diversity

MDPI

Article

Life Science—Microbial Culture Collections Data Integration Tasks

Alexander Vasilenko *, Galina Kochkina, Natalya Ivanushkina and Svetlana Ozerskaya

All-Russian Collection of Microorganisms (VKM), Pushchino Scientific Center for Biological Research of the Russian Academy of Sciences (PSCBR RAS), G.K. Skryabin Institute of Biochemistry and Physiology of Microorganisms Russian Academy of Sciences, 142290 Pushchino, Russia
* Correspondence: vanvkm@gmail.com; Tel.: +749-978-32402

Abstract: This paper presents interconnections between catalogs of microbial culture collections and biological databases inspected. Microbial Biological Resources Centers (mBRCs) provide Life Science (LS) and biotechnology with fit-for-use microbiological resources and related data of consistent quality. To optimize the services, facilitate cumulative research, make crosschecks, and avoid duplication of efforts, must ensure that the databases developed and maintained are interconnected with mBRC data. This research shows that, at present, connections are minimal. It proposes ways to plug the mBRC databases into the Life Science community. Such connections could open dialogue by making the mBRC data visible and accessible from the Life Science databases, and reciprocally making the Life Science database records visible and accessible from the mBRC-aggregated catalog. For this purpose, we inspected most of the databases discovered on the Internet. Each database was characterized by name, acronym, year of the last correction, uniform resource location (URL), area of practical use (health system, agriculture, etc.), presence of microbial data and database producer. The databases with microbial data were inspected in more detail in terms of the lists of the partner databases, the lists of ontologies used, the access format from computer programs, and database subjects. Our new metabase has collected 2667 Life Science databases, from which 1123 databases have microbial data.

Keywords: databases; metabase; data integration; fungal biodiversity; culture collections; bioeconomy; microorganisms; FAIR

Citation: Vasilenko, A.; Kochkina, G.; Ivanushkina, N.; Ozerskaya, S. Life Science—Microbial Culture Collections Data Integration Tasks. *Diversity* **2023**, *15*, 17. https://doi.org/10.3390/d15010017

Academic Editor: Ipek Kurtboke

Received: 31 October 2022
Revised: 29 November 2022
Accepted: 7 December 2022
Published: 22 December 2022

1. Introduction

Microbial culture collections, or rather the microbial Biological Resources Centres, help to ensure that microbiology is on solid ground by fulfilling three major roles [1]:

- mBRCs are long-term preservation facilities, and they provide well-characterized microorganisms as well as related data. By doing so, the mBRCs constitute the foundations of cumulative research because they ensure that experiments involving microorganisms can be repeated or exploited to generate new knowledge by any person that is skilled in the field;
- mBRCs are a professional source of taxonomy information, and they contribute to resolve nomenclatural issues related to microorganisms;
- mBRCs keep and study specific strains with unique characteristics, and these well-documented strains constitute libraries of strain-specific characters that are functionally useful in biotechnology.

Life Science Data repositories can be grouped in three categories:

1. Databases;
2. Publications;
3. Datasets.

The best structured data holdings in this list are the databases. This research studied the integration opportunities between them, more particularly the data integration of the mBRCs' microbial databases with the Life Science databases.

As a result of the disconnection between the mBRC and Life Science databases, this is not so easy to achieve:

- on Life Science side—to find the organism presented in a database several years ago so as to verify its properties and to do more advanced research,
- on mBRCs side—to select the most needful organisms to be preserved, the storage technology for them, to do appropriate microbial research and services for specific organisms kept. Forbes (https://www.forbes.com/sites/johncumbers/2022/09/12/white-house-inks-strategy-to-grow-trillion-dollar-us-bioeconomy/?sh=61e9bfa735e1 (accessed on 30 October 2022)) estimated that the USA bioeconomy was worth approximately USD 1 trillion in 2022 and that it will be worth over USD 30 trillion globally in the next two decades (message published on 12 September 2022, accessed on 10 December 2022). There are arguments that microbial systems comprise a big fraction of the bioeconomy, and the positive effect of the database integration task presented here could be estimated to be worth USD 1 billion per year, at least.

2. Materials and Methods

2.1. Life Science Databases That Were Inspected

The main sources were the following:

- FAIRsharing (1933 databases (https://fairsharing.org (accessed on 30 October 2022)), which were analyzed till October 2022);
- MB (1802 entries (http://metadatabase.org (accessed on 30 October 2022)), the last analysis was complited on 26 December 2015);
- Biosharing (724 databases, (https://www.biosharing.org/ (accessed on 30 October 2022)), 26 December 2015);
- BioMedBriges (814 databases, (http://wwwdev.ebi.ac.uk/fgpt/toolsui/ (accessed on 30 October 2022)), 27 December 2015);
- Pathguide (363 database names, 2013) (http://www.pathguide.org/ (accessed on 30 October 2022))
- ELIXIR list (579 entries, (https://bio.tools/?q=database (accessed on 30 October 2022)), 28 January 2016);
- ExPASy (85 + 665 databases, (http://www.expasy.org/old_links (accessed on 30 October 2022)), 12 February 2016;
- Bioinformatics Links Directory (621 databases);
- OBRC (http://www.hsls.pitt.edu/obrc/ (accessed on 30 October 2022)) (30 March 2017).

The additional sources include: the Nucleic Acids Research (NAR) journal (https://www.ncbi.nlm.nih.gov/pmc/journals/4/ (accessed on 30 October 2022)), the references of the databases in the main sources lists, the databases analyzed in the materials of the EOSC-Life project, and its Open Call subprojects.

2.2. Research Procedure

We collected 16 parameters for a database.

Eight parameters make the general description of any inspected Life Science database:

- A unique identifier. For example: BIODBCORE-000438.
- A database acronym that is used by the database producer. For example: dbSNP.
- A database name that is used by the database producer. For example: The Database of short genetic variation (single nucleotide polymorphism).
- A database URL. For example: http://www.ncbi.nlm.nih.gov/SNP/ (accessed on 30 October 2022)
- The access level. "Open" if the database records are available to read for free for anybody, but the "Restricted" otherwise.

