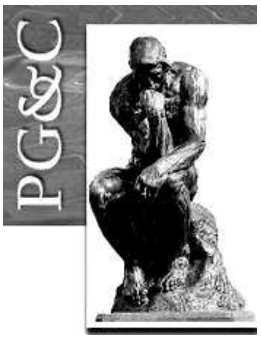


GESTÃO DE DADOS CIENTÍFICOS: NECESSIDADES E DESAFIOS DE UM PROGRAMA CIENTÍFICO PARA A AMAZÔNIA

Ronaldo Ferreira da Silva
Mestre em Gestão do Conhecimento e Tecnologia da Informação pela Universidade Católica de Brasília, Brasil.
Professor da Universidade Estadual de Goiás, Brasil.
E-mail: ronaldofsilva1@gmail.com

Edilson Ferneda
Doutor em Ciência da Computação pelo *Laboratoire d'Informatique, Robotique et Microélectronique de Montpellier*, França.
Professor da Universidade Católica de Brasília, Brasil.
E-mail: eferneda@gmail.com

Fernando William Cruz
Doutor em Ciência da Informação pela Universidade de Brasília, Brasil.
Professor da Universidade de Brasília, Brasil.
E-mail: fwcruz@gmail.com

Ana Paula Bernardi da Silva
Doutora em Engenharia Elétrica pela Universidade de Brasília, Brasil.
Professora da Universidade Católica de Brasília, Brasil.
E-mail: anap.bernardi@gmail.com

Luiza Beth Nunes Alonso
Doutora em Educação pela Universidade de Harvard, EUA.
Professora da Universidade Católica de Brasília, Brasil.
E-mail: luiza.alonso@hotmail.com

Resumo
A ciência contemporânea caracteriza-se pelo grande volume de dados, reflexo da crescente complexidade dos problemas de pesquisa, exigindo cada vez mais a colaboração entre grupos interdisciplinares e interinstitucionais. Nesse contexto, a Tecnologia da Informação vem propiciando novos ambientes que viabilizem essa dinâmica de trabalho. Este artigo apresenta um conjunto de dimensões conceituais com potencial de atender as necessidades para gestão do conhecimento e de dados científicos no contexto de um programa de pesquisa para a Amazônia. Um levantamento bibliográfico possibilitou a identificação de eixos temáticos e a análise dos dados coletados por meio de entrevistas semiestruturadas com pesquisadores, técnicos e dirigentes do programa de pesquisa possibilitou a identificação das dimensões conceituais e sua correlação com os eixos temáticos. O estudo apontou essas dimensões como fundamentais para uma efetiva gestão do conhecimento e de dados científicos produzidos no âmbito do programa e para o desenho de plataformas de *e-Science* que venham a dar suporte às atividades de pesquisa no programa estudado.

Palavras-chave: Gestão de Dados; e-Science; Curadoria Digital; Gestão de Conhecimento Científico.

SCIENTIFIC DATA MANAGEMENT:
NEEDS AND CHALLENGES OF A SCIENTIFIC PROGRAM FOR AMAZON REGION

Abstract
Contemporary science is characterized by a large volume of data due to the growing complexity of research problems which in turn requires increased collaboration between interdisciplinary and inter-institutional groups. In this context, the Information Technology has been providing new environments that enable this dynamic work. This dissertation analyzes electronic science (e-Science) platform components available today and lists a set of indicators, involving hybrid architectures, that are able to efficiently meet the specific scientific data management needs of a research program for Amazon region. The method used to analyze the collected data through semi-structured interviews focuses on the categorization of information, which in turn serves as the basis for the proposed indicators, using Content Analysis techniques.

Keywords: *Data Management; e-Science; Digital Curatorship; Management of Scientific Knowledge.*

1 INTRODUÇÃO

A importância da Amazônia brasileira para o ecossistema mundial em função de suas riquezas naturais tem atraído pesquisadores do mundo inteiro para realização de pesquisas vocacionadas para a interdisciplinaridade diante da complexidade de variáveis que envolvem as questões ligadas à natureza e suas interfaces endógenas e exógenas. Com objetivos que incluem a melhoria das condições de vida da população e a geração sustentável de riquezas na região várias instituições presentes na Amazônia participam simultaneamente de diversas redes de pesquisa científica, tais como: *(i)* Geoinformação para Gestão Ambiental (GEOMA); *(ii)* Rede de Saúde e condições de vida de povos Indígenas na Amazônia – Pronex; *(iii)* Rede Amazônica de Pesquisa e Desenvolvimento de Biocosméticos – RedeBio; *(iv)* Rede Biodiversidade e Biotecnologia da Amazônia Legal – Bionorte; e *(v)* Rede de pesquisa em Malária, das quais o Instituto de Pesquisas da Amazônia (INPA) participa.

O INPA[1] é um dos órgãos responsáveis pela realização de pesquisas sobre o ecossistema amazônico e sua dinâmica relacional com múltiplos agentes, registrando e acompanhando (*online*) as intervenções e seus efeitos e impactos nos recursos naturais e nas populações ribeirinhas na perspectiva da paisagem global (*landscape perspective*). Ao longo dos anos, o INPA tem fortalecido um corpo de pesquisadores locais e o estabelecimento de convênios e intercâmbio com pesquisadores nacionais e internacionais. Seus projetos de pesquisa atendem os requisitos acadêmicos de produção de conhecimento científico e de interface com a sociedade brasileira ao incluir pesquisas para elaboração de políticas de preservação e estudos sobre os impactos causados pelas ações do homem, como, por exemplo, em relação ao uso do solo nesta região.

Dentro do INPA, o LBA (*Large-scale Biosphere-Atmosphere Experiment in Amazonia*)[2], foco deste artigo, é um programa de pesquisa que tem como objetivo estudar e entender o funcionamento do ecossistema da região amazônica de forma interdisciplinar, com um olhar sobre o sistema amazônico como uma entidade regional no sistema Terra e as causas e efeitos das mudanças em curso na região. Entre as premissas desse programa está a consideração dos conceitos inerentes à Gestão de Conhecimento Científico (GCC), visto como um conjunto de processos envolvidos na aquisição, criação, compartilhamento, disseminação e utilização do conhecimento científico, para balizar sua estratégia de apoio aos trabalhos de pesquisa, desde a coleta de dados até a simulação e disponibilização dos resultados de experimentos realizados.

[1] https://www.gov.br/inpa.
[2] http://lba2.inpa.gov.br.

Cada grupo de pesquisa no LBA conta com uma estrutura individualizada de coleta de dados e metadados, adotando protocolos específicos para cada disciplina científica, mas cabe ao INPA a definição de plataformas redundantes para replicação dos dados em estruturas *ad hoc*, adquiridas em função da demanda. São também necessárias políticas de governança de dados, com impactos no ciclo de vida dos dados. Embora o INPA já possua uma infraestrutura de *e-Science* com condições de interagir e atender demandas dos grupos de pesquisa, constata-se que há carências não resolvidas. Um ponto comum das plataformas de *e-Science* é o fato de não serem de fácil manuseio, o que leva ao uso limitado de suas funcionalidades pelos seus usuários. Diante da grande variedade de plataformas de *e-Science* é preciso uma análise das necessidades dos grupos de pesquisa e as características das plataformas existentes, o que demanda um maior escrutínio daqueles que poderão ter acesso aos conteúdos depositados. Apesar dos ganhos providos por essas soluções, até o momento, o Programa LBA/INPA só disponibiliza seus dados de forma restrita e para grupos específicos.

Por outro lado, na última década, a viabilização de ciberinfraestruturas para *e-Science* no Brasil tem como exemplos o Sistema Nacional de Processamento de Alto Desempenho (SINAPAD)[3] e GridUNESP[4], iniciativas que possibilitam a gestão e o tratamento de dados em larga escala proporcionando agilidade e precisão nos resultados de pesquisas científicas. No entanto, mesmo no contexto limitado à gestão de dados, não foram encontrados trabalhos que apontem os requisitos de uma plataforma computacional que possibilite a integração do contexto e do ambiente da pesquisa, a conexão dos dados com os resultados disponibilizados, e o reuso destes dados em outras pesquisas.

O objetivo desse artigo é identificar elementos do LBA/INPA que sejam confluentes com os princípios da GCC a fim de gerar um conjunto de dimensões de análise que os auxiliem na escolha de plataformas atuais e futuras de *e-Science* para a Gestão de Dados Científicos (GDC), entendidos como um pré-requisito da GCC.

2 PROGRAMA DE PESQUISA LBA/INPA

A importância da Amazônia para o planeta é evidenciada por seu impacto em seu ecossistema e seus efeitos no clima, e subsequentes desdobramentos sociais, econômicos e políticos (United Nations, 2021). Pesquisas técnico-científicas contribuem para a criação de conhecimentos coadjuvantes no delineamento e implementação de políticas de preservação ao demonstrarem os impactos causados por intervenções técnico-humanas no uso do solo e da água e seus efeitos no ar. De acordo com informações da Coordenação de Pesquisas do INPA (COPES/INPA)[5], o conhecimento sobre a região por grupos de pesquisa, organizados em 4 grandes focos institucionais: *(i) Biodiversidade*; *(ii) Dinâmica Ambiental*; *(iii) Sociedade, Ambiente e Saúde*; e *(iv) Tecnologia e Inovação*.

Criado em 1952, o INPA[6] "*ao longo dos anos, vem realizando estudos científicos do meio físico e das condições de vida da região amazônica para promover o bem-estar humano e o desenvolvimento socioeconômico regional. Atualmente, o INPA é referência mundial em Biologia Tropical*".

Dentre a miríade de projetos de pesquisa ocorrendo na Amazônia observam-se dois grandes focos de investigação: *(i)* melhoria das condições de vida da população: e *(ii)* geração sustentável de riquezas na região. O LBA, é um programa *(i)* interdisciplinar extenso, *(ii)* criado em 1996 por meio de acordos internacionais de cooperação científica, e *(iii)* tem como

[3] https://www.lncc.br/sinapad.
[4] http://www.unesp.br/portal#!/gridunesp.
[5] http://pesquisa.inpa.gov.br/index.php/coordenacoes-de-pesquisa.
[6] https://dados.gov.br/organization/about/instituto-nacional-de-pesquisas-da-amazonia.

pressuposto o atendimento de objetivos relacionados com a Natureza, a Intervenção Humana e o Relacionamento entre ambos. (EMILIO; LUIZÃO, 2014; AVISSAR et al., 2002; KELLER et al., 2004)

O LBA/INPA tem como pressuposto teórico o fato de que as florestas tropicais de todo o mundo, e especialmente a da Bacia Amazônica, são submetidas a taxas regulares de desmatamento e a bruscas mudanças nos usos da terra. As pesquisas são relacionadas à percepção da Amazônia enquanto biosfera que envolve elementos naturais como solo e clima e seu impacto nas dimensões biológicas, químicas e físicas da região amazônica e do planeta Terra como um todo em uma perspectiva à longo prazo. Como um exemplo, a análise dos dados coletados constatou que os desmatamentos e queimadas aceleram o efeito estufa, alteram o mecanismo da formação de nuvens e podem modificar o regime e a distribuição das chuvas na Amazônia e em outras partes do país ou mesmo do continente. Uma das consequências apontadas pelo estudo foi a crise hídrica do Estado de São Paulo entre 2015 e 2016.

