8 Engenharia da informação: contribuições para a agricultura digital

Ivo Pierozzi Júnior
Marcos Cezar Visoli
Marcia Izabel Fugisawa Souza
Luiz Manoel Silva Cunha
Isaque Vacari
Tércia Zavaglia Torres

1 Introdução

A agricultura digital apresenta-se como um exercício de modelagem dos fenômenos e dos processos agropecuários, nas dimensões ambiental, econômica e social, por meio de artefatos computacionais e de Tecnologias de Informação e Comunicação (TIC), visando trazer para o setor agrícola facilidades de organização, acesso, uso, compartilhamento, disseminação e aplicação do conhecimento científico.

> São múltiplos os desafios de um mundo globalizado, sendo até mesmo difícil obter consenso sobre quais seriam os prioritários. Todavia, um deles – tornar o conhecimento acessível a todos – destaca-se como mais importante devido aos seus efeitos estruturantes. Em nenhuma outra época da história, a produção de conhecimentos foi tão intensa como nos dias de hoje, como também em nenhuma outra época a sua aplicação assumiu papel tão preponderante. Daí a importância da gestão do conhecimento, pois entre a sua produção e a sua utilização há uma cadeia de procedimentos complexos que podem ou não determinar o seu êxito operativo. Para alguns especialistas como Manuel Castells, a aplicação do

> conhecimento está na centralidade da revolução conceitual e operacional impulsionada pelos avanços da ciência e da tecnologia que se opera nas sociedades contemporâneas, e que atinge em velocidade sem precedentes todos os setores da vida humana. Importa, assim, pensar a utilização de conhecimentos, pavimentar caminhos para os seus diversos usos e assegurar a sua dimensão social e ética. (Defourny, 2006, p. 7).

O termo Engenharia da Informação, conforme apresentado por Martin e Finkelstein (1989), enuncia-se em três definições diferentes, mas convergentes conceitualmente. Em duas delas, ressalta-se a palavra "automatizadas":

1) "A aplicação de um conjunto interligado de técnicas formais de planejamento, análise, projeto e construção de sistemas de informações sobre uma organização como um todo ou em um de seus principais setores";
2) "Um conjunto interligado de técnicas automatizadas no qual são construídos modelos de organização, modelos de dados e modelos de processos em uma abrangente base de conhecimentos, a fim de serem usados para criarem e manterem sistemas de processamento de dados";
3) "Um conjunto de disciplinas automatizadas em nível de organização cuja finalidade é fornecer as informações certas, às pessoas certas e na hora certa."

A partir dessa perspectiva e desse contexto é que a Engenharia da Informação é apresentada no presente capítulo, conforme entendimento desta disciplina, como via de mapeamento, organização e representação do conhecimento agropecuário e no contexto da agricultura digital.

Na perspectiva da Gestão do Conhecimento (GC), o itinerário de entrega do conhecimento científico e a garantia de sua eficácia e eficiência como resposta às demandas da sociedade, incluindo aquelas afetas à agropecuária (alimentos, energia e fibras, fundamentalmente) são possibilitados pela Engenharia da Informação. Na Embrapa, a relação entre dados, informação e conhecimento já foi trabalhada (Pierozzi et al., 2017) para viabilizar a utilização de concepções teóricas e conceituais que alinham esses três níveis de organização da percepção humana sobre o mundo real e sua consequente transformação em tecnologias.

A aplicação do conhecimento, ou seja, sua apreensão e utilização como solução de problemas e desafios, passa pelo processo de tomada de decisão. Não existe a melhor decisão a ser tomada. Existe a decisão possível diante da capacidade de se identificar, reunir, processar e conjugar o maior volume possível de informações sobre um determinado assunto. Assim, enquanto disciplina de pesquisa, de desenvolvimento e de inovação, a Engenharia da Informação posiciona-se no ponto central para conjugar dados, originados

da prática da pesquisa agropecuária, ao conhecimento, este representando a oferta e a utilização dos resultados da atuação da Embrapa, modelados em TIC.

Soma-se aos desafios mencionados, o ritmo dinâmico e massivo de produção e oferta de conhecimento, acelerado pelo avanço e pelo suporte das TIC. Então, simultaneamente, emerge outro desafio no qual a qualidade do conhecimento oferecido configura-se igualmente como demanda social: o conhecimento que se busca passa a ser exigido como ambientalmente sustentável, economicamente viável e socialmente justo. Não é diferente no contexto da pragmática do conhecimento científico e, em particular, no contexto do conhecimento agropecuário.

A Embrapa tem manifestado, em seus exercícios constantes de planejamento estratégico, no qual revisa periodicamente sua missão, visão, objetivos e metas (Embrapa, 2018), constante preocupação com a entrega de conhecimento tecnológico para a sociedade, utilizando premissas de qualidade e efetividade. Sua recente incursão na implantação de modelos de gestão de pesquisa orientados à inovação corroboram a convergência e a coerência dessa intenção, em especial porque é uma empresa geradora de conhecimentos e competências e, portanto, uma empresa que aprende e evolui científica, tecnológica e organizacionalmente (Garcia; Salles Filho, 2009).

Diante dessa realidade, a Embrapa Informática Agropecuária insere a sua contribuição, visto que tem investido esforços para desenvolver e inovar metodologias e tecnologias e produzir conhecimentos dentro de suas competências em Computação e TIC.

É nesse contexto que se alinham e exploram as vias da agricultura digital, conceito e termo que se apresenta, no âmbito da PD&I e da C&T, não apenas como uma tendência mas, em especial, como um novo paradigma socioeconômico do setor agrícola, já que lança mão de TIC para dar fluxo aos dados de pesquisa e transmutá-los em informação, conhecimento e tecnologia para o produtor, por meio de *Internet of Things* (IoT), *Big Data, Cloud Computing, Machine Learning* etc. Esse paradigma inerente à agricultura digital conjuga-se com outros paradigmas contemporâneos, como a Economia da Informação ou Economia do Conhecimento, a Ciência de Dados e a Ciência Aberta (Porat; Rubin, 1977; Powell; Snellman, 2004; Pordes et al., 2007; Aalst, 2016).

A palavra "engenharia" tem se associado à computação (engenharia de software, de dados, do conhecimento etc.), como forma de expressar os processos de "construção" de artefatos computacionais que representam as coisas (entidades, fenômenos, processos) do mundo real para a linguagem de máquinas. Uma explicação possível para esse fenômeno linguístico de recombinações conceituais e terminológicas interdisciplinares é o entendimento de que o uso prático do conhecimento é um interminável processo, contínuo e dinâmico, de recombinação, análise, síntese e ressignificação conceituais

em contextos permanentemente emergentes. Daí a significação metafórica da palavra "engenharia".

Paralelamente, no mesmo itinerário de construção e desenvolvimento do conhecimento, outra reflexão conceitual tem associado as palavras "dado", "informação" e "conhecimento" (D-I-C), criando vários modelos de representação dessa relação e induzindo e ressignificando, a partir daí, as acepções do termo "engenharia". Recentemente, na proposição de um modelo de Governança de Dados e Informação para Conhecimento na Embrapa, uma concepção de modelo dessa relação, diferente das convencionais, foi apresentada com o intuito de facilitar a sua implantação organizacional e operacional em suporte aos processos corporativos de gestão de dados, informação e conhecimento (Pierozzi Junior et al., 2017). O modelo tem como referenciais teóricos e conceituais, além da noção da relação D-I-C, os ciclos de vida dos dados, da informação e do conhecimento, concebidos de forma conjugada e alinhada e representados como uma mandala.

Esse modelo também embasa uma abordagem ontológica (Mol, 2008), que serve como itinerário de construção do ôntico, ou seja, da entidade propriamente dita. O termo "entidade" é empregado como referência aos objetos ou artefatos computacionais a serem engenhados (softwares, aplicativos, sistemas de informação etc.), visto serem esses os objetos que concretizam operacionalmente o conhecimento científico e multidisciplinar, viabilizando sua aplicação na solução de problemas ou em resposta a demandas do setor agropecuário.

A partir dessa concepção geral, a Engenharia da Informação, enquanto área de conhecimento, disciplina ou proposta de processo produtivo de tecnologias e inovação, configurou-se como uma atraente opção conceitual e terminológica para reunir as competências, as tecnologias e as soluções executadas e produzidas pela Embrapa Informática Agropecuária ao longo de sua história e, principalmente, como uma opção oportuna para sistematizar o processo de transformação de dados científicos em conhecimento pragmático.

Para tanto, um metamodelo conceitual está sendo elaborado para organizar concepções de mundo nas esferas ambiental, agropecuária, social e econômica (Figura 1), visando o desenvolvimento de produtos computacionais em resposta a desafios e oportunidades para a agricultura digital. Outro metamodelo (Figura 2) reúne abordagens conceituais, metodológicas e tecnológicas que se alinham, a partir do conceito de Engenharia da Informação, como um constructo integrador de pragmáticas do conhecimento que são inerentes a várias ciências como as da Cognição, Informação e Computação.

Nas seções seguintes serão apresentadas, discutidas e contextualizadas ações de pesquisa e resultados que, no contexto da Engenharia da Informação, estão sendo desenvolvidos na Embrapa Informática Agropecuária.

