

EBOOK · VERSION 0.2 · NOVEMBER 2026

Biodadic

The science of the biological data lifecycle: why living data needs a discipline of its own.

Biodadic Project · Biodadica

biodadic.org

Contents

01 A field that already existed, without a name
The gap between biology and data

02 The name: Biodadic and Biodadica
Bio + dad + ic, and why the D stays in English

03 Why biological data is special
Twelve reasons for a discipline of its own

04 The framework: six stages
From production to visualization

05 The biological data record
The six pillars that travel with every value

06 Biodadic and neighboring fields
What changes compared with bioinformatics and others

07 A map of data areas
16 major areas and 196 topics, from design to preservation

08 The platform's 20 tools
Wiki, ontology, R and Python packages, model store, school...

09 The ecosystem and the path ahead
.org, .com, .net and the roadmap

– Glossary and references

Biodadic: the science of the biological data lifecycle. Preliminary version 0.2, November 2026. Biodadic Project (in Portuguese, Biodadica). Translated from the original Brazilian Portuguese edition.

A living document, revised with the community. Suggestions are welcome at biodadic.org.

How to cite: BIODADIC PROJECT. *Biodadic: the science of the biological data lifecycle*. Version 0.2. 2026. Available at: https://biodadic.org.

CHAPTER 01

A field that already existed, without a name

All life creates data. A sequenced genome, a microscopy image, a sensor reading in a cattle barn, a species occurrence recorded by a researcher in the field, a blood test result: all of these are biological data.

Over the last two decades, the volume of this data has grown faster than any lab imagined. Sequencing became cheap, sensors became small, and instruments started producing results automatically, at scale. With that came a huge amount of work that goes far beyond "analysis": data must be **produced** following a protocol, its quality **checked**, **organized** in reliable systems, **enriched** with context, **analyzed**, and what it reveals **communicated**.

Today that work is spread across several fields, each taking care of one piece:

- **bioinformatics** handles algorithms and computational analysis, mainly of sequences and molecules;
- **biostatistics** handles experimental design and inference;
- **biocuration** handles annotation and the organization of databases;
- **data engineering** handles flows, storage and integration, but usually without knowing the biology behind them;
- **data science** brings analysis and machine learning methods, but

treats biological data like any other data.

Nobody looks at the whole cycle. And it is precisely in the hand-offs between stages that the most information is lost: the collection time that was never written down, the read quality that was thrown away, the reference genome version nobody recorded, the consent form kept on paper, separate from the file.

> Biodadic is born to give a name and a method to this complete cycle: from the moment something alive is measured to the moment that measurement becomes reliable, reusable knowledge.

Biodadic is the discipline devoted to the complete lifecycle of biological data (production, quality control, engineering, enrichment, analysis and visualization) and to turning it into knowledge that is reliable, open and reusable.

CHAPTER 02

The name: Biodadic and Biodadica

A new field needs a word. We built this one piece by piece, the way languages have always named fields of knowledge.

bio	dad	ic
life	**data**	**science**
Greek βίος	Port. dado ← Lat. datum	Greek -ικός

Bio comes from the Greek *bíos*, life: organisms, cells, genomes, ecosystems, crops, herds, patients. **Dad** comes from *dado*, the everyday Brazilian Portuguese word for "data", which in turn comes from the Latin *datum*, "that which is given". And **-ic** is the suffix of fields of knowledge, as in genetic, electronic, informatic and statistic.

Together, the three parts say exactly what the field is: **the science of the data of life**.

Why a D and not a T

The obvious English translation would replace the D with the T of *data* and spell it "biodatic". We don't, for three reasons:

- **Datum means "dado".** Latin *datum* literally means "that which is given". In Portuguese that is *dado*, with a D. The Brazilian word is the most faithful translation of the Latin root.
- **A traceable origin.** The D is a signature. Anyone who finds *Biodadic* in a paper, anywhere in the world, will know the concept has Brazilian roots.
- **A name without noise.** In English, *biodata* already means "biographical data", as in a résumé. *Biodadic* is new, short and easy to search for.

> Just as the world says *bossa nova* and *capoeira*, the world says **Biodadic**: a Brazilian word, **Biodadica**, traveling abroad.

LANGUAGE	OFFICIAL SPELLING	NOTE
Portuguese	**Biodadica**	No accent. Pronounced bi·o·da·di·ca.
English	**Biodadic**	With a D. Pronounced /ˌbaɪ.oʊ ˈdæd.ɪk/.
Spanish	**Biodadic**	Same spelling as English.

CHAPTER 03

Why biological data is special

A sales figure doesn't change once it's recorded. A blood sample does. Biological data is a snapshot of something alive, and living things change, vary, err, evolve and belong to someone.

Each of the characteristics below exists, on its own, in other kinds of data. What makes biological data special is that **they all happen at the same time**. That combination is what justifies a discipline of its own.

01 · UNCERTAINTY

Every measurement carries uncertainty

Sequencers don't read DNA, they estimate it. Every sequenced base comes with an error probability, expressed on the Phred scale: Q20 means 1 error in 100 bases and Q30 means 1 in 1,000. Mass spectrometers, sensors and automated microscopes work the same way: they deliver a value and a confidence.

When the confidence is discarded and only the value is kept, the data loses half its meaning. In Biodadic, uncertainty is a mandatory field.

```
Q30 → 99.9% accuracy per base
```

02 · TIME

The clock keeps running after collection

Once removed from the organism, the material stays alive for a while: cells consume nutrients, RNA degrades, proteins change and gene expression responds to the stress of collection itself. A sample processed 2 hours after collection is no longer the one it was 10 minutes after.

