FGV SB

Projeto de criação e sustentação do Dataverse na FGV

FGV

Setembro de 2019

ESTRUTURA DA APRESENTAÇÃO

01

- A Plataforma Harvard Dataverse na FGV (Integração Repositório Digital, Rede de Pesquisa e TI)

+ *Desenvolvimento*
+ *Pesquisadores da FGV já utilizam essa plataforma*
+ *Pessoal envolvido*

02

- Questões de implementação do Dataverse

+ *Por que o Dataverse?*
+ *Benefícios*
+ *Iniciativa de instalação*
+ *Padrões de metadados*
+ *Plano de Gestão de Dados*
+ *Desafios*

03

- Como a ciência de dados está ligada ao Repositório de Dados

+ *Ciência de dados & Repositório de dados*

04

- Aspectos de um Repositório de dados

+ *Repositório sob o contexto de Governança de Dados*
+ *Ciclo de vida do dado de pesquisa*

05

- A Plataforma Harvard Dataverse

+ *Principais etapas previstas para a implementação*
+ *Arquitetura inicial proposta, acesso interno no início do projeto*

01

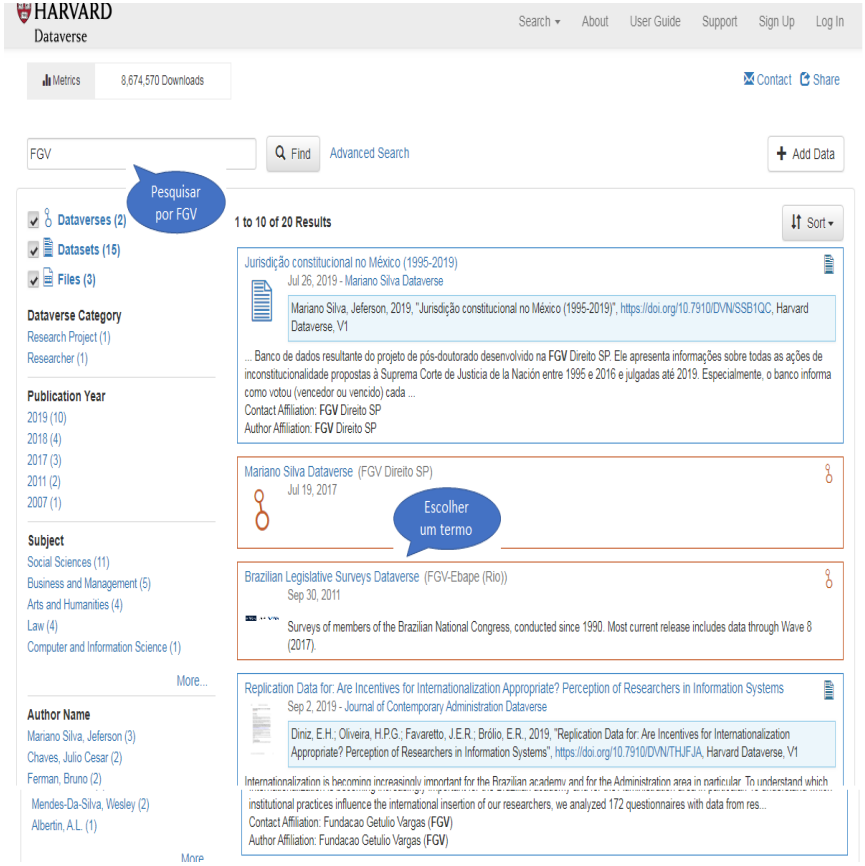
A Plataforma Harvard Dataverse na FGV (Integração Repositório Digital, Rede de Pesquisa e TI)