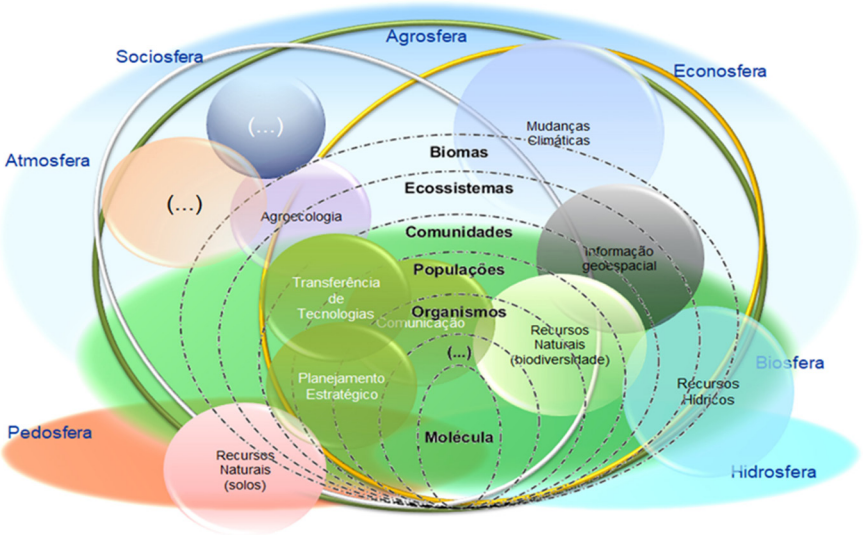

Figura 1. Representação conceitual multi, inter e transdisciplinar da agropecuária.

2 Sistemas de organização e representação do conhecimento

Para tornar o conhecimento acessível e utilizável, seja por agentes humanos, seja por agentes tecnológicos, é preciso organizá-lo (Soergel, 2009). Dessa forma, incluída a percepção tecnológica, a Engenharia da Informação pode ser entendida como uma área ou disciplina de conhecimento que permite construir um itinerário operacional, com suporte de computação e TIC, para que o conhecimento se torne acessível e utilizável.

A Enciclopédia de Organização do Conhecimento define os Sistemas de Organização do Conhecimento (SOC) como:

> […] um termo genérico usado para se referir a uma ampla gama de itens (por exemplo, títulos de assuntos, tesauros, esquemas de classificação e ontologias), que foram concebidos com relação a diferentes finalidades, em momentos históricos distintos. Eles são caracterizados por diferentes estruturas e funções específicas, maneiras variadas de se relacionar com a tecnologia e usados em uma pluralidade de contextos por diversas comunidades. No entanto, o que todos eles têm em comum é que foram projetados para apoiar a organização de conhecimento e informação, a fim de facilitar o gerenciamento e a recuperação. (Mazzocchi, 2019, p. 1, tradução nossa).

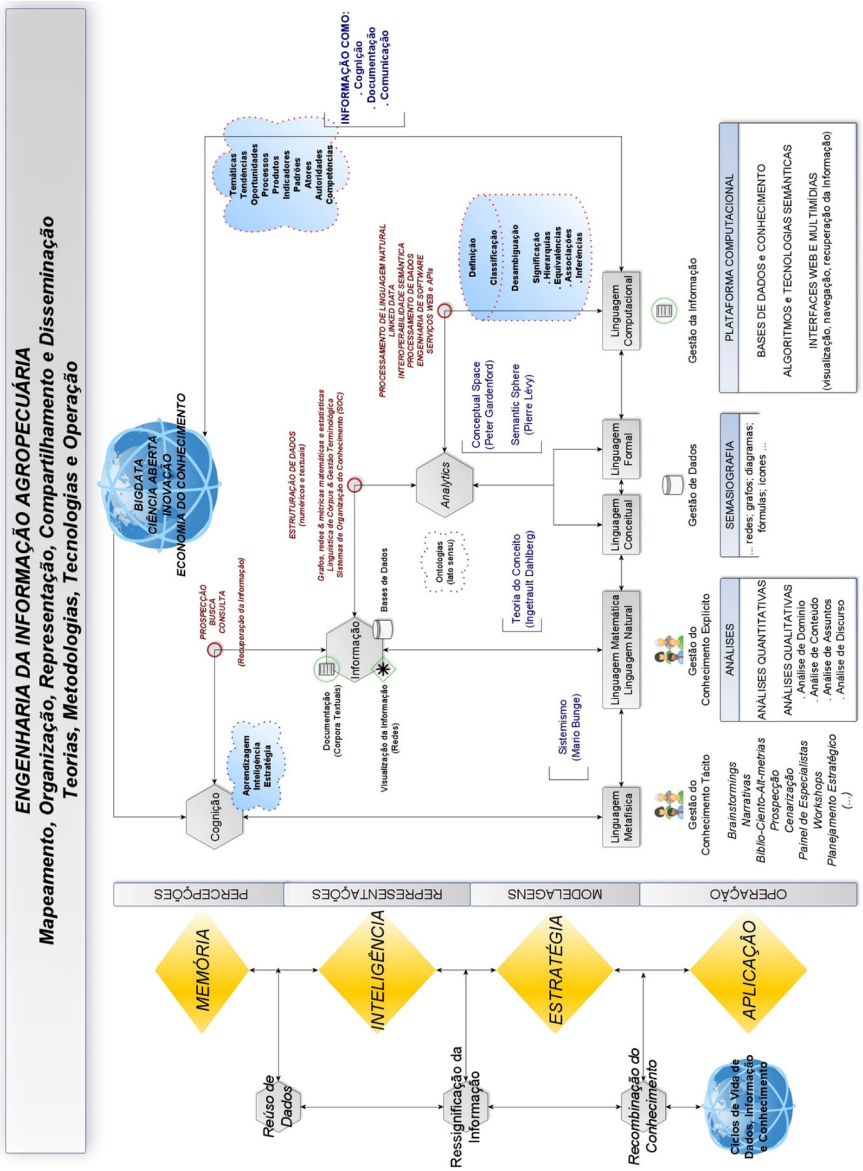

Figura 2. Metamodelo da Engenharia da Informação Agropecuária.

Também podem ser definidos como "[…] sistemas conceituais semanticamente estruturados que contemplam termos, definições, relacionamentos e propriedades dos conceitos" (Carlan; Medeiros, 2011, p. 54). O termo é a tradução para o português da expressão "Knowledge Organization System" (KOS), proposta pelo Networked Knowledge Organization Systems Working Group, na 1ª Conferência da ACM Digital Libraries, em 1998, em Pittsburgh, na Pennsylvania (Carlan; Medeiros, 2011, p. 54).

As Figuras 3 e 4 ilustram bem como os SOC podem ser entendidos e indicam como podem ser apreendidos e utilizados no âmbito da Engenharia da Informação.

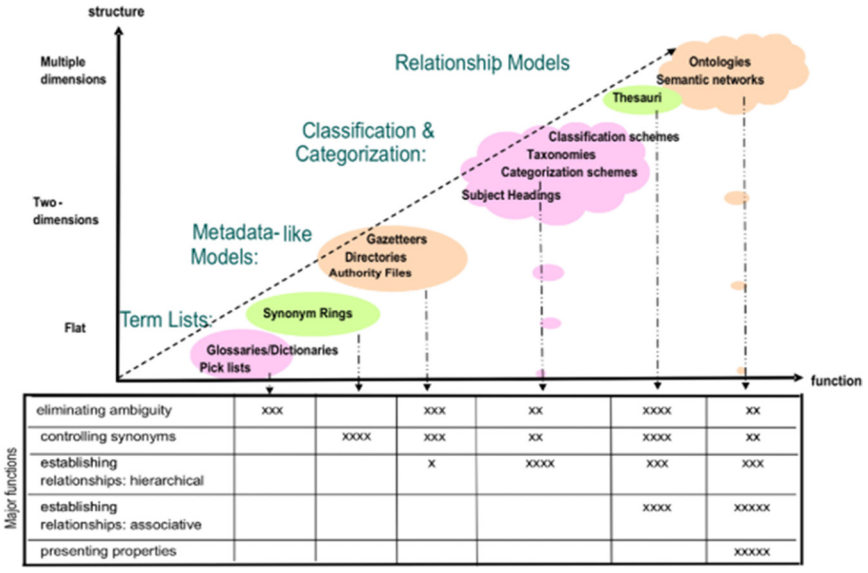

Figura 3.
Tipos de SOC.

Fonte: Zeng (2008).

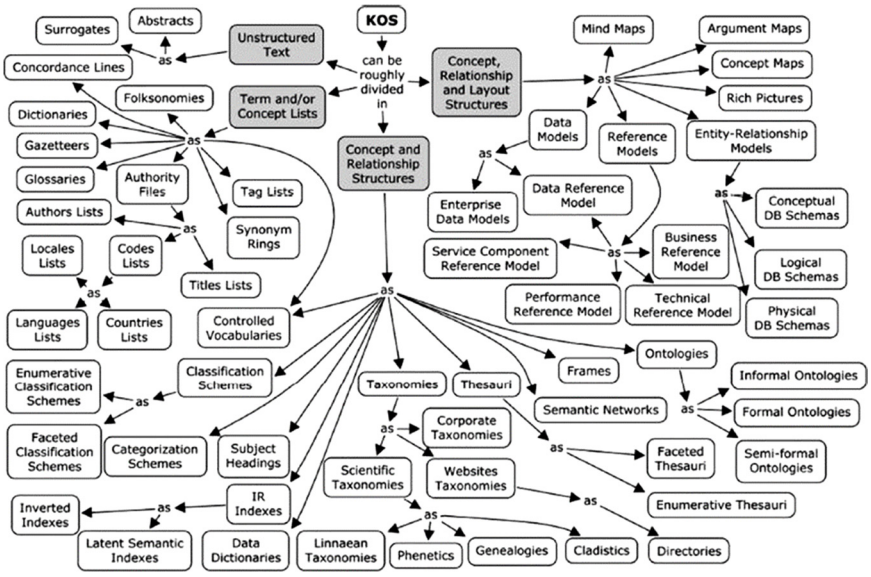

Figura 4.
Classificação dos SOC.

