

Report on MOSAICS Integrated Conveyance System

Summary

MOSAICS central objective is the development of an integrated conveyance system to manage the access and benefit sharing issues related to microbiological resources, in the context of the CBD and the enforcement of other relevant international rules.

The purpose of the project is to design procedures and documents that set the rules at the source or original provider and during transfers. MOSAICS looks for an effective uncomplicated 'lightweight' system to enable tracking of the biological resources from origin to the final user, through the use of Globally unique identifiers (GUIDs).

In the field of microbial genetic resources, strain labels are used as locally unique identifiers. These may consist of a culture collection acronym followed by a number, or may just be any name given to the strain by an individual researcher. At this given moment, there is no universal way to refer to all resources stored in the different culture collections or private collections, and thus no reliable way to detect if multiple digital resources present data on the same biological source.

Persistent unique identifiers for global use need to combine both the strain label and a persistent location were to retrieve the information.

An initiative in this context, StrainInfo.net, which is a cooperation pilot project between the Laboratory of Microbiology and the Department of Applied Mathematics, Biometrics and Process Control of the University of Ghent and the Information Network Centre of the Chinese Academy of Sciences in Beijing [7,8] operates through an Integrated Strain Database, a curated central repository that provides a complete and correct view on the synonymous labels assigned to biological specimen during their lifetime. The StrainInfo.net portal adds to the commonly used strain numbers a more persistent and dumb identifier; in order to incorporate it within a larger namespace that provides extended unicity.

Taking advantage of the StrainInfo.net project, which was available as an example within the framework of MOSAICS, a hypothetical model was build for assigning GUIDs to biological resources.

Preferentially, such an integrated database should be located at the World Data Centre for Micro organisms (WDCM; <http://www.wdcm.org>), which is the heart of the World Federation for Culture Collections (WFCC) and already retains an ID system for registrated culture collections and institutions.

In the framework of MOSAICS, we would recommend to microbial resource centres and microbial data providers or database owners to work towards registration at WDCM and assignment of persistent globally unique identifiers to their material items and data elements.

Unique identifiers do, by no means intend to replace the traditional labelling of strains, genes or other data elements, but allow incorporating them in a larger namespace that provides an extended unicity and interoperability.

Among the different models of identifier systems, Digital Object Identifiers (DOI's) seem to be the most appropriate system for tracking of microbial resources, with the strongest business plan and safest expectations for the future.

DOIs can be assigned to any identity, for use on digital networks. Information about a digital object may change over time, including where to find it, but its DOI will not change.

Context

The world is changing; people are more interested in the environment, worried about climate change, loss of biodiversity and many other matters. Appreciation of this has led to increased concern about and regulations for biological resources. Authorities are now demanding that accession history of a particular specimen be documented to ensure that each and every specimen was legally acquired. Nations value their biodiversity and are granted legal rights to it by the Convention of Biological Diversity (CBD, <http://www.biodiv.org>).

The March 2003 CBD “Open-Ended Inter-Session meeting on the Multi-Year Program of Work for the Conference of the Parties up to 2010” requested in its conclusions to work further on the processes, modalities and scopes of implementation of access and benefit sharing (ABS), including the “Bonn guidelines”; and also to check their effectiveness. This was reaffirmed during the third meeting of the working group on ABS in Bangkok in February 2005.

To abide by the CBD rules of access and to enable benefit sharing, the key issue is tracking of the (micro) biological resources.

MOSAICS objective

MOSAICS central objective is the development of an integrated conveyance system to manage the access and benefit sharing issues related to microbiological resources, in the context of the CBD, the drawing up of the “Bonn Guidelines”, and the enforcement of other relevant international rules. Such a system must have three features:

- provide reliable tools to evaluate the economic value of microbiological resources,
- utilise validated documents with standard provisions to enable tracking via an uncomplicated procedure,
- be widely used by microbiologists, working in culture collections, public or private research institutes, commercial companies, non profit organizations, etc.

In general, this integrated conveyance system must thus be compatible with the political framework, be easily applicable in most conditions, full the needs of both users and providers and be an attractive system enforceable without constraints.