That is why the collection time, the processing time and the interval between them are part of the data. An interval above the protocol's limit should trigger an alert, not be forgotten.

```
collected 08:14 · processed 10:41 · Δ 2h27 ⚠
```

03 · MOMENT

Time of day and stage of life matter

Circadian rhythm, fasting, season, developmental stage and reproductive cycle all change the results. The same individual gives different numbers at different moments. Without that context, a comparison may be measuring nothing but coincidence.

```
cortisol: high in the morning, low at night
```

04 · INTERFERENCE

Observing changes what is observed

Capturing an animal raises its stress, a biopsy triggers inflammation, and freezing and thawing break cells. The act of collecting changes the biology itself, so how it was done must be recorded as a condition of the data.

`temperature · preservative · freeze–thaw cycles`

05 · VARIABILITY

No two living beings are the same

There is variation between individuals, between cells in the same tissue and between repeats of an experiment. You must tell biological replicates (another individual) from technical replicates (the same sample measured again), and real effects from batch effects (differences caused by the day, the instrument or the operator). An average without its variation does not describe life.

`biological replicate ≠ technical replicate`

06 · UNREPEATABLE

Samples run out and moments don't return

Tissue is finite and perishable. Patients recover, crops are harvested, species can vanish. Often there is no second measurement: the data is the only record left of that organism at that moment. Losing it, or losing its context, means losing that moment of life forever.

`unique, unrepeatable material`

07 · REFERENCE

Knowledge itself gets new versions

Reference genomes are updated (for example, from GRCh37 to GRCh38 for humans), species get new names and genes are renamed. Even Excel has forced changes: genes such as SEPT1 and MARCH1 were automatically converted into dates and, in 2020, were renamed SEPTIN1 and MARCHF1. All biological data must state which version of knowledge it was interpreted against.

`SEPT1 → 1-Sep → SEPTIN1`

08 · PRIVACY

Biological data belongs to someone

A genome identifies a person and reveals information about their whole family, including relatives who never consented. Perfect anonymization of genetic data does not exist. Brazil's LGPD classifies genetic and biometric data as sensitive personal data, and Europe's GDPR does the same. Consent, purpose and access rules are part of the data, not paperwork beside it.

`LGPD art. 5 II · GDPR art. 9`

09 · SOVEREIGNTY

Biodiversity is heritage

In Brazil, access to genetic heritage and associated traditional knowledge is regulated by Law 13,123/2015 and must be registered in SisGen. Worldwide, it is governed by the Nagoya Protocol. Where a sample comes from has legal consequences and determines how the benefits of its use are shared. For a megadiverse country like Brazil, this is a central issue.

`Law 13,123/2015 · SisGen · Nagoya`

10 · SCALE

From molecule to ecosystem

Biological data spans scales (molecules, cells, organs, individuals, populations, biomes) and formats (sequences, images, time series, occurrence records, clinical text). Integrating all of it takes shared ontologies and identifiers, not just tables.

`genome · image · sensor · occurrence`

11 · PROVENANCE

A complete chain of custody

Data passes through many hands and machines before reaching a conclusion. Any step can introduce error or contamination. Without end-to-end traceability, results cannot be reproduced and errors cannot be located.

organism → sample → aliquot → library → run → file → analysis

who · when · where · how

12 · LONGEVITY

Value that lasts for centuries

Herbarium specimens collected more than 100 years ago now reveal the effects of climate change and preserve ancient DNA. Well-documented biological data can answer questions nobody has asked yet, as long as it is findable, accessible, interoperable and reusable: the FAIR principles.

FAIR principles

> Ordinary data describes things. Biological data describes something that was alive, and a moment later had already changed.

CHAPTER 04

The framework: six stages

Biodadic organizes everything that happens to a piece of biological data into six stages. Each stage has a goal and a set of questions whose answers must be recorded.

1 Production → 2 Quality → 3 Engineering → 4 Enrichment → 5 Analysis → 6 Visualization

1 Production

Generate data following a protocol: collection, sequencing, sensor readings, images, field observations.

- Which organism, tissue and individual?
- When was it collected, and when was it processed?
- With which protocol, instrument and operator?

2 Quality control

Check whether the data is reliable enough for its intended use, and record its uncertainty.

- What is the confidence of each measurement?
- Did the time to processing stay within the limit?
- Are there signs of contamination or batch effects?

3 Engineering

Store, move and integrate data in reproducible pipelines, using open formats.

- Where is the data, and in which format?
- Which pipeline version processed it?
- Can the data's path be rebuilt from scratch?

4 Enrichment

Give data context: annotation, curation, ontologies and links to other databases.

- Which standardized terms describe this data?
- Against which reference version was it annotated?
- Which other databases does it connect to?

5 Analysis

Extract knowledge: statistics, bioinformatics, models and artificial intelligence, taking uncertainty into account.

- Was measurement uncertainty propagated to the result?
- Was biological variation separated from technical variation?
- Can someone else reproduce the analysis?

6 Visualization and release

Communicate what was found and give the data back to the world, in a findable, reusable form.

- Does the chart also show uncertainty?
- Who can access the data, and under which license?
- Does the data follow the FAIR principles?

CHAPTER 05

The biological data record

Biodadic's central proposal is simple: a measured value never travels alone. It carries its own story, from the first record to the last reuse.

Value · Uncertainty · Time · Context · Provenance · Permissions

PILLAR	WHAT IT RECORDS	EXAMPLE
Value	The measurement itself.	a sequence, a concentration, an image
Uncertainty	The confidence of the measurement.	Q32; ± 0.4 mg/dL; limit of detection
Time	Collection, processing and the interval between them.	2h27min, over the 2 h limit
Context	Collection and storage conditions; organism; reference used.	4 °C, EDTA, GRCh38.p14
Provenance	The full path, from organism to file.	sample → library → run
Permissions	Consent, license, legal registrations.	consent form, LGPD/GDPR, SisGen

An example

Below is the record of a sequencing read from a human blood sample. Note the automatic alert: the interval between collection and processing exceeded 2 hours, and that is stored with the data, forever.

```
{
  "value":       "ATGGCGTACCTGA…",
  "uncertainty": "Q32 · 99.94% per base",
  "collected":   "2026-11-12T08:14-03:00",
  "processed":   "2026-11-12T10:41-03:00",
  "interval":    "2h27min",  // ⚠ over 2h
  "condition":   "4 °C · EDTA · 0 thaws",
  "organism":    "Homo sapiens · NCBITaxon:9606",
  "reference":   "GRCh38.p14",
  "provenance":  "sample BD-0192 → lib L07 → run R2311",
  "permission":  "consent 2026-118 · sensitive data"
}
```

Golden rules

- **No value without uncertainty.** If the confidence is unknown, that is recorded too.
- **Every timestamp has a time zone.** Dates without time zones become ambiguous in cross-country projects.
- **Every annotation cites its version.** Genome, ontology, pipeline and taxonomy.
- **Alerts are never erased.** A protocol deviation may invalidate an analysis, but it must never be hidden.
- **Permissions travel with the data.** If the file is copied, the rules of use go with it.

