áreas temáticas do Instituto; caracterizar as áreas temáticas do Instituto; definir as entidades que comporão o modelo de publicação ampliada e de *links* semânticos; estimar o nível de preservação digital aplicado aos dados pesquisa;

7.2.1.1 Mapeamento dos temas abordados nas pesquisas desenvolvidas no IEN

Os cadernos de áreas temáticas do IEN registram os projetos em desenvolvimento no Instituto propostos pelos próprios pesquisadores, conforme descrito no item 6.2.2.1 deste trabalho. A indexação dos cadernos de áreas temáticas – usando o padrão da base de dados INIS - fornece uma representação e um mapeamento dos assuntos tratados pelo

Instituto como objetos de pesquisa com um grau de fidedignidade adequado ao propósito do presente estudo. Como já mencionado em capítulo anterior, as áreas temáticas são as seguintes:

- Engenharia e tecnologia de reatores nucleares
- Radioquímica e química nuclear
- Realidade virtual aplicada na área nuclear
- Desenvolvimento e caracterização de materiais funcionais e estruturas para o setor nuclear
- Desenvolvimento de instrumentação nuclear e sistemas de controle
- Desenvolvimento de tecnologia para sistemas complexos
- Gestão do conhecimento nuclear
- Aplicação de técnicas nucleares na indústria, saúde e meio ambiente.

Os resultados do procedimento de indexação das áreas temáticas são apresentados no Quadro 4, que apresenta os descritores controlados de acordo com o Tesouro do INIS grafados em letras maiúsculas e os descritores livres usados para representar assuntos não constantes no Tesouro grafados em letras minúsculas.

Quadro 4: Assuntos cobertos pelas áreas temáticas do IEN

ÁREAS TEMÁTICAS			
ENGENHARIA E TECNOLOGIA DE REATORES NUCLEARES	RADIOQUÍMICA E QUÍMICA NUCLEAR	REALIDADE VIRTUAL APLICADA NA ÁREA NUCLEAR	DESENVOLVIMENTO E CARACTERIZAÇÃO DE MATERIAIS FUNCIONAIS E ESTRUTURAS PARA O SETOR NUCLEAR
computational fluid dynamics ECONOMICS FLOW RATE FLOW VISUALIZATION FLOWMETERS HEAT RECOVERY HEAT TRANSFER NATURAL CONVECTION NUCLEAR ENGINEERING NUCLEAR REACTORS REACTOR CORES REACTOR SAFETY REACTOR TECHNOLOGY	BIOLOGY CHEMICAL ANALYSIS DECONTAMINATION EDUCATION EQUIPMENT INTERFACES GEOCHEMISTRY HYDROLOGY INDUSTRY NUCLEAR CHEMISTRY MAN-MACHINE SYSTEMS RADIOCHEMISTRY SAMPLE PREPARATION SIMULATORS	CONTROL ROOMS EDUCATIONAL TOOLS EDUCATION EQUIPMENT INTERFACES HUMAN FACTOR ENGINEERING INDUSTRY LEARNIG NUCLEAR REACTORS PHYSICAL PROTECTION RADIATION PROTECTION RADIOACTIVE WASTE STORAGE	CERAMICS FUEL CYCLE LIQUID WASTES MEMBRANES NUCLEAR FUELS POLYMERS RADIOACTIVE WASTE STRESS ANALYSIS URANIUM DIOXIDE ULTRASONIC WAVES ULTRASONIC TESTING

REACTOR LICENSING SIMULATION SITE SELECTION TRANSIENTS TRANSPORT THEORY TWO-PHASE FLOW ULTRASONIC WAVES ULTRASONIC TESTING	SIMULATION STATISTICS RADIATION SOURCES	RADIATION MONITORING RADIATION ACCIDENTS SIMULATION Virtual reality	
ÁREAS TEMÁTICAS			
DESENVOLVIMENTO DE INSTRUMENTAÇÃO NUCLEAR E SISTEMAS DE CONTROLE	DESENVOLVIMENTO DE TECNOLOGIA PARA SISTEMAS COMPLEXOS	GESTÃO DO CONHECIMENTO NUCLEAR	APLICAÇÃO DE TÉCNICAS NUCLEARES NA INDÚSTRIA, SAÚDE E MEIOAMBIENTE
CONTROL SYSTEMS DIGITAL SYSTEMS DOSEMETERS FLUORIMETERS FUEL CYCLE EQUIPMENT INTERFACES MEASURING INSTRUMENTS multipurpose reactors NUCLEAR MEDICINE PROCESS CONTROL RADIATION MO NITORS RADIATION PROTECTION RESEARCH REACTORS SPECTROSCOPY	ACCIDENTS ACCIDENT MANAGEMENT ARTIFICIAL INTELLIGENCE CONTROL ROOMS EMERGENCY PLANS EQUIPMENT INTERFACES EVACUATION ERRORS HUMAN FACTORS HUMAN FACTORS ENGINEERING NEURAL NETWORKS MAN-MACHINE SYSTEMS REACTOR OPERATORS RELIABILITY RESEARCH REACTORS resilience engineering SIMULATORS FUZZY LOGIC SAFETY ANALYSIS	EDUCATION DATA DATA BASE MANAGEMENT INFORMATION DISSEMINATION INFORMATION RETRIEVAL INFORMATIO SYSTEMS NUCLEAR DATA COLLECTION KNOWLEDGE MANAGEMENT KNOWLEDGE PRESERVATION LIBRARIES RESEARCH PROGRAM	ATTENUATION ARTIFICIAL INTELLIGENCE BIOLOGICAL RADIATION EFFECTS COINCIDENCE METHODS ENVIRONMENT GAMMA RADIATION GAMMA SPECTROSCOPY FLUID FLOW INDUSTRY MINING NEURAL NETWORKS NEUTRON SCATTERING NONDESTRUCTIVE TESTING NUCLEAR MEDICINE NATURAL GAS PETROLEUM RADIATION DOSES RADIATION PROTECTION RADIOACTIVE SOURCES SEWAGE TRACER TECHNIQUES TWO-PHASE FLOW

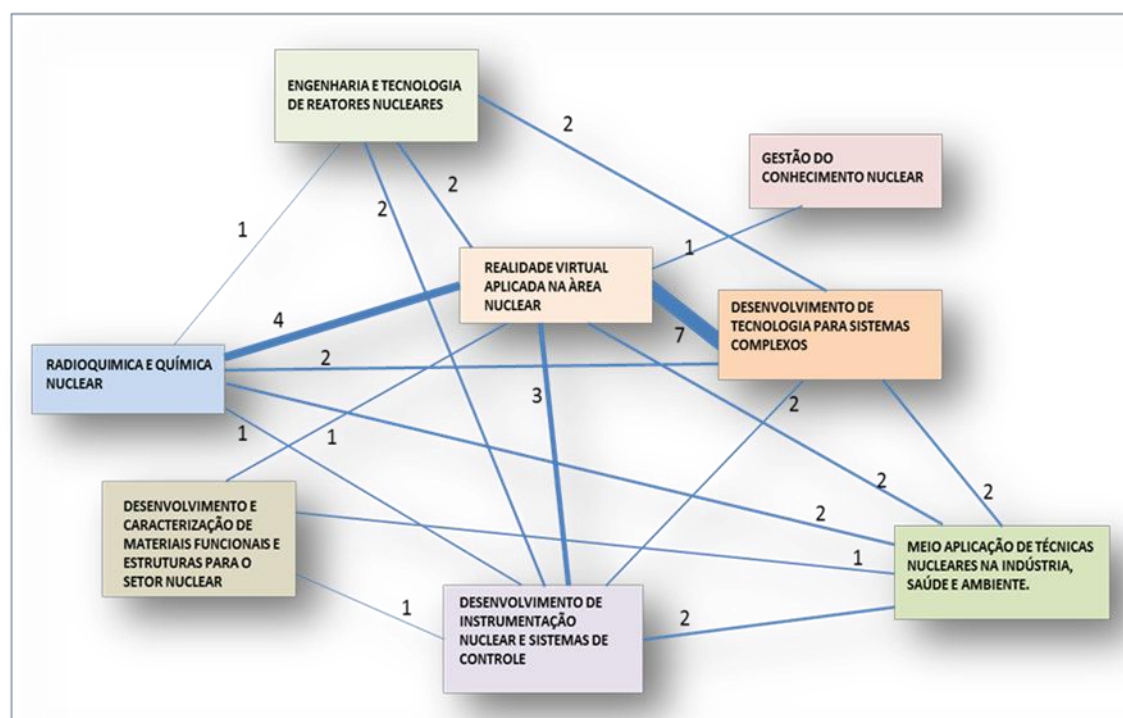
Fonte: A autora

A observação dos termos assinalados pelo indexador revela um número relativamente elevado de assuntos cobertos pelas atividades de pesquisa do IEN, que é – como observado anteriormente - uma característica intrínseca dos estudos em Ciências Nucleares, e, particularmente, em Engenharia Nuclear, que trata de sistemas complexos e

abrangentes, que pressupõe a reunião de experiências e de saberes provenientes de muitos domínios de conhecimento, que não se limitam somente ao conhecimento tecnológico e científico.

No entanto, a comparação entre esses assuntos indica também que há uma superposição importante entre os assuntos tratados pelas diferentes áreas temáticas. Essa interseção detectada pelo levantamento não indica absolutamente duplicação de atividades pelas diferentes áreas temáticas, mas sim que os assuntos são investigados em ambientes e contextos distintos, revelando concretamente o alto grau de inter e multidisciplinaridade da área de Ciências Nucleares. A Figura 20 procura representar a intensidade da intercessão entre as áreas temáticas.

Figura 20: Intensidade da Interseção entre as áreas temáticas



Fonte: A autora

Visões distintas interpretando os mesmos objetos de pesquisa, o que corresponde, em escala variada, em diferenças metodológicas, instrumentais e teóricas, configura uma ambientação propícia para se formular novas formas de diálogo que otimizem o potencial informacional dos produtos de pesquisa gerados, dado que esses produtos não guardam

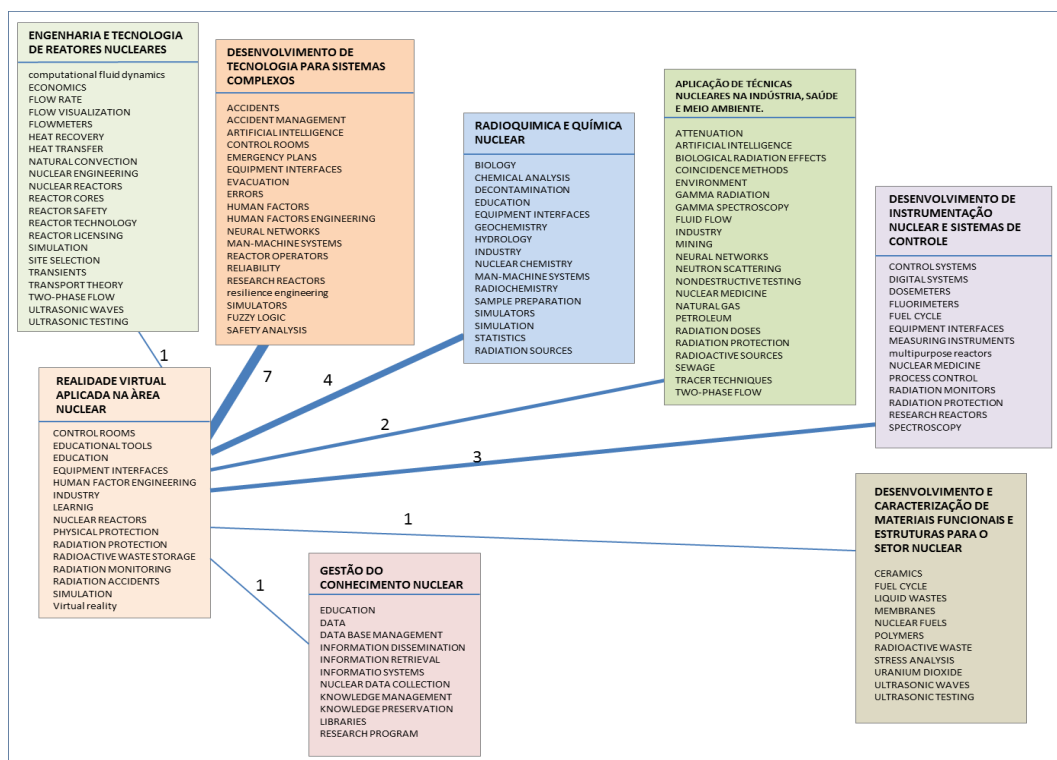
intrinsecamente os seus significados e interpretações. Estes são conferidos por processos extrínsecos de representação e de curadoria.

A partir daí, pode-se pressupor que o reuso e o compartilhamento de dados e informações num ambiente de pesquisa, caracterizados pela pluralidade de visão sobre esses recursos, abrem a possibilidade de se conceituar formas inéditas de agregações abstratas de produtos de pesquisa que sejam portadores de interpretações específicas, criando, dessa forma, novos constructos intelectuais que possuam os atributos mínimos dos recursos informacionais, ou seja, possam ser identificados e tenham sua autoria reconhecida. Esses novos constructos podem constituir formas de expressão que portem unidades de pensamento, hipóteses, opiniões, etc.

É oportuno enfatizar que o reuso e a interpretação de dados de pesquisa em diferentes contextos é um desafio importante na área de curadoria digital de dados de pesquisa e para a *eScience*, que tem que lidar com os enigmas colocados pela grande quantidade de dados produzidos pelas disciplinas científicas, que se enquadram no quarto paradigma, constituindo-se para ambas as áreas objetos essenciais de pesquisa.

Algumas áreas temáticas apresentam graus elevados de interseção com assuntos tratados por outras áreas. A área temática Realidade Virtual, por sua natureza aplicada, instrumental e abrangente, e por estar fortemente baseada em ferramentas computacionais, mantém um grau de interseção com outras áreas, conforme representado na Figura 21. As áreas aparentemente mais autocontidas geram produtos de pesquisa cujo potencial de reuso pode ser potencializado por meio de uma gestão dinâmica baseada em estruturas de banco de dados e de representação por metadados padronizados.

Figura 21: Intensidade da interseção entre a área temática Realidade Virtual e demais áreas



Fonte: A autora

Como será visto na seção seguinte, o IEN produz uma quantidade extraordinária de dados digitais de pesquisa, caracterizados pela heterogeneidade de mídias e formatos, cuja potencial de reuso entre as áreas é maior na medida em que esses dados possam ser revelados e tornados disponíveis através de ferramentas apropriadas. Dessa forma, se delineia uma possibilidade de cooperação entre as áreas temáticas por meio uso de dados digitais de pesquisa produzidos por uma determinada área e reusados por outra.

O reuso de dados de pesquisa em outros contextos e a otimização do seu potencial cognitivo, entretanto, implica na implantação de processos de gestão de dados e de formulação de estratégias que alavanquem o reuso. Para a solução do primeiro problema foi implantado o repositório de dados e informações, a Plataforma Carpe diEN; para o segundo, é colocado a proposta objeto da presente pesquisa, que é a formulação de publicações ampliadas para a área nuclear, que vincule semanticamente produtos de pesquisa, dados e publicações a outros atores importantes do fluxo de geração de conhecimento científico da área de Ciências Nucleares. A proposta implica na vinculação por *links* semânticos de produtos de pesquisa, além de outras entidades, independentes da

área temática que originalmente gerou esses produtos, intensificando, dessa forma, o diálogo entre as áreas. A ideia é que esses *links* semânticos possam ser estabelecidos pelos próprios pesquisadores por meio da oferta de interfaces específicas.

No contexto particular do IEN, fica claro que os cadernos de áreas temáticas – na sua condição de documento de projeto – têm um papel determinante no ciclo de pesquisa institucional. Isso indica que um elemento importante do modelo de publicação ampliada proposto para o IEN deve ser a *entidade* PROJETO.

7.2.1.2 Mapeamento de dados de pesquisa gerados no IEN

Dados digitais de pesquisa são intrinsecamente diversificados e heterogêneos. Têm origens distintas, servem a propósitos específicos e são gerados pelos mais diferentes processos; além do mais, a tecnologia digital exacerba esta disparidade criando a todo momento formas inéditas de formatação e apresentação dos dados. Deve-se considerar ainda que uma parcela importante dos dados de pesquisa é gerada por simulação e por outras aplicações computacionais, especialmente nas áreas de ciências exatas - o que caracteriza, portanto, os dados de pesquisa e a diversidade e heterogeneidade.

Os dados digitais produzidos como decorrência das atividades acadêmicas e de pesquisa do IEN não se distanciam das características acima. O amplo espectro de tipos de dados constatado no levantamento realizado é decorrência direta da multidisciplinaridade da área nuclear, reforçada pelo número e multiplicidade de projetos em desenvolvimento no IEN e de suas metodologias específicas para a geração dos seus dados.

O quadro 5 explicita, para melhor compreensão, as definições dos tipos de dados de pesquisa gerados pelo Instituto e foi elaborado com a finalidade de minimizar falhas na comunicação devido a uma possível ambiguidade conceitual, enquanto o quadro 6 exhibe os tipos de dados gerados por cada uma das áreas temáticas do IEN, revelando mais claramente sua multiplicidade;

Quadro 5: Definição dos diversos tipos de dados

TIPOS DE DADOS (quanto à natureza)		DEFINIÇÃO
Número	Medidas	Quantidade fixada por um padrão para determinar as dimensões ou o valor de uma grandeza da mesma espécie
	Resultados de levantamentos	Resultado de pesquisa prévia e mais ou menos aprofundada de um fenômeno, antes de se fazer um projeto, um programa, uma pesquisa científica etc. (coleta)
	Resultados de experimentos	Resultado de trabalho científico que se destina a verificar um fenômeno.
	Fórmulas	Expressão concisa e rigorosa, constituída em geral de símbolos, que resume um certo número de dados
	Equações	Redução de uma questão, um problema intrincado, a pontos simples e claros, para facilitar a obtenção de uma solução
	Algoritmos	Seqüência finita de regras, raciocínios ou operações que, aplicada a um número finito de dados, permite solucionar classes semelhantes de problemas
Multimídia	Imagens	Representação da forma ou do aspecto de ser ou objeto por meios artísticos
	Vídeo	Técnica de reprodução eletrônica de imagens em movimento
	Áudio	Sinal sonoro; som
	Animações	Ato ou efeito de imprimir movimento ou aceleração
	Filme	Seqüência de imagens registradas em filme cinematográfico ou videoteipe, para exibição em movimento ou não;
	Fotografia	Imagem obtida por arte ou processo de reprodução sobre uma superfície fotossensível (como um filme), pela ação de energia radiante, esp. a luz
Software	Bases de dados	Conjunto de dados inter-relacionados sobre determinado assunto, armazenados em sistemas de processamento de dados segundo critérios preestabelecidos (reúne)
	Simulações	Teste, experiência ou ensaio em que se empregam modelos para simular o ser humano, em especial em casos de grande perigo de vida
	Códigos nucleares	Programa de computador que representam as simulações matemáticas do núcleo do reator.
Visualização	Tabelas	Quadro sistemático de consulta de dados
	Gráficos	Curva num sistema de coordenadas, que representa uma função [A curva pode ser substituída por uma superfície, uma série de colunas etc.]
	Diagramas	Representação gráfica, por meio de figuras geométricas (pontos, linhas, áreas etc.), de fatos, fenômenos, grandezas, ou das relações entre eles
	Modelos em 3D	Modelo em formato tridimensional, que inclui a idéia de profundidade
	Modelos reduzidos	Esquema que possibilita a representação de um fenômeno ou conjunto de fenômenos físicos e eventualmente a previsão de novos fenômenos ou propriedades, tomando como base um certo número de leis físicas, em geral obtidas ou testadas experimentalmente
	Desenhos	Representação de seres, objetos, idéias, sensações, feita sobre uma superfície, por meios gráficos, com instrumentos apropriados
	Metadados	Dados que registram e preservam dados
	Questionários	Seqüência de perguntas feitas para servir de guia a uma investigação
	Entrevistas	Coleta de declarações tomadas para divulgação
	Anotações	Indicação escrita breve
	Normas	Aquilo que regula procedimentos ou atos;

Textuais	Padrões	Base de comparação, algo que o consenso geral ou um determinado órgão oficial consagrou como um modelo aprovado. objeto que serve de modelo para outro
	Certificados	Documento no qual se atesta a existência de certo fato e dele se dá ciência
	Caderno de laboratório	Ferramenta usada por pesquisadores de várias áreas para fazer anotações sobre a pesquisa quando executada em laboratórios.
	Transcrição	Passar para o papel ou equivalente (algo) que está sendo ouvido (p.ex., um texto de discurso, uma música etc.)
	Correspondências	Intercâmbio de mensagens, cartas etc. entre pessoas, promovido através de serviço próprio
	Diário	Escrito em que se registram os acontecimentos de cada dia
	Caderno de campo	Ferramenta usada por pesquisadores de várias áreas para fazer anotações quando executam trabalhos de campo. É um exemplo clássico de Fonte primária.
Artefatos	Espécimes	Exemplo, amostra, modelo
	Amostras	Pequena porção de alguma coisa dada para ver, provar ou analisar, a fim de que a qualidade do todo possa ser avaliada ou julgada
	Maquete	Representação em escala reduzida de uma obra de arquitetura ou engenharia a ser executada.
		Cenário em miniatura destinado a filmagens de estúdio, quando a obtenção de certas imagens, em ambientes ou paisagens reais, se torna muito difícil ou impraticável; reprodução em miniatura de edifícios, meios de transporte, paisagens etc., us. na simulação de peripécias impossíveis de filmar (p.ex., cenas de catástrofes)
Processos	Procedimentos operacionais padronizados	Procedimento que busca fazer com que um processo, independente da área, possa ser realizado sempre de uma mesma forma, permitindo a verificação de cada uma de suas etapas. Ele deve ser escrito de forma detalhada para a obtenção de uniformidade de uma rotina operacional, seja ela na produção ou na prestação de serviços.
	Workflows	Sequência de passos necessários para que se possa atingir a automação de processos de negócio , de acordo com um conjunto de regras definidas, envolvendo a noção de processos, permitindo que estes possam ser transmitidos de uma pessoa para outra de acordo com algumas regras.
	Protocolos	Planejamento que visa responder uma pergunta ou problema em evidência, definindo a estrutura da pesquisa, selecionando o tipo e o número de variáveis a serem estudadas, e analisando os resultados encontrados
	Teste	Exame crítico ou prova das qualidades de uma pessoa ou coisa
Outros	Phanton ou Manequim	UP Simulador de Tecido Material que possui as mesmas características que o tecido humano com relação à absorção e espalhamento da radiação ionizante.

Fonte: A autora

Quadro 6: Tipos de dados gerados pelas áreas temáticas do IEN

ÁREAS TEMÁTICAS DADOS DE PESQUISA		ENGENHARIA E TECNOLOGIA DE REATORES	REALIDADE VIRTUAL	RADIOQUÍMICA E QUÍMICA NUCLEAR	DESENV. DE TECNOLOGIA PARA SISTEMAS COMPLEXOS	DESENV. E CARACTERIZAÇÃO DE MATERIAIS FUNCIONAIS E MATERIAIS DE CONSTRUÇÃO	APLICAÇÕES DE TÉCNICAS NUCLEARES NA INDÚSTRIA, SAÚDE MEIO AMBIENTE	GESTÃO DO CONHECIMENTO NUCLEAR	DESENV. DE INSTRUMENTAÇÃO NUCLEAR E SISTEMAS DE CONTROLE
Número	Medidas	S	N	S	S	S	S	S	S
	resultados de levantamentos	S	S	S	N	S	S	S	S
	resultados de experimentos	S	S	S	N	S	S	N	S
	Fórmulas	S	S	S	S	S	S	S	S
	Equações	S	S	S	S	S	S	N	S
	Algoritmos	S	S	S	S	S	S	S	S
Multimídia	Imagens	S	S	S	S	S	S	S	N
	Vídeo	S	S	N	S	S	S	S	N
	Áudio	S	S	N	S	N	S	S	N
	animações	S	S	N	S	S	N	S	N
	Filme	S	S	N	S	N	N	S	N
	fotografia	S	S	S	S	S	S	S	N
Software	bases de dados	S	S	S	S	S	S	S	S
	simulações	S	S	S	S	S	S	S	S
	Códigos Nucleares	S	S	N	S	S	N	S	S
Visualização	tabelas	S	S	S	S	S	S	S	S
	gráficos	S	S	S	S	S	S	S	S
	diagramas	S	S	S	S	S	S	S	S
	modelos em 3D	S	S	S	S	S	S	S	S
	Modelos reduzidos	S	N	S	S	N	N	S	S
	Desenhos	S	S	S	S	S	S	S	S
Textuais	metadados	S	S	S	N	N	N	S	S
	questionários	N	S	N	S	S	N	S	S
	entrevistas	N	S	N	S	S	N	S	S
	anotações	S	S	S	S	S	S	S	S
	normas	S	S	S	S	S	S	S	S
	padrões	S	S	S	N	S	S	S	S
	certificados	S	S	S	S	S	S	S	N
	caderno de laboratório	S	S	S	S	S	S	S	S
	transcrição	S	S	N	S	S	N	S	N
	correspondências	S	S	S	S	S	S	S	N
	diário	S	N	N	N	N	N	S	N
	caderno de campo	S	N	S	S	N	S	S	N

Artefatos	espécimes	N	N	S	N	S	S	S	N
	amostras	S	N	S	N	S	S	S	S
	maquete	S	S	N	S	S	N	N	S
Processos	procedimentos operacionais padronizados	N	S	S	S	S	S	S	S
	workflows	S	S	S	S	S	S	S	S
	protocolos	S	S	S	S	S	S	S	S
	teste	S	S	S	S	S	S	S	S
Outros	Phantom ou manequim	n	N	n	n	S	n	n	n

Fonte: A autora

Uma constatação importante revelada pelo levantamento diz respeito à curadoria digital dos dados de pesquisa. Apesar da quantidade e diversidade de tipos de dados gerados pelos laboratórios do IEN, não há em contrapartida nenhuma estratégia de preservação desses recursos informacionais. Via de regra, eles são armazenados em computadores pessoais ou em mídias isoladas e estão suscetíveis a perdas por danos físicos e obsolescência tecnológica, e conseqüentemente não podem ser identificados, recuperados ou colocados disponíveis em rede.

Grande parte destes dados confere sustentação experimental aos conteúdos presentes nas publicações acadêmicas e são evidências imprescindíveis no fluxo de geração de conhecimento. O tratamento e registro dos dados em repositório e a vinculação dos dados aos projetos, pesquisadores e publicações – na forma de publicações ampliadas - criam um compromisso institucional de gestão e curadoria e abrem possibilidades reais para novos serviços de informação.

Voltando ao modelo de publicação ampliada que está sendo proposto, fica claro que os DADOS DE PESQUISA se constituem numa *entidade* imprescindível para a concepção do modelo, dado a sua importância como recurso informacional e a riqueza das relações que podem se estabelecer entre os dados e as demais *entidades*, como por exemplo, com as publicações acadêmicas em formato digital, que serão vistas a seguir.

7.2.1.3 Mapeamento de e-prints gerados no IEN

O mapeamento de e-prints considerou inicialmente as publicações acadêmicas tradicionais e já constantes como coleção nas comunidades existentes na Plataforma Carpe diEN, a saber:

- Artigos de Periódicos
- Capítulos de Livros
- Dissertações do PPGIEN
- Livros
- Objetos de Ensino
- Patentes
- Relatórios
- Softwares
- Teses e Dissertações produzidas por servidores do IEN
- Trabalhos / Apresentação de Congresso

Além das publicações acadêmicas mais tradicionais, como artigos de periódicos, livros, dissertações e teses, outras formas de publicação em formato digital são produzidas no Instituto, muitas vezes refletindo a tendência de utilização de novas mídias e redes sociais para a comunicação entre pesquisadores. Coletivamente essas publicações serão denominadas *e-prints*. A justificativa por esse nome é tornar o modelo mais próximo ao padrão de publicação ampliada preconizada pelo padrão SURF.

Na plataforma Carpe diEN, além de publicações acadêmicas, constam também as coleções “dados e conjuntos de dados de pesquisa” e “cadernos de áreas temáticas e outros projetos”, como pode ser observado em <<http://carpedien.ien.gov.br>>, mas que não foram citadas nesse item por constarem como outro tipo de entidade no modelo proposto.

O quadro 7 exhibe o resultado do levantamento relativo à produção de outros *e-prints* produzidos no IEN por área temática.