Tracking versus traceability

MOSAICS [1] has defined a procedure to implement the rules of access. It can be summarized as follows: register the source of the (micro) biological resource and trace it to its final destination.

The purpose of MOSAICS is to design standard documents that set the rules at the source or original provider and during transfers. These documents constitute the documenting part of the tracking system. In complement, MOSAICS looks for an effective uncomplicated ‘lightweight’ system to enable tracking of the biological resources from origin to the final use. During the different transfers, the biological resources are subjected to very different environments. Only when the transfers are restricted to moves and/or exchanges between

biological resources centres, which have a well-established quality management system and compatible legal systems, control and conveyance all along the way is possible and true traceability can be achieved. In all other cases, this total conveyance can not be guaranteed, since gaps in the quality control may occur, and the term TRACKING should be used, rather than traceability, referring to the follow up of the material from original provider to end user rather than registration of every single movement of the biological resource.

The MOSAICS project aspires to make recommendations on the use of unique identifiers to enable tracking of biological resources. These identifiers constitute the practical tool of the tracking system and should preferentially enable tracking through an electronic path, organised as a build in system, allowing *ex ante* and *ex post* conveyance.

Unique Identifiers

Focusing on the need for tracking and interoperability, some examples are given below, which clearly illustrate the need for 'unique identifiers' in different life science fields. This need has today been resolved in many different ways, for the different applications:

1. The Entomological Collection Network [2] attempts to tie the data derived from a specimen to that particular specimen by using attached barcodes. According to them, barcodes, while still expensive, allow the identification of individual specimens and greatly reduce the cost of subsequent data handling.
The problem of prospective data capture has been solved by one collection: INBio; barcodes are attached as part of the labelling process, data on locality, time, collector,... are captured when the print order for the label is generated (60-100 characters per label). The problem of retrospective data capture can be solved by using a similar approach. When researchers study previously collected specimens, they capture the specimen label data. Each specimen that is handled gets a unique number that links the specimen to an electronic data record.
2. Bar-coding of life (www.barcodinglife.com) is a project that uses DNA sequences as genetic barcodes. A 648 base-pair section of the mitochondrial cytochrome oxidase I (COI) gene has been shown to provide species-level resolution in varied animal phyla. The COI database could serve as the basis for a global identification system, a tool for taxonomists and a cost-effective system with which non-specialists can assign unidentified specimens to known species.
The consortium for the barcode of life is an international initiative of natural history museums, herbaria, other biodiversity research organisations, government organisations and private companies.
The weakness of this approach however, lays in the choice of the sequenced gene. Allowing powerful discrimination within the animal Phyla, the COI system does not necessarily work for other groups. For Archeae and Bacteria for instance, the method is complicated by the fact that horizontal transfer of DNA has heavily impacted their genomes.
3. The International Plant Exchange Network (IPEN, <http://www.bgci.org.uk/abs/ipen>) is an exchange system for botanic gardens for non-commercial purposes according to the CBD. The objectives are to comply with the obligations of the CBD and system transparency to countries of origin. All plant material supplied by an IPEN member

therefore needs to be accompanied by an IPEN number that remains connected with the material and its derivatives through all generations to come. With the aid of this number it is possible to track where and under which conditions the plant entered the network.

4. The World Data Centre for Micro organisms (WDCM, <http://www.wdcm.org>) assigns unique identifiers to the different registered culture collections. Collection acronyms followed by a number, are used to refer to a certain strain and act as such as an identifier at the strain level. When strains are being exchanged between collections however, synonym acronyms may cause great confusion. This problem is resolved by WDCM, by assigning a unique identifier to the collections or institutes that provide the strains.
5. The Organisation for Economic Cooperation and Development (OECD, www.oecd.org) documented on the designation of unique identifiers for transgenic plants [3]. They describe the unique identifier as being a key attributed to a biotech product, which could unlock information from a range of databases, as well as a harmonised unique entry point enabling information management related to that product. So the unique identifier should facilitate the ability to cross reference information in different databases, and improve access to and management of information by regulators and other interested stakeholders.
6. DNA or protein sequence databases make use of accession numbers as the identifier for a given sequence.