Fonte: Souza et al. (2012).

O processo de organização e representação do conhecimento sempre fez parte da história da humanidade (Martins; Moraes, 2015). Esse processo configura-se como uma ação multi e interdisciplinar inerente a todas as ciências

e as culturas. Assim, tanto quando Aristóteles[1] definiu as dez categorias do ser como categorias metafísicas, que classificam palavras em relação ao nosso conhecimento do ser, quanto hoje, na modernidade plural na qual estamos submersos, necessitamos de métodos de organização e representação do conhecimento para compreendermos o mundo à nossa volta (Mendonça, 2005).

Ocorre que, para representar o conhecimento, é necessário, antes de qualquer ação, organizá-lo, ou seja, precisamos ordená-lo em um determinado sistema para que possamos compreendê-lo (Martins; Moraes, 2015). O processo de organização de objetos é um processo classificatório e relacional, que exige do homem a capacidade cognitiva de associar ideias, criar ordem e sentido nas suas experiências, usando a interpretação do mundo, a atribuição de significados e a estruturação das ideias. A representação do conhecimento é, portanto, parte de uma totalidade que, pela percepção e pela razão, busca a formulação de conceitos abstratos sobre a realidade a qual pertence (Martins; Moraes, 2015).

Pode-se argumentar que o mundo como objeto do conhecimento humano existe como um mundo interpretado que é completamente infundido com significado. A cognição humana não pode ver fatos simples sem que esses façam parte de sua estrutura de significação (Tuomi, 1999). Daí a importância dos sistemas de organização e representação de conhecimento, sobretudo quando se inserem no escopo do desenvolvimento de soluções de Tecnologias da Informação e Comunicação (TIC) para a agricultura digital.

No contexto do Grupo de Pesquisa de Engenharia da Informação da Embrapa Informática Agropecuária, esses sistemas são entendidos como objetos "performados", ou seja, construídos e executados de forma alinhada a seus contextos, como objetos que inserem realidades múltiplas, respondendo diretamente às peculiaridades da modelagem de sistemas complexos. No âmbito do conhecimento agropecuário brasileiro, os SOC apresentam-se como sistemas organizativos e representativos do conhecimento desse domínio, descrevendo associações entre os múltiplos elementos interdependentes que atuam na produção do conhecimento (Latour, 2012).

Eles devem ser percebidos como espaços que integram os contextos social, cultural, ambiental e econômico da agropecuária brasileira em relações de interdependência que, para além da ação, consideram o processo de produção de conhecimento no escopo de uma realidade articulada e híbrida, com capilaridades complexas que envolvem agentes humanos e não humanos.

Nesse sentido, o sistema de representação e organização de conhecimento na Embrapa é um sistema aberto, que se configura mais como um mapa constitutivo de uma rede de atores que se autoinfluenciam por estarem em

1 As dez categorias do conhecimento são utilizadas na classificação e na representação do pensamento humano, ou seja, os pensamentos tornaram-se ponto de partida para as representações.

permanente interação – redesenhando novos percursos – do que como um território fechado, limitado e estático.

Enquanto sistemas abertos que possuem agentes (humanos e não humanos), que se articulam e se transformam mutuamente, os SOC na Embrapa têm sido construídos e empregados em diversas dimensões de domínios de conhecimentos, resultando em ontologias fragmentadas e dispersas no tempo e no espaço, cujo potencial de representação prejudica-se em função da ausência de percepções de mais alto nível. Assim, a utilização desses artefatos conceituais da forma como têm sido concebidos e utilizados, em suporte à modelagem computacional do conhecimento agropecuário, ainda enfrenta dificuldades de coerência e convergência, pois os SOC deveriam representar múltiplas ontologias advindas de uma variedade de métodos e práticas empregados pelos pesquisadores, que movem o conhecimento da agropecuária brasileira na Embrapa (Baum et al., 2020). Dessa forma, SOC concebidos sob a ótica da política ontológica (Mol, 2008) são compreendidos como objetos performáticos que possuem maior aderência ao paradigma da complexidade à qual a agricultura brasileira está submersa.

Assim, a agropecuária como um objeto de conhecimento não é um tema passivo à espera de ser percebido sob o ponto de vista de uma infindável série de perspectivas. Ao contrário, é um objeto vivo e aberto que se constitui por meio das práticas científicas, através das quais é manipulado para ser melhor compreendido (Baum et al., 2020).

Essa racionalidade justificou a necessidade de engenhar um sistema de organização e representação de conhecimento para a Embrapa que tivesse concepção ontológica pluralista, o que implica em afirmar que é um sistema orientado e "[…] favorável à coexistência de uma variedade de explicações, suposições, métodos, metodologias, abordagens, teorias" (Baum et al., 2020, p. 14), que juntas performatam a agricultura como objeto aberto, exposto às suas múltiplas relações, interações, capilaridades e, consequentemente, à sua própria inteireza.

2.1 Agrotermos: vocabulário controlado da Embrapa

O conhecimento é uma experiência intelectual pessoal e, dessa forma, o termo "transferência de conhecimento" pode até ter sentido conceitual, mas na prática não possui sentido operacional. Tal transferência, na verdade, ocorre por meio de um processo de codificação da energia cerebral em linguagem natural, e se manifesta via comunicação entre um agente "emissor" e um agente "receptor". A humanidade tem executado esse processo tão naturalmente que, por vezes, nem se considera que existem outros meios possíveis de codificação do conhecimento, como símbolos, sons, cheiros, texturas etc. A verdade é que, fundamentalmente, quase a totalidade do conhecimento humano tem sido codificada em linguagem natural falada ou escrita e, mais recentemente,

de forma digital, o que ainda mantém sua natureza original. Essa naturalidade humana de representação está baseada, certamente, na preponderância da percepção visual em detrimento dos outros sentidos.

Boa parte dos SOC, então, são concebidos, construídos e executados por meio dessa via: baseados no léxico, que, por sua vez, tecnologicamente, pode ser modelado por meio de métodos e ferramentas de Processamento de Linguagem Natural (PLN) e, assim, codificado graficamente.

Vocabulários controlados são SOC que colecionam e organizam palavras ou, no campo das especialidades científicas, termos. Com base na Teoria do Conceito, termos denotam conceitos, mas são apenas um dos vértices do triângulo que representa um determinado conceito. Outro vértice é o referente, ou seja, aquilo que no mundo real a mente humana percebe. Por fim, o terceiro vértice são as propriedades que se podem atribuir ao referente e que, finalmente, orientam a escolha de um elemento léxico na linguagem natural que melhor sintetize a percepção do referente (Dahlberg, 1978). Segundo a mesma autora, conceitos são considerados como unidades do conhecimento. Dessa forma, contextualiza-se o papel dos vocabulários controlados enquanto recursos facilitadores na gestão de instituições, que se confrontam com produção, acesso e compartilhamento de D-I-C, inclusive em escalas de volume e fluxo de *Big Data*. E na mesma lógica, vocabulários controlados, atualmente, beneficiam-se imensamente da Engenharia da Informação para serem concebidos, geridos e mantidos como sistemas abertos e dinâmicos, em conformidade com premissas de melhores práticas de representação do conhecimento e suas aplicações pragmáticas.

Até recentemente, a Embrapa utilizava, em seus processos corporativos de gestão de D-I-C, vocabulários controlados externos, concebidos e geridos em contextos não perfeitamente alinhados ao seu próprio repertório de conteúdos referentes à agropecuária tropical, o que dificultava enormemente o alinhamento desse acervo de conhecimento a várias fases dos ciclos de vida D-I-C, ainda mais quando desdobradas em escalas globais. No nível dos processos de catalogação, indexação, recuperação, acesso e disseminação de D-I-C, problemas de consistência, ambiguidade e falta de interoperabilidade entre sistemas de informação acumularam-se na mesma proporção que tal acervo crescia exponencialmente.

O Agrotermos foi então construído, por meio da reunião das terminologias em língua portuguesa presentes em tesauros agrícolas nacionais e internacionais, com base metodológica e tecnológica na iniciativa Global Agricultural Concept Space (GACS) (Research Data Alliance, 2020). Em decorrência, a Embrapa foi reconhecida como curadora do conteúdo em língua portuguesa, para a variante brasileira, pelo grupo editor do Agrovoc (FAO, 2020) – o vocabulário controlado da Food and Agriculture Organization (FAO) da Organização das Nações Unidas (ONU).

Atualmente, o Agrotermos está composto por aproximadamente 245 mil termos. Pelas vias da Engenharia da Informação, com utilização de metodologias e ferramentas de PLN, Linguística de Corpus e modelagem semântica, está sendo preparado para expandir sua funcionalidade tecnológica de recurso terminológico para um nível de espaço conceitual do conhecimento agropecuário brasileiro. Para que alcance tal potencial, o Agrotermos está sendo trabalhado organizacionalmente, por um grupo de trabalho permanente da Embrapa, o Gtermos, responsável por sua concepção, curadoria e gestão, dentro do contexto da Política de Governança de Dados e Informação para Conhecimento, já implantada na Embrapa e que contribui para a intenção e para os esforços de trazer a agricultura digital para a realidade do setor agropecuário brasileiro.