Quadro 7: Resultado do levantamento de e-prints do IEN por área temática

EPRINT ÁREAS TEMÁTICAS		REALIDADE VIRTUAL	RADIOQUÍMICA E QUÍMICA NUCLEAR	DESENV. DE TECNOLOGIA PARA SISTEMAS NUCLEARES	DESENV. E CARACTERIZAÇÃO DE MATERIAIS FUNCIONAIS E NUCLEARES	APLICAÇÕES DE TÉCNICAS NUCLEARES NA INDÚSTRIA, SAÚDE E MEIO AMBIENTE	GESTÃO DO CONHECIMENTO NUCLEAR	DESENV. DE INSTRUMENTAÇÃO NUCLEAR E SISTEMAS DE CONTROLE
Relatórios técnicos	S	S	S	S	S	S	S	S
Relatórios de pesquisa	S	S	S	S	S	S	S	S
Pedido de financiamento	S	S	S	S	S	S	S	S
Formulários	S	S	S	S	S	S	S	S
Pedidos de consentimento	S	X	S	S	S	S	S	S
Blogs	N	N	N	N	S	N	S	N
Redes Sociais	S	N	N	S	S	N	S	N
Relatório de pesquisa em andamento	S	S	S	S	S	S	S	S
Comissão de ética	N	S	N	S	N	N	N	N
Projeto de pesquisa	S	S	S	S	S	S	S	S
Pedido de financiamento	S	S	S	S	S	S	S	S
Avaliação dos financiadores	S	S	S	S	S	S	N	N
Avaliação dos pares	S	S	S	S	S	S	N	S

Fonte: A autora

No Quadro acima o que se pode observar é que no IEN, além dos dados de pesquisa, existem outros tipos de documentos digitais considerados importantes e que também registram a memória da pesquisa desenvolvida, merecendo serem acoplados à publicação ampliada. Pode-se perceber ainda que há no Instituto uma tendência no uso de blogs e redes sociais para fins de divulgação científica. Esse fato revela que de alguma forma há uma necessidade por parte do pesquisador de compartilhar suas pesquisas, o que

pode ser ampliado a partir de novos modelos de publicação que sendo *open source* disseminem esses resultados, permitam identificação de novos pares e a criação de novas parcerias.

Sobre a forma de condução dessa pesquisa, é importante destacar que alguns modelos, como por exemplo, o CERIF, consideram os produtos de pesquisa – dados de pesquisa e publicações acadêmicas (*e-prints*) – como uma única *entidade*. Entretanto, para a presente proposta de modelo foram consideradas *entidades* independentes pelos seguintes motivos: manutenção de um maior grau de aderência ao padrão SURF; para a realização de tratamento diferenciado aplicado aos dados em relação aos *e-prints* no ambiente do IEN; e a possibilidade de explicitar com maior detalhe a semântica das relações entre dados e *e-prints*, que é o foco central do modelo proposto.

Nessa perspectiva soma-se ao conjunto de *entidades* de pesquisa do modelo definidas a *entidade* E-PRINTS. Nessa direção define-se um conjunto de *entidades* chamadas coletivamente de objetos de pesquisa, composto de *e-print*, dado de pesquisa e projeto. A seguir serão explicitados os atores que interagem sobre esses objetos num ambiente de pesquisa e relacionamentos importantes para o modelo: pessoas e organizações.

7.2.1.4 Atores: organização e pessoa

É necessário considerar que as pesquisas transcorrem como uma atividade institucionalizada, que envolvem diversas organizações como laboratórios, departamentos, institutos de pesquisa, órgãos financiadores, instituições parceiras, além de organizações que estabelecem políticas de C&T e órgãos internacionais, entre tantas outras. Complementarmente, é indispensável considerar a interação da organização de pesquisa com os demais setores da sociedade, por exemplo: a educação, nos processos de divulgação científica; a indústria, quando da transferência de tecnologia; e as instâncias políticas, governamentais e jurídicas, na definição de marcos legais e controles sociais. Assim sendo, torna-se importante assinalar ORGANIZAÇÃO como uma *entidade* essencial para o modelo proposto para o IEN.

Por fim, cumpre destacar a infinidade de papéis que um indivíduo pode protagonizar num ambiente de pesquisa: autor, coordenador, diretor, orientador, avaliador,

etc. Assim sendo, uma representação para *entidade* que represente esses papéis torna-se essencial para o modelo, para tal postula-se a *entidade* PESSOA.

Cabe observar que a vinculação entre pessoas e organizações esta fundamentada no fato de que a história acadêmica das pessoas, a experiência e o contexto organizacional na qual a pesquisa se desenrola são fatores determinantes que influenciam diretamente nos resultados da pesquisa.

Outras *entidades* importantes para o mundo da pesquisa podem ser consideradas, como por exemplo, EVENTO - que é incluída entre as entidades do CERIF. No entanto, na presente proposta essa entidade não foi considerada inicialmente, posto que evento se materializa por meio de publicações de anais de congressos, resumos ou apresentações, podendo ser classificado juntamente com e-prints. É preciso observar que o modelo é capaz de incorporar entidades quando se tornarem necessário para a representação de realidades específicas. A idéia de entidades de partida, ou entidades essenciais, é adotada pelos principais sistemas de informações, como o CERIF. Seguindo esse princípio, as entidades de partida que deve constar no modelo de publicação ampliada para a área de ciências nucleares, especificamente para o IEN são as seguintes:

- DADO DE PESQUISA
- EPRINT
- PROJETO
- PESSOA
- ORGANIZAÇÃO

7.2.1.5 Entidades e relações

A concepção de um modelo de publicação ampliada para a área nuclear, instanciada por uma especificação voltada para o ambiente de pesquisa do Instituto de Engenharia Nuclear da CNEN, deve seguir de perto alguns padrões, recomendações e pressupostos já estabelecidos, sendo os principais os seguintes:

- a) Recomendações sobre a criação de publicações ampliadas da *SURF Foundation: Enhanced Publications: Linking Publications and Research Data in Digital Repositories* (2009b)
- b) Relatórios da EROCRIS: *CERIF 1.3 Full Data Model (FDM): Introduction and Specifications* (2007) e *CERIF 1.3 Semantics: Research Vocabulary* (2010)
- c) Vocabulários, taxonomias e ontologias consagradas e de ampla aplicação nas áreas científicas, por exemplo: CITO - *Citation Typing Ontology*, DoCo - *Document Components Ontology*, SKOS - *Simple Knowledge Organization System*, CERIF etc.

Complementarmente devem ser considerados os estudos de Serguei Parinov (2012, 2013) e de Serguei Parinov e Mikhail Kogalovisky (2014) em torno dos conceitos de *link* semântico e de suas possibilidades para o desenvolvimento de sistemas de informação mais avançados e flexíveis. Entretanto, é importante observar que os desenvolvimentos propostos por Parinov têm uma forte influência do modelo de Camada semântica proposto pelo CERIF e que será objeto da discussão que se segue.

O CERIF é, de uma forma geral, um padrão para intercâmbio de informações sobre pesquisa científica recomendado pela União Européia a seus membros. Este padrão compreende um conjunto de documentos que especificam modelos conceituais que apóiam a gestão e o intercâmbio de informações sobre pesquisa. Duas características primordiais do CERIF são importantes para o presente estudo:

- A formalização geral das relações e tipos de termos para o domínio da pesquisa científica;
- O conceito de objetos ou *entidades* de pesquisa com atributos tais como PROJETO, PESSOA, ORGANIZAÇÃO.

Com a versão CERIF 2006 melhorias substanciais foram incorporadas ao modelo anterior, especialmente no que concerne à incorporação de atributos semânticos. A implementação da chamada “Camada semântica” (*SemanticLayer*) tornou o modelo mais flexível e escalável para aplicações em ambientes muito heterogêneos, característicos das áreas científicas. A camada semântica do CERIF é considerada “um instrumento simples, mas poderoso que permite a representação de diversos tipos de relação (...)”. Essa camada admite o registro e o armazenamento de valores semânticos que são portados ou

referenciados pelos *links*, que por sua vez conectam as entidades do domínio e conferem valor semântico a agregação que se forma. A simplicidade do modelo contrasta com a sua capacidade de expressar as relações entre objetos de pesquisa, entre os atores que interagem com os objetos de pesquisa e entre os objetos e os atores. O vocabulário de pesquisa estabelece uma semântica formal que pode ser aplicada amplamente em todo o universo da pesquisa científica.

No contexto do CERIF, a relação ou *links* entre entidades CERIF são chamadas de *Link Entities*. O conceito de *Link Entities* é considerado o elemento mais robusto do modelo CERIF, dado que conecta sempre duas entidades do domínio – por exemplo, PESSOA_PROJETO, ORGANIZAÇÃO_PUBLICAÇÃO – constituindo um leque de relações possíveis entre as entidades definidas para o domínio considerado. Cada registro de *link* no sistema deve portar a semântica da ligação através de uma referência à Camada semântica do CERIF, o que tem uma importância crítica no modelo semântico do CERIF.

O padrão especifica ainda um formato de intercâmbio de dados CERIF em XML – CERIF 1.3 – XML: Specification Document (EUROCRIS, 2012b). O exemplo abaixo mostra uma representação em CERIF/XML do registro de um *link* PESSOA_PUBLICAÇÃO, cuja relação entre a pessoa e a publicação é de “primeiro autor”. O registro XML a seguir inclui referência ao vocabulário semântico utilizado: CERIF – 1.2 *Semantics: Research Vocabulary*.

```
<cfPers_ResPubl>
  <cfPersId>person-brigitte-joerg</cfPersId>
  <cfResPublId>publ-analytic-information-service-era</cfResPublId>
  <cfClassId>FirstAuthor</cfClassId>
  <cfClassSchemeld>cf2008-1.2_CERIF_Semantics</cfClassSchemeld>
  <cfStartDate>2008-01-01T00:00:00-00:00</cfStartDate>
  <cfEndDate>2008-12-31T00:00:00-00:00</cfEndDate>
  <cfFraction>0.25</cfFraction>
</cfPers_ResPubl>
```

Na ótica do relatório da SURF Foundation, expressa pelos autores Woutersen-Windhouwer e Brandsma (2009, p.21) temos a seguinte constatação:

O número de objetos de pesquisa na Internet está crescendo rapidamente. Integração de toda essa informação científica por *links* torna-se necessária para manter a publicação eficiente e para se garantir o controle sobre todo o processo. Portanto, as publicações devem oferecer estes *links* que as transformem em Publicações Ampliadas.

Quando este mesmo relatório introduz a questão do significado, caracterizando-o como um avanço em relação à *linkage* implementado por um URL, despido de semântica, que simplesmente direciona o leitor para outro recurso, ele apresenta o seguinte argumento:

Ao invés de inserir um URL, como quase todos os editores o fazem nos dias de hoje, é também possível fornecer *links* com um ‘significado’ no qual a conexão lógica de termos estabelece interoperabilidade entre as partes de uma Publicação Ampliada” (WOUTERSEN-WINDHOUSER E BRANDSMA, 2009, p.59)

Logo em seguida conclui que ontologias devem ser utilizadas em ambientes de Publicações Ampliadas para dotar os *links* de valor semântico.

A aproximação do modelo de publicação ampliada preconizado pelo Relatório da SURF Foundation com os pressupostos do padrão CERIF, abre uma perspectiva motivadora de se dispor de uma infraestrutura semântica padronizada – implementada pelo CERIF *Semantic Layer* - para conferir significados aos *links* que vinculam os recursos agregados por uma publicação ampliada.

Indo nessa direção, a compatibilização com os pressupostos básicos do CERIF e do SURF formam os alicerces para a composição do modelo de relações semânticas e para enriquecimento semântico das publicações ampliadas para a área de ciências nucleares, particularmente para o IEN.

Os relacionamentos ou *link* semântico entre as entidades definidas para o ambiente de pesquisa IEN – chamada no escopo do CERIF de *Link Entities*– é o ponto chave para viabilização do modelo proposto. Um *link* semântico sempre conecta duas entidades que têm papel relevante no universo da pesquisa científica, sejam elas resultados de pesquisa, pessoas, instituições, projetos ou qualquer outro ator ou objeto de pesquisa identificados como necessários à representação do domínio onde ocorrem as atividades científicas.

As relações que se estabelecem entre os diversos objetos de pesquisa e atores – ou seja, as entidades de pesquisa – podem ter os seus significados atribuídos por diferentes ontologias já existentes e incluem:

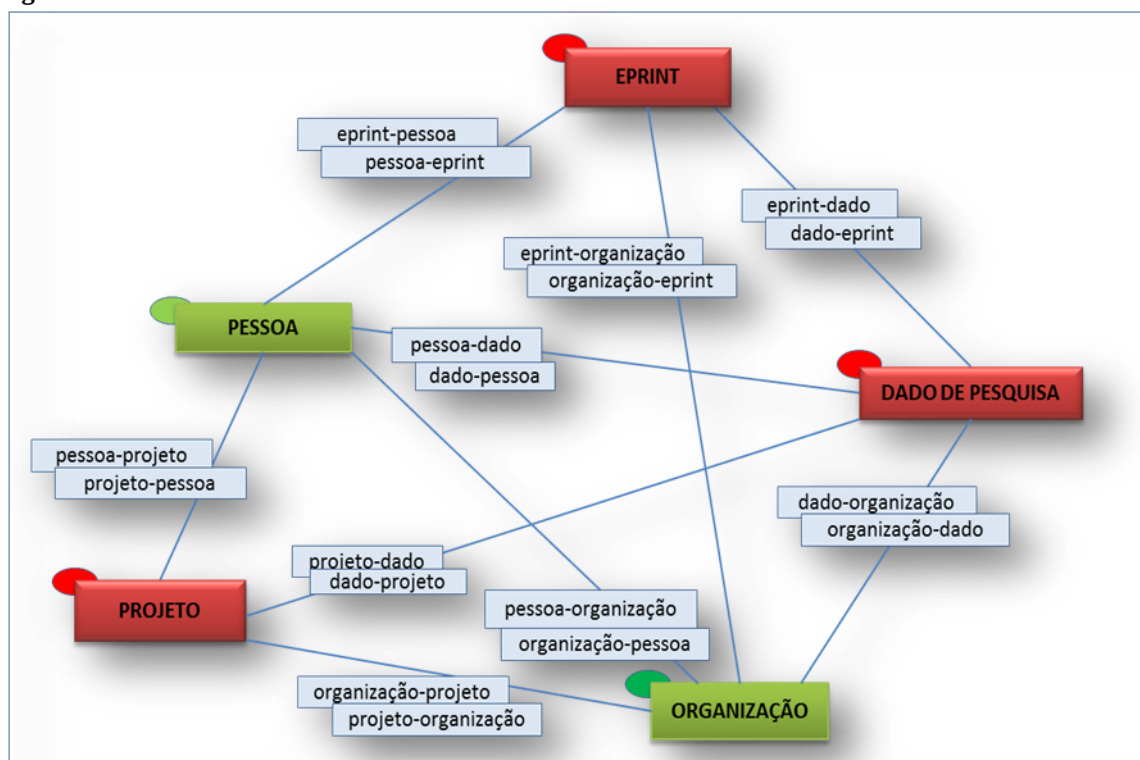
- Relacionamentos entre os vários objetos de pesquisa, como *e-prints* (artigos, teses, etc.), dados de pesquisa, comentários, anotações, projetos, avaliações etc.;

- Relacionamentos entre os atores – pessoas e organizações - que têm papel relevante no cenário científico, tais como pesquisadores, orientadores, membros de banca, departamentos, laboratórios, órgãos de fomento a pesquisa etc.;
- Relacionamento entre os diversos objetos de pesquisa de um lado e o conjunto de atores do outro, por exemplo, dados de pesquisa e laboratório.

Conforme enfatiza Parinov (2012, p.1), os sistemas de informações mais avançados – como os ambientes informacionais definidos pelo modelo CRIS – operam virtualmente com o mesmo conjunto de entidades de pesquisa: “pessoa”, “organização”, “projeto”, “produtos de pesquisa”, “eventos” e alguns outros. A seleção do conjunto de entidades para o modelo aproxima-se desse padrão que varia ligeiramente com o propósito e conceito de cada sistema.

A Figura 22 representa uma visão abstrata dos *links* semânticos que indicam as classes de relacionamentos entre as entidades identificadas na seção anterior como mais relevantes no ambiente de pesquisa do IEN.

Figura 22: Classes de relacionamentos



Fonte: A autora baseado em EUROCRIS(2010a)

Nesse domínio específico, as entidades são definidas operacionalmente como:

- EPRINT – objetos de pesquisa em formato digital usado para comunicar formalmente e também informalmente resultados de atividades de pesquisa acadêmica, como artigos, teses e dissertações. Incluem novos documentos e mídias eletrônicas como blogs e registros em redes sociais.
- DADO DE PESQUISA – objetos de pesquisa criados experimentalmente, teoricamente ou por simulação. Incluem um conjunto amplo de objetos gerados por derivação, inferência, análises, anotações e reformatação de outros objetos de informação.
- PROJETO – objeto de pesquisa constituído de documentos que registram parâmetros técnicos, científicos, administrativos, de planejamento e orçamentário voltado para o desenvolvimento de pesquisa científica e acadêmica.
- PESSOA – atores envolvidos nos fluxos de pesquisa científica tanto em áreas afins, como pesquisadores e professores, como também em áreas de gestão, suporte e financiamento de atividades de pesquisa, como coordenadores, diretores.
- ORGANIZAÇÃO – instituições ou unidades de instituições, como laboratórios, departamentos e cursos onde se desenrolam atividades acadêmicas ou de pesquisa, ensino, fomento ou planejamento de pesquisa científica.

Adotando o padrão especificado pelo CERIF para dar nomes físicos aos *links* semânticos, temos que o nome do *link* é composto pelos nomes herdados das duas entidades relacionadas, incluindo o prefixo IEN - que identifica o domínio onde as relações se estabelecem; a ordem dos nomes das entidades indica a entidade “Fonte” e a entidade “alvo” da relação. Nessa direção, o nome do *link* semântico tem a seguinte forma: ienNome da Entidade 1_Nome da Entidade 2.

Dado que no ambiente IEN foram consideradas inicialmente as entidades: EPRINTS, DADOS, PESSOAS, ORGANIZAÇÃO, PROJETOS, temos como ponto de partida as classes de *links* semânticos apresentados no quadro 8, abaixo.

Quadro 8: Classes de links semânticos

ENTIDADES	EPRINT	DADO	PESSOA	ORGANIZAÇÃO	PROJETO
LINKS SEMÂNTICOS	EPRINT_EPRINT EPRINT_DADO EPRINT_PESSOA EPRINT_ORG EPRINT_PROJETO	DADO_DADO DADO_EPRINT DADO_PESSOA DADO_ORG DADO_PROJETO	PESSOA_PESSOA PESSOA_EPRINT PESSOA_DADO PESSOA_ORG PESSOA_PROJETO	ORG_ORG ORG_EPRINT ORG_DADO ORG_PESSOA ORG_PROJETO	PROJETO_PROJETO PROJETO_EPRINT PROJETO_DADO PROJETO_PESSOA PROJETO_ORG

Fonte: A autora

O CERIF reforça a idéia de que cada *link* porta significados na medida em que há uma referência objetiva a sua camada semântica, composto por classes de *links* e por esquemas de classificação. É necessário incluir, portanto, no modelo proposto, um tipo de ancoragem que permita conferir significados aos *links*. Isto é realizado por meio de uma TAXONOMIA DE RELAÇÕES que apresente referência às ontologias e vocabulários científicos já existentes ou ao vocabulário IEN, quando a relação não for identificada nos vocabulários científicos mais conhecidos. As entidades identificadas e as relações definidas pela Taxonomia de Relações formam um espaço de relações científicas IEN.

De acordo com Jorg et al. (2009a, 2009b), os vocabulários semânticos são entendidos como uma coleção de significados que representam diferentes aspectos de uma classe específica de relações entre objetos de pesquisa e atores. Estes vocabulários podem estar expressos por ontologias e taxonomias e devem ser tratados por computador para serem consultados e apresentados de forma conveniente para edição, registro de novas relações e composição de publicações ampliadas.

Conforme proposto por Parinov (2013), a formalização da relação binária entre os objetos de pesquisa pode ser visualizada através de uma matriz bidimensional, onde as colunas e linhas correspondem à lista completa das entidades científicas consideradas no domínio em cujas relações estão sendo mapeadas. Cada uma das células dessa matriz contém um ou mais vocabulários semânticos que definem os significados possíveis para o par de entidades que forma o *link* semântico.

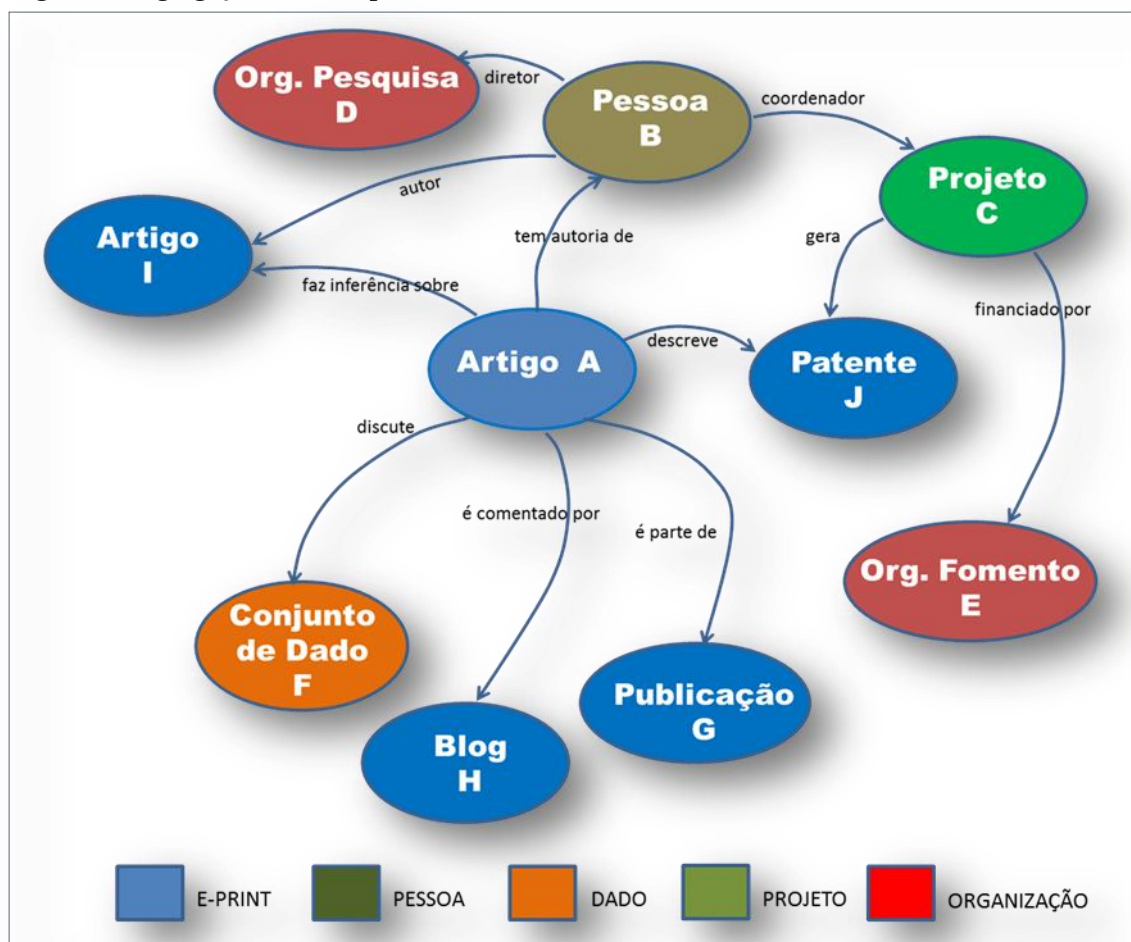
No modelo proposto, cada célula da matriz contém no mínimo um vocabulário semântico, indicando que a Taxonomia de Relações científicas gerada por esse mecanismo cobre todos os objetos do dado domínio.

Nessa direção, pressupondo que os *links* semânticos expressam o relacionamento entre dois objetos, resta determinar que classe de relacionamento científico deve ser aplicado para cada combinação, e a que vocabulário semântico pertence cada classe aplicada. Esta reposta é dada pela Taxonomia de Relações.

À guisa de ilustração, considera-se o seguinte recorte que se desenrola num espaço hipotético de recursos, atores e relações. o artigo E-PRINT “A” tem como autor a PESSOA “B”, que por sua vez é coordenador do PROJETO “C”; a PESSOA “B” trabalha na ORGANIZAÇÃO “D” como diretor; o PROJETO “C” é financiado pela ORGANIZAÇÃO “E”; considera-se que o E-PRINT “A” abre uma discussão sobre o conjunto de dados “F”(DADO “F”) que é apresentado em formato multimídia; “A” é um item da publicação E-PRINT “G” e é comentado no blog E-PRINT “H” do autor “B” e faz inferências sobre o artigo E-PRINT “I”; o PROJETO “C” gerou a patente E-PRINT “J” que é descrita por “A”.

A Figura 23 apresenta uma representação gráfica de algumas das possíveis relações semânticas entre as entidades do espaço postulado. Uma das características importantes que deve ser observada é que uma publicação individual, por exemplo, o E-PRINT “A”, está agora inserida em uma rede semântica contextualizada pelos *links* entre diversos outros recursos e atores. A formalização dessa rede abre possibilidade da implementação de uma rede navegável e interoperável.

Figura 23: Agregação formada pelas entidades e links semânticos



Fonte: A autora

O Quadro 9 organiza os elementos que tomam parte da agregação formada pelas entidades e *links* semânticos, incluindo as classes de relacionamentos e vocabulários semânticos que foram aplicados para cada par “FONTE_ALVO”, definindo, de uma forma padronizada, o significado que porta cada *link*.

Quadro 9: Elementos da agregação formada por entidades e links semânticos

ENTIDADE		CLASSE DE RELACIONAMENTO	RELAÇÃO	VOCABULÁRIO
FONTE	ALVO			
A	B	E-PRINT_PESSOA	<i>tem autoria de</i>	VOC-IEN
A	F	E-PRINT_DADO	<i>discute</i>	CITO
A	G	E-PRINT_E-PRINT	<i>é parte de</i>	DoCo
A	H	E-PRINT_E-PRINT	<i>concorda com</i>	CITO
A	I	E-PRINT_E-PRINT	<i>é continuação de</i>	DATA CITE
B	C	PESSOA_PROJETO	<i>coordenador</i>	CERIF SemanticVocabulary
B	D	PESSOA_ORG	<i>diretor</i>	CERIF SemanticVocabulary
B	I	PESSOA_E-PRINT	<i>autor</i>	CERIF SemanticVocabulary
C	E	PROJETO_ORG	<i>financiado por</i>	VOC-IEN
C	J	PROJETO_E-PRINT	<i>gera</i>	VOC-IEN
A	J	E-PRINT_E-PRINT	<i>descreve</i>	CITO

Fonte: A autora

As relações são estabelecidas em condições distintas de abstração e subjetividade. Enquanto algumas relações têm seus significados explicitados precisamente tanto pelo censo comum, pelo uso cotidiano, como pelo seu grau de formalização dado pelos instrumentos terminológicos, por exemplo: “autoria”, “citação” e “coordenador”; outras relações se dão em níveis mais altos de abstração e incluem opiniões, julgamentos, inferências, hipóteses, comentários. Essas relações, apesar do grau de subjetividade que portam, proporcionam um alto grau de sofisticação ao modelo, na medida em que permitem a criação de agregações de objetos informacionais mais expressivos, representativos e interativos, espelhando mais precisamente os diálogos possíveis do universo da pesquisa científica no geral, e mais especificamente no IEN.

Considerando essa questão, foi necessário incorporar no modelo proposto uma categorização das relações que levasse em conta os tipos de relação e os seus níveis de abstração. O Quadro 10 apresenta uma proposta de classificação.