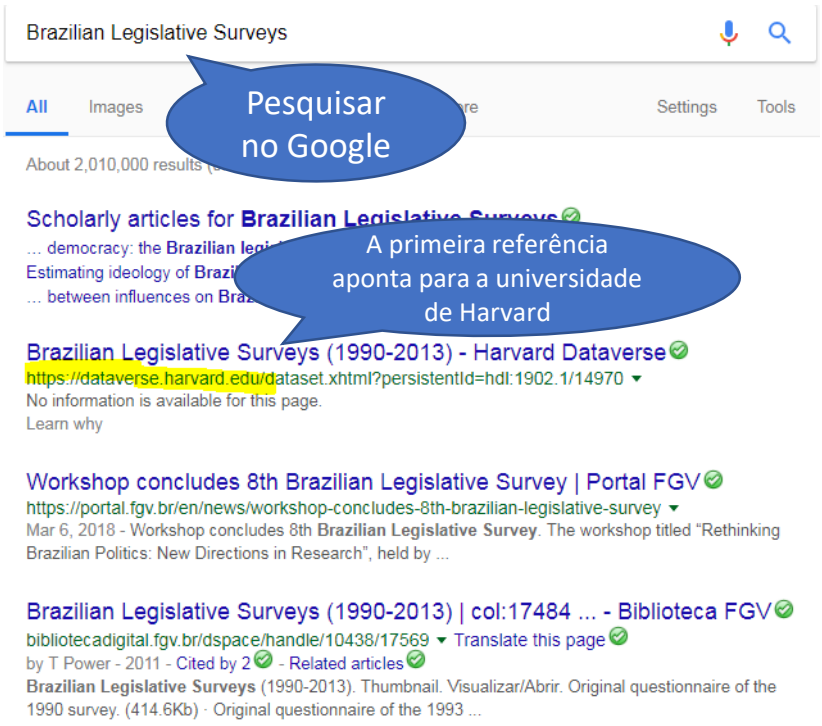
DESENVOLVIMENTO

- Com a necessidade de armazenamento dos dados das pesquisas aplicadas da FGV, surgiu a demanda da criação do Repositório de dados e o envolvimento da Biblioteca com a Rede de Pesquisa e a TI.

- A disponibilização da plataforma Dataverse na FGV será a referência principal aos pesquisadores da FGV, que atualmente buscam soluções de repositórios externos.

PESQUISADORES DA FGV JÁ UTILIZAM ESSA PLATAFORMA





PESSOAL ENVOLVIDO

- Criação de um Grupo Multidisciplinar

Maioria das universidades envolvidas com open science estão caminhando com um grupo de trabalho multidisciplinar.

Na FGV essa equipe é formada pela Rede de Pesquisa, Biblioteca e TI.

Esse grupo facilita o caminho diante de possíveis questões burocráticas.

O grupo está composto da seguinte forma:

- Rede de Pesquisas - 1 analista
- Biblioteca Digital – 3 bibliotecárias
- TIC – 1 analista

- No início de 2018 começou a elaboração das diretrizes e orçamento para o projeto

02

QUESTÕES DE IMPLEMENTAÇÃO DO DATAVERSE

POR QUE O DATAVERSE?