Para alcançar seus objetivos o Programa LBA busca o avanço e a confluência de conhecimentos em sete áreas gerais: *(i)* Física do Clima; *(ii)* Armazenamento e Trocas de Carbono; *(iii)* Biogeoquímica; *(iv)* Química Atmosférica; *(v)* Hidrologia e Química das Águas; *(vi)* Mudanças dos Usos da Terra e da Cobertura Vegetal; e *(vii)* Dimensões Humanas das Mudanças Climáticas. Essas áreas se compõem em três grandes grupos: *(i)* interação biosfera-atmosfera, *(ii)* ciclo hidrológico e *(iii)* dimensões sócio-políticas e econômicas das mudanças ambientais. Por sua vez, as pesquisas em curso foram concentradas em três focos integradores: *(i)* o ambiente amazônico em mudança; *(ii)* a sustentabilidade dos serviços ambientais e sistemas de produção terrestres e aquáticos; *(iii)* variabilidade das mudanças climáticas e hidrológicas – retroalimentação, adaptação e mitigação. Essa distribuição em áreas de pesquisa mostra a relevância de trabalhos colaborativos, de compartilhamento de dados e evidencia os desafios de integração social aos resultados dos estudos e ao desenvolvimento regional sustentável, com a consequente geração de um grande volume de dados e a necessidade de infraestrutura tecnológica para a gestão desses dados.

Diante da dimensão e importância do Programa LBA, no INPA, sua infraestrutura tecnológica e computacional necessita de planejamento e investimentos que garantam a governança dos dados produzidos e suporte à rede de cientistas que atuam nos diversos projetos de pesquisa vinculados. Desta forma, a estrutura de gerenciamento do Programa está organizada em duas gerências: científica e operacional. Ambas mantêm comunicação e compartilham os planos de execução, acompanhando cada projeto de pesquisa em andamento, os dados coletados, suas análises, publicação de resultados e armazenamento para disseminação em larga escala. Os dados são classificados como dados operacionais e científicos que são armazenados, para consumo posterior por aplicativos e sistemas de gestão do Programa ou aplicativos de domínio específico, conforme as especificidades e finalidade de utilização dos dados primários produzidos.

Para garantir a efetiva gestão e organização dos dados, cada pesquisa do Programa adota protocolos específicos para cada disciplina científica. Os cientistas detêm de uma infraestrutura flexível, desde que os dados produzidos possuem uma cópia na estrutura do Programa. Isso garante a perenidade dos dados e metadados com qualidade, bem como os devidos prazos de retenção e publicação dos mesmos e as ações que envolvem a aquisição, coleta, qualidade, garantia, uso e reuso dos dados são feitos conforme o modo de trabalho de cada pesquisador.

Devido a alguns experimentos que produzem elevada volumetria de dados, já na ordem de *Petabytes*, tanto a infraestrutura em nível de projeto quanto do LBA/INPA como um todo, precisam constantemente alocar recursos financeiros para investimentos em atualizações e acréscimos da própria infraestrutura de hardware e software. Por esse motivo,

embora o INPA possua uma infraestrutura já disponível, é uma necessidade premente de revisar os requisitos e elencar cenários possíveis, eficientes e financeiramente viáveis para gestão dos dados produzidos. Há ainda, paralelamente a gestão dos dados científicos, dados operacionais que precisam ser gerenciados para garantir o controle da logística e a total atendimento das demandas dos usuários dos sistemas *online*.

Em relação aos pesquisadores do Programa, esses atuam em diferentes grupos e com diferentes funções ao mesmo tempo. Isso evidencia a existência de uma grande rede de relacionamento e colaboração científica nesses grupos. Essa característica do Programa é importante, pois é composto por um corpo de cientistas de diversas instituições e países. Desta forma, a governança de dados tem papel essencial para as instituições e no processo de tomada de decisões. Os requisitos de um sistema dessa natureza consideram as composições dos grupos e subgrupos, o gerenciamento dos grupos, de repositórios de arquivos, com configuração de nível de acessos que refletem as políticas de dados internas dos grupos de modo a não conflitarem com as institucionais.

No ano de 2012, o INPA iniciou a implantação de um novo sistema para GDC. Consiste em um sistema de informação modular que visa contemplar as necessidades dos diferentes tipos de usuários, sejam esses acadêmicos ou gestores ambientais. Sua estrutura conceitual é baseada na ideia de transformar informação em conhecimento, utilizando a melhor tecnologia disponível. O sistema de gestão usa bancos de dados relacionais para armazenar e analisar todas as entradas de informações. Os dados brutos são documentados por meio de um sistema de metadados, compatível com diversas especificações, como o *Ecological Metadata Language*[7] e o padrão ISO-19115 (INTERNATIONAL ORGANIZATION FOR STANDARDIZATION, 2014). O núcleo do sistema é composto por um repositório de modelos (chamado *ModeleR*) que é capaz de documentar e executar diferentes tipos de modelos e procedimentos analíticos. Além disso, há sistemas que conseguem processar grandes quantidades de dados automaticamente provenientes desse repositório.

Mais recentemente o INPA identificou funcionalidades básicas de interesse de projetos científicos que envolveram armazenamento e compartilhamento de dados heterogêneos, de diversos volumes numa infraestrutura tecnológica de alto desempenho que faz parte do SINAPAD. Mais especificamente, a infraestrutura computacional desenhada para atendimento às pesquisas contempla um Centro Nacional de Processamento de Alto Desempenho (CENAPAD) e uma nuvem privada que é utilizada para processamento de dados oriundos dos diversos projetos desenvolvidos pelo INPA e instituições parceiras. Essa infraestrutura dá acesso aos pesquisadores a diversos sistemas através de tecnologia *Single-Sign-On* – SSO (SOUZA, 2017), que permite a autenticação centralizada.

Além disso, a infraestrutura disponibilizada tem sido útil para o LBA na medida em que permite operar com grandes quantidades de dados geradas por estações meteorológicas, torres de fluxo, ou outros dados atmosféricos coletados instrumentalmente. De fato, toda a informação processada é visualizada em um portal *Web* por meio de gráficos dinâmicos e a descrição dos dados é feita por meio de ferramentas de análise *Online Analytical Processing* (OLAP), que auxiliam na criação de cubos multidimensionais.

3 GESTÃO DO CONHECIMENTO CIENTÍFICO

Para Köche (2011, p. 29):

> O conhecimento científico surge da necessidade de o homem não assumir uma posição meramente passiva, de testemunha dos fenômenos, sem poder de ação ou controle dos mesmos. Cabe ao homem, otimizando o uso

[7] https://eml.ecoinformatics.org

> da sua racionalidade, propor uma forma sistemática, metódica e crítica da sua função de desvelar o mundo, compreendê-lo, explicá-lo e dominá-lo.

Este conhecimento, que permite a humanidade encontrar soluções para seus problemas cada vez mais complexos, há algum tempo passa por transformações na sua concepção e divulgação, influenciadas principalmente pelas TICs, que possibilitam tanto uma escala de dados primários cada vez maior quanto um aumento exponencial da produção científica.

> O conhecimento científico é produzido pela investigação científica, através de seus métodos. Resultante do aprimoramento do senso comum, o conhecimento científico tem sua origem nos seus procedimentos de verificação baseados na metodologia científica. É um conhecimento objetivo, metódico, passível de demonstração e comprovação. O método científico permite a elaboração conceitual da realidade que se deseja verdadeira e impessoal, passível de ser submetida a testes de falseabilidade. Contudo, o conhecimento científico apresenta um caráter provisório, uma vez que pode ser continuamente testado, enriquecido e reformulado. Para que tal possa acontecer, deve ser de domínio público (FONSECA, 2002, p. 11).

O conhecimento resultante da pesquisa científica parte do princípio de que não há verdade absoluta. Creswell e Creswell (2021) defendem essa característica ao descrever as concepções filosóficas do conhecimento científico dentro suposições pós-positivistas, que têm governado as alegações sobre o que garante o conhecimento. Os processos da GC, amparados pelo aperfeiçoamento de técnicas e avanços de ferramentas de TI, podem ter um papel de destaque na produção de novos conhecimentos, haja vista a necessidade de reavaliação constante dos resultados da pesquisa científica, submetendo os dados a um ciclo evolutivo de testes de forma longitudinal.

A evidente passagem da escassez para uma grande quantidade de conhecimento produzido em um pequeno intervalo de tempo implicou na necessidade de métricas e processos para gerenciá-lo. As práticas de GC, bem definidas e aceitas para gerir o conhecimento organizacional, só recentemente começam a ser aplicadas em ambientes de produção de conhecimento científico (SANTOS; MENEZES, 2019; BAKER; MAYERNIK, 2020; SAMPAIO, 2007). Trabalhos como os de Martinelli et al. (2017) corroboram os achados de Leite e Costa (2007), que indicam serem poucas as iniciativas efetivas, os estudos ou os modelos de GCC no meio acadêmico. Santos e Santos (2015), no entanto, se contrapõem a esse juízo e afirmam que existem diversas práticas de GC adotadas no ambiente de pesquisa, porém, observam que se trata de iniciativas pessoais por parte dos pesquisadores, sem qualquer tipo de coordenação ou gestão com finalidade específica.

Leite e Costa (2007) contribuíram para essa temática ao propor um modelo conceitual para a GCC no contexto acadêmico, com ênfase na comunicação científica. De acordo com estes autores, ao propor estudar e explorar um novo assunto, cientistas recorrem aos trabalhos correlatos por meio da pesquisa bibliográfica de seus pares, disponibilizados por meio da comunicação científica. Contudo, o conhecimento científico não se limita àquele publicado em periódicos e eventos científicos; há ainda o conhecimento tácito que o pesquisador acumula com a experiência e carrega consigo. Segundo Collins (2001, p. 71):

> [...] o conhecimento científico tácito [...] refere-se ao que pode ser entendido como o conhecimento ou habilidade que pode ser passada entre cientistas por contatos pessoais, mas não pode ser exposto ou passado em fórmulas, diagramas, descrições verbais ou instruções para ação.

