



**CINCO ANOS DE GESTÃO DE DADOS ESPACIAIS NA EMBRAPA:
DIAGNÓSTICO E PERSPECTIVAS**

Ana Claudia Meciano **Maria**¹; Jaudete **Daltio**²; Cristina **Criscuolo**³

Nº 23509

RESUMO – Lançado em 2018 e construído a partir do uso de softwares livres, o GeolInfo é atualmente a Infraestrutura de Dados Espaciais institucional da Embrapa. O GeolInfo é responsável tanto pela catalogação e pelo armazenamento quanto pela disponibilização/publicação desses dados, a partir dos padrões internacionais de compartilhamento de dados estabelecidos pela Open Geospatial Consortium (OGC) e dos princípios Free Open Source (FOSS). Ele está disponível para diferentes públicos (interno e externo) e integra-se com a Infraestrutura Nacional de Dados Espaciais (Inde). O GeolInfo agrega dados de 26 Unidades Descentralizadas da Embrapa que são produtoras e usuárias de dados geográficos. Após cinco anos de implantação e cerca de 12.000 conjuntos de dados catalogados, o GeolInfo passa atualmente por um processo de avaliação e diagnóstico, que visa nortear o planejamento de ações futuras sobre sua evolução e manutenção. Este trabalho tem o objetivo de apresentar esse diagnóstico e suas implicações na proposição de uma nova arquitetura para o GeolInfo. Trata-se de análises feitas nos metadados dos dados catalogados na plataforma, com a finalidade de compreender seu perfil de utilização.

Palavras-chave: catálogo, diagnóstico, infraestrutura nacional de dados espaciais, metadados.

1 Autora, Bolsista Embrapa: Graduação em Geografia, Unicamp, Campinas-SP; ana.maria@colaborador.embrapa.br

2 Orientadora, Analista da Embrapa Territorial, Campinas-SP; jaudete.daltio@embrapa.br

3 Orientadora, Pesquisadora da Embrapa Territorial, Campinas-SP



ABSTRACT – Launched in 2018 and built using open-source software, GeolInfo is currently Embrapa's institutional Spatial Data Infrastructure. GeolInfo is responsible for cataloging, storing, and for the provision/publication of this data, based on international data sharing standards established by the Open Geospatial Consortium (OGC) and the Free Open Source Software (FOSS) principles. It is available to various audiences (internal and external) and integrates with the Brazilian National Spatial Data Infrastructure (Inde). GeolInfo aggregates data from 26 of Embrapa's Decentralized Units, which produce and use geographic data. Five years into its full implementation and around 12,000 catalogued datasets later, GeolInfo is currently undergoing an evaluation and diagnostic process aimed at guiding future planning for its evolution and maintenance. This work aims to present this diagnosis and its implications in the proposition of a new architecture for GeolInfo. It involves analyses conducted on the metadata of the data cataloged in the platform, with the purpose of understanding its usage profile.

Keywords: catalog, diagnosis, national spatial data infrastructure, metadata.

1. INTRODUÇÃO

Dados geográficos representam fatos, objetos e fenômenos associados a uma localização sobre a superfície terrestre, em um certo instante ou período de tempo (Câmara et al., 1996). Podem ser “de referência”, quando proporcionam informações genéricas e servem como base para a construção de outros dados, ou “temáticos”, quando se referem a um determinado fenômeno ou tema específico. Representam, hoje, uma parcela significativa dos dados disponíveis na web e, visando ampliar sua utilização e seu compartilhamento, vários países iniciaram a definição das chamadas Infraestrutura de Dados Espaciais (IDE) (Brasil, 2010).

No Brasil, o Decreto nº 6.666¹ institui a Infraestrutura Nacional de Dados Espaciais (Inde) como:

o conjunto integrado de tecnologias; políticas; mecanismos e procedimentos de coordenação e monitoramento; padrões e acordos, necessários para facilitar e ordenar a geração, o armazenamento, o acesso, o compartilhamento, a disseminação e o uso dos dados geoespaciais de origem federal, estadual, distrital e municipal (Brasil, 2010).

