EBOOK · VERSÃO 0.2 · NOVEMBRO DE 2026

Biodadica

A ciência do ciclo de vida dos dados biológicos: por que o dado vivo precisa de uma disciplina própria.

Projeto Biodadica · Biodadic

biodadica.org

Sumário

01 Uma área que já existia, sem nome
A lacuna entre a biologia e os dados

02 O nome: Biodadica e Biodadic
Bio + dad + ica, e por que o D fica no inglês

03 Por que o dado biológico é especial
Doze razões para uma disciplina própria

04 O framework: seis etapas
Da produção à visualização

05 A ficha do dado biológico
Os seis pilares que acompanham cada valor

06 Biodadica e as áreas vizinhas
O que muda em relação à bioinformática e às outras

07 Mapa de áreas de dados
16 grandes áreas e 196 tópicos, do projeto à preservação

08 As 20 ferramentas da plataforma
Wiki, ontologia, pacotes R e Python, loja de modelos, escola...

09 O ecossistema e o caminho
.org, .com, .net e o roadmap

— Glossário e referências

Biodadica: a ciência do ciclo de vida dos dados biológicos. Versão preliminar 0.2, novembro de 2026. Projeto Biodadica (em inglês, Biodadic).
Documento vivo, revisto com a comunidade. Sugestões em biodadica.org.
Como citar: PROJETO BIODADICA. *Biodadica: a ciência do ciclo de vida dos dados biológicos*. Versão 0.2. 2026. Disponível em: https://biodadica.org.

CAPÍTULO 01

Uma área que já existia, sem nome

Toda vida gera dados. Um genoma sequenciado, uma imagem de microscopia, a leitura de um sensor no curral, a ocorrência de uma espécie registrada por um pesquisador em campo, o resultado de um exame de sangue: tudo isso é dado biológico.

Nas últimas duas décadas, o volume desses dados cresceu numa velocidade que nenhum laboratório imaginava. O sequenciamento ficou barato, os sensores ficaram pequenos e os equipamentos passaram a gerar resultados automaticamente, em escala. Com isso surgiu um trabalho enorme que vai muito além de "analisar": é preciso **produzir** o dado com protocolo, **verificar** a sua qualidade, **organizá-lo** em sistemas confiáveis, **enriquecê-lo** com contexto, **analisá-lo** e **comunicar** o que ele revela.

Esse trabalho hoje está espalhado por várias áreas, cada uma cuidando de um pedaço:

- a **bioinformática** cuida dos algoritmos e da análise computacional, principalmente de sequências e moléculas;
- a **bioestatística** cuida do desenho dos experimentos e da inferência;
- a **biocuradoria** cuida da anotação e da organização de bases de dados;
- a **engenharia de dados** cuida dos fluxos, do armazenamento e

da integração, mas em geral sem conhecer a biologia por trás;
- a **ciência de dados** traz métodos de análise e aprendizado de máquina, mas trata o dado biológico como qualquer outro.

Ninguém olha para o ciclo inteiro. E é justamente nas passagens de uma etapa para a outra que mais se perde informação: a hora da coleta que não foi anotada, a qualidade da leitura que foi descartada, a versão do genoma de referência que ninguém registrou, o consentimento que ficou num papel separado do arquivo.

> A Biodadica nasce para dar nome e método a esse ciclo completo: do momento em que algo vivo é medido até o momento em que essa medida vira conhecimento confiável e reutilizável.

Biodadica é a disciplina dedicada ao ciclo de vida completo dos dados biológicos (produção, controle de qualidade, engenharia, enriquecimento, análise e visualização) e à sua transformação em conhecimento confiável, aberto e reutilizável.

CAPÍTULO 02

O nome: Biodadica e Biodadic

Uma área nova precisa de uma palavra. Construímos esta peça por peça, como a língua portuguesa sempre fez com as áreas do saber.

bio	dad	ica
vida	**dado**	**ciência**
grego βίος	port. dado ← lat. datum	grego -ική

Bio vem do grego *bíos*, vida: organismos, células, genomas, ecossistemas, lavouras, rebanhos, pacientes. **Dad** vem de *dado*, a palavra que usamos todos os dias no Brasil, que por sua vez vem do latim *datum*, "aquilo que foi dado". E **-ica** é o sufixo das áreas do conhecimento, presente em física, química, genética, estatística e informática.

Juntas, as três partes dizem exatamente o que a área é: **a ciência dos dados da vida**.

Em inglês: Biodadic, com D

A tradução natural para o inglês seria trocar o D pelo T de *data* e escrever "biodatic". Não fazemos isso, por três razões:

- **Datum é "dado".** O latim *datum* significa literalmente "o que foi

dado". Em português isso se diz *dado*, com D. A palavra brasileira é a tradução mais fiel da raiz latina.

- **A origem fica marcada.** O D é uma assinatura. Quem encontrar *Biodadic* num artigo, em qualquer lugar do mundo, saberá que o conceito tem raiz brasileira.
- **Um nome sem ruído.** Em inglês, *biodata* já significa "dados biográficos", como num currículo. *Biodadic* é inédito, curto e fácil de buscar.

> Assim como o mundo diz *bossa nova* e *capoeira*, o mundo diz **Biodadic**: uma palavra brasileira, **Biodadica**, que viaja.

IDIOMA	GRAFIA OFICIAL	OBSERVAÇÃO
Português	**Biodadica**	Sem acento. Pronúncia: bi·o·da·di·ca.
Inglês	**Biodadic**	Com D. Pronúncia: /ˌbaɪ.oʊˈdæd.ɪk/.
Espanhol	**Biodadic**	Mesma grafia do inglês.

CAPÍTULO 03

Por que o dado biológico é especial

Um número de vendas não muda depois de registrado. Uma amostra de sangue muda. O dado biológico é o retrato de algo vivo, e o que é vivo muda, varia, erra, evolui e pertence a alguém.

Cada uma das características abaixo existe, isoladamente, em outros tipos de dado. O que torna o dado biológico especial é que **todas acontecem ao mesmo tempo**. É essa combinação que justifica uma disciplina própria.

01 · INCERTEZA

Toda medida carrega uma incerteza

Sequenciadores não leem o DNA, eles estimam. Cada base sequenciada vem com uma probabilidade de erro, expressa na escala Phred: Q20 significa 1 erro em 100 bases e Q30, 1 erro em 1.000. Espectrômetros, sensores e microscópios automatizados funcionam do mesmo jeito: entregam um valor e uma confiança.

Quando a confiança é descartada e só o valor é guardado, o dado perde metade do seu significado. Na Biodadica, a incerteza é um campo obrigatório.

`Q30 → 99,9% de acerto por base`

02 · TEMPO

O relógio não para depois da coleta

Depois de retirado do organismo, o material continua vivo por um tempo: as células consomem nutrientes, o RNA se degrada, as proteínas se modificam e a expressão gênica responde ao estresse da própria coleta. Uma amostra processada 2 horas depois de coletada já não é a mesma de 10 minutos depois.