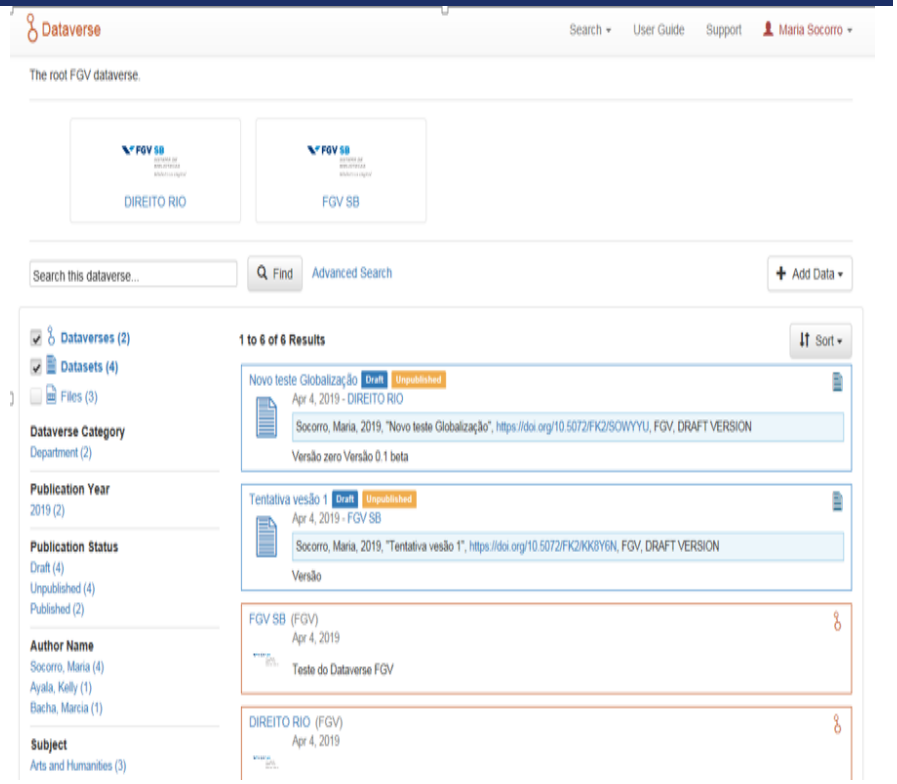
- O DSpace foi avaliado porém é um repositório genérico, não contemplando as questões em torno dos dados em si, como por exemplo as visualizações dos dados antes do download;
- Por tratar-se de uma plataforma de software livre, que atende aos protocolos da ciência aberta a interoperabilidade com outros repositórios de dados fica garantida;
- A Rede Cariniana (IBICT) dá suporte IBICT;
- Permite versionamento e preserva todas as versões;
- A preservação e acesso a longo prazo estão garantidos na identificação persistente;
- Ao depositar dados de pesquisa em um repositório Dataverse (incluindo Harvard Dataverse), os pesquisadores podem cumprir os requisitos da agência de financiamento para planos de gerenciamento de dados, o que não ocorre com o DSPACE, por exemplo;
- As publicações científicas são cada vez mais referenciadas a dados. Atualmente estamos perdendo referências de pesquisadores da FGV para universidade de Harvard.
- A plataforma de repositório de dados traz os meios necessários para a replicabilidade da pesquisa
- Propicia o uso dos princípios FAIR: to be **F**indable, **A**ccesible, **I**nteroperable, **R**eusable

BENEFÍCIOS

- **Gestão de dados segura.** O Dataverse cria condições de utilização e estabelecimento de restrições, caso se deseje limitar o uso ou acesso aos dados.
- **Partilha efetiva.** Dataverse é uma forma conveniente para divulgar seus dados. A equipe pode trabalhar on-line e sincronizar as alterações quando uma versão estiver concluída.
- **Acompanhamento das mudanças.** Dataverse oferece maior controle sobre gestão de mudanças para um projeto sem substituir qualquer parte desse projeto.
- **Acesso a longo prazo e preservação.** A plataforma oferece identificação persistente para o estudo e garante uma proteção confiável contra obsolescência de dados.
- **Organização e compatibilidade.** A plataforma conta com padrões de metadados para maximizar a compatibilidade e busca no sistema.
- **Economia de tempo.** Dataverse tem uma interface fácil de usar para carregar e pesquisar os dados.
- **Aumento da visibilidade de pesquisa.** A plataforma aumenta o reconhecimento acadêmico para o trabalho científico.

INICIATIVA DE INSTALAÇÃO

- A universidade de Harvard incentiva a adoção dessa plataforma em outras universidades, provendo vasta documentação de instalação e configuração
- No Brasil, o IBICT foi a primeira instituição que adotou essa plataforma. A FGV tem a oportunidade de ser uma referência no compartilhamento de bases “abertas” ligadas a pesquisa.
- Acabamos de instalar um piloto
- Verificar a utilização do Daverse em:
  - https://dataverse.harvard.edu



PADRÕES DE METADADOS

Dentre os listados no re3Data.org (setembro de 2019), o mais utilizado era o Dublin Core:

- Dublin Core (277)
- DDI - Data Documentation Initiative (161)
- DataCite Metadata Schema (142)

PLANO DE GESTÃO DE DADOS

O PGD é um documento de até duas páginas criado no início do projeto, descreve os dados e metadados, organizados, armazenados, preservados e compartilhados com o objetivo de facilitar a compreensão e a reutilização dos dados posteriormente.