Quadro 10: Proposta de classificação dos tipos de relações para publicações ampliadas

TIPO DE RELAÇÃO	SUBCLASSES E EXEMPLOS		
LÓGICA Relação que revela hierarquias de gênero-espécie.	é um é subclasse de é subgrupo de		
ÔNTICA Revelam relação do objeto no mundo e caracterizam-se pela contiguidade no tempo e no espaço ou pela conexão de causa efeito. (Dahlberg, 1978b)	MERONÍMICA Relação entre as partes de uma publicação	é formado de é parte de adjacente, envolve	
	FUNCIONAL relação que revela a função de uma entidade sobre a outra.	CITAÇÃO	revela o tipo de citação entre os documentos
		DERIVAÇÃO	revela fase e estágios de um documento ou processo de pesquisa no tempo
		PROPRIEDADE	revela atributos, características, papel ou propriedade de um documento
		INFLUÊNCIA	revela o impacto, efeito ou ação de um objeto ou processo sobre o outro

Fonte: A autora

Nessa direção, uma taxonomia de relações define e explicita as classes de relações que podem existir sobre um conjunto de objetos de pesquisa num determinado domínio de aplicação. Para a explicitação de entidades e relações adequadas ao IEN, as relações foram estruturadas na forma de uma Taxonomia de Relações Científicas do IEN, apresentadas no Apêndice A onde se indica o vocabulário e/ou ontologias de onde são originadas.

É preciso observar que uma parcela estimável de classes de relacionamento, consideradas como necessárias para composição de publicações ampliadas para o IEN, não foram identificadas nas ontologias consultadas e foram incluídas no Vocabulário IEN

(VOC_IEN – Apêndice C), construído para essa finalidade, isto é, para suprir a carência dos demais vocabulários em relação às necessidades do IEN. O Quadro 11 apresenta uma amostra com as relações consideradas mais significativas para os exemplos apresentados na presente tese. No apêndice D é possível verificar a taxonomia completa.

Quadro 11: Extrato da Taxonomia de Relações Científicas

CLASSE DE RELAÇÃO	RELAÇÃO	CLASSIFICAÇÃO	ONTOLOGIA/ VOCAB. SEMÂNTICO
EPRINT_EPRINT	Cita	Citação	CITO
	É citado por	Citação	CITO
	É refutado por	Citação	CITO
	Revisa	Citação	CITO
	E parte de	Meronímica	DoCo
EPRINT_DADO	Cita como Fonte	Citação	CITO
	Concorda com	Citação	CITO
	Discute	Citação	CITO
	Documenta	Influência	Data Cite
EPRINT_PESSOA	Avaliado por	Influência	VOC_IEN
	É orientado por	Propriedade	VOC_IEN
EPRINT_PROJETO	É resultado de	Meronímica	DoCo
	É citado por	Citação	CITO
EPRINT_ORG	É financiado por	Propriedade	VOC_IEN
	É propriedade de direito	Propriedade	CERIF
DADO_DADO	Atualiza	Derivação	CITO
	É versão de	Derivação	Data Cite
	É suplementado por	Influência	Data Cite
DADO_EPRINT	É figura de	Meronímica	DoCo
	Confirma	Citação	CITO
	Usado em	Influência	CERIF
DADO_PESSOA	É compilado por	Derivação	CITO; Data Cite
	É curado por	Propriedade	VOC_IEN
DADO_PROJETO	Embasa	Influência	CITO
	É apêndice de	Meronímica	DoCo
DADO_ORG	Financiado por	Propriedade	VOC_IEN
	É creditado por	Citação	CITO
PESSOA_PESSOA	Supervisiona	Influência	CERIF
	Administra	Influência	CERIF
	Orienta	Influência	VOC_IEN
PESSOA_EPRINT	É autor de	Propriedade	VOC_IEN
	É revisor de	Propriedade	CERIF
	cita	Citação	CITO
PESSOA_DADO	Publica	Propriedade	CERIF
	Administra	Influência	CERIF
PESSOA-PROJETO	Supervisiona	Influência	CERIF
	É revisor de	Propriedade	CERIF
PESSOA_ORG	Administra	Influência	CERIF
	Trabalha para	Influência	VOC_IEN
PROJETO_PROJETO	É parte de	Partitiva	DoCo
	É atualizado por	Derivação	CITO
PROJETO_EPRINT	Gera	Derivação	VOC_IEN
	É descrito por	Influência	CITO
PROJETO_PESSOA	Avaliado por	Influência	VOC_IEN
	É atualizado por	Derivação	CITO
PROJETO_ORG	É financiado por	Propriedade	VOC_IEN
	Usa dado de	Influência	CITO
	Obtém apoio de	Influência	CITO
ORG_ORG	É financiadora de	Propriedade	VOC_IEN
	É publicadora de	Propriedade	VOC_IEN
ORG_EPRINT	É editor de	Propriedade	CERIF
	É propriedade de direito	Propriedade	CERIF
	É autor institucional de	Propriedade	CERIF
ORG_DADO	Publica	Propriedade	CERIF
	É proprietário de direito	Propriedade	CERIF
	É autor institucional de	Propriedade	CERIF
	Credita	Influência	CITO
ORG_PESSOA	É financiadora de	Propriedade	VOC_IEN
	Credita	Influência	CITO
ORG_PROJETO	Apóia	Influência	CITO
	É proprietário de direito	Propriedade	CERIF

Fonte: A autora

No extrato de relações apresentadas no quadro 11 pode-se observar a relação, a categoria na qual se enquadra dentro da Taxonomia e a ontologia que deu origem à relação. Apesar de todo esforço realizado para o levantamento dessas relações não se pode considerar que todas as possibilidades tenham sido esgotadas. Algumas mais específicas poderão surgir conforme necessidade do autor no momento da criação. Para uma compreensão mais clara de como essa taxonomia de relações será utilizada na construção de uma publicação ampliada e como a retroalimentação dessa taxonomia pode ser feita, a seção a seguir vem apresentar algumas propostas de interfaces que vão mostrar desde a primeira ação na construção de uma publicação ampliada, passando pelo estabelecimento de relações até a visualização da publicação de forma integrada.

7.3 PROPOSTA DE INTERFACES PARA PUBLICAÇÃO AMPLIADA

Dado a dinâmica vertiginosa do mundo da pesquisa contemporânea, especialmente nos ambientes de *eScience*, as relações não são fixas e não podem ser completamente pré-definidas. Dessa forma, os pesquisadores e gestores dos sistemas de informação para a pesquisa devem dispor de ferramentas amigáveis automatizadas para expandir, de forma controlada e sob alguma forma de moderação, os vocabulários semânticos associados com cada tipo de relação. Esse parece ser um pressuposto relevante para os sistemas de informação científicos mais avançados. Considerando essa condição, as interfaces de registro de relações fazem parte dos resultados do estudo, pois revelam o processo de construção de uma publicação ampliada e por isso serão discutidas nesta seção.

7.3.1 *Interface de Autoria/Edição de publicações ampliadas*

A criação de publicações ampliadas deve se tornar parte do ambiente natural de trabalho do cientista nuclear, dado a multiplicidade de objetos digitais de pesquisa gerados no decorrer das atividades de pesquisa, o grau de heterogeneidade desses objetos e, sobretudo, o potencial latente de reuso que possuem, conforme constatado pelo levantamento de dados da presente pesquisa.

O Relatório do DRIVER II (VEHAAR, 2008) – discutido anteriormente no capítulo 3 - argumenta a favor de que ferramentas simples devem ser desenvolvidas e tornadas

disponíveis para que os membros das comunidades acadêmicas e científicas possam, eles mesmos, arquivar seus dados, a sua descrição e a relação entre eles em repositórios especificamente desenvolvidos para esse fim. A idéia acolhida pelo modelo proposto é apresentada na forma de uma camada sobreposta ao repositório de dados do IEN, aqui denominada “Interface de criação e edição de publicações ampliadas”, apresentada na Figura 24, mais adiante.

O Relatório de Vehaar (2008) identifica uma série de requisitos fundamentais para a composição de objetos digitais complexos que assegurem a sua capacidade de intercâmbio, preservação e acesso, conforme se espera das publicações ampliadas. Tomando os pressupostos de padronização preconizados pelo Relatório como base para a especificação das interfaces de autoria de publicações ampliadas para a área nuclear, torna-se necessário a conformidade com dois itens de fundamental importância para o conceito de publicação ampliada que está sendo proposto. São eles:

- Deve ser possível registrar a autoria da publicação ampliada e dos itens que a compõem.
- Deve ser possível registrar as relações entre os recursos *Web* que fazem parte da publicação ampliada.

Considerando o primeiro item, é necessário que uma distinção clara entre o autor da publicação ampliada e os autores das partes que a compõem possa ser bem estabelecida. Autores de um *e-print* e de um particular conjunto de dados são agentes responsáveis pelo conteúdo intelectual desses recursos, enquanto o “autor da publicação ampliada como um todo é o agente que decidiu combinar esses vários recursos na forma de um único objeto composto” (VEHAAR, 2008, p.107).

Tomando em conta o segundo item, o modelo proposto estipula que além de registrar os dados de pesquisa e as informações que os contextualizam – expressos por metadados - o pesquisador possa também estabelecer ligações entre esses dados e os diversos outros objetos de pesquisa através de interfaces *Web*, criando objetos digitais compostos com valores semânticos entre eles. Esse objeto composto que resulta da agregação configura outra publicação cujo autor é o pesquisador que reuniu os objetos de pesquisa e estabeleceu as relações semânticas entre eles.

Dessa forma, *links* entre objetos informacionais que portem significados semânticos podem ser criados diretamente por pesquisadores formando agregações na forma de

publicações ampliadas. Essas agregações podem ser registradas com indicação explícita de quem é o autor da agregação e responsável por explicitar semanticamente conhecimento, opiniões profissionais, comentários ou hipóteses científicas reveladas pelos *links* semânticos (PARINOV, 2013). O registro das relações é acompanhado pela referência à taxonomia de relações, que o vinculam a ontologias e vocabulários científicos que conferem significado a este registro.

A agregação formada pelos *links* pode ser depositada por seus autores no Repositório Digital como um recurso independente e identificado por URI própria, cujo acesso depende das políticas praticadas pela instituição, mas que no modelo proposto deve ser livre e disponível via interface *Web*.

7.3.2 Fluxo de autoria de publicação ampliada

Esta seção vem apresentar o fluxo que um autor deve seguir na criação de uma publicação ampliada, mostrando as ações necessárias desde a identificação do objeto, passando pelo estabelecimento de relações, até a adição de comentários conforme necessidade do autor.

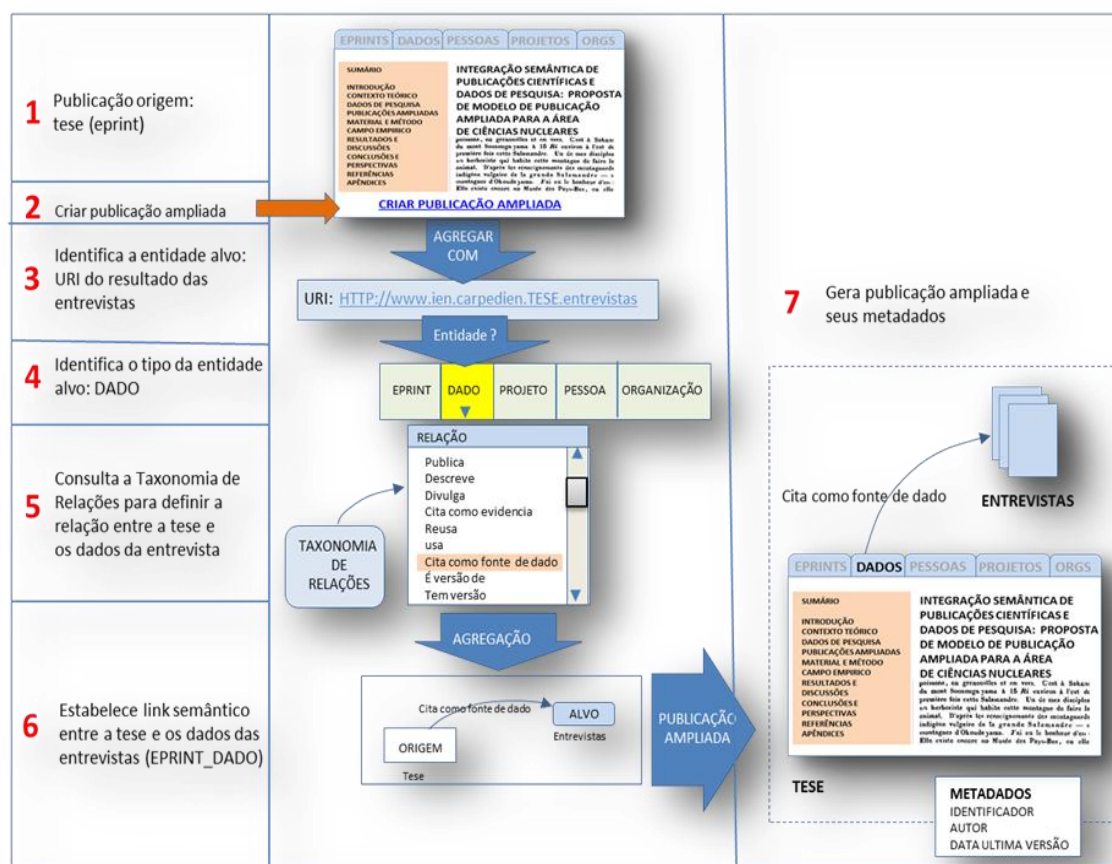
A Figura 25 representa o fluxo de ações requeridas para o registro de relações entre objetos de pesquisa no ambiente do repositório do IEN. Essas ações são controladas pela interface de autoria/edição de publicações ampliadas. Para melhor compreensão, foi tomado como exemplo os objetos de pesquisa gerados pelo presente estudo, ou seja: a tese publicada (E-PRINT) e o resultado das entrevistas (DADO). As ações requeridas para compor uma publicação ampliada são as seguintes:

- Ação 1 – a interface exibe um objeto de pesquisa (*e-print*, dado, projeto) acessado via mecanismos de recuperação do repositório. No exemplo é exibido pelo sistema a presente tese.
- Ação 2 - a interface oferece a opção da criação de uma publicação ampliada ancorada no objeto apresentado.
- Ação 3 – a interface solicita a identificação, na forma de URI, do recurso que se quer vincular à tese. Esse recurso é denominado entidade-alvo.
- Ação 4 – a interface solicita identificar o tipo da entidade-alvo: e-PRINT, DADO, PROJETO, PESSOA ou ORGANIZAÇÃO.

- Ação 5 – identificado o tipo da entidade-alvo, é definido o tipo de relação - no exemplo EPRINT_DADO - a partir daí a interface consulta a Taxonomia de Relações e exibe as relações de partida já definidas. Caso as relações exibidas não contemplem a representação que o pesquisador tem em mente, há opção de registrar uma nova relação.
- Ação 6 – a interface confirma o estabelecimento do *link* semântico entre as entidades origem e alvo apresentando uma representação gráfica da agregação. No caso do exemplo, o *link* semântico entre a tese e os dados é “cita como Fonte de dados” e tem o significado definido pela ontologia correspondente.
- Ação 7 – gera a publicação ampliada e seus metadados essenciais: autor, identificação e data da última versão.
- Ação 8 – opcionalmente o autor é solicitado a explicar ou a fazer comentários sobre o valor semântico da agregação que ele está criando (não representado na Figura).

O sistema registra também os dados mínimos sobre o autor que permitam informá-lo sobre possíveis mudanças ocorridas nos objetos agregados e da necessidade de revisar o valor semântico do *link*. A Figura 24 a seguir mostra o fluxo das ações de construção da publicação ampliada em uma proposta de interface de autoria e edição de publicação.

Figura 24: Interface de criação e edição de publicações ampliadas



Fonte: A autora

Quando a relação que o pesquisador quer estabelecer não se encontra representada na Taxonomia de Relações, que é, por princípio, uma taxonomia de partida, a interface do sistema abre a possibilidade do pesquisador registrar uma nova relação. O fluxo de registro de uma nova relação encontra-se ilustrado na Figura 25. As ações necessárias são as seguintes.

- Ação 1 – o pesquisador assinala a opção de “criar nova relação”.
- Ação 2 – a interface solicita informar o tipo de relação, no exemplo a relação assinalada automaticamente é EPRINT_DADO.
- Ação 3 – a interface solicita entrar como a relação, por exemplo: “confirmado por”.
- Ação 4 – a interface solicita informar se a relação foi extraída de alguma ontologia ou vocabulário conhecido pelo pesquisador.
- Ação 5 – opcionalmente o pesquisador é solicitado a justificar ou fazer comentários sobre o registro da nova relação.

Figura 25: Interface do fluxo de registro de nova relação

1 Criar nova relação ▶

2 CRIAÇÃO DE RELAÇÃO

TIPO DE RELAÇÃO	
ORIGEM	ALVO
☒ EPRINT	☐ EPRINT
☐ DADO	☒ DADO
☐ PROJETO	☐ PROJETO
☐ PESSOA	☐ PESSOA
☐ ORG	☐ ORG

3 **RELAÇÃO**

Entre com a relação proposta

4 **ONTOLOGIA**

Indique se a relação pertence a alguma ontologia/vocabulário (pode indicar mais de um)

☐ CERIF ☐ CITO ☐ DoCo ☐ DataCite

☐ Outra ☐ Nenhuma

5 **JUSTIFICATIVA**

SUBMETER

Fonte: A autora

Considerando que tanto as agregações que caracterizam as publicações ampliadas como as relações semânticas podem ser criadas no ambiente do repositório de modo descentralizado, os procedimentos de registro desses recursos implicam também na implantação de dispositivos de moderação e validação – que incluem filtros éticos e gestão de direitos - para determinar quais os recursos que devem estar disponíveis publicamente.

Além disso, as relações podem ser registradas de forma independente da criação da publicação ampliada. As relações que forem validadas pelos gestores do repositório vão incrementar a Taxonomia de Relações.

Dessa forma, alinham-se aos requisitos mais gerais os requisitos específicos que definem a aplicação no âmbito do IEN, que implica na adoção de *links* entre os

componentes da publicação ampliada que possam ser portadores de significado, ou seja, *links* semânticos que estejam associados a vocabulários e ontologias científicas de relevância para os pesquisadores no domínio do Instituto. A Taxonomia de Relações é o instrumento concebido para cumprir esse papel.

A pesquisa tornou claro que é possível e, muitas vezes necessário, registrar as relações semânticas como recurso informacional, identificadas por URI, com autoria reconhecida e associada a algum vocabulário semântico de partida, tornando-as publicamente disponíveis para uso no ambiente do Repositório. As relações tornadas disponíveis se tornam ferramentas que permitem aos pesquisadores reconstruírem significados e delinearem novas unidades de pensamento a partir da vinculação de objetos de pesquisa já registrados, definindo um novo objeto intelectual. Os cientistas podem trabalhar com os objetos de pesquisa como blocos de encaixe, expressando sua opinião e estabelecendo novas formas, novos objetos intelectuais que podem ser compartilhados e dessa forma definindo novos modelos para ideias de pesquisa e para problemas reais.

A padronização das relações, concretizada pela Taxonomia de Relações, e a disponibilidade pública via interface *Web* dos seus termos e estrutura, abrem a possibilidade inédita para que a cada pesquisador individualmente ou em grupo expresse sua opinião sobre os múltiplos relacionamentos que podem existir entre as entidades do seu domínio específico de pesquisa e possa compartilhar a sua experiência. Isso possivelmente colabora para a criação de um ambiente propício à criatividade científica.

Como exemplo de um desdobramento prático o pesquisador pode construir novas unidades de pensamento e novas formulações padronizadas para áreas específicas, por exemplo, um tipo de publicação ampliada para a área de realidade virtual voltado para treinamento de operadores de reatores nucleares que agregue cartilha, aulas, blogs, ambientes virtuais, instrutor, curso, etc., que poderia ser renderizada na forma de multimídia.

7.3.3 *Interface de Apresentação de publicações ampliadas*

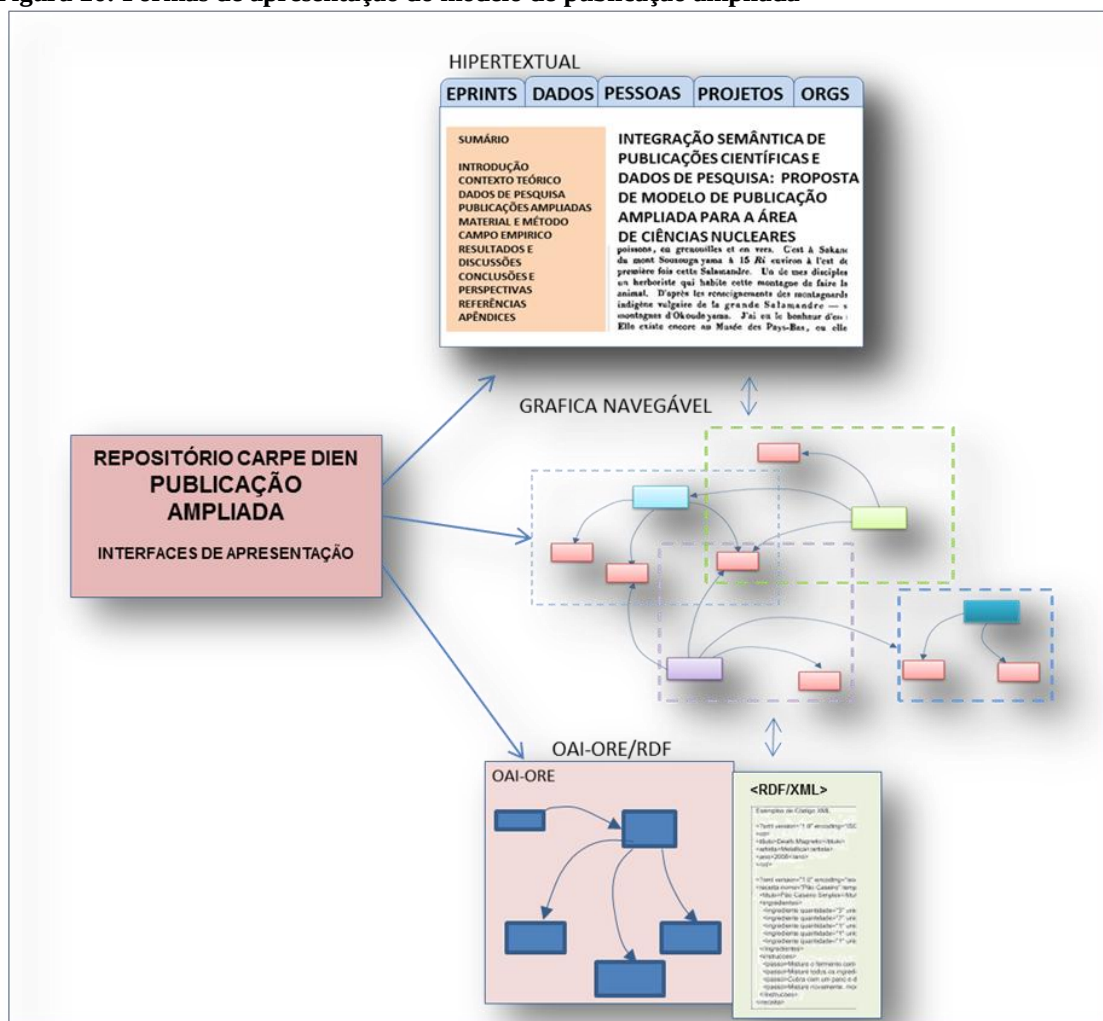
A visualização via interface *Web* das agregações e dos valores semânticos dos *links* é uma finalização importante para o modelo proposto. A possibilidade de haver uma manifestação gráfica da agregação que seja navegável é um complemento que pode contribuir para o avanço dos serviços oferecidos pelos repositórios digitais, principalmente

para aqueles que gerenciam dados de pesquisa e *e-prints*, que é precisamente o caso do repositório do IEN.

No ambiente do repositório Carpe diEN, quando um item de informação for recuperado, será informado se faz parte da agregação de uma ou mais publicações ampliadas; caso positivo, será oferecida a possibilidade de recuperar as agregações ou apenas o item que interessar ao usuário.

O modelo de publicação ampliada pode se manifestar em três tipos de formas de visualização: hipertextual, gráfica e no formato OAI-ORE/RDF. Essas manifestações estão representadas na Figura 26, e descritas a seguir.

Figura 26: Formas de apresentação do modelo de publicação ampliada



Fonte: A autora

A) Apresentação hipertextual

Está ancorada num *e-print* reformatado num padrão de documento hipertextual que apresenta abas indicando as entidades alvo que compõem a publicação ampliada. Para cada entidade alvo é apresentado um menu com as relações semânticas existentes que são clicáveis e conduzem à entidade alvo correspondente.

B) Apresentação gráfica

Mostra numa visão gráfica, a totalidade das agregações da publicação ampliada explicitadas, incluindo as vinculações existentes entre as diversas publicações ampliadas que estão vinculadas entre si. O gráfico permite a navegação entre a

totalidade dos recursos agregados, compreendendo a navegação entre publicações ampliadas, formando, assim, uma rede semântica clicável.

C) Apresentação em XML, no formato de agregação OAI-ORE

Como o OAI-ORE define padrões para a descrição e intercâmbio de agregações de recursos *Web*, o formato pode ser usado para descrever publicações ampliadas e facilitar o intercâmbio entre instituições. Além disso, a serialização em RDF permite o *harvesting* da descrição via o protocolo OIA-PMH (WOUTERSEN-WINDHOUWER; BRANDSMA, 2009, p.137)

É importante destacar que em todas as manifestações, os seguintes dados são apresentados de acordo com cada tipo de entidade que está sendo agregada:

- Identificação da publicação ampliada;
- Identificação de cada objeto que compõe a publicação ampliada;
- Autoria da publicação ampliada;
- Autoria de cada componente;
- Dados sobre o versionamento da publicação ampliada e dos objetos componentes;
- Relações semânticas entre pares de objetos.

Com o intuito de melhor compreensão, a Figura 27 exhibe o exemplo de uma manifestação hipertextual de uma possível publicação ampliada ancorada na presente tese. Os objetos de pesquisa e alguns dos atores envolvidos na agregação são apresentados com os respectivos *links* semânticos.

Uma questão que também se coloca nesta tese é a necessidade de curadoria dos dados de pesquisa que deverão estar integrados na formação de uma publicação ampliada. Esta necessidade, porém extrapola os limites de um domínio, como é o caso das Ciências Nucleares, e requer diretrizes que possam ser seguidas em âmbito nacional.

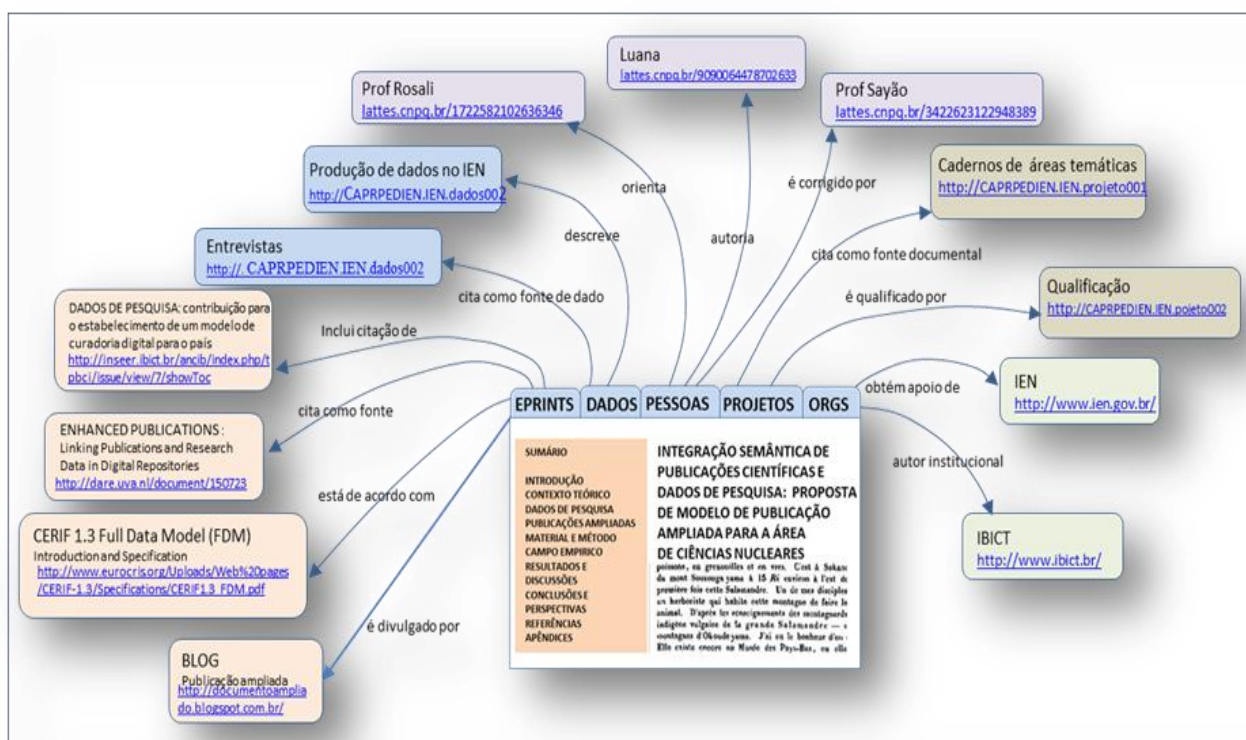
Essa necessidade se justifica por que o modelo de publicação proposto por esta tese é, em uma primeira instância, um objeto digital complexo, podendo ser formado por meio de relacionamentos entre recursos distribuídos em diversos repositórios digitais, saindo assim do alcance de uma única instituição. Uma publicação ampliada, enquanto um objeto digital complexo, pode se configurar como a combinação de um *e-print* depositado em um repositório institucional, dados de pesquisa depositados em um repositório de dados e

ainda outros documentos depositados em qualquer outra base de dados de informação técnico-científica.