Por isso a hora da coleta, a hora do processamento e o intervalo entre elas fazem parte do dado. Um intervalo acima do limite do protocolo deve gerar um alerta, e não ser esquecido.

```
coleta 08:14 · processamento 10:41 · Δ 2h27 ⚠
```

03 · MOMENTO

A hora do dia e a fase da vida importam

O ritmo circadiano, o jejum, a estação do ano, a fase de desenvolvimento e o ciclo reprodutivo mudam os resultados. O mesmo indivíduo dá números diferentes em momentos diferentes. Sem esse contexto, uma comparação pode estar medindo só uma coincidência.

```
cortisol: alto de manhã, baixo à noite
```

04 · INTERFERÊNCIA

Observar altera o que é observado

Capturar um animal eleva o estresse, uma biópsia dispara inflamação e congelar e descongelar rompe células. O próprio ato de coletar muda a biologia, e a forma como foi feito precisa ficar registrada como condição do dado.

`temperatura · conservante · ciclos de congelamento`

05 · VARIABILIDADE

Nenhum ser vivo é igual a outro

Há variação entre indivíduos, entre células de um mesmo tecido e entre repetições de um experimento. É preciso separar a réplica biológica (outro indivíduo) da réplica técnica (a mesma amostra medida de novo) e o efeito real do efeito de lote (diferenças causadas pelo dia, pelo equipamento ou pelo operador). Uma média sem a sua variação não descreve a vida.

`réplica biológica ≠ réplica técnica`

06 · IRREPETÍVEL

A amostra acaba e o momento não volta

Um tecido é finito e perecível. O paciente se cura, a lavoura é colhida, a espécie pode desaparecer. Muitas vezes não existe uma segunda medição: o dado é o único registro que sobra daquele organismo naquele instante. Perdê-lo, ou perder o seu contexto, é perder aquele momento da vida para sempre.

`material único e irrepetível`

07 · REFERÊNCIA

O próprio conhecimento muda de versão

Genomas de referência são atualizados (por exemplo, do GRCh37 para o GRCh38 no caso humano), espécies mudam de nome e genes são renomeados. Até o Excel já forçou mudanças: genes como SEPT1 e MARCH1 eram convertidos automaticamente em datas e, em 2020, ganharam nomes novos (SEPTIN1, MARCHF1). Todo dado biológico precisa dizer contra qual versão do conhecimento foi interpretado.

`SEPT1 → 1-set → SEPTIN1`

08 · PRIVACIDADE

O dado biológico pertence a alguém

Um genoma identifica a pessoa e revela informações da família inteira, inclusive de parentes que nunca deram consentimento. Não existe anonimização perfeita para dados genéticos. A LGPD classifica dados genéticos e biométricos como dados pessoais sensíveis, e o GDPR europeu faz o mesmo. Consentimento, finalidade e regras de acesso fazem parte do dado, e não são burocracia ao lado dele.

`LGPD, art. 5º, II · GDPR, art. 9`

09 · SOBERANIA

A biodiversidade é patrimônio

No Brasil, o acesso ao patrimônio genético e ao conhecimento tradicional associado é regulado pela Lei 13.123/2015 e precisa ser cadastrado no SisGen. No mundo, o tema é regulado pelo Protocolo de Nagoya. A origem de uma amostra tem consequências legais e define como os benefícios do seu uso são repartidos. Para um país megadiverso como o Brasil, esse é um tema central.

`Lei 13.123/2015 · SisGen · Nagoya`

10 · ESCALA

Da molécula ao ecossistema

Dados biológicos atravessam escalas (moléculas, células, órgãos, indivíduos, populações, biomas) e formatos (sequências, imagens, séries temporais, registros de ocorrência, textos clínicos). Integrar tudo isso exige ontologias e identificadores comuns, e não apenas tabelas.

`genoma · imagem · sensor · ocorrência`

11 · PROVENIÊNCIA

Uma cadeia de custódia completa

O dado passa por várias mãos e máquinas antes de chegar a uma conclusão. Cada etapa pode introduzir erro ou contaminação. Sem rastrear o caminho de ponta a ponta, não há como reproduzir o resultado nem encontrar onde algo deu errado.

organismo → amostra → alíquota → biblioteca → corrida → arquivo → análise

`quem · quando · onde · como`

12 · LONGEVIDADE

Um valor que dura séculos

Exsicatas de herbário coletadas há mais de 100 anos hoje revelam efeitos das mudanças climáticas e guardam DNA antigo. Um dado biológico bem documentado pode responder perguntas que ainda nem foram feitas, desde que seja encontrável, acessível, interoperável e reutilizável: os princípios FAIR.

`princípios FAIR`

> Dados comuns descrevem coisas. Dados biológicos descrevem algo que estava vivo, e que no instante seguinte já era outro.

CAPÍTULO 04

O framework: seis etapas

A Biodadica organiza tudo o que acontece com um dado biológico em seis etapas. Cada etapa tem um objetivo e um conjunto de perguntas que precisam ter resposta registrada.

1 Produção → 2 Qualidade → 3 Engenharia → 4 Enriquecimento → 5 Análise → 6 Visualização

1 Produção

Gerar o dado com protocolo: coleta, sequenciamento, leitura de sensores, imagens, observações em campo.

- Qual organismo, qual tecido, qual indivíduo?
- Quando foi coletado e quando foi processado?
- Com qual protocolo, equipamento e operador?

2 Controle de qualidade

Verificar se o dado é confiável o bastante para o uso pretendido, e registrar a sua incerteza.

- Qual a confiança de cada medida?
- O intervalo até o processamento ficou dentro do limite?
- Há sinais de contaminação ou efeito de lote?

3 Engenharia

Armazenar, transportar e integrar o dado em fluxos (pipelines) reprodutíveis, com formatos abertos.

- Onde o dado está e em qual formato?
- Qual versão do pipeline o processou?
- O caminho do dado pode ser refeito do zero?

4 Enriquecimento

Dar contexto ao dado: anotação, curadoria, ontologias e ligação com outras bases.

- Quais termos padronizados descrevem este dado?
- Contra qual versão de referência foi anotado?
- A quais outras bases ele se conecta?

5 Análise

Extrair conhecimento: estatística, bioinformática, modelos e inteligência artificial, levando a incerteza em conta.

- A incerteza das medidas foi propagada até o resultado?
- A variação biológica foi separada da técnica?
- A análise pode ser reproduzida por outra pessoa?

6 Visualização e disponibilização

Comunicar o que foi descoberto e devolver o dado ao mundo, de forma encontrável e reutilizável.

- O gráfico mostra também a incerteza?
- Quem pode acessar o dado, e com qual licença?
- O dado segue os princípios FAIR?

CAPÍTULO 05

A ficha do dado biológico

A proposta central da Biodadica é simples: o valor medido nunca anda sozinho. Ele viaja junto com a sua história, do primeiro registro até o último reúso.