A distinção entre conhecimento tácito e conhecimento explícito é dada por Ryle (1949), que diferencia "saber o que" de "saber como", e Polanyi (1966), que introduziu a ideia de conhecimento tácito.

Na produção do conhecimento científico, "existe um acoplamento entre rigidez e liberdade na produção científica, e cada assunto de pesquisa segue uma dialética entre razão e intuição do pesquisador" (AMARAL, 2017).

As redes de colaboração possuem relevante contribuição na difusão de conhecimento tácito, pois permitem sua externalização e, posteriormente, sua conversão em conhecimento explícito (OLIVEIRA, 2018). Segundo Chin Jr et al. (2002):

> [...] o projeto e execução de atividades científicas são altamente repetitivos e cientistas geralmente mantêm o modelo projeto e informações sobre a execução dos processos científicos em suas mentes e em notas não oficiais, como cadernos e papéis avulsos. Com isto, surge a necessidade de - além de representar o dado, a informação e o conhecimento - representar como foi o processo de criação deles, bem como as competências necessárias para a sua criação.

Evidencia-se, portanto, que no âmbito acadêmico há também os conhecimentos tácito e explícito dentro da dimensão epistemológica e, portanto, os modelos de GC definidos por diversos autores clássicos (NONAKA; TUKEUCHI, 1997; STEWART, 1998; TERRA, 2000) podem ter aplicabilidade na GCC.

A GCC ainda não possui modelos definidos, que levem em consideração a heterogeneidade do ciclo de produção do conhecimento científico desde a sua concepção até a comunicação. Em alguns dos trabalhos existentes, consideram-se preceitos da GC organizacional. Para Ferreira (2010, p. 17):

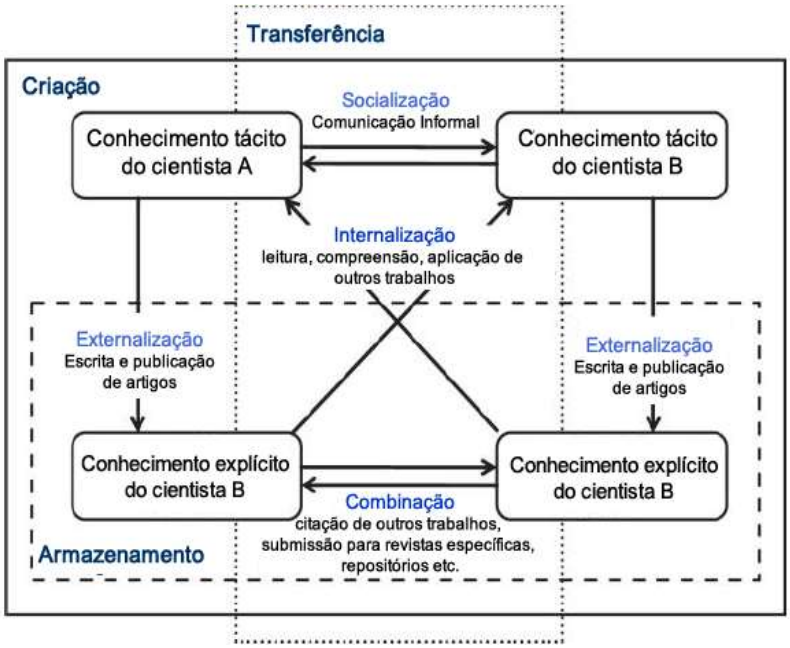
> A GCC deve promover o estabelecimento de condições que possibilitem a esses atores a criação e o compartilhamento do conhecimento científico, proporcionando meios para que os cientistas possam se comunicar, colaborar, e uma interação entre diversas disciplinas na construção de novos conhecimentos. A colaboração impulsiona a criação de novos conhecimentos, uma vez que promove a interação entre os indivíduos e o compartilhamento do conhecimento existente, contribuindo para uma maior rapidez no desenvolvimento dos trabalhos científicos.

No âmbito acadêmico, a socialização do conhecimento é viabilizada por meio da interação possibilitada pelas redes de colaboração. Neste contexto, as plataformas de *e-Science* têm papel relevante, possibilitando a criação de novos conhecimentos, de natureza complexa e interdisciplinar, por meio da troca de experiências.

Bernius (2010) serve-se do modelo SECI (Socialização, Externalização, Combinação, Internalização), proposto por Nonaka e Tukeuchi (1997), para propor um processo de produção do conhecimento científico e propõe um *framework* de GC em processos de pesquisa científica ilustrado na Figura 1. Afirma ele:

> O modelo envolve dois cientistas para exemplificar o uso do *framework* e o processo de conversão do conhecimento. Para cada cientista, o conhecimento tácito que pode ser explicitado e o conhecimento já explicitado pelo próprio cientista é compreendido. As bordas entre os quatro tipos de conhecimento resultantes (conhecimento tácito e explícito do cientista A, conhecimento tácito e explícito do cientista B) representam os diferentes modos de conversão.

Figura 1 – *Framework* de gestão do conhecimento em processos de pesquisa científica



Fonte: Bernius (2010)

Bernius (2010) explica ainda que a escrita e publicação são os processos que os cientistas utilizam para explicitar e, em uma segunda etapa, disseminar os resultados da pesquisa. Se a informação descrita no artigo não tiver já sido abordada em estudos anteriores, pode-se denotar como novo conhecimento explícito. Portanto, do ponto de vista de um cientista, a externalização compreende principalmente os processos de escrita e publicação de um artigo. Em contraposição, a socialização descreve atividades de transferência de conhecimento tácito (ainda não explicado) entre cientistas. Essa comunicação informal, por exemplo, ocorre no grupo de pesquisa com troca de experiências. A última etapa é a internalização do conhecimento científico criado, que na ciência implica em ler e compreender o trabalho de outros colegas cientistas. Ainda para Bernius (2010), "Se o cientista A lê e entende o trabalho do cientista B, esse entendimento é convertido em conhecimento tácito". O quarto modo de conversão é uma combinação. O conhecimento explícito combinado de diferentes cientistas normalmente pode ser encontrado no nível do diário. Por exemplo, a compilação de um problema específico publicado em uma revista, ou de um repositório, representa uma combinação de conhecimento explícito criado por vários pesquisadores. A citação pode ser vista como um processo que suporta o modo de combinação.

O *framework* enfatiza a comunicação entre cientistas, caracterizando, de acordo com Newman (2001), uma rede de pesquisa, uma das formas de possibilitar o trabalho em colaboração. Neste contexto, a *e-Science* possui papel relevante oferecendo importantes contribuições e aparato tecnológico para a GCC.

4 GESTÃO DE DADOS CIENTÍFICOS

Dados, no contexto da Ciência da Informação, são, segundo Davenport e Prusak (2003, p. 6), fatos objetivos e distintos relativos a um fenômeno. Os mesmos autores definem informação como uma mensagem com dados para influenciar a opinião do receptor. "Conhecimento é a experiência condensada, informação contextualizada, *insight*, crenças e valores" (DAVENPORT; PRUSAK, 2003, p. 6).

No que concerne à pesquisa científica contemporânea, os dados apresentam-se em escalas cada vez maiores, levando a necessidade de infraestruturas tecnológicas para seu gerenciamento. Para Mattoso at. al. (2008) "produzir conhecimento científico atualmente implica, dentre outras características, ubiquidade e distribuição, visando ao desenvolvimento e execução de soluções com alto desempenho, baseadas em reutilização, gerência de dados e experimentos".

Sayão e Sales (2015) descrevem a importância dos dados científicos na tomada de decisões, sendo capazes de oferecer respostas com um alto índice de confiabilidade. Diante da complexidade dos problemas, houve também uma mudança sobre o aspecto temporal da análise dos dados. Sayão e Sales (2015, p. 5) afirmam ainda:

> [...] os dados eram considerados somente na sua configuração final, sem considerar os seus ciclos de vida, versões e linhagens e contexto, eram descartados ou armazenados em mídias ou em servidores sem a devida gestão quando os projetos eram concluídos. Quase sempre eram tragados silenciosamente pelo tempo: pela obsolescência tecnológica e pela fragilidade das mídias digitais.

GDC tornou-se primordial para os avanços da ciência, permitindo que o mesmo conjunto de dados primários possa ser analisado sob outras perspectivas e principalmente de forma interdisciplinar, com o apoio de redes colaborativas com a participação de diversos cientistas. Partindo do princípio de que o principal insumo (e muitas vezes produto) de uma instituição científica são os dados, a correta utilização e preservação para novas análises, com técnicas e metodologias modernas, torna-se essencial.

Dados de pesquisa são classificados quanto à (SAYÃO; SALES, 2015, p. 7-9): *(i)* origem: dados observacionais, dados computacionais e dados experimentais; *(ii)* natureza: números, imagens, vídeos ou áudios, softwares, algoritmos, equações, animações ou modelos e simulações e *(iii)* fase da pesquisa: dados brutos ou preliminares (*raw data*), dados derivados e dados canônicos ou dados referenciais.

Dados observacionais são dados obtidos por meio da observação direta e relacionam-se com tempos e espaços específicos (p.ex., a erupção de um vulcão). Por estar ligados ao tempo, é impossível coletar o mesmo tipo de amostra mais de uma vez. Dados computacionais são extraídos de modelos ou artefatos computacionais ou simulações. Por serem originados em um ambiente controlado, esses dados são fáceis de serem reproduzidos ou repetidos. Dados experimentais são dados provenientes de bancada controlada de laboratórios (p.ex., uma reação química) e podem ser reproduzidos com exatidão.

5 CURADORIA DIGITAL

A Curadoria Digital é uma área interdisciplinar que congrega conhecimentos e necessidades de diversas áreas do conhecimento, porém possui suas definições sustentadas na Ciência da Computação e na Ciência da Informação. De acordo com Dutra e Macedo (2016), "[...] a ideia de curadoria nos remete quase que imediatamente ao termo utilizado de maneira mais tradicional por museus e bibliotecas, especialmente em relação a coleções de artefatos físicos." A definição cunhada por Abbot (2008) traz a convergência com a TI:

> [Curadoria Digital é] o conjunto das todas as atividades existentes no gerenciamento de dados, desde o planejamento da sua criação, passando pela digitalização (transformação digital) ou criação, procurando assegurar a disponibilidade e adequação para a recuperação e reuso futuro destes dados.