- The year of the last correction. It could be the last year presented in records of the database or in the news messages or in copyright. It could be also the current year if there is a clear message that this database is still being curated. For example: 2022.
- The developer/Owner. For example: USA, NCBI; USA, National Library of Medicine, National Institutes of Health.
- A comment. For example: *Escherichia coli*.

Eight additional parameters are used for databases with microbial data only:

- A practical domain. Here the used values refer to: patent, health (mostly human), pharmacology, agriculture, bioremediation/biodegradation, veterinary, food preparation, winemaking, baking, brewing, biofuel, and other kinds of biotechnology. These help in navigating the collected databases.
- The microbial level. Here the used values refer to: MO, SP, and ST. The value ST is used if at least one database message provides information on a specific microorganism strain. The value SP is used if there are no strains, but there is information on at least one microorganism species discovered. The value MO is used if there are no species, but there is some kind of information on the microorganisms discovered.
- The properties. This shows the types of data discovered in this database. The used values include: cell, chemistry, disease, DNA, drugs, enzyme, gene, genome, image, immunology, interactome, lipid, metabolite, microbiome, molecules, pathogen, pathways, peptide, proteomics, publications, RNA, taxonomy, and toxicology. On average, the number of properties and keywords assigned to a database is between six and seven. For example: DNA, gene, genome, proteomics, publications, and RNA.
- The orientation. If the database is focused on some kind of microorganisms. For example: fungi.
- Search by. Data types used in indexing between database partners. It makes a tools in integration technology. With these tools two big communities of Life Science databases were discovered, each interconnected inside like, such as in a LOD Cloud but with no obligation to be open.
- Ontologies list. For example: SO.
- Partner databases. Shows integration of this specific database into the other databases. The external database obtains partner status for this specific database for the following:
  - (1) It is mentioned on its WEB pages in data interchange;
  - (2) If this specific database use some data from that external database as a data source;
  - (3) There are messages with fields values from this external database;
  - (4) There is a message that this is the database partner;
  - (5) When these databases have common datacuration. Example: Assembly, BioProject, BioSample, ClinVar, dbGaP, dbMHC, dbSTS, dbVar, Ensembl, GenBank, Homologene, IGSR, MapViewer, NCBI Gene, Nucleotide, OMIM, PMC, Protein, PubChem Substance, PubMed, RDP, RefSeq, UniGene, UniProtKB.
- Program interface. Mostly according to descriptions in documents of this specific database. Example: ELIXIR WEB UI, Entrez Programming Utilities (E-Utils). It makes the tools in integration technology.

In this research the parameters "Orientation", "Properties," and "Practical domain" help in the navigation of the collected databases community. The parameters: "Partner databases" and "Ontologies list" show the integration of this specific database into the other databases. The parameters: "Search by" and "Program interface" make tools in integration technology.

3. Results

3.1. Microbial Databases—Microbial Culture Collection (CC) Interconnection

The total number of Life Science database names or references identified in this study was more than 14,800. The total number of database references inspected manually exceeded 5500.

We collected 2667 visible online Life Science databases, from which 1123 had microbial data (Figure 1). These 1123 databases are contingently represented by an ellipse.

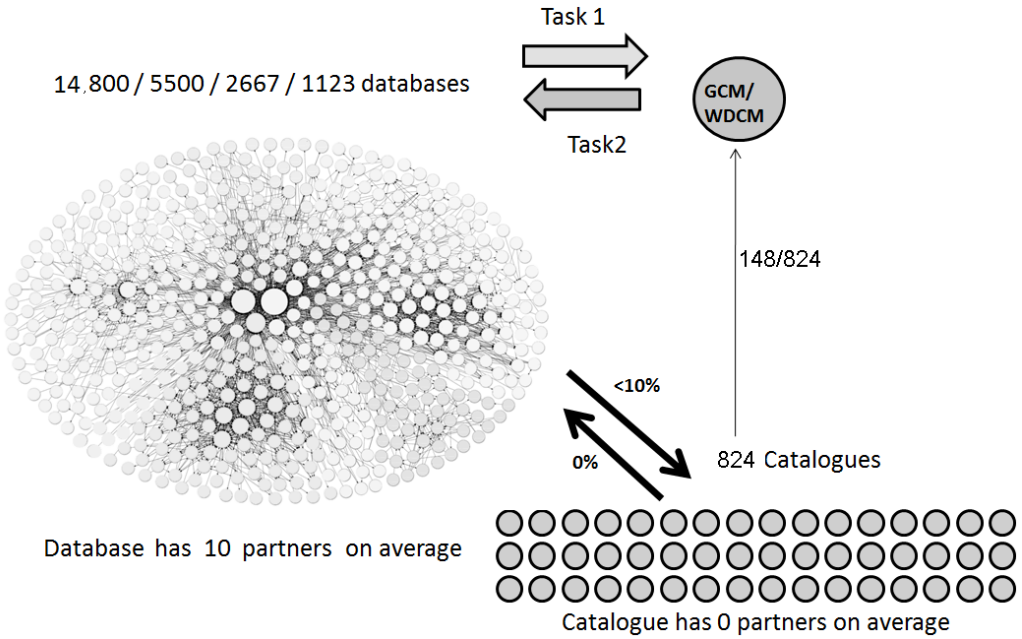

Figure 1. Microbial databases—CC interconnection.

As of December 2022, the World Data Centre for Microorganisms (WDCM) collected information related to 824 microbial culture collections (CC or mBRC) from all over the world, and published online the Culture Collections Information Worldwide (CCINFO) information system [2]. These collections are presented by the circles on bottom right side of the figure.

From these 824 microbial culture collections, 148 collections shared their catalog data and made them accessible via the Global Catalog of Microorganisms data portal GCM/WDCM (http://gcm.wfcc.info/ (accessed on 30 October 2022)) [2].

On average, information in each catalog had no connections to other catalogs or to other databases.