¹ DOU de 28/11/2008, p. 57.



Embora existam na literatura diversas definições para o termo “infraestrutura de dados espaciais”, a definição adotada no Brasil tem como componente central a interoperabilidade, traduzida como mecanismo de publicação, acesso e integração por diferentes perfis de usuários e aplicações e provida por meio de geosserviços.

Os principais atores nacionais que atuam na produção de dados geoespaciais são as entidades governamentais (nas esferas federal, estaduais ou municipais), as universidades, os institutos de pesquisa, as empresas da iniciativa privada e empresarial. A sociedade também participa intensamente desse processo de geração de dados geoespaciais, por meio da atuação de entidades do terceiro setor, associações e de cidadãos em geral. Todos esses entes são convidados a participar da Inde, com o intuito de colaborar entre si para a produção e o consumo de dados geoespaciais, o que, além de otimizar tempo e recursos, também contribui para o avanço do conhecimento.

Nessa ampla gama de produtores de dados, as instituições governamentais são chamadas a fazer parte da Inde de forma prioritária, uma vez que o dado geoespacial produzido a partir de recursos provenientes da União precisa estar acessível para a sociedade e para os tomadores de decisão que deles farão amplos usos. A Empresa Brasileira de Pesquisa Agropecuária (Embrapa) se insere neste perfil. Vinculada ao Ministério da Agricultura e Pecuária (Mapa), a Embrapa tem por missão “viabilizar soluções de pesquisa, desenvolvimento e inovação para a sustentabilidade da agricultura, em benefício da sociedade brasileira” (Embrapa, 2020). Nesse amplo contexto, é essencial que suas equipes atuem na produção e disponibilização de dados espaciais de forma a subsidiar estudos relacionados ao setor agropecuário.

Segundo o Plano de Ação da Inde (Brasil, 2010), a implementação de uma infraestrutura de dados espaciais (IDE) envolve componentes oriundos de três dimensões distintas: organizacional, técnica e humana. O processo de gestão dos dados espaciais da Embrapa utilizou o Plano de Ação como referência e foi iniciado em 2013 e lançado em 2018. Atualmente, o GeoInfo é a infraestrutura de dados espaciais institucional da Embrapa, utilizado tanto para o armazenamento de dados restritos ao público interno como para publicação de dados geoespaciais da empresa para o público externo por meio da Inde.



O objetivo deste trabalho é apresentar um diagnóstico do GeoInfo como solução corporativa da Embrapa. Após cinco anos de sua implantação e utilização, o GeoInfo é hoje adotado como ferramenta oficial para armazenar bases de dados e ativos cartográficos resultantes dos projetos de pesquisa conduzidos pela instituição. O GeoInfo também é utilizado como parâmetro para o monitoramento de adoção desses resultados, visando mensurar seu impacto e alcance.

Este diagnóstico é baseado em análises nos metadados dos dados catalogados na plataforma, com a finalidade de compreender o perfil de utilização do GeoInfo. Espera-se que este diagnóstico possa contribuir com o planejamento de ações futuras sobre sua evolução e manutenção.

1.1. Histórico

A Embrapa dispõe de uma estrutura capilarizada com 43 centros de pesquisa distribuídos pelo Brasil, dos quais aproximadamente 30 contam com equipes dedicadas à produção de dados geoespaciais. Antes da implantação do GeoInfo, todo o acervo da Embrapa apresentava-se distribuído nesses centros, em mapotecas, bibliotecas, servidores, dispositivos variados (pendrives, HDs externos) e em estações de trabalho (desktops). A gestão desses dados era feita de forma heterogênea, muitas vezes sem catalogação com metadados, versionamento e disponibilização para usuários internos ou externos. Essas dificuldades na gestão de dados espaciais na Embrapa impactavam no desenvolvimento de pesquisas colaborativas e resultavam, muitas vezes, em retrabalho.