Valor | Incerteza | Tempo | Contexto | Proveniência | Permissões

PILAR	O QUE REGISTRA	EXEMPLO
Valor	A medida em si.	uma sequência, uma concentração, uma imagem
Incerteza	A confiança da medida.	Q32; ± 0,4 mg/dL; limite de detecção
Tempo	Coleta, processamento e o intervalo entre eles.	2h27min, acima do limite de 2h
Contexto	Condições da coleta e da conservação; organismo; referência usada.	4 °C, EDTA, GRCh38.p14
Proveniência	O caminho completo, do organismo ao arquivo.	amostra → biblioteca → corrida
Permissões	Consentimento, licença, cadastros legais.	TCLE, LGPD, SisGen

Um exemplo

Abaixo, a ficha de uma leitura de sequenciamento de uma amostra de sangue humano. Note o alerta automático: o intervalo entre a coleta e o processamento passou de 2 horas, e isso fica gravado junto com o dado, para sempre.

```
{
  "valor":          "ATGGCGTACCTGA…",
  "incerteza":      "Q32 · 99,94% por base",
  "coleta":         "2026-11-12T08:14-03:00",
  "processamento": "2026-11-12T10:41-03:00",
  "intervalo":      "2h27min",  // ⚠ acima de 2h
  "condicao":       "4 °C · EDTA · 0 descongelamentos",
  "organismo":      "Homo sapiens · NCBITaxon:9606",
  "referencia":     "GRCh38.p14",
  "proveniencia":  "amostra BD-0192 → lib L07 → run R2311",
  "permissao":      "TCLE 2026-118 · LGPD: sensível"
}
```

Regras de ouro

- **Nada de valor sem incerteza.** Se a confiança não é conhecida, isso também é registrado.
- **Todo tempo tem fuso horário.** Datas sem fuso viram ambiguidade em projetos entre países.
- **Toda anotação cita a sua versão.** Genoma, ontologia, pipeline e taxonomia.
- **Alertas não se apagam.** Um desvio de protocolo pode invalidar uma análise, mas nunca deve ser escondido.
- **A permissão viaja com o dado.** Se o arquivo for copiado, as regras de uso vão junto.

CAPÍTULO 06

Biodadica e as áreas vizinhas

A Biodadica não substitui nenhuma área existente. Ela costura o que hoje está separado e assume a responsabilidade pelo ciclo inteiro.

ÁREA	FOCO PRINCIPAL	O QUE A BIODADICA ACRESCENTA
Bioinformática	Algoritmos e análise computacional de sequências e moléculas.	Começa antes (na coleta) e termina depois (no reúso); vale para qualquer dado vivo.
Bioestatística	Desenho de experimentos e inferência.	Garante que a incerteza e o contexto cheguem intactos até a análise.
Biologia computacional	Modelos de sistemas biológicos.	Fornece dados com proveniência para alimentar os modelos.
Biocuradoria	Anotação e organização de bases.	Liga a curadoria à produção e à qualidade de origem.
Engenharia de dados	Fluxos, armazenamento e integração.	Traz as regras próprias do dado vivo: tempo, variabilidade, ética.
Ciência de dados	Métodos de análise e aprendizado de máquina.	Impede que o dado biológico seja tratado como um dado qualquer.

Onde a Biodadica se aplica

- **Saúde:** laboratórios clínicos, biobancos, pesquisa com pacientes.
- **Agro e pecuária:** genética de sementes e rebanhos, sensores no campo, sanidade animal.
- **Biodiversidade:** coleções, herbários, monitoramento de espécies e ecossistemas.
- **Biotecnologia:** bioprocessos, fermentação, engenharia de organismos.
- **Ensino:** formação de profissionais que entendem ao mesmo tempo de biologia e de dados.

CAPÍTULO 07

Mapa de áreas de dados

Uma lista organizada das disciplinas, práticas, arquiteturas e padrões de dados que se aplicam ao dado biológico, do planejamento do projeto até a preservação por décadas. Cada tópico traz um exemplo de uso na biologia.

A versão on-line, com busca, está em **biodadica.org/areas**.

01

ANTES DO DADO · 10 TÓPICOS

Planejamento e projeto

Decidir o que será medido, como, por quem, com quais permissões e quanto vai custar guardar.

TÓPICO	O QUE É · EXEMPLO NA BIOLOGIA
Plano de Gestão de Dados (PGD / DMP)	Documento que descreve como os dados serão gerados, documentados, guardados, compartilhados e preservados. `ex.: PGD exigido por agências como a FAPESP`
Desenho experimental	Estrutura do experimento: tratamentos, controles, blocos, aleatorização e repetições. `ex.: ensaio de campo em blocos casualizados`
Cálculo amostral e poder estatístico	Quantas unidades são necessárias para detectar um efeito com confiança. `ex.: número de animais por grupo aprovado pela CEUA`
Pré-registro de estudos	Registro público das hipóteses e da análise antes da coleta. `ex.: ensaio clínico registrado no ReBEC ou no ClinicalTrials.gov`

Levantamento de requisitos e casos de uso	Quem vai usar o dado, para quê e com qual frequência. `ex.: painel de sanidade para o veterinário do rebanho`
Modelagem conceitual de dados	Entidades e relações do domínio antes de qualquer banco. `ex.: Indivíduo → Amostra → Alíquota → Ensaio`
Checklists de metadados mínimos	Conjunto mínimo de informações exigido por tipo de experimento. `ex.: MIxS (sequências), MIAME (microarranjos), MIAPE (proteômica)`
Aprovações éticas e legais	Autorizações necessárias antes de coletar. `ex.: CEP/CONEP, CEUA, SISBIO, cadastro no SisGen`
Orçamento de dados	Custo de gerar, processar, armazenar e preservar. `ex.: quanto custa guardar 500 genomas por 10 anos`
Gestão de riscos do projeto	Perda de amostras, falha de equipamento, vazamento de dados. `ex.: plano B para falha do freezer −80 °C`

02

ORIGEM DO DADO · 12 TÓPICOS

Amostragem e coleta

O momento em que algo vivo vira registro. Grande parte da qualidade do dado é decidida aqui.

TÓPICO	O QUE É · EXEMPLO NA BIOLOGIA
Amostragem probabilística	Aleatória simples, estratificada, sistemática e por conglomerados. `ex.: sorteio estratificado de propriedades rurais por região`
Amostragem ecológica	Parcelas, transectos, armadilhas, pontos de escuta, captura-recaptura. `ex.: transecto de aves em fragmento de Mata`

	Atlântica
Esforço amostral	Quanto foi amostrado e se foi suficiente. ex.: curva de acumulação de espécies
Amostragem temporal e longitudinal	Coletas repetidas no tempo, no mesmo indivíduo ou local. ex.: coorte com coletas a cada 6 meses
Variáveis pré-analíticas	Tudo o que acontece entre a coleta e a análise. ex.: tempo até o processamento (alerta acima de 2 h), temperatura, anticoagulante
Cadeia fria e logística	Transporte e conservação da amostra. ex.: gelo seco, nitrogênio líquido, registrador de temperatura
Identificação de amostras	Códigos únicos e legíveis por máquina. ex.: código de barras, QR code, RFID, IGSN
Georreferenciamento	Coordenadas, sistema de referência e precisão. ex.: SIRGAS 2000 / WGS84, incerteza em metros
Coleta digital em campo	Formulários móveis que validam na hora da coleta. ex.: ODK, KoboToolbox
Espécime-testemunho (voucher)	Exemplar depositado em coleção para confirmar a identificação. ex.: exsicata depositada em herbário
Biobancos e biorrepositórios	Guarda organizada de material biológico com seus dados. ex.: biobanco de soro com consentimento amplo
Ciência cidadã	Coleta feita pelo público, com validação. ex.: registros no iNaturalist

03

ORIGEM DO DADO · 14 TÓPICOS

Geração e aquisição de dados

Instrumentos, sensores e sistemas que transformam a amostra em dado digi-

tal.