Neste sentido, propor diretrizes para uma política de curadoria digital para o país significa estabelecer em esfera nacional, normas e padrões comuns que propiciem a cooperação entre as instituições de pesquisa e a interoperabilidade entre esses sistemas que abrigam os recursos digitais de pesquisa (*e-prints*, dados de pesquisa, etc). Dito de outra maneira, esses recursos digitais precisam estar tratados adequadamente para que possam servir como elementos para serem agregados na formação de uma publicação ampliada.

Apesar de essa pesquisa estar voltada para aplicação em um domínio específico, a proposição do modelo deverá considerar explicitamente as estruturas envolvidas na formulação de políticas de C&T no país e aspectos geopolíticos, como legislação, financiamento, pesquisa, educação, capacitação, infraestruturas tecnológicas e de gerenciamento de C&T etc. A infraestrutura concebida como contextualização do modelo proposto serve também para prover uma contribuição para a formação de um ambiente de curadoria para os países. Assim, enquanto uma política nacional não é estabelecida, as diretrizes propostas podem servir como elementos norteadores para as instituições locais que pretendem dar algum tipo de tratamento aos dados de pesquisa, sejam para fins de preservação, de gestão, de disseminação via novos modelos de publicação ou qualquer outra finalidade.

Figura 27: Exemplo de publicação ampliada e respectivas relações semânticas



Fonte: A autora

Textualmente pode-se descrever a publicação ampliada apresentada na Figura 27 do exemplo da seguinte forma: A tese (E-PRINT) tem autoria de Luana (PESSOA); foi desenvolvida a partir do projeto de qualificação (PROJETO) no âmbito do programa de pós-graduação do IBICT (ORGANIZAÇÃO) e com o apoio institucional do IEN (ORGANIZAÇÃO); teve como orientador a Profa. Rosali (PESSOA) e contou com correções do Prof. SAYÃO (PESSOA); a tese descreve o levantamento realizado sobre a produção de dados do IEN (DADOS) e utilizou também como Fonte de dados entrevistas com pesquisadores (DADOS) e os cadernos de áreas temáticas do IEN (PROJETO); a tese está baseada no relatório “Enhanced Publications: Linking Publications and Research Data in Digital Repositories” (E-PRINT) e está de acordo com o padrão “CERIF 1.3 Full Data Model (FDM): Introduction and Specification” (E-PRINT); além disso inclui trecho do artigo “Dados de pesquisa: contribuição para o estabelecimento de um modelo de curadoria digital para o país” (E-PRINT); a pesquisa é divulgada pelo blog “Documentos Ampliados” (E-PRINT).

Tomando um grau a mais de formalização, traduzindo as relações presentes (sublinhadas) por meio da Taxonomia de Relações, e identificando as entidades por meio dos seus URIs o quadro 12, apresenta todos os elementos que compõem a publicação ampliada, exemplificada na Figura 27.

Quadro 12: Elementos que compõem o exemplo da publicação ampliada da pesquisa-tese

PUBLICAÇÃO AMPLIADA	
AUTOR: Luana Farias Sales	
TITULO: Pesquisa tese sobre publicações ampliadas para a área de ciências nucleares	
IDENTIFICADOR:CARPEDIEN.IEN.pesquisa_tese_luana	
DATA DA ÚLTIMA VERSÃO: 23junho de 2014	

ENTIDADE FONTE	
AUTOR: Luana Farias Sales	
TITULO: Integração semântica de publicações científicas e dados de pesquisa: proposta de modelo de publicação ampliada para a área de ciências nucleares	
TIPO: EPRINT	

ENTIDADE SALVO					
ENTIDADE	TIPO	IDENTIFICADOR	CLASSE DE RELACIONAMENTO	TAXONOMIA DE RELAÇÕES	
				RELAÇÃO	ONTLOGIA/VOCABULÁRIO
Luana	PESSOA	lattes.cnpq.br/9090064478702633	EPRINT_PESSOA	Autoria	VOC_IEN
Profa. Rosali	PESSOA	lattes.cnpq.br/1722582102636346	EPRINT_PESSOA	Orienta	CITO
Prof. Sayão	PESSOA	lattes.cnpq.br/3422623122948389	EPRINT_PESSOA	É corrigido por	CITO
Produção de dados do IEN	DADO	CAPRPEDIEN.IEN.dados001	EPRINT_DADO	Descreve	CITO
Entrevistas	DADO	CAPRPEDIEN.IEN.dados002	EPRINT_DADO	Cita como Fonte de dados	CITO
Caderno de áreas temáticas	PROJETO	CAPRPEDIEN.IEN.projeto001	EPRINT_PROJETO	Cita como Fonte documental	CITO
Projeto de qualificação	PROJETO	CAPRPEDIEN.IEN.projeto002	EPRINT_PROJETO	Resulta em	VOC_IEN
IEN	ORG	www.ien.gov.br	EPRINT_ORG	Obtém apoio de	CITO
IBICT	ORG	www.ibict.br	EPRINT_ORG	Autor institucional	CERIF
Dados de pesquisa: contribuição p/ o estabelecimento de um modelo de	EPRINT	http://inseer.ibict.br/ancib/index.php/tpbci/issue/view/7/showToc	EPRINT_EPRINT	Inclui trecho de	CITO

curadoria digital para o país					
Enhanced Publications : Linking Publications and Research Data in Digital Repositories	EPRINT	http://dare.uva.nl/document/150723	EPRINT_EPRINT	É base de	Data Cite
CERIF 1.3 Full Data Model (FDM) Introduction and Specification	EPRINT	http://www.eurocris.org/Uploads/W eb%20pages/CERIF-1.3/Specifications/CERIF1.3_FDM.pdf	EPRINT_EPRINT	Concorda com	CITO

Fonte: A autora

Este Quadro é apenas outra maneira de apresentar a publicação ampliada, mas poderia se configurar também como um formulário de entrada de dados para composição de uma publicação ampliada em um sistema de informação. A subseção a seguir vem mostrar que o modelo proposto, além de integrar informações de diversos repositórios, por meio das relações, pode também oferecer uma série de serviços, o que expande ainda mais as vantagens de uso dessa nova forma de publicação.

7.3.4 Serviços integrados ao modelo de publicações ampliadas

Muito além dos serviços básicos de busca e recuperação oferecidos pelos repositórios digitais, a camada de registro de relações e de edição, autoria e apresentação de publicações ampliadas expande a gama de serviços que pode ser oferecido aos pesquisadores e à sociedade em geral.

O modelo proposto pelo estudo tem um olhar mais atento sobre as questões de dados de pesquisa. O fato dos dados formarem agregações na forma de publicação ampliada cria um compromisso institucional de armazená-los e preservá-los por longo prazo, para que seja mantida a integridade desse novo objeto de informação. A característica de variar com o tempo adiciona um grau a mais de complexidade aos dados de pesquisa e exige uma gestão dinâmica que só é possível por meio das ferramentas, padrões e boas práticas oferecidas pelas metodologias da curadoria digital de dados de pesquisa. Portanto, a curadoria digital é um serviço importante que deve estar disponível pelos sistemas de informação para a pesquisa.

Dessa forma, os dados de pesquisa que antes estavam armazenados em mídias e servidores isolados e sujeitos a perdas irreversíveis, têm a sua visibilidade multiplicada

quando agregados às publicações científicas por meio de *links* semânticos. Este fato determinante amplia o potencial cognitivo – entendido como a capacidade do objeto ser interpretado agora e no futuro – e, naturalmente, o seu potencial de reuso por todas as áreas temáticas, quando se fala especificamente do IEN. A oferta dos dados de pesquisa, tratados tecnicamente por metadados apropriados e contextualizados pelos demais objetos agregados pela publicação ampliada, constitui um novo serviço desenvolvido sobre os dados de pesquisa.

Em relação à geração de indicadores de atividade científica, Parinov e Kogalovsky (2014) - no âmbito de suas pesquisas sobre *links* semânticos voltados para estudos cientométricos – esclarecem que: “o monitoramento e o processamento dos *links* semânticos criados nos contextos de um repositório digital produz uma Fonte de dados para novos estudos cientométricos [...]”. Complementarmente aos indicadores já existentes, os dados provenientes desses repositórios permitem uma inspeção profunda sobre as características de impacto e de uso de dados e informações por cientistas e organizações, gerando novas modalidades de indicadores. A cadeia de relações que é exposta pelas ligações semânticas entre as diversas entidades, produtos de pesquisa e atores, expõe mais claramente a dinâmica dos processos envolvidos. Uma melhor visualização dos usos dos produtos de pesquisa podem aprimorar o ciclo de comunicação científica e dar maior consistência às métricas de avaliação da atividade de pesquisa.

A cadeia de ligações semânticas entre as entidades envolvidas mostra com clareza os produtos gerados por instituições, laboratórios, projetos e pesquisadores; os desdobramentos em artigos, livros, materiais didáticos, *software* e patentes; os financiadores; as entidades intervenientes externas e ainda o possível retorno social das atividades de pesquisa. O processamento por computador dessas informações pode criar instrumentos de gestão inéditos, que se compatibilizam com os serviços preconizados pelo padrão CRIS – *Current Research Information System*.

A apresentação dos produtos de pesquisa e de seus principais atores agregados semanticamente exibem uma contextualização que pode estar na forma de serviços de informação úteis para as esferas sócias fora do mundo científico e acadêmico, ou para instituições ou pessoas que tenham alguma responsabilidade política, gerencial, filosófica, educacional ou financeira em relação à pesquisa científica, ou para o cidadão – que contribui com seus impostos para a manutenção de toda a estrutura - que deseja entender mais claramente o que se passa no misterioso e complexo mundo da ciência.

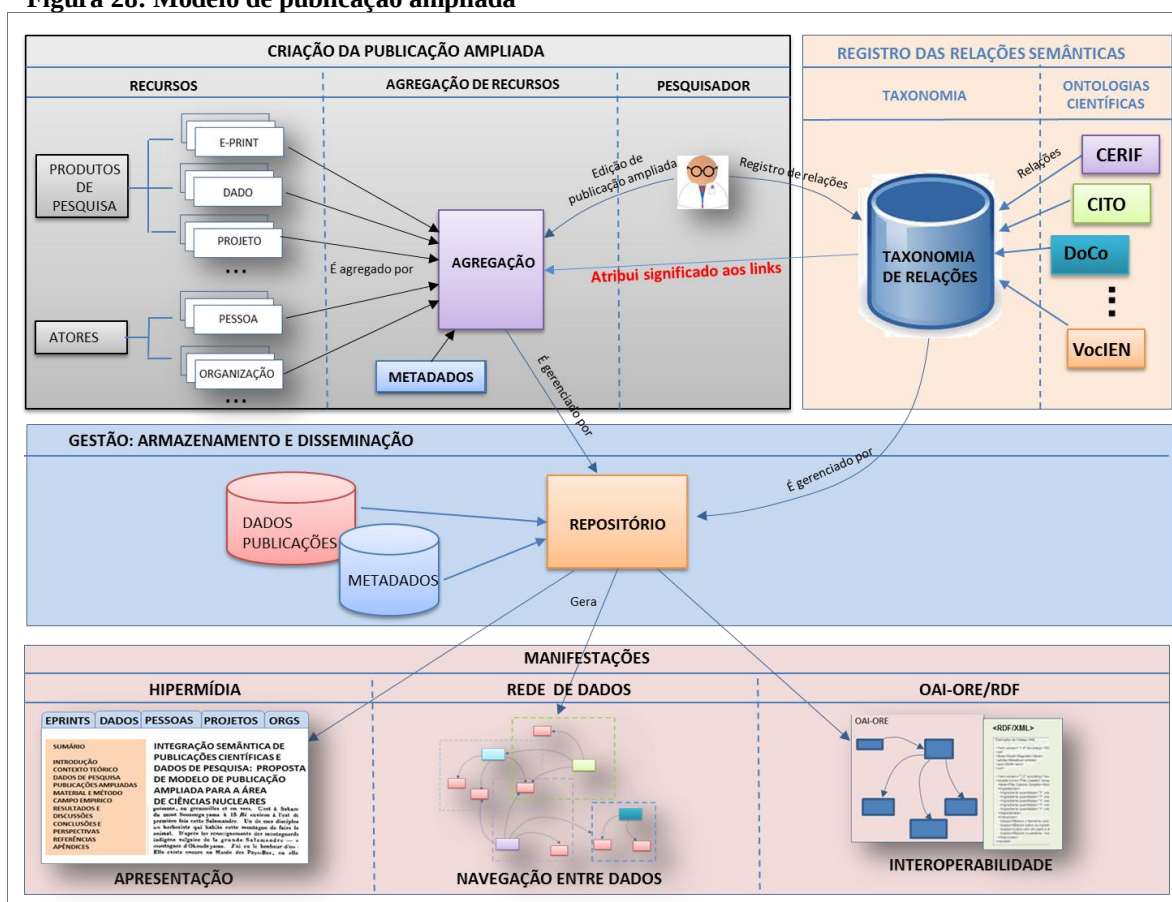
Por fim, é importante considerar que os serviços podem ser criados de acordo com as necessidades específicas de cada área ou de cada instituição. Assim, a seção a seguir apresenta o modelo de publicação ampliada construído para o contexto do IEN.

7.4 MODELO DE PUBLICAÇÃO AMPLIADA PARA O IEN

Nesse ponto em que todas as entidades, relações, interfaces e fluxos estão estabelecidos, a abordagem descritiva adotada até aqui para a formulação de um modelo conceitual de publicação ampliada para a área de ciências nucleares - instanciada pelo IEN - pode ser expressa por representações gráficas. Estas representações se encaminham na direção de ilustrar mais claramente que se confirma a hipótese de que se pode formular publicações ampliadas que expressem com mais intensidade a complexidade e a sofisticação dos processos de pesquisa da área nuclear, e dos seus vários atores envolvidos, através de processos que permitam conferir significados consistentes às relações entre os diversos recursos agregados.

A Figura 28 representa – com as limitações de uma linguagem não formal – a arquitetura de um sistema voltado para a criação, gestão e apresentação de uma publicação ampliada genérica. Junta-se ao modelo a camada de registro de relações semânticas que adiciona qualidade de expressar semanticamente as relações entre os recursos de informação que estão agregados, que é uma contribuição importante da presente pesquisa. O dispositivo que padroniza as relações e as estruturas em forma de uma taxonomia toma emprestado as relações já padronizadas das principais ontologias científicas, garantindo consistência na atribuição de semântica aos *links*, tornando-os *links* semânticos, enquanto abre também a possibilidade de interoperabilidade com outros sistemas.

Figura 28: Modelo de publicação ampliada

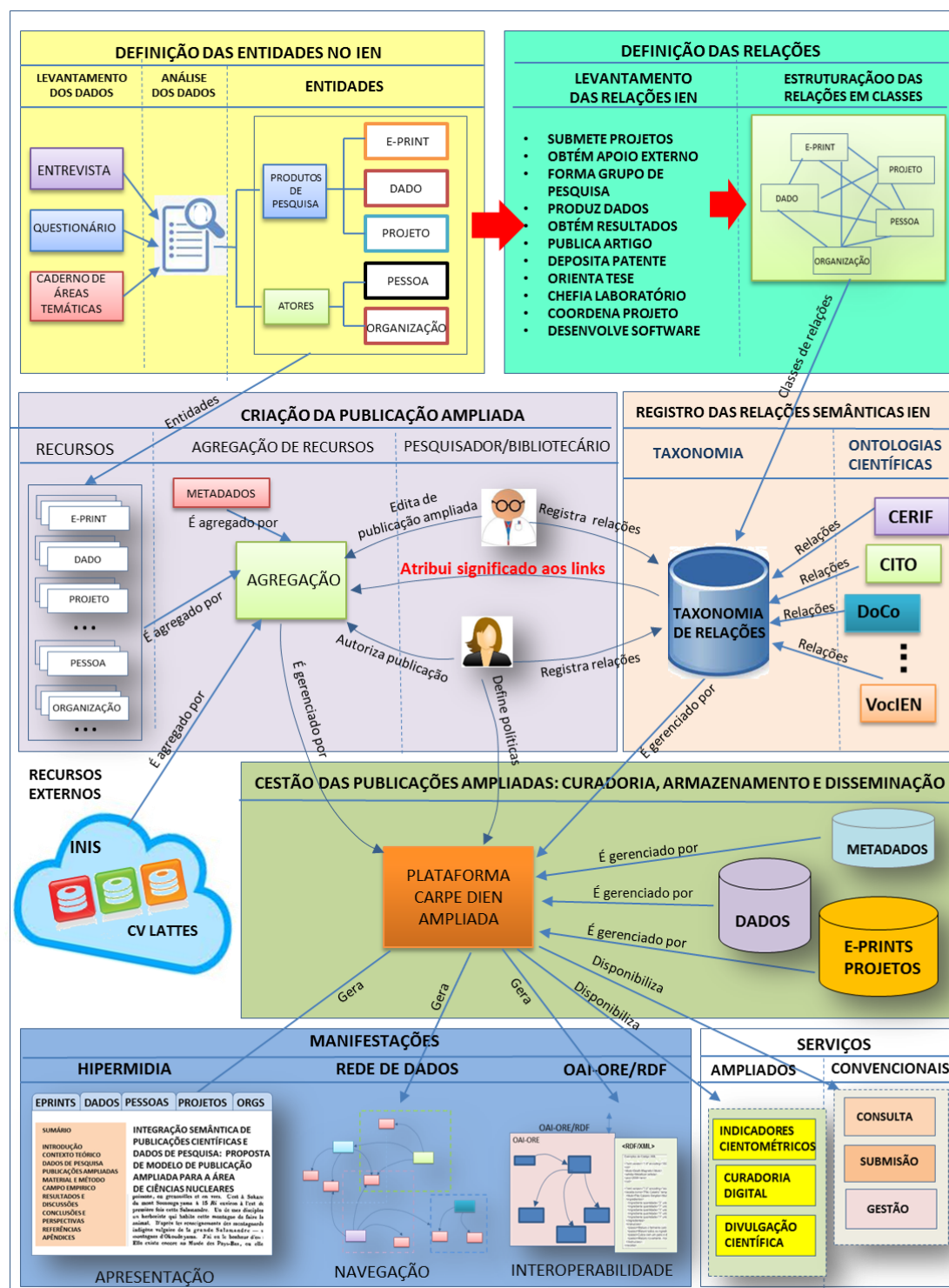


Fonte: A autora

Esta primeira Figura 28 ilustra a base para a definição dos fluxos na criação de publicações ampliadas que atendam as especificidades da área nuclear representada pelo IEN, na qualidade de campo empírico.

Conforme já discutido anteriormente e demonstrado pelo levantamento de dados, a publicação ampliada para o IEN deve atender algumas especificidades em termos de produtos de pesquisa e atores envolvidos e, sobretudo, as relações semânticas que se estabelecem entre essas entidades. Essas especificidades devem estar padronizadas na forma de uma taxonomia. Dessa forma, tomando como base a representação da Figura 28, é proposto o modelo representado pela Figura 29, onde novas camadas são adicionadas para expressar os requisitos do IEN.

Figura 29: Fluxo de construção de publicação ampliada para o IEN



Fonte: A autora

Analisando brevemente cada uma delas, tem-se:

1. DEFINIÇÃO DAS ENTIDADES DO IEN – representa o processo de análise do levantamento de dados (questionário, entrevistas, cadernos temáticos, conversas registradas, etc.) para a definição das entidades de partida que farão parte do modelo; para isso são considerados também os padrões adotados pelos principais sistemas voltados para a pesquisa.
2. DEFINIÇÃO DAS RELAÇÕES ENTRE AS ENTIDADES – a partir do levantamento dos fluxos de pesquisa do IEN é definido um conjunto de relações que se estabelecem entre produtos de pesquisa e atores; essas relações são estruturadas duas a duas tomando como base o padrão CERIF; o resultado dessa operação é o conjunto de CLASSES DE RELACIONAMENTOS que servirão de base para a estruturação de uma taxonomia.
3. REGISTROS DAS RELAÇÕES SEMÂNTICAS DO IEN – as relações são então formalizadas tomando como base as ontologias científicas mais conhecidas e as relações específicas do IEN denominada VOC-IEN; essas relações são estruturadas, tomando como base as CLASSES DE RELACIONAMENTOS, na forma de uma taxonomia denominada “TAXONOMIA DE RELAÇÕES”. Esta operação, como já vista, é realizada através de uma interface própria, definida em 7.3.2, Figura 25, pelo pesquisador e/ou bibliotecário.
4. CRIAÇÃO DA PUBLICAÇÃO AMPLIADA - os recursos informacionais (dados, *e-prints*, projetos) e os atores (pessoas e organizações) são agregados e a relação entre eles é assinalada pelo pesquisador tomando como base a Taxonomia de Relações; recursos externos, como dados da Plataforma Lattes e da base de dados do INIS podem ser ligados via URI; o bibliotecário tem a prerrogativa de liberar a agregação; esta operação é realizada por uma interface própria definida em 7.3.2, figura 24, chamada de interface de criação e edição de publicação ampliada; são assinalados nessa camada os metadados mínimos necessários à gestão e à recuperação da publicação ampliada.

5. GESTÃO DE PUBLICAÇÃO AMPLIADA – nessa camada as publicações ampliadas são gerenciadas tendo em vista o armazenamento, a preservação, a recuperação e o acesso; isso inclui também os metadados associados à publicação. Para o IEN, essa gestão será realizada pelo ambiente Carpe diEN ampliado como novas interfaces para operar com as publicações ampliadas.
6. MANIFESTAÇÕES E SERVIÇOS – as publicações ampliadas podem se manifestar de diversas maneiras: em formato de hipermídia voltado para a apresentação; em formato de rede de dados em que várias publicações ampliadas estão interligadas formando uma rede semântica ; e padrão de objetos compostos OAI-ORE/RDF voltados para a interoperabilidade e compartilhamento com outros sistemas e compatibilização com a *Web* semântica; o sistema deve também prover serviços básicos de acesso, recuperação e submissão e ainda proporcionar serviços inovadores a partir dos recursos que estão agregados pela publicação ampliada, como por exemplo, análises cientométricas e geração de indicadores.

A exequibilidade que o modelo reafirma está diretamente ligada à questão da viabilidade técnica de se implementar o conceito proposto com as ferramentas tecnológicas disponíveis no Instituto de Engenharia Nuclear – como banco de dados, linguagem de *software*, computadores em rede e armazenamento em massa seguro. Dado que já existe em ambiente próprio proporcionado pelo repositório de dados e informações Carpe diEN, essa questão pode ser superada. Outra questão importante está relacionada com a necessidade de se construir um sistema que esteja em consonância com os padrões internacionais de interoperabilidade, representação e preservação, e que esteja próximo aos novos modelos de informação para a pesquisa. Esse requisito foi tomado em consideração todo o tempo. Por fim é necessário *expertise* para fazer a gestão de um sistema complexo, o que está acima das questões tecnológicas e redimensiona o papel do profissional da informação.

A metodologia baseada em ontologias consagradas, enriquecida com relações de um vocabulário próprio mostrou-se adequada ao propósito de conferir significado aos *links*, posto que esses instrumentos oferecem um vasto espectro de relações que cobrem quase que a totalidade de valores semânticos de partida necessários aos sistemas científicos. Mas para isso houve a necessidade de estruturação e classificação dessas

relações tendo em vista a futura implementação em base de dados, o que foi realizado no decorrer da pesquisa.

Por fim, ao adicionar a arquitetura clássica de um repositório institucional – composto de módulos de gestão, armazenamento, recuperação e acesso – uma camada que permite “*linkar*” os recursos, proporciona um ganho sensível na capacidade dos recursos agregados transmitirem com mais precisão informação e conhecimento; mas atribuir valores semânticos, padronizados e provenientes de vocabulários científicos a esse *links*, significa oferecer ao pesquisador uma nova ferramenta, que extrapola os limites dos sistemas convencionais, sem abdicar deles, para expressar a sofisticação do seu trabalho de pesquisa.

O objetivo último da presente tese foi precisamente provar que as relações semânticas constituem uma ferramenta adequada para ser incorporada aos sistemas de informação para a pesquisa, como um dispositivo essencial para se alcançar um novo patamar de desenvolvimento para esses sistemas, mais próximos do usuário, da pesquisa contemporânea e da sociedade.

8 CONCLUSÃO E PERSPECTIVAS FUTURAS

O ambiente onde a presente pesquisa se desenvolveu está sob as condições de um paradigma científico emergente, conhecido como *eScience* ou 4º paradigma, que é em si um instigante desafio para a Ciência da Informação. Paradoxalmente, isto não significa que os problemas sejam realmente novos para a Ciência da Informação, como o é para outras disciplinas como Ciência da Computação. Porém estes problemas devem se repensados sob uma ótica de investigação nova, buscando desenvolver soluções teóricas práticas para a integração e a interoperabilidade entre os sistemas de informação para a pesquisa científica, que tenham como referencial um usuário que detém conhecimentos profundos das tecnologias digitais e que tem acesso as ferramentas computacionais avançadas.

Nesse contexto de mudanças velozes, impulsionados pelas tecnologias digitais e redes de computadores, antigos problemas que são objetos de pesquisa que deram origem à Ciência da Informação – como a explosão da informação e a recuperação de informação – passaram a ter sua complexidade ampliada e redefinida sob essas novas condições. O mais significativo dos problemas, e o que cotidianamente nos assombra, é a constatação de que um obstáculo importante ao avanço da ciência contemporânea não é a escassez de dados, mas sim o seu oposto, ou seja, o excesso de dados e informações que estão disponíveis à procura de significados de que precisam para transmitir conhecimento ao longo do espaço e do tempo. Para isso não basta que esses recursos sejam publicados na *Web*, não basta somente a tecnologia, é necessário fundamentalmente que sejam tratados, descritos, integrados, ou seja, que passem por processos técnicos que confirmem semântica e estrutura a eles. Os *bits* não falam por si próprios e precisam para isso de esquemas de representação. Mas isso sempre fez parte do elenco de conhecimentos, metodologias e ferramentas da Ciência da Informação. Este fenômeno do nosso tempo, potencializado por dispositivos, ferramentas, sensores, satélites, escalas, aceleradores de partículas de amplitude internacional, *software* de simulação que produzem as torrentes de dados que esboçam as novas frentes de pesquisa da *eScience* e do *Big Data*, recolocam também, de forma contundente, questões críticas para a Ciência da Informação.

O problema conceitualmente mais relevante de que tratou esta pesquisa foi o de investigar como integrar de forma semântica os sistemas de informação tradicionais baseado em publicações – impressas ou digitais – a um universo informacional cujos dados

de pesquisa, com sua natureza diversificada, complexa, heterogênea e distribuída, são determinantes. O que fica evidente é que a riqueza dos resultados da pesquisa contemporânea não se esgota potencialmente num texto, ou num arquivo em formato PDF. O texto não consegue contar toda a história. Os produtos de pesquisa de hoje têm movimento, interação, estão em 3D, estão distribuídos por muitos repositórios, estão em bases de dados e dependem de programas de computador e equipamentos específicos para serem executados e apresentados, têm versões, manifestações e variam no tempo. O problema tratado aqui, resumidamente, foi como conectar semanticamente esses dois paradigmas: um orientado por documentos textuais e unidimensionais e outro orientado por dados distribuídos, compostos e assentados sobre uma infinidade de mídias digitais. A partir dessa questão abre-se um mundo de novas pesquisas em que a ótica e o referencial histórico, epistemológico, teórico e prático da Ciência da Informação prevalecem fortemente como um ferramental importante.