A implantação de um processo de gestão dos dados apoiado por uma IDE no escopo institucional (o GeoInfo) foi iniciada em 2013 com a participação de diversos centros de pesquisa da empresa. Partiu-se de um inventário sobre os dados geoespaciais existentes na empresa, o qual revelou uma grande diversidade de produtos, que abrangia: conjuntos de dados temáticos, relacionados à agricultura e temas afins; dados resultantes de coletas de campo, em formato digital georreferenciado ou em planilhas eletrônicas; imagens de satélites, fotografias aéreas em formato analógico e digital; mapas e produtos de terceiros; entre outros.

Esse levantamento resultou na definição da tipologia de dados espaciais da Embrapa, relacionados aos aspectos de “produtor”, “produção” e “disponibilização de dados na Inde”.



Segundo Pinto et al. (2017), o acervo de dados espaciais da Embrapa é composto por: dados primários (resultantes da coleta e sistematização de dados científicos ou brutos); dados derivados com aporte intelectual (gerados por meio da integração de dados de diferentes naturezas a partir do esforço intelectual de seu corpo técnico); e dados espacializados de fontes externas (dados obtidos de diversas fontes que, embora apresentem a componente espacial, não configuram-se em informações inéditas e passíveis de disponibilização para a sociedade). O processo de gestão dos dados espaciais da Embrapa foi descrito por Drucker et al. (2017) e Pinto et al. (2017).

Em atendimento aos pilares Tecnologia e Padrões sugeridos pela Inde (Daltio et al., 2012), definiu-se o uso dos softwares GeoNode² e GeoNetwork³, ambos alinhados aos requisitos estabelecidos internacionalmente pela Open Geospatial Consortium (OGC) e aos princípios Free Open Source (FOSS). O GeoNode é uma plataforma voltada para a gestão e publicação de dados geoespaciais que reúne projetos de software de código aberto e mecanismos de persistência. O GeoNetwork é uma aplicação de catalogação especializada em metadados geográficos. Ambas as plataformas atuam de forma integrada na solução GeoInfo: os dados armazenados no GeoNode têm seus metadados catalogados no GeoNetwork. Essa integração permite mecanismos elaborados de recuperação (buscas avançadas) atrelados à publicação dos dados via protocolos web que são padrões para geoserviços.

A arquitetura do GeoInfo é formada atualmente por 22 instâncias do GeoNode, como mostra a Figura 1, que atuam como nós dedicados para hospedagem e publicação de dados de Unidades da Embrapa (algumas Unidades utilizam juntas uma instância central). O GeoNetwork é utilizado como catálogo único para os metadados e agrega todo o acervo de dados da empresa. Ambos os elementos são essenciais na integração com a Inde, seja no compartilhamento dos metadados, seja na disponibilização dos dados.