TÓPICO	O QUE É · EXEMPLO NA BIOLOGIA
Sequenciamento de DNA/RNA	Leituras curtas e longas, com qualidade por base. `ex.: Illumina, Oxford Nanopore, PacBio → FASTQ`
Ômicas	Medições em larga escala de moléculas. `ex.: genômica, transcriptômica, proteômica, metabolômica, epigenômica, metagenômica`
Single-cell e espacial	Medições célula a célula ou com posição no tecido. `ex.: scRNA-seq, transcriptômica espacial`
Imagem biológica	Microscopia, histologia digital, fenotipagem por imagem. `ex.: lâminas digitais em OME-TIFF`
Imagem médica	Exames de imagem com metadados clínicos. `ex.: tomografia em DICOM`
Sensoriamento remoto e drones	Imagens aéreas e de satélite. `ex.: NDVI de lavoura, mapeamento de vegetação`
Sensores e IoT	Leituras contínuas e automáticas. `ex.: coleira GPS, sensor de solo, estação meteorológica, ordenha robotizada`
Instrumentos laboratoriais	Equipamentos analíticos e seus arquivos brutos. `ex.: espectrometria de massas, citometria, qPCR, ELISA`
Bioacústica	Gravação e identificação de sons de animais. `ex.: gravadores autônomos para anfíbios`
Dados clínicos	Prontuário eletrônico, exames laboratoriais, desfechos. `ex.: resultados de hemograma ligados à amostra`
LIMS	Sistema de gestão de informação do laboratório. `ex.: rastreio de amostras, ensaios e laudos`

Caderno eletrônico de laboratório (ELN)	Registro digital do que foi feito na bancada. `ex.: protocolo, desvios e observações do dia`
Integração com instrumentos	Captura automática de arquivos e metadados das máquinas. `ex.: middleware que envia a corrida do sequenciador ao LIMS`
Dados sintéticos e simulados	Dados gerados por modelo para teste ou privacidade. `ex.: coorte sintética para treinar pipeline`

04

CONFIANÇA · 12 TÓPICOS

Qualidade de dados

Medir, registrar e melhorar a confiança em cada dado. Na Biodadica, a incerteza é um campo obrigatório.

TÓPICO	O QUE É · EXEMPLO NA BIOLOGIA
Dimensões de qualidade	Completude, exatidão, consistência, atualidade, unicidade e validade. `ex.: % de amostras sem hora de coleta`
Escores de confiança	Probabilidade de erro associada à medida. `ex.: Phred (Q30), FDR em proteômica, limite de detecção`
Controle de qualidade de sequências	Avaliação e limpeza de leituras. `ex.: FastQC, MultiQC, remoção de adaptadores`
Controles e padrões	Referências conhecidas para checar o processo. `ex.: controles positivo e negativo, brancos, spike-ins, material de referência`
Efeito de lote e normalização	Separar variação técnica da biológica. `ex.: correção com ComBat`
Detecção de contaminação	Identificar material que não deveria estar ali. `ex.: DNA humano em amostra ambiental`

Detecção de outliers	Valores incomuns: erro ou biologia? `ex.: peso de animal 10× acima da média`
Perfilamento de dados (profiling)	Retrato estatístico de cada campo. `ex.: distribuição, nulos e valores distintos por coluna`
Validação por regras e esquemas	Testes automáticos sobre os dados. `ex.: Great Expectations, Pandera, JSON Schema`
Não conformidades e desvios	Registro permanente de fugas ao protocolo. `ex.: amostra processada 2h27 após a coleta`
Integridade de dados (ALCOA+)	Dados atribuíveis, legíveis, contemporâneos, originais e exatos, além de completos, consistentes, duradouros e disponíveis. `ex.: registros de estudo clínico`
Normas de qualidade	Acreditação de laboratórios e boas práticas. `ex.: ISO 15189, ISO/IEC 17025, BPL`

05

FLUXO · 15 TÓPICOS

Engenharia de dados

Mover, transformar e integrar dados com fluxos (pipelines) reprodutíveis e escaláveis.

TÓPICO	O QUE É · EXEMPLO NA BIOLOGIA
Ingestão em lote e em fluxo	Carga periódica ou contínua de dados. `ex.: lote noturno do LIMS; fluxo contínuo de sensores`
Streaming e mensageria	Eventos em tempo real. `ex.: Apache Kafka, MQTT para IoT`
ETL e ELT	Extrair, transformar e carregar, em ordens diferentes. `ex.: dbt transformando dados clínicos no warehouse`

Orquestração	Agendar e encadear tarefas. `ex.: Apache Airflow, Prefect, Dagster`
Workflows científicos	Pipelines de análise portáveis e reprodutíveis. `ex.: Nextflow (nf-core), Snakemake, CWL, WDL, Galaxy`
Contêineres	Empacotar o software com suas dependências. `ex.: Docker, Apptainer/Singularity`
HPC e nuvem	Computação de alto desempenho e elástica. `ex.: cluster SLURM, AWS, Google Cloud, Azure`
Formatos de arquivo biológicos	Formatos padrão por tipo de dado. `ex.: FASTQ, BAM/CRAM, VCF, FASTA, GFF, mzML, AnnData, OME-Zarr`
Formatos analíticos	Formatos colunares e de matrizes grandes. `ex.: Parquet, Arrow, HDF5, Zarr`
Compressão especializada	Reduzir o volume sem perder informação. `ex.: CRAM com referência`
Transferência de grandes volumes	Mover terabytes com segurança. `ex.: Globus, Aspera, checksums`
APIs e integração	Acesso programático a dados e serviços. `ex.: REST, GraphQL, APIs GA4GH (htsget, DRS)`
Captura de mudanças (CDC)	Replicar alterações de um banco em outro. `ex.: Debezium`
Contratos de dados	Acordo formal sobre o formato e a qualidade entregues. `ex.: contrato entre laboratório e equipe de análise`
Infraestrutura como código	Infraestrutura descrita em arquivos versionados. `ex.: Terraform, Ansible`

06

GUARDA · 13 TÓPICOS

Armazenamento e bancos de dados

Onde e como o dado fica guardado, de acordo com a sua forma e o seu uso.

TÓPICO	O QUE É · EXEMPLO NA BIOLOGIA
Bancos relacionais	Tabelas com relações e integridade. `ex.: PostgreSQL com amostras, indivíduos e ensaios`
Bancos de documentos	Registros flexíveis em JSON. `ex.: MongoDB com fichas de campo variáveis`
Bancos de grafos	Relações como cidadãs de primeira classe. `ex.: Neo4j com redes de interação proteína–proteína`
Bancos de séries temporais	Leituras indexadas no tempo. `ex.: TimescaleDB, InfluxDB para sensores`
Bancos geoespaciais	Geometrias e consultas espaciais. `ex.: PostGIS com ocorrências de espécies`
Bancos analíticos colunares	Consultas rápidas sobre grandes volumes. `ex.: DuckDB, ClickHouse`
Bancos vetoriais	Busca por similaridade em embeddings. `ex.: embeddings de proteínas para busca de homólogos`
Triplestores (RDF/SPARQL)	Dados ligados a ontologias. `ex.: grafo de conhecimento biomédico`
Bancos genômicos especializados	Estruturas otimizadas para variantes e matrizes. `ex.: TileDB-VCF`
Armazenamento de objetos	Arquivos grandes acessados por URL. `ex.: S3, MinIO com arquivos FASTQ e imagens`
Camadas quente, fria e de arquivo	Custo de acordo com a frequência de acesso. `ex.: dados brutos antigos em arquivo frio ou fita`

Backup e recuperação	Cópias e plano para desastres. `ex.: regra 3-2-1, testes de restauração`
Modelagem física e índices	Particionamento, índices e desempenho. `ex.: partição por ano de coleta`

07

ORGANIZAÇÃO · 13 TÓPICOS

Arquiteturas analíticas

Como os dados são organizados para análise, relatórios e decisões.