CHAPTER 06

Biodadic and neighboring fields

Biodadic doesn't replace any existing field. It stitches together what is now separate and takes responsibility for the whole cycle.

FIELD	MAIN FOCUS	WHAT BIODADIC ADDS
Bioinformatics	Algorithms and computational analysis of sequences and molecules.	Starts earlier (at collection) and ends later (at reuse); covers any living data.
Biostatistics	Experimental design and inference.	Ensures uncertainty and context reach the analysis intact.
Computational biology	Models of biological systems.	Provides data with provenance to feed the models.
Biocuration	Annotation and organization of databases.	Connects curation to production and source quality.
Data engineering	Flows, storage and integration.	Brings the rules specific to living data: time, variability, ethics.
Data science	Analysis and machine learning methods.	Prevents biological data from being treated like any other data.

Where Biodadic applies

- **Health:** clinical laboratories, biobanks, patient research.
- **Agriculture and livestock:** seed and herd genetics, field sensors, animal health.
- **Biodiversity:** collections, herbaria, monitoring of species and ecosystems.
- **Biotechnology:** bioprocesses, fermentation, organism engineering.
- **Education:** training professionals who understand both biology and data.

CHAPTER 07

A map of data areas

An organized list of the data disciplines, practices, architectures and standards that apply to biological data, from project planning to preservation for decades. Each topic includes an example from biology.

The searchable online version is at **biodadic.org/areas**.

01

BEFORE THE DATA · 10 TOPICS

Planning and design

Decide what will be measured, how, by whom, with which permissions and how much it will cost to keep.

TOPIC	WHAT IT IS · EXAMPLE IN BIOLOGY
Data Management Plan (DMP)	Document describing how data will be generated, documented, stored, shared and preserved. e.g. DMP required by funders such as FAPESP
Experimental design	Structure of the experiment: treatments, controls, blocks, randomization and replicates. e.g. randomized block field trial
Sample size and statistical power	How many units are needed to detect an effect with confidence. e.g. number of animals per group approved by the ethics committee
Study pre-registration	Public record of hypotheses and analysis before collection. e.g. clinical trial registered in ReBEC or ClinicalTrials.gov

Requirements and use cases	Who will use the data, for what and how often. e.g. herd health dashboard for the veterinarian
Conceptual data modeling	Domain entities and relationships before any database. e.g. Individual → Sample → Aliquot → Assay
Minimum metadata checklists	Minimum information required for each type of experiment. e.g. MIxS (sequences), MIAME (microarrays), MIAPE (proteomics)
Ethical and legal approvals	Authorizations required before collecting. e.g. human research ethics (CEP/CONEP), animal ethics (CEUA), SISBIO, SisGen registration
Data budget	Cost of generating, processing, storing and preserving. e.g. cost of keeping 500 genomes for 10 years
Project risk management	Sample loss, equipment failure, data leaks. e.g. plan B for a −80 °C freezer failure

02

DATA ORIGIN · 12 TOPICS

Sampling and collection

The moment something alive becomes a record. Much of the data's quality is decided here.

TOPIC	WHAT IT IS · EXAMPLE IN BIOLOGY
Probability sampling	Simple random, stratified, systematic and cluster sampling. e.g. stratified draw of farms by region
Ecological sampling	Plots, transects, traps, point counts, capture–recapture. e.g. bird transect in an Atlantic Forest fragment
Sampling effort	How much was sampled and whether it was enough. e.g. species accumulation curve

Temporal and longitudinal sampling	Repeated collections over time, from the same individual or site. `e.g. cohort sampled every 6 months`
Pre-analytical variables	Everything that happens between collection and analysis. `e.g. time to processing (alert above 2 h), temperature, anticoagulant`
Cold chain and logistics	Sample transport and storage. `e.g. dry ice, liquid nitrogen, temperature loggers`
Sample identification	Unique, machine-readable codes. `e.g. barcode, QR code, RFID, IGSN`
Georeferencing	Coordinates, reference system and precision. `e.g. SIRGAS 2000 / WGS84, uncertainty in meters`
Digital field collection	Mobile forms that validate at the moment of collection. `e.g. ODK, KoboToolbox`
Voucher specimens	Specimen deposited in a collection to confirm identification. `e.g. herbarium voucher`
Biobanks and biorepositories	Organized storage of biological material and its data. `e.g. serum biobank with broad consent`
Citizen science	Collection by the public, with validation. `e.g. iNaturalist records`

03

DATA ORIGIN · 14 TOPICS

Data generation and acquisition

Instruments, sensors and systems that turn a sample into digital data.