A Curadoria Digital surgiu como uma nova prática interdisciplinar que visa estabelecer diretrizes para a gestão disciplinada de informação (HIGGINS, 2008). Diversos autores definem ainda como etapa importante na Curadoria Digital, a agregação de valores em objetos digitais ao longo da sua preservação (LEE; TIBBO, 2007; ABBOTT, 2008; SAYÃO; SALLES, 2012). Não obstante, Abbott (2008) descreve as etapas da Curadoria Digital como

> [...] gestão de dados, desde o planejamento da sua criação, passando pelas boas práticas na digitação, na seleção dos formatos e na documentação, e na garantia dele estar disponível e adequado para ser descoberto e reusado no futuro.

O Digital Curation Centre (DCC)[8], iniciativa que propôs métricas consideradas referências em preservação digital, advoga que Curadoria Digital "envolve a manutenção, a preservação e a agregação de valor a dados de pesquisa durante o seu ciclo de vida". As contribuições do DCC incluem um modelo que assegura contemplar todas as etapas do processo, com a finalidade de obter sucesso na preservação digital e, consequentemente, êxito em seus objetivos, que permeiam principalmente a disponibilidade e reuso de dados científicos. Sua revista eletrônica *International Journal of Digital Curation* (IJDC)[9] destina-se exclusivamente à publicação de artigos científicos relacionados à Curadoria Digital.

Sayão e Salles (2012) confirmam em seus estudos que o modelo proposto pelo DCC está orientado para o planejamento das atividades de curadoria nas organizações ou consórcios ajudando a garantir que todos os passos do ciclo serão cumpridos.

O modelo de curadoria do DCC engloba todas as fases da preservação de objetos digitais, incluindo ações para cada um dos elementos chave desse processo: dados, objetos digitais e bases de dados. Essas ações são classificadas quanto à sua ocorrência dentro do processo de preservação, classificando em ações sequenciais e ações ocasionais, englobando as seguintes etapas:

1. **Conceitualização**: conceber e planejar a criação de objetos digitais, incluindo métodos de captura de dados e opções de armazenamento.
2. **Criação**: produzir objetos digitais e atribuir metadados arquivísticos administrativos, descritivos, estruturais e técnicos.
3. **Acesso e uso e reuso**: assegurar que os usuários possam acessar facilmente objetos digitais no dia-a-dia. Alguns objetos digitais podem estar disponíveis publicamente, enquanto outros podem ter o acesso restrito.
4. **Avaliação e seleção**: avaliar os objetos digitais e selecionar aqueles que necessitam de curadoria e preservação de longo prazo. Aderir a orientações documentadas, políticas e requisitos legais.
5. **Eliminação**: eliminar objetos digitais não selecionados para preservação e conservação a longo prazo. Orientação documentada, políticas e requisitos legais podem exigir a destruição segura desses objetos.
6. **Ingestão/utilização**: transferir objetos digitais para um arquivo, repositório digital confiável, data center ou similar, aderindo novamente a orientações documentadas, políticas e requisitos legais.
7. **Ação de preservação**: realizar ações para assegurar a preservação e retenção a longo prazo da natureza autorizada dos objetos digitais.
8. **Reavaliação**: devolver objetos digitais que falham nos procedimentos de validação para posterior avaliação e seleção.

[8] http://www.dcc.ac.uk.
[9] http://www.ijdc.net.

9. **Armazenamento**: manter os dados de forma segura, conforme descrito em padrões relevantes.
10. **Transformação**: criar novos objetos digitais do original, por exemplo, pela migração para uma forma diferente.

Sayão e Salles (2012) destacam dentre as atividades da preservação digital, o reuso de dados disponibilizados de forma aberta por outros cientistas, podendo até ser replicados, gerando novos resultados de pesquisas.

6 ABORDAGEM METODOLÓGICA

A presente pesquisa seguiu as seguintes etapas: *(i)* revisão da literatura; *(ii)* visita *in loco*; *(iii)* entrevistas semiestruturadas; *(iv)* processamento e análise das informações obtidas nas entrevistas e *(v)* análise e discussão dos resultados. Com a revisão da literatura buscou-se identificar: *(i)* o estado da arte das práticas e metodologias no contexto da GDC e GCC e *(ii)* as características das plataformas tecnológicas para GDC e GCC. Os critérios de seleção dos artigos para o embasamento teórico desta pesquisa levaram em consideração o atributo temporal da publicação, fazendo um recorte dos trabalhos publicados nos últimos quinze anos, classificados por sua relevância (aferidos pelas ferramentas de buscas dos portais das bases de dados científicas consultadas), considerando o número de citações. As bases utilizadas foram *IEEExplorer, Web Of Science, Springer, Science Direct*, *SciELO e* BDTD. As buscas conjugaram os termos *"pesquisa científica + interdisciplinaridade"* e *"arquiteturas de e-Science para Gestão de Dados Científicos"* e *"Curadoria Digital"*, pesquisados nos idiomas português e inglês.

A visita *in loco* na sede do INPA buscou observar o aparato de hardware e software disponível para a comunidade científica do Programa LBA/INPA e a percepção dos pesquisadores com relação a sua disponibilidade e efetividade para GDC, desde a sua concepção até a publicação dos resultados. A técnica de observação adotada seguiu o que preconiza Silva e Fossá (2015), ao se buscar informações utilizando os sentidos no processo de alcançar aspectos da realidade que, à primeira vista, poderiam ser incompreensíveis.

A partir do referencial teórico considerado, foram identificados quatro eixos temáticos que se configuraram como fundamentais para as práticas de *e-Science*: *(i)* trabalhos colaborativos em grupos de pesquisa; *(ii)* governança, gestão e política de dados; *(iii)* infraestrutura de TIC; *(iv)* desenvolvimento da pesquisa e publicação dos resultados. Essa organização temática do estudo, em conjunto com a pré-análise documental da estrutura e organização do LBA/INPA, realizada por meio de consultas em sítios, publicações e repositório institucional, auxiliaram na elaboração das perguntas e dos grupos para aplicação dos questionários nas entrevistas. Com a análise desses dados, foi possível conhecer a estrutura organizacional do objeto de estudo, grupos de pesquisas, infraestrutura tecnológica, projetos desenvolvidos dentro das áreas de estudo ligadas ao Programa e como tem se dado a evolução e desafios enfrentados pela comunidade científica (pesquisadores, corpo técnico e dirigentes) envolvidas LBA/INPA.

As entrevistas semiestruturadas tiveram como objetivo registrar a percepção, por meio de perguntas subjetivas, das necessidades da comunidade científica do LBA/INPA com relação à infraestrutura de hardware e software disponível, além de políticas institucionais para gestão e governança de dados científicos, e identificar os desafios dos dirigentes e da equipe técnica quanto à implantação e gestão dos recursos informáticos para infraestrutura de *e-Science*. O objetivo das entrevistas foi identificar as reais demandas por infraestrutura tecnológica para prover GDC e GCC, de forma a identificar um conjunto de dimensões conceituais inerentes e necessárias em plataformas computacionais para *e-Science* que possam apoiar o Programa. Buscou-se também identificar cenários possíveis, com ênfase na eficiência, economicidade e perenidade, para auxiliar no desenvolvimento da ciência no

contexto da região Amazônica.

A formação dos grupos de entrevistados levou em consideração o tipo de atuação no Programa LBA/INPA. O Grupo 1 foi formado por quatro pesquisadores e o Grupo 2 por dois técnicos e dois dirigentes envolvidos direta ou indiretamente com a implementação de ferramentas de tecnologia da informação. Antes das entrevistas, foram apresentados seus objetivos e os conceitos de *e-Science*, *Cloud Computing* (computação em nuvem) e *Big Data* (dados em larga escala). Para o Grupo 1, as entrevistas seguiram o seguinte roteiro:

1. Quais disciplinas subsidiam seus projetos de pesquisas dentro do LBA e como eles são originados?
2. A infraestrutura tecnológica atende às suas demandas e expectativas, e consequentemente contribui para a execução de suas pesquisas? Há limitações?
3. Atualmente, devido à complexidade e à natureza interdisciplinar dos problemas da sociedade contemporânea, as pesquisas estão exigindo um esforço conjunto de diversos cientistas, ensejando em trabalho colaborativo envolvendo diversas áreas, a infraestrutura tecnológica do LBA permite o trabalho colaborativo, além de subsidiar outras atividades relacionadas, como comunicação, *feedback*, compartilhamento e discussões entre os cientistas envolvidos no projeto?
4. A curadoria digital garante a preservação dos para reuso dos dados de pesquisa em estudos futuros, podendo produzir novos resultados. A infraestrutura tecnológica do LBA permite a preservação de dados em longo prazo? Caso afirmativo, qual o modelo adotado?

enquanto para o Grupo 2, o roteiro foi:

1. O LBA tem abarcado uma robusta infraestrutura tecnológica, com relevantes investimentos para prover o armazenamento, processamento, compartilhamento e segurança de dados científicos primários. A arquitetura computacional criada é suficiente para atender as demandas do projeto a curto, médio e longo prazo? Como?
2. A computação em nuvem tem mostrado ser eficiente para solução de problemas que envolvem *Big Data*, oferecendo principalmente serviços de armazenamento, processamento e segurança de forma elástica e escalável. A nuvem pode ser pública, privada ou híbrida. Soluções de nuvem podem auxiliar na gestão de dados do LBA? Caso exista, quais seriam os fatores limitantes em cada um dos tipos citados?
3. Com base em informações da Diretoria Tecnológica do LBA, recentemente o programa optou por utilizar o sistema LINARIA para gestão dos dados científicos. Isso evidencia a tendência do LBA/INPA prover soluções de nuvem privada para solução dos problemas do programa. A tendência em implementar nuvem privada, e ainda considerando o grande volume de dados gerados por diversas fontes, não elevará os custos de manutenção do programa a longo prazo? Existe um plano para isso?
4. Nos últimos anos têm surgido diversas arquiteturas de *e-Science* (termo utilizado para denotar ciência com uso de Computação de Alto Desempenho) globais, estas são disponibilizadas para a comunidade científica, dentre elas: SINAPAD, E-Infrastructure Shared Between Europe and Latin America (EELA)[10], Rede Galileu entre outras, além de *frameworks* que utilizam recursos da Web semântica que auxiliam na composição semântica do conhecimento científico, com o uso de ontologias. O LBA faz uso de algumas destas arquiteturas? Existe alguma limitação?
5. O que é essencial em uma solução de nuvem, para atender integralmente todas as demandas do LBA e da comunidade científica global que trabalha no projeto?