TÓPICO	O QUE É · EXEMPLO NA BIOLOGIA
Data warehouse	Repositório central, integrado e histórico para análise. `ex.: warehouse com todos os ensaios do laboratório`
Data mart	Recorte do warehouse para uma área ou pergunta. `ex.: mart de sanidade animal; mart de resistência antimicrobiana`
Data lake	Dados brutos em qualquer formato, guardados como chegaram. `ex.: lake com FASTQ, imagens e planilhas`
Lakehouse	Lake com transações, esquema e versões de tabela. `ex.: Delta Lake, Apache Iceberg`
Arquitetura medalhão	Camadas bronze (bruto), prata (limpo) e ouro (pronto). `ex.: leituras → variantes filtradas → painel de variantes`
Modelagem dimensional (Kimball)	Tabelas de fatos e dimensões, em estrela ou floco de neve. `ex.: fato: medição; dimensões: indivíduo, tempo, local, método`
Abordagem corporativa	Warehouse normalizado que alimenta os marts.

(Inmon)	`ex.: modelo integrado de uma rede de laboratórios`
Data Vault	Hubs, links e satélites para histórico e auditoria. `ex.: histórico completo de reanotações`
OLAP e cubos	Análise multidimensional agregada. `ex.: casos por espécie × região × mês`
Data mesh e produtos de dados	Dados como produtos, sob responsabilidade de cada domínio. `ex.: domínio 'genômica' publicando o seu produto de variantes`
Data fabric	Camada de integração e metadados sobre várias fontes. `ex.: catálogo unificado de várias instituições`
Ambientes de pesquisa seguros (TRE)	Analisar dados sensíveis sem tirá-los do lugar. `ex.: análise federada de genomas de pacientes`
Modelos comuns de dados	Esquemas padronizados para integrar fontes. `ex.: OMOP CDM, i2b2, HL7 FHIR`

08

CONTEXTO · 13 TÓPICOS

Curadoria e enriquecimento

Dar significado ao dado: nomes corretos, termos padronizados e ligações com o conhecimento existente.

TÓPICO	O QUE É · EXEMPLO NA BIOLOGIA
Biocuradoria	Revisão especializada e anotação de dados biológicos. `ex.: curadoria de função gênica a partir da literatura`
Anotação genômica e funcional	Atribuir genes, funções e vias. `ex.: termos da Gene Ontology em genes novos`
Ontologias	Vocabulários formais com relações lógicas. `ex.: GO, ChEBI, Uberon, NCBITaxon, ENVO, OBI,`

	HPO, Plant Ontology
Terminologias clínicas	Códigos padronizados para saúde. ex.: SNOMED CT, LOINC, CID-10/11
Taxonomia e nomenclatura	Validar e atualizar nomes científicos. ex.: Flora e Funga do Brasil, GBIF Backbone
Padrões de metadados	Campos e formatos acordados pela comunidade. ex.: Darwin Core, ABCD, MIxS, ISA-Tab, DataCite, Bioschemas
Mapeamento de identificadores	Ligar o mesmo objeto entre bases. ex.: UniProt ↔ Ensembl ↔ NCBI Gene
Resolução de entidades e deduplicação	Encontrar registros que são a mesma coisa. ex.: mesmo espécime cadastrado em dois herbários
Integração de dados heterogêneos	Unir fontes com esquemas diferentes. ex.: clínica + genoma + ambiente
Dados ligados e web semântica	Publicar dados com URIs e RDF. ex.: links entre ocorrências e taxonomia
Mineração de texto e NLP	Extrair fatos de artigos e laudos. ex.: extrair espécie e local de artigos antigos
Grafos de conhecimento	Rede de entidades e relações para consulta e inferência. ex.: gene–doença–fármaco
Curadoria assistida por IA	Sugestões automáticas validadas por especialistas. ex.: LLM sugerindo termos de ontologia

09

RASTRO · 12 TÓPICOS

Versionamento, proveniência e reprodutibilidade

Saber exatamente qual versão do dado, do código e do conhecimento produziu cada resultado.

TÓPICO	O QUE É · EXEMPLO NA BIOLOGIA

Versionamento de código	Histórico de mudanças em scripts e pipelines. `ex.: Git, GitHub, GitLab`
Versionamento de dados	Versões de arquivos e tabelas grandes. `ex.: DVC, lakeFS, git-annex, Dolt`
Versões de referência	Registrar genoma, ontologia e taxonomia usados. `ex.: GRCh37 → GRCh38; release da GO`
Versionamento de esquemas	Migrações controladas do banco. `ex.: Flyway, Alembic`
Viagem no tempo (time travel)	Consultar uma tabela como ela era no passado. `ex.: Iceberg/Delta: estado do dado na data do artigo`
Versão semântica de datasets	Numerar versões com significado. `ex.: v2.1.0: correção de 12 amostras`
Proveniência formal	Modelos que descrevem a origem e as transformações. `ex.: W3C PROV, ISO 23494, RO-Crate`
Linhagem de dados	Caminho de cada campo, da fonte ao relatório. `ex.: OpenLineage`
Ambientes reprodutíveis	Fixar versões de software. `ex.: conda, Docker, Nix, renv`
Documentos executáveis	Texto, código e resultado juntos. `ex.: Jupyter, Quarto, R Markdown`
Rastreamento de experimentos de ML	Parâmetros, dados e métricas de cada treino. `ex.: MLflow, Weights & Biases`
Identificadores persistentes	Nomes estáveis para dados, pessoas e instituições. `ex.: DOI, ARK, ORCID, ROR, RRID, IGSN`

10

DEVOLUÇÃO AO MUNDO · 13 TÓPICOS

Depósito, publicação e preservação

Entregar o dado à comunidade de forma citável e garantir que ele dure décadas.

TÓPICO	O QUE É · EXEMPLO NA BIOLOGIA
Repositórios de sequências	Depósito obrigatório para sequências publicadas. `ex.: INSDC (GenBank, ENA, DDBJ), SRA`
Repositórios de domínio	Bases especializadas por tipo de dado. `ex.: GEO, ArrayExpress, PRIDE, MetaboLights, PDB, BOLD`
Biodiversidade	Publicação de ocorrências e coleções. `ex.: GBIF, SiBBr, speciesLink`
Acesso controlado	Dados sensíveis liberados por comitê. `ex.: EGA, dbGaP, comitê de acesso (DAC)`
Repositórios generalistas	Qualquer tipo de dado, com DOI. `ex.: Zenodo, Figshare, Dryad, OSF, Dataverse institucional`
Citação de dados	Dados citados como publicações. `ex.: DOI do dataset na lista de referências`
Data papers	Artigos que descrevem um conjunto de dados. `ex.: Scientific Data, GigaScience`
Licenças de dados	Regras claras de reúso. `ex.: CC0, CC BY 4.0, ODbL`
Embargo	Período antes da liberação pública. `ex.: liberação após a publicação do artigo`
Princípios FAIR e CARE	Dados reutilizáveis e respeito a povos indígenas. `ex.: FAIR + CARE em dados de comunidades tradicionais`
Repositórios confiáveis	Certificação de repositórios. `ex.: CoreTrustSeal, princípios TRUST`
Preservação digital	Garantir leitura e integridade ao longo de décadas.

	ex.: modelo OAIS (ISO 14721), checksums de fixidez, formatos abertos
Políticas de retenção e descarte	Quanto tempo guardar e como apagar. ex.: dados brutos mantidos por 10 anos

11

CONHECIMENTO · 20 TÓPICOS

Análise e ciência de dados

Transformar dados em conhecimento, levando a incerteza e a variação biológica em conta.