Os PGDs incluem os seguintes itens:

- Descrição do projeto
- Autores e afiliações
- Tipos de dados produzidos
- Metadados
- Regras para distribuição e reutilização
- Planos de conservação e arquivamento

DESAFIOS

- Assinatura de serviços externos
  - DOI
  - ORCID
- Registrar o repositório no Re3data
- Obter investimento para a infra-estrutura tecnológica;
- Elaborar e executar treinamentos para equipe e pesquisadores;
- Equipe destacada para o projeto;
- Dar apoio e suporte aos especialistas relativo aos formatos, padrões, etc;
- Estabelecimento do projeto e apoio das unidades da FGV
- Criar uma política para o repositório

03

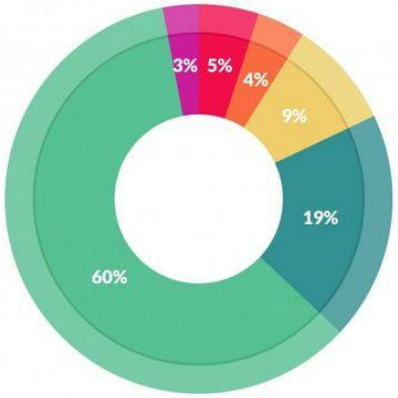
Como a Ciência de Dados está ligada ao Repositório de Dados

CIÊNCIA DE DADOS & REPOSITÓRIO DE DADOS

Qual a importância de um repositório de dados na era de ciência de dados?

Resultados de uma entrevista com 80 cientistas de dados:

A preparação dos dados consome 79% do tempo dos cientistas de dados. Sendo 60% para organização e 19% para a coleta.

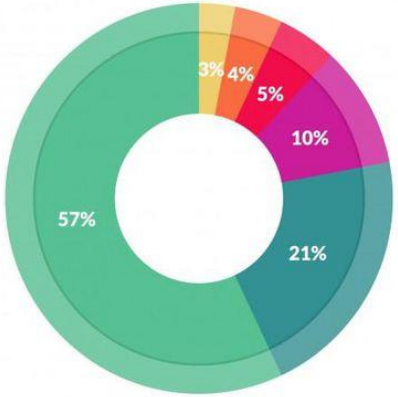


What data scientists spend the most time doing

- Building training sets: 3%
- Cleaning and organizing data: 60%
- Collecting data sets; 19%
- Mining data for patterns: 9%
- Refining algorithms: 4%
- Other: 5%

A etapa de "limpeza" e organização dos dados é considerada *menos agradável* por 57% dos entrevistados.

Nesse contexto, as disciplinas de governança e de catálogo de dados são destacadas, justamente por endereçarem questões que vão de encontro às dificuldades sentidas pelos cientistas de dados.



What's the least enjoyable part of data science?

- Building training sets: 10%
- Cleaning and organizing data: 57%
- Collecting data sets: 21%
- Mining data for patterns: 3%
- Refining algorithms: 4%
- Other: 5%

Fonte: Forbes, 2016.

04

Aspectos de um repositório de dados

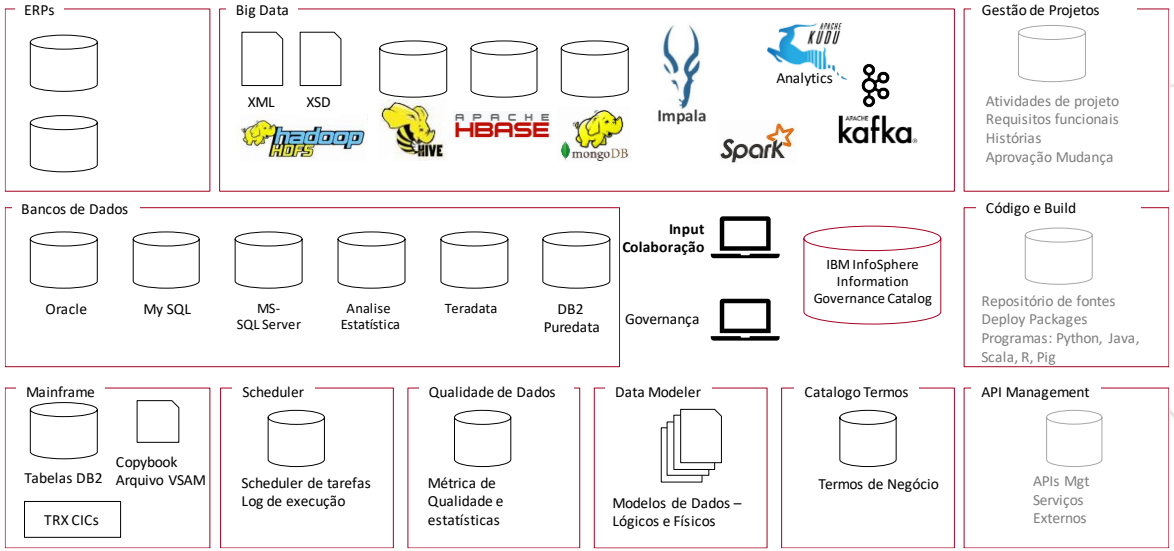
REPOSITÓRIO SOB O CONTEXTO DE GOVERNANÇA DE DADOS