As entrevistas tiveram duração média aproximada de 25 minutos. Foram gravadas e,

[10] https://www-eela.ceta-ciemat.es.

posteriormente, transcritas para análise. As entrevistas foram organizadas em um *corpus* não estruturado e segmentado (um segmento para cada entrevistado). O processamento das entrevistas foi realizado de forma automatizada com a utilização do software Iramuteq[11], com o qual foram utilizadas seguintes técnicas: *(i)* Classificação Hierárquica Descendente (CHD) (REINERT, 1990) para a classificação dos segmentos de texto em função dos seus respectivos vocabulários; *(ii)* Análise Fatorial de Correspondências (AFC), para a representação em um plano cartesiano do conjunto de palavras e variáveis associadas a cada uma das classes da CHD (HAIR JUNIOR et al, 1998), *(iii)* Análise de Similitude (MARCHAND; RATINAUD, 2012), para a representação na forma de um grafo das conexões entre as palavras; e *(iv)* nuvem de palavras (HEIMERL et al., 2014), para o agrupamento e organização gráfica das palavras em função da sua frequência.

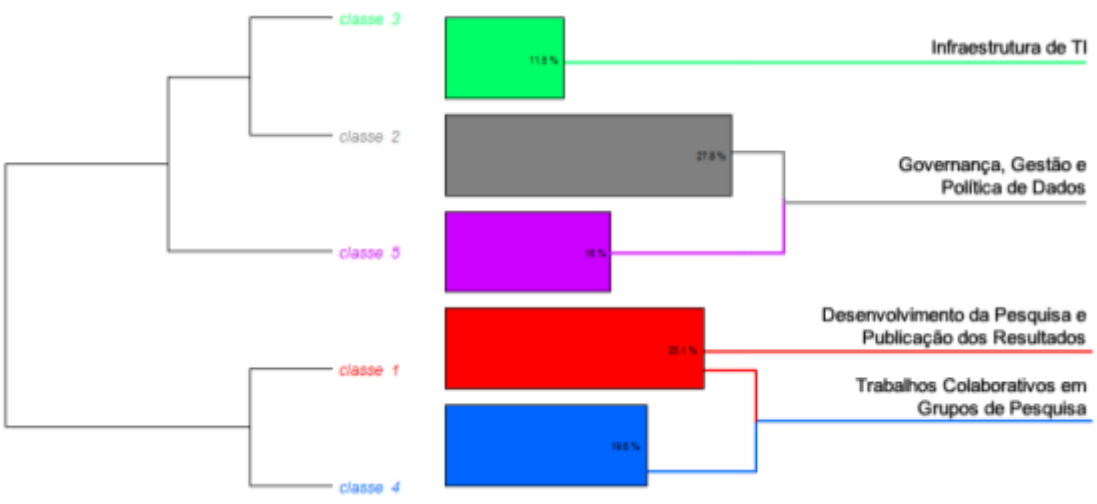
7 RESULTADOS

Os resultados deste trabalho são apresentados a seguir na seguinte sequência: (*i)* processamento e análise dos textos das entrevistas*; (ii)* descrição das dimensões conceituais pertinentes às necessidades do Programa LBA/INPA; e *(iii)* apresentação das necessidades relacionadas à GDC e GCC no Programa LBA/INPA.

7.1. Processamento das entrevistas

A CHD evidenciou uma proximidade das formas lematizadas com os eixos temáticos sugeridos a partir do referencial teórico – *trabalhos colaborativos em grupos de pesquisa; governança, gestão e política de dados; infraestrutura de TIC; desenvolvimento da pesquisa e publicação dos resultados* – classificando o *corpus* em cinco *clusters*. O dendrograma mostra a classificação e a frequência dos termos lematizados e agrupados em cada classe, e a relação entre elas (Figura 2).

Figura 2 – Dendrograma



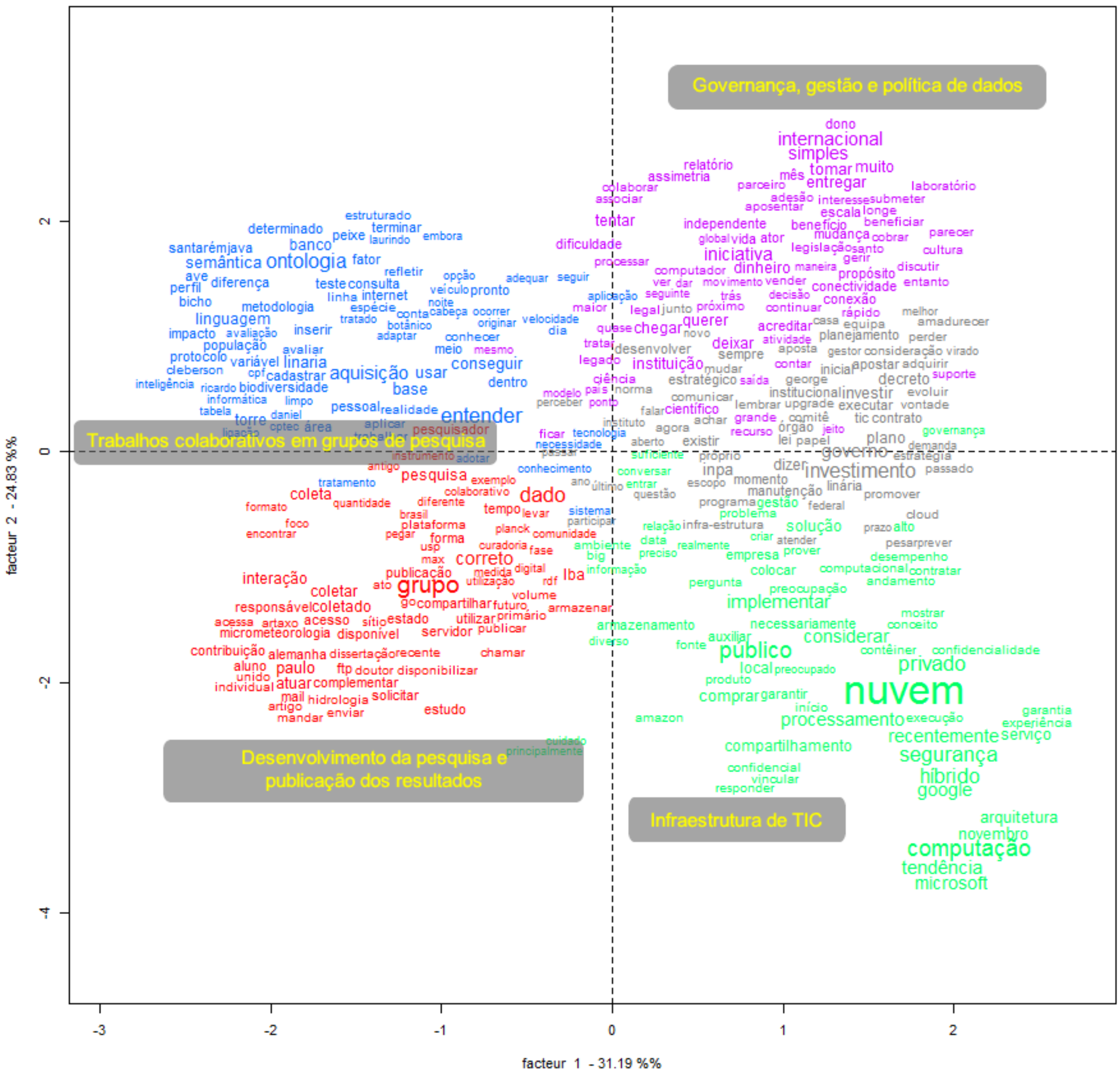
Fonte: Gerado pelo Iramuteq com edição dos autores

A análise pós-fatorial (Figura 3) mostra os termos agrupados em um plano cartesiano e a associação com os eixos temáticos, as duas técnicas são complementares. Por esse motivo, o padrão de cores foi mantido, mostrando o conteúdo extraído de cada classe e associado aos

[11] http://www.iramuteq.org.

quadrantes.

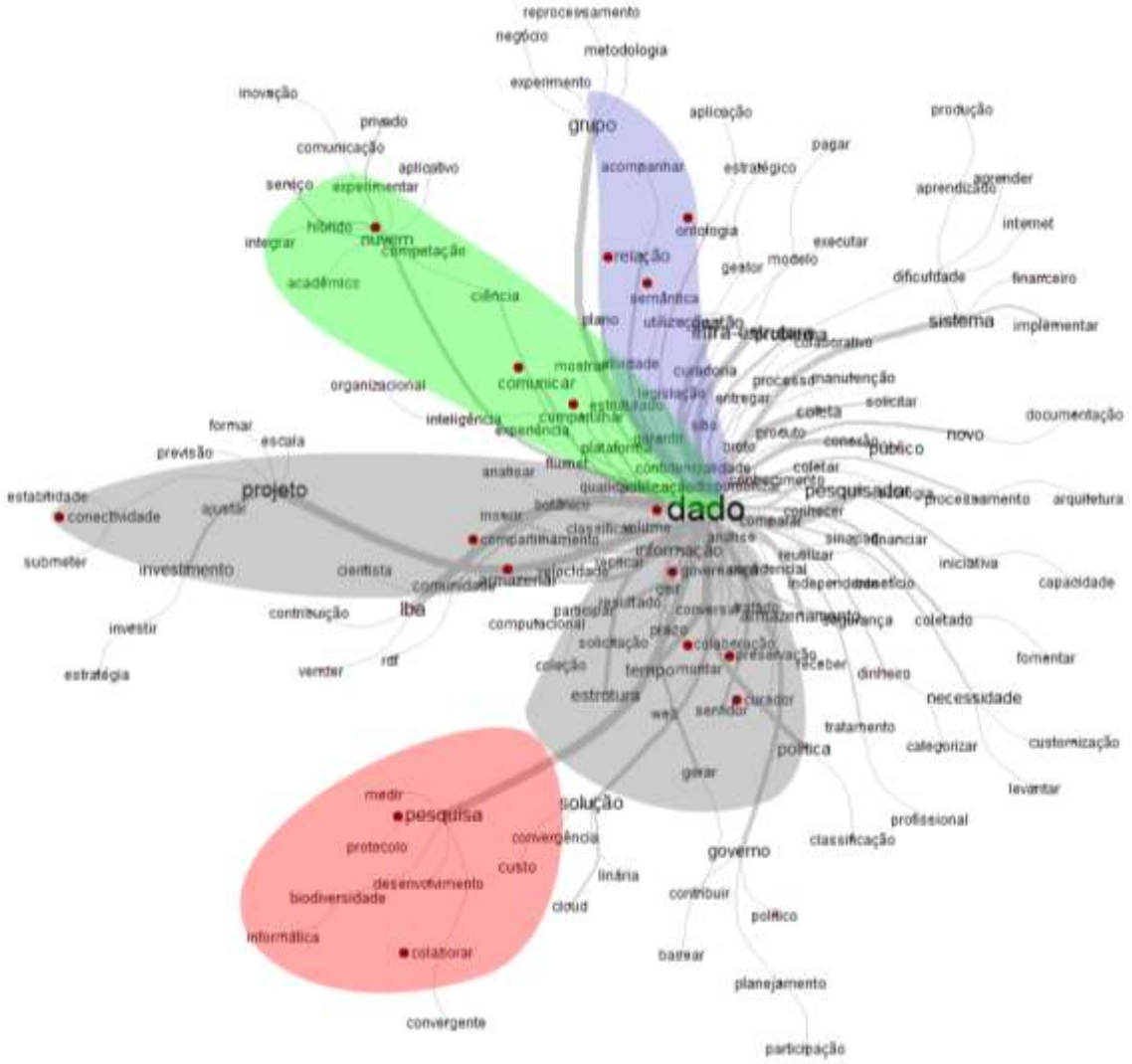
Figura 3 – Análise Pós-Fatorial



Fonte: Gerado pelo Iramuteq com edição dos autores

A análise de similitude mostra que o termo dado é o nó principal ligando as palavras identificadas nos segmentos de texto, conforme mostra o grafo ilustrado na Figura 4. Desta forma, a gestão dos dados, como afirma Abbott (2008), possibilita a interação entre os grupos de pesquisas envolvido. As áreas coloridas foram inseridas posteriormente para realçar as ligações entre palavras consideradas relevantes.

