Unifying the identification of biomedical entities with the Bioregistry

Charles Tapley Hoyt1, Meghan Balk2, Tiffany J. Callahan3, Daniel Domingo-Fernández4,5, Melissa A. Haendel3, Harshad B. Hegde6, Daniel S. Himmelstein7, Klas Karis1, John Kunze8, Tiago Lubiana9, Nicolas Matentzoglu10, Julie McMurry3, Sierra Moxon6, Christopher J. Mungall6, Adriano Rutz11,12, Deepak R. Unni6,13, Egon Willighagen14, Donald Winston15 & Benjamin M. Gyori1

The standardized identification of biomedical entities is a cornerstone of interoperability, reuse, and data integration in the life sciences. Several registries have been developed to catalog resources maintaining identifiers for biomedical entities such as small molecules, proteins, cell lines, and clinical trials. However, existing registries have struggled to provide sufficient coverage and metadata standards that meet the evolving needs of modern life sciences researchers. Here, we introduce the Bioregistry, an integrative, open, community-driven metaregistry that synthesizes and substantially expands upon 23 existing registries. The Bioregistry addresses the need for a sustainable registry by leveraging public infrastructure and automation, and employing a progressive governance model centered around open code and open data to foster community contribution. The Bioregistry can be used to support the standardized annotation of data, models, ontologies, and scientific literature, thereby promoting their interoperability and reuse. The Bioregistry can be accessed through https://bioregistry.io and its source code and data are available under the MIT and CC0 Licenses at https://github.com/biopragmatics/bioregistry.

Introduction

One of the key challenges in creating and maintaining FAIR (Findable, Accessible, Interoperable, and Reusable) data1-3 in the life sciences is the standardized identification of entities ranging from chemicals, proteins, and diseases to patents and publications. These entities are typically curated in identifier resources (e.g., ontologies and databases) such as the Chemical Entities of Biomedical Interest (ChEBI)4, UniProt5, and PubMed that assign to each entity a local unique identifier (i.e., accession number). Each resource defines an internally consistent pattern for its entities’ local unique identifiers, such as the combination of numbers and letters found in UniProt identifiers (e.g., P0DP23) or the simple numbers found in PubMed identifiers (e.g., 29175850). Uniform resource identifiers (URIs) (e.g., https://www.uniprot.org/uniprot/) and compact resource identifiers (CURIEs) (e.g., uniprot:P0DP23) have become the predominant syntaxes used in the life sciences for identifying entities that encode both the resource from which the entity originates and its local unique identifier. URIs encode the resource with a URI prefix (e.g., https://www.uniprot.org/uniprot/) while CURIEs encode it with a prefix (e.g., uniprot).

1Laboratory of Systems Pharmacology, Harvard Medical School, Boston, USA. 2University of Arizona, Tucson, USA. 3University of Colorado Anschutz Medical Campus, Aurora, USA. 4Department of Bioinformatics, Fraunhofer SCAI, Sankt Augustin, Germany. 5Enveda Biosciences, Boulder, USA. 6Lawrence Berkeley National Laboratory, Berkeley, USA. 7Related Sciences, Denver, USA. 8California Digital Library, University of California, Berkeley, USA. 9School of Pharmaceutical Sciences, University of São Paulo, São Paulo, Brazil. 10Semanticity, London, UK. 11School of Pharmaceutical Sciences, University of Geneva, Geneva, Switzerland. 12Institute of Pharmaceutical Sciences of Western Switzerland, University of Geneva, Geneva, Switzerland. 13European Molecular Biology Laboratory, Heidelberg, Germany. 14Department of Bioinformatics - BiGCaT, NUTRIM, Maastricht University, Maastricht, Netherlands. 15Polyneme LLC, New York, USA. *e-mail: benjamin_gyori@hms.harvard.edu
However, even when using URIs and CURIEs, a number of challenges remain in establishing consistency and interoperability. Namely, several different incompatible URIs and CURIEs can be used to refer to the same entity. For example, for the local identifier P0DPE2.3 for the entry in UniProt about the Calmodulin-1 protein, it can be represented by at least seven distinct URIs and three distinct CURIEs (see Supplementary Tables 1 and 2 for details). This problem is compounded when attempting to integrate multiple resources, a cornerstone of modern computational life sciences. For example, genomic data from the HUGO Gene Nomenclature Committee (HGNC) can not be readily integrated with biochemical reactions data from Rhea because HGNC uses the prefix ec-code and Rhea uses the prefix EC when referring to entities in the Enzyme Commission identifier resource. Similarly, many biomedical resources construct local unique identifiers for the same entity in the Enzyme Commission identifier resource differently, e.g., 1.4 in IntEnz, 1.4. -.-. in the Gene Ontology (GO), and 1.4.+. in ChEBI for Oxidoreductases acting on the CH-NH2 group of donors.

In order to standardize the usages of URIs and CURIEs and therefore enable their interoperability, a registry is needed containing canonical, validatable definitions of identifier resources that, for each resource, includes a prefix, a URI prefix, a local unique identifier pattern, and other associated metadata. Registries thus capture for each identifier resource how to construct, parse, and interchange canonical URIs and CURIEs. A registry can be used by external biomedical resources to standardize the way they reference entities (e.g., database cross-references appearing in ontologies) to promote integration with other resources, as well as by consumers to navigate prefixes and their associated metadata. Multiple registries have been previously built for this purpose, but they each suffer from substantial gaps in their coverage of known resources and the metadata captured about these resources. They also lack interoperability among each others’ entries, for example, the NCBI GenBank is prefixed as taxonomy in Identifiers.org and as NCBI_TAXON in BioPortal. These issues are exacerbated by shortcomings in existing registries’ governance and curation workflows, which impede their ability to stay current, trustworthy, and engage the community as the landscape of life science resources rapidly evolves. These issues include, that they (1) are built on private infrastructure within an institution; (2) are maintained by small, private groups that - due to under-funding - struggle to respond to requests; (3) lack adequate support for external contributions; and (4) are neither versioned nor archived. As an alternative to general-purpose registries, numerous projects (e.g., GO, Cellosaurus, NCBI GenBank) have created their own registries, however, these each only cover identifier resources relevant for the given project, and use standards that are only internally consistent to the project. Finally, several services act as registries but are by design limited in scope to include a selected set of resources and provide incomplete metadata necessary to promote the standardization of references. These include the Open Biomedical Ontologies (OBO) Foundry, the Ontology Lookup Service (OLS), and BioPortal. A detailed survey of the governance and maintenance models for existing registries can be found in Supplementary Table 3. Overall, the content of any one registry does not reflect the evolving landscape of biomedical resources and thus satisfy user needs.