O estabelecimento de uma linguagem comum

Um repositório possibilita a navegabilidade entre um campo abstrato e um concreto:

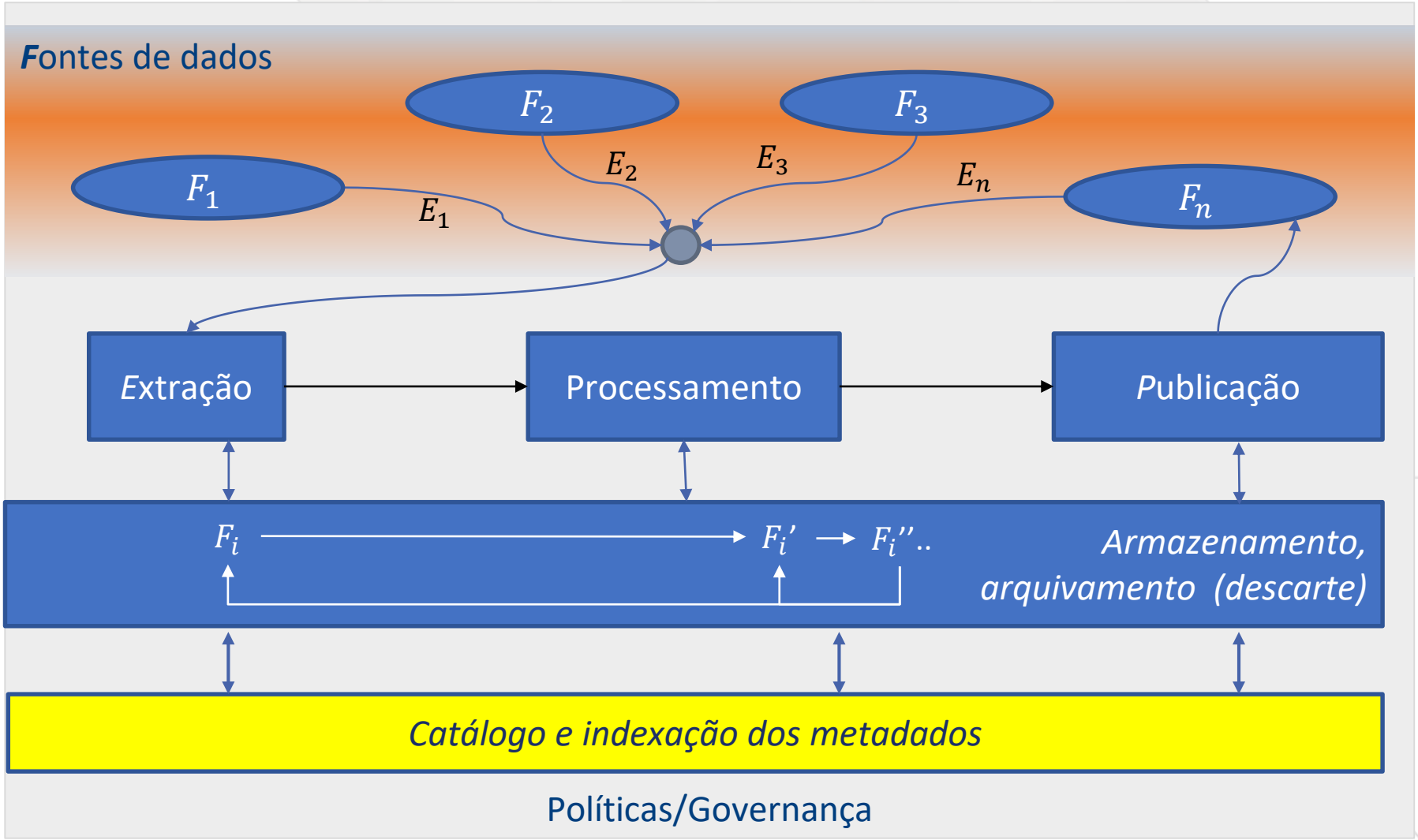
O universo abstrato prevê a gestão conjunta de políticas, regras e significância de termos.