In [3], it was presented that approximately only 10% of the information on microorganisms in Life Science databases referred to strains in microbial culture collections. It was also discovered that:

- There were mostly no links to microbial catalogs,
- The strains interconnection service called Histri in the former Straininfo system was currently not available.

For these reasons we indicated the "<10%" connection from Life Science to microbial strains catalogued in Culture Collections.

In other words, more than 90% of the information related to properties of microorganisms had no reliable link to the place of long-term preservation and the supply of the microorganisms: the microbial culture collections. This meant that there is no assurance that verification, cross-checking, and further studies on the research outcomes made by the use of microorganisms could be performed, since these particular genetic resources were lost. This decreased the efficiency of research in Life Science and Biotechnology, and jeopardized cumulative research.

Such situations require action to obtain the microorganisms "from paper to bench", that is, making the genetic resources presented in the databases and referenced in scientific papers effectively available for further cumulative research and unlimited in time. We structured the research into complementary parts:

Part 1 focused on making mBRC data visible and accessible via the Life Science databases. Its solution appeared technically possible, as we will show further in this paper and may provide for financial advantage.

Similarly, Part 2 found the knowledge collected in Life Science databases and related it to the specific microorganism. It appeared more complicated, technically, but could provide scientific benefits.

The natural framework for these tasks was the integration contracts between the database producers and mBRCs; we will thus talk about "Life Science databases that are partners of the mBRCs system".

Practically, the adjective "accessible" data refers to two formats: accessible to human access and accessible by computer programs. This led to the specification following of parts 1 and 2:

- 1a: to make mBRC data visible and accessible via partners of Life Science databases for human access.
- 1b: to make mBRC data visible and accessible via partners of Life Science databases for computer programs.
- 2a: to make the records for specific microorganisms stored by partners of the Life Science database visible and accessible for human access.
- 2b: to make the records for specific microorganisms stored by partners of the Life Science database visible and accessible by computer.

Potentially, when every culture collection makes a separate contract with each database, this makes 148 × 1123 = 16,6204 contracts which are neither practical nor cost-efficient. Our technical proposal is described in the scheme of Figure 2.

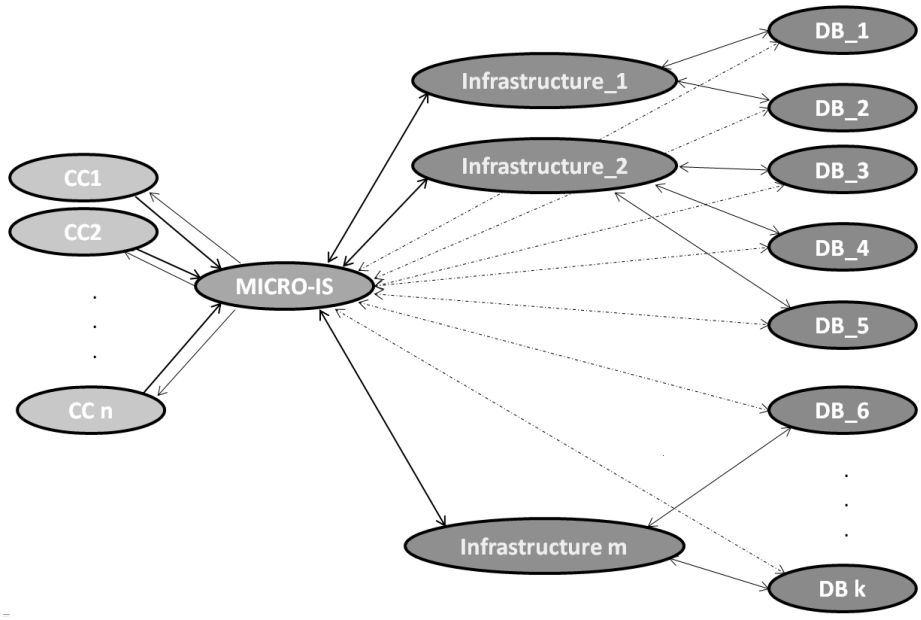

Figure 2. General integration schema.

We proposed to compile copies of CC catalogs into one database similarly to the most recent development, WDCM/GCM, or older systems such as former StrainInfo [4] or CABRI [5]. In this research, we call it MICRO-IS. Integration contracts are to be signed between infrastructures that produce or control databases to be integrated and MICRO-IS.

Based on the analyses of Life Science databases, on the next pages, we introduced two parameters of the database integration levels, and showed that four integration contracts with some specific database producers could be sufficient to shift MIRCO-IS from zero integration level into the group of leaders.

The annotated list of the Life Science databases made a new metabase with all 2667 Life Science databases discovered, and eight fields of search, or 1123 databases with microbial

data and 16 fields of search. The first proposal for this metabase name was MONALISa Databases (Micro-Organisms related fraction of All Life Sciences Databases).

We can conduct a search in this metabase, or collect statistics. Table 1 shows an example of how many databases have specific types of data.

Table 1. Keywords yielded by a search on "Properties" in metabase.

Specific Type of Data	How Many Databases Have It
Gene	826
Proteomics	716
Publications	625
Image	517
RNA	389
DNA	395
Genome	355
Enzyme	361
Cell	316
Chemistry	297
Pathways	270
Disease	263
Interactome	226
Taxonomy	219
Drugs	195
Peptide	195
Molecular data	183
Immunology	167
Metabolite	166
Toxicology	156
Pathogen	151
Lipids	132
Microbiome	45

The "Gene" value or, in other words, the genetic data was on top of the list of the most frequent value for "Properties". In other words, 826 databases have "genes" in the field "Properties".

Information on proteins came in second position (716 databases). The leader in this group was the UniprotKB database. More than one half (625) of the databases had references to publications. The most popular were the links to PubMed (PubMed ID). Taxonomic data are mentioned in 219 databases. From these 219 databases, the most popular data provider was NCBI Taxonomy.