To overcome the limitations of existing registries, a new approach for building a biomedical registry is necessary, which ideally fulfills the three requirements: (1) integrative, (2) open, and (3) community-driven. First, an integrative registry re-uses, improves, and extends on existing registries. Given that existing registries define conflicting standards (i.e., assign conflicting prefixes, URI prefixes, or other metadata for the same identifier resource) and therefore lack interoperability, this necessitates alignment and harmonization among resources in each registry. Second, an open registry makes its underlying data and associated code available under permissive licenses in a public, version-controlled repository, and relies on free, publicly available infrastructure for semi-automated quality control and deployment. Third, a community-driven registry solicits contributions from community members and provides an appropriate technical platform and governance structure to support this. This technical platform needs to support discussion and feedback as well as quality assurance workflows tightly coupled to the underlying open data, code, and infrastructure. Overall, these properties are expected to promote the sustainability and longevity of the registry.

To address these limitations, we introduce the Bioregistry: an integrative, open, and community-driven metaregistry. The Bioregistry integrates content from existing registries, semi-automatically identifies equivalences between records in existing registries, and resolves conflicts between them using a novel workflow. The result of this alignment makes the Bioregistry a metaregistry (i.e., a registry of registries) which, for each resource, maintains cross-registry mappings to serve as an interoperability layer between conflicting standards. The Bioregistry also includes substantial manual curation for resources not appearing in any pre-existing registry, and additional curation extends and improves on the metadata associated with resources in these registries with curation practices heavily influenced by the most similar existing registries (e.g., Identifiers.org and Prefix Commons). As a result of this process, the Bioregistry expands substantially on the content of each individual pre-existing registry (e.g., 81% over Identifiers.org) as well as all aligned registries combined, and incorporates feedback from the maintainers and developers of existing registries. The Bioregistry also provides a higher granularity data model compared to any existing registry thereby better supporting integration. The Bioregistry is built using open source code, open data, and leverages public infrastructure and automation to support its maintenance and extension. Further, it has well-defined contribution guidelines and a multi-institutional governance model that enables contributions directly from the broader community to support the project’s longevity.

The Bioregistry (0.5.132) integrates and aligns content previously curated in 23 external registries, and in total, contains 1,504 individual records. These records extend on each prior registry (as compared to e.g., 838 records in Prefix Commons and 774 in Identifiers.org), as well as all aligned registries combined: 192 of the Bioregistry’s 1,504 records are novel, i.e. they do not appear in any existing registry. The Bioregistry also adds novel curated metadata for 969 of the remaining 1,312 records (74% of all records). A summary of the content captured in Bioregistry is provided in Table 1. We provide detailed metrics and comparison to other resources in the Results section.
The Bioregistry has already been integrated with several projects aimed at data integration, knowledge assembly, and semantically annotated publications, including BridgeDb, PheKnowLator, Manubot, Biomappings, SSSOM, INDRA, and the OBO Foundry. The Bioregistry is available through an interactive web portal (https://bioregistry.io), an OpenAPI-documented web service (https://bioregistry.io/apidocs), a Python software package (https://pypi.org/project/bioregistry), and a Docker image (https://hub.docker.com/r/biopragmatics/bioregistry). All underlying data, code, and governance documentation are accessible through GitHub (https://github.com/biopragmatics/bioregistry) under the MIT and CC0 licenses and archived on Zenodo.

### Results

#### Bioregistry data model.

The Bioregistry uses a granular and extensible data model to represent records of identifier resources. Required fields for each record include a canonical prefix, a human-readable label, a homepage, and a description. The data model also allows for multiple optional fields including the license, version, prefix synonyms for the resource, and capturing whether the resource is deprecated or proprietary. Each record can further include an example local unique identifier, a regular expression pattern for validating local unique identifiers, and a URI format string for constructing URIs from local unique identifiers. Records describing ontologies can include optional download links for associated OBO, OWL, and OBO Graph JSON artifacts. To keep contributions traceable and provide attribution, each record captures the submitter and reviewer who contributed to the entry. Records can also be grouped into collections for better contextualization, such as prefixes useful for the Semantic Web (e.g., DC, FOAF, RDF, RDFS).

A comparison between the data model and various properties of the Bioregistry and external registries in Table 2 demonstrates the heterogeneity of metadata standards in external registries and the flexibility of the Bioregistry to represent more granular metadata. For example, this enables the Bioregistry to represent deprecated and obsolete records for posterity, such as hgnc.genefamily (the HGNC Gene Family resource) which was replaced by hgnc.genegroup, and casp (Eschmeyer's Catalog of Fishes) which is used by the Teleost Taxonomy Ontology and Vertebrate Taxonomy Ontology but was itself never published.

Importantly, the Bioregistry captures not only individual resource records but also semantic relationships between records (including e.g., `depends on`, which asserts that one resource uses terms from another, such as GO depends on ChEBI, and mappings of each resource to external registries where it appears. These additional relations constitute the Bioregistry's `metaregistry`, a term meant to represent the fact that it creates links among previously incompatible resources through a set of cross-registry mappings.

The Bioregistry data model is described in further detail in Supplementary Section 3.

#### Integration and harmonization of existing registries.

The Bioregistry imports records from the 23 external registries described in Table 2. We divided registries into distinct groups to which we applied different import policies ranging from registries imported entirely to ones from which metadata for only select records are imported (see Methods). The Bioregistry uses a multi-stage process in which registries are sequentially imported such that a record from a given registry is either (1) aligned with an existing Bioregistry record and a cross-registry mapping is created (2) added as a new record, or (3) set aside for manual curation (see Methods). The key challenge in this import process is aligning (i.e., finding equivalences between) records since the external registries' records are partially overlapping but inconsistent. This inconsistency stems from high heterogeneity among existing registries in the usage of capitalization (e.g., go vs. GO), punctuation (e.g. ec-code vs eccode), abbreviations (e.g., flybase vs. fb), and even different vocabulary (e.g., intenz vs. eccode) to represent the same resource, which results in fragmentation and lack of interoperability. A novel contribution of the Bioregistry is that it explicitly represents the results of its alignment procedure as equivalence mappings between its own records and records in external registries. This constitutes a network of cross-registry mappings (with the Bioregistry as the hub), creating a metaregistry. The Bioregistry's alignment procedure recovers a total of 7,150 such cross-registry mappings thereby connecting resources across existing registries that were previously disconnected. Importantly, the Bioregistry does not import nor redistribute the resources described by external registries, which themselves may have restrictive licenses (e.g., ICD-10). Further, rather than pushing enriched metadata back to the existing registries, the Bioregistry instead makes its content available under a permissive license in many formats (described in the section on Exported artifacts) that can be individually processed by each external registry's maintainers to best suit their needs.

We additionally curated 423 synonyms used when referring to identifier resources outside registries such as in OBO Foundry ontology database cross-references. These synonyms support the registry alignment workflow and broaden the ability of Bioregistry to standardize references to identifier resources beyond just external registries.

