Food and Agriculture Organization of the United Nations

The International Treaty ON PLANT GENETIC RESOURCES FOR FOOD AND AGRICULTURE

Guidelines for the optimal use of Digital Object Identifiers as permanent unique identifiers for germplasm samples- v.1

23 September 2016

1. Introduction

These guidelines are based on a broad consultative process and describe the main features and benefits of Digital Object Identifiers (DOIs) associated to germplasm samples and a set of basic rules for users to determine when to assign them[1]. This text also provides background on the development of DOIs and references to previous discussions.

It should be emphasized that the DOI is associated to the physical sample, not to the description of the sample. DOIs may also be assigned to publications or datasets created from the sample, and these should be related to the sample's DOI. However, such DOIs are not the subject of this document and should not be confused with the sample's DOI.

2. Background

Several communities have highlighted the importance of creating and adopting permanent unique identifiers (PUIs) for germplasm samples, to overcome the shortcomings of existing systems of identification. For modern regulated cultivars, many countries have quality control systems in place to ensure that one cultivar name corresponds to one genetic entity. In these cases, the cultivar name thus serves as an effective identifier for legal purposes, at least within the territory/ies where the cultivar is regulated. On the other hand, for traditional varieties and wild species, one name typically encompasses a wide range of genetic entities. A traditional variety is often genetically heterogeneous within a seed lot, variable between seed harvested from different locations, and variable from year to year. Thus in these cases the cultivar name cannot be used to identify a genetic entity. In some cases, such as for Basmati rice in India and Pakistan, regulations specify a set of traits that must be displayed to record a sample as Basmati, but even so the term spans a wide range of cultivars and genetic entities.

Genebanks often conserve multiple samples that have the same cultivar name because they are genetically distinct[2] . Regardless of the cultivar name, genebanks attempt to maintain the genetic composition of accessions unchanged from sample to sample and from generation to generation. Where the sample originally received is genetically heterogeneous, the genebank

---

[1] The first version of the document "*Data required for the assignation of Digital Object Identifiers in the Global Information System- v.1*" is made available at the same time of these guidelines at http://www.fao.org/3/a-bp767e.pdf

[2] https://www.genesys-pgr.org/explore?filter=%7B%22institute.code%22:%5B%22PHL001%22%5D,%22alias%22:%5B%22malagkit%22%5D%7D#

may choose to split it into homogeneous groups to be conserved as independent accessions, or keep it as a heterogeneous population for conservation as a single accession (Sackville Hamilton *et al.* 2002). Accession identifiers are assigned to enable each accession to be identified uniquely, and rigorous quality standards are followed to ensure that samples of the accession remain true to type (FAO 2014).

Genebanks have tried for decades to encourage breeders and other users to identify samples by accession identifier rather than variety name. However, many authors continue to use the names varieties in their publications without identifying the specific strain or source of the sample they used[3] , and breeders may even use just the name of a species in their pedigree information. The accession identifier as a unique identifier for accessions is effective within the context of the genebank concerned, but has not gained acceptance as a germplasm identifier outside the scope of the genebanks.

Molecular and informatics technologies require an even higher level of precision for the identification of the genetic entity analysed. Samples of landraces and wild relatives are typically subjected to some process of genetic purification, e.g. by one or more generations of single seed descent, to ensure sufficient homozygosity in the plant being genotyped. As with genebanks, although molecular laboratories have rigorous sample identification and tracking processes in place within the laboratory, the connection between the material and the information about the material is often lost on publication, through the absence of an accepted sample identification system. This is becoming an increasingly acute problem, as multiple laboratories need to collaborate to gain the full benefits of modern technologies. This requires a global sample identification system that the laboratories can share to ensure that they are working on the same sample.

Additionally, the International Treaty on Plant Genetic Resources for Food and Agriculture (the Treaty) brings a number of legal issues that require better identification of germplasm samples[4]. For example, holders of germplasm in the Multilateral System are encouraged to declare what material they have available. Providers are required to report what they have provided and to make available passport data and other associated non-confidential information. Recipients are required to make available the same type of information arising from their use of the material. In cases of dispute, the FAO may be required to investigate the transfers and use specific samples. Inclusion of material in the Treaty's Multilateral System of Access and Benefit-sharing (MLS) is based on samples, not on cultivars or on any other assemblage of entities.

Following recognition of the growing urgency to resolve the above issues, in January 2015, the Expert Consultation on the Global Information System on Plant Genetic Resources for Food and Agriculture organized by the International Treaty, discussed a paper on technical options to facilitate the establishment of data links[5] and decided to convene during 2015 a task force to further compare the properties of a number of established PUIs – Digital Object

---

[3] https://mikejackson1948.wordpress.com/2015/04/10/whats-in-a-name-im-on-a-germplasm-id-crusade/

[4] The '*Vision paper on the development of the Global Information System'* presents a succinct analysis of the legal requirements in Figure 1. http://www.fao.org/3/a-mo290e.pdf

[5] *Technical Options to Facilitate the Establishment of Data Links in the Field of Plant Genetic Resources for Food and Agriculture: Permanent Unique Identifiers*, http://www.fao.org/3/a-be643e.pdf

Identifiers (DOIs)[6], Life Science Identifiers (LSIDs), Archival Resource Keys (ARKs) – as candidates for identifying germplasm samples, as well as best practices and methodologies for their deployment as an essential element for the implementation of the Global Information System (GLIS)[7]. LSIDs were already losing traction and appear to have been largely discarded (Soiland-Reyes & Williams 2016) while some of the most interesting features of the ARK architecture had not been implemented yet.