3 Gestão de dados de pesquisa

As primeiras décadas do século XXI têm se caracterizado por um crescimento explosivo na capacidade humana de adquirir, armazenar e comunicar dados digitais. Na perspectiva científica, o conceito de "ciência intensiva em dados" ou, ainda, "e-Science" (Borgman, 2007; Gray, 2009), vem se consolidando como realidade em inúmeros campos do conhecimento, muitos deles relevantes à pesquisa agropecuária.

Avanços tecnológicos têm permitido maior precisão e cobertura na aquisição de dados. Alguns exemplos são aplicações da Internet das Coisas (IoT), que vem tornando o uso de sensores uma realidade no campo; as crescentes possibilidades de imageamento de áreas rurais com o uso de drones (também conhecidos como Veículos Aéreos Não Tripulados - VANTs); outras aplicações de geotecnologias; e a evolução das áreas da bioinformática e da nanotecnologia.

A conjuntura atual também vem sendo chamada de Era do "Big Data", caracterizada pelos 5 Vs: Volume, Velocidade, Variedade, Veracidade e Valor (Mcafee; Brynjolfsson, 2012). Essa grande quantidade de dados, para ser útil, deve ser bem gerida, recuperável e passível de acesso, compreensão e integração. Esse cenário tem provocado transformações na maneira como dados, informações e conhecimento são criados e utilizados: para lidar de modo inteligente e ágil com o "dilúvio" de dados, novas habilidades são requeridas, de modo a garantir a preservação, a integração e o reuso dos dados.

Gestão de dados de pesquisa (GDP) é uma disciplina que reúne um conjunto de atividades consideradas essenciais ao planejamento, implantação e execução de estratégias, procedimentos e práticas voltadas ao gerenciamento efetivo de dados. Existem várias abordagens para se compreender a GDP; uma delas remete a diferentes concepções de modelos de ciclo de vida da

gestão do dado (CVGD), os quais fornecem uma visão da dinâmica do dado, desde a sua geração até o seu reuso.

Ciclo de vida do dado é definido pelo DataONE (2020b) como uma representação em alto nível dos estágios envolvidos na gestão e na preservação dos dados, com vista ao seu uso e reuso (Figura 5). Neste capítulo, essa definição do DataONE é tomada como referência, em razão da adequabilidade a adaptações e da versatilidade de interpretação de suas definições e conceitos. Esse ciclo de vida é composto de oito etapas: planejar, coletar, assegurar a qualidade, descrever, preservar, descobrir, integrar e analisar, as quais são descritas sucintamente, com base em Sayão e Sales (2015), Strasser et al. (2015), Araújo et al. (2019) e DataONE (2020b). Essas etapas envolvem ações cíclicas que possibilitam: a) construção de um plano de gestão dos dados voltado ao atendimento da política de dados da instituição de pesquisa; b) coleta dos dados visando a garantia da sua usabilidade e reutilização a longo prazo; c) garantia assegurada aos conjuntos de dados para que possam ser usados e sejam reprodutíveis; d) descrição dos dados de forma precisa e detalhada, adotando-se padrão de metadados, taxonomias e vocabulários controlados; e) preservação dos dados mediante armazenamento apropriado em data centers, com vistas à garantia de interoperabilidade, recuperação e busca; f) descoberta dos dados potencialmente úteis, que, tendo sido descritos por metadados, podem ser facilmente encontrados; g) integração dos dados provenientes de fontes distintas, que, combinados, geram novos conjuntos de dados passíveis de uso; h) análise dos dados propiciada pelos novos conjuntos de dados gerados na integração, com o intuito de fornecer informações relevantes para futuras pesquisas.

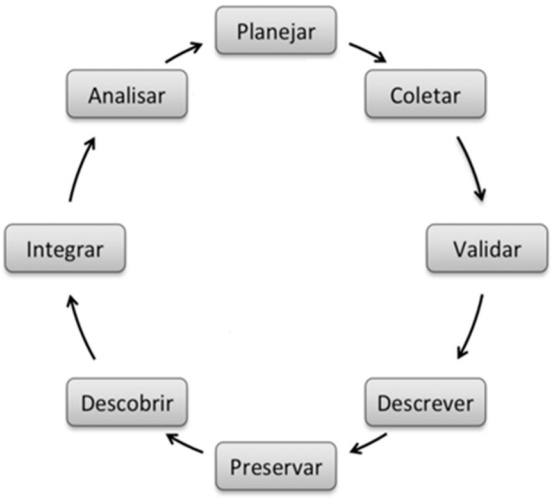

Figura 5. Ciclo de Vida de Dados.

Fonte: Adaptado de DataOne (2020b).

A GDP baseada no ciclo de vida do dado remete para a necessidade de adoção de melhores práticas, que são definidas como "[…] métodos ou enfoques que são reconhecidos por uma comunidade como sendo corretos ou mais apropriados para aquisição, gerenciamento, análise e compartilhamento de dados" (Sayão; Sales, 2015, p. 81). Tais práticas devem orientar as pessoas sobre como trabalhar efetivamente com seus dados em cada etapa de seu ciclo de vida, auxiliando, assim, no mapeamento dos processos envolvidos no CVGD (DataONE, 2020a). Como exemplo de melhores práticas como as preconizadas pelo DataONE, na etapa planejar, recomenda-se: criação, gestão e documentação dos dados; definição dos tipos e formato de dados a serem produzidos etc. Para a etapa descrever, por exemplo, recomenda-se: atribuir nomes a arquivos de modo que descrevam e reflitam o seu conteúdo; descrever formato para localização geoespacial e temporal do dado; adotar taxonomias padrões para descrição de quaisquer conjuntos de dados; adotar vocabulário controlado especializado; definir conjunto de elementos metadados etc.

No entanto, a recomendação de melhores práticas, por si só, não garante eficiência e eficácia na gestão de dados e, de acordo com Veiga (2019, p. 15), "não basta compartilhar dados, eles precisam ser FAIR". Faz-se necessário, pois, associar tais melhores práticas aos princípios FAIR (acrônimo inglês para as palavras *Findable*, *Accessible*, *Interoperable*, *Reusable*), para que dados de pesquisa, além de serem bem geridos, sejam também passíveis de se tornarem encontráveis, acessíveis, interoperáveis e reusáveis (Wilkinson et al., 2016). Os quatro princípios FAIR são compostos de 15 elementos que contribuem para complementar e enriquecer as informações essenciais às oito etapas do ciclo de vida do dado, ampliando as possibilidades de localização, acesso, interoperabilidade e reuso.

Os princípios FAIR aplicam-se a quaisquer objetos de pesquisa, para que se tornem disponíveis e entendíveis a humanos e a máquinas, garantindo transparência, reprodutibilidade e reutilização, além de proporcionarem a devida e adequada citação das informações geradas pela ciência intensiva em dados (Wilkinson et al., 2016). Os princípios FAIR, inclusive, nortearam a concepção da Política de Governança de Dados, Informação e Conhecimento da Embrapa (Embrapa, 2019), de modo a aproximar pesquisadores e demais sujeitos da pesquisa aos conjuntos de dados disponíveis em repositórios e plataformas de dados.

3.1 Metadados e catalogação de dados

O objeto de trabalho da catalogação e dos metadados, neste tópico, é essencialmente o dado. Dados é uma palavra que possui vários significados, abrangendo os genéricos e os especializados, a depender do contexto em que é empregada. De acordo com Semeler e Pinto (2019, p. 113), "[…] dados

significa uma peça única de informação" e, por sua vez, "os dados de pesquisa são o resultado de qualquer investigação sistemática que envolva processos de observação, experimentação ou simulação de procedimentos de pesquisa científica".

Os metadados e a catalogação de dados são necessários para que os dados de pesquisa possam ser "[…] identificáveis, citáveis, visíveis, recuperáveis, interpretáveis, contextualizáveis, interoperáveis e reutilizáveis onde quesitos de consistência e procedência são considerados" (Semeler; Pinto, 2019, p. 116). Cabe destacar, ainda, a necessidade de se considerar o contexto da gestão de dados de pesquisa e do ciclo de vida dos dados nos quais se inserem os metadados e a catalogação de dados, que além de serem duas subetapas importantes da etapa descrição, devem estar alinhadas aos princípios FAIR.

Diante da necessidade de se ampliarem os mecanismos de representação de dados e informações para melhor gerenciá-los, e da consequente complexidade que envolve a definição de seus atributos, metadados não podem mais ser definidos como sendo tão somente "dados sobre os dados". Na atualidade, tal definição é considerada uma expressão que não ajuda no entendimento do que significa exatamente metadados (Sayão; Sales, 2015). O que expande esse entendimento e amplia o seu domínio de aplicação é a definição dada por Riley (2017, p. 1), ao considerar metadados como sendo a informação que criamos, armazenamos e compartilhamos para descrever as coisas, e que nos permite interagir com essas coisas para obter o conhecimento necessário.

Metadados possibilitam a exploração de outras dimensões e facetas do dado, que ao serem reveladas pela catalogação passam a contribuir para a melhoria da gestão e da qualidade, favorecendo a descoberta das coleções de dados para a comunidade científica. Tais dimensões trazem à tona a necessidade de criação de novos elementos metadados, capazes de ampliar e enriquecer o esquema de metadados adotado. Metadados são indispensáveis para que, no futuro, os conteúdos digitais possam ser acessados e interpretados. Sem metadados, de acordo com Gray (2009), citado por Sayão e Sales (2016, slide 83), os usuários

> […] não saberão os detalhes de como os dados foram obtidos e preparados: 1) como os instrumentos foram projetados e construídos; 2) quando, onde e como os dados foram coletados; e, 3) não terão uma descrição dos processos que levaram aos dados derivados, que são tipicamente usados para análises científicas.