² <https://geonode.org/>

³ <https://geonetwork-opensource.org/>

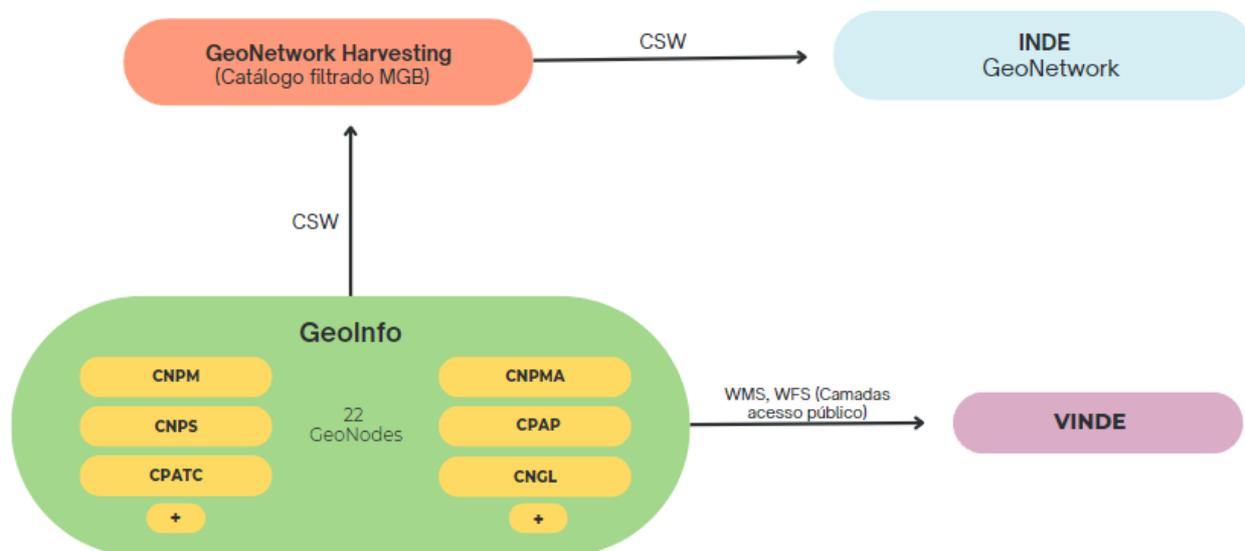


Figura 1. Arquitetura atual do Geoinfo.

Por sua natureza, o Geoinfo está em constante aprimoramento e evolução, tanto em termos tecnológicos como na revisão de seus processos, e foram identificados potenciais pontos de melhoria nessa arquitetura. Considerando-se o atual momento da solução, passado seu período de implantação, optou-se por um diagnóstico mais profundo, que pudesse guiar as próximas ações de manutenção e evolução. A próxima seção descreve esse diagnóstico.

2. MATERIAL E MÉTODOS

O diagnóstico foi produzido a partir de um *dashboard*⁴ público elaborado pelos membros do Comitê Gestor do Geoinfo na Embrapa com o uso do software Tableau⁵ a partir dos dados dos GeoNodes. O *dashboard* oferece um panorama geral do uso da solução e é composto por dois painéis, (i) Visão geral e (ii) Evolução, navegáveis através dos botões no topo. O painel Visão geral é mostrado na Figura 2.

⁴ <https://public.tableau.com/app/profile/marcos.fernando/viz/geoinfo01/Histria1>

⁵ <https://www.tableau.com/>



Figura 2. Dashboard de diagnóstico, painel Visão geral.

O gráfico de barras, no canto superior esquerdo, apresenta a evolução histórica da catalogação dos dados. O GeolInfo foi implantado institucionalmente há cinco anos, porém o cadastramento de artefatos foi iniciado em algumas Unidades da Embrapa no ano 2013, durante o projeto piloto da criação da IDE-Embrapa. Tais dados também foram considerados no diagnóstico. Na Figura 2, o tamanho das barras apresenta a quantidade de registros de dados catalogados, enquanto o gradiente de cores representa a quantidade de Unidades da Embrapa responsáveis por essa catalogação.

Na sequência (canto inferior direito), o primeiro gráfico setorial permite qualificar a natureza desse volume de dados em duas categorias: camadas de dados/*dataset* (dados que podem ser acessados diretamente via geosserviços); e documentos/*document* (dados que precisam de apoio de ferramentas de geoprocessamento para sua integração espacial). O segundo gráfico setorial, na sequência do *dashboard*, e o gráfico de árvore discorrem sobre o nível de acesso dos dados: cerca de 60% dos dados da plataforma são públicos, porém apenas 17% dos dados estão publicados na Inde.

O gráfico de barras central disponível na Figura 2 apresenta a distribuição dessas camadas por Unidade da Embrapa (quantitativo representado tanto no tamanho da barra quanto na intensidade da cor). O gráfico de barras mais à direita mostra o volume de camadas catalogadas

por usuário e permite identificar, dentro das Unidades, os principais usuários catalogadores de dados (potenciais produtores). Todos os elementos gráficos desse painel podem ser usados como filtro. É possível, portanto, selecionar uma Unidade, em um determinado ano, para listar usuários. Ou, ainda, identificar as que mais contribuem com dados públicos, ou na Inde.

A Figura 3 mostra o painel “Evolução”, que permite o acompanhamento temporal (mês a mês) da catalogação de dados. Este painel pode ser filtrado por ano, Unidade e usuário. A intenção dessa análise é verificar a existência de um padrão temporal de pico de publicação, ou a curva desse processo ao longo do ano.