Microbiome data were the least cited values. The list of 45 databases with microbiom data is: Biology Reference, BioSamples, BioSystems, Bookshelf, EMBL, EMBL-EBI, ENA, Espacenet, Europe PMC, ForestScience Current Database, GO Database, GONUTS, GoPubMed, HGTree, HOMD, HPMCD, IMG, IMG/M, IMG/VR, MEDLINE, MeSH, Microbiome, NARCIS, NCBI, NFSD, NLM Catalog, Nowomics, OMIM, OMIM (1), OReFiL, PANGAEA, PLOS One, PMC, PSP, Pubget, PubMed, PubMed Health, QIAGEN, RefSeq, ScienceDirect, SRA, TACONIC, UniProtKB, VetMed Resource, and WikiGenes.

Another structure in the Life Science databases communities is the database groups of the largest database producers.

The absolute leader was BESC, with BioCyc database community (http://www.biocyc.org/biocyc-pgdb-list.shtml)—20,025 databases (as of 25 August 2022). There were three database groups in this BioCyc:

- Tier 1 databases: EcoCyc, MetaCyc, HumanCyc, AraCyc, YeastCyc.
- Tier 2 had 63 databases generated by the PathoLogic program, with subsequent curation conducted manually.
- Tier 3 had 19,936 databases generated by the PathoLogic program with no manual review of the pathway predictions, nor subsequent curation.

The second database producer was EMBL-EBI, we discovered 97 databases in its list including ArrayExpress, ASD, ASTD, ATD, BioModels, BioSamples, Cellular Phenotype Db, ChEBI, ChEMBL, CluSTr, CSA, DGVa, DNAtraffic, DrugPort, e!Ensembl S. cerevisiae, e!EnsemblBacteria, e!EnsemblCat, e!EnsemblChicken, e!EnsemblChimpanzee, e!EnsemblCow, e!EnsemblDog, e!EnsemblFugu, e!EnsemblFungi, e!EnsemblGenomes, e!EnsemblGorilla, e!EnsemblHorse, e!EnsemblMetazoa, e!EnsemblMouse, e!EnsemblPig, e!EnsemblPlants, e!EnsemblProtists, e!EnsemblRabbit, e!EnsemblZebrafish, EGA, EMBL, EMBL-EBI, EMDB, ENA, Ensembl, Enzyme Portal, Enzyme Structures, EVA, Expression Atlas, FunTree, GeneDB, GWAS Catalog, HGNC, HipSci, IGSR, IMEx, IMGT/HLA, IntAct, IntEnz, InterPro, IPD, IPD-ESTDAB, IPD-HPA, IPD-KIR, IPD-MHC, logRECOORD, MACiE, MEROPS, MetaboLights, Metal MACiE, MicroCosm, MIRIAM collection, MTBLS, NRNL1, NRNL2, NRPL1, NRPL2, OLDERADO, PANDIT, PDBe, PDBe EM Resources, PDBeChem, PDBsum, Pfam, Pfam, PhenoDigm, PICR, PRIDE, PROCOGNATE, Reactome, RECOORD, Rfam, RNAcentral, SAS, SRS@EMBL-EBI, SureChEMBL, TreeFam, UniChem, UniProt-GOA, UniSave, VASCO, and VectorBase.

The third largest was NCBI with databases including Assembly, BioProject, BioSample, BioSystems, Bookshelf, CCDS, CDD, ClinGen, ClinVar, Clone DB, COGs, dbEST, dbGaP, dbMHC, dbProbe, dbSNP, dbSTS, dbVar, Dengue virus database, ECRbase, Genbank, Gene, Genetic Codes, Genome, GEO, GEO DataSets, GEO Profiles, GSS, GTR, Histone, HIV-1, Homologene, IBIS, Influenza Virus Resource, MapViewer, MedGen, MEDLINE, MeSH, MMDB, NCBI, NCBI taxonomy, NCBI Trace Archives, NLM Catalog, Nucleotide, OMIM, Organelle genomes, PMC, PopSet, Probe, Protein, Protein Clusters, PubChem, PubChem BioAssay, PubChem Compound, PubChem Substance, PubMed, PubMed Health, RefSeq, RefSeqGene, Retroviruses, SKY/M-FISH and CGH, SPARCLE, SpliceInfo, SRA, Structure, TPA, UniGene, UniVec, Viral genomes, and Virus Variation.

The NCBI list of databases was smaller, but its microbial fraction was larger than that in EMBL-EBI. One more structure in the Life Science databases made interconnection communities (Figure 3).