Figura 4 – Grafo do resultado da análise de similitude aplicada ao *corpus*



Fonte: Gerado com Iramuteq com edição dos autores

Os pontos coloridos reforçam as ligações de palavras que indicaram a seleção das dimensões conceituais. Os pontos destacados com as cores verde e cinza estão relacionados à análise que originaram as dimensões *armazenamento de dados; segurança, conectividade; gestão, governança e política de dados; curadoria digital e compartilhamento e replicação de dados,* associados respectivamente aos eixos que remetem a *Infraestrutura de TI e Governança, Gestão e Política de Dados*; os pontos destacados em cor vermelha descrevem as coocorrências do eixo *Desenvolvimento da Pesquisa e Publicação dos Resultados* contemplando as dimensões *colaboração científica* e *interdisciplinaridade*; já a cor lilás representa as conexões alusivas à *relação semântica*, relacionada ao eixo *Trabalhos Colaborativos em Grupos de Pesquisa*. Conforme ilustra a Figura 3, a análise (CHD) aplicada sob o *corpus* sugere ligações que não se restringem aos limites dos quadrantes do plano cartesiano, podendo uma dimensão estar presente em mais de um eixo temático, conforme os critérios de análise, organização e critérios deste trabalho.

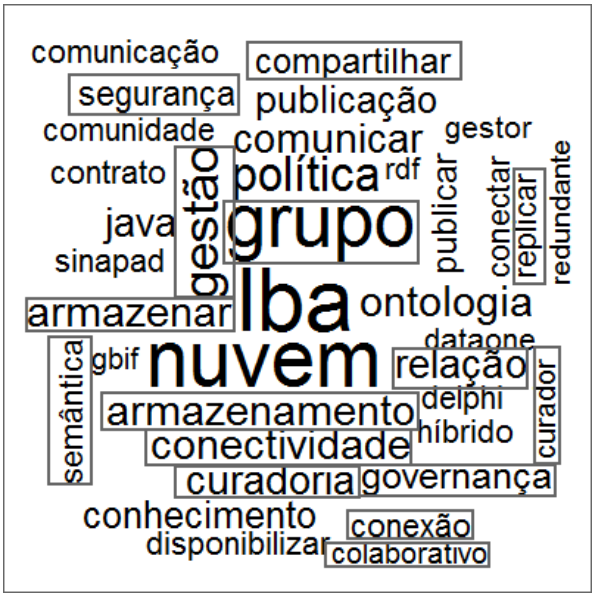
Em determinados casos uma palavra está relacionada a mais de uma outra palavra. Por exemplo, o termo *governança* está ligado à *colaboração* e a *pesquisa*. Isso reflete a existência

de colaboração entre grupos de pesquisa relacionados a diversas disciplinas científicas abrangidas pelo Programa, que coletam dados primários em diferentes projetos de pesquisa. Neste aspecto, os resultados vão ao encontro da observação de Oliveira (2018), para quem "pesquisas compartilhadas em diferentes grupos enriquecem os resultados de várias maneiras". Os cientistas entrevistados apontaram a necessidade de estabelecer uma política de gestão e governança de dados para otimização de recursos em relação a aquisição e utilização destes dados, possibilitando a reutilização e análises diferentes sob o mesmo *dataset*, além da garantia (*data assurance*), já que quando o cientista não estiver mais atuando no Projeto ou na instituição, os dados continuarão disponíveis para outros pesquisadores. Estas relações são comprovadas com a conexões entre os termos *colaboração, preservação* e *curadoria*.

Além da análise de similitude aplicada ao *corpus*, o agrupamento considerado para o apontamento da dimensão conceitual correspondente levou em consideração o conhecimento tácito dos cientistas, apontado Barnius (2010) em seu *Framework* de gestão do conhecimento em processos de pesquisa científica (Figura 1). Este conhecimento foi aferido na visita *in loco* e entrevistas.

A nuvem de palavras (Figura 5), gerada com base nas formas ativas e complementares presentes no *corpus,* destacou que a conexão entre palavras apontada no grafo de similitude (Figura 4) possui um padrão de frequência semelhante. Para a construção da nuvem, a palavra *dado*, nó central do grafo, foi retirado para não ofuscar as demais palavras.

Figura 5 – Nuvem de palavras



Fonte: Criado com o Iramuteq com edição dos autores

7.2 Identificação do conjunto de dimensões conceituais

Considerando as associações e ligações identificadas por meio do conjunto de análises aplicadas ao *corpus* e inferências corroboradas por meio da visita *in loco*, foram identificas nove dimensões conceituais necessárias para uma efetiva GDC do Programa LBA/INPA: *(i)* Compartilhamento de Dados, *(ii)* Conectividade e *(iii)* Segurança, essas dimensões foram identificadas na análise de similitude com conexão entre os termos *Compartilhamento*, *Conectividade* e *Estabilidade*; *(iv)* Governança e Gestão de Dados, *(v)* Armazenamento e

Replicação de Dados, *(vi)* Curadoria Digital, estas dimensões foram identificadas com a conexão entre os termos *Governança*, *Armazenamento* e *Preservação*; *(vii)* Relação Semântica, identificada com a conexão entre os termos *Relação*, *Semântica* e *Ontologia*; *(viii)* Colaboração Científica e *(ix)* Interdisciplinaridade, identificadas com a conexão entre *Pesquisa* e *Colaboração*. No Quadro 3 são apresentadas as descrições conceituais para as dimensões propostas.

Quadro 3 – Descrição dos conceitos relativos às dimensões propostas

Identificação na análise de similitude	Dimensão considerada	Descrição	Fonte bibliográfica
Compartilhamento ⇒ Conectividade ⇒ Estabilidade [Cor verde na Fig. 4]	*Compartilhamento de Dados*	Liberação de dados primários de pesquisas para o uso de outros cientistas.	Borgman (2012)
	Conectividade	O fator conectividade, neste contexto, refere-se a interconexão de diversos equipamentos tanto de coleta em sítios como de armazenamento para uso e reuso de dados primários como dados derivados. Essa conectividade significa uma maior quantidade de dados, recolhidos a partir de mais lugares, com muitas maneiras de aumentar a eficiência e melhorar a proteção e segurança.	Fabricio et al. (2016)
	Segurança (SLA)	*Service Level Agreement* ou Contrato de Nível de Serviço. Contrato assinado pelo fornecedor e cliente de serviços de nuvem pública, com a finalidade de garantir, dentre outros requisitos, confidencialidade, disponibilidade, Qualidade do Serviço (QoS).	Dastjerdi et al. (2011).
Governança ⇒ Armazenamento ⇒ Preservação [Cor cinza na Fig. 4]	*Governança e Gestão de Dados*	Governança de dados é uma estrutura que orienta e estabelece estratégias, políticas e objetivos com a finalidade de gerenciar os dados, como se fossem qualquer outro recurso de uma organização. GDC é um termo geral capaz de cobrir a organização, a estrutura, o armazenamento e o cuidado da informação gerada durante o processo de pesquisa.	Loftis (2014) Universidade de Oxford (2013)
	Armazenamento e Replicação de Dados	A replicação é útil na melhoria da disponibilidade de dados. O caso mais relevante é a replicação do banco de dados inteiro em cada *site* no sistema distribuído, criando assim um banco de dados distribuído totalmente replicado.	Elmasri e Navathe (2011) Appel (2014)

Identificação na análise de similitude	Dimensão considerada	Descrição	Fonte bibliográfica
		Isso garante também maior segurança.	
	Curadoria Digital	Conjunto das todas as atividades existentes no gerenciamento de dados, desde o planejamento da sua criação, passando pela digitalização (transformação digital) ou criação, procurando assegurar a disponibilidade e adequação para a recuperação e reuso futuro destes dados.	Abbot (2008)
Relação ⇒ Semântica ⇒ *Ontologia* [Cor lilás na Fig. 4]	*Relação Semântica*	As relações semânticas são um importante componente para organização do conhecimento, sendo a unidade básica entre dois conceitos. A relação semântica, neste aspecto, tem por finalidade criar uma nova instância de um conhecimento (ou dado), possibilitando a expansão informacional sobre determinado *corpus*, ou ainda sua relação com dados primários. No contexto da ciência da informação, Khoo e Na (2006) consideram relações semânticas como relações significativas entre dois ou mais conceitos, entidades ou conjunto de entidades, podendo ainda fazer referência a relações entre conceitos mentais, entre elementos lexicais e entre parágrafos.	Hjørland (2003) Khoo e Na (2006)
Pesquisa ⇒ Colaboração [Cor vermelha na Fig. 4]	*Colaboração Científica*	Há colaboração científica quando dois ou mais cientistas trabalham juntos em um projeto de pesquisa e compartilham recursos intelectuais, econômicos ou físicos.	Vanz e Stumpf (2010)
	Interdisciplinaridade	A interdisciplinaridade está imbricada com as demais dimensões, uma vez que se deve considerar, além da comunidade científica, diversos outros atores da sociedade.	Minayo (2007, 2010) Philippi Jr e Silva Neto (2010) Alonso et al. (2011)

Fonte: Elaborado pelo autores

As dimensões *Segurança (SLA)* e *Interdisciplinaridade* foram inferidas e associadas, respectivamente, à *Compartilhamento* e *Colaboração* em função da natureza do Projeto

INPA/LBA e da infraestrutura utilizada, aferidas na visita *in loco* e entrevistas.

