

Nota Técnica

Nova Arquitetura de Dados da Astec

Nº 06

Presidência

Astec – Assessoria Técnica

Novembro de 2017

Lucas Ferreira Mation

Nilo Luiz Saccaro Junior

Raimundo da Rocha



Governo Federal
Ministério do Planejamento, Desenvolvimento e Gestão
Ministro Dyogo Henrique de Oliveira

ipea Instituto de Pesquisa
Econômica Aplicada

Fundação pública vinculada ao Ministério do Planejamento, Desenvolvimento e Gestão, o Ipea fornece suporte técnico e institucional às ações governamentais – possibilitando a formulação de inúmeras políticas públicas e programas de desenvolvimento brasileiros – e disponibiliza, para a sociedade, pesquisas e estudos realizados por seus técnicos.

Presidente

Ernesto Lozardo

Diretor de Desenvolvimento Institucional

Rogério Boueri Miranda

Diretor de Estudos e Políticas do Estado, das Instituições e da Democracia

Alexandre de Ávila Gomide

Diretor de Estudos e Políticas Macroeconômicas

José Ronaldo de Castro Souza Júnior

Diretor de Estudos e Políticas Regionais, Urbanas e Ambientais

Alexandre Xavier Ywata de Carvalho

Diretor de Estudos e Políticas Setoriais de Inovação e Infraestrutura, interino

Rogério Boueri Miranda

Diretora de Estudos e Políticas Sociais

Lenita Maria Turchi

Diretor de Estudos e Relações Econômicas e Políticas Internacionais

Sérgio Augusto de Abreu e Lima Florêncio Sobrinho

Assessora-chefe de Imprensa e Comunicação

Regina Alvarez

Ouvidoria: <http://www.ipea.gov.br/ouvidoria>

URL: <http://www.ipea.gov.br>

NOVA ARQUITETURA DE DADOS DA ASTEC

Lucas Ferreira Mation¹

Nilo Luiz Saccaro Junior²

Raimundo da Rocha³

Resumo

Esta nota técnica apresenta o novo padrão de documentação, organização e guarda de bases de dados (pesquisas domiciliares, registros administrativos) usadas nas pesquisas do Instituto de Pesquisa Econômica Aplicada (Ipea). As bases obtidas devem ser informadas à Assessoria Técnica da Presidência do Ipea (Astec), que as registrará no Banco de Bases Estatísticas (BBE) e as salvará na pasta no *storage* principal de dados (*storage6/bases*). Essa pasta terá duas subpastas: DADOS, que conterá os dados em si, divididos nas pastas PUBLICO e RESTRITO; e DOCUMENTACAO, que conterá toda documentação, dicionários e programas de importação. A pasta de documentação terá controle de versão e acesso de leitura a todos os usuários na rede do Ipea. Assim, os pesquisadores poderão investigar as variáveis disponíveis em cada base antes de solicitar acesso aos dados restritos. Atualmente, as bases de dados originais estão dispersas em diversos *storages*. Ao longo do próximo semestre, a Astec, juntamente com a Coordenação-Geral de Tecnologia da Informação e Comunicações (CGTIC), farão a unificação dessas informações no *storage* principal.

Nota Técnica

O Instituto de Pesquisa Econômica Aplicada (Ipea) sempre trabalhou com grandes volumes de dados. Essa tradição vem sendo ampliada desde o final de 2016 com a participação do Ipea nos esforços de interoperabilidade entre bases de dados do governo federal e a criação do IpeaDATA-lab, laboratório de integração de registros administrativos do Ipea. Essas iniciativas, juntamente com a simplificação de procedimentos para a obtenção de bases de dados trazida pelo Decreto nº 8.789, de 29 de junho de 2016, que dispõe sobre o compartilhamento de bases de dados na administração pública federal, fizeram com que o Ipea passasse a receber um conjunto ainda maior de bases de dados de acesso público e de acesso restrito. Desde então o volume de registros administrativos disponíveis no Ipea já dobrou.

A Assessoria Técnica da Presidência do Ipea (Astec) é a área responsável por solicitar, armazenar, tratar e conceder o acesso interno a tais bases de dados⁴. As bases de dados que chegam ao

¹ Técnico em Planejamento e Pesquisa, Chefe da Astec do Ipea.

² Técnico em Planejamento e Pesquisa da Astec do Ipea.

³ Auxiliar Técnico da Astec do Ipea.

⁴ Portaria nº 324, de 21 de julho de 2010. Dispõe sobre a gestão de bases de dados de suporte a estudos e pesquisas finalísticas hospedadas no Ipea.