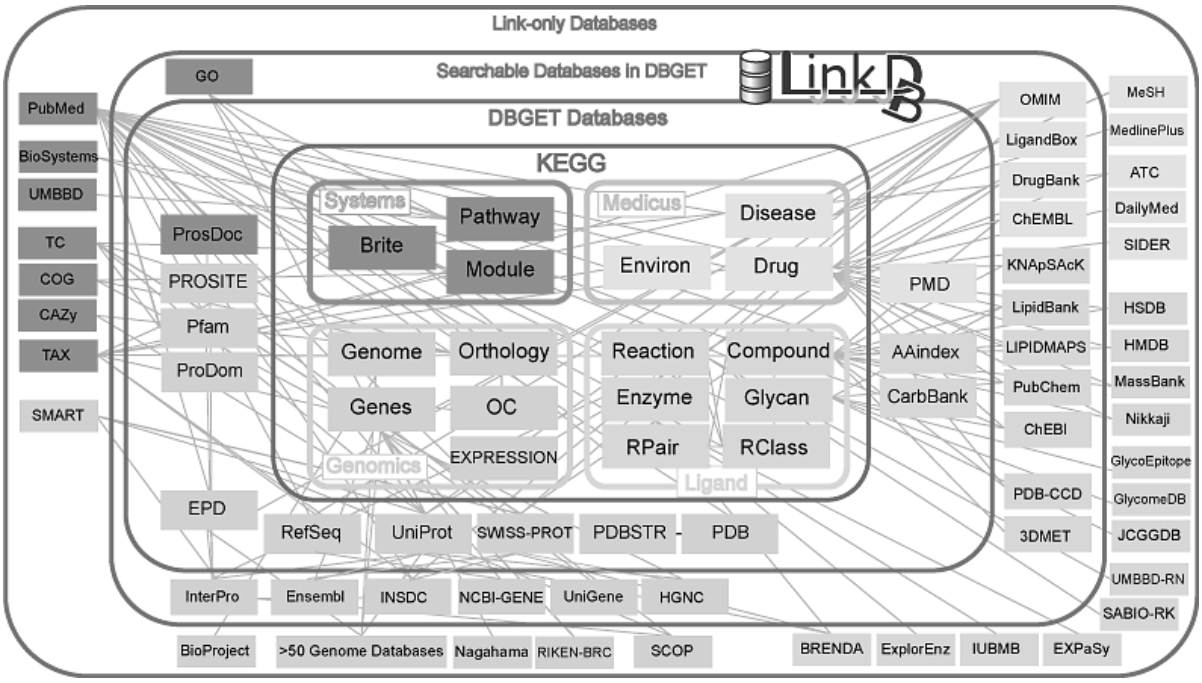

Figure 3. Example of database interconnection. DBGET materials (Japan).

Using our new metabase we introduced two integration parameters:

The connection number—CN—the number of database partners found in each microbial Life Science database,
The attraction number—AN—The number of microbial databases indicating this database to be a partner.

To calculate these parameters, we drew the "Interconnection matrix" (Figure 4), where the columns show databases with microbial data, and the lines list all the Life Science databases collected.

	3D Lectin	3D RIBOSC	5S RNA Da	A.pernix		UniPROBE	UniProt-G	UniProtKB	UniRef		YMPL	YPM	Zif-BASE		AN
2P2Idb	0	0	0	0		0	0	0	0		0	0	0		1
3D Genome Browser	0	0	0	0		0	0	0	0		0	0	0		1
3D Lectin	0	0	0	0		0	0	0	0		0	0	0		0
3D RIBOSOMAL MODIFICATIO	0	0	0	0		0	0	0	0		0	0	0		3
3DBIONOTES	0	0	0	0		0	0	0	0		0	0	0		0
3Dee	0	0	0	0		0	0	0	0		0	0	0		0
3DEM Benchmark	0	0	0	0		0	0	0	0		0	0	0		0
3D-footprint	0	0	0	0		0	0	0	0		0	0	0		1
3DID	0	0	0	0		0	0	0	0		0	0	0		6
PubChem	0	0	0	0		0	0	0	0		0	0	0		57
PubChem BioAssay	0	0	0	0		0	0	0	0		0	0	0		16
PubChem Compound	0	0	0	0		0	0	0	0		0	0	0		31
PubChem Substance	0	0	0	0		0	0	0	0		0	0	0		25
pubget	0	0	0	0		0	0	0	0		0	0	0		2
Public Data Portal	0	0	0	0		0	0	0	0		0	0	0		1
PubMed	0	0	0	1		0	0	1	0		0	0	1		565
PubMed Health	0	0	0	0		0	0	0	0		0	0	0		6
YPM	0	0	0	0		0	0	0	0		0	0	0		1
ZFIN	0	0	0	0		0	1	1	0		0	0	0		20
Zif-BASE	0	0	0	0		0	0	0	0		0	0	0		0
ZOBODAT Vespoidea	0	0	0	0		0	0	0	0		0	0	0		1
ZorapteraSF	0	0	0	0		0	0	0	0		0	0	0		1
ИИПС	0	0	0	0		0	0	0	0		0	0	0		0
CN	5	8	4	10		6	148	161	15		2	3	5		

Figure 4. Partial picture of the Interconnection matrix.

The cell in line I, colon J, has a value of one if the microbial database J has a database I in the list of its partners; otherwise, it has value zero.

The CN is the last bottom line of the matrix, AN is the right column. Top line—acronyms of microbial databases; left column—acronyms of all Life Science databases.

One of the most integrated databases—UniProtKB (produced by SIB) has the following parameters: CN = 161 and AN = 501; PubMed (NCBI, NIH) has: CN = 7 and AN = 565. The integration level in the microbial culture collections is mostly zero.

3.2. Microbial Databases with the Largest Number of Partners

From the structured programming technology [6], we know that it is very difficult to control more than seven processes. Most databases have regular corrections and updates, and as a result, a database partner is, in fact, a process.

Thus, it is difficult to have more than seven database partners. However, in our research, 482 databases indicated eight or more database partners. Table 2 presents 35 databases with 38 partner databases or more. The most popular databases are presented in Table 3.

To be stable, these database communities must have internal synchronization tools and potentially indicate good integration schema to us. The first names in the list could be the best.

We began this study with UniProtKB database integration partners with the calculated attraction number (AN).

They had the largest value of AN. Most of them were microbial for some reason. It was highly desirable to have them on our integration list. For this reason, we summed up the database attraction numbers for the large database producers (Table 4).

Table 2. Microbial databases with more than 37 partners (connection number—CN).

Database Name	Number of Its Partners
Pathguide	364
UniProtKB	161
iProClass	159
COL	153
UniProt-GOA	148
OReFiL	143
EcoliWiki	129
GeneCards	128
PIR	91
UCD 2D-PAGE	73
Hits	63
SWISS 2DPage	61
PubChem	57
PubChem BioAssay	57
PubChem Compound	57
PubChem Substance	57
dbProbe	55
NCBI	55
PiroplasmsDB	53
E!EnsemblGenomes	52
EMBL	49
ENA	49
SBKB	49
EcoGene	46
SGD	46
Gene	45
InterMitoBase	44
EMBL-EBI	43
OMIM	43
MetaCyc	42
Guide to Pharmacology	40
MalaCards	40
NCBI Taxonomy	38
OpenHelix	38
ViralZone	38

Table 3. The most popular databases.