A solução proposta para essa integração é um conceito novo de publicação que considera os dados de pesquisa como um recurso valioso, mas que ainda está fortemente ancorado nas publicações tradicionais como artigos e teses. O que se propôs investigar foi como ampliar essa publicação tradicional, enriquecendo-a com dados de pesquisa. A ideia inerente à concepção da publicação ampliada é precisamente criar elos que liguem os conteúdos dos repositórios institucionais e temáticos, que armazenam as publicações científicas mais tradicionais, com os conteúdos dos repositórios de dados de pesquisa.

A vinculação entre *e-prints* e dados abriu oportunidades de investigação de inúmeras questões, tais como: como conferir semântica as ligações entre dados e *e-prints*? Como ampliar elementos que compõem a publicação ampliada com os atores que interagem no processo de investigação científica? Como formalizar essas relações usando ontologias científicas conhecidas? Como disponibilizar essa metodologia na forma de interfaces *Web* e taxonomia para que o pesquisador monte e remonte e publique as suas publicações ampliadas, como um novo artefato intelectual, expressando ideias, opiniões, hipóteses, ou seja, reinterpretando os dados?

A área nuclear, especialmente a área de Engenharia Nuclear do IEN se mostrou adequada como campo empírico: interdisciplinar por natureza, a área produz e usa uma quantidade extraordinária de dados de pesquisa que são extremamente diversificados e distribuídos e esta implantando um repositório cujo formato de metadados contempla *e-prints* e dados de pesquisa.

Na trajetória em direção a uma publicação científica mais rica semanticamente fica evidente que o fazer científico cria diferentes tipos de relações entre os produtos de pesquisa que podem ser considerados recursos de informação se devidamente padronizados – na forma de ontologias, por exemplo - e incorporados aos sistemas de informação tradicionais. Como se observou, algumas dessas relações são claras, por exemplo, autoria. No entanto, as mais ricas semanticamente não estavam explícitas e existiam somente de forma subjetiva e precisavam ser reveladas e padronizadas para domínios específicos. Disponibilizar uma ferramenta para que o pesquisador formalizasse a sua opinião sobre os múltiplos relacionamentos que existiam entre os objetos de pesquisa, abre novas possibilidades para a criatividade científica traduzida na forma de publicações *Web* que podem ser consideradas como redes semânticas navegáveis. Dito de outra maneira, a possibilidade do pesquisador criar *links* semânticos entre objetos de pesquisa oferece a oportunidade de se fazer uma intervenção intelectual e interagir sobre algo que na publicação tradicional só poderia ser vista pelo leitor de maneira fragmentada. Isso muda a visão do pesquisador sobre o que é uma publicação científica.

Neste sentido, a hipótese defendida no início desta pesquisa de que uma publicação pode ser ampliada a partir de um modelo que ligue os dados de pesquisa ao documento final por meio de relações que expressem uma semântica consistente, se confirma. Além disso, soma-se a essa constatação a possibilidade das relações semânticas estabelecidas na formação de uma publicação ampliada abrirem uma oportunidade para que os cientistas possam trabalhar com os objetos de pesquisa como blocos de montar do tipo “lego”, criando objetos intelectuais que podem ser compartilhados em áreas menos científicas e mais técnicas e educacionais. Por exemplo: um tipo de publicação ampliada para treinamento de operadores de reatores nucleares baseado em realidade virtual agregando semanticamente manuais, dados, ferramenta de *software*, ambientes virtuais, tipos de usuários e instrutores.

Os percursos em torno da composição da publicação ampliada para a área nuclear tornou aparente alguns conceitos, padrões e tendências internacionais voltados para sistemas de informações para a pesquisa que parecem ser importantes para o país, mas que, entretanto, se mantinham quase desconhecidos. Por conta disso, uma contribuição desta tese foi contextualizar essas referências tecnológicas e de padronização, além de considerar as facetas políticas, gerenciais e éticas, como parte da pesquisa articulando-as na forma de requisitos para um modelo de curadoria digital de dados de pesquisa para o país. Porém,

foram muitos os estudos que contribuíram para que a ideia de publicação ampliada estivesse devidamente contextualizada como um artefato intelectual.

Em primeiro lugar, a constatação de um fenômeno marcante que mudou a visão sobre a importância dos dados de pesquisa: os dados que estavam subjacentes às publicações eram apenas subprodutos dos processos de pesquisa e normalmente descartados no final dos projetos, enquanto hoje são considerados recursos essenciais, que devem estar abertos e interpretáveis em prol do progresso científico. Um conjunto de metodologias coletivamente chamado de curadoria digital de dados de pesquisa busca manter esses dados vivos, reutilizáveis, íntegros e autênticos. Este fato tem desdobramentos importantes na comunicação científica, pois esses mesmos dados que vão apoiar a revisão por pares, a veracidade das pesquisas e, sobretudo, a reprodutibilidade dos experimentos, têm um fundamento crítico nas ciências exatas. Além do mais, novas pesquisas de qualidade podem ser desenvolvidas, com a segurança necessária, a partir desses dados, que estão instrumentalizados para serem transportados e interpretados em novos domínios. Em âmbito geral, pode-se observar que uma nova relação se estabelece entre os pesquisadores na medida em que um pesquisador, para desenvolver seus projetos, pode depositar toda a confiança nos dados levantados por outro, distante no tempo e no espaço. Por outro lado, a característica de poder variar dinamicamente no tempo, registrando versões, manifestações – por exemplo, um artigo impresso e sua versão em vídeo –, linhagens de dados, quando analisado à luz da Comunicação Científica, pode ser considerada como um novo veículo de disseminação da informação científica. Como outros tipos de publicação, se estabelece no ciclo da comunicação científica com a finalidade de certificar o conhecimento científico por meio da comunicação autorizada e validada, e ainda tornando as coleções de dados digitais de pesquisa como parte importante da memória das atividades de pesquisa da instituição que as gerou.

Outro pressuposto básico que foi encarado no decorrer da pesquisa foi que a proposta deveria se compatibilizar com o percurso da *Web* na direção de um sistema com mais significado, cujas assertivas pudessem ser processadas por pessoas e por programas. O padrão OAI-ORE define as regras para isso, considerando questões chave como interoperabilidade e estruturação em triplas RDF. Por esse motivo a investigação caminhou na direção de normas e padrões internacionais que pudessem dar ao modelo proposto subsídios para que, no futuro, se tornasse adequado à proposta da *Web* semântica, que se diferencia da *Web* atual justamente por ampliar o conceito da *Web*, que inicialmente

era voltada apenas para documentos, para uma *Web* voltada para a ligação semântica de dados distribuídos.

A pesquisa realizada demonstra ser factível a criação de diálogos com significados explícitos e padronizados entre publicações acadêmicas e um universo povoado por dados de pesquisa, permite o planejamento de sistemas de informação mais próximos das necessidades e da lógica do pesquisador atual sem, entretanto, fazê-lo abdicar dos sistemas tradicionais de publicação, centrados em periódicos.

Em termos de recuperação de informação, a implementação do conceito de publicação ampliada proporciona o acesso integrado a uma publicação e a todos os produtos e atores vinculados, poupando o tempo do pesquisador; os dados, quando recuperados em conjunto, são contextualizados pela própria publicação, mantendo seu sentido original, mas podendo ser reutilizados para novas pesquisas de forma mais precisa; cria-se o compromisso por parte das instituições de pesquisa em preservar os dados de pesquisa, no sentido de manter íntegras as publicações ampliadas que referenciam esses dados; os dados podem ser citados e recuperados isoladamente e como um todo, de acordo com a finalidade da pesquisa e o desejo do usuário pesquisador; os dados recuperados em conjunto com as publicações permitem verificar a veracidade e a reprodutibilidade da pesquisa e dão mais confiabilidade aos seus resultados.

Dar visibilidade aos dados disponibilizando-os na forma de publicação ampliada oferece condições para a prática das recomendações da chamada “Ciência Aberta”, cuja preocupação primordial é tornar a atividade de pesquisa mais transparente, mais colaborativa e mais eficiente. A concepção de ciência aberta tem muitas faces e muitos significados, porém o mais importante deles é o que reconhece, primordialmente, que o conhecimento científico é patrimônio da humanidade e, que, portanto, devem estar disponíveis livremente para que as pessoas – cientistas ou não - possam usá-lo, reusá-lo e distribuí-lo sem constrangimentos tecnológicos, econômicos, sociais ou legais. Esta também é a ideia que corporifica a proposta do modelo descrito nesta tese. Para a Ciência Aberta, o modelo de publicação ampliada pode ser encarado como uma ferramenta importante para abrir e disseminar os dados de forma consolidada e contextualizada, garantindo a permanência de seus significados e a citação dos mesmos, assegurando ao pesquisador o reconhecimento da autoria e, aos que reusam os dados, a certeza da legitimidade da pesquisa.

No que diz respeito ao movimento do acesso livre, a publicação ampliada vem atender a necessidade de padrões mais ágeis e mais dinâmicos e organicamente mais próximos das comunidades científicas, estendendo também o movimento de livre acesso aos dados científicos. Esses dados são recursos que constituem uma fatia importante do estoque de conhecimento acumulado pelo trabalho de pesquisa, fazendo parte da memória científica e devem ser disponibilizados não apenas à comunidade de pesquisadores, mas também a toda sociedade, como um retorno sobre o que vem sendo feito com os recursos públicos destinados à pesquisa científica.

No caso do modelo de publicação ampliada construída para o contexto do IEN, o modelo também relaciona patentes, material didático, divulgação científica, etc., tornando-se uma evidência do retorno social que se espera das atividades de pesquisa. Assim, cabe ressaltar que o modelo de publicação ampliada proposto intrinsecamente pode gerar o compromisso por parte das instituições e dos pesquisadores em preservar e manter a integridade da memória da pesquisa científica.

É importante registrar que a investigação desta temática ainda é recente e alguns de seus desdobramentos ainda são imprevisíveis, mas o que se espera idealmente é a integração e a interoperabilidade dos sistemas de informação para a pesquisa e concepção de publicações acadêmicas mais próximas da dinâmica da pesquisa atual. Além do modelo que está sendo proposto, um dos objetivos – não revelado formalmente – foi criar um pretexto para iniciar as discussões e pesquisas no país em torno das diversas vertentes abertas por esses fenômenos renovados surgidos no seio da Ciência da Informação, especialmente da Comunicação Científica.

Delinear uma agenda de pesquisa para o futuro que estimule a continuidade dessas investigações é um exercício arriscado, posto que o ambiente é mutável e imprevisível, mas há muitas questões pendentes cuja relevância e o espectro de aplicações exigem soluções rápidas. A começar pelo próprio ambiente de *eScience* e de como se conferir significado, além da capacidade de transportar conhecimento à torrente de dados gerados otimizando o seu potencial cognitivo. Este é um problema que está sendo equacionado pela ideia – tão antiga e tão recente – da curadoria digital de dados de pesquisa, cujas estratégias mais proeminentes passam por esquemas de metadados, agora ampliados para gerenciar e estruturar objetos digitais. A partir daí surgem questões importantes relacionadas à qualidade, seleção, fidedignidade, autenticidade e preservação, além de questões sobre o conflito latente entre acesso e privacidade e às demandas relativas aos

direitos autorais e ao reconhecimento, citação e recompensa para os autores dos dados.

Os processos de curadoria digital se desenrolam em ambientes tecnológicos que precisam ser mais explorados pela Ciência da Informação, são os chamados repositórios digitais confiáveis. São ambientes que estão sujeitos a uma forte padronização, à certificação e a uma gestão rigorosa em que a regra principal é o compromisso de longo prazo. Este tipo de repositório é de suma importância para a preservação de dados de pesquisa, mas também são aplicados aos arquivos digitais, aos depósitos legais das bibliotecas nacionais, aos originais digitais dos acervos museológicos, ou seja, garantem a preservação dos tesouros digitais.

Somado a isso, é importante continuar acompanhando novas formas de troca de informação que se estabelecem a cada dia entre os cientistas, como, por exemplo, as trocas proporcionadas pelos protocolos OAI-PMH, OAI-ORE e pela própria *Web*. Este é um fenômeno importante, porque a inovação que eles portam são desencadeadas pelo usuário/pesquisador, como são os repositórios digitais temáticos e institucionais. Em torno dessas questões há um mundo de novas tecnologias, padrões e protocolos, centrados em esquemas de metadados, que vão permear as redes de repositórios de um futuro próximo.

Nesse ambiente em mutação, surgem novos sistemas de informação e padrões - concebidos especialmente para o mundo da pesquisa - mais abrangentes, mais integrados e mais próximos das articulações do fazer científico, como é o CRIS e o CERIF. Estes sistemas envolvem projetos, financiadores, recursos, dados e outras informações científicas, e criam padrões, ontologias, procedimentos para trocas dessas informações. Esses padrões precisam ser investigados à luz das necessidades do país, mas considerando a sua amplitude internacional.

No que diz respeito à incorporação das publicações ampliadas ao ciclo convencional de comunicação científica, há ainda um longo caminho a percorrer para que seja aceita como uma publicação que possa ser submetida normalmente a qualquer periódico científico. Há disciplinas em que o produto principal de pesquisa é uma base de dados, um conjunto de dados, entretanto, mesmo nessas disciplinas o periódico científico e o artigo permanecem como principal veículo de comunicação científica. Um desafio importante, que vai definir a continuidade e sobrevivência da ideia de publicação ampliada e a sua integração e apropriação pelos sistemas convencionais, como base de dados, repositórios e OPACs. Como isso impacta os padrões e procedimentos estabelecidos por uma área tão solidamente consolidada como é a Recuperação de Informação? As métricas

tão bem equacionadas – como revocação e precisão - continuam válidos num ambiente cujos registros são mutáveis com o tempo e cuja configuração pode ser redefinida pelo usuário? É provável que tenhamos que redesenhar novos parâmetros de avaliação. É importante também que o modelo proposto seja implementado empiricamente e que seja replicado em outros domínios para verificar sua aplicabilidade e consistência.

É fato que num tema novo questionamentos surjam a todo tempo, mas esse fato apenas motiva a continuação dessas investigações em outro patamar. Apenas para registrar alguns desses questionamentos pode-se citar a necessidade de investigação sobre questões referentes aos direitos autorais em uma publicação ampliada, tais como: a quem pertence o direito sobre os dados de pesquisa? (a aqueles que irão compor a publicação ampliada?) Os direitos são do autor? Do grupo de pesquisa? (já que muitas vezes são várias pessoas produzindo o dado) Da instituição? Da financiadora? Das Editoras? Para as publicações tradicionais existem as políticas das editoras, mas e para os dados? Os pesquisadores estão muito receosos de liberarem seus dados, mas se a pesquisa é financiada com recursos públicos não seria um direito das instituições terem acesso aos dados e saber o que está sendo pesquisado? Divulgar os dados também não seria uma forma de dar retorno para a sociedade sobre o investimento do dinheiro público em pesquisa?

Soma-se a isso a possibilidade de ampliação da publicação via relações entre os conteúdos temáticos da pesquisa. Se por um lado, o modelo aqui proposto contemplou a relação entre objetos de pesquisa, por outro lado, as investigações devem continuar buscando uma forma de continuar ampliando as publicações, porém, agora voltadas para os significados de suas temáticas. Isso implica em trabalhar na expansão do modelo para que o autor possa estabelecer relações durante a própria construção do texto científico, ou ainda que, em um sistema, um conjunto de temas relacionados possa auxiliar na construção de uma nova temática de pesquisa.

No que tange ao IEN, enquanto instituição de pesquisa, a presente tese trouxe à tona o problema da necessidade de curadoria dos dados de pesquisa, tendo em vista que, muitos desses dados estavam se perdendo nos computadores dos pesquisadores por estarem armazenados precariamente. Este problema está sendo soliconado em termos práticos com a construção de repositório configurado para receber esses dados e em termos teóricos, com a criação de um grupo de pesquisa em Gestão do Conhecimento Nuclear, com a publicação de diversos artigos sobre o tema, apresentações de trabalhos em congressos e convites para palestras. Trazer à tona um problema até então adormecido abre

espaço para discussão sobre diversas questões que envolvem acesso aberto aos dados, formas de compartilhamento, armazenamento e preservação, garantia de autoria, normas e padrões de tratamento, entre outros. Percebe-se no IEN uma mudança de comportamento por parte dos pesquisadores em relação aos cuidados que os dados devem receber que se revela a partir da melhor compreensão sobre a importância e fragilidade dos dados colocados por essa pesquisa.

As últimas palavras dessa tese são para registrar uma constatação que já estava delineada nas abstrações da autora, mas que foi se consolidando no decorrer da pesquisa. Não obstante a forte presença de outras disciplinas no estudo de novos conceitos de publicações científicas para o ambiente da *eScience* - que tem a tecnologia como um fator determinante, principalmente as provenientes da Ciência da Computação - as teorias, metodologias e práticas postuladas pela Ciência da Informação e pela Biblioteconomia prevalecem como um componente ordenador e estruturante, e estão fortemente subjacentes a todos os estudos e aplicações da área, especialmente quando envolvem questões de representação e significado. Mesmo quando pensamos em disciplinas afins que lidam tradicionalmente com documento e patrimônio, como a Arquivologia e a Museologia, ainda assim muitos conceitos estão presentes dando sentido à tecnologia, como por exemplo, autenticidade, integridade, acervo, curadoria e memória. O que se conclui, finalmente, é que os novos avanços aqui relatados, dependem cada vez mais de uma Ciência da Informação sólida nos seus princípios e teorias, mas que seja capaz de abrir diversos canais de interlocução com a tecnologia que permeia o mundo atual da ciência, contribuindo para o seu progresso.

REFERÊNCIAS

- AALBERSBERG, Ijsbrand Jan; DUNHAN, Judson; KOERS, Hylke. *Connecting scientific articles with research data: new directions in online scholarly publishing*. 2011. Disponível em: <<http://www2.nict.go.jp/isd/ISDS-contents/wds-kyoto-2011.org/pdf/IS704.pdf>>. Acesso em: 19 maio 2013.
- ABBOTT, Daisy. *What is digital curation?* Edinburgh, UK : Digital Curation Centre, 2008. Disponível em: <http://www.era.lib.ed.ac.uk/bitstream/1842/3362/3/Abbott%20What%20is%20digital%20curation_%20_%20Digital%20Curation%20Centre.doc>. Acesso em: 20 dez. 2011.
- AGUIAR, Sueli Marques. *Bases de dados científicos de primatas*. [mensagem pessoal] Mensagem recebida por: <lfs@gmail.com>. Em: 5 dez. 2011.
- ALMEIDA, Elizabeth; TAUHATA, Luiz. *Física nuclear*. Rio de Janeiro: Guanabara Dois, 1981.
- ALTMAN, Micah; KING, Gary. A proposed standard for the scholarly citation of quantitative data. *D-lib Magazine*, v. 13, n. 3/4, 2007.
- AMBINDER, Débora Motta. *Artigos científicos digitais na Web: novas experiências para apresentação, acesso e leitura*. Dissertação (Mestrado em Ciência da Informação) UFF – PPGCI, Niterói, 2012.
- BEAGRIE, Neil. Digital curation for science, digital libraries, and individuals. *International Journal of Digital Curation*, v. 1, n. 1, p. 3-16, 2008.
- BEAGRIE, Neil; POTHEN, P. The digital curation: Digital archives, libraries and eScience seminar. *Ariadne*, v. 30. 2001. Disponível em: <<http://www.ariadne.ac.uk/issue30/digital-curation/>>. Acesso em: 19 maio 2013.
- BELL, Gordon; HEY, Tony; SZALAY, Alex. Beyond the data deluge. *Science*, v. 323, n. 5919, p. 1297-1298, 2009.
- BERLIN. *Declaration on Open Access to Knowledge in the Sciences and Humanities*. Berlin, 2003. Disponível em: <http://www.zim.mpg.de/openaccess-berlin/berlin_declaration.pdf>. Acesso em: 20 dez. 2011.
- BERNERS-LEE, T; HENDLER, J.; LASSILA, O. The semantic. *Web. Scientific American*, v. 284, n. 5, p. 28-37, May 2001. Disponível em: <<http://www.scientificamerican.com/2001/0501issue/0501berners-lee.html>>. Acesso em: 15 jun. 2004.
- BIOLCHINI, Jorge Calmon de Almeida. Semântica e cognição em bases de conhecimento: do vocabulário controlado à ontologia. *Datagramazero: Revista de Ciência da Informação*, v.2, out. 2001.

BORGMAN, Christine L. *Research Data* : who will share what, with whom, when and why. (RatSWD Working Paper n. 161. Oct. 2010). Disponível em: <http://sydney.edu.au/research/data_policy/resources/ANDS_Borgman_2010_research_data.pdf>. Acesso em: 19 maio 2013.

BORKO, Harold. Information science: what is it?. *American Documentation*, v. 19, n. 1, p. 3-5, 1968.

BREITMAN, Karin Koogan. *Web semântica: a internet do futuro*. Rio de Janeiro: LTC, 2005.

BREURE, Leen; VOORBIJ, Hans; HOOGERWERF, Maarten. Rich internet publications: show what you tell. *Journal of Digital Information*, v. 12, n. 1, 2011. Disponível em: <<http://journals.tdl.org/NotTheJodi1/article/viewArticle/1606/1738>>. Acesso em: 19 maio 2013.

BURT, Patricia; KINNUCAN, Mark. Information models and modeling techniques for information systems. *Annual Review of Information Science and Technology*, v. 25, p. 175-208, 1990.

BUSH, Vannevar. As we may think. *Atlantic Monthly*, July 1945. Disponível em: <<http://www.theatlantic.com/doc/194507/bush>>. Acesso em: 30 jun. 2011.

CALLAGHAN, S. A. et al.. Data Publication in the meteorological sciences: the OJIMS project. *Geophysical Research Abstracts*, v. 12, 2010. Disponível em: <<http://meetingorganizer.copernicus.org/EGU2010/EGU2010-8750.pdf>>. Acesso em: 30 jun. 2011.

CAMPOS, Maria Luiza de Almeida. *A organização de unidades do conhecimento em hipertextos: o modelo conceitual como um espaço comunicacional para realização da autoria*. Rio de Janeiro, 2001b. 186 p. Tese (Doutorado em Ciência da Informação)-IBICT/UFRJ/ECO.

_____. O papel das definições na pesquisa em ontologia. *Perspectivas em Ciência da Informação*, v. 15, p. 10-20, 2010.

CAPLAN, Priscilla. *Understanding PREMIS*. Washington D.C.: Library of Congress, 2009. Disponível em: <<http://www.loc.gov/standards/premis/understanding-premis.pdf>>. Acesso em: 30 dez. 2011.

CATARINO, Maria Elisabete; SOUZA, Terezinha Batista de. Descriptive representation in the semantic Web context. *Transinformação*, v. 24, n. 2, p. 77-90, 2012.

CAVALLI, Nicola. *Overlay publications: a functional overview of the concept*. 2009. Disponível em: <http://elpub.scix.net/data/works/att/91_elpub2009.content.pdf>. Acesso em: 19 maio 2013.

CESAR JUNIOR, Roberto Marcondes. Do mundo aos dados e dos dados ao conhecimento. 2011 In: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (orgs.). *O quarto paradigma: descobertas científicas na era da eScience*. São Paulo : Oficina do Texto, 2011, 263 p.

CHEUNG, Kwok et al.. SCOPE: a scientific compound object publishing and editing system. *International Journal of Digital Curation*, v. 3, n. 2, p. 4-18, 2008.

CLARKE, Stella G. Dextre. Thesaural relationships. In: BEAN, A.; GREEN, Rebecca (Ed.) *Relationships in the organization of knowledge*. Dordrecht: Kluwer, 2001, p. 37-51.

CONSULTATIVE COMMITTEE FOR SPACE DATA SYSTEM - CCSDS. *Reference Model for an Open Archival Information System (OAIS)*. Blue book (CCSDS 650.0-B-1). Washington, DC: CSDS, 2002. Disponível em: <<http://public.ccsds.org/publications/archive/650x0b1.pdf>>. Acesso em: 30 dez. 2011.

COUGO, Paulo. *Modelagem conceitual e projeto de banco de dados*. Rio de Janeiro: Campus, 1997.

DAHLBERG, I. A referent-oriented analytical concept theory of interconcept. *International Classification*, Frankfurt, v. 5, n. 3, p. 142-150, 1978a.

_____. *Ontical structures and universal classification*. Bangalore: Sarada Ranganathan Endowment, 1978b.

DE ROURE, David; HENDLER, James A. E-Science: the grid and the semantic Web. *IEEE Intelligent Systems*, v. 19, n. 1, p. 65-71, 2004. Disponível em: <http://www.computing.surrey.ac.uk/courses/csm23/Papers/IEEE_eScience_the_grid_and_the_semantic_Web.pdf>. Acesso em: 19 maio 2013.

DE ROURE, David; JENNINGS, Nicholas R.; SHADBOLT, Nigel R. *Research agenda for the semantic grid: a future eScience infrastructure*. 2001. Disponível em: <<http://users.ecs.soton.ac.uk/dder/semgrid.pdf>>. Acesso em: 19 maio 2013.

_____. The semantic grid: a future eScience infrastructure. In: BERMAN, Fran; FOX, Geoffrey; HEY, Anthony J. G. (Ed.). *Grid computing: making the global infrastructure a reality*. Chichester: Wiley, 2003. p. 437-470.

DEMO, Pedro. *Introdução à metodologia da ciência*. São Paulo: Atlas, 1987.

DIGITAL CURATION CENTER – DCC. *Curation Lifecycle Model*. 2008. Disponível em: <<http://www.dcc.ac.uk/resources/curation-lifecycle-model>>. Acesso em 20 maio 2013.

DODEBEI, Vera L. D. *Tesouro: linguagem de representação da memória documentária*. Niterói: Intertexto, 2002. 120p.

ECO, H. *Como se faz uma tese*. São Paulo: Editora Perspectiva. 1977.

ERIKSSON, Henrik. An annotation tool for semantic documents. *Lecture Notes in Computer Science*, v. 4519, p. 759-768, 2007b.

_____. The semantic-document approach to combining documents and ontologies. *International Journal of Human-Computer Studies*, v. 65, n. 7, p. 624-639, 2007a.

ERWAY, Ricky; LAVOIE, Brian. *The economics of data integrity*. Ohio: OCLC, 2012. Disponível em:

<<http://www.Webjunction.org/content/dam/research/publications/library/2012/erway-dataintegrity.pdf>>. Acesso em: 21 set. 2013

ESCHENBACH, Carola; HEYDRICH, Wolfgang. Classical mereology and restricted domains. *International Journal of Human Computer Studies*, v. 43, p. 723-740, 1995. Disponível em: <<http://www.nsf.gov/pubs/2005/nsb0540/>>. Acesso em: 18 abr. 2009.

EUROCRIS. *CERIF – 1.3 Full Data Model*: Model Introduction and Specification. 2012a. Disponível em: <http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.3/Specifications/CERIF1.3_FDM.pdf>. Acesso em: 2 abr. 2014.

EUROCRIS. CERIF 2008 – 1.3 Semantics: Research Vocabulary. 2010. Disponível em: <http://www.eurocris.org/Uploads/Web%20pages/CERIF2008/Release_1.2/CERIF2008_1.2_Semantics.pdf> Acesso em: 2 abr. 2014

EUROCRIS. CERIF – 1.3 XML Schema Files. 2012b. Disponível em: <<http://www.eurocris.org/Uploads/Web%20pages/CERIF-1.3/XML-EXAMPLES/>> Acesso em: 2 abr. 2014

FEDORA COMMONS. Fedora Digital Object Relationships. In: **Fedora Project**. 2008. Disponível em: <<<http://www.fedora-commons.org/documentation/3.0/userdocs/digitalobjects/introRelsExt.html>> Acesso em: 17 jun. 2014.

FELBER, H. *Terminology manual*. Paris: UNESCO, 1984.

FERRAZ, Terezine Arantes. *A informação na área nuclear e a estrutura de trabalhos científicos*. Rio de Janeiro: CNPq/IBBD, 1975.

GIL, Antônio Carlos. *Métodos e técnicas de pesquisa social*. 6. ed. São Paulo: Atlas, 2008.

GINSPARG, Paul. Winners and losers in the global research village. *The Serials Librarian*, v. 30, n. 3-4, p. 83-95, 1997.

GRAY, Jim et al.. Scientific data management in the coming decade. *ACM SIGMOD Record*, v. 34, n. 4, p. 34-41, 2005.