TOPIC	WHAT IT IS · EXAMPLE IN BIOLOGY

DNA/RNA sequencing	Short and long reads, with per-base quality. `e.g. Illumina, Oxford Nanopore, PacBio → FASTQ`
Omics	Large-scale measurement of molecules. `e.g. genomics, transcriptomics, proteomics, metabolomics, epigenomics, metagenomics`
Single-cell and spatial	Cell-by-cell measurements or with position in the tissue. `e.g. scRNA-seq, spatial transcriptomics`
Biological imaging	Microscopy, digital histology, image-based phenotyping. `e.g. digital slides in OME-TIFF`
Medical imaging	Imaging exams with clinical metadata. `e.g. CT scans in DICOM`
Remote sensing and drones	Aerial and satellite imagery. `e.g. crop NDVI, vegetation mapping`
Sensors and IoT	Continuous, automatic readings. `e.g. GPS collar, soil sensor, weather station, robotic milking`
Laboratory instruments	Analytical equipment and their raw files. `e.g. mass spectrometry, flow cytometry, qPCR, ELISA`
Bioacoustics	Recording and identifying animal sounds. `e.g. autonomous recorders for amphibians`
Clinical data	Electronic health records, lab tests, outcomes. `e.g. blood count results linked to the sample`
LIMS	Laboratory information management system. `e.g. tracking samples, assays and reports`
Electronic lab notebook (ELN)	Digital record of what was done at the bench. `e.g. the day's protocol, deviations and notes`

Instrument integration	Automatic capture of files and metadata from machines. `e.g. middleware sending the sequencer run to the LIMS`
Synthetic and simulated data	Model-generated data for testing or privacy. `e.g. synthetic cohort to train a pipeline`

04

TRUST · 12 TOPICS

Data quality

Measure, record and improve trust in each piece of data. In Biodadic, uncertainty is a mandatory field.

TOPIC	WHAT IT IS · EXAMPLE IN BIOLOGY
Quality dimensions	Completeness, accuracy, consistency, timeliness, uniqueness and validity. `e.g. % of samples without collection time`
Confidence scores	Error probability attached to a measurement. `e.g. Phred (Q30), FDR in proteomics, limit of detection`
Sequence quality control	Assessment and cleaning of reads. `e.g. FastQC, MultiQC, adapter trimming`
Controls and standards	Known references to check the process. `e.g. positive and negative controls, blanks, spike-ins, reference material`
Batch effects and normalization	Separate technical from biological variation. `e.g. ComBat correction`
Contamination detection	Identify material that shouldn't be there. `e.g. human DNA in an environmental sample`
Outlier detection	Unusual values: error or biology? `e.g. an animal weighing 10× the average`

Data profiling	Statistical portrait of each field. e.g. distribution, nulls and distinct values per column
Rule and schema validation	Automated tests on the data. e.g. Great Expectations, Pandera, JSON Schema
Nonconformities and deviations	Permanent record of departures from protocol. e.g. sample processed 2h27 after collection
Data integrity (ALCOA+)	Attributable, legible, contemporaneous, original and accurate data, plus complete, consistent, enduring and available. e.g. clinical study records
Quality standards	Laboratory accreditation and good practice. e.g. ISO 15189, ISO/IEC 17025, GLP

05

FLOW · 15 TOPICS

Data engineering

Move, transform and integrate data with reproducible, scalable pipelines.

TOPIC	WHAT IT IS · EXAMPLE IN BIOLOGY
Batch and streaming ingestion	Periodic or continuous data loading. e.g. nightly LIMS batch; continuous sensor stream
Streaming and messaging	Real-time events. e.g. Apache Kafka, MQTT for IoT
ETL and ELT	Extract, transform and load, in different orders. e.g. dbt transforming clinical data in the warehouse
Orchestration	Scheduling and chaining tasks. e.g. Apache Airflow, Prefect, Dagster
Scientific workflows	Portable, reproducible analysis pipelines. e.g. Nextflow (nf-core), Snakemake, CWL, WDL, Galaxy

Containers	Package software with its dependencies. e.g. Docker, Apptainer/Singularity
HPC and cloud	High-performance and elastic computing. e.g. SLURM cluster, AWS, Google Cloud, Azure
Biological file formats	Standard formats per data type. e.g. FASTQ, BAM/CRAM, VCF, FASTA, GFF, mzML, AnnData, OME-Zarr
Analytical formats	Columnar and large-array formats. e.g. Parquet, Arrow, HDF5, Zarr
Specialized compression	Reduce volume without losing information. e.g. reference-based CRAM
Large data transfer	Move terabytes safely. e.g. Globus, Aspera, checksums
APIs and integration	Programmatic access to data and services. e.g. REST, GraphQL, GA4GH APIs (htsget, DRS)
Change data capture (CDC)	Replicate changes from one database to another. e.g. Debezium
Data contracts	Formal agreement on delivered format and quality. e.g. contract between the lab and the analysis team
Infrastructure as code	Infrastructure described in versioned files. e.g. Terraform, Ansible

06

STORAGE · 13 TOPICS

Storage and databases

Where and how data is kept, according to its shape and its use.

TOPIC	WHAT IT IS · EXAMPLE IN BIOLOGY
Relational databases	Tables with relationships and integrity. e.g. PostgreSQL with samples, individuals and

	`assays`
Document databases	Flexible JSON records. `e.g. MongoDB with variable field forms`
Graph databases	Relationships as first-class citizens. `e.g. Neo4j with protein–protein interaction networks`
Time-series databases	Readings indexed by time. `e.g. TimescaleDB, InfluxDB for sensors`
Geospatial databases	Geometries and spatial queries. `e.g. PostGIS with species occurrences`
Columnar analytical databases	Fast queries over large volumes. `e.g. DuckDB, ClickHouse`
Vector databases	Similarity search over embeddings. `e.g. protein embeddings to search for homologs`
Triplestores (RDF/SPARQL)	Data linked to ontologies. `e.g. biomedical knowledge graph`
Specialized genomic databases	Structures optimized for variants and arrays. `e.g. TileDB-VCF`
Object storage	Large files accessed by URL. `e.g. S3, MinIO with FASTQ files and images`
Hot, cold and archive tiers	Cost according to access frequency. `e.g. old raw data in cold storage or tape`
Backup and recovery	Copies and a disaster plan. `e.g. 3-2-1 rule, restore tests`
Physical modeling and indexes	Partitioning, indexes and performance. `e.g. partition by collection year`

07

ORGANIZATION · 13 TOPICS

Analytical architectures

How data is organized for analysis, reporting and decisions.