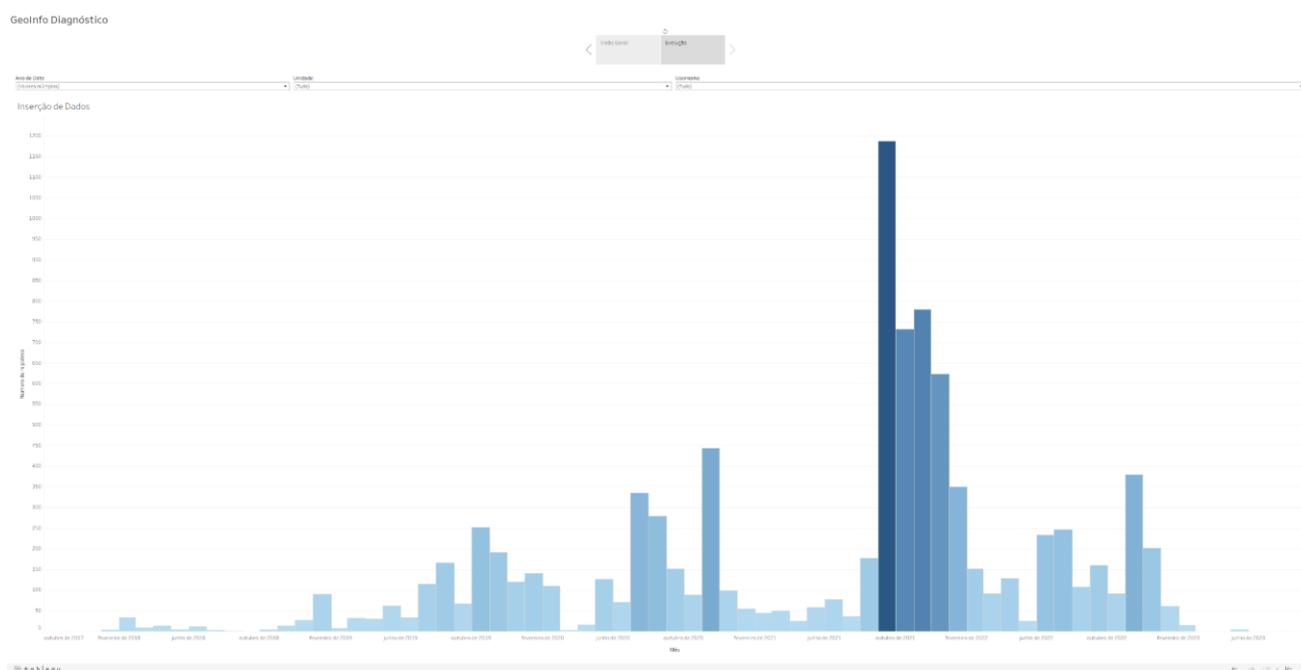


Figura 3. Dashboard de diagnóstico, painel Evolução.

3. RESULTADOS E DISCUSSÃO

O trabalho feito no diagnóstico permitiu a verificação de algumas situações:

- O ano de maior catalogação de artefatos foi 2021 (com o registro de 3.851 camadas e atuação de 22 Unidades), três anos após a implantação do GeoInfo.
- O volume de dados catalogados não apresenta um padrão ao longo do tempo e não está diretamente relacionado com maior participação de Unidades no processo. O ano 2015, por



exemplo, contou com expressiva participação de Unidades (21 Unidades) quando comparado a 2019 (maior volume de camadas, porém apenas 12 Unidades atuantes).