GREEN, Rebecca. Relationships in the organization of knowledge. An overview. In: BEAN, A.; GREEN, Rebecca (Ed.). *Relationships in the organization of knowledge*. Netherlands: Springer, 2001. p. 3-18. (Information Science and Knowledge Management, v. 2)

GROSS, Bernhard. *Sistema Internacional de Informações Nucleares*: tendências atuais e futuras da informação científica. Rio de Janeiro: CNEN, 1967. 32p.

GUIZZARDI, Giancarlo. The role of foundational ontologies for conceptual modeling and domain ontology representation. In: INTERNATIONAL BALTIC CONFERENCE ON DATABASES AND INFORMATION SYSTEMS, 7. Vilnius, Lithuania: IEEE, 2006. p. 17-25.

HACHEM, Nabil I. et al.. *GaeaPN: A Petri Net Model for the Management of Data and Metadata Derivations in Scientific Experiments*. 1994. (Worcester Polytechnic Institute Computer Science Department Technical Report WPI-CS-TR-94, v. 1). Disponível em: <<ftp://ftp.cs.wpi.edu/pub/techreports/pdf/94-1.pdf>>. Acesso em: 19 maio 2013

HARMSZE, Frédérique-Anne Pacifique. *A modular structure for scientific articles in an electronic environment*. 2000. Disponível em: <<http://dare.uva.nl/document/2621>>. Acesso em: 2 abr. 2014

HARVEY, Douglas Ross. *Digital curation: a how-to-do-it manual*. London: Facet, 2010.

HEDRICK, Terry E. Justifications for the sharing of social science data. *Law and Human Behavior*, v. 12, n. 2, p. 163-171, 1988.

HEY, Tony; TREFETHEN, Anne E. Cyberinfrastructure for e-Science. *Science*, v. 308, n. 5723, p. 817-821, 2005.

_____. e-Science and its implications. *Philosophical Transactions of the Royal Society of London A: Mathematical, physical and engineering sciences*, v. 361, n.1809, p.1809-1825, 2003.

HEY, Tony; HEY, Jessie. e-Science and its implications for the library community. *Library Hi Tech*, v.24, n.4, p.515-528, 2006.

HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin. Jim Gray on e-Science: a transformed scientific method. In: HEY, Tony; TANSLEY, Stewart; TOLLE, Kristin (Org.). *The Fourth Paradigm: Data-Intensive Scientific Discovery*, 2009. Disponível em: <http://research.microsoft.com/en-us/collaboration/fourthparadigm/4th_paradigm_book_jim_gray_transcript.pdf>. Acesso em: 20 dez. 2011.

HJØRLAND, B. Domain analysis in information science: eleven approaches – traditional as well as innovative. *Journal of Documentation*, v.58, n.4, p. 422 – 462, 2002.

HUMPHREY, C. *e-Science and the Life Cycle of Research*. 2006. Disponível em: <datalib.library.ualberta.ca/~humphrey/lifecycle-science060308.doc>. Acesso em: 13 jan. 2012.

HUNTER, Jane. Scientific publication packages: a selective approach to the communication and archival of scientific output. *The International Journal of Digital Curation*, v. 1, n. 1, 2006. Disponível em: <<http://www.ijdc.net/index.php/ijdc/article/view/8/4>>. Acesso em: 13 jan. 2012.

INSTITUTO BRASILEIRO DE INFORMAÇÃO EM CIÊNCIA E TECNOLOGIA - IBICT. *Repositórios Digitais*. 2012. Disponível em: <<http://www.ibict.br/informacao-para-ciencia-tecnologia-e-inovacao%20/repositorios-digitais>>. Acesso em 2 abr. 2014.

INSTITUTO DE ENGENHARIA NUCLEAR - IEN. *Comunicado DIEN: cadernos de áreas temáticas de P & D do IEN/CNEN*. [chamada Institucional]. Recebida por lsales@ien.gov.br. Em 19 fev. 2013.

INTERNATIONAL FEDERATION OF LIBRARY ASSOCIATIONS AND INSTITUTIONS - IFLA. *Functional Requirements for Bibliographic Records: final report*. 2009. Disponível em: <http://www.ifla.org/files/assets/cataloguing/frbr/frbr_2008.pdf>. Acesso em: 19 maio 2013

IVANOVIĆ, Dragan; SURLA, Dušan; RACKOVIĆ, Miloš. A CERIF data model extension for evaluation and quantitative expression of scientific research results. *Scientometrics*, v. 86, n. 1, p. 155-172, 2011.

JORG, B. et al.. *CERIF 2008—1.0 Full Data Model (FDM): Introduction and specification*. 2009a. 43p. Disponível em: <http://www.eurocris.org/Uploads/Web%20pages/CERIF2008/CERIF2008_1.0_FDM.pdf>. Acesso em: 04 abr. 2014.

_____. *CERIF 2008—1.0 XML Data Exchange Format Specification*. 33 p. 2009b. Disponível em: <http://www.eurocris.org/Uploads/Web%20pages/CERIF2008/CERIF2008_1.0_XML.pdf>. Acesso em: 16 fev. 2010.

KALLINIKOS, Jannis; AALTONEN, Aleks; MARTON, Attila. A theory of digital objects. *First Monday*, v.15, n.6, p.1-17, 2010.

KENNEY, Anne R.; RIEGER, Oya Y.; ENTLICH, Richard. Levando la teoria a La práctica: tutorial de digitalización de imágenes. 2001. Disponível em: <<http://www.library.cornell.edu/preservation/tutorial-spanish/contents.html>> Acesso em 5 jul. 2011.

KERR, Peter; REDDINGTON, Fiona; WILKINSON, Max. Digital curation: where do we go from here. *Ariadne*, v. 45, 2005.

KIETZ, Jörg-Uwe et al.. A method for semi-automatic ontology acquisition from a corporate intranet. In: EKAW'00: WORKSHOP ON ONTOLOGIES AND TEXT, 2000. *Proceedings ...* Juan-Les-Pins: 2000.

KIRCZ, Joost G. New practices for electronic publishing 1: Will the scientific paper keep its form? *Learned Publishing*, v. 14, n. 4, p. 265-272, 2001. Disponível em: <www.kra.nl/Website/Artikelen/learned-publ1.pdf>. Acesso em: 11 ago. 2013

_____. New practices for electronic publishing 2: New forms of the scientific paper. *Learned Publishing*, v. 15, n. 1, p. 27-32, 2002. Disponível em: <www.kra.nl/Website/Artikelen/learnedpubl2.pdf>. Acesso em: 11 ago. 2013

LAGOZE, Carl et al.. *Object re-use & exchange: A resource-centric approach*. 2008. (arXiv preprint arXiv:0804.2273). Disponível em: <<http://arxiv.org/ftp/arxiv/papers/0804/0804.2273.pdf>>. Acesso em: 5 set. 2013.

LAGOZE, Carl; VAN DE SOMPEL, Herbert. Compound information objects: The OAI-ORE perspective. Open Archives Initiative. 2007. Disponível em: <<http://www.openarchives.org/ore/documents/CompoundObjects-200705.html>>. Acesso em 4 abr. 2014.

LICKLIDER, J.C.R. *Libraries of the future*. Cambridge, Mass.: MIT Press, 1965.

LOPATENKO, Andrei. *Information retrieval in current research information systems*. (arXiv preprint cs/0110026). 2001. Disponível em: <<http://arxiv.org/ftp/cs/papers/0110/0110026.pdf>>. Acesso em: 11 ago. 2013.

LYNCH, Clifford. The shape of the scientific article in the developing cyberinfrastructure. *CTWatch Quarterly*, v. 3, n. 3, 2007.

LYON, Liz. *Dealing with data; role, rigths, responsibilities and relationships consultancy report*. p. 1-65, jun. 2007. Disponível em: <http://opus.bath.ac.uk/412/1/dealing_with_data_report-final.pdf>. Acesso em: 19 maio 2013.

MARCONDES, Carlos Henrique et al.. Bases ontológicas e conceituais para um modelo do conhecimento científico em artigos biomédicos. *RECIIS*, v. 3, n. 1, 2009.

_____. Um modelo semântico de publicações eletrônicas. A semantic model for electronic publishing. *Liinc em Revista*, v. 7, n. 1, 2011.

MAYERNIK, Matthew et al.. The data conservancy instance infrastructure and organization service for research data curation. *D-Lib Magazine*, v. 18, n. 9/10, Sep./Oct. 2012.

MEY, Eliane S. A. *Introdução à catalogação*. Brasília: Briquet de Lemos, 1995.

MOTTA, D. F. *Método relacional como nova abordagem para a construção de tesauros*. Rio de Janeiro: SENAI/ DN, 1987.

MYLOPOULOS, John et al.. Telos: representing knowledge about information systems. *ACM Transactions on Information Systems*, v. 8, n. 4, p. 325-362, 1990.

NATIONAL INFORMATION STANDARD ORGANIZATION - NISO. *Understanding Metadata*. Bethesda, MD: NISO Press, 2004. Disponível em: <<http://www.niso.org/publications/press/UnderstandingMetadata.pdf>> Acesso em: 30 set. 2009.

NATIONAL RESEARCH COUNCIL - NRC. *A Question of Balance: Private Rights and the Public Interest in Scientific and Technical Databases*. Washington, DC: National Academy Press. 1999. Disponível em: <<http://www.nap.edu>>. Acesso em: 19 maio 2013.

NATIONAL SCIENCE FOUNDATION - NSF. *Cyberinfrastructure vision for 21st century Discovery*. March 2007. Disponível em: <http://escience.caltech.edu/workshop/CI_Vision_March07.pdf>. Acesso em: 19 maio 2013.

NEELAMEGHAN, A; MAITRA, R. *Non-hierarchical associative relationships among concepts: Identification and typology*. Bangalore: FID/CR Secretariat, 1978. (Part A of FID/CR report no. 18)

NELSON, Ted. *Dream machines*. South Bend, IN: The distributors, 1974.

NEŠIĆ, Saša. Semantic document model to enhance data and knowledge interoperability. In: DEVEDŽIC, Vladan; GAŠEVIC, Dragan (Eds.). *Web 2.0 e Semantic Web*. Netherlands: Springer, 2009, p. 135-160. (Annals of Information Systems, v. 6).

NEŠIĆ, Saša et al.. Search and Navigation in Semantically Integrated Document Collections. In: *SEMAPRO 2010: THE FOURTH INTERNATIONAL CONFERENCE ON ADVANCES IN SEMANTIC PROCESSING*, 2010. p. 55-60.

NOVO, Hildenise Ferreira. *A elaboração de taxonomia: princípios classificatórios para domínios interdisciplinares*. Orientador: Maria Luiza de Almeida Campos. Niterói, 19 mar 2007. 172f. Dissertação (Mestrado em Ciência da Informação) - IBICT/UFF.

ONLINE COMPUTER LIBRARY CENTER/ RESEARCH LIBRARY GROUP OCLC/RLG. *Implementing preservation repositories for digital materials: current practice and emerging trends in the cultural heritage - A Report by the PREMIS Working Group*. September 2004, 66 p. Disponível em: <<http://www.oclc.org/research/activities/past/orprojects/pmwg/surveyreport.pdf>>. Acesso em: 12 dez. 2011.

ONLINE COMPUTER LIBRARY CENTER/ RESEARCH LIBRARY GROUP - OCLC/RLG. *Data dictionary for preservation metadata: final report of the PREMIS Working Group*. 2005. Disponível em: <<http://www.oclc.org/research/activities/past/orprojects/pmwg/premis-final.pdf>>. Acesso em: 30 dez. 2011.

ORGANIZAÇÃO PARA A COOPERAÇÃO E DESENVOLVIMENTO ECONÔMICO - OCDE. *Principles and guidelines for access to research data from public data*. 2007. Disponível em: <<http://www.oecd.org/dataoecd/9/61/38500813.pdf>> Acesso em: 17 fev. 2012.

OTLET, Paul. *Traite de documentation: le livre sur le livre, theorie et pratique*. Bruxelles: Editiones Mundaneum, 1934.

PALMER, Carole L.; WEBER, Nicholas M.; CRAGIN, Melissa M. The analytic potential of scientific data: understanding re-use value. *ASIST 2011*, v. 9, n. 13, Oct. 2011.

PARINOV, Sergey. *Open repository of semantic linkages*. Moscow: 2012.

_____. Semantic fragment of a research e-infrastructure: necessary information objects, tools and services. *Russian Digital Libraries Journal*, v. 16, n. 1, 2013.

PARINOV, Sergey; KOGALOVISKY, Mikhail. Semantic linkage in research information systems as new data source for scientometric studies. *Scientometrics*, v. 98, n. 2, p. 927-943, 2014.

PEREIRA, Maria de Nazaré. *O Laboratório de P&D no terceiro milênio*. Pré-print de palestra proferida no encontro da ABEC. Petropolis, RJ: ABEC, 19 dez. 1999. 12p.

PÉREZ-GONZÁLEZ, Lourdes. *Modelo/s de coste para la preservación de los datos científicos en la e-ciencia*. 2010. Disponível em: <<http://eprints.rclis.org/8555/1/Perez.pdf>>. Acesso em: 20 set. 2013.

PINFIELD, S.; JAMES, H. The digital preservation of e-Prints. *D-Lib Magazine*, v. 9, n. 9, 2003.

PINHEIRO, Lena Vania R. Comunidades científicas e infraestrutura tecnológica no Brasil para uso de recursos eletrônicos de comunicação e informação na pesquisa. *Ciência da Informação*, v. 32, n. 3, p. 62-73, 2003.

RAMALHO, R. *Web semântica: aspectos interdisciplinares da gestão de recursos informacionais no âmbito da Ciência da Informação*. Marília: UNESP, 2006.

RANGANATHAN, S. R. *Prolegomena to library classification*. Bombay: Asia Publishing House, 1967. 640 p.

RESEARCH LIBRARY GROUP / ONLINE COMPUTER LIBRARY CENTER - RLG/OCLC. *Trusted digital repositories: attributes and responsibilities*. Mountain View, CA. May 2002. Disponível em: <<http://www.oclc.org/content/dam/research/activities/trustedrep/repositories.pdf?urlm=161690>>. Acesso em: 2 abr. 2014.

RODRIGUES, Eloi et al. *Os repositórios de dados científicos: estado da arte*. 2010. (Relatório D-24: RCAAP). Disponível em: <http://projeto.rcaap.pt/index.php?option=com_remository&Itemid=2&func=startdown&id=271&lang=pt>. Acesso em: 2 abr. 2014.

RUDIO, F. V. *Introdução ao projeto de pesquisa científica*. 32. ed. Petrópolis: Vozes, 2004.

RUUSALEPP, Raivo. *Infrastructure planning and data curation: a comparative study of international approaches to enabling the sharing of research data*. DCC Report commissioned by JISC, 2008. Disponível em: <<http://www.dcc.ac.uk/sites/default/files/documents/publications/reports/Data-SharingReport.pdf>>. Acesso em: 21 set. 2013.

SALES, Luana Farias. Modelo triádico de relações para aplicação em ontologias. In: SEMINÁRIO BRASILEIRO DE ONTOLOGIAS, 1. *Anais...* Niterói: UFF, 2008. Disponível em: <<http://www.uff.br/ontologia/artigos/13.pdf>>. Acesso em: 26 jun. 2012.

_____. *Ontologias de domínio: estudo das relações conceituais e sua aplicação*. Dissertação (Mestrado em Ciência da Informação) - IBICT/UFF. Niterói, 14 set. 2006. 139 f.

_____. *Plataforma Carpe diem: uma ferramenta para gestão do conhecimento nuclear* [apresentação]. Rio de Janeiro: 18 jul. 2013. Disponível em: <<http://carpedien.ien.gov.br/handle/ien/593>>. Acesso em: 3 abr. 2014.

SALES, Luana Farias; SAYÃO, Luís Fernando. Inovações tecnológicas: grandes pensadores e seu reflexo nas bibliotecas. In: FERREIRA, Sueli Mara Soares Pinto; TARGINO, Maria das Graças (Org.). *Conhecimento: custódia e acesso*. São Paulo: SIBiUSP, 2012. p. 46-71.

SALES, Luana Farias; SAYÃO, Luís Fernando; MOTTA, Dilza Fonseca da. *Modelagem de relações conceituais para a área nuclear*. Recife, 2012. p. 182-187. Disponível em: <http://ceur-ws.org/Vol-938/ontobras-most2012_paper16.pdf>. Acesso em 19 maio 2013.

SAMPAIO, Paulo Augusto Berquó de. *Gestão de P & D no IEN: 2013-2017* [apresentação]. Rio de Janeiro: 16 set. 2013. Disponível em: <<http://prezi.com/noq7uesdaxgm/gestao-de-pd-no-ien/>>. Acesso em 3 abr. 2014.

SAYÃO, Luís Fernando. Modelos teóricos em Ciência da Informação: abstração e método científico. *Ciência da Informação*, v. 30, n. 1, p. 82-91, 2001.

_____. Uma outra face dos metadados: informações para gestão da preservação digital. *Encontros Bibli: Revista Eletrônica de Biblioteconomia e Ciência da Informação*, v. 15, n. 30, p. 1-31, 2010. Disponível em: <<http://periodicos.ufsc.br/index.php/eb/issue/view/1412>>. Acesso em: 30 dez. 2011.

SAYÃO, Luís Fernando; SALES, Luana Farias. Dados de pesquisa: contribuição para o estabelecimento de um modelo de curadoria digital para o país. *Tendências da Pesquisa Brasileira em Ciência da Informação*, v. 6, n. 1, 2013.

SERINGHAUS, Michael R.; GERSTEIN, Mark B. Publishing perishing? Towards tomorrow's information architecture. *BMC Bioinformatics*, v. 8, n. 1, p. 17, 2007.

SHETH, A.; ARPINAR, I. B.; KASHYAP, V. Relationships at the heart of semantic Web: modeling, discovering, and exploiting complex semantic relationships. In: NIKRAVESH, M. et al.. *Enhanceing the power of the Internet*. Berlin: Springer, 2003. (Studies in Fuzziness and Soft Computing, v. 139).

SHOTTON, David. Semantic publishing: the concept the coming revolution in scientific journal publishing. *Learned publishing*, v.22, n.2, p. 85-94, 2009. Disponível em: <<http://www.ploscompbiol.org/article/info%3Adoi%2F10.1371%2Fjournal.pcbi.1000361>>. Acesso em: 19 maio 2013.

SIEBER, Joan E. Data sharing. *Law and Human Behavior*, v. 12, n. 2, p. 199-206, 1988.

SOEHNER, Catherine; STEEVES, Catherine; WARD, Jennifer. *E-Science and data support services: A study of ARL member institutions*. Washington, DC: Association of Research Libraries, 2010. Disponível em: <<http://arl.nonprofitsoapbox.com/storage/documents/publications/escience-report-2010.pdf>>. Acesso em: 19 maio 2013.

SOUZA, Rosali Fernandez de. Universo de Ciência e Tecnologia: organização e representação em classificações do conhecimento. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO - ENANCIB, 13. Rio de Janeiro, 2012. Disponível em: <<http://www.eventosecongressos.com.br/metodo/enancib2012/arearestrita/pdfs/19371.pdf>>. Acesso em: 20 maio 2013.

STANTON, Joffrey M. Education for eScience professionals: job analysis, curriculum guidance, and program consideration. *Journal of Education for Library and Information Science*, v.52, n.2, Apr. 2011.

TAYLOR, J. M. The UK eScience programme [Powerpoint presentation]. In: ESCIENCE LONDON MEETING. Sept. 2001.

TELLO, A. L. Ontologías em la Web Semántica. In: JORNADAS DE INGENIERIA WEB 01. 2002. Disponível em: <http://www.anobium.es/docs/gc_fichas/doc/68ERfhjkmv.pdf>. Acesso em: 30 out. 2004.

UHLIR, Paul F. Information Gulags, Intellectual Straightjackets, and Memory Holes: Three Principles to Guide the Preservation of Scientific Data. *Data Science Journal*, v. 9, p. ES1-ES5, 2010. Disponível em: <https://www.jstage.jst.go.jp/article/dsj/9/0/9_Essay-001-Uhlir/_pdf>. Acesso em: 5 set. 2013

VAN DE SOMPEL, Herbert et al.. Rethinking scholarly communication. *D-Lib Magazine*, v. 10, n. 9, 2004.

VAN DEN HEUVEL, Charles; RAYWARD, W. Boyd. Facing interfaces: Paul Otlet's visualizations of data integration. *Journal of the American Society for Information Science and Technology*, v. 62, n. 12, p. 2313-2326, 2011.

VERHAAR, Peter. Report on object models and functionalities. In: PLACE, Thomas et al. (Eds.). *DRIVER II*. [S.l: s.n], 2008.

VERNOOY-GERRITSEN, Marjan. *Emerging standards for enhanced publications and repository technology: survey on technology*. Amsterdam: University Press, 2009a. (Driver and Surf studies).

_____. *Enhanced Publications: linking publications and research data in digital repositories*. 2009b

W3C Brasil. *Semantic Web*. 2001. Disponível em: <<http://w3c.org/2001/ws>>. Acesso em: 10 jul. 2006.

WEGNER, Peter; GOLDIN, Dina. *Mathematical models of interactive computing*. [S.l: s.n], 1999.

WOUTERSEN-WINDHOUWER, Saskia; BRANDSMA, Renze. *Enhanced publications: state of the art*. v. 83, part 1, p.19-91, 2009. Disponível em: <http://www.driverrepository.eu/component/option,com_jdownloads/Itemid>. Acesso em: 4 abr. 2014

WULF, William A. *The national collaboratory—a white paper*. Appendix A: In Toward a national collaboratory. Unpublished report of a national science foundation. Invitational workshop held at Rockefeller University. 1989. p.1.

WÜSTER, E. L'étude scientifique générale de la terminologie, zone frontalière entre la linguistique, la logique, l'ontologie, L'informatique et les sciences des chose. In: RONDEAU, G.; FELBER, E. (Org.). *Textes choisis de terminologie*. Québec: GIRSERM, 1981, p. 57-114. (Fondements Théoriques de la Terminologie, v. I).

HUGE, Hai; SUN, Yunchuan. The schema theory for semantic link network. *Future Generation Computer Systems*, v. 26, n. 3, p. 408-420, 2010.

APENDICE A - PROPOSTA DE CLASSIFICAÇÃO DE RELAÇÕES PARA PUBLICAÇÃO AMPLIADA

Relação Lógica – Relação que revela hierarquias de gênero-espécie. Ex: É um; É subclasse de; É subgrupo de
 Relação Ôntica - revelam relação do objeto no mundo e caracterizam-se pela contiguidade no tempo e no espaço ou pela conexão de causa efeito. (Dahlberg, 1978b)

Relação Meronímica – Relação entre as partes de uma publicação. Ex: É formado de (Data cite), É parte de (Data cite), Adjacente, Envolve (VOC_IEN) *isConstituentOf* <hasConstituent> , <hasMember> <isMemberOf> , <isSubsetOf> <hasSubset> , <hasCollectionMember> (OAI-ORE)

Relação Funcional – Relações que revelam a função de uma entidade sobre a outra.

Relação de citação – revela o tipo de citação entre os documentos.

Relação de derivação – revela fase e estágios de um documento ou processo de pesquisa no tempo.

Relação de propriedade – revela atributos, características, papel ou propriedade de um documento.

Relação de influência – revela o impacto, efeito ou ação de um objeto ou processo sobre o outro.

Relação de Citação	Relação de Derivação	Relação de Propriedade	Relação de Influência
Cita (CITO)	Atualiza (CITO)	É anotação de (OAI-ORE)	Administra (CERIF)
Cita como autoridade (CITO)	Compila (CITO)	É autor de (VOC_IEN)	Afeta ((VOC_IEN)
Cita como evidencia (CITO)	Continua (VOC-IEN)	É autor institucional de (CERIF)	Analisa (VOC_IEN)
Cita como Fonte (CITO)	Deriva (CITO)	É certificadora de (VOC_IEN)	Apóia (CITO)
Cita como Fonte de dados (CITO)	É atualizado por (CITO)	É curado por (VOC_IEN)	Avalia (VOC_IEN)
Cita como Fonte documental (CITO)	é base de (CITO, data cite)	É descrição de (OAI-ORE)	Avaliado por (VOC_IEN)
Cita como leitura recomendada (CITO)	É baseado em (Data cite, CITO)	É editor de (CERIF)	Coloca em (VOC_IEN)
Cita como metadado (CITO)	É compilado por (CITO, Data cite)	É financiado por (VOC_IEN)	Corrige (CITO)
Cita como relacionado (CITO)	É continuação de (Data cite)	É financiadora de (VOC_IEN)	Credita (CITO)
Cita como solução potencial (CITO)	É corrigido por (CITO)	É medida de (VOC_IEN)	Cria (VOC_IEN)
Cita para informar (CITO)	É derivação de (OAI-ORE)	É metadado de (CITO)	Da suporte para (CITO)
Concorda com (CITO)	É estágio de (VOC_IEN)	É metadado de (OAI-ORE)	Descreve (CITO)
Confirma (CITO)	É estendido por (CITO)	É orientado por (VOC_IEN)	Destrói (VOC_IEN)
Contem afirmações de (CITO)	É fase de (VOC_IEN)	É par de (VOC_IEN)	Diagnostica (VOC_IEN)
Discorda com (CITO)	É nova versão de (Data cite)	É parceiro de (VOC_IEN)	Divulga (VOC_IEN)
Discute (CITO)	é revisado (CITO)	É proprietário de direito (CERIF)	Documenta (Data cite)
É citado como autoridade por (CITO)	É Revisão de (VOC_IEN)	É publicado por (VOC_IEN)	É requerido por (Data cite)
É citado como evidencia por (CITO)	É substituído por (Data cite)	É publicadora de (VOC_IEN)	É agente (VOC_IEN)
É citado como Fonte de dado para (CITO)	É uma forma original de (Data cite)	É revisor de (CERIF)	É Causado por (VOC_IEN)
É citado como Fonte documental por (CITO)	É uma forma variante de (Data cite)	É tradutor de (CERIF)	É dependente de (OAI-ORE)
É citado como informação por (CITO)	É versão de (Data cite)	Ocorre em (VOC_IEN)	É descrito por (CITO)
É citado como leitura recomendada (CITO)	Gera (VOC_IEN)	Propriedade de (VOC_IEN)	É diretor de (VOC_IEN)
E citado por (CITO, Data cite)	Revisa (VOC_IEN)	Publica (CERIF)	É divulgado por (VOC_IEN)
É confirmado por (CITO)	Revisado por (VOC_IEN)	qualifica (CITO)	É documentado por (CITO)
É creditado por (CITO)	Substitui (Data cite)	Reivindica direito (CERIF)	É gerado por (VOC_IEN)
É criticado por (CITO)	Tem derivação (OAI-ORE)	Tem anotação para (OAI-ORE)	É levantado por (VOC_IEN)
	Tem versão (Data cite)	Tem autoria de (VOC_IEN)	É planejado por (VOC_IEN)
	Transforma em (VOC_IEN)		É produto de (VOC_IEN)
			É produzido por (VOC_IEN)
			É qualificado por (CITO)
			É suplementado por (Data cite)
			É usado por ((VOC_IEN)
			Embasa (CITO)

É referenciado por (Data cite) é refutado por (CITO) Está de acordo com (CITO) Inclui citação de (CITO) Inclui trecho de (CITO) Referencia (Data cite) Responde a (CITO) Retrata (CITO) Tem resposta (CITO)		Tem descrição (OAI-ORE) Tem equivalente (OAI-ORE) Tem metadado para(OAI-ORE)	Escreve (VOC_IEN) Estende / continua (CITO) Estuda em (VOC_IEN) Fornece informação para (CITO) Fornece método para (CITO) Fornece trecho para (CITO) Indica (CITO) Interage com (VOC_IEN) Levanta (VOC_IEN) Mede (VOC_IEN) Método de (VOC_IEN) Normaliza (VOC_IEN) Obtém apoio de (CITO) Orienta (VOC_IEN) Pesquisa (VOC_IEN) Planeja (VOC_IEN) Produz (VOC_IEN) Prover afirmação para (CITO) Prover conclusão para (CITO) Prover dado para (CITO) Realiza (VOC_IEN) Recebe informação de (VOC_IEN) Recomenda (VOC_IEN) Requer (Data cite) Resulta em (VOC_IEN) Resultado da (VOC_IEN) Reusa (VOC_IEN) Serve de suporte (CITO) Supervisiona (CERIF) Tem dependente (OAI-ORE) Trabalha para (VOC_IEN) Traz (VOC_IEN) Usa (VOC_IEN) Usa conclusão de (CITO) Usa dado de (CITO) Usa método de (CITO) Usado em (VOC_IEN) Valida (VOC_IEN) Verifica (VOC_IEN) Verificados por (VOC_IEN)
--	--	---	--

APÊNDICE B – FORMULÁRIO PARA LEVANTAMENTO DE DADOS E DOCUMENTOS PRODUZIDOS NO IEN

Nome da Área: _____

1) Indique os dados utilizados ou gerados em sua área de acordo com a classificação abaixo:

Quanto à origem – os dados podem ser gerados para diferentes propósitos e por diferentes processos	SIM	NÃO	Digitais
Observacionais			
Computacionais ou de Simulação			
Experimentais			
Registros governamentais e de negócios			
Derivados ou Compilados			

TIPOS DE DADOS (quanto à natureza)		SIM	NÃO	FORMATO
Número	medidas			
	resultados de levantamentos			
	resultados de experimentos			
	fórmulas			
	equações			
	algoritmos			
Multimídia	Imagens			
	vídeo			
	áudio			
	animações			
	filme			
	fotografia			
Software	bases de dados			
	simulações			
	Códigos nucleares			
Visualização	tabelas			
	gráficos			
	diagramas			
	modelos em 3D			
	Modelos reduzidos			
	Desenhos			
Textuais	metadados			
	questionários			
	entrevistas			
	anotações			
	normas			
	padrões			
	certificados			
	caderno de laboratório			

	transcrição			
	correspondências			
	diário			
	caderno de campo			
Artefatos	espécimes			
	amostras			
	maquete			
Processos	procedimentos operacionais padronizados			
	workflows			
	protocolos			
	teste			
Outros				

2) Indique no formulário abaixo outros dados e informações relacionados com a pesquisa que também registram a memória das atividades realizadas por sua área temática

	TIPOS DE DADOS	SIM	NÃO	FORMATO
Documentos	relatórios técnicos			
	relatórios de pesquisa			
	pedido de financiamento			
	formulários			
	Pedidos de consentimento			
	Blogs			
	Redes Sociais			
	Relatório de pesquisa em andamento			
	comissão de ética			
	Projeto de pesquisa			
	Pedido de financiamento			
	Avaliação dos financiadores			
	Avaliação dos pares			
	Outros			

3) Esses dados e informações são preservados de alguma maneira? Como?