Por meio da AFC verificou-se a correspondência dos termos nos segmentos de texto. Quantitativamente, considerando os agrupamentos sugeridos na análise de similitude, cada dimensão possui uma frequência de participação nos segmentos de textos, conforme aponta o Quadro 4.

Quadro 4 – Frequência relativa (AFC) de dimensões em cada entrevista

Dimensão	***Entrevista***					
	n_1	n_2	n_3	n_4	n_5	n_6
Gestão e Governança de Dados	5,40	0,00	6,88	1,41	0,00	6,53
Relação Semântica	2,31	0,00	0,86	1,98	0,85	2,42
Armazenamento e Replicação de Dados	3,08	6,16	1,72	6,21	2,54	2,9
Curadoria Digital	1,54	3,08	2,14	0,56	5,93	0,97
Compartilhamento de Dados	0,00	7,70	1,29	0,28	3,39	0,00
Conectividade	0,00	0,00	3,01	0,00	0,00	4,11
Colaboração Científica	0,00	1,00	0,00	0,00	0,00	2,00

Fonte: Elaborado pelo autores

Observa-se que, embora as formas possuam uma frequência relativamente baixa, existe uma correlação entre as dimensões distribuídas em cada segmento de texto, possuindo desvio padrão de 22,25%. Isso também fica evidente no grafo construído a partir da análise de similitude (Figura 3).

A aplicação de cada dimensão conceitual ao programa LBA é apresentada conforme a seguir:

- *Compartilhamento de Dados*. Os dados produzidos pelo Programa LBA/INPA são originados por projetos e experimentos em locais distintos, compartilhados por diversas instituições e organismos nacionais e internacionais. O compartilhamento desses dados deve primariamente seguir uma política de aquisição, uso e reuso, garantindo a confidencialidade necessária à pesquisa, conforme determinado pelas instituições e grupos envolvidos e a disponibilidade para pesquisas futuras. Posteriormente, é fortemente recomendada a disponibilização desses dados em formato aberto.
- *Armazenamento e Replicação de Dados*. A infraestrutura para armazenamento dos dados deve levar em consideração os diversos grupos de pesquisa vinculados às áreas de pesquisa e aos inúmeros dispositivos de coletas que capturam variados tipos de dados, incluindo, entre outros, textos, séries temporais de dados de diversos tipos e imagens. A replicação de dados leva em consideração o acesso por grupos de pesquisa de instituições parceiras, mas não conectadas por alguma plataforma digital, e ainda a limitação de banda de internet, que impede a utilização integral de plataformas como DataONE, a infraestrutura do SINAPAD ou do supercomputador Santos Dumont.
- *Governança e Gestão de Dados*. A governança e a gestão incluem uma série de atividades sobre os dados de pesquisa, além de políticas de aquisição, uso, qualidade e garantia de disponibilidade (*data acquisition, data assurance e data quality*) para as diversas instituições envolvidas no projeto. As regras instituídas devem levar em consideração as seguintes ações: planejamento, coleta, garantia de disponibilidade, descrição (metadados), submissão, preservação, descoberta, integração, análise e

publicação. Essas regras são essenciais para a garantia de que os dados originados pelos diversos projetos de pesquisas estarão disponíveis no futuro para novas pesquisas.

- *Relação Semântica*. A relação semântica foi elencada devido a necessidade de permitir o acesso aos dados primários por meio dos resultados de pesquisa, considerando o contexto e esforço de coleta, aproximando os resultados da pesquisa à origem do problema estudado. Considera ainda a relação entre dados coletados por grupos de pesquisa e experimentos distintos dentro do Programa LBA/INPA e a criação de ontologias com a finalidade de criar sistemas de categorização.
- *Segurança* (SLA). Um Contrato de Nível de Serviço (SLA - *Service Level Agreement*) é necessário, pois uma das preocupações em ambos os grupos entrevistados foi quanto à garantia de confidencialidade dos dados. A equipe de TI mostrou preocupação quanto ao uso de uma solução de nuvem pública, enquanto os pesquisadores informaram que é importante garantir o compartilhamento de dados entre as instituições parceiras mediante uma política de dados que garanta o sigilo. Desta forma, o SLA deve prioritariamente ser discutido tanto na contração de serviços, como na realização de parcerias.
- *Colaboração científica*. As pesquisas realizadas no âmbito do Programa LBA/INPA possuem natureza interdisciplinar para a busca de soluções para problemas complexos. Interdisciplinaridade e complexidade são premissas que exigem a criação de redes de pesquisa para realização de trabalhos colaborativos. Existe essa prática, porém realizada conforme critérios particulares dos pesquisadores, utilizando metodologias, ferramentas próprias e individualizadas nos grupos. É preciso, portanto, o estabelecimento de políticas institucionais nesse sentido.
- *Conectividade*. A rede de experimentos instalada, o grande volume de dados existentes ou em vias de coleta e as oportunidades criadas pela HPC (*High-performance computing*), faz com que o Programa LBA/INPA demande uma conectividade estável e de alta velocidade, tanto com os equipamentos que coletam dados instalados em sítios remotos, como em *grids* instalados no país e em outros continentes para armazenamento de dados.
- *Interdisciplinaridade*. No contexto deste trabalho, a interdisciplinaridade está imbricada com as demais dimensões, uma vez que se deve considerar, além da comunidade científica, diversos outros atores da sociedade.

8 CONSIDERAÇÕES FINAIS

A finalidade deste trabalho foi propor um conjunto de dimensões conceituais necessárias para a GDC e GCC em um programa de pesquisa coordenado por uma instituição de pesquisa brasileira. Considerou-se que, em um ambiente computacional, onde esse conjunto de dimensões esteja habilitado, tenha-se um cenário ideal, desde a produção e coleta de dados até a publicação dos resultados da pesquisa, e promovendo a GCC suportada por recursos computacionais.

Foi realizado um amplo levantamento bibliográfico com a finalidade de verificar na literatura aspectos relativos à gestão e utilização de dados científicos e sua preservação digital. Isso permitiu verificar os avanços da ciência habilitada pela tecnologia.

Pelo estudo ter sido realizado no contexto de um programa de pesquisa específico, a visita *in loco* permitiu uma análise e avaliação detalhada sobre a atividade científica dos diversos grupos de pesquisa e da infraestrutura tecnológica oferecida para suporte a estas atividades, e a percepção dos atores envolvidos (pesquisadores, técnicos e gestores) sobre os recursos tecnológicos existentes e necessários para subsidiar o trabalho científico. Outros pontos também foram observados, a exemplo das formas como se dá o trabalho colaborativo

e a relação entre os pesquisadores e outros profissionais do Instituto. Por meio de entrevistas semiestruturadas foi possível mapear as características necessárias de ambientes computacionais e de infraestrutura de *e-Science* para o programa.

A análise automática dos dados coletados por meio das entrevistas possibilitou a explicitação da relação dos eixos temáticos identificados durante a fase de levantamento bibliográfico com as dimensões conceituais identificadas como essenciais para a gestão conhecimento e de dados científicos produzidos no âmbito do programa de pesquisa estudado.

AGRADECIMENTOS

Os autores agradecem a valiosa contribuição do Dr. José Laurindo Campos dos Santos, do Instituto Nacional de Pesquisas da Amazônia, sem a qual esta pesquisa não teria sido possível.

REFERÊNCIAS

ABBOT, D. **What is digital curation?** DCC Briefing Papers: Introduction to Curation. Edinburgh: Digital Curation Centre, 2008. Disponível em: https://www.dcc.ac.uk/guidance/briefing-papers/introduction-curation/what-digital-curation. Acesso em: 12 jan. 2021.

ALONSO, L.; SALLANTIN, J.; FERNEDA, E. ; LUZEAUX, D. Scientific Knowledge Management Anchored on Socioenvironmental Systems. **TripleC**, v. 9, n. 2, p. 610–623, 2011.

AMARAL, L. Q. O processo de validação do conhecimento científico. **Jornal da USP**, 24 abr 2017. Disponível em: http://jornal.usp.br/artigos/o-processo-de-validacao-do-conhecimento-cientifico. Acesso em: 25 jan. 2022.

APPEL, A. L. **A e-Science e as atuais práticas de pesquisa científica**. Dissertação (Mestrado em Ciência da Informação) – IBICT-UFRJ, Rio de Janeiro, 2014. Disponível em: https://ridi.ibict.br/bitstream/123456789/872/1/Pesquisa_Andre_Appel_2014-06-26_final.pdf. Acesso em: 17 dez. 2021.

AVISSAR, R.; SILVA-DIAS, P. L.; SILVA-DIAS, M. A. F.; NOBRE, C. The Large-Scale Biosphere-Atmosphere Experiment in Amazonia (LBA): Insights and future research needs, **Journal of Geophysical Research**, v. 107, n. D20, 2002. https://doi.org/10.1029/2002JD002704.

BAKER K. S. MAYERNIK, M. S. Disentangling knowledge production and data production. **Ecosphere**, v. 11, n. 7, 2020. https://doi.org/10.1002/ecs2.3191.

BERNIUS, S. The impact of open access on the management of scientific knowledge. **Online Information Review**, v. 34, n. 4, p. 583–603, 2010. https://doi.org/10.1108/14684521011072990.

BORGMAN, C. L. The conundrum of sharing research data. **Journal of the American Society for Information Science and Technology**, v. 63, n. 6, p. 1059–1078, 2012. https://doi.org/10.1002/asi.22634.

CHIN JR., G.; LEUNG, L. R.; SCHUCHARDT, K.; GRACIO, D. New paradigms in problem solving environments for scientific computing. In: INTERNATIONAL CONFERENCE ON INTELLIGENT

USER INTERFACES, 7, 2002, San Francisco (USA). **Proceedings [...]**. Association for Computing Machinery (ACM), 2002. pp. 39-46. https://doi.org/10.1145/502716.502726.

COLLINS, H. M. Tacit Knowledge, Trust and the Q of Sapphire. Social Studies of Science, v. 31, n. 1, p. 71-85, 2001.

COSTA, M. M. **Diretrizes para uma política de gestão de dados científicos no Brasil**. Tese (Doutorado em Ciência da Informação) – Universidade de Brasília, Brasília, 2007. Disponível em: https://repositorio.unb.br/handle/10482/24895. Acesso em: 02 fev. 2001.

CRESWELL, J. W.; CRESWELL, J. D. **Projeto de Pesquisa**: Métodos qualitativo, quantitativo e misto. 5a ed. Penso Editora, Porto Alegre, 2021.