No universo concreto, os bancos de dados, arquivos, processos, indicadores de desempenho e de qualidade são implementados de forma subordinada aos conceitos do universo abstrato.



Fonte:MD2

CICLO DE VIDA DO DADO DE PESQUISA



05

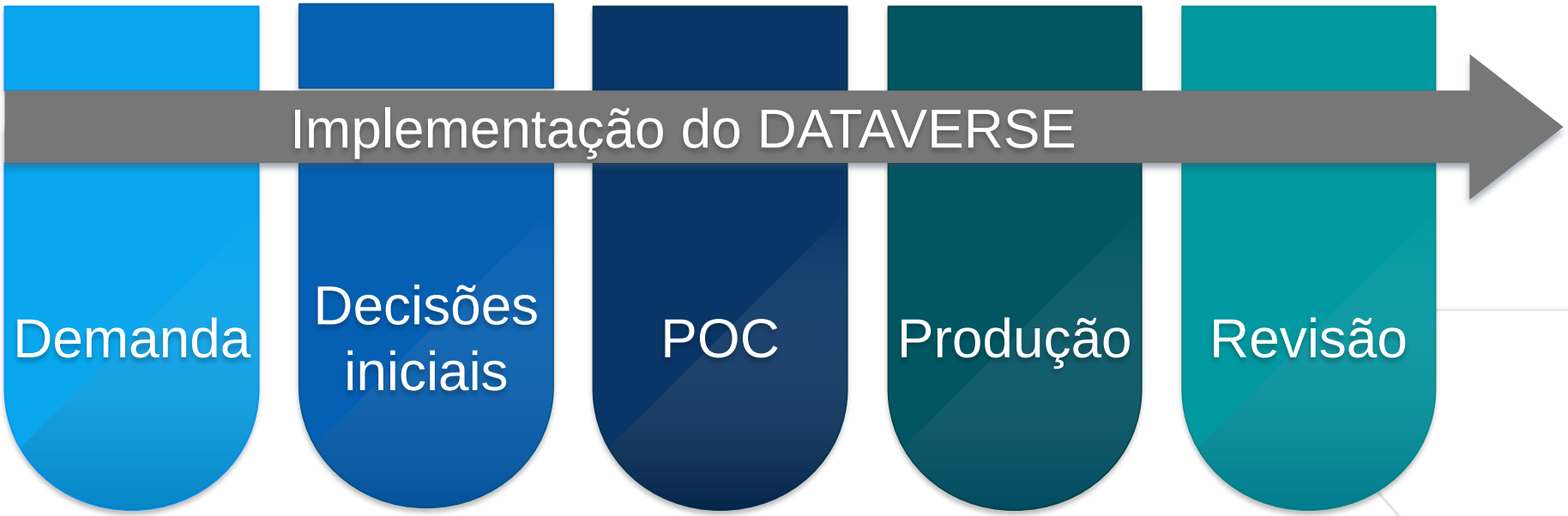
A Plataforma Harvard Dataverse

PRINCIPAIS ETAPAS PREVISTAS PARA IMPLEMENTAÇÃO

DOI, ~R$ 6.000/ano.

ORCID, ~R$ 16.500/ano.

Como software livre, a plataforma é sempre passível de contribuições, além de revisões em relação à demanda inicial.



Verificação de demanda junto a comunidade acadêmica.

Duas decisões precisam ser feitas antes do início da implementação: 1) formas de autenticação; 2) uso interno ou externo.

Prova de conceito do ambiente, com um grupo restrito de pessoas. Pode ser necessária interação com a comunidade de Harvard.