Database Name	Number of Databases That Refer to It
PubMed	565
UniProtKB	501
NCBI Taxonomy	275
RCSB PDB	255
Genbank	239
Gene	229
KEGG	199
RefSeq	193
EC	187
Pfam	182
InterPro	160
Protein	157
Ensembl	152

Table 3. *Cont.*

Database Name	Number of Databases That Refer to It
Nucleotide	125
OMIM	109
SGD	99
ENA	80
HGNC	80
PROSITE	76
CAS	71
IntAct	71
Reactome	70
ChEBI	69
FlyBase	63
MEDLINE	61
UniGene	61
BioGRID	59
PubChem	57
GEO	56
NCBI	56
SMART	56
DrugBank	55
MGI	55
PIR	55
PMC	55
SCOP	55
COGs	54
Genome	53
DIP	51
STRING	51
ENZYME	50
KEGG Pathway	49
HPRD	48
WormBase	48
BioProject	47

Table 4. Integration candidates.

Priority No	Producer	Databases	AN
1	NCBI	70	2909
2	EMBL-EBI	97	1209
3	SIB	37	762
4	Kyoto University	19	348
5	Instute Paster	18	148
6	BioCyc	9378	133
7	InterMine	16	20
	1 + 2 + 3 + 4	133	5228
	1 + 3 + 4	92	4019
	2 + 3 + 4	78	2319
	Total	1116	8870

3.3. Connection Strategy

The sum of AN values in 1123 microbial Life Science databases totalized 8870. The sum of AN values in the NCBI microbial databases was 2909; which is, 33% of the total. That was the best result from all of the producers.

If MICRO-IS had four integration contracts with (1) NCBI, (2) EMBL-EBI, (3) SIB, (4), and Kyoto University, then according to the line "1 + 2 + 3 + 4", the number of microbial

databases integrated would be 133, and the attraction sum for them would be 5228 (59% of the maximum possible). This could make a good integration solution with four contract partners as referenced in Figure 2.

The list of the 133 partner databases connected via these contracts comprised the following:

- ArrayExpress, Assembly;
- BioModels, BioProject, BioSample, BioSamples, BioSystems, Bookshelf;
- CDD, Cellosaurus, ChEBI, ChEMBL, COGs, CSA;
- dbEST, dbProbe, dbSNP, Dengue virus database, DNAtraffic, DrugPort;
- e!Ensembl, e!Ensembl Saccharomyces cerevisiae, e!EnsemblBacteria, e!EnsemblFungi, e!EnsemblGenomes, e!EnsemblProtists, EMBL, EMBL-EBI, EMDB, ENA, Ensembl, ENZYME, Enzyme Structures, EPD, EVA, Expression Atlas;
- Genbank, Gene, GeneDB, Genetic Codes, Genome, GEO, GEO DataSets, GEO Profiles, GSS;
- HAMAP, Hits, HIV-1, Homologene;
- IMEx, Influenza Virus Resource, IntAct, InterPro;
- KEGG, KEGG BRITE, KEGG DISEASE, KEGG GENES, KEGG GENOME, KEGG GLYCAN, KEGG LIGAND, KEGG MEDICUS, KEGG MODULE, KEGG Organisms, KEGG ORTHOLOGY, KEGG PATHWAY;
- MACiE, MapViewer, MedGen, MEDLINE, MEROPS, MeSH, MetaboLights, MIA-PEGelDB, MMDB, MTBLS;
- NCBI, NCBI taxonomy, NCBI Trace Archives, neXtProt, NLM Catalog, Nucleotide;
- OMA, OMIM, OpenFlu, Organelle genomes;
- PathComp, PathPred, PathSearch, PaxDB, PDBe, PDBe EM Resources, PDBsum, Pfam, PICR, PMC, PMP, PomBase, PopSet, PRIDE, Probe, PROSITE, Protein, Protein Clusters, Protein Spotlight, Proteomes, PubChem, PubChem BioAssay, PubChem Compound, PubChem Substance, PubMed, PubMed Health;
- Reactome, RefSeq, RefSeqGene, Retroviruses, Rfam, Rhea, RNAcentral;
- SPARCLE, SpliceInfo, SRA, Structure, SugarBind, SWISS-2DPAGE, SWISS-MODEL, SwissVar;
- UniGene, UniProt-GOA, UniProtKB, UniRef;
- Viral genomes, ViralZone, Virus Variation.

3.4. Possible Tasks 1 and 2 Solutions

One possible path for Task 1a is to ask the partner infrastructures to replace the text string with the microorganism's name that answers the user's request by the link to the MICRO-IS program with microorganism's name in the parameter.

If this program obtained the strain number, it found the specific strain in the MICRO-IS catalog and sent back the strain's parameters presented by the mBRC that keeps it.

If the program received the species name only, it found the strains with this species name (arrows in the left-center) and sent back passports for each of them. However, this makes a potential solution only: there is no evidence that the system will provide the passport of that specific microorganism that was searched for in the database.

The problems were, firstly, that there was no certainty that the integration contract partner infrastructures would accept this schema; secondly, some databases provided the microbial name with a link, but not to the strain passport in the microbial culture collection. For this reason, we need to study more sophisticated solutions. The research is not finished yet.

It also appeared feasible to provide the names/synonyms processing system, like in MycoBank (http://www.mycobank.org/BioloMICS.aspx?Table=Mycobank&Rec=18759&Fields=All (accessed on 30 October 2022)), Look at an example of its page for *Penicillium cyaneofulvum*, but for all the microorganism types.