DAVENPORT, T. H.; PRUSAK, L. **Conhecimento Empresarial**: como as organizações gerenciam o seu capital intelectual. 10. ed. Rio de Janeiro: Campus, 2003.

DUTRA, M. L.; MACEDO, D. D. J. Curadoria Digital: Proposta de um modelo para curadoria digital em ambientes Big Data baseado numa abordagem semi-automática para a seleção de objetos digitais. **Informação & Informação**, v. 21, n. 2, p. 143-169, 2016. https://doi.org/10.5433/1981-8920.2016v21n2p143.

ELMASRI, R.; NAVATHE, S. B. **Sistema de Banco de Dados**. 6a ed. Editora Pearson, 2011.

EMILIO, T. LUIZÃO, F. (Orgs). **Cenários para a Amazônia**: Clima, Biodiversidade e Uso da Terra. Manaus (AM): Editora INPA, 2014. Disponível em: https://www.academia.edu/12087225/CEN%C3%81RIOS_PARA_A_AMAZ%C3%94NIA_CLIMA_BIODIVERSIDADE_E_USO_DA_TERRA. Acesso em 14 jan. 2021.

FABRICIO, M. A.; BEHRENS F.; BIANCHINI, D. Monitoramento de Equipamentos Elétricos para Manutenção Preditiva utilizando IoT. In: BRAZILIAN TECHNOLOGY SYMPOSIUM, 1, 2016, Campinas. **Proceedings [...]**. Faculdade de Engenharia Elétrica e de Computação da Unicamp, 2016. Disponível em: https://www.lcv.fee.unicamp.br/images/BTSym-16/proceedings/pa49-16-edited.pdf. Acesso em: 17 abr. 2021.

FERREIRA, M. A. **Estudo sobre a utilização de ferramentas de colaboração em redes de pesquisa científica**. Dissertação (Mestrado em Gestão do Conhecimento e da Tecnologia da Informação) – Universidade Católica de Brasília, Brasília, 2010. Disponível em: https://bdtd.ucb.br:8443/jspui/handle/123456789/1315. Acesso em: 23 mai. 2021.

FONSECA, J. J. S. **Metodologia da Pesquisa Científica**. Universidade Estadual do Ceará, Apostila, 2002.

HAIR JUNIOR., J. F.; ANDERSON, R. E.; TATHAM, R. L.; BLACK, W. C. **Multivariate data analysis with readings**. 4a ed. New Jersey: Prentice Hall, 1998.

HEIMERL, F.; LOHMANN, S.; LANGE, S.; ERTL, T. Word Cloud Explorer: text analytics based on word clouds. In: HAWAII INTERNATIONAL CONFERENCE ON SYSTEM SCIENCES, 47, 2014, Waikoloa (EUA), **Proceedings [...]**. University of Hawai'i at Mānoa, 2014. pp. 6-9. Disponível em: https://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=6758829. Acesso em: 16 mar. 2021.

HIGGINS, S. The DCC Curation Lifecycle. **The International Journal of Digital Curation**, v. 3, n. 1, 2008. https://doi.org/10.2218/ijdc.v3i1.48.

HJØRLAND, B. Fundamentals of knowledge organization. **Knowledge Organization**, v. 30, n. 2, p. 87–111, 2003.

INTERNATIONAL ORGANIZATION FOR STANDARDIZATION. **ISO 19115- I: Geographic information – Metadata – Part 1: Fundamentals**. ISO, 2014.

KELLER, M.; DIAS, M. A. S.; NEPSTAD, D. C.; ANDREAE, M. O. **The Large-Scale Biosphere-Atmosphere Experiment in Amazonia**: Analyzing Regional Land Use Change Effects. Ecosystems and Land Use Change. Geophysical Monograph Series 153, American Geophysical Union, 2004.

KÖCHE, J. C. **Fundamentos de Metodologia Científica**: Teoria da ciência e iniciação à pesquisa. 14. Ed. Petrópolis (RJ): Editora Vozes, 2011.

KHOO, C. S. G.; NA, J. Semantic relations in Information Science. **Annual Review of Information Science and Technology**, v. 40, p. 157–228, 2006. https://doi.org/10.1002/aris.144040011.

LEE, C.; TIBBO, H. Digital curation and trusted respositories: steps toward success. **Journal of Digital Information**, v. 8, n. 2, 2007. Disponível em: http://journals.tdl.org/jodi/article/viewArticle/229/183. Acesso em: 17 dez. 2020.

LEITE, F. C. L.; COSTA, S. M. S. Gestão do conhecimento científico: proposta de um modelo conceitual com base em processos de comunicação científica. **Ciência da Informação**, v. 36, n. 1, p. 92-107, 2007. https://doi.org/10.18225/ci.inf.v36i1.1189

MARCHAND, P.; RATINAUD, P. L'analyse de similitude appliqueé aux corpus textueles: les primaires socialistes pour l'election présidentielle française. In: JOURNÉES INTERNATIONALES D'ANALYSE STATISTIQUE DES DONNÉES TEXTUELLES (JADT), 11, 2012, Liège (Belgique). **Actes […]**. Université de Liège, 2012. pp. 687-699.

MARTINELLI, S. G.; CORRÊA, A. C.; SCHUCH JUNIOR, V. F.; LOPES, L. F. D. Gestão do conhecimento científico em universidades: mapeamento dos processos de desenvolvimento de projetos. In: COLÓQUIO INTERNACIONAL DE GESTÃO UNIVERSITÁRIA, 17, Mar del Plata, Argentina, **Anais** [...]. 2017. Disponível em: http://www.sigmees.com.br/files/GESTAO_DO_CONHECIMENTO_EM_UNIV-SUSTENTAB_IES_CIGU_2016.pdf. Acesso em: 14 dez. 2020.

MATTOSO, M.; WERNER, C.; TRAVASSOS, G. H.; BRAGANHOLO, V.; MURTA, L. Gerenciando Experimentos Científicos em Larga Escala. In: CONGRESSO DA SOCIEDADE BRASILEIRA DE COMPUTAÇÃO / SEMINÁRIO INTEGRADO DE SOFTWARE E HARDWARE, 18, 2008, Belém (PA). **Anais** [...]. SBC, 2008. Disponível em: https://www.researchgate.net/publication/228709654_Gerenciando_Experimentos_Cientificos_em_Larga_Escala/link/0912f510674d0c7b8a000000/download. Acesso em: 23 jan. 2021.

MINAYO, M. C. S. **O desafio do conhecimento**: pesquisa qualitativa em Saúde. 10. ed. São Paulo: HUCITEC, 2007.

MINAYO, M. C. S. Transdisciplinaridade, interdisciplinaridade e complexidade. **Emancipação**, v. 10, n. 2, p. 435–442, 2010. Disponível em: https://revistas.uepg.br/index.php/emancipacao/article/view/1937. Acesso em: 22 jan. 2021.

NEWMAN, M. E. J. The structure of scientific collaboration networks. **Proceedings of the National Academy of Sciences of the United States of America**, v. 98, n. 2, p. 404–409, 2001. https://doi.org/10.1073/pnas.98.2.40.

NONAKA, I.; TAKEUCHI, H. **Criação do Conhecimento na Empresa**: Como as empresas japonesas geram a dinâmica da inovação. Rio de Janeiro (RJ): Elsevier, 1997.

OLIVEIRA, E. H. C. Redes de colaboração em pesquisa e intercâmbio de conhecimento científico. **Revista Pan-Amazônica de Saúde**, v. 9, n. 4, p. 9-11, 2018. http://dx.doi.org/10.5123/s2176-62232018000400001.

PHILIPPI JR, A.; SILVA NETO, A. J. **Interdisciplinaridade em Ciência, Tecnologia & Inovação**. Barueri (SP): Editora Manole, 2010.

REINERT, M. ALCESTE, une méthodologie d'analyse des données textuelles et une application: *Aurélia* de Gerard de Nerval. **Bulletin de Méthodologie Sociologique**, v. 28, p. 24-54, 1990. https://doi.org/10.1177/075910639002600103.

RIBES, D.; LEE, C. P. Sociotechnical studies of cyberinfrastructure and e-research: current themes and future trajectories. **Computer Supported Cooperative Work**, v. 19, n. 3-4, p. 231-244, 2010. https://doi.org/10.1007/s10606-010-9120-0.

SAMPAIO, J. O. **METHEXIS: uma abordagem de apoio à Gestão do Conhecimento para ambientes de e-Science**. Tese (Doutorado em Ciências) – Universidade Federal do Rio de Janeiro, Rio de Janeiro, 2007.

SANTOS, D. B. G.; SANTOS, J. C. Applicability of knowledge management practices at scientific research environment. In: IBERIAN CONFERENCE ON INFORMATION SYSTEMS AND TECHNOLOGIES (CISTI), 10, 2015, Aveiro (Portugal). **Proceedings** […]. IEEE, 2015. https://doi.org/10.1109/CISTI.2015.7170556.

SANTOS, T. S.; MENEZES, A. M. F. Gestão do conhecimento científico como síntese interdisciplinar: interfaces teórico-conceituais entre a gestão do conhecimento, a comunicação científica e a comunicação organizacional. **PontodeAcesso**, v. 13, n. 3, p. 167–183., 2019. https://doi.org/10.9771/rpa.v13i3.34899.

SAYÃO, L. F.; SALES, L. F. **Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores**. Rio de Janeiro. CNEN/IEN, 2015.

SILVA, A. H.; FOSSÁ, M. I. T. Análise de conteúdo: exemplo de aplicação da técnica para análise de dados qualitativos. **Qualit@s**, v. 17, n. 1, 2015.

SOUZA, M. O que é SSO ou Single Sign-On? **e-trust**, 2017, Disponível em: https://www.e-trust.com.br/o-que-e-sso-ou-single-sign-on. Acesso em: 17 fev. 2022.

STEWART, T. A. **Capital Intelectual**. Rio de Janeiro: Campus, 1998.

TERRA, C. **Gestão do Conhecimento:** o grande desafio empresarial. Negócio Editora, 2000.

UNITED NATIONS. **Amazon Assessment Report 2021**. New York: United Nations Sustainable Development Solutions Network, 2021. https://www.theamazonwewant.org/amazon-assessment-report-2021/. Acesso em: 14 jul. 2022.

VANZ, S. A. S.; STUMPF, I. R. C. Colaboração científica: revisão teórico-conceitual. **Perspectivas em Ciência da Informação**, v. 15, n. 2, p. 42–55, 2010.

Recebido em/Received: 24/09/2022 | Aprovado em/Approved: 21/03/2023