### Table 1. Overview statistics of the Bioregistry version 0.5.132 (2022-10-17).

| Category                          | Count |
|----------------------------------|-------|
| External registries imported     | 23    |
| Individual resources represented | 1,504 |
| Cross-registry resource mappings | 7,150 |
| Contributors                     | 36    |

---

3Scientific
We investigated how the content of the integrated Bioregistry compares to each individual external registry that it imports and aligns (Fig. 1a). It covers several registries (BioContext14, CHEMINF18, Crop Ontology Curation Tool (CropOCT)19, OBO Foundry27, OLS, Name-to-Thing (N2T)25, Identifiers.org24) almost entirely (over 85% of the external registries’ records are mapped to a Bioregistry record) while significantly expanding on the content of each of them from a minimum of +94% for Identifiers.org to a maximum of +7,076% for CHEMINF. The Bioregistry is able to align a smaller proportion of records in external registries such as FAIRSharing22 (33.3%), BioPortal16 (31.3%), Aber-OWL12 (23.8%), and Wikidata (20.7%) due to several characteristics of each registry. For example, many records in FAIRSharing 22 do not refer to identifier resources, Wikidata contains many records lacking a biological scope, and Aber-OWL and BioPortal contain ontologies of heterogeneous quality which are queued for on-demand inclusion rather than automated ingestion into the Bioregistry (see Methods). Despite lower coverage of their entries, the Bioregistry still substantially expands on these external registries between a minimum of +86% for Aber-OWL to a maximum of +6,000% for EcoPortal22.

We then investigated the frequency of appearance of each identifier resource in multiple registries (Fig. 1b). We found that only 7 resources appeared in more than 13 of the 23 external registries (including well-known resources such as GO, ChEBI, and NCBITaxon), and no resource appeared in more than 17, further illustrating the fragmented state of existing registries and the benefits of an integrative registry in having improved coverage. Further, the Bioregistry contains 192 novel prefixes not available in any other registry (Fig. 1b, green bar) that were curated for a diverse set of reasons. For example, we systematically reviewed cross-references in OBO Foundry ontologies, found a number of prefixes referring to resources for which no external registry contained an

| Registry          | Citation | Metadata Model | Name | Homepage | Description | Example LUID | LUID Pattern | Provider | All Providers | All Prefixes | License | Version | Contact | Capabilities and Qualities | Structured Data | Bulk Data | No Auth. | Required | Permissive | License | Prefix Search | Prefix Provider | Resolve CURIEs | Lookup CURIES |
|-------------------|----------|----------------|------|----------|-------------|--------------|--------------|-----------|-------------|-------------|---------|---------|---------|-----------------------------|----------------|-----------|---------|----------|------------|---------|-------------|--------------|--------------|--------------|
| Bioregistry       | 40       | Y              | Y    | Y        | Y           | o            | o            | o         | o         | o           | Y       | Y       | Y       | Y                          | Y              | Y         | Y       | Y        | Y          | Y       | Y           | Y            | Y            | Y            |
| AberOWL           | 12       | Y              | Y    | Y        | o           | o            | o            | Y         | Y         | Y           | Y       | Y       | Y       | Y                          | Y              | Y         | Y       | Y        | Y          | Y       | Y           | Y            | Y            | Y            |
| AgroPortal        | 13       | Y              | Y    | o        | Y           | o            | o            | Y         | Y         | Y           | Y       | Y       | Y       | Y                          | Y              | Y         | Y       | Y        | Y          | Y       | Y           | Y            | Y            | Y            |
| BioContext        | 14       |                 |      | o        |             |              |              |           |           |             |         |         |         | Y                          | Y              | Y         | Y       | Y        | Y          | Y       | Y           | Y            | Y            | Y            |
| Biolink           | 15       | Y              | Y    | o        | Y           |              |              |           |           |             |         |         |         | Y                          | Y              | Y         | Y       | Y        | Y          | Y       | Y           | Y            | Y            | Y            |
| BioPortal         | 16       | Y              | Y    | o        | Y           |              |              |           |           |             |         |         |         | Y                          | Y              | Y         | Y       | Y        | Y          | Y       | Y           | Y            | Y            | Y            |
| Cellosaurus       | 17       | Y              |      | Y        | Y           |              |              |           |           |             |         |         |         | Y                          | Y              | Y         | Y       | Y        | Y          | Y       | Y           | Y            | Y            | Y            |
| CHEMINF           | 18       | Y              |      | Y        | Y           | Y            |              |           |           |             |         |         |         | Y                          | Y              | Y         | Y       | Y        | Y          | Y       | Y           | Y            | Y            | Y            |
| CropOCT           | 19       | Y              | Y    | o        | Y           |              |              |           |           |             |         |         |         | Y                          | Y              | Y         | Y       | Y        | Y          | Y       | Y           | Y            | Y            | Y            |
| EcoPortal         | 20       | Y              | o    | Y        | o           | o            | o            | o         | Y         | Y           | Y       | Y       | Y       | Y                          | Y              | Y         | Y       | Y        | Y          | Y       | Y           | Y            | Y            | Y            |
| EDAM              | 21       | Y              |      | Y        | Y           |              |              |           |           |             |         |         |         | Y                          | Y              | Y         | Y       | Y        | Y          | Y       | Y           | Y            | Y            | Y            |
| FAIRSharing       | 22       | Y              | Y    | Y        |              |              |              |           |           |             |         |         |         | Y                          | Y              | Y         | Y       | Y        | Y          | Y       | Y           | Y            | Y            | Y            |
| GO                | 11       | Y              | o    | o        | o           | o            | o            |           |           |             |         |         |         | Y                          | Y              | Y         | Y       | Y        | Y          | Y       | Y           | Y            | Y            | Y            |
| HL7               | 23       | Y              | o    | o        |             |              |              |           |           |             |         |         |         | Y                          | Y              | Y         | Y       | Y        | Y          | Y       | Y           | Y            | Y            | Y            |
| Identifiers.org   | 24       | Y              | Y    | Y        | Y           | Y            |              |           |           |             |         |         |         | Y                          | Y              | Y         | Y       | Y        | Y          | Y       | Y           | Y            | Y            | Y            |
| N2T               | 25       | Y              | Y    | Y        | Y           | Y            |              |           |           |             |         |         |         | Y                          | Y              | Y         | Y       | Y        | Y          | Y       | Y           | Y            | Y            | Y            |
| NCBI              | 26       | Y              | Y    |          | Y           |              |              |           |           |             |         |         |         | Y                          | Y              | Y         | Y       | Y        | Y          | Y       | Y           | Y            | Y            | Y            |
| OBO Foundry       | 27       | Y              | Y    | Y        |              |              |              |           |           |             |         |         |         | Y                          | Y              | Y         | Y       | Y        | Y          | Y       | Y           | Y            | Y            | Y            |
| OLS               | 28       | Y              | o    | Y        | o           | o            | o            | o         | Y         | Y           | Y       | Y       | Y       | Y                          | Y              | Y         | Y       | Y        | Y          | Y       | Y           | Y            | Y            | Y            |
| OntoBee           | 29       | Y              | Y    | Y        | o           | o            |             |           |           |             |         |         |         | Y                          | Y              | Y         | Y       | Y        | Y          | Y       | Y           | Y            | Y            | Y            |
| Prefix Commons    | 30       | Y              | o    | o        | o           | o            | o            | o         | o         | o           |         |         |         | Y                          | Y              | Y         | Y       | Y        | Y          | Y       | Y           | Y            | Y            | Y            |
| re3data           | 31       | Y              | Y    | o        |              |              |              |           |           |             |         |         |         | Y                          | Y              | Y         | Y       | Y        | Y          | Y       | Y           | Y            | Y            | Y            |
| UniProt           | 32       | Y              | o    | o        | o           | o            | o            | o         | o         | o           |         |         |         | Y                          | Y              | Y         | Y       | Y        | Y          | Y       | Y           | Y            | Y            | Y            |