Metadados também são indispensáveis à interoperabilidade técnica e semântica, ou seja, sem eles os repositórios e as plataformas de dados não poderão intercambiar dados e informações. Metadados são constituídos por elementos descritivos bem definidos, por exemplo: autor, título, descrição,

assunto, palavra-chave, identificador, produtor, tipos de dados, condições de acesso, termo de uso das coleções etc., formando, a partir da catalogação de dados, um corpo de informações capaz de contextualizar os dados quanto à proveniência, à história, à natureza, ao propósito e a outros aspectos.

A adoção de metadados enriquecidos traz benefícios diretos para a gestão de dados, impactando positivamente no arquivamento e na preservação, bem como na interoperabilidade e na recuperação de conjuntos de dados de pesquisa. Dados somente serão úteis para análise se tiverem sido descritos por metadados de qualidade, e para que isso aconteça, a melhor recomendação é adotar os princípios FAIR ao catalogá-los.

Ainda no que se refere a metadados, e de acordo com Veiga (2019, p. 18-22), faz-se necessário destacar, sucintamente, os elementos-chave que devem orientar a adoção dos princípios FAIR, especialmente quanto ao aspecto descritivo dos metadados, quanto a: "a) elementos metadados para identificadores únicos e persistentes tanto para dados como para o conjunto de dados; b) conjunto de dados utilizando metadados enriquecidos com pluralidade de atributos precisos e relevantes; c) elemento metadado que indique clara e explicitamente os identificadores persistentes, tanto do conjunto de dados como também do próprio metadado nos repositórios e plataformas de dados; e) metadados registrados ou indexados em recursos de identificação que ofereçam capacidade de busca; f) metadados utilizando protocolos padronizados de comunicação para facilitar a recuperação dos dados via metadados, inclusive; g) disponibilidade de metadados para acesso, mesmo que os dados não estejam acessíveis e disponíveis; h) elemento metadado para representação do conhecimento por meio de linguagem formal e pelo uso de taxonomias e vocabulários controlados de acordo com os princípios FAIR, especializados e padronizados por área específica do domínio; i) elemento metadado para referências qualificadas de conjunto de dados e outros objetos de pesquisa derivados, que se interconectam, assegurando interligações semânticas entre eles, e que sejam linkáveis para outros conjuntos de dados; j) metadados com riqueza de atributos com alto nível de detalhes para que permitam ao pesquisador avaliar a possibilidade de reuso e adequação às suas necessidades; k) elemento metadado com informações inequívocas, definindo claramente quem pode ter acesso aos dados, com que finalidade e sob quais condições; l) elemento metadado que especifique a proveniência dos dados, o que subsidiará o pesquisador ao decidir sobre a utilidade dos dados ou metadados e ao atribuir crédito ao produtor dos dados; m) adoção de metadados deve estar alinhada com padrões relevantes e específicos da comunidade e da área de pesquisa".

No âmbito da agricultura digital, metadados descritos de acordo com os princípios FAIR contribuirão diretamente para a descoberta e o reuso dos dados por outros pesquisadores e instituições de pesquisa.

3.2 Plataforma Dataverse

O relatório intitulado "Acesso aberto a dados de pesquisa no Brasil: soluções tecnológicas - relatório 2018" (Rocha, 2018) apresenta os resultados do projeto de pesquisa Rede de Dados de Pesquisa Brasileira (RDP Brasil), que identifica, explora e analisa em profundidade três soluções tecnológicas (Dataverse, DSpace e CKAN) para a construção de repositório de Acesso Aberto a Dados de Pesquisa (AADP).

Com base no modelo OASIS (Open Access and Scholarly Information System) – composto por 56 critérios, classificados em Representação do Ambiente do Repositório, Representação dos Conjuntos de Dados, Descrição e Documentação dos Conjuntos de Dados, Produção dos Conjuntos de Dados, Armazenamento a Longo Prazo e Planejamento da Preservação, Acesso e Uso dos Conjuntos de Dados e Uso, Desenvolvimento e Manutenção do Software –, conclui-se que as tecnologias Dataverse e DSpace possuem recursos para configuração de vários tipos de repositório de dados, incluindo hierarquias organizacionais, temáticas e políticas de dados distintas para grupos ou unidades de pesquisa, com esquemas de metadados e suporte a licenças de uso. Por sua vez, o software CKAN é uma boa alternativa quando usado como serviço de publicação e de acesso, com a submissão e a preservação digital sendo realizadas por outros ambientes de repositório.

A Plataforma Dataverse é um software livre para armazenamento, publicação e compartilhamento de dados (Dataverse, 2020b). Traz facilidades para representar cenários que são compostos por diversas entidades hierárquicas, como universidades, instituições, laboratórios, grupos de pesquisa e departamentos, provendo autonomia para implementar os detalhes da gestão de dados, como a definição de quem pode criar, autorizar a publicação ou acessar conjuntos de dados, estabelecer licenças e definir que o uso dos dados somente pode ser feito mediante solicitação.

A plataforma utiliza esquemas de metadados (compatíveis com DDI Lite, DDI Codebook, Dublin Core, DataCite, VORResource, ISA-Tab), gerencia versões de conjuntos de dados, identifica unicamente conjuntos de dados (considerando versões) de forma universal e persistente (DOI ou Handle System), disponibiliza metadados de citação e uma estrutura para citação que envolve a verificação da imutabilidade do material citado. Também viabiliza o armazenamento de documentos complementares junto a conjunto de dados, adiciona ferramentas de visualização e exploração de dados, permite a customização de sua interface, bem como proporciona a colheita de metadados a partir do protocolo Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH).

Funcionalidades adicionais podem ser incorporadas, como o suporte ao armazenamento de grandes volumes de dados; o suporte à visualização e à exploração de dados; a melhoria no mecanismo de indexação e busca e nos sistemas de autenticação de usuários.

3.3 Metadados na Plataforma Dataverse: a experiência da Embrapa Informática Agropecuária

Como dito anteriormente, a Plataforma Dataverse é uma aplicação web, dedicada para o compartilhamento, a preservação, a citação, a exploração e a análise de dados de pesquisa, que hospeda vários dataverses compostos por conjuntos de dados, os quais são tratados por meio de metadados (Dataverse, 2020b). A Plataforma Dataverse foi escolhida pela Embrapa Informática Agropecuária para apoiar na gestão dos conjuntos de dados de pesquisa, tendo como projeto-piloto o Laboratório Multiusuário de Bioinformática (LMB) (Embrapa Informática Agropecuária, 2020). Portanto, é nesse ambiente da Plataforma Dataverse dedicado ao LMB que ocorre a experiência relatada no tocante aos metadados (Dataverse, 2020a).

Assim como ocorre em diversas plataformas e repositórios de dados, os conjuntos de dados (*datasets*) na Plataforma Dataverse são inseridos utilizando-se um formulário para cadastro, o qual contém inúmeros campos, correspondentes a elementos metadados. Alguns desses campos são de preenchimento obrigatório, mas a maioria é opcional. Além disso existem conjuntos adicionais de metadados que podem ser adicionados, para domínios específicos de dados. Os conjuntos de metadados seguem padrões estabelecidos, garantindo a interoperabilidade com outras plataformas.

Tomando por base os princípios FAIR, foram conduzidos estudos e testes, envolvendo os usuários e os produtores dos dados do projeto-piloto do LMB, em busca de definição de elementos (termos) metadados para atender às especificidades dos usuários do projeto-piloto. Desse trabalho obtiveram-se definições quanto a: a) metadados básicos, sendo alguns obrigatórios, bem como metadados complementares, não obrigatórios; b) ferramentas para descrição de metadados: normas de descrição, taxonomia, tesauros e vocabulários controlados. Essas definições são importantes para a geração de metadados enriquecidos, que são aqueles que utilizam ferramentas de descrição, como taxonomia, tesauros e vocabulário controlado, específicas do domínio, associadas às informações de proveniência, trazendo clareza semântica quando comparados ao metadado (original) básico. De acordo com Lira (2014, p. 43):

> Um conjunto de metadados enriquecidos deve apresentar algumas características como: (i) maior quantidade de atributos semânticos…; (ii) facilidade de interpretação e processamento do conteúdo dos datasets; (iii) termos de vocabulários padrões associados […].

Os metadados na Plataforma Dataverse, a partir da experiência com criação de *dataverses* e *datasets* para gestão de dados de pesquisa no LMB, estão

espelhados nos princípios FAIR, sobretudo para torná-los ricos em destaques e com pluralidade de atributos precisos e relevantes.

3.4 Catalogação de dados na Plataforma Dataverse

O processo de catalogação dos conjuntos de dados do projeto-piloto do LMB inicia-se com a atividade de autodepósito, a qual é feita pelo pesquisador, o proprietário do dataset ou outra pessoa designada para tal. O momento do autodepósito do *dataverse* ou *dataset* na Plataforma Dataverse corresponde à fase de pré-catalogação, na qual são preenchidos os campos correspondentes aos elementos metadados básicos, sendo obrigatório informar: título, autor, contato, descrição e assunto.