- As camadas de dados representam atualmente (2023) 56% do acervo do GeoInfo. Pouco mais da metade dos dados é de acesso público (55%) e 17% está publicada na Inde.
- Os documentos representam 43% dos dados cadastrados no GeoInfo e precisam ser avaliados, diante da adequabilidade do uso do GeoNode para armazenamento de dados desta natureza.
- A maior parte das Unidades participantes do GeoInfo dispõe de um volume de acervo inferior a 250 artefatos, e pode-se questionar, em termos técnicos, a necessidade de uma infraestrutura dedicada.
- A utilização de usuários não nominados durante a catalogação (usuário “embrapa”) dificulta não só o mapeamento dos produtores de dados como também o contato com os responsáveis pelo registro do conteúdo para eventuais necessidades de correções.
- É preciso considerar múltiplos aspectos para qualificar o acervo de dados de cada Unidade no GeoInfo. A análise deve extrapolar o critério “volume de camadas” como sinônimo de importância ou visibilidade. Quando se considera o critério “existência de dados públicos ou publicados na Inde”, por exemplo, a contribuição que cada Unidade da Embrapa oferece ao sistema é consideravelmente diferente.
- Não existe um calendário uniforme de cadastramento de novas camadas pelas Unidades ao longo do ano. Cada Unidade tem uma dinâmica própria e apresenta variações sazonais no uso.

Junto à equipe de manutenção e evolução das ferramentas GeoNode e GeoNetwork, foi elencado o principal gargalo a ser endereçado: o alto custo de manutenção e de atualização de versões de software, por conta da arquitetura descentralizada em múltiplos nós de hospedagem de dados. Além disso, percebeu-se que a existência de múltiplos pontos de entrada traz complexidade considerável tanto para usuários internos quanto para externos, como na Inde, pois não oferece uma visão unificada da arquitetura do GeoInfo na Embrapa, e sim fragmentada por meio de suas diversas Unidades ou centros de pesquisa.

Iniciou-se, com isso, o processo de definição de uma nova arquitetura para o GeoInfo, composta por apenas uma instância de GeoNode e uma instância de GeoNetwork e utilizando como base as versões mais atualizadas e estáveis de cada uma das ferramentas. O histórico do perfil de uso (volume de dados, usuários, Unidades e periodicidade de catalogação) está sendo



considerado no dimensionamento computacional dos servidores dedicados para hospedar a nova versão do GeolInfo.

De forma complementar, está sendo conduzido um diagnóstico detalhado do acervo, em cada Unidade participante. O processo foi iniciado pelas Unidades de menor volume (abaixo de 250 camadas/documentos cadastrados), seja o conteúdo de acesso público ou restrito. Atualmente já foram analisadas cerca de 2.000 camadas e documentos provenientes de 15 Unidades. Esta análise dos dados e metadados revela que há quantidade considerável de camadas que necessitam de algum tipo de ajuste – há muitos campos cadastrados em dissonância com as orientações e diretrizes do processo de gestão dos dados espaciais da Embrapa.

Os apontamentos foram classificados em três níveis de severidade: "moderado", "crítico" e "muito crítico". Apontamentos "moderados" referem-se essencialmente a preenchimentos incorretos de metadados, à ausência de campos obrigatórios ou à não observância de padronizações. As orientações para o preenchimento dos metadados estão disponíveis para os funcionários e colaboradores da Embrapa, e sua observância é essencial para estabelecer uma referência normativa para produtores e consumidores de geoinformação. Apontamentos "críticos" tratam de questões que impedem a correta interpretação e uso dos dados: ausência de legendas, erros de projeção cartográfica ou ausência de descrição dos atributos. Já os apontamentos "muito críticos" tratam de registros com ausência significativa de atributos, o que impossibilita a compreensão do conteúdo, e registros que apresentam erros durante o processo de acesso/download.

Os membros do Comitê Gestor do GeolInfo na Embrapa estão conduzindo o processo de migração do acervo atual para a nova versão do GeolInfo a partir desses subsídios, de forma a estabelecer um procedimento a ser adotado nesse processo e que possa ser adequado às especificidades de cada Unidade. A migração está em fase inicial, com a homologação do novo ambiente de software, validação da nova arquitetura e considerando, sempre que possível, os apontamentos de catalogação. Pretende-se, dessa forma, contribuir para a qualidade do acervo e revisar o processo de catalogação de dados estabelecido, para torná-lo ainda mais compatível com a dinâmica de trabalho dos produtores de dados espaciais da Embrapa.