TOPIC	WHAT IT IS · EXAMPLE IN BIOLOGY
Data warehouse	Central, integrated, historical repository for analysis. e.g. warehouse with all of a lab's assays
Data mart	Subset of the warehouse for an area or question. e.g. animal health mart; antimicrobial resistance mart
Data lake	Raw data in any format, kept as it arrived. e.g. lake with FASTQ, images and spreadsheets
Lakehouse	A lake with transactions, schemas and table versions. e.g. Delta Lake, Apache Iceberg
Medallion architecture	Bronze (raw), silver (clean) and gold (ready) layers. e.g. reads → filtered variants → variant dashboard
Dimensional modeling (Kimball)	Fact and dimension tables, in star or snowflake schemas. e.g. fact: measurement; dimensions: individual, time, place, method
Enterprise approach (Inmon)	Normalized warehouse feeding the marts. e.g. integrated model for a laboratory network
Data Vault	Hubs, links and satellites for history and audit. e.g. full history of re-annotations
OLAP and cubes	Aggregated multidimensional analysis. e.g. cases by species × region × month
Data mesh and data products	Data as products, owned by each domain. e.g. the 'genomics' domain publishing its variant product

Data fabric	Integration and metadata layer across many sources. e.g. unified catalog across institutions
Trusted research environments (TRE)	Analyze sensitive data without moving it. e.g. federated analysis of patient genomes
Common data models	Standardized schemas to integrate sources. e.g. OMOP CDM, i2b2, HL7 FHIR

08

CONTEXT · 13 TOPICS

Curation and enrichment

Give data meaning: correct names, standardized terms and links to existing knowledge.

TOPIC	WHAT IT IS · EXAMPLE IN BIOLOGY
Biocuration	Expert review and annotation of biological data. e.g. gene function curation from the literature
Genomic and functional annotation	Assign genes, functions and pathways. e.g. Gene Ontology terms for new genes
Ontologies	Formal vocabularies with logical relationships. e.g. GO, ChEBI, Uberon, NCBITaxon, ENVO, OBI, HPO, Plant Ontology
Clinical terminologies	Standardized health codes. e.g. SNOMED CT, LOINC, ICD-10/11
Taxonomy and nomenclature	Validate and update scientific names. e.g. Flora e Funga do Brasil, GBIF Backbone
Metadata standards	Community-agreed fields and formats. e.g. Darwin Core, ABCD, MIxS, ISA-Tab, DataCite, Bioschemas
Identifier mapping	Link the same object across databases. e.g. UniProt ↔ Ensembl ↔ NCBI Gene
Entity resolution and	Find records that are the same thing.

deduplication	e.g. the same specimen registered in two herbaria
Heterogeneous data integration	Combine sources with different schemas. e.g. clinical + genome + environment
Linked data and semantic web	Publish data with URIs and RDF. e.g. links between occurrences and taxonomy
Text mining and NLP	Extract facts from papers and reports. e.g. extract species and locality from old papers
Knowledge graphs	Network of entities and relationships for querying and inference. e.g. gene–disease–drug
AI-assisted curation	Automatic suggestions validated by experts. e.g. an LLM suggesting ontology terms

09

TRAIL · 12 TOPICS

Versioning, provenance and reproducibility

Know exactly which version of the data, the code and the knowledge produced each result.

TOPIC	WHAT IT IS · EXAMPLE IN BIOLOGY
Code versioning	History of changes to scripts and pipelines. e.g. Git, GitHub, GitLab
Data versioning	Versions of large files and tables. e.g. DVC, lakeFS, git-annex, Dolt
Reference versions	Record the genome, ontology and taxonomy used. e.g. GRCh37 → GRCh38; GO release
Schema versioning	Controlled database migrations. e.g. Flyway, Alembic
Time travel	Query a table as it was in the past. e.g. Iceberg/Delta: the data as of the paper's date

Semantic versioning of datasets	Meaningful version numbers. e.g. v2.1.0: 12 samples corrected
Formal provenance	Models describing origin and transformations. e.g. W3C PROV, ISO 23494, RO-Crate
Data lineage	Path of each field from source to report. e.g. OpenLineage
Reproducible environments	Pin software versions. e.g. conda, Docker, Nix, renv
Executable documents	Text, code and results together. e.g. Jupyter, Quarto, R Markdown
ML experiment tracking	Parameters, data and metrics of each training run. e.g. MLflow, Weights & Biases
Persistent identifiers	Stable names for data, people and institutions. e.g. DOI, ARK, ORCID, ROR, RRID, IGSN

10

GIVING BACK · 13 TOPICS

Deposit, publication and preservation

Hand the data to the community in a citable form and make sure it lasts for decades.

TOPIC	WHAT IT IS · EXAMPLE IN BIOLOGY
Sequence repositories	Mandatory deposit for published sequences. e.g. INSDC (GenBank, ENA, DDBJ), SRA
Domain repositories	Specialized databases per data type. e.g. GEO, ArrayExpress, PRIDE, MetaboLights, PDB, BOLD
Biodiversity	Publishing occurrences and collections. e.g. GBIF, SiBBr, speciesLink
Controlled access	Sensitive data released by a committee.

	e.g. EGA, dbGaP, data access committee (DAC)
Generalist repositories	Any type of data, with a DOI. e.g. Zenodo, Figshare, Dryad, OSF, institutional Dataverse
Data citation	Data cited like publications. e.g. dataset DOI in the reference list
Data papers	Articles describing a dataset. e.g. Scientific Data, GigaScience
Data licenses	Clear reuse rules. e.g. CC0, CC BY 4.0, ODbL
Embargo	Period before public release. e.g. release after the paper is published
FAIR and CARE principles	Reusable data and respect for Indigenous peoples. e.g. FAIR + CARE for data from traditional communities
Trustworthy repositories	Repository certification. e.g. CoreTrustSeal, TRUST principles
Digital preservation	Keep data readable and intact for decades. e.g. OAIS model (ISO 14721), fixity checksums, open formats
Retention and disposal policies	How long to keep and how to delete. e.g. raw data kept for 10 years

11

KNOWLEDGE · 20 TOPICS

Analysis and data science

Turn data into knowledge, accounting for uncertainty and biological variation.