Persistent Globally Unique Identifiers (GUID)

In the examples mentioned above however, the role of the identifiers is restricted in most of the cases to the assignment of a **LABEL** to the material or data involved. The identifiers are developed for well-defined purposes and for use within a small context, scope or application and are only locally unique.

Within a global context, it is not only harmonisation that is of major importance, but also the need for more persistent globally unique identifiers.

In comparison, on the web, simple uniform resource locators (URLs) have the role of just identifying the location of a given digital object. URLs have largely been used to identify and access web-based digital resources, but they are by no means reliable or persistent and therefore uniform resource identifiers (URIs), Persistent URLs (PURLS) and other systems are now used.

Persistent unique identifiers for global use need to combine both the label and the persistent location were to retrieve the information. Besides assigning an ID to a resource, also the resource, for which the information can change over time, must be assigned to the ID.

The publishing sector has progressed the use and management of persistent identifiers more than any other discipline: the need to unify in one scheme the document management, digital libraries, copyright registration, object based software and to enable core interoperability, integration of disparate sourced data and the ability to trace ownership to manage rights led to the launch of the DOI initiative (Digital Object Identifier, www.doi.org) [4], which has been

the first system for persistent and actionable identification and interoperable exchange of managed information in the digital environment (see further).

Following the DOI initiative, many other systems have been reinvented. One of them, the Life Science Identifier system (LSID, <http://lsid.sourceforge.net>) [5], has been specifically adapted for life sciences. The LSID is a URN specification from the Interoperable Informatics Infrastructure (I3C), with members from life science Companies, academic labs and vendors as IBM, and the Object Managing group (OMG).

I3C (<http://www.i3c.com>) was launched in early 2001 at the request of the life science community, with the mission to coordinate disparate efforts around the world and to drive data and tool interoperability across the value chain towards the goal of accelerating basic research, drug discovery and development. This initiative however, seems to have disappeared again today. This proves that an appropriate business plan is as necessary as the technical tools to ensure permanent operation.

Persistent unique identifiers for microbial resources

In the field of microbial genetic resources, strain labels are used as locally unique identifiers. These may consist of a culture collection acronym followed by a number, or may just be any name given to the strain by an individual researcher.

There is thus, at this given moment, no universal way to refer to all resources stored in the different culture collections or private collections, and thus no reliable way to detect if multiple digital resources present data on the same biological source.

Furthermore, taxonomic names or strain numbers have constituted the key link between different databases, however, in the absence of a single comprehensive database of organism names and synonym strain numbers, individual databases lack an easy means of linking information.

According to Garrity and Lyons [6], a resolution system is required that can handle the complex relationships between biological names and the entities they denote. They believe that an implementation of the DOI system may provide the most robust and future-proof solution and they are developing a model for assigning DOIs to prokaryotic taxa as a test case. Though the definition of a taxon may change and its nomenclature may be redefined, the DOI will persist; leaving a forward-pointing trail that can be used to reliably locate digital and physical resources. This is of course also extensible to the level of individual genes.

Another initiative in this context, StrainInfo.net, is a cooperation pilot project between the Laboratory of Microbiology and the Department of Applied Mathematics, Biometrics and Process Control of the University of Ghent and the Information Network Centre of the Chinese Academy of Sciences in Beijing [7,8]. The StrainInfo.net portal envisions the establishment of a technology platform that works towards the use of multi-perspective integrated information in a broadened context. At the heart of this portal lays an Integrated STRAIN database, a curated central repository that provides a complete and correct view on the synonymous labels assigned to biological specimen during their lifetime. This data repository is constructed automatically through the seamless integration of label equivalence information as it is disseminated through the online catalogues of CC and BRC's.

Almost all BRC's keep track of the history of their resources, from the point of deposit, back to the initial point of isolation. This linear information however, only gives a fragmented view on the complete trace the strains have followed. With the help of the StrainInfo.net portal

acting as an information broker between all online catalogue entries of the BRCs that have a given strain in their holdings, it gets straightforward to manually extract all history information of a strain as it is fragmentarily recorded over all these data sources.