Table 2. An overview of the data models, capabilities, and qualities of the 23 external registries covering biomedical ontologies, controlled vocabularies, and databases imported by the Bioregistry. A “Y” means the field is required. A “o” means it is part of the schema, but not required or incomplete on some entries. A blank cell means that it is not part of the metadata schema. Two glossaries describing the elements in the data model (left) and capabilities and qualities (right) appearing as column headers in this table can be found respectively in Supplementary Sections 3.1 and 8.1.
entry, and added them to the Bioregistry. Several novel prefixes were suggested by external contributors who were themselves the maintainers of the corresponding resource. Another subset of these entries were added by members of the community who encountered them and were then motivated to create an entry for it in the Bioregistry.

Next, we examined NCBITaxon, one of the resources that appears in the largest number of existing registries. This identifier resource appears in 17 external registries under 10 different prefixes including taxon, taxon-onomy, and NCBITaxon (Fig. 1c, not all shown). In addition, the Bioregistry curates 9 prefix synonyms (e.g., NCBI_Taxon_id, uniprot.taxonomy, NCBI_taxid) that appear in various non-registry biomedical resources, demonstrating the high heterogeneity of usages for a given identifier resource. Such cross-registry mappings and synonyms in the Bioregistry enable it to act as an interoperability layer to standardize across a large number of external registries and non-registry resources.

**Web portal for interactive and programmatic use.** The contents of the Bioregistry can be browsed interactively through the web portal at https://bioregistry.io shown in Fig. 2. The portal implements a powerful search feature to help users look up prefixes and CURIEs they encounter in various databases, ontologies, and
other biomedical resources. The search feature extends to not only the prefix, synonyms, title, and description of each record, but also all of the corresponding fields in linked records from external registries (Fig. 2a). The full prefix list can be browsed (Fig. 2b) and each prefix page organizes and contextualizes all information available from novel curation in the Bioregistry as well as imported from external registries (Fig. 2c,d). Notably, it links to external registry pages when mappings are available. For example, the page for NCBITaxon (https://bioregistry.io/ncbitaxon) links to a large number of external registry pages (Fig. 2e). In addition to the data in the Bioregistry, this page constructs example URIs for first-party providers, third-party providers (e.g., OntoBee29, OLS), and external resolvers (e.g., Identifiers.org, N2T) using a combination of information stored about external registries and programmatic logic in the underlying Bioregistry Python package (Fig. 2f). The web portal provides several other features including generating pages for each of the external registries integrated into the Bioregistry that show their various properties and functionalities, facilitating curating and displaying user-generated collections of prefixes such as the list of Semantic Web prefixes at https://bioregistry.io/collection/0000002, and listing contributors. The portal also implements a resolver that allows for the uniform construction of URIs from CURIEs that are automatically redirected to the appropriate location based on the URI format string annotated to the CURIE's prefix. The Bioregistry's resolver uses the URI scheme https://bioregistry.io/<prefix>:<local-unique-identifier>, similar to the resolver schemes used by Identifiers.org and N2T. Bioregistry also makes available a programming language-agnostic RESTful interface that gives access to all functionality (e.g., search, autocompletion, record retrieval, URI generation) and is documented with OpenAPI/Swagger at https://bioregistry.io/apidocs. The underlying data used to generate each page can be downloaded in a
variety of formats including JSON, YAML, TSV, RDF, and others (where applicable). Finally, the portal serves as a hub for links to the code, data, documentation, and narrative surrounding the Bioregistry.

As a companion to the main Bioregistry site, a static site is automatically generated and deployed to https://bioregistry.github.io/bioregistry using GitHub’s infrastructure. Notably, this includes a health report that runs a weekly check for which of the Bioregistry’s resources’ homepages are still accessible (i.e., do not return HTTP 404 or other connection errors) and which resources’ URI format strings are still valid. The site both provides a high-level summary of which resources have recently become invalid as well as a detailed, color-coded table reflecting the statuses of all Bioregistry records. Ultimately, this site can help more systematically monitor and improve the maintenance of biomedical resources.

Exported artifacts for data integration and reusability. The Bioregistry GitHub repository contains the root content of the database (i.e., the registry, metaregistry, and collections) as JSON files in a version-controlled setting, serving as the single source of truth. In addition, it makes available several derived artifacts that are meant to facilitate integration with downstream systems and resources. These exported artifacts are regenerated daily, and made available via the web portal at http://bioregistry.io/download and are archived on Zenodo.

In addition to the native JSON format, the Bioregistry data is made available as a set of YAML and TSV files to facilitate reuse. Further, equivalence mappings between resources in external registries are exported into the Simple Standard for Sharing Ontological Mappings (SSSOM) format. SSSOM is a standard for sharing mappings between different namespaces that we use to represent mappings between resources appearing in different registries, such as the relations exemplified in Fig. 1c.

The Bioregistry also provides a number of artifacts to facilitate integration with Semantic Web contexts and linked open data. First, we constructed an RDF schema for the Bioregistry that reuses elements from common Semantic Web vocabularies (e.g., DC, FOAF) and creates its own elements in the bioregistry.schema vocabulary described at https://bioregistry.io/schema. All components of the Bioregistry (i.e., the registry, metaregistry, collections) were jointly exported into RDF under this schema in several commonly used formats including N-Triples, Turtle and JSON-LD. This allows the Bioregistry to be loaded using triple stores (e.g., Virtuoso) or programming libraries (e.g. Python’s RDFLib) and subsequently queried with SPARQL. We also assembled a network derived from the RDF export that can be browsed interactively on the Network Data Exchange (NDEX) available at https://bioregistry.io/ndex:860647c4-f7c1-11ec-ac45-0a135e8baf.