TOPIC	WHAT IT IS · EXAMPLE IN BIOLOGY
Descriptive and inferential statistics	Summaries, tests and confidence intervals. e.g. weight gain compared between diets

Mixed and hierarchical models	Fixed and random effects, repeated measures. `e.g. animal as a random effect`
Bayesian statistics	Uncertainty expressed as a distribution. `e.g. prevalence estimate with prior knowledge`
Multivariate analysis	Many variables at once. `e.g. PCA, PERMANOVA, cluster analysis`
Survival analysis	Time to an event. `e.g. Kaplan–Meier, Cox regression`
Time series	Patterns and forecasting over time. `e.g. seasonality of dengue cases`
Spatial statistics	Patterns in space. `e.g. kriging, spatial autocorrelation`
Statistical genetics	Association between genotype and phenotype. `e.g. GWAS, heritability, genomic selection`
Sequence bioinformatics	Alignment, assembly, variants, expression. `e.g. BWA, GATK, DESeq2`
Phylogenetics and evolution	Evolutionary relationships. `e.g. phylogenetic trees, molecular clock`
Machine learning	Supervised and unsupervised models. `e.g. classify tissues, predict phenotypes`
Deep learning	Neural networks for images, sequences and structures. `e.g. AlphaFold, protein language models`
Computer vision	Automated image analysis. `e.g. cell counting, species identification from photos`
Causal inference	Separate cause from correlation. `e.g. Mendelian randomization`

Mathematical modeling	Models of biological processes. `e.g. SIR epidemic models, population dynamics, pharmacokinetics`
Niche modeling	Potential species distribution. `e.g. MaxEnt under climate scenarios`
Systems biology	Networks and models of metabolic pathways. `e.g. SBML, metabolic flux analysis`
Meta-analysis	Combine results from several studies. `e.g. average effect of a treatment`
Uncertainty propagation	Carry measurement error through to the result. `e.g. confidence interval that accounts for per-base quality`
Validation and data leakage	Evaluate models without fooling yourself. `e.g. cross-validation by animal, not by sample`

12

KNOWLEDGE · 9 TOPICS

Visualization and communication

Show what the data says, and how certain it is.

TOPIC	WHAT IT IS · EXAMPLE IN BIOLOGY
Dashboards and BI	Up-to-date indicators for decisions. `e.g. Metabase, Apache Superset, Power BI`
Scientific plots	Figures for papers. `e.g. ggplot2, matplotlib, Plotly`
Genome browsers	View reads and annotations along the genome. `e.g. IGV, JBrowse, UCSC Genome Browser`
Network visualization	Interaction graphs. `e.g. Cytoscape`
Maps	Distributions in space. `e.g. QGIS, web maps`

3D structures	Proteins and molecules. e.g. PyMOL, Mol*
Uncertainty visualization	Show error and variation, not just averages. e.g. intervals, distributions, confidence bands
Reports and data storytelling	Communicate with decision-makers. e.g. automated report for the farmer
Accessibility	Colors and formats readable by everyone. e.g. colorblind-safe palettes

13

RESPONSIBILITY · 13 TOPICS

Governance, ethics and security

Who decides, who can access, and how to protect people, biodiversity and institutions.

TOPIC	WHAT IT IS · EXAMPLE IN BIOLOGY
Data governance	Roles, policies and responsibilities. e.g. owner, data steward and custodian of each database
Data catalog	Searchable inventory of data and metadata. e.g. CKAN, DataHub, OpenMetadata
Master data management (MDM)	A single version of key entities. e.g. single registry of individuals and species
Personal data protection	Legal bases, data subject rights, data protection officer (DPO). e.g. LGPD, GDPR, HIPAA
Sensitive personal data	Stricter rules for genetic and health data. e.g. LGPD art. 5 II; GDPR art. 9
Anonymization and pseudonymization	Reduce re-identification risk. e.g. k-anonymity, differential privacy

Consent	Broad, specific or dynamic. `e.g. consent form with an option for future use`
Information security	Confidentiality, integrity and availability. `e.g. encryption, access control (RBAC/ABAC), auditing`
Genetic heritage and benefit sharing	Rules for access to biodiversity. `e.g. Brazilian Law 13,123/2015, SisGen, Nagoya Protocol, DSI`
Indigenous data sovereignty	Communities' control over their data. `e.g. CARE principles`
Biosecurity and dual use	Prevent misuse of sensitive information. `e.g. sequences of high-risk pathogens`
Regulatory compliance	Requirements for clinical and regulated research. `e.g. Good Clinical Practice, 21 CFR Part 11, ALCOA+`
AI ethics	Bias, transparency and explainability. `e.g. a model trained on a single population`

14

SUSTAINMENT · 8 TOPICS

Data operations (DataOps and MLOps)

Keep everything running with quality, controlled cost and continuous improvement.

TOPIC	WHAT IT IS · EXAMPLE IN BIOLOGY
DataOps	Agile practices and automation for data flows. `e.g. continuous delivery of new tables`
MLOps	Deploy and maintain models in production. `e.g. automatic retraining with new data`
Data observability	Monitor volume, freshness, schema and distribution. `e.g. alert when a sensor stops reporting`

Data tests and CI/CD	Test pipelines on every change. `e.g. a test that blocks samples without a collection date`
SLAs and SLOs	Commitments on timing and quality. `e.g. report within 48 h`
Incident management	Responding to failures and fixes. `e.g. reprocess a corrupted run`
FinOps	Cloud cost control. `e.g. cost per processed genome`
Sustainable computing	Reduce energy consumption. `e.g. carbon footprint of pipelines`

15

COMMON LANGUAGE · 11 TOPICS

Interoperability and standards

Norms and standards that let data from different places talk to each other.