Na fase seguinte ocorre a catalogação propriamente dita, começando pela revisão do preenchimento prévio dos conteúdos relativos aos elementos metadados básicos. A descrição catalográfica dos *dataverse* e *dataset* ocupa-se essencialmente do preenchimento dos elementos metadados complementares, a ser feito pelo especialista do domínio, preferencialmente pelo proprietário do *dataset*. No entanto, essa atividade requer conhecimento de normas e padrões da biblioteconomia, emprestando maior significado e qualidade aos conjuntos de dados.

O emprego de técnicas de catalogação para descrever e representar dados ou quaisquer objetos de pesquisa, tomando por base um conjunto estruturado de elementos metadados FAIR, é essencial para garantir a interoperabilidade técnica e semântica, o compartilhamento, o uso e o reuso dos dados de pesquisa, cabendo a responsabilidade técnica ao bibliotecário e/ou ao cientista da informação.

4 Qualidade de dados de pesquisa

O alto padrão de qualidade no arquivamento e na manutenção de dados é amplamente reconhecido por diversos segmentos da sociedade. Na agricultura digital, a qualidade dos dados torna-se particularmente importante para que o processo de tomada de decisões, as atividades de planejamento e outras possam ter um elevado nível de assertividade. Com base nessa constatação, e no impacto causado nos diferentes tipos de negócios, o tema Qualidade de Dados (DQ, do inglês Data Quality) é visto como um pilar forte no processo de gestão de dados. Isso tem despertado, cada vez mais, o interesse de pesquisadores de diferentes áreas do conhecimento em investigar e aprofundar os estudos sobre o tema.

Como apresentado na literatura, há diferentes definições para DQ. Essas variações estão relacionadas ao contexto em que DQ está sendo discutido e ao grau de necessidade com que o usuário trata a qualidade. Importante destacar

a definição do Data Management Body Knowledge (DMBOK) para DQ: "[…] o planejamento, implementação e controle de atividades que aplicam técnicas de gerenciamento de qualidade de dados, a fim de garantir que sejam adequados ao consumidor e atendam às necessidades dos consumidores de dados" (Knight, 2017, p. 1 - tradução nossa).

Para uma organização, seja ela pública ou privada, voltada para negócios ou pesquisas, o arquivamento, a manutenção e a recuperação de dados de alta qualidade trazem oportunidades para a formulação de melhores estratégias de negócio, facilidades para que tomadas de decisões tenham o maior nível de acerto, e melhores condições de obtenção de vantagem competitiva. A falta desse nível de qualidade, além de dificultar a obtenção das vantagens mencionadas, contribui para a elevação dos custos de processamento dos dados. Além disso, diminui o nível de satisfação do cliente, quando ele recebe o resultado de um serviço ou uma informação requerida, como resposta a ação de pesquisa (Jaya et al., 2017).

O tema QD não deve ser tratado de forma independente, porque dados de baixa qualidade levam a conclusões enganosas e caras, o que pode levar à desconfiança ou à perda de credibilidade da equipe de dados. Problemas encontrados na cultura organizacional de uma instituição também afetam o processo de coleta e a qualidade dos dados.

A QD pode ser desenvolvida sob as abordagens qualitativa e quantitativa (Vancauwenbergh, 2019); na qualitativa, categorias (por exemplo, aferição, contextualização, representação e acesso) são estabelecidas e dimensões/características são associadas a cada uma delas. Por sua vez, na abordagem quantitativa, a DQ está pautada na adequação dos dados para servir a um propósito em um determinado contexto, isto é, em operações de tomadas de decisões e/ou de planejamento.

4.1 Medidas para avaliação da qualidade de dados

A avaliação da qualidade dos dados é um processo crucial dentro do gerenciamento da qualidade dos dados, por compreender diferentes etapas envolvendo vários grupos de pessoas de uma organização. O objetivo desse tipo de avaliação é identificar dados contendo algum tipo de erro e medir o impacto de vários processos de negócios orientados por dados.

O processo pode ser iniciado na fase de coleta de dados, incluindo diretrizes mínimas para garantia da qualidade, contribuindo para a redução da quantidade de trabalho a ser executada na fase de qualificação/preparação de dados (Pyle, 1999) e propiciando, assim, melhores condições para a realização das análises de dados.

A DQ pode ser avaliada utilizando-se medidas subjetivas e/ou estabelecidas por cálculos computacionais. Na prática, quando um processo de avaliação da qualidade dos dados é iniciado, medidas básicas, tais como número

de dados faltantes, quantidade de dados apresentando erro de tipografia, quantidade de dados fora do padrão, medidas estatísticas básicas e análises visuais, são utilizadas. A adoção dessas medidas dá uma noção preliminar de quão bom é o nível da qualidade dos dados. Em algumas situações, o uso de uma única medida não é suficiente para o alcance do resultado desejado, daí a utilização, em conjunto, de outras medidas.

Além das medidas mencionadas para a avaliação da qualidade dos dados, outras são descritas por Cichy e Rass (2019). Elas estão relacionadas a *frameworks* disponíveis para o estabelecimento desse tipo de avaliação. O emprego dessas medidas ocorre quando há interesse em aprofundar o nível da avaliação. Quanto maior se apresenta o nível da qualidade, maior será o nível de acurácia dos resultados gerados, o que traz maiores possibilidades para tomadas de decisões mais assertivas. Outra contribuição importante de se ter dados de qualidade é contribuir para a descoberta de conhecimentos em base de dados, conhecimentos estes que estão inseridos nos dados e não visualizados, num primeiro momento, pelo usuário (Fayyad et al., 1996).

4.2 Gestão da qualidade de dados - DQM

A DQ é um dos problemas cruciais para se medir e analisar ciência, tecnologia e inovação corretamente, o que permite o monitoramento adequado da eficiência da pesquisa, da produtividade e até da tomada de decisões estratégicas (Fan; Geerts, 2017).

Normalmente, dados apresentam algum tipo de inconsistência – alguns estão duplicados, incompletos, imprecisos e obsoletos. Para torná-los aptos a produzirem resultados de qualidade, após processados e analisados, lança-se mão do processo de Gestão da Qualidade de Dados (DQM, do inglês Data Quality Management).

O objetivo principal de DQM é remover todos e quaisquer problemas encontrados, elevando a qualidade dos dados e permitindo que eles contribuam para a adição de valor aos processos de negócio e/ou produzam respostas qualificadas para as questões endereçadas (Vancauwenbergh, 2019).

A DQM envolve a execução de várias tarefas, definições de parâmetros e associações de valores a eles e estabelecimento de um fluxo de trabalho. Tudo isso deve ser registrado, atualizado e recuperado com facilidade. Para suportar todo esse trabalho, diferentes *frameworks* encontram-se disponíveis, com destaque para: DAMA DMBOK´s Data Governance Model (Barbieri, 2013); EWSolutions' EIM Maturity Model (Smith, 2009); e Oracle's Data Quality Management Process (Oracle, 2009).

Esses modelos estão centrados em três elementos básicos, que são os metadados associados aos dados, os processos voltados para registro, organização e (re)uso de dados, e o contexto organizacional em relação aos dados. A qualidade de cada elemento e a interação entre eles, em última análise,

determinam a qualidade e, portanto, o verdadeiro valor do patrimônio de dados de uma organização. A permissão para descrever metadados compreensíveis por toda a organização e alinhados com os processos, as estratégias e os objetivos de negócios dessa organização é recurso disponível nesses modelos. Além disso, tais modelos fornecem meios para relatar os fatores críticos de sucesso, elementos úteis para o desenvolvimento de estratégias eficazes de gerenciamento de DQ.

4.3 Fatores críticos de sucesso para implantação de Gestão de Qualidade de Dados

De acordo com Milosevic e Patanakul (2005, p. 183), fatores críticos de sucesso (CFS) são "[…] características, condições ou variáveis que podem ter um impacto significante no sucesso de uma organização ou de um projeto quando adequadamente sustentado, mantido ou gerenciado". Santos (2015) apresenta 20 CFS aplicáveis à DQM que, na visão de Milosevic e Patanakul (2005), formam quatro grandes grupos: a) operacional; b) gerenciamento; c) governança; e d) capacitação. O grupo operacional foca nos processos operacionais envolvidos em coleta, armazenamento, análise e segurança dos dados, todos eles altamente interdependentes. O grupo gerenciamento congrega os processos gerenciais, que são oriundos do grupo operacional e que visam, principalmente, o alinhamento da qualidade dos dados com as metas da organização em relação aos dados e aos resultados das análises de dados. O terceiro grupo, governança, abrange os processos de governança associados à DQM. Esses processos podem ser apresentados pela alta gerência da organização como compromisso prioritário para a implantação de DQM, estimulando uma mudança de cultura em toda a organização voltada para esse tópico. Por último, o grupo de capacitação é considerado primordial para investir em um programa de DQM, mesmo que a empresa possua uma estrutura para capacitar seus colaboradores nas ações de caráter operacional, de gerenciamento e de governança. O objetivo principal desse grupo é informar as pessoas sobre a importância dos dados qualitativos para a organização. Além do treinamento visando a implementação sistemática do DQ, em toda a organização, uma ação de acompanhamento contínuo das capacitações deve ser instituída. Isso permitirá ajustes rápidos, em casos de identificação de erros, e adequações nas regras de negócios.