Penicillium cyaneofulvum	
Summary:	*Penicillium cyaneofulvum* Biourge, La Cellule 33: 174 (1923)
Synonymy:	=*Penicillium brunneorubrum* Dierckx, Annales de la Société Scientifique de Bruxelles 25 (1): 88 (1901) =*Penicillium griseoroseum* Dierckx, Annales de la Société Scientifique de Bruxelles 25 (1): 86 (1901) =*Penicillium chrysogenum* Thom, U.S.D.A. Bureau of Animal Industry Bulletin 118: 58 (1910) =*Penicillium baculatum* Westling, Svensk Botanisk Tidskrift 4: 139 (1910) =*Penicillium notatum* Westling, Arkiv før Botanik 11 (1): 95 (1911) =*Penicillium chlorophaeum* Biourge, La Cellule 33: 271 (1923) =*Penicillium meleagrinum* Biourge, La Cellule 33: 147 (1923) =*Penicillium roseocitreum* Biourge, La Cellule 33: 184 (1923) =*Penicillium flavidomarginatum* Biourge, La Cellule 33: 150 (1923) =*Penicillium fluorescens* Laxa, Zentralblatt für Bakteriologie und Parasitenkunde Abteilung 2 86 (5–7): 160–165 (1932) =*Penicillium camerunense* R. Heim, Bull. Acad. R. Belg. Cl. Sci.: 42 (1949) =*Penicillium aromaticum* f. microsporum Romankova, Uchenn. Zap. Leningr. Univ. Zhadanov: 102 (1955) =*Penicillium harmonense* Baghd., Novosti Sistematiki Nizshikh Rastenii 5: 102 (1968)
Current name:	*Penicillium chrysogenum* Thom, U.S.D.A. Bureau of Animal Industry Bulletin 118: 58 (1910)
Classification:	*Fungi, Dikarya, Ascomycota, Pezizomycotina, Eurotiomycetes, Eurotiomycetidae, Eurotiales, Trichocomaceae, Penicillium*

In the MICRO-IS we proposed the same schema for bacteria and archaea based on the List of Prokaryotic names with Standing in Nomenclature (LPSN). For protists, microalgae and viruses, the same schema has not been tested yet.

The visible difficulty was in the taxonomies used in databases (Table 5). The right column is the number of references to popular taxonomical databases from all 1123 databases with microbial data, and the central column is from the fraction of 219 databases that present taxonomy. The Table 5 indicates that more than 50% of Life Science databases used NCBI Taxonomy, while microbial culture collections used the less popular IF, LPSN and MycoBank. The WDCM study showed not so large differences in these taxonomies; nevertheless, interconnection tables could be desirable.

Table 5. Taxonomy in Life Science databases.

Taxonomy	**References in the**	
	Databases with Taxonomical Data	**Databases with Microbial Data**
NCBI	**115**	**267**
GBIF	16	16
IF	7	9
COL	6	6
LPSN	2	2
MycoBank	2	5

The general solution for Task 1b appeared trivial. An example of this would be that the connection to the VKM catalog database with MySQL interface was initiated by mysql_real_connect function.

The connection tools for Tasks 2a and 2b are presented in the documents of the following:

- NCBI—Entrez Programming Utilities (E-Utils);
- EMBL-EBI—RESTful Web Services interface, Semantic WEB, RDF, SPARQL endpoint;
- SIB—RESTful Web Services interface, Semantic WEB;
- Kyoto University—KEGG API, LinkDB.

We inspected them and the desirable solutions seemed possible, as well as very complicated in the case of EMBL-EBI and SIB. The connection to EMBL-EBI was studied in EOSC-Life project materials and reported in EOSC-Life workshops.

4. Discussion

The main subject of the data integration tasks presented here lies in Open Access systems. For examples of the key materials published, see [7–10]. Detailed technical presentation of this action field could also be found in [9,11,12]. Open access promises very efficient data processing; however practically, it receives restrictions for commercial, safety, competition, and political reasons. A practical response to these restrictions initiated FAIR data processing method in 2015, and initially referenced in [13]. The most mature review of this action field is possibly presented in [14]. Practical usage of FAIR in the tasks of this paper was presented at the ECCO conferences; see [15,16].

The main tool constructed for the big data integration tasks was Semantic Web. The most official presentation is the W3C report (https://www.w3.org/standards/semanticweb/ (accessed on 30 October 2022)). This method was presented in a large list of publications; for examples, see [12,17]. In its practical usage for the tasks presented above, Semantic WEB must be implemented in most of the databases integrated. The most advanced integration solutions in the microbial domain were discovered inside NCBI Life Science databases with good connection to ATCC mBRC (kindly look at https://www.ncbi.nlm.nih.gov/ (accessed on 30 October 2022)) and in EMBL-EBI (https://www.ebi.ac.uk/services (accessed on 30 October 2022)), but the absolute majority of Life Science databases did not indicate the usage of these technologies. Therefore, we had to construct the data integration schema with extensive usage of the technologies mentioned above, and using publication examples: [18,19]. This integration schema was reported in MIRRI and EOSC-Life workshops. Currently, there are some privacy restrictions on its public presentation. However, we hope that later on, this will become the subject of additional and more technical article. In [3], there is also a reference to the Python script (kindly constructed by Mikhail Vasilenko) that discovered microorganism strains of the VKM microbial culture collection (http://www.vkm.ru/ (accessed on 30 October 2022)) in the Life Science databases of four databases producers mentioned here.

This research and this database integration effort were initiated inside the MIRRI project (2012–2015) and have not been finished even now. The main interconnection between records in the Life Science databases and strains in CC catalogs for human access and computers is not finished yet. The integration contracts with the main database producers are still not signed, but this effort was included in the EOSC-Life project and in the list of the MIRRI infrastructure actions.

5. Conclusions

According to our statistics, in this research we gathered the world's largest collection of Life Science databases, and the most structured examples of them. This collection provides a new metabase, which helps us in data integration tasks. The solutions for these tasks appear technically possible, and they promise financial and scientific success.

Author Contributions: Conceptualization, S.O. and A.V.; methodology, S.O. and A.V.; software, A.V.; validation, G.K. and N.I.; formal analysis, S.O.; investigation, A.V.; resources, S.O.; data curation, S.O., G.K. and N.I.; writing—original draft preparation, S.O., G.K. and N.I.; writing—review and editing, G.K. and N.I.; visualization, S.O.; supervision, S.O.; project administration, S.O. All authors have read and agreed to the published version of the manuscript.