Finally, the Bioregistry makes available several Semantic Web contexts that each map a set of prefixes (e.g., chebi) to a corresponding URI prefix (e.g., http://purl.obolibrary.org/obo/CHEBI_). These are derived from the root data using a set of policies for choosing a preferred prefix and URI format for resources in the registry. In addition to a general purpose context encompassing all of Bioregistry, we make available application-specific semantic contexts for integration with the OBO Foundry, and a context limited to prefixes useful for general Semantic Web resources.

Maintenance model and governance. In contrast to the maintenance and governance structures employed by existing registries, the Bioregistry takes an alternative approach relying on open data, open code, open infrastructure, automated testing, and automated updating. Similar models have been adopted with great success in existing large collaborative projects such as the OBO Foundry. We accomplished this through several steps. First, the bioregistry data is stored and versioned using GitHub (see Methods). Second, anyone can propose additions or changes either directly by submitting a pull request to the Bioregistry repository or by filling out an appropriate issue template that triggers an automated generation of a pull request. Both create an open forum for discussion that invites a wide variety of stakeholders to engage (implementation details in Methods). Third, using GitHub allows for the technical implementation of quality control and quality assurance workflows that are coupled to pull requests in order to ensure that all changes meet a predefined set of standards, which are described explicitly and publicly in the contribution guidelines (https://github.com/biopragmatics/bioregistry/blob/main/docs/CONTRIBUTING.md) as well as implicitly in the implementation of the quality assurance workflow. In addition to the open data, open code, open infrastructure philosophy, the Bioregistry project has sought out community guidance on how to establish a governance model that is more robust to the fluctuation of funding and personnel who are actively working on and moderating the project. This has resulted in the establishment of a Review Team and a Development Team as well as a public minimal governance model (https://github.com/biopragmatics/bioregistry/blob/main/docs/GOVERNANCE.md) that describes how to induct new members, how to remove members, who respectively has the technical authority and community responsibility to facilitate and ultimately judge changes to the underlying database and make changes to the code base, and how to improve the governance model over time. These teams have been initially seeded with members from diverse scientific backgrounds, locations, and institutions to further promote the durability of the project. These guidelines also include a liberal policy on authorship to further demonstrate the project’s commitment to inclusivity.

Use cases and integrations
Here, we highlight several projects and standards that have already adopted various functionalities of the Bioregistry.

Supporting interoperable data annotation. Several projects use the Bioregistry to create prefix maps, or mappings between prefixes (e.g., uniprot) and their corresponding URI prefixes (e.g., https://www.uniprot.org/uniprot/). These support the the interoperability of data annotations and the conversion between URIs and CURIEs in Semantic Web applications. The Simple Standard for Sharing Ontological Mappings (SSSOM) is a metadata standard for various mappings (e.g., equivalences) between ontology and database terms and an associated toolset (https://github.com/mapping-commons/sssom-py) based on LinkML (https://linkml.io) for
loading, validating, and converting SSSOM content. The standard is meant to encourage higher quality curation in biomedical ontologies which often lack important metadata such as the mapping type, a standardized prefix and local unique identifier for the subject and object terms, provenance about how the mapping was generated, and provenance about who generated the mapping. The default prefix map used in validation is generated by the Bioregistry following the procedure described in Supplementary Section 6.1.

Manubot36 is a tool for open collaborative writing that aims to bring automation, customizability, and transparency to scholarly publishing. With Manubot, users write manuscripts using markdown with special support for citation by persistent identifiers represented as CURIEs such as ![@doi:10.1371/journal.pcbi.1007128] (which can then be automatically turned into a full citation). Embedding CURIEs in manuscripts is especially valuable when referring to resources that are not citable manuscripts such as clinical trials (e.g., in a review of COVID-1946). Manubot initially added support for 700 CURIE prefixes by incorporating Identifiers.org but later switched to the Bioregistry which at the time added support for an additional 365 prefixes. Besides being more comprehensive, the Bioregistry's open contribution model allowed for addressing several longstanding issues with Identifiers.org including invalid regular expression patterns, missing prefixes, as well as inconsistencies due to some namespaces being redundantly embedded in identifiers.

We discuss the plans and considerations for adopting the Bioregistry for interoperable data annotation in further software and resources including the Biolink Model19 and the Alliance of Genome Resources46 in Supplementary Section 7.

Validation and quality control of entity references. Several projects use the Bioregistry to standardize or validate prefixes and local unique identifiers and promote interoperability and reusability. Biomappings31 (https://github.com/biopragmatics/biomappings) is a repository for curated and predicted mappings between equivalent (or otherwise related) biomedical entities in different identifier resources. It contains several workflows for generating predicted mappings using Gilda48 and provides a web-based curation interface for reviewing predicted mappings and adding novel ones. Biomappings ensures data integrity by validating all prefixes and local unique identifiers in the repository using the Bioregistry. Further, the curation interface uses the Bioregistry to generate links to a web page describing each biomedical entity, making curation easier. Biomappings also generates a network-based summary of its content that uses the Bioregistry to provide links to identifier resources and to resolve CURIEs.

The Integrated Network and Dynamical Reasoning Assembler (INDRA)19 assembles biomedical knowledge from multiple databases combined with text mining of scientific publications to construct executable models. When performing assembly, INDRA maintains references to biomedical entities that are grounded to one or more identifier resources. It uses the Bioregistry to first check that the prefixes used in these groundings are standardized, and then to validate the associated unique local identifier according to the pattern provided by the Bioregistry. Further, the curation interface uses the Bioregistry to generate links to a web page describing each biomedical entity, making curation easier. Biomappings also generates a web-based summary of its content that uses the Bioregistry to provide links to identifier resources and to resolve CURIEs.

The Phenotype Knowledge Translator (PheKnowLator)35 ecosystem constructs FAIR biomedical knowledge graphs using ontologies and reasoning with the addition of non-ontological data sources. PheKnowLator uses the Bioregistry to standardize references in CURIEs and URIs from both data types to provide semantically consistent results for downstream use cases. The Bioregistry helps overcome significant challenges posed by ontologies (e.g., changing namespaces over time, data that is not from an ontology that does not provide valid namespaces or URIs), and their integration. The Bioregistry API has become a vital component of the build process and is used to standardize URIs for all entities and triples. It has also provided new opportunities to extend PheKnowLator’s testing harness. Overall, the inclusion of Bioregistry has improved the PheKnowLator ecosystem and the knowledge graphs it produces. Similarly, the Graph Representation eArning, Predictions and Evaluation (GRAPE) software package (https://github.com/AnacletoLAB/grape) uses the Bioregistry to normalize the identifiers in several networks and knowledge graphs (including PheKnowLator).