Ipea têm de ser cadastradas no Banco de Bases Estatísticas (<http://intranet.ipea.gov.br/?id=1.3.22>). Entretanto, apesar de catalogadas, devido a limitações de espaço nos *Storages*⁵ as bases de dados encontravam-se dispersas em diversos locais da rede. Paradoxalmente, a ausência de uma partição do *Storage* com maior capacidade de armazenamento para centralização dos dados acarreta uma maior pulverização e replicação das bases, o que consome ainda mais espaço de disco nos *Storages* da rede. Por fim, numa organização com tantos colaboradores, projetos e bases de dados existe uma tendência natural à dispersão da informação, ao aumento da “entropia informacional”.

Neste documento, apresentamos a nova arquitetura de armazenamento da informação no Ipea, que busca superar esses desafios e assessorar todo o Ipea em sua missão de auxiliar o governo federal no desenho, execução e avaliação dos resultados de políticas públicas. Conforme detalhado na próxima seção, os dados serão armazenados na [\\storage6\bases\](#), seguindo uma estrutura de pastas que garante a restrição de acesso aos dados sigilosos e assegura a transparência no acesso aos metadados dessas mesmas bases.

A compra pelo Ipea de um novo *storage 150* com *terabytes*, que ocorreu no final de 2016, permite uma reorganização de bases que antes estavam pulverizadas em várias áreas da rede. Desse modo, iniciando uma melhor gestão dos dados já existentes no órgão. Teve início em setembro de 2017 a migração de algumas bases de dados para o novo *Storage*, ação que será paulatinamente ampliada para mais bases de dados até a conclusão, prevista para julho de 2018.

Catálogo e armazenamento dos dados

As bases de microdados obtidas pelo Ipea, especialmente quando se tratarem de registros administrativos, devem ser comunicadas à Astec. Essa assessoria, por sua vez, catalogará as bases no Banco de Bases Estatísticas (BBE), e manterá os originais no *Storage* central de dados. Isso é necessário a fim de que todos os pesquisadores do Instituto tenham conhecimento de todas as bases que o Ipea possui e possam eventualmente utilizá-las em pesquisas específicas. A comunicação e a solicitação de cadastro das bases podem ser realizadas via e-mail para astec.bbe@ipea.gov.br.

Estrutura das pastas

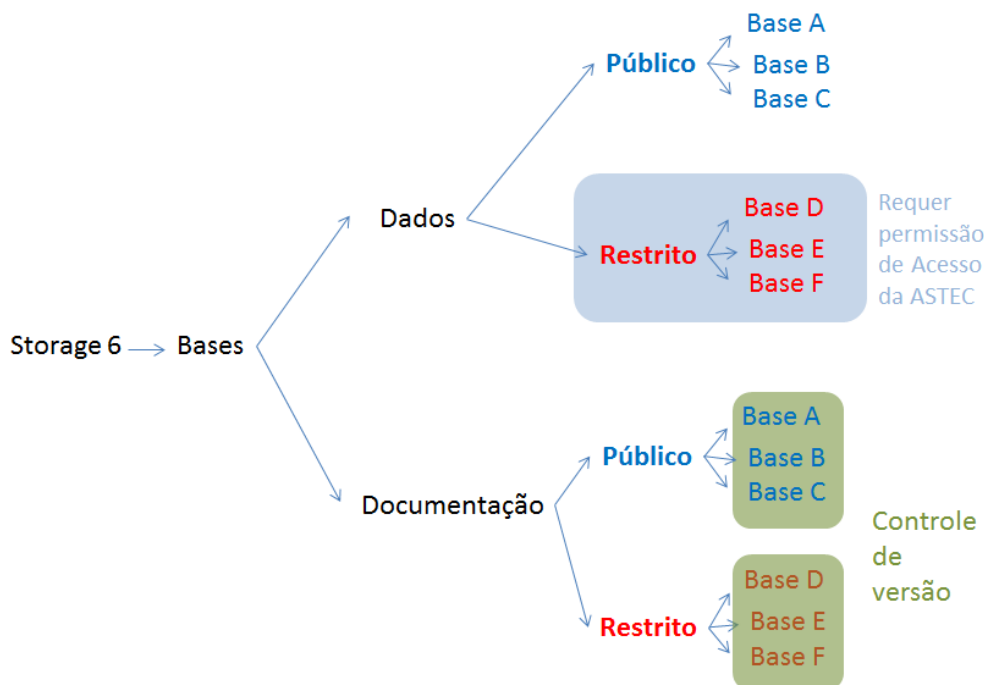
Todas as bases serão armazenadas em [\\storage6\](#). Dentro da pasta [\\storage6\bases\](#) há outras duas pastas, `\dados\` e `\documentação\`. A primeira contém os dados propriamente ditos (dados originais e dados importados em diversos formatos), enquanto a segunda contém as informações e documentos relativos aos dados (ofícios de solicitação do Ipea e de respostas dos órgãos cedentes, formulários específicos que se fizerem necessários, metadados, programas de importação, dicionários, tradutores, entre outros). Ambas as pastas se subdividem novamente em pastas `\publico\` e `\restrito\`. Essas, por sua vez, se subdividem em pastas que compõem as diferentes bases de dados da Astec, correspondentes entre si em ambas as divisões. Dessa