TOPIC	WHAT IT IS · EXAMPLE IN BIOLOGY
GA4GH	Global standards for genomic and health data. `e.g. Phenopackets, Beacon, VRS, htsget`
HL7 FHIR	Exchange of health information. `e.g. a lab result as a FHIR resource`
DICOM	Medical images and their metadata. `e.g. veterinary imaging exams`
OMOP CDM	Common model for observational health data. `e.g. hospital network with federated analyses`
OBO Foundry	Principles and collection of open biological ontologies. `e.g. interoperable ontologies`
Darwin Core and ABCD	Biodiversity data. `e.g. publishing to GBIF`

MIxS and MIBBI checklists	Minimum metadata per experiment type. e.g. environmental metagenome sample
SBML and CellML	Systems biology models. e.g. shareable metabolic model
ISO 20387	Requirements for biobanks. e.g. accredited biobank
ISO 23494	Provenance in biotechnology. e.g. tracing from sample to data
Bioschemas	Web page markup for search engines. e.g. dataset findable in Google Dataset Search

16

ORGANIZATION · 8 TOPICS

People, methods and maturity

The roles, working methods and reference models that structure a data field.

TOPIC	WHAT IT IS · EXAMPLE IN BIOLOGY
Professional roles	Who does what. e.g. data engineer, data scientist, bioinformatician, biocurator, data steward, BI analyst, data architect, ML engineer, DPO
DAMA-DMBOK	Data management body of knowledge. e.g. knowledge areas of data management
CRISP-DM	Process for data science projects. e.g. business understanding → deployment
Maturity models	Where the organization is and where to go. e.g. data management maturity assessment
Data literacy	Training so everyone uses data well. e.g. courses in the style of The Carpentries
Communities of	Exchange among professionals.

practice	`e.g. the Biodadic forum (.net)`
Data project management	Planning, delivery and prioritization. `e.g. roadmap of data products`
Open science	Transparency across the whole cycle. `e.g. open code, data and protocols`

CHAPTER 08

The platform's 20 tools

The Biodadic platform will bring together, in a single ecosystem, everything needed to produce, document, analyze, share and learn about biological data. These are the planned tools, organized by the project's three domains.

All tools are in the planning stage. Follow them at **biodadic.org/tools**.

.org Open knowledge

Public reference, open standards and free data tools.

01 · WIKI — PHASE 2

Biodadic Wiki

The open, collaborative encyclopedia of the field: concepts, methods, good practices and case studies, written and reviewed by the community.

✓ community peer-reviewed articles ✓ version history for every page
✓ EN and PT side by side

biodadic.org/wiki

02 · ONTOLOGY — PHASE 2

Biodadic Ontology

The formal vocabulary of the biological data lifecycle (collection, interval, uncertainty, provenance), in OWL and aligned with the OBO Foundry ontologies.

✓ persistent identifiers per term ✓ mappings to OBI, ENVO and Uberon
✓ versioned releases

biodadic.org/ontology

03 · ONTOLOGY TOOL — PHASE 2

Ontology browser

Search for a term and find the right standardized code across several ontologies at once, with suggestions for annotating spreadsheets.

✓ search in EN and PT ✓ spreadsheet column annotation
✓ AI term suggestions, validated by curators

`biodadic.org/ontology/search`

04 · GLOSSARY — PHASE 2

EN ↔ PT Glossary

Biodadic terms in English and Portuguese, with definition, origin and examples. The entry point for newcomers.

✓ short, citable definitions ✓ equivalents across languages
✓ links to the ontology

`biodadic.org/glossary`

05 · SPECIFICATION — PHASE 1

Biological Data Record standard

The open specification of the record that accompanies each piece of data: value, uncertainty, time, context, provenance and permissions.

✓ JSON Schema and LinkML ✓ examples per data type
✓ compatible with Darwin Core and MIxS

`biodadic.org/data-tools/record`

06 · DATA TOOL — PHASE 2

Record validator

Upload a spreadsheet or file and get a report: missing fields, dates without time zones, intervals over the limit and invalid terms.

✓ runs in the browser, without sending sensitive data ✓ PDF report
✓ rules configurable per protocol

`biodadic.org/data-tools/validator`

07 · R PACKAGE PHASE 2

biodadica R package

Create, validate and export records in R. Integrates with the tidyverse and Bioconductor.

✓ ficha(), validar(), exportar() ✓ uncertainty propagation ✓ protocol alerts

biodadic.org/data-tools/r

08 · PYTHON PACKAGE PHASE 2

biodadica Python package

The same record in Python, with pandas and AnnData integration and a command-line interface.

✓ bd.Ficha and bd.validar() ✓ read FASTQ/VCF with metadata
✓ bdd command in the terminal

biodadic.org/data-tools/python

09 · API PHASE 3

API and connectors

An open API to register, query and validate records from LIMS, electronic lab notebooks and pipelines.

✓ REST + OpenAPI ✓ LIMS and ELN connectors ✓ alert webhooks

biodadic.org/data-tools/api

.com Platform and products

Where data is planned, processed, deposited and transformed.

10 · BLOG PHASE 1

Biodadic Blog

News, tutorials, interviews and stories from people who work with biological data in Brazil and around the world.

✓ step-by-step tutorials ✓ real lab cases ✓ monthly newsletter

biodadic.com/blog

11 · MODEL STORE PHASE 3

Statistical model store

Ready-to-use statistical and machine learning models, tested and documented with their own record.

✓ growth curves, GWAS, niche models ✓ free and commercial models
✓ model cards with training data and limits

biodadic.com/store

12 · PROTOCOLS PHASE 2

Protocol library

Versioned collection and lab protocols, with mandatory metadata and limits (such as the maximum time to processing) already defined.