5 Considerações finais

O capítulo teve como objetivo apresentar um relato das ações de pesquisa e dos resultados que foram executados e que estão sendo desenvolvidos na Embrapa Informática Agropecuária, no contexto da Engenharia da

Informação. Com isso, pretende-se alinhar esse trabalho às frentes de atuação da agricultura digital, bem como agregar valor à pragmática do conhecimento científico, oferecendo tecnologias mais facilmente percebidas e assimiláveis por seus potenciais usuários.

Essas ações, em curso no Grupo de Pesquisa de Engenharia da Informação, acontecem no domínio de três naturezas da informação (cognitiva, documentária e comunicativa) e por meio de artefatos computacionais que operacionalizam os processos constituintes dos ciclos de vida dos dados, da informação e do conhecimento (D-I-C). A Ciência da Computação é a principal fonte geradora dessas ações, conseguindo conjugar e complementar aportes metodológicos e tecnológicos originados em outros domínios de conhecimento, produzindo desdobramentos inter, multi e transdisciplinares com a Agronomia, a Ecologia, a Matemática, a Economia, a Sociologia e toda a gama de interseções imagináveis. Inserida e articulada nesse universo de interações entre diferentes áreas de conhecimento, a Engenharia da Informação apresenta-se como uma alternativa para a operacionalização de estratégias, visando maior alinhamento entre as ações desenvolvidas nas áreas de P&D e o processo de inovação da Embrapa.

A Embrapa Informática Agropecuária, ao inaugurar a linha de pesquisa Engenharia da Informação, reorganiza, orienta e resgata suas competências na direção de um reposicionamento de suas ações de PD&I, considerando a perspectiva do processo de inovação implantado na empresa. Em especial, no tocante ao enfrentamento dos atuais desafios de pesquisa para a consolidação da transformação digital na agricultura, a Engenharia de Informação apresenta-se capaz de contribuir de forma efetiva com aportes conceituais, metodológicos, processuais e, sobretudo, no suporte ao desenvolvimento de artefatos, objetos e ferramentas computacionais com qualidade, segundo um processo de engenharia desenhado dentro de uma concepção ontológica pluralista. Em outras palavras, a Engenharia da Informação amplia as chances de os diversos atores que circunscrevem o fenômeno "agricultura brasileira" perceberem-na como um objeto que admite múltiplas explicações, suposições, métodos, metodologias, abordagens, teorias etc. Ademais, os esforços e as iniciativas em Engenharia da Informação, envidados até o presente momento, mostram-se alinhados às tendências e às oportunidades contemporâneas de desenvolvimento e aplicações computacionais e TIC na agricultura digital.

A partir das heurísticas viabilizadas pela Engenharia da Informação, os artefatos computacionais de representação de D-I-C podem ser apreendidos para além de suas funcionalidades imediatas (Pierozzi Junior et al., 2018). No entanto, as contribuições da Engenharia da Informação podem ser traduzidas e materializadas sob diferentes perspectivas e exemplificadas na forma de repositórios e bancos de dados que viabilizam o trabalho colaborativo; na catalogação, indexação e recuperação inteligente de informação; na mineração,

uso, reuso e gestão de dados; na ressignificação da informação e interoperabilidade com outros sistemas; na descoberta de conhecimento; no acesso facilitado e controlado, comunicação, compartilhamento, aprendizagem e inteligência coletiva. Uma vez que a agricultura digital é fundamentalmente baseada em conteúdo digital, a partir de dados obtidos por meio da Internet das Coisas, preconiza-se que a Engenharia da Informação promoverá facilidades e melhorias na construção de artefatos computacionais que atendam interesses de usuários em diferentes segmentos da agropecuária brasileira

Outra leitura é possível: a) quando o **dado** é trabalhado nas perspectivas de classificação, significação e acesso, diz-se que o que estão sendo trabalhadas são suas propriedades **cognitivas**; b) quando a **informação** é trabalhada nas perspectivas de catalogação, indexação e recuperação, diz-se que o que estão sendo trabalhadas são suas propriedades **documentárias**; c) quando o **conhecimento** é trabalhado nas perspectivas de visualização ou linguagens de máquina, diz-se que o que estão sendo trabalhadas são suas propriedades de **comunicação** (disseminação). Além dessas propriedades, devem ser consideradas aquelas trabalhadas e herdadas dos níveis precedentes de dados e de informação, respectivamente. O resultado disso é um movimento cíclico e contínuo de retroalimentação, que ocorre quando o conhecimento comunicado retorna como insight para uma nova rodada dos ciclos dos dados e da informação.

6 Referências

AALST, W. van der. **Processing mining**: data science in action. Berlin: Springer-Verlag, 2016. 467 p. DOI: 10.1007/978-3-662-49851-4.

ARAÚJO, D. G. de; ALMEIDA LLARENA, M. A.; SIEBRA, S. de A.; DIAS, G. A. Contribuições para a gestão de dados científicos: análise comparativa entre modelos de ciclo de vida dos dados. **Liinc em Revista**, v. 15, n. 2, p. 32-51, nov. 2019. DOI: 10.18617/liinc.v15i2.4686.

BARBIERI, C. **Uma visão sintética e comentada do Data Management Body of Knowledge (DMBOK)**. Belo Horizonte: Fumsoft, 2013. 46 p.

BAUM, C.; MARASCHIN, C.; MARKUART, E. N. Política ontológica como abordagem para as relações intercientíficas. **Psicología, Conocimiento y Sociedad**, v. 9, n. 2, p. 8-30, nov. 2019; abr. 2020. Disponível em: http://www.scielo.edu.uy/pdf/pcs/v9n2/1688-7026-pcs-9-02-6.pdf. Acesso em: 16 maio 2020.

BORGMAN, C. L. **Scholarship in the digital age**: information, infrastructure and internet. Cambridge, MA: MIT Press, 2007. DOI: 10.7551/mitpress/7434.001.0001.

CARLAN, E.; MEDEIROS, M. B. B. Sistemas de organização do conhecimento na visão da Ciência da Informação. **Revista Ibero-Americana de Ciência da Informação**, v. 4, n. 2, p. 53-73, ago./dez. 2011. DOI: 10.26512/rici.v4.n2.2011.1675.

CICHY, C.; RASS, S. An overview of data quality frameworks. **IEEE Access**, v. 7, p. 24634-24648, Feb 2019. DOI: 10.1109/ACCESS.2019.2899751.

DAHLBERG, I. Teoria do conceito. **Ciência da Informação**, v. 7, n. 2, p. 101-107, dez. 1978. Disponível em: http://revista.ibict.br/ciinf/article/view/115/115. Acesso em: 19 maio 2020.

DATAONE. **Best practices**. Albuquerque, NM: University of New Mexico, 2020a. Disponível em: https://www.dataone.org/best-practices. Acesso em: 27 maio 2020.

DATAONE. **Data life cycle**. Albuquerque, NM: University of New Mexico, 2020b. Disponível em: https://www.dataone.org/data-life-cycle. Acesso em: 2 maio 2020.

DATAVERSE. **GenClima**. Campinas: Embrapa Informática Agropecuária, 2020a. Disponível em: https://www.dataverse-h.cnptia.embrapa.br/dataverse/umip. Acesso em: 2 maio 2020.

DATAVERSE. **Harvard Dataverse**. Cambridge, MA: Harvard College, 2020b. Disponível em: https://dataverse.harvard.edu/. Acesso em: 02 maio 2020.

DEFOURNY, V. Apresentação. In: TARAPANOFF, K. (org.). **Inteligência, informação e conhecimento em corporações**. Brasília, DF; Ibict: Unesco, 2006. p. 7.

EMBRAPA INFORMÁTICA AGROPECUÁRIA. **Laboratório multiusuário de bioinformática**. Campinas, 2020. Disponível em: https://www.embrapa.br/informatica-agropecuaria/lmb. Acesso em: 2 jun. 2020.

EMBRAPA. Política de Governança de Dados, Informação e Conhecimento da Embrapa. **Boletim de Comunicações Administrativas**, v. 45, n. 16, p. 1-19, abr. 2019. 19 p. (Manual de normas da Embrapa).

EMBRAPA. **Visão 2030**: o futuro da agricultura brasileira. Brasília, DF, 2018. 212 p. Disponível em: https://www.embrapa.br/visao/o-futuro-da-agricultura-brasileira. Acesso em: 15 maio 2020.

FAN, W.; GEERTS, F. Foundations of data quality management. **Synthesis Lectures on Data Management**, v. 4, n. 5. p. 1-227, July 2017. DOI: 10.2200/S00439ED1V01Y201207DTM030.

FAO. **AGROVOC**. Rome: FAO-AIMS, 2020. Disponível em: http://aims.fao.org/vest-registry/vocabularies/agrovoc. Acesso em: 20 maio 2020.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37-54, fall 1996.

GARCIA, A. E. B.; SALLES FILHO, S. L. M. Trajetória institucional de um instituto público de pesquisa: o caso do Ital após 1995. **Revista de Administração Pública**, v. 43, n. 3, p. 661-693, maio/jun. 2009. DOI: 10.1590/S0034-76122009000300007.

GRAY, J. Jim Gray on eScience: a transformed scientific method. In: HEY, T.; TANSLEY, S.; TOLLE, K. (ed.). **The fourth paradigm**: data-intensive scientific discovery. Redmond, WA: Microsoft Research, 2009. p. xvii-xxxi.

JAYA, I.; SIDI, F.; ISHAK, I.; AFFENDEY, L. S.; JABAR, M. A. A review of data quality research in achieving high data quality within organization. **Journal of Theoretical and Applied Information Technology**, v. 95, n. 12, p. 2647-2657, 2017.