⁵ *Storage* é um hardware projetado especificamente para armazenamento de dados, que permite o acesso a partir dos terminais de uma rede interna.

forma, enquanto os dados de cada base estão dentro de \dados\, sua documentação correspondente está dentro de \documentação\, conforme esquematiza a figura a seguir.

FIGURA 1

Storage 6 – Arquitetura dos Dados



Todas as pastas possuem acesso livre para os pesquisadores autenticados dentro da rede do Ipea, exceto a pasta \DADOS\RESTRITO\ (bem como suas subpastas), indicada pelo quadrado azul na Figura 1. Apenas esse acesso precisa ser concedido pelo Chefe da Astec. Assim, os pesquisadores podem saber exatamente o tipo e abrangência dos dados armazenados, antes mesmo de terem acesso a eles e sem a necessidade de consulta à Astec. Essa é uma forma de aliar agilidade, transparência e responsabilidade com as bases de dados de acesso restrito concedidas ao Ipea por outros órgãos.

As pastas que contém as bases de dados públicas e restritas dentro da pasta \documentação\ possuem controle de versão (área em verde na Figura1). Esse controle é especialmente importante porque permite a recuperação de versões anteriores em caso de erro, bem como o acompanhamento do processo de manipulação dos dados, seguindo as boas práticas da Gestão da Informação. Dentro de cada pasta há ainda arquivos LEIAME.TXT, com orientações aos pesquisadores para a leitura, a manipulação e a integração das bases de dados.

Cada uma das pastas onde estão as bases de dados (exemplificadas na Figura 1 como bases A, B, C, D, E e F) corresponde ainda a um projeto dentro do grupo de GitLab https://gitlab.com/ASTEC_IPEA/, onde estão os repositórios para o tratamento, manipulação e leitura de cada dado armazenado pelo IpeaData-lab. O GitLab (https://gitlab.com/users/sign_in) é um gerenciador de repositório de software, que fornece uma plataforma de hospedagem de código-fonte com controle de versão. Constitui um software livre, e sua plataforma permite que programadores, utilitários ou qualquer usuário cadastrado contribuam em projetos privados

e/ou *Open Source*. No grupo ASTEC_IPEA do GitLab, haverá um projeto para cada base de dados. Dessa maneira, a pasta <\\storage6\bases\documentação\publico\base A>, por exemplo, referente à Base A na Figura 1, corresponde ao projeto https://gitlab.com/ASTEC_IPEA/Base A/.

É importante que não se confunda os dados de acesso restrito aqui tratados com os dados da sala de sigilo do Ipea. A sala de sigilo não está conectada a nenhuma rede, e o acesso a ela depende de procedimentos específicos determinados pela Portaria Ipea nº 078, de 20 de maio de 2014. Já os dados de acesso restrito são dados que não podem ser compartilhados externamente por inexistir autorização expressa do órgão ou entidade responsável (Art. 8º, parágrafo 2º do Decreto nº 8.789/2016), sendo por isso necessário um grau maior de controle sobre o seu acesso do que sobre o acesso a dados públicos.

Esse esforço de organização de dados e de acesso a dados visa a atingir objetivos de interoperabilidade e de integração de políticas e ações do Governo Federal. Esses objetivos esbarram em novos desafios impostos pelo crescimento da estrutura computacional governamental e pela imensa quantidade de dados disponíveis, que precisam ser levados em consideração de maneira organizada e inteligente. O Ipea tem os requisitos para se tornar um *player* importante na solução desses desafios, influenciando de maneira decisiva o futuro das políticas públicas brasileiras.

Missão do Ipea

Aprimorar as políticas públicas essenciais ao desenvolvimento brasileiro por meio da produção e disseminação de conhecimentos e da assessoria ao Estado nas suas decisões estratégicas

ipea Instituto de Pesquisa
Econômica Aplicada

MINISTÉRIO DO
**PLANEJAMENTO,
DESENVOLVIMENTO E GESTÃO**