TÓPICO	O QUE É · EXEMPLO NA BIOLOGIA
Estatística descritiva e inferencial	Resumo, testes e intervalos de confiança. ex.: comparação de ganho de peso entre dietas
Modelos mistos e hierárquicos	Efeitos fixos e aleatórios, medidas repetidas. ex.: animal como efeito aleatório
Estatística bayesiana	Incerteza expressa como distribuição. ex.: estimativa de prevalência com conhecimento prévio
Análise multivariada	Muitas variáveis ao mesmo tempo. ex.: PCA, PERMANOVA, análise de agrupamentos
Análise de sobrevivência	Tempo até um evento. ex.: Kaplan–Meier, regressão de Cox
Séries temporais	Padrões e previsão ao longo do tempo. ex.: sazonalidade de casos de dengue
Estatística espacial	Padrões no espaço. ex.: krigagem, autocorrelação espacial
Genética estatística	Associação entre genótipo e fenótipo. ex.: GWAS, herdabilidade, seleção genômica
Bioinformática de sequências	Alinhamento, montagem, variantes, expressão. ex.: BWA, GATK, DESeq2

Filogenia e evolução	Relações evolutivas. `ex.: árvores filogenéticas, relógio molecular`
Aprendizado de máquina	Modelos supervisionados e não supervisionados. `ex.: classificar tecidos, prever fenótipos`
Aprendizado profundo	Redes neurais para imagens, sequências e estruturas. `ex.: AlphaFold, modelos de linguagem de proteínas`
Visão computacional	Análise automática de imagens. `ex.: contagem de células, identificação de espécies por foto`
Inferência causal	Separar causa de correlação. `ex.: randomização mendeliana`
Modelagem matemática	Modelos de processos biológicos. `ex.: SIR em epidemiologia, dinâmica populacional, farmacocinética`
Modelagem de nicho	Distribuição potencial das espécies. `ex.: MaxEnt com cenários climáticos`
Biologia de sistemas	Redes e modelos de vias metabólicas. `ex.: SBML, análise de fluxo metabólico`
Meta-análise	Combinar resultados de vários estudos. `ex.: efeito médio de um tratamento`
Propagação de incerteza	Levar o erro das medidas até o resultado. `ex.: intervalo de confiança que considera a qualidade por base`
Validação e vazamento de dados	Avaliar modelos sem enganar a si mesmo. `ex.: validação cruzada por animal, e não por amostra`

12

CONHECIMENTO · 9 TÓPICOS

Visualização e comunicação

Mostrar o que os dados dizem, e com que grau de certeza.

TÓPICO	O QUE É · EXEMPLO NA BIOLOGIA
Painéis e BI	Indicadores atualizados para decisões. `ex.: Metabase, Apache Superset, Power BI`
Gráficos científicos	Figuras para artigos. `ex.: ggplot2, matplotlib, Plotly`
Navegadores genômicos	Visualizar leituras e anotações no genoma. `ex.: IGV, JBrowse, UCSC Genome Browser`
Visualização de redes	Grafos de interações. `ex.: Cytoscape`
Mapas	Distribuições no espaço. `ex.: QGIS, mapas web`
Estruturas 3D	Proteínas e moléculas. `ex.: PyMOL, Mol*`
Visualização de incerteza	Mostrar erro e variação, não só médias. `ex.: intervalos, distribuições, bandas de confiança`
Relatórios e narrativa com dados	Comunicar para quem decide. `ex.: relatório automático para o produtor rural`
Acessibilidade	Cores e formatos legíveis para todos. `ex.: paletas seguras para daltonismo`

13

RESPONSABILIDADE · 13 TÓPICOS

Governança, ética e segurança

Quem decide, quem pode acessar e como proteger pessoas, biodiversidade e instituições.

TÓPICO	O QUE É · EXEMPLO NA BIOLOGIA
Governança de dados	Papéis, políticas e responsabilidades.

	ex.: dono, curador (data steward) e custodiante de cada base
Catálogo de dados	Inventário pesquisável de dados e metadados. ex.: CKAN, DataHub, OpenMetadata
Gestão de dados mestres (MDM)	Uma única versão de entidades-chave. ex.: cadastro único de indivíduos e de espécies
Proteção de dados pessoais	Bases legais, direitos do titular, encarregado (DPO). ex.: LGPD, GDPR, HIPAA
Dados pessoais sensíveis	Regras reforçadas para dados genéticos e de saúde. ex.: LGPD art. 5º, II; GDPR art. 9
Anonimização e pseudonimização	Reduzir o risco de identificação. ex.: k-anonimato, privacidade diferencial
Consentimento	Amplo, específico ou dinâmico. ex.: TCLE com opção de uso futuro
Segurança da informação	Confidencialidade, integridade e disponibilidade. ex.: criptografia, controle de acesso (RBAC/ABAC), auditoria
Patrimônio genético e repartição de benefícios	Regras de acesso à biodiversidade. ex.: Lei 13.123/2015, SisGen, Protocolo de Nagoya, DSI
Soberania de dados indígenas	Controle das comunidades sobre os seus dados. ex.: princípios CARE
Biossegurança e uso dual	Evitar mau uso de informações sensíveis. ex.: sequências de patógenos de alto risco
Conformidade regulatória	Requisitos para pesquisa clínica e regulada. ex.: Boas Práticas Clínicas, 21 CFR Part 11, ALCOA+
Ética em IA	Viés, transparência e explicabilidade. ex.: modelo treinado só em uma população

14

SUSTENTAÇÃO · 8 TÓPICOS

Operações de dados (DataOps e MLOps)

Manter tudo funcionando com qualidade, custo controlado e melhoria contínua.

TÓPICO	O QUE É · EXEMPLO NA BIOLOGIA
DataOps	Práticas ágeis e automação para fluxos de dados. `ex.: entrega contínua de novas tabelas`
MLOps	Colocar e manter modelos em produção. `ex.: retreino automático com dados novos`
Observabilidade de dados	Monitorar volume, frescor, esquema e distribuição. `ex.: alerta quando o sensor para de enviar`
Testes de dados e CI/CD	Testar pipelines a cada mudança. `ex.: teste que barra amostras sem data de coleta`
SLAs e SLOs	Compromissos de prazo e qualidade. `ex.: laudo em até 48 h`
Gestão de incidentes	Resposta a falhas e correções. `ex.: reprocessar corrida corrompida`
FinOps	Controle de custo em nuvem. `ex.: custo por genoma processado`
Computação sustentável	Reduzir consumo de energia. `ex.: pegada de carbono de pipelines`

15

LINGUAGEM COMUM · 11 TÓPICOS

Interoperabilidade e padrões

Normas e padrões que permitem que dados de lugares diferentes conversem.