4. CONCLUSÕES

Após cinco anos de implantação e uso, o GeolInfo é atualmente a infraestrutura de dados espaciais institucional da Embrapa. Uma solução consolidada, amplamente adotada e parte integrante dos processos de entrega dos dados espaciais e cartográficos dos projetos de pesquisa conduzidos pela instituição. Desde a sua implantação em 2018, o GeolInfo desempenha a dupla função: ser o repositório de dados espaciais da Embrapa e conectar a empresa à Infraestrutura Nacional de Dados Espaciais (Inde).

Construído a partir de softwares livres, a solução está em constante manutenção e evolução. O GeolInfo passa atualmente por um momento de transformações, e o diagnóstico apresentado neste trabalho irá subsidiar as ações nesse processo. Apesar do custo humano envolvido, acredita-se ser uma ação necessária para reafirmar o status alcançado pela solução dentro da Embrapa e seu posicionamento estratégico junto aos usuários e parceiros.

5. REFERÊNCIAS

BRASIL. Ministério do Planejamento, Orçamento e Gestão. Comissão Nacional de Cartografia. **Plano de ação para implantação da Infraestrutura Nacional de Dados Espaciais**. [Brasília: MPOG], 2010. Disponível em: <https://inde.gov.br/pdf/PlanoDeAcaoINDE.pdf>. Acesso em: 10 maio 2022.

CÂMARA, G.; CASANOVA, M. A.; HEMERLY, A. S.; MAGALHÃES, G. C.; MEDEIROS, C. M. B. **Anatomia de sistemas de informações geográficas**. Campinas: Unicamp, 1996. 197 p.

DALTIO, J.; SPADOTTO, C. A.; SANTOS, B. B. de O.; FOIS, N. S.; MARTINHO, P. R. R.; TRAJANO, S. R. R. da S.; HOLLER, W. A. **Infraestrutura Nacional de Dados Espaciais: aspectos computacionais**. Campinas: Embrapa Gestão Territorial, 2012. 14 p. (Embrapa Gestão Territorial. Comunicado Técnico, 1).

DRUCKER, D. P.; PINTO, D. M.; FIDALGO, E. C. C.; CUSTÓDIO, D. de O.; VICTORIA, D. de C.; ALMEIDA, B. T. de; SIMÕES, M.; MACHADO, C. R. de L.; SANTOS, V. V. dos; DART, R. de O.; AGLIO, M. L. D.; RASCHE, F.; LAFORET, M. R. C.; BETTIOL, G. M.; PEREIRA, S. E. M.; BRANDÃO, Z. N.; GARRASTAZU, M. C.; FILIPPINI ALBA, J. M.; COSTA, F. A. da; TORRES, R. C.; DOMPIERI, M. H. G.; SAMPAIO, S. M. N.; MARTINS, E. C.; PIEROZZI JUNIOR, I.; SILVA, G. B. S. da; GALINARI, G.; FIORINI, F. B.; TAKEMURA, C. M.; CRUZ, S. A. B. da; HOLLER, W. A.; OLIVEIRA, L. H. M. de. GeolInfo - Infraestrutura de dados espaciais abertos para a pesquisa agropecuária. **Reciis: Revista Eletrônica de Comunicação, Informação & Inovação em Saúde**, Rio de Janeiro, v. 11, p. 1-17, 2017.

EMBRAPA. **VII Plano Diretor da Embrapa 2020-2030**. Brasília, DF, 2020. 31 p.

PINTO, D. M.; BRANDÃO, V. V. dos S.; DRUCKER, D. P.; DOMPIERI, M. G.; RASCHE, F.; HOLLER, W. A.; BETTIOL, G. M.; CUSTODIO, D. O.; VICTORIA, D. C.; GONÇALVES, L. de M. P. B. Descrição dos dados da pesquisa geoespacial: a experiência da Embrapa. **Reciis: Revista Eletrônica de Comunicação, Informação & Inovação em Saúde**, Rio de Janeiro, v. 11, p. 1-16, 2017.