About FAIRness of microbial resource data

Paolo Romano

IRCCS Ospedale Policlinico San Martino, Genoa, Italy

Introduction

Curators of microbial Culture Collections (CCs) manage fundamental data about microorganisms and their biological features. This information is mainly made available through distinct databases, not always in agreement with existing common standards. Some efforts have been made to integrate the data and make it accessible in a coherent way. These include the Global Catalogue of Microorganisms (GCM) (1,2) and the Common Access to Biological Resources and Information (CABRI) Network Services (3,4).

Making available this information in the landscape of biomedical information would be important for many applications, in many fields. Integrated databases are just the first step towards an effective integration of this information because requires the actual possibility to exchange data and make it available for downstream data elaborations.

In this view, various methods and tools have been proposed. Semantic integration of information, through advanced methods that include the adoption of common vocabularies and ontologies, the setting up of unique and persistent identifiers of data, and common methods for programmatic access to data. In this context, the "FAIR Guiding Principles" for scientific data (5) propose the minimum requirements for an information system to be included in the landscape of integrated biomedical databases.

In this work, still in its infancy, we are examining the current features of CABRI in the light of the FAIR approach and try to define a possible path towards more FAIR systems for microbial resource information. The work is being extended to incorporate other microbial resource information systems and to see if common efforts can be carried out in order to improve the FAIRness of microbial resource data.

Methods

CABRI Network Services are one of the main output of the homonymous project funded by the EU from 1996 to 1999. They currently integrate information on 28 collections of microbial resources (bacteria, fungi, yeasts, phages, plasmids). The CABRI IT system is based on the well-known SRS software. The integration of catalogues in CABRI was implemented by defining common data sets and flat file formats for data sharing. Three distinct CABRI datasets were defined for each of the involved microbial resource respectively including information strictly requested for the identification of the resource (Minimum Data Set, MDS), information useful for an appropriate description of the main features of the resource (Recommended Data Set, RDS), and all remaining information (Full Data Set, FDS). Each information in the data sets is described in terms of data input and authentication rules within the CABRI Guidelines for Catalogue Production [6]. Rules include a description of the contents of each field and of its syntax, along with the input process. For many fields, an appropriate reference vocabulary or database is included. Some works are currently undergoing in the context of MIRRI (Microbial Resource Research Infrastructure) for the definition of a standard format for data exchange and of the Minimum Information about Biological Resources (MIaBRe). CABRI Network Services are accessible through the standard SRS interface and through a simplified form, the CABRI Simple Search. Programmatic access is also provided through the so called IST Bioinformatics Web Services (IBWS), that are currently under deep revision.

In the context of this analysis, we have concentrated our attention on FAIRness of CABRI Network Services, i.e. to the respect of its features to the FAIR principles. The FAIRness metrics developed by the FAIR Metrics Group was adopted [7].

Results

The following aspects were considered in the analysis: identifiers and their characteristics, availability of metadata and their characteristics, interfaces and their characteristics, overall adoption of domain and non-domain community standards, adoption of standard languages for knowledge representation. Preliminary results are reported in table 1.

Metric	Analysis
FM-F1A Identifier uniqueness	Each microbial resource has its collection specific identifier, which is usually composed by the acronym of the collection followed by the reference number or code for the resource in the collection. All acronyms are different and reference numbers are not reused, so the identifier is unique.
FM-F1B Identifier persistence	Catalogues are current: they only include information on resources that are available for distribution. Dismissed or lost or removed resources are not available anymore and the related information is lost.
FM-F2 Machine readability of metadata	Some work is undergoing for implementing metadata in the output in the form designed by schema.org and Bioschemas, but the majority of information is not described by them and this reduces the advantages of this method, at least by now.
FM-F3 Resource identifier in metadata	Data and metadata are included in a unique database. Reference from metadata to data is implicit.
FM-F4 Indexed in a searchable resource	CABRI is based on SRS and as such it is not indexable by google and other search engines. The contents of CABRI are however indexed by google through the HyperCatalogue, that is constituted by a set of HTML indexes of the resources, mainly by taxonomy and name. Indexes include links to the SRS pages where the detailed information is shown for each resource.
FM-A1.1 Access protocol	Information is available in HTML in an open web site. HTTP is used for retrieving data.
FM-A1.2 Access authorization	Access is completely free. No authorization is required neither to access nor to download the description of microbial resources.
FM-A2 Metadata longevity	Metadata is lost along with data when the resource is no more available for distribution.
FM-I1 Use a knowledge representation language	There currently is no way to express the data by using a devoted knowledge representation language.
FM-I2 Use FAIR vocabularies	CABRI adopted since its constitution various public reference vocabularies and taxonomies, as defined by the CABRI Guidelines for the Catalogue Production []. Some internal reference list is also used. None ontology is presently used.
FM-I3 Use qualified references	No qualified references are used
FM-R1.1 Accessible usage license	An explicit license for data access is not defined. It is assumed that access is free and reuse is permitted, but this information is not made explicit.

FM-R1.2 Detailed provenance	The field related to Literature is meant to provide provenance of the most important characteristics of the microbial resource, but it is not clearly specified which information was derived from the cited reference and which was determined by the staff of the collection. Provenance of the material is provided by reporting its History (initial collector, centers and collections that have conserved the material over time before the current collection) and Origin (place, specified by means of country, region, town or geographic coordinates).
FM-R1.3 Meets community standard	The information is structured according to the CABRI and OECD Guidelines.

References

1. Linhuan, W.; Global catalogue of microorganisms (gcm): a comprehensive database and information retrieval, analysis, and visualization system for microbial resources, BMC Genomics 2013, 14:933
2. GCM: http://gcm.wfcc.info/
3. P. Romano, M. Kracht, M.A. Manniello, G. Stegehuis and D. Fritze, The role of informatics in the coordinated management of biological resources collections, Applied Bioinformatics. 2005;4(3):175-86.
4. CABRI Network Services: http://www.cabri.org/
5. Wilkinson MD, Dumontier M et al. The FAIR Guiding Principles for scientific data management and stewardships. Scientific Data 2016, 3:160018. https://www.nature.com/articles/sdata201618 .
6. CABRI Guidelines for Catalogue Production: http://www.cabri.org/guidelines/catalogue/CPcover.html
7. Wilkinson MD, Sansone S-A, Schultes E, Doorn P, da Silva Santos LOB, Dumontier M. A design framework and exemplar metrics for FAIRness. Scientific Data. Nature Publishing Group; 2018;5:180118.