Contextualizing entities with website links. Several projects use the Bioregistry to generate and resolve URIs within their APIs or user-facing websites in order to provide additional context to the entities they refer to. BridgeDb44 is a web service that maps between local unique identifiers from different identifier resources representing equivalent entities (e.g., P0DP23 in UniProt and Q17855525 in Wikidata for the Calmodulin-1 protein). The Bioregistry has been integrated in BridgeDb’s Java and R clients as well as Bacting49 to enable lookup based on standardized CURIE, to enable creating internal BridgeDb identifier objects via standardized CURIEs, and to generate URIs resolvable through the Bioregistry web application.

The Bioregistry has also been used in several websites to generate URLs for human genes, protein complexes, and other entities. For example, the DUB Portal19 is a website summarizing experimental analyses of deubiquitinating enzymes and uses the Bioregistry to link to human genes in HGNC and protein families in the FamPlex vocabulary51. The website for FamPlex (https://sorgerlab.github.io/famplex) also uses the Bioregistry to standardize and link references for human genes; equivalence mappings to InterPro52, Medical Subject Headings (MeSH)39, GO, Complex Portal44, NextProt35 for protein families and complexes; and references for publications in PubMed and PubMed Central. Similarly, the interactive user interface for the BERN26 named entity recognition platform standardizes its biomedical entities and generates links using the Bioregistry.

Unified access to external registries. Because of its integration of external registries, the Bioregistry is also useful for unified access to their respective data. The OBO Foundry27 facilitates the coordinated development of biomedical ontologies through a set of guiding principles and community organization. Its associated repository (https://github.com/OBOFoundry/OBOFoundry.github.io) stores the structured metadata about each ontology, including their preferred prefix, title, homepage, description, and usages. The Bioregistry is used to support
the standardization and maintenance of this metadata in several ways described in detail in Supplementary Section 6.2.

Discussion
We presented the Bioregistry, an integrative registry of biomedical identifier resources. The Bioregistry takes a novel approach to curation by importing and harmonizing data from external registries that can be further improved and extended with novel curation. It relies on an open data, open code, open infrastructure philosophy combined with a novel governance strategy to foster community contributions and engagement, and ensure its longevity and adoption. It uses public infrastructure for quality assurance, distribution, and deployment to promote transparency, reduce cost, and uncouple its long-term maintenance from a specific institution, funding source, or group of maintainers. While the Bioregistry demonstrates higher coverage and metadata granularity than other registries, it also explicitly encourages reuse and redistribution via its highly permissive CC0 license.

Limitations. Entries in the Bioregistry that represent identifier resources, their preferred CURIE prefix, and other metadata are integrated semi-automatically from external registries (such as Identifiers.org and N2T) or manually curated directly in the Bioregistry. The design choice that the Bioregistry semi-automatically imports and aligns content from external registries is important for maintaining broad coverage, and to distribute curation effort across multiple projects. However, this still poses challenges for consistency. Namely, the Bioregistry has limited ability to enforce guidelines and conventions in other registries. For instance, there are differing views in the community on stylistic choices in the capitalization of preferred CURIE prefixes (e.g., chebi vs CHEBI or ChEBI) for identifier resources. Drawing on other registries can also lead to future conflicts where multiple registries choose the same CURIE prefix for two different identifier resources, creating a situation that has to be retroactively arbitrated in the Bioregistry (an example is given in Supplementary Section 5.2). Nevertheless, the Bioregistry maintains guidelines (https://github.com/biopragmatics/bioregistry/blob/main/docs/CONTRIBUTING.md#submitting-new-prefixes) for creating new identifier resource prefixes, which, if followed, can mitigate these issues. Further, the purview of the Bioregistry does not extend to directly advising and mentoring creators of new identifier resources to make good choices in their identifier schemes. Creators of such resources can rely on recommendations such as those suggested by McMurry et al.6.

Adopting the Bioregistry’s standard for prefixes, CURIEs, and URIs in a new resource is straightforward. However, applying it retroactively to an existing resource can pose challenges. It may require updating the data and associated code in the resource itself as well as in downstream consumers of the resource. This can manifest in several ways, including updating non-standard synonyms (e.g. many ontologies use MSHE as a non-standard prefix for MeSH), updating non-standard construction of CURIEs (e.g., using redundant prefixes as prescribed by Identifiers.org like in GO:GO:0006915), or updating non-standard URIs (e.g., switching all ORCID URIs to use the https protocol). If such changes are not feasible in the resource, it is still possible to implement mappings to the Bioregistry or create custom exports following the Bioregistry standard, potentially broadening the resource’s interoperability.

The Bioregistry provides a solution for standardizing references to individual entries in identifiers resources. However, it is often the case that multiple identifiers resources contain entries representing equivalent entities (e.g., multiple disease ontologies representing the same disease) leading to redundancy when integrating disparate resources, such as when constructing knowledge graphs. Determining which identifier resource to prioritize when representing an entity that appears in multiple resources is beyond the scope of the Bioregistry. Nevertheless, the Bioregistry can contribute to the standardization of the cross-references between equivalent entities in different identifiers resources (cross-references, in practice, often use non-standard CURIEs and URIs) thereby helping redundancy resolution among them.

While the Bioregistry is limited to resources of interest to researchers in the life sciences, its methodology and technological implementation could extend to other scientific areas. Ultimately, the Bioregistry could serve as a template for the creation of domain-specific metaregistries in other areas or be the basis for the creation of a metaregistry spanning multiple scientific domains.

Future work on the bioregistry. Following the initial development, deployment, and early adoption of the Bioregistry, two ongoing challenges remain. The first is to be responsive in the maintenance, enrichment, and extension of the content in the registry to best reflect the reality of the ever-changing landscape of biomedical identifier resources. While this has not been realized by previous registries, the Bioregistry’s combination of technical infrastructure and governance model will enable this effort in a sustainable way. Further, we plan to develop tools and curation practices to proactively, systematically identify new resources to add to the Bioregistry.

The second is to build and maintain a community of users. This entails continuing to engage multiple groups of users and stakeholders. This includes curators, and consumers of biomedical resources, as well as groups designing automated data- or knowledge-extraction and aggregation systems. Serving the needs of these communities requires identifying their challenges, and improving the Bioregistry’s data model, tooling, and content accordingly. It also entails facilitating discussion between a diverse set of individuals and offering training for usage of the Bioregistry and its philosophy. To this end, the authors plan to organize a set of recurring community workshops (following an initial workshop held in 2021 (https://biopragmatics.github.io/workshops/WPCI2021) around the topics of identifier resources and registries.