TÓPICO	O QUE É · EXEMPLO NA BIOLOGIA
GA4GH	Padrões globais para dados genômicos e de saúde.

	ex.: Phenopackets, Beacon, VRS, htsget
HL7 FHIR	Troca de informações de saúde. ex.: resultado laboratorial como recurso FHIR
DICOM	Imagens médicas e seus metadados. ex.: exames de imagem veterinária
OMOP CDM	Modelo comum para dados observacionais de saúde. ex.: rede de hospitais com análises federadas
OBO Foundry	Princípios e coleção de ontologias biológicas abertas. ex.: ontologias interoperáveis
Darwin Core e ABCD	Dados de biodiversidade. ex.: publicação no GBIF
MIxS e checklists MIBBI	Metadados mínimos por tipo de experimento. ex.: amostra ambiental de metagenoma
SBML e CellML	Modelos de biologia de sistemas. ex.: modelo metabólico compartilhável
ISO 20387	Requisitos para biobancos. ex.: biobanco acreditado
ISO 23494	Proveniência em biotecnologia. ex.: rastreio da amostra ao dado
Bioschemas	Marcação de páginas web para buscadores. ex.: dataset encontrável no Google Dataset Search

16

ORGANIZAÇÃO · 8 TÓPICOS

Pessoas, métodos e maturidade

Os papéis, os métodos de trabalho e os modelos de referência que estruturam uma área de dados.

TÓPICO	O QUE É · EXEMPLO NA BIOLOGIA

Papéis profissionais	Quem faz o quê. `ex.: engenheiro de dados, cientista de dados, bioinformata, biocurador, data steward, analista de BI, arquiteto de dados, engenheiro de ML, DPO`
DAMA-DMBOK	Corpo de conhecimento em gestão de dados. `ex.: áreas de conhecimento de gestão de dados`
CRISP-DM	Processo de projetos de ciência de dados. `ex.: entendimento do negócio → implantação`
Modelos de maturidade	Onde a organização está e para onde ir. `ex.: avaliação de maturidade em gestão de dados`
Letramento em dados	Formação para todos usarem dados bem. `ex.: cursos estilo The Carpentries`
Comunidades de prática	Troca entre profissionais. `ex.: fórum Biodadica (.net)`
Gestão de projetos de dados	Planejamento, entregas e priorização. `ex.: roadmap de produtos de dados`
Ciência aberta	Transparência em todo o ciclo. `ex.: código, dados e protocolos abertos`

CAPÍTULO 08

As 20 ferramentas da plataforma

A plataforma Biodadica vai reunir, num só ecossistema, tudo o que é preciso para produzir, documentar, analisar, compartilhar e aprender sobre dados biológicos. Estas são as ferramentas planejadas, organizadas pelos três domínios do projeto.

Todas as ferramentas estão em fase de planejamento. Acompanhe em **biodadica.org/ferramentas**.

.org Conhecimento aberto

Referência pública, padrões abertos e ferramentas de dados gratuitas.

01 · WIKI — FASE 2

Wiki Biodadica

A enciclopédia aberta e colaborativa da área: conceitos, métodos, boas práticas e estudos de caso, escritos e revisados pela comunidade.

✓ artigos revisados por pares da comunidade
✓ histórico de versões de cada página ✓ PT e EN lado a lado

biodadica.org/wiki

02 · ONTOLOGIA — FASE 2

Ontologia Biodadica

O vocabulário formal do ciclo de vida do dado biológico (coleta, intervalo, incerteza, proveniência), em OWL e alinhado às ontologias da OBO Foundry.

✓ identificadores persistentes por termo
✓ mapeamentos para OBI, ENVO e Uberon ✓ releases versionadas

biodadica.org/ontologia

03 · FERRAMENTA DE ONTOLOGIA — FASE 2

Navegador de ontologias

Busque um termo e encontre o código padronizado certo em várias ontologias ao mesmo tempo, com sugestões para anotar planilhas.

✓ busca em PT e EN ✓ anotação de colunas de planilhas
✓ sugestão de termos com IA, validada por curadores

`biodadica.org/ontologia/busca`

04 · GLOSSÁRIO — FASE 2

Glossário PT ↔ EN

Os termos da Biodadica em português e inglês, com definição, origem e exemplos. A porta de entrada para quem chega na área.

✓ definições curtas e citáveis ✓ equivalências entre idiomas
✓ link para a ontologia

`biodadica.org/glossario`

05 · ESPECIFICAÇÃO — FASE 1

Padrão da Ficha do Dado Biológico

A especificação aberta da ficha que acompanha cada dado: valor, incerteza, tempo, contexto, proveniência e permissões.

✓ JSON Schema e LinkML ✓ exemplos por tipo de dado
✓ compatível com Darwin Core e MIxS

`biodadica.org/data-tools/ficha`

06 · DATA TOOL — FASE 2

Validador de fichas

Envie uma planilha ou arquivo e receba um relatório: campos faltando, datas sem fuso horário, intervalos acima do limite e termos inválidos.

✓ roda no navegador, sem enviar dados sensíveis ✓ relatório em PDF
✓ regras configuráveis por protocolo

`biodadica.org/data-tools/validador`

07 · PACOTE R — FASE 2

Pacote R biodadica

Crie, valide e exporte fichas no R. Integra com o tidyverse e com o Bioconductor.

✓ ficha(), validar(), exportar() ✓ propagação de incerteza
✓ alertas de protocolo

`biodadica.org/data-tools/r`

08 · PACOTE PYTHON — FASE 2

Pacote Python biodadica

A mesma ficha em Python, com integração ao pandas, ao AnnData e a uma interface de linha de comando.

✓ bd.Ficha e bd.validar() ✓ leitura de FASTQ/VCF com metadados
✓ comando bdd no terminal

`biodadica.org/data-tools/python`

09 · API — FASE 3

API e conectores

Uma API aberta para registrar, consultar e validar fichas a partir de LIMS, cadernos eletrônicos e pipelines.

✓ REST + OpenAPI ✓ conectores para LIMS e ELN ✓ webhooks de alertas

`biodadica.org/data-tools/api`

.com Plataforma e produtos

Onde os dados são planejados, processados, depositados e transformados.

10 · BLOG — FASE 1

Blog Biodadica

Notícias, tutoriais, entrevistas e histórias de quem trabalha com dados biológicos no Brasil e no mundo.

✓ tutoriais passo a passo ✓ casos reais de laboratórios ✓ newsletter mensal

`biodadica.com/blog`

11 · LOJA DE MODELOS FASE 3

Loja de modelos estatísticos

Modelos estatísticos e de aprendizado de máquina prontos para usar, testados e documentados com a sua própria ficha.

✓ curvas de crescimento, GWAS, modelos de nicho
✓ modelos gratuitos e comerciais
✓ cartão de modelo com dados de treino e limites

biodadica.com/store

12 · PROTOCOLOS FASE 2

Biblioteca de protocolos

Protocolos de coleta e laboratório versionados, com os metadados obrigatórios e os limites (como o tempo máximo até o processamento) já definidos.

✓ DOI para cada versão ✓ campos que viram regras do validador
✓ comentários e adaptações da comunidade

biodadica.com/protocol

13 · PLANEJAMENTO FASE 2

Gerador de Plano de Gestão de Dados

Responda a algumas perguntas e receba um Plano de Gestão de Dados pronto para editais de fomento.