✓ a DOI for every version ✓ fields that become validator rules
✓ community comments and adaptations

biodadic.com/protocol

13 · PLANNING PHASE 2

Data Management Plan generator

Answer a few questions and get a Data Management Plan ready for funding calls.

✓ templates for funding agencies ✓ export to Word and PDF ✓ storage budget

biodadic.com/tool-data/dmp

14 · PLANNING PHASE 2

Sample size calculator

Calculate sample size and statistical power for common biological designs, with a justification ready for your proposal.

✓ common tests and mixed models ✓ sampling effort in ecology
✓ text ready for ethics committees

biodadic.com/tool-data/sampling

15 · WORKFLOWS PHASE 3

Pipeline catalog

Ready, reproducible pipelines that already produce the Biodadic record for each result.

✓ Nextflow and Snakemake ✓ ready-made containers
✓ automatic quality reports

biodadic.com/tool-data/pipelines

16 · DATA DEPOSIT PHASE 3

Biodadic Repository

Deposit datasets with a complete record, DOI and license, ready to be cited and submitted to international repositories.

✓ DOI per version ✓ open, embargoed or controlled access
✓ easy submission to GBIF, SRA and Zenodo

biodadic.com/tool-data/repository

.net Network and learning

The place for people: talk, learn and meet.

17 · FORUM PHASE 3

Biodadic Forum

The place to ask, answer and share experience about biological data, from the field to the code.

✓ categories by area of the map ✓ answers marked as solutions
✓ search integrated with the wiki

biodadic.net/forum

18 · LEARNING MANAGEMENT SYSTEM PHASE 3

Biodadic School (LMS)

Online courses in learning paths, from beginner to advanced, with hands-on exercises on real data and certificates.

✓ paths: collection, quality, engineering, analysis
✓ exercises with the R and Python packages ✓ verifiable certificates

biodadic.net/school

19 · EVENTS PHASE 3

Events and hackathons

Webinars, workshops, curation sprints and hackathons to build the tools together with the community.

✓ public calendar ✓ recordings in the School ✓ challenges with open data

`biodadic.net/events`

20 · COMMUNITY PHASE 3

Expert directory

Find people and labs by field, to collaborate, hire or ask for mentoring.

✓ ORCID-linked profiles ✓ search by area of the map ✓ mentoring program

`biodadic.net/people`

CHAPTER 09

The ecosystem and the path ahead

Biodadic will grow as an open ecosystem with three fronts, each with its own purpose.

.org

biodadic.org · biodadica.org

Open knowledge. The public reference for the field: wiki, ontology and data tools.

.com

biodadic.com · biodadica.com

Platform and products. Blog, store, protocols and tools to store, standardize and transform biological data.

.net

biodadic.net · biodadica.net

Network and learning. A forum for the community and a school for those who want to learn.

Roadmap

PHASE	DELIVERABLE
0 · Manifesto (now)	The name, the definition, this ebook, the map of areas and the invitation to the first people.
1 · Founding paper	The framework published with a DOI: scope, taxonomy and use cases.
2 · Wiki and ontology	An open vocabulary and glossary in English and Portuguese.
3 · Platform	The first biological data tools, the forum and the school.

Join us

A new field is written by those who arrive first. If you produce, curate, transform or analyze data from living things, Biodadic is yours.

Bioinformaticians, biologists, agronomists, health professionals, data engineers, statisticians, students and educators: come help define the vocabulary, the tools and the community.

biodadic.org

APPENDIX

Glossary

Aliquot

Portion of a sample set aside for a specific use.

Batch effect

Difference in results caused by technical conditions (day, instrument, reagent, operator) rather than biology.

Chain of custody / provenance

Record of every step, person and instrument the data passed through.

Consent form

Document signed by research participants agreeing to take part (in Brazil, the TCLE).

FAIR

Principles for data that is Findable, Accessible, Interoperable and Reusable.

Library (sequencing)

Set of DNA or RNA fragments prepared to be loaded into the sequencer.

Ontology

Controlled, structured vocabulary that defines terms and the relationships between them.

Phred (scale)

Measure of the error probability of a sequenced base. $Q = -10 \cdot \log_{10}(P)$. Q30 equals 1 error in 1,000.

Reference genome

Standard sequence of a species used to compare and annotate other sequences. It is updated in versions.

Biological and technical replicates

A biological replicate repeats the experiment with another individual or sample; a technical replicate measures the same sample again.

SisGen

Brazil's National System for the Management of Genetic Heritage and Associated Traditional Knowledge, created under Law 13,123/2015.

APPENDIX

References

1. Ewing B, Green P. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Research*. 1998;8(3):186–194.
2. Wilkinson MD et al. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*. 2016;3:160018.
3. Leek JT et al. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*. 2010;11:733–739.
4. Ziemann M, Eren Y, El-Osta A. Gene name errors are widespread in the scientific literature. *Genome Biology*. 2016;17:177.
5. Bruford EA et al. Guidelines for human gene nomenclature. *Nature Genetics*. 2020;52:754–758.
6. Schneider VA et al. Evaluation of GRCh38 and de novo haploid genome assemblies demonstrates the enduring quality of the reference assembly. *Genome Research*. 2017;27:849–864.
7. Meineke EK, Davis CC, Davies TJ. The unrealized potential of herbaria for global change biology. *Ecological Monographs*. 2018;88(4):505–525.
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. Brazil. Law No. 13,709 of 14 August 2018. General Personal Data Protection Law (LGPD).
10. Brazil. Law No. 13,123 of 20 May 2015. Access to genetic heritage and associated traditional knowledge.
11. European Union. Regulation (EU) 2016/679. General Data Protection Regulation (GDPR).
12. Convention on Biological Diversity. Nagoya Protocol on Access to Genetic Resources and the Fair and Equitable Sharing of Benefits Arising from their Utilization. 2010.



All life creates data. Biodadic is the science of caring for it.

Invented in Brazil · biodadic.org