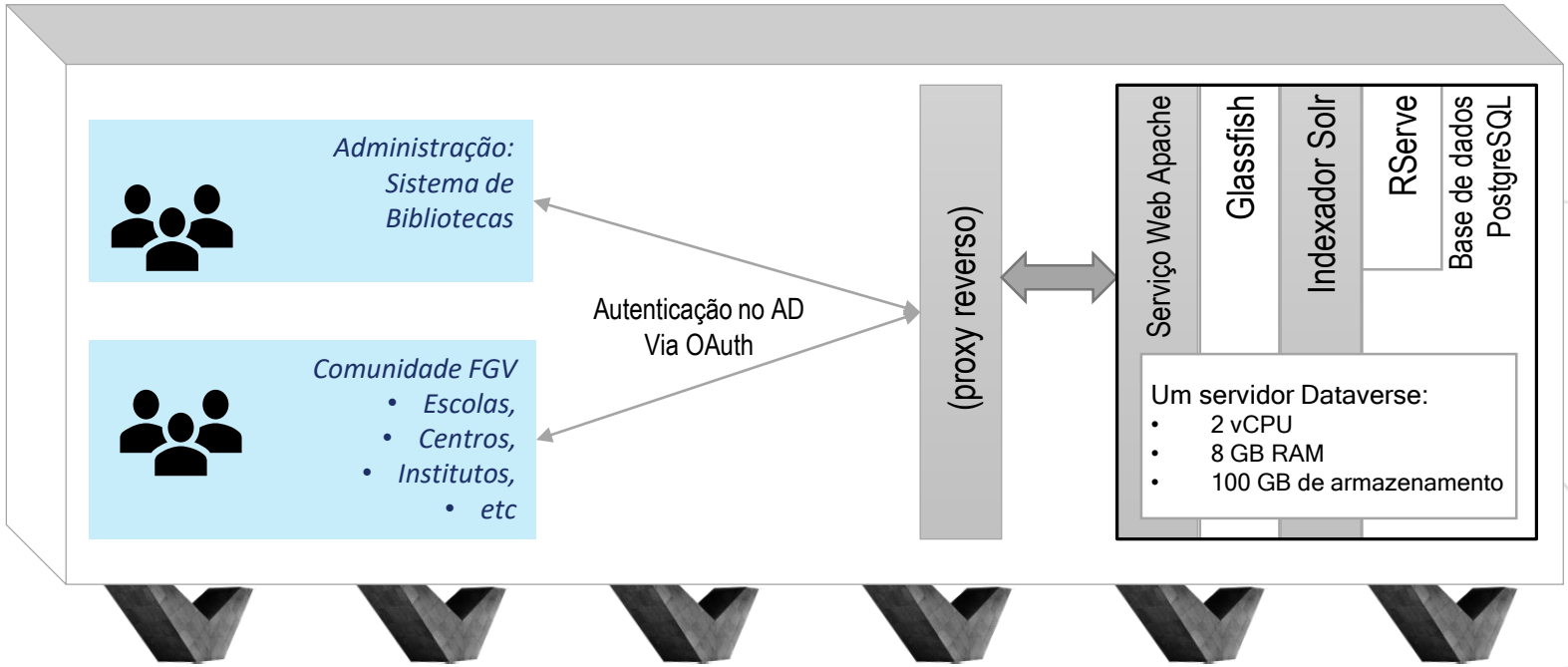
A entrada em produção marca o suporte oficial da instituição à plataforma, buscando sempre o engajamento da comunidade acadêmica.

ARQUITETURA INICIAL PROPOSTA,
acesso interno no início do projeto

Dificuldades iniciais:

Autenticação integrada, a forma de autenticação precisa ser debatida, pois dependendo da opção, há custo e mão de obra especializada para implementação.

A diferença entre possuir um repositório próprio e usar o de Harvard precisa ser compreendida pela comunidade acadêmica. Mais de 600 universidades usam repositório próprio, indexadas pelo portal de Harvard.



Obrigada!

Equipe:

TI

Júlio César Chaves: julio.chaves@fgv.br
Iuri Ferreira: iuri.ferreira@fgv.br
Felipe Vinhas: felipe.vinhas@fgv.br

Biblioteca Digital

Maria do Socorro Almeida: maria.socorro@fgv.br
Márcia Bacha: marcia.bacha@fgv.br
Kelly Ayala: kelly.ayala@fgv.br