Future vision. We envision the Bioregistry could more broadly be used to promote and support the standardized annotation of data, models, ontologies, and scientific literature. First, the growing body of data being made available through publications and data repositories often lack standardized annotations to their records (e.g., columns in a table, nodes/edges in knowledge graphs). If adopted by the diverse set of curators, developers,
maintainers, and users of life science tools and resources, the Bioregistry could provide a consistent way of annotating these data to make them more FAIR, especially facilitating reuse.

Second, we envision the Bioregistry supporting the standardization of structured metadata associated with models and networks derived from data such as mechanistic models (e.g., in the BioModels database\textsuperscript{57}), network-based models (e.g., in NDEx), knowledge graphs (e.g., those described by Bonner et al.\textsuperscript{38}), and machine learning models (e.g., such as those trained by GRAPE (https://github.com/AnacletoLAB/grape)) in order to promote their interoperability and reuse. For example, despite the recent proliferation of biomedical knowledge graphs\textsuperscript{58}, there has been little convergence on standardized syntax or semantics for identifying nodes and edges. The Bioregistry can support this effort both on a low level and also by integrating into higher-level tools that generate and exchange graph-like data such as KGX and Biolink (see Supplementary Section 7.2) that support a larger variety of use cases. More generally, we believe that this will enable the growing audience of machine learning researchers who are interested in combining biomedical datasets in novel ways.

This vision aligns well with the recommendations from a recent assessment of the reproducibility of such models\textsuperscript{59} that highlighted the more general importance of annotation using high-quality controlled vocabularies like GO and ChEBI.

Third, though biomedical ontologies have proven invaluable for data annotation, key ontologies still suffer from a lack of standardization of cross-references\textsuperscript{2}, making it difficult to merge and reason across ontologies and other structured data sources. Given that ontologies are often curated in public version-controlled repositories in standardized formats (e.g., OBO, OWL), the Bioregistry could be used to support their semi-automated standardization and maintenance in order to both reduce curation burden and potentiate their value in data integration scenarios.

Finally, we envision the potential adoption of the Bioregistry by academic publishers to support the standardized annotation of named entities in the text provided by authors (e.g., the BioFactoid\textsuperscript{60}), and thereby decrease the need for doing expensive and error-prone post-processing like automated named entity recognition (NER) on publications to create structured representations.

**Methods**

The Bioregistry repository tightly couples the data to a Python package that facilitates loading, accessing, and modifying the root data files. It provides several high-level data structures and workflows for accessing and reasoning over the Bioregistry and external registries’ integrated data, that support the quality assurance workflows (described above), the web application (described above), the alignment workflows (described below), the generation of derived artifacts (described above) and other user-facing functionality such as prefix standardization, CURIE standardization, and URI parsing. Full documentation for the Python software package can be found at https://bioregistry.readthedocs.io.

**Alignment.** While manual curation of mappings to external registries is feasible when adding novel prefixes to the Bioregistry, the frequency of updates to external registries motivated the development and application of periodic automated and semi-automated alignment. We first stratify all external registries into three categories based on their available metadata, biomedical scope, focus on assigning global prefixes to resources, and governance. The first group with metadata availability, a biomedical scope, and focus on assigning global prefixes contains the Identifiers.org, the OBO Foundry, OLS, N2T. The second group contains registries such as GO, NCBI, UniProt, Cellosaurus, and FAIRSharing that contain entries that do not correspond to identifier spaces which are excluded from the import. It additionally included registries like BioContext and BioPortal because of insufficient metadata that often made it impossible to determine what identifier resource the metadata refers to. The third group contains registries with minimal metadata or lack of biomedical focus such as Prefix.cc. The alignment algorithm first generates a lookup table based on the canonical prefix, preferred prefix, and all prefix synonyms (see Supplementary Section 3 for details on the data model and Supplementary Fig. 2 for a schematic diagram of this process) for each resource in the Bioregistry. The prefix policies and automated quality assurance checks in the Bioregistry ensure that there are no collisions in this lookup table. For each external registry, the data are downloaded, normalized, and exactly one field is annotated as the external prefix (see Supplementary Fig. 1). All Bioregistry prefixes that already have been mapped to an external prefix in the external registry are removed from the lookup table. Similarly, all external prefixes that already have mappings are not considered for new mappings. Each external prefix that matches an entry in the lookup table is assigned an automated mapping. A manually curated list of incorrect mappings and collisions are used to post-process the automated mappings and remove false negative mappings (see Supplementary Section 5.2). External prefixes that could not be mapped to a Bioregistry prefix are handled based on their stratification. For the first group of registries, the prefix is added as a new record to the Bioregistry. For the second group of registries, the prefix is added to a curation sheet along with its relevant metadata (e.g., title, homepage, example identifier) for later manual curation. For the third group of registries with minimal metadata, no report is made.

**Promoting sustainability and longevity through automation.** The Bioregistry is hosted on GitHub (https://github.com/biopragmatics/bioregistry) to take advantage of its public, cloud-based version control, collaboration, and workflow management platforms. The single source of truth data (i.e., root data) for the Bioregistry is stored in version control. This implicitly versions all minor changes with git commit hashes and allows git tags to be used to mediate releases, which are automatically archived and re-distributed on both GitHub and Zenodo (https://zenodo.org).

The Bioregistry uses GitHub Actions as a continuous integration service to run code and data quality assurance to promote the maintainability and integrity of the resource (see Supplementary Section 4.1). They further enable workflows for automatically generating pull requests and notifying reviewers to enable non-technical
users to make submissions to the resource. The Bioregistry further uses GitHub Actions as a continuous delivery and continuous deployment system to run the aforementioned alignment workflows, generate derived artifacts, release code to the Python Package Index (PyPI), containerize code on Docker Hub (https://hub.docker.com), deploy the web application to Amazon Web Services (https://aws.amazon.com) on a daily basis (see Supplementary Section 4.2). Combined, the continuous integration, delivery, and deployment services allow contributors and consumers of the Bioregistry to more easily propose improvements, review them as a community, and see them reflected in the data and website without the need for manual intervention by the project team. Using an entirely free, public, and open public infrastructure to do so promotes longevity and sustainability by mitigating the monetary requirements. Further, the technical requirements of the deployment of the web service and hosting are also minimized such that hosting costs around $33/year and compute costs around $27/year (see Supplementary Section 4.3).

Data availability

All data analyzed during this study are available on GitHub at https://github.com/biopragmatics/bioregistry. The data specific to the version of Bioregistry used in this article (v0.5.132) is archived on Zenodo40.

Code availability

The source code for the Bioregistry is available at https://github.com/biopragmatics/bioregistry under the MIT License. The source code specific to the version of Bioregistry used in this article (v0.5.132) is archived on Zenodo40.