4) Utilizam algum sistema ou banco de dados para levantarem seus dados e informações? Qual?

APENDICE C – RELAÇÕES CONCEITUAIS APLICADAS AO IEN

TIPO DE RELAÇÃO	RELAÇÃO VOC_IEN	DEFINIÇÃO
Relação de Citação	Divulga	Relação entre o mecanismo que divulga e o que é divulgado
	É usado por	Relação entre um objeto e um processo ou pessoa que o utiliza
	É divulgado por	Relação entre uma divulgação e a pessoa ou o meio de divulgação utilizado.
	Resenha	Relação entre um documento original e sua resenha
	Resume	Relação entre um documento original e seu resumo
Relação de Derivação	É estágio anterior de	Relação entre as etapas de uma pesquisa.
	É estágio posterior de	Relação entre as etapas de uma pesquisa.
	É revisão de	Relação entre as versões de um documento.
	Gera	Relação que revela o resultado de uma pesquisa.
	Revisa	Relação que se dá entre um revisor e algo revisto.
	Transforma em	Relação que revela a transformação de algo por algum processo ou alguém
	É técnica de	Relação entre uma técnica e objetivo para o qual ela serve.
	É reusado por	Relação entre um dado reusado e quem o reutilizou
	É transformado por	Relação entre um dado ou objeto que é transformado um processo ou pessoa que o transformou.
Relação de Influência	Afeta	Relação entre um objeto ou processo que é afetado por outro
	Analisa	Relação entre uma pessoa que analisa e um objeto ou processo de pesquisa
	Ocorre em	Relação entre uma processo ou fenômeno e o espaço onde ele ocorre
	Avaliado por	Relação entre um objeto ou processo e seu avaliador
	É agente de	Relação entre algo que exerce alguma ação ou produz algum efeito sobre outrem.
	Coloca em	Relação entre algo que é colocado e onde é colocado
	É Causado por	Relação entre uma coisa ou processo e o que a causou.
	É gerado por	Relação que liga o resultado de uma pesquisa ao processo ou pessoa que o gerou
	É levantado por	Relação entre um levantamento e pessoa ou instrumento que foi utilizado para tal.
	É planejado por	Relação entre um processo e que o planejou
	É produto de	Relação entre um processo e seu resultado
	É produzido por	Relação entre um resultado e o processo ou pessoa que o gerou.
	Pesquisa	Relação entre um pesquisador e seu objeto de pesquisa.
	Planeja	Relação que liga o planejador ao seu planejamento
	Produz	Relação que liga o produtor ao que é produzido
	Escreve	Relação entre quem escreve e o que é escrito.
	Estuda em	Relação de um aluno com sua Instituição de ensino
	Levanta	Relação entre a pessoa que faz o levantamento e o resultado deste levantamento
	Interage com	Relação entre duas pessoas, duas substâncias, dois elementos ou qualquer outras coisas que interajam entre si
	Mede	Relação entre uma pessoa ou instrumento e a medida encontrada
	Método de	Relação entre um método e o seu resultado
	Normaliza	Relação entre uma pessoa ou norma utilizada e sua normalização
	Orienta	Relação entre um orientador e uma pessoa ou projeto acadêmico orientado por ele
	Recebe informação de	Relação entre alguém que recebe e a pessoa que fornece a informação
	Recomenda	Relação entre uma pessoa que recomenda e algo que é recomendado
	Resulta em	Relação entre um processo e seu resultado
	É recomendado por	Relação entre algo que é recomendado e a pessoa que o recomendou
	Reusa	Relação que liga um pesquisador aos dados reusados de outra pesquisa
	Traz	Relação entre o transporte e que é transportado
	Usa	Relação entre uma pessoa ou processo e o que é utilizado no desenvolvimento de uma pesquisa
	Usado em	Relação entre uma substancia, um processo ou instrumento que é utilizado em uma pesquisa
	Valida	Relação entre o validador e método ou processo que é validado.
	Verifica	Relação entre o verificador e o método ou processo que é verificado
	Verificado por	Relação entre o que é verificado e o verificador
	Realiza	Relação entre quem realiza e o que é realizado
	Trabalha para	Relação entre uma pessoa e a Instituição para a qual trabalha
	Avalia	Relação entre quem avalia e o que é avaliado.
	É curador de	Relação que revela quem é o responsável pela curadoria do objeto.
	É inovação para	Relação entre uma inovação e objetivo para o qual ela serve.
	Fornecer informação para	Relação entre uma pessoa ou documento que fornece informação para outra pessoa ou documento
	Apóia	Relação entre uma pessoa, ou Instituição que apóia uma pesquisa
	É apoiado por	Relação entre uma pesquisa e uma pessoa ou Instituição que apóia o seu desenvolvimento.
	Causa	Relação entre um coisa ou processo que causa outra coisa ou processo

	É causado por	Relação entre uma coisa ou processo e o que a causou
	É medido por	Relação entre uma medida e o instrumento utilizado para tal
	Impacta	Relação entre uma coisa ou processo e outra impactada por ela.
	É diretor de	Relação que revela a função de um pessoa em uma Instituição ou que exerce sob outras pessoas.
	É impactado por	Relação que revela o impacto de uma coisa sobre a outra
Relação de Propriedade	Diagnostica	Relação entre processo, pessoa ou equipamento que diagnostica algo e o resultado do diagnóstico
	Revisado por	Relação que se dá entre algo que foi revisto e quem o revisou.
	É certificador de	Relação entre quem ou o que certifica e um objeto ou pesquisa
	É curado por	Relação entre o objeto e o responsável por sua curadoria digital.
	É autor de	Relação entre o responsável intelectual da obra e sua obra
	É financiado por	Relação que revela quem é o responsável pelo financiamento de uma pesquisa
	É financiadora de	Relação entre o objeto ou pesquisa financiada e seu financiador
	É medida de	Relação que revela que uma unidade serve como medida para algo
	É orientado por	Relação entre um aluno do PPGIEN e um professor doutor que o orienta.
	É par de	Pessoa que pesquisa na mesma temática de outra, podendo ser um avaliador, um colega de área, não sendo necessariamente um co-autor
	É parceiro de	Pessoa que pesquisa na mesma temática de outra, sendo co-autor em diversos trabalhos
	É publicado por	Relação que liga um documento ao responsável por sua publicação
	É publicadora de	Relação que liga um documento ao responsável por sua publicação
	Propriedade de	Relação que estabelece quem tem a propriedade intelectual sobre determinado objeto ou pesquisa
	Cria	Relação que revela o responsável pela criação de algo
	Tem autoria de	Relação entre a obra e seu responsável intelectual.
	Destrói	Relação entre algo que destrói e algo que foi destruído
	É programado por	Relação entre um <i>software</i> e que o desenvolveu
	Programa	Relação entre o programador e o <i>software</i>
	É produzido por	Relação entre um produto e o processo ou pessoa que o produziu
	Modela	Relação entre o modelizador e o objeto ou situação modelada
	Diagnostica	Relação entre processo, pessoa ou equipamento que diagnostica algo e o resultado do diagnóstico
	É Modelado por	Relação entre o objeto modelado e seu modelizador
	É criado por	Relação entre um objeto ou processo e quem o criou.

APÊNDICE D - PROPOSTA DE TAXONOMIA DE RELAÇÕES PARA USO EM PUBLICAÇÕES AMPLIADAS

CLASSE DE RELAÇÃO	RELAÇÃO	CLASSIFICAÇÃO	ONTOLOGIA/ VOCAB. SEMÂNTICO
EPRINT_EPRINT	Cita	Citação	CITO
	É citado por	Citação	CITO
	Discorda	Citação	CITO
	Discute	Citação	CITO
	É citado como evidência por	Citação	CITO
	É citado como Fonte documental por	Citação	CITO
	É citado como Fonte de dado por	Citação	CITO
	É citado como informação por	Citação	CITO
	É citado como leitura recomendada	Citação	CITO
	É confirmado por	Citação	CITO
	É creditado por	Citação	CITO
	É criticado por	Citação	CITO
	Descreve	Citação	CITO
	E referenciado por	Citação	CITO
	É refutado por	Citação	CITO
	Esta de acordo com	Citação	CITO
	Inclui citação de	Citação	CITO
	Referencia	Citação	CITO
	Responde a	Citação	CITO
	Retrata	Citação	CITO
	Tem resposta	Citação	CITO
	Atualiza	Derivação	CITO
	Compila	Derivação	CITO
	Deriva	Derivação	CITO
	É atualizado por	Derivação	CITO
	É base de	Derivação	CITO, DATACITE
	É baseado em	Derivação	CITO, DATACITE
	É compilado por	Derivação	CITO, DATACITE
	É continuação de	Derivação	CITO, DATACITE
	É continuado por	Derivação	CITO, DATACITE
	É corrigido por	Derivação	DATACITE
	É estendido por	Derivação	CITO
	cita como Fonte documental	Citação	CITO
	Cita como Fonte	Citação	CITO
	Cita como leitura recomendada	Citação	CITO
	Cita como relacionado	Citação	CITO
	Cita como solução potencial	Citação	CITO
	Cita como Fonte de dados	Citação	CITO
	Cita para informar	Citação	CITO
	Credita	Influência	CITO
	Da suporte para	Influência	CITO
	Divulga	Influência	VOC_IEN
	Documenta	Influência	DATA CITE
	Descreve	Influência	CITO
	É descrito por	Influência	CITO
	É documentado por	Influência	CITO
	Fornecer informação para	Influência	CITO
	Fornecer método para	Influência	CITO
	Fornecer trecho para	Influência	CITO
	Recomenda	Influência	VOC_IEN
	Prover afirmação para	Influência	CITO
	Prover conclusão para	Influência	CITO
	Prover dado para	Influência	CITO
	Usa conclusão de	Influência	CITO
	Usa dado de	Influência	CITO
	É nova versão de	Derivação	DATA CITE
	é revisado	Derivação	CITO
	É Revisão de (VOC_IEN)	Derivação	VOC-IEN

	Usa método de	Influencia	CITO
	É divulgado por	Influencia	VOC_IEN
	É substituído por	Derivação	DATA CITE
	É uma forma original de	Derivação	DATA CITE
	É uma forma variante de	Derivação	DATA CITE
	É versão de	Derivação	DATA CITE
	Gera	Derivação	VOC_IEN
	Revisa	Derivação	VOC_IEN
	Revisado por	Derivação	VOC_IEN
	Tem versão	Derivação	DATA CITE
	Substitui	Derivação	DATA CITE
	Transforma em	Derivação	VOC_IEN
	É derivação de	Derivação	OAI-ORE
	Tem derivação	Derivação	OAI-ORE
	Corrige	Influencia	CITO
	Concorda com	Citação	CITO
	Confirma	Citação	CITO
	É refutado por	Citação	CITO
	Revisa	Citação	CITO
	Inclui trecho de	Citação	CITO
	E parte de	Meronímica	DoCo
	Contém afirmações de	Citação	CITO
	É continuação de	Derivação	DATA CITE
EPRINT_DADO	Cita como Fonte	Citação	CITO
	Cita	Citação	CITO
	Cita como relacionado	Citação	CITO
	Cita como solução potencial	Citação	CITO
	Cita para informar	Citação	CITO
	Concorda com	Citação	CITO
	Discute	Citação	CITO
	Documenta	Influencia	DATA CITE
	Cita como evidencia	Citação	CITO
	Confirma	Citação	CITO
	Discute	Citação	CITO
	Referencia	Citação	CITO
	É baseado em	Derivação	CITO
	Compila	Derivação	CITO
	Analisa	Influencia	VOC_IEN
	Credita	Influencia	CITO
	Divulga	Influencia	VOC_IEN
	Descreve	Influencia	CITO
	Fornecer método para	Influencia	VOC_IEN
	Recomenda	Influencia	VOC_IEN
	Reusa	Influencia	VOC_IEN
	Valida	Influencia	VOC_IEN
EPRINT_PESSOA	Avaliado por	Influência	VOC_IEN
	É orientado por	Propriedade	VOC_IEN
	Tem autoria de	Propriedade	VOC_IEN
	Cita	Citação	CITO
	Cita como autoridade	Citação	CITO
	Cita como Fonte	Citação	CITO
	Concorda com	Citação	CITO
	Contem afirmações de	Citação	CITO
	Discorda com	Citação	CITO
	É citado como autoridade	Citação	CITO
	É citado como evidencia por	Citação	CITO
	É citado como Fonte documental por	Citação	CITO
	É citado como informação por	Citação	CITO
	É citado como leitura recomendada por	Citação	CITO
	É citado por	Citação	CITO, DATA CITE
	É criticado por	Citação	CITO
	É atualizado por	Citação	CITO
	É compilado por	Citação	CITO
	E referenciado por	Citação	CITO
	Referencia	Citação	CITO
	É continuado por	Derivação	DATA CITE

	É corrigido por	Derivação	DATA CITE
	É estendido por	Derivação	DATA CITE
	Revisado por	Derivação	DATA CITE
	É curado por	Propriedade	VOC_IEN
	Propriedade de	Propriedade	VOC_IEN
	Qualifica	Propriedade	CITO
	Tem autoria de	Propriedade	VOC_IEN
	É descrição de	Propriedade	OAI-ORE
	Tem metadado para	Propriedade	OAI-ORE
	Avaliado por	Influencia	VOC-IEN
	Descreve	Influencia	CITO
	É descrito por	Influencia	CITO
	É gerado por	Influencia	VOC_IEN
	É levantado por	Influencia	VOC-IEN
	É planejado por	Influencia	VOC-IEN
	É produzido por	Influencia	CITO
	É qualificado por	Influencia	CITO
	É usado por	Influencia	VOC-IEN
	Prover Afirmação para	Influencia	CITO
	Usa dado de	Influencia	CITO
	Verificado por	Influencia	VOC-IEN
	É divulgado por	Influencia	CITO
	Usa método de	Influencia	VOC-IEN
EPRINT_PROJETO	É resultado de	Meronímica	DoCo
	Cita	Citação	CITO
	Cita como Fonte	Citação	CITO
	Concorda com	Citação	CITO
	Contém afirmações de	Citação	CITO
	É citado como Fonte documental	Citação	CITO
	É citado como evidencia por	Citação	CITO
	É citado como informação por	Citação	CITO
	Deriva	Derivação	CITO
	É baseado em	Derivação	CITO, DATACITE
	É compilado por	Derivação	CITO, DATACITE
	É continuação de	Derivação	DATACITE
	É uma forma variante de	Derivação	DATACITE
	É versão de	Derivação	DATACITE
	É derivação de	Derivação	OAI-ORE
	É descrição de	Derivação	OAI-ORE
	Analisa	Influencia	VOC-IEN
	Divulga	Influencia	VOC-IEN
	Descreve	Influencia	CITO
	É descrito por	Influencia	CITO
	É gerado por	Influencia	VOC-IEN
	É produto de	Influencia	VOC-IEN
	É referenciado por	Citação	CITO
	Está de acordo com	Citação	CITO
	Inclui citação de	Citação	CITO
	Inclui trecho de	Citação	CITO
	Referencia	Citação	CITO
	Indica	Influencia	CITO
	Fornece informação para	Influencia	CITO
	Fornece trecho para	Influencia	CITO
	Prover afirmação para	Influencia	CITO
	Usa dado de	Influencia	CITO
	Usa método de	Influencia	CITO
	É divulgado por	Influencia	CITO, DATA CITE
	É citado por	Citação	CITO
EPRINT_ORG	Cita	Citação	CITO
	É financiado por	Propriedade	VOC_IEN
	Cita	Citação	CITO
	Cita como autoridade	Citação	CITO
	Cita como evidencia	Citação	CITO
	Cita como Fonte de dados	Citação	CITO
	Cita como metadado	Citação	CITO
	É criticado por	Citação	CITO
	É atualizado por	Derivação	CITO
	É curado por	Propriedade	CITO
	É financiado por	Propriedade	VOC-IEN

	É publicado por	Propriedade	VOC-IEN
	Qualifica	Propriedade	CITO
	Tem autoria de	Propriedade	VOC-IEN
	É produto de	Influencia	VOC-IEN
	É referenciado por	Citação	CITO
	Referencia	Citação	CITO
	Indica	Influencia	CITO
	Fornecer informação para	Influencia	CITO
	Fornecer método para	Influencia	CITO
	Fornecer trecho para	Influencia	CITO
	Obtém apoio de	Influencia	CITO
	Recomenda	Influencia	VOC-IEN
	Prover dado para	Influencia	CITO
	Usa método de	Influencia	CITO
	Usa dado de	Influencia	VOC-IEN
	É divulgado por	Influencia	VOC-IEN
	É descrição de	Influencia	OAI-ORE
	Analisa	Influencia	VOC-IEN
	Divulga	Influencia	VOC-IEN
	Descreve	Influencia	CITO
	É descrito por	Influencia	CITO
	É propriedade de direito	Propriedade	CERIF
	É gerado por	Influencia	VOC-IEN
DADO_DADO	Atualiza	Derivação	CITO
	Cita	Citação	CITO
	Confirma	Citação	CITO
	É citado por	Citação	CITO, DATA CITE
	É confirmado por	Citação	CITO
	Atualiza	Derivação	CITO
	Compila	Derivação	CITO
	Deriva	Derivação	CITO
	E atualizado por	Derivação	CITO
	E base de	Derivação	CITO
	E baseado em	Derivação	CITO-DATA CITE
	E compilado por	Derivação	CITO, DATA CITE
	E continuação de	Derivação	CITO, DATA CITE
	E continuado por	Derivação	CITO, DATA CITE
	E corrigido por	Derivação	DATA CITE
	E estendido por	Derivação	DATA CITE
	E nova versão de	Derivação	DATA CITE
	E revisado por	Derivação	CITO
	E revisão de	Derivação	CITO
	E substituído por	Derivação	DATA CITE
	É forma original de	Derivação	DATA CITE
	E forma variante de	Derivação	DATA CITE
	E versão de	Derivação	DATA-CITE
	Gera	Derivação	VOC-IEN
	Tem versão	Derivação	DATA CITE
	Substitui	Derivação	DATA CITE
	Transforma em	Derivação	VOC-IEN
	E derivação de	Derivação	OAI-ORE
	Tem derivação	Derivação	OAI-ORE
	E metadado de	Propriedade	CITO
	E descrição de	Propriedade	OAI-ORE
	Tem descrição	Propriedade	OAI-ORE
	Tem metadado	Propriedade	OAI-ORE
	É anotação de	Propriedade	OAI-ORE
	Tem anotação para	Propriedade	OAI-ORE
	Tem equivalente	Propriedade	OAI-ORE
	Analisa	Influencia	VOC-IEN
	Apóia	Influencia	CITO
	Corrige	Influencia	CITO
	Destrói	Influencia	VOC-IEN
	Documenta	Influencia	DATA-CITE
	Descreve	Influencia	CITO
	E requerido por	Influencia	DATA CITE
	E dependente de	Influencia	OAI-ORE
	Tem dependente	Influencia	OAI-ORE
	E descrito por	Influencia	CITO
	E documentado por	Influencia	CITO

	Requer	Influencia	DATA CITE
	Resulta em	Influencia	VOC-IEN
	Reusa	Influencia	VOC-IEN
	E gerado por	Influencia	VOC-IEN
	Estende	Influencia	DATA CITE
	Indica	Influencia	CITO
	Embasa	Influencia	CITO
	Tem equivalente	Derivação	OAI-ORE
	É versão de	Derivação	DATA CITE
	É suplementado por	Influência	DATA CITE
DADO_EPRINT	É figura de	Meronímica	DOCO
	Prover dado para	Influencia	CITO
	É citado como evidencia	Citação	CITO
	E citado como informação por	Citação	CITO
	E citado por	Citação	CITO
	E criticado por	Citação	CITO
	E referenciado por	Citação	CITO
	Esta de acordo com	Citação	CITO
	E base de	Derivação	CITO, DATA CITE
	E compilado por	Derivação	CITO, DATA CITE
	E publicado por	Derivação	OAI-ORE
	E metadado de	Derivação	CITO
	Tem metadado para	Derivação	OAI-ORE
	E anotação de	Derivação	OAI-ORE
	Tema notação para	Derivação	OAI-ORE
	E metadado de	Derivação	OAI-ORE
	E publicado por	Derivação	VOC-IEN
	Apóia	Influência	CITO
	E avaliado por	Influência	VOC-IEN
	Credita	Influência	CITO
	Da suporte para	Influência	CITO
	Descreve	Influência	CITO
	E requerido por	Influência	CITO
	E descrito por	Influência	CITO
	E documentado por	Influência	CITO
	E produto de	Influência	VOC-IEN
	E gerado por	Influência	VOC-IEN
	E qualificado por	Influência	CITO
	Embasa	Influência	CITO
	Fornecer afirmação para	Influência	CITO
	Confirma	Citação	CITO
	Valida	Influência	VOC-IEN
	E divulgado por	Influência	VOC-IEN
	Serve de suporte	Influência	CITO
	Usado em	Influência	CERIF
	Usa dado de	Influência	VOC-IEN
DADO_PESSOA	É compilado por	Derivação	CITO; DATA CITE
	Cita.cita como autoridade	Citação	CITO
	Cita como metadado	Citação	CITO
	E citado como autoridade	Citação	CITO
	E citado como evidencia por	Citação	CITO
	e citado como informação por	Citação	CITO
	E citado como leitura recomendada por	Citação	CITO
	E citado por	Citação	CITO
	E confirmado por	Citação	CITO
	E creditado por	Citação	CITO
	E criticado por	Citação	CITO
	E referenciado por	Citação	CITO
	E refutado por	Citação	CITO
	Esta de acordo com	Citação	CITO
	E atualizado por	Derivação	CITO
	E continuado por	Derivação	CITO, DATA CITE
	E corrigido por	Derivação	CITO, DATA CITE
	E estendido por	Derivação	DATA CITE
	Revisado por	Derivação	CITO
	E curado por	Propriedade	VOC-IEN
	E metadado de	Propriedade	CITO
	E publicado por	Propriedade	VOC-IEN
	Tem autoria de	Propriedade	VOC-IEN

	E descrição de	Propriedade	OAI-ORE
	Tem descrição	Propriedade	OAI-ORE
	E metadado de	Propriedade	OAI-ORE
	E anotação de	Propriedade	OAI-ORE
	Avaliado por	Influência	VOC-IEN
	E requerido por	Influência	DATA CITE
	E descrito por	Influência	CITO
	E documentado por	Influência	CITO
	E gerado por	Influência	VOC-IEN
	E levantado por	Influência	VOC-IEN
	E planejado por	Influência	VOC-IEN
	E produzido por	Influência	VOC-IEN
	E qualificado por	Influência	CITO
	E usado por	Influência	VOC-IEN
	Fornecer informação para	Influência	CITO
	Verificado por	Influência	VOC-IEN
	E divulgado por	Influência	VOC-IEN
	É curado por	Propriedade	VOC_IEN
DADO_PROJETO	Embasa	Influência	CITO
	Confirma	Citação	CITO
	É citado como evidencia	Citação	CITO
	E citado como Fonte de dado para	Citação	CITO
	É citado como informação por	Citação	CITO
	É citado por	Citação	CITO
	É criticado por	Citação	CITO
	É referenciado por	Citação	CITO
	Está de acordo com	Citação	CITO
	É base de	Derivação	CITO/ DATACITE
	É financiado por	Propriedade	VOC-IEN
	É metadado de	Propriedade	CITO
	É publicado por	Propriedade	VOC-IEN
	É descrição de	Propriedade	OAI-ORE
	Tem metadado para	Propriedade	OAI-ORE
	É anotação de	Propriedade	OAI-ORE
	Afeta	Influencia	VOC-IEN
	Apóia	Influencia	CITO
	Avaliado por	Influencia	VOC-IEN
	Coloca em	Influencia	VOC-IEN
	Descreve	Influencia	CITO
	É requerido por	Influencia	DATA CITE
	É descrito por	Influencia	CITO
	É documentado por	Influencia	CITO
	É levantado por	Influencia	VOC-IEN
	É gerado por	Influencia	VOC-IEN
	É planejado por	Influencia	VOC-IEN
	É produto de	Influencia	VOC-IEN
	É suplementado por	Influencia	VOC-IEN
	É usado por	Influencia	VOC-IEN
	Embasa	Influencia	CITO
	Fornecer informação para	Influencia	CITO
	Prover dado para	Influencia	CITO
	Usado em	Influencia	VOC-IEN
	É divulgado por	Influencia	VOC-IEN
	É apêndice de	Meronímica	DoCo
DADO_ORG	Financiado por	Propriedade	VOC_IEN
	Cita como metadado	Citação	CITO
	É citado como autoridade	Citação	CITO
	É citado como evidencia por	Citação	CITO
	É citado como Fonte de dado para	Citação	CITO
	É citado como informação por	Citação	CITO
	É citado por	Citação	CITO
	Fornecer informação para	Influencia	CITO
	É referenciado por	Citação	CITO
	É atualizado por	Derivação	CITO
	É compilado por	Derivação	CITO, DATA CITE
	É corrigido por	Derivação	DATA CITE
	É estendido por	Derivação	DATA CITE
	Revisado por	Derivação	CITO
	É curado por	Propriedade	VOC-IEN
	Usado em	Propriedade	VOC-IEN