The StrainInfo.net portal adds to the commonly used strain numbers a more persistent and dumb identifier; in order to incorporate it within a larger namespace that provides extended unicity. As a result of incorporating the unique identifiers maintained by the StrainInfo.net portal within well-established global network identification infrastructures, the flexibility and interoperability of the identifier will no longer be related to its use to indicate biological resources but can possibly be linked to additional services that work completely independent of the StrainInfo.net system.

Very recently, the International Nucleotide Sequence Databases (EMBL, GenBank and DDBJ) decided to promote the inclusion of specimen identifiers in the sequence database. The identifier will be composed of the institution code and the specimen number; however the format has yet to be fixed. The database of institutions will be maintained by GenBank/NCBI, which will therefore rely on stable databases of the directory of institutions, as is the CCINFO database from WDCM.

Use of Persistent Unique Identifiers in the context of Access and Benefit Sharing

In summary, we can state that in the post-genomic era, a larger and larger portion of the value of any software application will consist of its ability to interoperate with other programs so that it reasons across a wider range of results. This additional value must be provided in an uncomplicated 'lightweight' way.

In the framework of MOSAICS, we would recommend to microbial resource centres and microbial data providers or database owners to work towards the assignment of persistent globally unique identifiers to their material items and data elements. These identifiers are of major importance for the appropriate management of both resources and related information and form the key element, rather than certificates of origin, for the tracking of biological resources. This is a prerequisite for reliable access and benefit sharing issues.

Certificates of origin only describe the source of the material and do not allow as such tracking transfers or exchanges of material or related information. However, these certificates stay of course compatible and, more important, complementary with the identifier system. By assigning an identifier to these documents too, it will be possible to link them automatically to the corresponding material and data elements.

Unique identifiers do, by no means intend to replace the traditional labelling of strains, genes or other data elements, but allow incorporating them in a larger namespace that provides an extended unicity and interoperability.

In principle, different kinds of identifier systems can be used, although it would be preferential to retain only one system in order to achieve maximal harmonisation and global uniformity.

In practice, to take advantage of an identifier system requires 2 pieces of software: a client piece within an informatics application and a server piece associated with the actual data. Once database owners have a database in place, the information would be sent to an authority, which contains a list of all available data resources. An application that wants to access the

data needs client software. The application makes a request to the authority, which returns a document that includes the location of the data and metadata that contain the practical information.

The recommendations made within this framework fit well with these made by the World Health Organisation (WHO, www.who.org) through the International Agency for Research on Cancer. They aim to work towards the defining of a persistent and unique identification and coding system for their medical data files.

However, as described extensively above, identifiers in the context of MOSAICS must be unique and must also be persistent to ensure their role in reliable tracking of the material for access and benefit sharing issues.

Digital Object Identifiers (DOI's) as the most appropriate system for tracking of microbial resources?

DOIs can be assigned to any identity, for use on digital networks. Information about a digital object may change over time, including where to find it, but its DOI will not change. Using DOIs as identifiers makes IP in a networked environment much easier and more convenient and allows the construction of automated services and transactions. It provides a system for persistent and actionable identification and interoperable exchange of managed information on digital networks [4, 9].

The system is managed by the International DOI Foundation (IDF, founded in 1998), an open membership consortium including both commercial and non-commercial partners, and has recently been accepted for standardisation by ISO. The IDF provides implementation through agreed standards of governance and scope, policy as well as a technical infrastructure and a social infrastructure; it is a central authority and maintenance agency [4, 9].

The DOI system was built using several existing standard-based components, notably the Handle resolution system and the indecs Data Dictionary, which have been brought together and further developed to provide a consistent system. The Handle system enables resolution to multiple associated data [4, 9].

The system consists of several components:

- a specified standard numbering syntax
- a resolution service
- a data model (metadata tools) allowing interoperability
- procedures for the implementation

Any existing numbering schemes and any existing metadata schemes can be used within the DOI system. E.g. 10.2245/LMG 3654. [4, 9]

DOIs are widespread used now, with over 17 million DOI's assigned, from over 1000 naming authorities.