✓ modelos para agências de fomento ✓ exporta para Word e PDF
✓ orçamento de armazenamento

biodadica.com/tool-data/pgd

14 · PLANEJAMENTO FASE 2

Calculadora amostral

Calcule o tamanho da amostra e o poder estatístico para desenhos comuns em biologia, com a justificativa pronta para o projeto.

✓ testes comuns e modelos mistos ✓ esforço amostral em ecologia
✓ texto pronto para CEUA/CEP

biodadica.com/tool-data/amostragem

15 · WORKFLOWS FASE 3

Catálogo de pipelines

Pipelines prontos e reprodutíveis que já geram a ficha Biodadica de cada resultado.

✓ Nextflow e Snakemake ✓ contêineres prontos
✓ relatórios de qualidade automáticos

biodadica.com/tool-data/pipelines

16 · DEPÓSITO DE DADOS FASE 3

Repositório Biodadica

Deposite conjuntos de dados com ficha completa, DOI e licença, prontos para citação e para envio aos repositórios internacionais.

✓ DOI por versão ✓ acesso aberto, embargado ou controlado
✓ envio facilitado para GBIF, SRA e Zenodo

biodadica.com/tool-data/repositorio

.net Rede e formação

O lugar das pessoas: conversar, aprender e se encontrar.

17 · FÓRUM FASE 3

Fórum Biodadica

O lugar para perguntar, responder e trocar experiências sobre dados biológicos, do campo ao código.

✓ categorias por área do mapa ✓ respostas marcadas como solução
✓ busca integrada à wiki

biodadica.net/forum

18 · AMBIENTE VIRTUAL DE APRENDIZAGEM FASE 3

Escola Biodadica (AVA)

Cursos on-line em trilhas, do básico ao avançado, com exercícios práticos usando dados reais e certificados.

✓ trilhas: coleta, qualidade, engenharia, análise
✓ exercícios com os pacotes R e Python ✓ certificados verificáveis

biodadica.net/escola

19 · EVENTOS FASE 3

Eventos e hackathons

Webinars, oficinas, maratonas de curadoria e hackathons para construir as ferramentas junto com a comunidade.

✓ agenda pública ✓ gravações na Escola ✓ desafios com dados abertos

biodadica.net/eventos

20 · COMUNIDADE FASE 3

Diretório de especialistas

Encontre pessoas e laboratórios por área de atuação, para colaborar, contratar ou pedir mentoria.

✓ perfis ligados ao ORCID ✓ busca por área do mapa ✓ programa de mentoria

biodadica.net/pessoas

CAPÍTULO 09

O ecossistema e o caminho

A Biodadica vai crescer como um ecossistema aberto, com três frentes, cada uma com o seu propósito.

.org

biodadica.org · biodadic.org

Conhecimento aberto. A referência pública da área: wiki, ontologia e ferramentas de dados (data tools).

.com

biodadica.com · biodadic.com

Plataforma e produtos. Blog, loja, protocolos e ferramentas para armazenar, padronizar e transformar dados biológicos.

.net

biodadica.net · biodadic.net

Rede e formação. Fórum para a comunidade e escola para quem quer aprender.

Roadmap

FASE	ENTREGA
0 · Manifesto (agora)	O nome, a definição, este ebook, o mapa de áreas e o convite às primeiras pessoas.
1 · Artigo fundador	Publicação do framework com DOI: escopo, taxonomia e casos de uso.
2 · Wiki e ontologia	Vocabulário aberto e glossário em português e inglês.
3 · Plataforma	Primeiras ferramentas de dados biológicos, fórum e escola.

Participe

Uma área nova se escreve com quem chega primeiro. Se você produz, cuida, transforma ou analisa dados de seres vivos, a Biodadica é sua.

Bioinformatas, biólogos, agrônomos, profissionais de saúde, engenheiros de dados, estatísticos, estudantes e educadores: venham ajudar a definir o vocabulário, as ferramentas e a comunidade.

biodadica.org

APÊNDICE

Glossário

Alíquota

Parte de uma amostra separada para um uso específico.

Biblioteca (de sequenciamento)

Conjunto de fragmentos de DNA ou RNA preparados para entrar no sequenciador.

Cadeia de custódia / proveniência

Registro de todos os passos, pessoas e equipamentos por onde o dado passou.

Efeito de lote

Diferença nos resultados causada por condições técnicas (dia, equipamento, reagente, operador), e não pela biologia.

FAIR

Princípios para dados encontráveis (Findable), acessíveis (Accessible), interoperáveis (Interoperable) e reutilizáveis (Reusable).

Genoma de referência

Sequência padrão de uma espécie usada para comparar e anotar outras sequências. É atualizado em versões.

Ontologia

Vocabulário controlado e estruturado que define termos e as relações entre eles.

Phred (escala)

Medida da probabilidade de erro de uma base sequenciada. $Q = -10 \cdot \log_{10}(P)$. Q30 equivale a 1 erro em 1.000.

Réplica biológica e réplica técnica

A biológica repete o experimento com outro indivíduo ou outra amostra; a técnica mede de novo a mesma amostra.

SisGen

Sistema Nacional de Gestão do Patrimônio Genético e do Conhecimento Tradicional Associado, previsto na Lei 13.123/2015.

TCLE

Termo de Consentimento Livre e Esclarecido, assinado por quem participa de uma pesquisa.

APÊNDICE

Referências

1. EWING, B.; GREEN, P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*, v. 8, n. 3, p. 186–194, 1998.
2. WILKINSON, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, v. 3, 160018, 2016.
3. LEEK, J. T. et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, v. 11, p. 733–739, 2010.
4. ZIEMANN, M.; EREN, Y.; EL-OSTA, A. Gene name errors are widespread in the scientific literature. *Genome Biology*, v. 17, 177, 2016.
5. BRUFORD, E. A. et al. Guidelines for human gene nomenclature. *Nature Genetics*, v. 52, p. 754–758, 2020.
6. SCHNEIDER, V. A. et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*, v. 27, p. 849–864, 2017.
7. MEINEKE, E. K.; DAVIS, C. C.; DAVIES, T. J. The unrealized potential of herbaria for global change biology. *Ecological Monographs*, v. 88, n. 4, p. 505–525, 2018.
8. ISO 20186-1:2019. Molecular in vitro diagnostic examinations — Specifications for pre-examination processes for venous whole blood — Part 1: Isolated cellular RNA.
9. BRASIL. Lei nº 13.709, de 14 de agosto de 2018. Lei Geral de Proteção de Dados Pessoais (LGPD).
10. BRASIL. Lei nº 13.123, de 20 de maio de 2015. Dispõe sobre o acesso ao patrimônio genético e ao conhecimento tradicional associado.
11. UNIÃO EUROPEIA. Regulamento (UE) 2016/679. Regulamento Geral sobre a Proteção de Dados (GDPR).
12. CONVENÇÃO SOBRE DIVERSIDADE BIOLÓGICA. Protocolo de Nagoya sobre Acesso a Recursos Genéticos e Repartição Justa e Equitativa dos Benefícios Derivados de sua Utilização. 2010.



Toda vida gera dados. Biodadica é a ciência de cuidar deles.

Inventada no Brasil · biodadica.org