	É metadado de	Propriedade	CITO
	É publicado por	Propriedade	VOC-IEN
	Term autoria de	Propriedade	VOC-IEN
	É descrição de	Propriedade	OAI-ORE
	E qualificado por	Influencia	CITO
	Avaliado por	Influencia	VOC-IEN
	É requerido por	Influencia	DATA CITE
	E descrito por	Influencia	CITO
	É documentado por	Influencia	VOC-IEN
	É gerado por	Influencia	VOC-IEN
	É levantado por	Influencia	VOC-IEN
	É planejado por	Influencia	VOC-IEN
	É produzido por	Influencia	VOC-IEN
	É usado por	Influencia	VOC-IEN
	E divulgado por	Influencia	VOC-IEN
	É creditado por	Citação	CITO
	Supervisiona	Influencia	CERIF
	Cita	Citação	CITO
	Cita como autoridade	Citação	CITO
	Cita como Fonte	Citação	CITO
	Concorda com	Citação	CITO
	Discorda com	Citação	CITO
	Discute	Citação	CITO
	É citado como autoridade por	Citação	CITO
	E citado por	Citação	CITO
	E creditado por	Citação	CITO
	E criticado por	Citação	CITO
	E referenciado por	Citação	CITO
	E refutado por	Citação	CITO
	Esta de acordo com	Citação	CITO
	E orientado por	Propriedade	VOC_IEN
	E parceiro de	Propriedade	VOC_IEN
	E par de	Propriedade	VOC_IEN
	Avaliado por	Influência	VOC_IEN
	Indica	Influência	CITO
	Fornecer informação para	Influência	CITO
	Interage com	Influência	VOC_IEN
	Recebe informação de	Influência	VOC_IEN
	Usa dado de	Influência	CITO
	Usa método de	Influência	CITO
PESSOA_EPRINT	Supervisiona	Influência	CERIF
	Trabalha para	Influência	VOC-IEN
	Avalia	Influência	VOC-IEN
	Administra	Influencia	CERIF
	Orienta	Influencia	VOC_IEN
	É autor de	Propriedade	VOC_IEN
	Descreve	Influencia	CITO
	Cita como Fonte	Citação	CITO
	Concorda com	Citação	CITO
	Cita como evidencia	Citação	CITO
	Cita como Fonte documental	Citação	CITO
	Cita como leitura recomendada	Citação	CITO
	Confirma	Citação	CITO
	Discute	Citação	CITO
	discorda com	Citação	CITO
	É citado como autoridade por	Citação	CITO
	É citado por	Citação	CITO
	É criticado por	Citação	CITO
	É creditado por	Citação	CITO
	Esta de acordo com	Citação	CITO
	Referencia	Citação	CITO
	Atualiza	Derivação	CITO
	Revisa	Derivação	VOC-IEN
	É certificadora de	Propriedade	VOC-IEN
	É descrito por	Influencia	CITO
	É editor de	Influencia	CERIF
	É metadado de	Propriedade	CITO
	E proprietário de direito de	Propriedade	CITO
	Documenta	Influencia	DATA CITE
	E qualificado por	Influencia	CITO

	Pesquisa	Influencia	VOC-IEN
	Planeja	Influencia	VOC-IEN
	Produce	Influencia	VOC-IEN
	Indica	Influencia	CITO
	Prover afirmação para	Influencia	CITO
	Prover conclusão para	Influencia	CITO
	Usa	Influencia	VOC-IEN
	Usa conclusão de	Influencia	CITO
	Usa dado de	Influencia	CITO
	Usa método de	Influencia	CITO
	Valida	Influencia	VOC-IEN
	Verifica	Influencia	VOC-IEN
	Avalia	Influencia	VOC-IEN
	Escreve	Influencia	VOC-IEN
	Estende	Influencia	CITO
	Fornecer informação para	Influencia	CITO
	Fornecer método para	Influencia	VOC-IEN
	Levanta	Influencia	VOC-IEN
	Normaliza	Influencia	VOC-IEN
	Orienta	Influencia	VOC-IEN
	Recomenda	Influencia	VOC-IEN
	E tradutor de	Propriedade	CERIF
	Publica	Propriedade	CERIF
	Qualifica	Propriedade	CITO
	Revindica direito	Propriedade	CERIF
	Tem autoria de	Propriedade	VOC-IEN
	Analisa	Influencia	VOC-IEN
	Cria	Influencia	VOC-IEN
	Corrige	Influencia	CITO
	Credita	Influencia	CITO
	Divulga	Influencia	VOC-IEN
	É revisor de	Propriedade	CERIF
	Cita	Citação	CITO
PESSOA_DADO	Publica	Propriedade	CERIF
	Cita	Citação	CITO
	Cita como Fonte de dados	Citação	CITO
	Cita como metadado	Citação	CITO
	Cita como relacionado	Citação	CITO
	Cita para informar	Citação	CITO
	Concorda com	Citação	CITO
	Confirma	Citação	CITO
	Referencia	Citação	CITO
	Atualiza	Derivação	CITO
	Compila	Derivação	CITO
	Gera	Derivação	CITO
	Revisa	Derivação	VOC_IEN
	Substitui	Propriedade	DATA CITE
	Transforma	Propriedade	VOC-IEN
	É certificadora de	Propriedade	VOC-IEN
	É autor de	Propriedade	VOC-IEN
	E editor de	Propriedade	VOC-IEN
	É metadado de	Propriedade	CITO
	E revisor de	Propriedade	CERIF
	Qualifica	Propriedade	CITO
	Tem anotação para	Propriedade	OAI-IRE
	Analisa	Influência	VOC-IEN
	Cria	Influência	VOC-IEN
	Corrige	Influência	CITO
	Credita	Influência	CITO
	Destrói	Influência	VOC-IEN
	Divulga	Influência	VOC
	Documenta	Influência	VOC
	Descreve	Influência	CITO
	Pesquisa	Influência	VOC-IEN
	Planeja	Influência	VOC-IEN
	Produce	Influência	VOC-IEN
	Indica	Influência	CITO
	Levanta	Influência	VOC-IEN
	Recomenda	Influência	VOC-IEN
	Reusa	Influência	VOC-IEN

	Valida	Influência	VOC-IEN
	Usa	Influência	VOC-IEN
	Verifica	Influência	VOC-IEN
	Avalia	Influência	VOC-IEN
	Administra	Influência	CERIF
PESSOA-PROJETO	Supervisiona	Influência	CERIF
	Cita	Citação	CITO
	Cita como evidencia	Citação	CITO
	.cita como Fonte	Citação	CITO
	Cita como Fonte documental	Citação	CITO
	Cita como leitura recomendada	Citação	CITO
	Cita como relacionado	Citação	CITO
	Cita como solução potencial	Citação	CITO
	Cita para informar	Citação	CITO
	Concorda com	Citação	CITO
	Confirma	Citação	CITO
	Discorda com	Citação	CITO
	É citado como autoridade por	Citação	CITO
	É citado por	Citação	CITO
	É referenciado por	Citação	CITO
	Está de acordo com	Citação	CITO
	Referencia	Citação	CITO
	Atualiza	Derivação	DATA CITE
	Gera	Derivação	VOC-IEN
	Revisa	Derivação	VOC-IEN
	Substitui	Derivação	DATA CITE
	É certificadora de	Propriedade	VOC-IEN
	É autor de	Propriedade	VOC-IEN
	É proprietário de direito de	Propriedade	CERIF
	Descreve	Propriedade	CITO
	É tradutor de	Propriedade	CERIF
	Publica	Propriedade	CERIF
	Qualifica	Propriedade	CITO
	Reivindica direito	Propriedade	CERIF
	É metadado de	Propriedade	CITO
	Analisa	Influencia	VOC-IEN
	Cria	Influencia	VOC-IEN
	Corrige	Influencia	CITO
	Credita	Influencia	VOC-IEN
	Divulga	Influencia	VOC-IEN
	Documenta	Influencia	DATA CITE
	Corrige	Influencia	CITO
	Divulga	Influencia	VOC-IEN
	Documenta	Influencia	DATA CITE
	É agente de	Influencia	VOC-IEN
	É qualificado por	Influencia	CITO
	Planeja	Influencia	VOC-IEN
	Produz	Influencia	VOC-IEN
	Indica	Influencia	CITO
	Escreve	Influencia	VOC-IEN
	Continua	Influencia	VOC-IEN
	Fornece informação para	Influencia	CITO
	Fornece método para	Influencia	CITO
	Levanta	Influencia	VOC-IEN
	Normaliza	Influencia	VOC-IEN
	Orienta	Influencia	VOC-IEN
	Recomenda	Influencia	DATA CITE
	Prover dado para	Influencia	CITO
	Usa	Influencia	VOC-IEN
	Usa conclusão de	Influencia	CITO
	Usa dado de	Influencia	VOC-IEN
	Usa método de	Influencia	CITO
	Valida	Influencia	VOC-IEN
	Verifica	Influencia	VOC-IEN
	Realiza	Influencia	VOC-IEN
	Avalia	Influencia	VOC-IEN
	Supervisiona	Influencia	CERIF
	É revisor de	Propriedade	CERIF
PESSOA_ORG	Administra	Influência	CERIF
	Cita	Citação	CITO

	Cita como autoridade	Citação	CITO
	Cita como Fonte de dados	Citação	CITO
	Cita como metadado	Citação	CITO
	É citado por	Citação	CITO
	É creditado por	Citação	CITO
	Responde a	Citação	CITO
	É financiado por	Propriedade	VOC-IEN
	Analisa	Influencia	VOC-IEN
	Avaliado por	Influencia	VOC-IEN
	Diagnostica	Influencia	VOC-IEN
	Divulga	Influencia	VOC-IEN
	Documenta	Influencia	VOC-IEN
	Descreve	Influencia	CITO
	É dependente de	Influencia	OAI-ORE
	É qualificado	Influencia	CITO
	Pesquisa	Influencia	VOC-IEN
	Planeja	Influencia	VOC-IEN
	Indica	Influencia	CITO
	Estuda	Influencia	VOC-IEN
	Fornecer informação para	Influencia	CITO
	Interage com	Influencia	CITO
	Obtém apoio de	Influencia	CITO
	Recebe informação de	Influencia	VOC-IEN
	Recomenda	Influencia	DATA CITE
	Prover dado para	Influencia	CITO
	Usa dado de	Influencia	VOC-IEN
	Usa método de	Influencia	CITO
	Supervisiona	Influencia	CERIF
	Avalia	Influencia	VOC-IEN
	É diretor de	Influencia	VOC-IEN
	Trabalha para	Influência	VOC_IEN
PROJETO_DADO	Cito	Citação	CITO
	cita como evidencia	Citação	CITO
	cita como Fonte de dados	Citação	CITO
	Cita como metadado	Citação	CITO
	Cita para informar	Citação	CITO
	Discute	Citação	CITO
	É confirmado por	Citação	CITO
	É creditado por	Citação	CITO
	Atualiza	Derivação	CITO
	Compila	Derivação	CITO
	Deriva	Derivação	CITO
	É baseado em	Derivação	DATA CITE, CITO
	Gera	Derivação	VOC-IEN
	Revisa	Derivação	VOC-IEN
	Substitui	Derivação	DATA CITE
	Tem descrição	Derivação	OAI-ORE
	Afeta	Influencia	VOC-IEN
	Cria	Influencia	VOC-IEN
	Corrige	Influencia	CITO
	Credita	Influencia	CITO
	Divulga	Influencia	VOC-IEN
	Documenta	Influencia	DATA CITE
	Descreve	Influencia	CITO
	É dependente de	Influencia	OAI-ORE
	É suplementado por	Influencia	DATA CITE
	Está de acordo com	Influencia	CITOC-IEN
	Referência	Influencia	DATA CITE
	Pesquisa	Influencia	VOC-IEN
	Planeja	Influencia	VOC-IEN
	Produz	Influencia	VOC-IEN
	Indica	Influencia	CITO
	Estende	Influencia	CITO
	Levanta	Influencia	VOC-IEN
	Requer	Influencia	DATA-CITE
	Resulta	Influencia	VOC-IEN
	Reusa	Influencia	VOC-IEN
	Usa	Influencia	VOC-IEN
	Valida	Influencia	VOC-IEN
	Verifica	Influencia	VOC-IEN

	Avalia	Influencia	VOC-IEN
	É parte de	meronimics	DoCo
	Cita	Citação	CITO
	Cita como Fonte	Citação	CITO
	Cita como Fonte de dados	Citação	CITO
	Cita como Fonte	Citação	CITO
	Cita como relacionado	Citação	CITO
	Cita para informar	Citação	CITO
	Concorda com	Citação	CITO
	Confirma	Citação	CITO
	Contem afirmações de	Citação	CITO
	Discorda	Citação	CITO
	É citado como Fonte de dado para	Citação	CITO
	É citado como Fonte documental por	Citação	CITO
	É citado como informação por	Citação	CITO
	É citado por	Citação	CITO
	É confirmado por	Citação	CITO
	É creditado por	Citação	CITO
	É criticado por	Citação	CITO
	É referenciado	Citação	CITO
	É refutado por	Citação	CITO
	Esta de acordo com	Citação	CITO
	Inclui citação de	Citação	CITO
	Inclui trecho de	Citação	CITO
	Referencia	Citação	CITO
	Deriva	Derivação	CITO
	Afeta	Influência	VOC-IEN
	É continuação de	Derivação	DATA CITE
	É continuado por	Derivação	DATA CITE
	É estágio de	Derivação	VOC-IEN
	É estendido por	Derivação	CITO
	É fase de	Derivação	VOC-IEN
	É nova versão de	Derivação	DATA-CITE
	É revisão de	Derivação	VOC-IEN
	É substituído por	Derivação	DATA-CITE
	É uma forma original de	Derivação	DATA-CITE
	É uma forma variante de	Derivação	DATA-CITE
	É versão de	Derivação	DATA-CITE
	Substitui	Derivação	DATA CITE
	É derivação de	Derivação	OAI-ORE
	Tem equivalente	Derivação	OAI-ORE
	Credita	Influencia	CITO
	Dá suporte para	Influencia	CITO
	Documenta	Influencia	DATA CITE
	Descreve	Influencia	CITO
	É requerido por	Influencia	DATA CITE
	É dependente de	Influencia	OAI-ORE
	É descrito por	Influencia	CITO
	É documentado por	Influencia	CITO
	É gerado por	Influencia	VOC-IEN
	É planejado por	Influencia	CITO
	É produzido por	Influencia	VOC-IEN
	É produto de	Influencia	VOC-IEN
	É produzido por	Influencia	VOC-IEN
	Indica	Influencia	CITO
	Fornecer método para	Influencia	CITO
	Recomenda	Influencia	CITO
	Prover dado para	Influencia	CITO
	Usa dado de	Influencia	CITO
	Usa método de	Influencia	CITO
	É atualizado por	Derivação	CITO
PROJETO_PROJETO	Gera	Derivação	VOC_IEN
	Cita	Citação	CITO
	Cita como Fonte	Citação	CITO
	Cita como Fonte documental	Citação	CITO
	Cita como leitura recomendada	Citação	CITO
	Cita como relacionado	Citação	CITO
	Cita para informar	Citação	CITO
	Concorda com	Citação	CITO
	Confirma	Citação	CITO
PROJETO_EPRINT			

	Contem afirmações de	Citação	CITO
	É citado como Fonte documental por	Citação	CITO
	E citado como informação por	Citação	CITO
	E citado por	Citação	CITO
	E criticado por	Citação	CITO
	E referenciado por	Citação	CITO
	Esta de acordo com	Citação	CITO
	Inclui citação de	Citação	CITO
	Inclui trecho de	Citação	CITO
	Referencia	Citação	CITO
	Deriva	Derivação	CITO
	É base de	Derivação	CITO
	É baseado em	Derivação	DATA CITE, CITO
	É corrigido por	Derivação	DATA CITE, CITO
	É estendido por	Derivação	CITO
	É uma forma original de	Derivação	DATA CITE
	É uma forma variante de	Derivação	DATA CITE
	Tem versão	Derivação	DATA CITE
	Tem derivação	Derivação	OAI-ORE
	Credita	Influencia	CITO
	Dá suporte	Influencia	CITO
	É documentado por	Influencia	CITO
	Pesquisa	Influencia	VOC-IEN
	Produz	Influencia	VOC-IEN
	Indica	Influencia	CITO
	Embasa	Influencia	VOC-IEN
	Fornece informação para	Influencia	CITO
	Fornece método para	Influencia	CITO
	Fornece trecho para	Influencia	CITO
	Resulta em	Influencia	VOC-IEN
	Prover afirmação para	Influencia	CITO
	Prover conclusão para	Influencia	CITO
	Prover dado para	Influencia	CITO
	Usa conclusão de	Influencia	CITO
	Usa dado de	Influencia	CITO
	Usa método de	Influencia	CITO
	É divulgado por	Influencia	VOC-IEN
	É descrito por	Influência	CITO
PROJETO_PESSOA	Avaliado por	Influencia	VOC_IEN
	Cita	Citação	CITO
	Cita como autoridade	Citação	CITO
	É citado por	Citação	CITO
	É confirmado por	Citação	CITO
	É creditado por	Citação	CITO
	Cita como Fonte	Citação	CITO
	Cita como metadado	Citação	CITO
	Contém afirmações de	Citação	CITO
	É citado como Fonte documental por	Citação	CITO
	É criticado por	Citação	CITO
	É referenciado por	Citação	CITO
	É refutado por	Citação	CITO
	E compilado por	Derivação	CITO, DATA CITE
	É continuado por	Derivação	DATA CITE
	É corrigido por	Derivação	CITO
	Revisado por	Derivação	VOC-IEN
	É orientado por	Propriedade	VOC-IEN
	Descreve	Influencia	CITO
	É requerido por	Influencia	DATA CITE
	É descrito por	Influencia	CITO
	É documentado por	Influencia	CITO
	É gerado por	Influencia	VOC-IEN
	É levantado por	Influencia	VOC-IEN
	É planejado por	Influencia	VOC-IEN
	É produzido por	Influencia	VOC-IEN
	É qualificado por	Influencia	CITO
	Fornece informação para	Influencia	CITO
	Prover dado para	Influencia	CITO
	Usa método de	Influencia	CITO
	Verificado por	Influencia	VOC-IEN
	É divulgado por	Influencia	VOC-IEN

PROJETO_ORG	É atualizado por	Derivação	CITO
	É financiado por	Propriedade	VOC_IEN
	Cita	Citação	CITO
	Cita como autoridade	Citação	CITO
	Cita como Fonte de dados	Citação	CITO
	Cita como metadado	Citação	CITO
	É confirmado por	Citação	CITO
	É creditado por	Citação	CITO
	É referenciado por	Citação	CITO
	Esta de acordo com	Citação	CITO
	Referencia	Citação	CITO
	É atualizado	Derivação	CITO
	É financiado por	Propriedade	VOC-IEN
	Propriedade de	Propriedade	VOC-IEN
	Ocorre em	Propriedade	VOC-IEN
	Tem autoria	Propriedade	VOC-IEN
	É descrição de	Propriedade	OAI-ORE
	Avaliado por	Influencia	VOC-IEN
	Divulga	Influencia	VOC-IEN
	Descreve	Influencia	CITO
	É requerido por	Influencia	OAI-ORE
	É dependente de	Influencia	OAI-ORE
	É descrito por	Influencia	CITO
	É documentado por	Influencia	CITO
	É gerado por	Influencia	VOC-IEN
	É planejado por	Influencia	VOC-IEN
	É produto de	Influencia	VOC-IEN
	É produzido por	Influencia	VOC-IEN
	Fornece informação para	Influencia	VOC-IEN
	Prover dado para	Influencia	CITO
	Verificados por	Influencia	VOC-IEN
	É divulgado por	Influencia	CITO
	Usa dado de	Influência	CITO
	Obtém apoio de	Influência	CITO
ORG_ORG	É financiadora de	Propriedade	VOC_IEN
	Cita	Citação	CITO
	Cita como autoridade	Citação	CITO
	Cita como Fonte de dados	Citação	CITO
	Cita com relacionado	Citação	CITO
	Cita para informar	Citação	CITO
	É citado como autoridade por	Citação	CITO
	É citado como Fonte de dado para	Citação	CITO
	É citado por	Citação	CITO
	É certificadora de	Propriedade	VOC-IEN
	É financiado por	Propriedade	VOC-IEN
	É parceiro de	Propriedade	VOC-IEN
	Administra	Influencia	CERIF
	Apóia	Influencia	CITO
	Credita	Influencia	CITO
	Dá suporte para	Influencia	CITO
	Divulga	Influencia	VOC-IEN
	É dependente de	Influencia	OAI-ORE
	É qualificado por	Influencia	CITO
	Fornece informação para	Influencia	CITO
	Fornece método para	Influencia	CITO
	Interage com	Influencia	VOC-IEN
	Obtém apoio de	Influencia	CITO
	Recebe informação de	Influencia	VOC-IEN
	Recomenda	Influencia	VOC-IEN
	Prover dado para	Influencia	CITO
	Usa dado de	Influencia	CITO
	Usa método de	Influencia	CITO
	Valida	Influencia	VOC-IEN
	É publicadora de	Propriedade	VOC_IEN
ORG_EPRINT	É editor de	Propriedade	CERIF
	Cita	Citação	CITO
	Cita como Fonte	Citação	CITO
	Cita como Fonte documental	Citação	CITO
	Cita para informar	Citação	CITO
	É citado como autoridade por	Citação	CITO

	É citado como Fonte de dado para	Citação	CITO
	É citado por	Citação	CITO
	É criticado por	Citação	CITO
	É referenciado por	Citação	CITO
	É refutado por	Citação	CITO
	Referencia	Citação	CITO
	Está de acordo com	Citação	CITO
	É atualizado por	Derivação	CITO
	Gera	Derivação	VOC-IEN
	É autor institucional de	Propriedade	CERIF
	É editor de	Propriedade	CERIF
	É financiadora de	Propriedade	VOC-IEN
	É parceiro de	Propriedade	VOC-IEN
	É publicadora de	Propriedade	VOC-IEN
	Verifica	Influencia	VOC-IEN
	é divulgado por	Influencia	VOC-IEN
	Publica	Propriedade	CERIF
	Qualifica	Propriedade	CITO
	Reivindica dire ito	Propriedade	CERIF
	Tem autoria de	Propriedade	VOC-IEN
	É metadado de	Propriedade	OAI-ORE
	Avaliado por	Influencia	VOC-IEN
	Dá suporte para	Influencia	CITO
	Divulga	Influencia	VOC-IEN
	É descrito por	Influencia	CITO
	É documentado por	Influencia	CITO
	Produz	Influencia	VOC-IEN
	Indica	Influencia	CITO
	Embasa	Influencia	CITO
	Fornece informação para	Influencia	CITO
	Recomenda	Influencia	VOC-IEN
	Reusa	Influencia	VOC-IEN
	Prover afirmação para	Influencia	CITO
	Prover dado para	Influencia	CITO
	Usa dado de	Influencia	CITO
	Usa método de	Influencia	CITO
	Valida	Influencia	CITO
	É propriedade de direito	Propriedade	CERIF
	É autor institucional de	Propriedade	CERIF
ORG_DADO	Atualiza	Derivação	CITO
	Compila	Derivação	CITO
	Gera	Derivação	VOC-IEN
	Revisa	Derivação	VOC-IEN
	Substitui	Derivação	DATA CITE
	Transforma	Derivação	VOC-IEN
	cita	Citação	CITO
	É certificadora de	Propriedade	VOC-IEN
	É financiadora de	Propriedade	VOC-IEN
	É editor de	Propriedade	VOC-IEN
	É publicadora de	Propriedade	VOC-IEN
	Publica	Propriedade	VOC-IEN
	Qualifica	Propriedade	CITO
	Reivindica direito	Propriedade	CERIF
	Tem autoria de	Propriedade	VOC-IEN
	Analisa	Influencia	VOC-IEN
	Avaliado por	Influencia	VOC-IEN
	Dá suporte para	Influencia	CITO
	Divulga	Influencia	VOC-IEN
	Descreve	Influencia	CITO
	É dependente	Influencia	OAI-OIRE
	Produz	Influencia	VOC-IEN
	Indica	Influencia	CITO
	Embasa	Influencia	CITO
	Levanta	Influencia	VOC-IEN
	Publica	Propriedade	CERIF
	Requer	Influencia	DATA CITE
	Reusa	Influencia	VOC-IEN
	Usa	Influencia	VOC-IEN
	Valida	Influencia	VOC-IEN
	Verifica	Influencia	VOC-IEN

	Avalia	Influencia	VOC-IEN
	É proprietário de direito	Propriedade	CERIF
	É autor institucional de	Propriedade	CERIF
	Credita	Influência	CITO
ORG_PESSOA	É financiadora de	Propriedade	VOC_IEN
	Cito	Citação	CITO
	Cita como autoridade	Citação	CITO
	Cita como Fonte de dados	Citação	CITO
	Cita para informar	Citação	CITO
	Cita como relacionado	Citação	CITO
	Cita para informar	Citação	CITO
	É citado como autoridade por	Citação	CITO
	Concorda com	Citação	CITO
	Confirma	Citação	CITO
	É citado como autoridade por	Citação	CITO
	É citado como Fonte de dado para	Citação	CITO
	É citado como informação por	Citação	CITO
	É citado por	Citação	CITO
	É criticado por	Citação	CITO
	É referenciado por	Citação	CITO
	Esta de acordo com	Citação	CITO
	Referencia	Citação	CITO
	Responde a	Citação	CITO
	É atualizado por	Derivação	CITO
	É certificadora de	Propriedade	VOC-IEN
	É financiado por	Propriedade	VOC-IEN
	Qualifica	Propriedade	CITO
	Administra	Influencia	CERIF
	Apóia	Influencia	CITO
	Avaliado por	Influencia	VOC-IEN
	Valida	Influencia	VOC-IEN
	Dá suporte para	Influencia	CITO
	Divulga	Influencia	VOC-IEN
	É depende de	Influencia	OAI-ORE
	E descrito por	Influencia	CITO
	É documento por	Influencia	CITO
	É planejado por	Influencia	VOC-IEN
	É qualificado por	Influencia	CITO
	Indica	Influencia	CITO
	Fornece informação para	Influencia	CITO
	Interage com	Influencia	VOC-IEN
	Supervisiona	Influencia	CERIF
	Recebe informação de	Influencia	VOC-IEN
	Requer	Influencia	DATA CITE
	Prover dado para	Influencia	CITO
	Usa conclusão	Influencia	CITO
	Avalia	Influencia	VOC-IEN
	Usa método de	Influencia	CITO
	Credita	Influência	CITO
ORG_PROJETO	Apóia	Influência	CITO
	cita	Citação	CITO
	Cita como Fonte de dados	Citação	CITO
	Cita como Fonte documental	Citação	CITO
	Cita como leitura recomendada	Citação	CITO
	Cita como solução potencial	Citação	CITO
	Cita para informar	Citação	CITO
	Confirma	Citação	CITO
	Discute	Citação	CITO
	É citado como autoridade por	Citação	CITO
	É citado como Fonte de dado para	Citação	CITO
	É citado por	Citação	CITO
	É creditado por	Citação	CITO
	É referenciado por	Citação	CITO
	Está de acordo com	Citação	CITO
	Referencia	Citação	CITO
	Atualiza	Derivação	CITO
	Gera	Derivação	VOC-IEN
	Revisa	Derivação	VOC-IEN
	Substitui	Derivação	DATA CITE
	É autor institucional de	Propriedade	CERIF

É certificadora de	Propriedade	VOC-IEN
É financiadora de	Propriedade	VOC-IEN
É metadado de	Propriedade	CITO
É financiado por	Propriedade	VOC-IEN
É proprietário de direito	Propriedade	CERIF
É publicadora de	Propriedade	VOC-IEN
Qualifica	Propriedade	CITO
Tem descrição	Propriedade	OAI-ORE
Tem metadado para	Propriedade	OAI-ORE
Tem anotação para	Propriedade	OAI-ORE
Administra	Influencia	CERIF
Analisa	Influencia	VOC-IEN
Avaliado por	Influencia	VOC-IEN
Cria	Influencia	VOC-IEN
Corrige	Influencia	CITO
Credita	Influencia	CITO
Dá suporte para	Influencia	CITO
Divulga	Influencia	VOC-IEN
Documenta	Influencia	DATA-CITE
Descreve	Influencia	CITO
É agente	Influencia	VOC-IEN
É descrito por	Influencia	CITO
É documentado por	Influencia	CITO
É planejado por	Influencia	VOC-IEN
Planeja	Influencia	VOC-IEN
Produz	Influencia	VOC-IEN
Indica	Influencia	CITO
Embasa	Influencia	CITO
Estende	Influencia	CITO
Fornecer informação para	Influencia	CITO
Fornecer método para	Influencia	CITO
Normaliza	Influencia	VOC-IEN
Recebe informação de	Influencia	VOC-IEN
Recomenda	Influencia	VOC-IEN
Requer	Influencia	VOC-IEN
Reusa	Influencia	DATA CITE
Prover dado para	Influencia	VOC-IEN
Usa conclusão de	Influencia	CITO
Usa dado de	Influencia	CITO
Usa método de	Influencia	CITO
Valida	Influencia	CITO
Verifica	Influencia	VOC-IEN
Realiza	Influencia	VOC-IEN
É divulgado por	Influencia	VOC-IEN
Serve de suporte	Influencia	CITO
Supervisiona	Influencia	CERIF
Avalia	Influencia	VOC-IEN
É proprietário de direito	Propriedade	CERIF