Examples of projects using DOI applications for scientific data are the German National Library of Science and Technology (TIB) and the Names for Life project.

Like domain name registration, DOI assignment requires a fee and agreement to follow the defined standard and rules. This makes the system managed. The model selected for a long-

term position of the DOI organisation is a body that is not reliant on external sources, but is a self-funding system that can be supported in its perpetuity from its own resources (fee for participation, not for use of a DOI once issued).

Registration agencies hold a 'franchise' on the DOI. In exchange for a fee to the IDF, and a commitment to follow the ground rules of the DOI system, they are free to build their own offerings to a particular community (fee is based on number of DOI's assigned). [10]

Among the different models of identifier systems, the DOI system seems to be the one with the strongest business plan and safest expectations for the future.

A possible scenario, as recommended by MOSAICS

Taking advantage of the StrainInfo.net project, which was available as a model for testing within the framework of MOSAICS, a hypothetical model was build for assigning GUIDs to biological resources.

Preferentially, such an integrated database should be located at the World Data Centre for Micro organisms (WDCM; <http://www.wdcm.org>). WDCM, which is the heart of the World Federation for Culture Collections (WFCC) already retains an ID system for registered culture collections and institutions.

MOSAICS thus recommends that WDCM further develops the existing system of collection registration and makes it fit into a well established and funded global system, which can also function as a registration authority for GUIDs.

Also for biological resources (BR) other than microbial resources, a similar procedure would be recommended, such that most *ex situ* BR could be managed through one global system.

The participation of professional networks and federations must than be organized on a modular concept, allowing gradual connection of collections, institutuions and scientists to a compatible permanent system, without making them interdependent.

At the level of microbial resources, MOSAICS suggests that the WFCC Board, including the WDCM Director, monitor the registration and GUID assignment and set up a program to improve the (tracking) system, helping culture collections to scope with the technical and administrative hurdles. Such a development, possible at the level of the WFCC, could serve as a model for other professional networks.

An appropriate management system has to be appointed, stipulating the role and responsibility of culture collections and individual researchers within this system, by deposit, transfer or exchange of microbial resources and information.

The WFCC should delegate representatives to foster proper communication in this debate.

MOSAICS suggests close collaboration with IUMS and IUBS to define a collaborative scenario between WFCC and the scientific community at large.

References

- [1] MOSAICC stands for “Micro-Organisms Sustainable use and Access regulation International Code of Conduct “. The MOSAICC Code of Conduct now available on internet (www.belspo.be/bccm/mosaicc) is the result of five successive drafts improved through dialogue between MOSAICC partners and other experts.
- [2] Thompson, F.C. (1994). Bar Codes for Specimen Data Management. *Insect Collection News*, 9, 2-4.
- [3] OECD Documents: (ENV/JM/MONO (2001)5; 2001 and ENV/JM/MONO (2002)7; 2004).
- [4] Paskin, N. (2005). The DOI Handbook. Edition 4.2.0, International DOI Foundation, Inc.
- [5] Life Science Identifiers RFP response, OMG Document lifesci/2003-12-02.
- [6] Garrity, G.M., Lyons, C. (2003). Future-proofing Biological Nomenclature. *OMICS: a journal of Integrative Biology*, 7, 31-33.
- [7] Dawyndt, P., Vancanneyt, M., De Meyer, H. & Swings, J. (2005). Knowledge accumulation and resolution of data inconsistencies during the integration of microbial information sources. *IEEE Transactions on knowledge and data engineering*, vol. 17, 8, 1111-1126.
- [8] Dawyndt, P., De Baets, B., Zhou, X., Ma, J. & Swings, J. (2005). StrainInfo.net: Holding a wealth of downstream information on microbial resources right in our hands. In preparation.
- [9] Paskin, N. (1993). A 2003 Progress Report. *D-Lib Magazine*, 9.
- [10] Paskin, N. (2005). Digital Object Identifiers for Scientific Data. *Data Science Journal*, 4, 1-8.