KNIGHT, M. What is data quality? In: DATAVERSITY. **Dataversity.net**. Studio City, CA; Dataversity Education, 2017. Disponível em: https://www.dataversity.net/what-is-data-quality/. Acesso em: 19 maio 2020.

LATOUR, B. Reagregando o social: uma introdução à Teoria do Ator-Rede. Salvador: Edufba; Bauru: Edusc, 2012. 399 p. Resenha de: SEGATA J. **Ilha R. Antr.**, Florianópolis, v. 14, n. 2, p. 238-243, jul./dez. 2012. DOI: 10.5007/2175-8034.2012v14n1-2p238.

LIRA, M. A. B. de. **Uma abordagem para enriquecimento semântico de metadados para publicação de dados abertos**. 2014. 95 p. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal de Pernambuco, Centro de Informática, Recife.

MARTIN, J.; FINKELSTEIN, C. **Information engineering**. Englewood Cliffs, NJ: Prentice Hall, 1989.

MARTINS, G. K.; MORAES, J. B. E. Organização e representação do conhecimento: institucionalização como disciplina científica no âmbito da Ciência da Informação. In: ENCONTRO NACIONAL DE PESQUISA EM PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO - ENANCIB, 16., 2015, João Pessoa, PB. **Anais**... João Pessoa: UFPb, 2015. Disponível em: http://www.ufpb.br/evento/index.php/enancib2015/enancib2015/paper/viewFile/3162/1030. Acesso em: 16 maio 2020.

MAZZOCCHI, F. Knowledge organization system (KOS). In: HJORLAND, B.; GNOLI, C. (ed.). **Encyclopedia of Knowledge Organization**. Alberta: University of Alberta, 2019. Disponível em: http://www.isko.org/cyclo/kos. Acesso em: 16 maio 2020.

MCAFEE, A.; BRYNJOLFSSON, E. Big data: the management revolution. **Harvard Business Review**, v. 90, n. 10, p. 61-68, Oct 2012.

MENDONÇA, E. S. A organização e a representação do conhecimento no tempo. **Revista de Ciências Humanas**, n. 38, p. 277-94, out. 2005.

MILOSEVIC, D.; PATANAKUL, P. Standardized project management may increase development projects success. **International Journal of Project Management**, v. 23, n. 3, p. 181-192, Apr 2005. DOI: 10.1016/j.ijproman.2004.11.002.

MOL, A. Política ontológica: algumas ideias e várias perguntas. In: NUNES, A.; ROQUE, R. (ed.). **Objectos impuros**: experiências em estudos sobre a ciência. Porto: Ed. Afrontamento, 2008. p. 63-106.

ORACLE. **Oracle® Warehouse Builder**: user's guide - 11g release 1(11.1). Redwood City, CA, 2009. 764 p. Disponível em: https://docs.oracle.com/cd/B31080_01/doc/owb.102/b28223.pdf. Acesso em: 16 maio 2020.

PIEROZZI JUNIOR, I.; BERTIN, P. R. B.; MACHADO, C. R. de L.; SILVA, A. R. da. Towards semantic knowledge maps applications: modelling the ontological nature of data and information governance in a R&D organization. In: THOMAS, C. (ed.). **Ontology in information science**. Rijeka: InTech, 2017. p. 83-104. DOI: 10.5772/67978.

PORAT, M. U.; RUBIN, M. R. **The information economy**. Washington, DC: US Govt. Print Off, 1977.

PORDES, R.; PETRAVICK, D.; KRAMER, B.; OLSON, D.; LIVNY, M.; ROY, A.; AVERY, P.; BLACKBURN, K.; WENAUS, T. The open science grid. **Journal of Physics**: conference series, v. 78, p. 1-15, 2007. DOI: 10.1088/1742-6596/78/1/012057.

POWELL, W. W.; SNELLMAN, K. The knowledge economy. **Annual Review of Sociology**, v. 30, p. 199-220, Aug 2004. DOI: 10.1146/annurev.soc.29.010202.100037.

PYLE, D. **Data preparation for data mining**. San Francisco, CA: Morgan Kaufmann, 1999. 560 p.

RESEARCH DATA ALLIANCE. **Global Agricultural Concept Space (GACS)**. [S. l.]: European Commission; National Science Foundation, 2020. Disponível em: https://agrisemantics.org/#GACShome. Acesso em: 20 maio 2020.

RILEY, J. **Understanding metadata**: what is metadata, and what is it for? Bethesda, MD: NISO Press, 2017. 49 p. Disponível em: https://groups.niso.org/apps/group_public/download.php/17446/Understanding%20Metadata.pdf. Acesso em: 01 maio 2020.

ROCHA, R. P. da. (coord.). **Acesso aberto a dados de pesquisa no Brasil:** soluções tecnológicas - relatório 2018. Porto Alegre: UFRGS, 2018. 75 p. Disponível em: http://hdl.handle.net/10183/185126. Acesso em: 14 maio 2020.

SANTOS, M. P. da C. dos. **Fatores críticos de sucesso na gestão da qualidade dos dados**. 2015. 55 p. Dissertação (Mestrado em Gestão de Sistemas de Informação) – Lisbon School of Economics & Management, Lisboa, Portugal.

SAYÃO, L. F.; SALES, L. F. **Guia de gestão de dados de pesquisa para bibliotecários e pesquisadores**. Rio de Janeiro: CNEN, 2015. 93 p.

SAYÃO, L. F.; SALES, L. F. **Guia de gestão de dados de pesquisa**: [minicurso]. Rio de Janeiro: CNEN, 2016. 196 slides.

SEMELER, A. R.; PINTO, A. L. Os diferentes conceitos de dados de pesquisa na abordagem da biblioteconomia de dados. **Ciência da Informação**, v. 48, n. 1, p. 113-129, jan./abr. 2019.

SMITH, A. Enterprise information management maturity: data governance's role. **EIMInsight Magazine**, v. 3, n. 1, jan. 2009. Disponível em: http://www.eiminstitute.org/library/eimi-archives/volume-3-issue-1-january-2009-edition/EIM-Maturity. Acesso em: 16 maio 2020.

SOERGEL, D. Digital libraries and knowledge organization. In: KRUK, S. R.; McDANIEL, B. (ed.). **Semantic digital librarie**s. Berlin: Springer, 2009. p. 9-39. DOI: 10.1007/978-3-540-85434-0_2.

SOUZA, R. R.; TUDHOPE, D.; ALMEIDA, M. B. Towards a taxonomy of KOS: dimensions for classifying knowledge organization systems. **Knowledge Organization,** v. 39, n. 3, p. 179-192, 2012. DOI: 10.5771/0943-7444-2012-3-179.

STRASSER, C.; COOK, R.; MICHENER, W.; BUDDEN, A. **Primer on data management**: what you always wanted to know. [S. l.]: California Digital Library, 2015. 12 p. (DataONE best practices primer).

TUOMI, I. Data is more than knowledge: implications of the reversed knowledge hierarchy for knowledge management and organizational memory. **Journal of Management Information Systems**, v. 16, n. 3, p. 103-117, 1999. DOI: 10.1080/07421222.1999.11518258.

VANCAUWENBERGH, S. Data quality management. In: KUNOSIC, S.; ZEREM, E. (ed.). **Scientometrics recent advances**. London: Intechopen Limited, 2019. p. 1-15. DOI: 10.5772/intechopen.86819.

VEIGA, V. **Gestão de dados de pesquisa FAIR**: dando um JUMP em seus dados. In: ENCONTRO DA REDE SUDESTE DE REPOSITÓRIOS INSTITUCIONAIS, 1., 2019, Rio de Janeiro. **Anais**... Rio de Janeiro: Fiocruz/Icit/UFRJ, 2019. 59 p. Disponível em: https://www.arca.fiocruz.br/handle/icict/33343. Acesso em: 27 abr. 2020.

WILKINSON, M. D.; DUMONTIER, M.; AALBERSBERG, J. J.; APPLETON, G.; AXTON, M.; BAAK, A.; BLOMBERG, N.; BOITEN, J.-W.; SANTOS, L. B. da S.; BOURNE, P. E.; BOUWMAN, J.; BROOKES, A. J.; CLARK, T.; CROSAS, M.; DILLO, I.; DUMON, O.; EDMUNDS, S.; EVELO, C. T.; FINKERS, R.; GONZALEZ-BELTRAN, A.; GRAY, A. J. G.; GROTH, P.; GOBLE, C.; GRETHE, J. S.; HERINGA, J.; HOEN, P. A. C 't; HOOFT, R.; KUHN, T.; KOK, R.; KOK, J.; LUSHERM, S. J.; MARTONE, M. E.; MONS, A.; PACKER, A. L.; PERSSON, B.; ROCCA-SERRA, P.; ROOS, M.; SCHAIK, R. van; SANSONE, S.-A.; SCHULTES, E.; SENGSTAG, T.; SLATER, T.; STRAWN, G.; SWERTZ, M. A.; THOMPSON, M.; LEI, J. van der; MULLIGEN, E. van; VELTEROP, J.; WAAGMEESTER, A.; WITTENBURG, P.; WOLSTENCROFT, K.; ZHAO, J.; MONS, B. The FAIR guiding principles for scientific data management and stewardship. **Scientific Data**, v. 3, article number 160018, 2016. DOI: 10.1038/sdata.2016.18.

ZENG, M. L. Knowledge Organization Systems (KOS). **Knowledge Organization**, v. 35, n. 2-3, p. 160-182, 2008. DOI: 10.5771/0943-7444-2008-2-3-160.