Received: 16 July 2022; Accepted: 26 October 2022;
Published online: 19 November 2022

References

1. Mark, D. et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3, 160018, https://doi.org/10.1038/sdata.2016.18 (2016).
2. Jacobsen, A. et al. FAIR Principles: Interpretations and Implementation Considerations. Data Intelligence 2, 10–29, https://doi.org/10.1162/dint_r_00024 (2020).
3. Samantha, L. et al. Sharing biological data: why when and how. FEBS Letters 595, 847–863, https://doi.org/10.1002/1873-3468.14067 (2021).
4. Hastings, J. et al. ChEBI in 2016: Improved services and an expanding collection of metabolites. Nucleic Acids Research 44 D1214–D1219 https://doi.org/10.1093/nar/gkx1031 (2016).
5. Bateman, A. et al. UniProt: the universal protein knowledgebase in 2021. Nucleic Acids Res 49, D480–D489, https://doi.org/10.1093/nar/gkaa1100 (2021).
6. McLurk, J. A. et al. Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. PLoS Biol 15, 1–18, https://doi.org/10.1371/journal.pbi.2001144 (2017).
7. Kristian, A. et al. A review of the new HGNC gene family resource. Human Genomics 10, 6, https://doi.org/10.1186/s40246-016-0062-6 (2016).
8. Bansal, P. et al. Rhea, the reaction knowledgebase in 2022. Nucleic Acids Res 50, D693–D700, https://doi.org/10.1093/nar/gkb1016 (2022).
9. Barrett, A. J. Enzyme Nomenclature. Recommendations 1992. Eur. J. Biochem. 232, 1, https://doi.org/10.1111/j.1432-1033.1995.tb20774.x (1995).
10. Fleischmann, A. et al. IntEnz, the integrated relational enzyme database. Nucleic Acids Res 32, D434–D437, https://doi.org/10.1093/nar/gkh1119 (2004).
11. The Gene Ontology Consortium. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. 47, D330–D338, https://doi.org/10.1093/nar/gkt1055 (2019).
12. Hoehndorf, R., Slater, L., Schofield, P. N. & Gkoutos, G. V. Aber-OWL: a framework for ontology-based data access in biology. BMC Bioinformatics 16, 26, https://doi.org/10.1186/s12859-015-0456-9 (2015).
13. Jonquet, C. et al. AgroPortal: A vocabulary and ontology repository for agronomy. Comput. Electron. Agric. 144, 126–143, https://doi.org/10.1016/j.compag.2017.01.012 (2018).
14. Mungall, C. et al. prefixcommons/biocontext: v0.1.0. Zenodo https://doi.org/10.5281/zenodo.6800248 (2022).
15. Unni, D. R. et al. Biolink Model: A universal schema for knowledge graphs in clinical, biomedical, and translational science. Clin. Transl. Sci. https://doi.org/10.1111/ctrs.13102 (2022).
16. Whetzel, P. L. et al. BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications. Nucleic Acids Res 39, W541–5, https://doi.org/10.1093/nar/gkaa469 (2011).
17. Bairach, A. The CelloSaurus, a Cell-Line Knowledge Resource. J. Biomol. Tech. 29, 25–38, https://doi.org/10.7171/jbt.18-2902-002 (2018).
18. Hastings, J. et al. The chemical information ontology: Provenance and disambiguation for chemical data on the biological semantic web. PLoS One 6, 1–13, https://doi.org/10.1371/journal.pone.0025513 (2011).
19. Arnaud, E. et al. The Ontologies Community of Practice: A CGAIR Initiative for Big Data in Agrifood Systems. Patterns 1, 100105, https://doi.org/10.1016/j.partern.2020.100105 (2020).
20. Kechagioglou, X., Vaira, L., Tommassino, P., Fiore, N. & Basset, A. EcoPortal: An Environment for FAIR Semantic Resources in the Ecological Domain. In JOWO (Bolzano, Italy, 2021).
21. Ison, J. et al. EDAM: an ontology of bioinformatics operations, types of data and identifiers, topics and formats. Bioinformatics 29, 1325–1332, https://doi.org/10.1093/bioinformatics/btq113 (2013).
22. Sansone, S.-A. et al. FArSharing as a community approach to standards, repositories and policies. Nat. Biotechnol. 37, 358–367, https://doi.org/10.1038/s41587-019-0080-8 (2019).
23. Bender, D. & Sartipi, K. H.L. FHHR: An Agile and RESTful approach to healthcare information exchange. In Proceedings of the 26th IEEE International Symposium on Computer-Based Medical Systems, 326–331, https://doi.org/10.1109/CBMS.2013.6627810 (2013).
24. Juty, N., Le Novère, N. & Latibeaudière, C. Identifiers.org and MIRIAM Registry: Community resources to provide persistent identification. Nucleic Acids Res 40, 580–586, https://doi.org/10.1093/nar/gkx1097 (2012).
25. Wimalaratne, S. M. et al. Uniform resolution of compact identifiers for biomedical data. Sci. Data 5, 1–8, https://doi.org/10.1038/sdata.2018.29 (2018).
26. Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J. & Sayers, E. W. GenBank. Nucleic Acids Res 44, D67–72, https://doi.org/10.1093/nar/gkv1276 (2016).
ACKNOWLEDGEMENTS

CTH, KK, BMG were funded under the Defense Advanced Research Projects Agency (DARPA) Young Faculty Award [W911NF-20-1-0255]. EW received funding by NWO grant [203.001.121]. TL received funding from the São Paulo Research Foundation (FAPESP) grant 2019/26284-1. CJM, MAH, and JM were supported by a NIH Office of the Director Grant #5R24OD011883. The authors would like to acknowledge the developers, maintainers, and curators of each of the external registries referenced throughout this manuscript and used in the Bioregistry, whose prior work enabled ours. More detailed “live” acknowledgments for these registries can be found at https://bioregistry.io/acknowledgements.

AUTHOR CONTRIBUTIONS

C.T.H. and B.M.G. conceived and implemented the software and resource, analyzed the results, and wrote the manuscript. C.J.M., J.K., J.M. and M.A.H. contributed to the conception and design of the resource and software. A.R., B.M.G., C.T.H., C.J.M., D.D.F., D.S.H., D.W., M.B., N.M., S.M., TL and T.J.C. performed data curation for the resource, B.M.G., C.T.H., D.R.U., E.W., H.B.H., K.K., N.M. and S.M. contributed to the software. C.J.M.,
D.S.H., D.R.U., E.W., J.M., N.M., S.M., T.J.C. co-wrote the manuscript. All authors reviewed and edited the manuscript.

**Competing interests**
DDF received salary from Enveda Biosciences.

**Additional information**
**Supplementary information** The online version contains supplementary material available at https://doi.org/10.1038/s41597-022-01807-3.

**Correspondence** and requests for materials should be addressed to B.M.G.

**Reprints and permissions information** is available at www.nature.com/reprints.

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

[Open Access] This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022