Funding: This research has received funding from the Ministry of Science and Higher Education of the Russian Federation under grant agreement No. 075-15-2021-1051.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors thank Philippe Desmeth, WFCC president in 2011—2017, for his steadfast confidence in our final results and his correction of this article stile, Mikhail Vasilenko for his Python script and EOSC-Life team for the friendly atmosphere in the job.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

BESC	BioEnergy Science Center
CABRI	Common Access to Biological Resources and Information
CC	microbial Culture Collection
CCINFO	Culture Collections Information Worldwide
EMBL-EBI	The European Bioinformatics Institute
GCM	Global Catalog of Microorganisms
LPSN	List of Prokaryotic names with standing in nomenclature
LS	Life Science
Mbrc	Microbial Biological Resources Centre
NAR	Nucleic Acids Research journal
NCBI	National Center for Biotechnology Information
SIB	Swiss Institute of Bioinformatics
URL	Uniform Resource Locator
VKM	All-Russian Collection of Microorganisms
WDCM	World Data Center for Microorganisms

References

1. Biological Resource Centres Underpinning the future of Life Sciences and Biotechnology. OECD, Paris, France. 2001. Available online: https://read.oecd-ilibrary.org/science-and-technology/biological-resource-centres_9789264193550-en#page1 (accessed on 25 October 2022).
2. Wu, L.; Sun, Q.; Desmeth, P.; Sugawara, H.; McCluskey, K.; Smith, D.; Vasilenko, A.; Lima, N.; Ohkuma, M.; Robert, V.; et al. World data centre for microorganisms: An information infrastructure for the exploration and utilization of microbial strains preserved worldwide. *Nucleic Acids Res.* **2017**, *45*, D611–D618. [CrossRef] [PubMed]
3. Vasilenko, A.N.; Stupar, O.S.; Kochkina, G.A.; Ivanushkina, N.E.; Ozerskaya, S.M. Comparison of Microbial Diversity in Life Science Databases and in the Culture Collection. *Microbiology* **2023**, *92*, 94–95, in press.
4. Verslyppe, B. *StrainInfo: From Microbial Information to Microbiological Knowledge*; Ghent University, Faculty of Sciences: Ghent, Belgium, 2012; Available online: http://hdl.handle.net/1854/LU-4337118 (accessed on 25 October 2022).
5. Romano, P.; Kracht, M.; Manniello, M.A.; Stegehuis, G.; Fritze, D. The role of informatics in the coordinated management of biological resources collections. *Appl. Bioinform.* **2005**, *4*, 175–186. [CrossRef] [PubMed]
6. Dijkstra, E.W. Notes on Structured Programming. In *EUT Report. WSK, Dept. of Mathematics and Computing Science, Report 70-WSK-03*, 2nd ed.; Technische Hogeschool Eindhoven: Eindhoven, The Netherlands, 1970.
7. Sakkas, N.; Yfanti, S. Open data or open access? The case of building data. *Acad. Lett.* **2021**, 3629. [CrossRef]
8. What Is Open? Available online: https://okfn.org/opendata/ (accessed on 25 October 2022).
9. Auer, S.R.; Bizer, C.; Kobilarov, G.; Lehmann, J.; Cyganiak, R.; Ives, Z. DBpedia: A Nucleus for a Web of Open Data. *Semant. Web* **2007**, *4825*, 722–735. [CrossRef]
10. Kassen, M. A promising phenomenon of open data: A case study of the Chicago open data project. *Gov. Inf. Q.* **2013**, *30*, 508–513. [CrossRef]
11. Science Commons. Available online: https://creativecommons.org/about/program-areas/open-science (accessed on 25 October 2022).
12. Lehmann, J.; Isele, R.; Jakob, M.; Jentzsch, A.; Kontokostas, D.; Mendes, P.N.; Hellmann, S.; Morsey, M.; van Kleef, P.; Auer, S.; et al. DBpedia: A large-scale, multilingual knowledge base extracted from Wikipedia. *Semant. Web* **2015**, *6*, 167–195. [CrossRef]
13. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; Blomberg, N.; Boiten, J.W.; da Silva Santos, L.B.; Bourne, P.E.; et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **2016**, *3*, 160018. [CrossRef] [PubMed]
14. Collins, S.; Genova, F.; Harrower, N.; Hodson, S.; Jones, S.; Laaksonen, L.; Mietchen, D.; Petrauskaité, R.; Wittenburg, P. *Turning FAIR into Reality—Final Report and Action Plan from the European Commission Expert Group on FAIR Data*; Report; Publications Office of the European Union: Brussel, Belgium, 2018.
15. Vasilenko, A.; Stupar, O.; Kochkina, G.; Ozerskaya, S. Data Intergration with Life Science Databases: Gathering of Technology. In *Proceedings of the Conference Proceedings of XXXVII Annual Meeting of the European Culture Collections' Organisation (ECCO)*; Moscow, Russia, 13–15 September 2018; p. 125.

16. Vasilenko, A.; Robert, V.; Coronado, J.M.L.; Stupar, O.; Kochkina, G.; Ozerskaya, S.; Casaregola, S. FAIR options in mBRC specifics. In Proceedings of the Conference Proceedings of XXXVIII Annual Meeting of the European Culture Collections' Organisation (ECCO), Turin, Italy, 12–14 June 2019; pp. 57–58.
17. Berners-Lee, T.; Hendler, J.; Lassila, O. The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities. *Sci. Am* **2001**. Available online: http://web.dfc.unibo.it/buzzetti/IUcorso2006-07/materiali/bl-engl.pdf (accessed on 30 October 2022). [CrossRef]
18. Vasilenko, A.; Ozerskaya, S.; Stupar, O.; Romano, P.; Wu, L.; Evtushenko, L.; Smith, D.; Ma, J. Life Science Databases Interconnection Data Specifications. In Proceedings of the Abstract book of IUMS 2017, 15th International Congress of Bacteriology and Applied Microbiology, Singapore, 17–21 June 2017; p. 164.
19. Vasilenko, A.; Ozerskaya, S.; Stupar, O.; Romano, P.; Wu, L.; Evtushenko, L.; Smith, D.; Ma, J. Data Integration Opportunities for BRC Catalogs and Life Science Databases. In Proceedings of the Abstract book of 14th International conference on culture collections. ICCC14, Singapore, 17–22 July 2017; p. 12.