As a result of this analysis DOIs were selected as the best technical option and a global survey was conducted "*Global survey on descriptors required to register material in the Global Information System*" to identify the minimum set of descriptors required. The Report of the survey[8] showed that there was a strong consensus on a strategic key set of descriptors to be considered mandatory and few more to be proposed as highly recommended.

Following the request of the Sixth Session of the Governing Body of the International Treaty on Plant Genetic resources for Food and Agriculture (ITPGRFA)[9], its Secretariat concluded an agreement with a DOI Registration Agency at the end of 2015 to be able to offer them for free to the plant genetic resources community.

Selection of DOIs, the descriptors definition, and the development of the technical infrastructure were just a first step. More important and less obvious is to determine *what entity* needs to be identified. Clear guidelines are needed for users to know when to register DOIs and for what germplasm.

3. Main features

The above scenarios all share the following features in common:

1. There is no need for DOIs to track samples through internal workflows within a laboratory or field or genebank (on the principle that existing systems work, so inventing a new one is not needed): no plot numbers, no test numbers, no cross numbers, no test tube numbers, no internal tracking numbers, no generation identifiers, no seed packet identifiers. Users will of course need these systems for their own process controls, but they would not need permanent unique identifiers for these purposes. Consequently the design and implementation of the system should not, at least initially, include provision for internal process controls, although the system could be extended to include it in future.
2. DOIs are needed for samples transferred between organizations.
3. The need is to identify samples of germplasm that may have no existing definition of their genetic or phenotypic composition.
4. The need is to identify each sample by its holder as well as by its genetic identity.

---

[6] ISO 26324 defines the syntax for a DOI name. The DOI name does not replace an identifier used in another scheme and can be used in conjunction with another identifiers.
[7] http://www.fao.org/plant-treaty/areas-of-work/global-information-system/en/
[8] Report and Analysis of the Global Survey on Descriptors required for PGRFA material, 2015. FAO: http://www.fao.org/3/a-bp470e.pdf
[9] Resolution 3/2015, *The Vision and the Programme of Work on the Global Information System*. http://www.fao.org/3/a-bl140e.pdf

5. Therefore, a DOI remains attached to a sample and all associated subsamples and progeny created as part of the holder's internal workflows and management of that initial sample, provided that those subsamples and progeny remain under the holder's management and, to the holder's knowledge, genetically the same, and the holder does not choose to use DOIs for finer internal tracking.
6. Therefore when a sample is transferred from a provider to a recipient, the sample received by the recipient must acquire a new DOI, distinct from the DOI of the provider's parental sample.
7. Moreover, if the holder believes that a subsample or progeny is genetically different from the original (whether unintentionally through processes such as drift, natural selection, natural mutation, cross-pollination, seed admixture, or mislabelling; or intentionally through processes such as crossing, induced mutation, transgenic modification, selecting a plant with a specific trait, or purifying an accession through single seed descent), the DOI of the original sample must not be used for the changed sample and a new one must be obtained if the sample is transferred outside the holder's organization.

The International DOI Foundation (IDF)[10] and its Global DOI System provide an infrastructure for persistent unique identification of objects of any type. The acronym DOI refers to "Digital Identifier of an Object". McMurry *et al.* 2015 list 10 simple rules for identifiers for web-based life science data. When the identifier is used to identify a physical object rather than just a record in a database, an additional requirement is critical: there must be an unbreakable pairing between the identifier and the object it identifies. The object must always be labelled with its identifier. If it's moved to a new container, the label must move with it.

An essential component of workflow management systems are carefully defined standard operating procedures ensuring that the label always stays with the object. Critical steps always need checks and backups. These are intended as more reliable objective alternatives to human memory; but no matter how good these objective alternatives are, human memory of the correct identifier associated with an object is always a useful backup. To this end, identifiers used locally within a genebank or laboratory should be brief (as brief as possible to enable unique identification of all the objects being identified), and may even contain semantic meaning or pronounceable names that help the operator remember accurately. This is directly contrary to the rules of McMurry *et al.* 2015.

This makes the requirements of local sample identification different from global permanent unique identification. It leads to the conclusion that, not only do PUIs not need to address local sample identification and process control (as in the first bullet above), but more strongly they should not do so. This leads to the conclusion that PUIs must be independent from any local system of identifying accessions, generations, seed lots etc., and should only be used in the context of transferring materials between provider and recipient. Existing identification systems should continue unchanged: genebank managers should continue to use their accession identifiers for local use; breeders should continue to follow their existing standards for naming crosses, selections, mutants etc. The only modification required is to add

[10] https://www.doi.org/

functionality to create a PUI when the context requires it – and therefore not to add PUIs for all samples.

As part of the services of the Global Information System on PGRFA, the Secretariat of the Treaty has set up a server to assign DOIs for germplasm samples. Any holder of germplasm can register with the Secretariat to use the service for free, regardless of whether the germplasm is held under the Multilateral System of the Treaty, and regardless of whether the holder is in a country that is Party to the Treaty.

4. Basic rules for assigning and using DOIs

The above considerations lead to the following suggested rules for DOIs:

Assigning DOIs

1. Germplasm holders are encouraged to obtain DOIs for any samples that they wish to make available to others.
    - This could be a public declaration that the sample is available, in advance of actually making it available; alternatively the holder could be in the process of actually providing a sample that does not yet have a DOI, and obtains a DOI to facilitate the process. In the first case, holders may obtain DOIs simultaneously for all living samples in their collections. In the second case, holders may obtain DOIs simultaneously for all samples in the planned transfer.
    - The sample could be available under the Multilateral System of the Treaty, accessible to anyone willing to accept the associated terms and conditions; or available to only specific collaborators at the discretion of the provider; and if the latter, then it may be completely outside the Multilateral System of the Treaty.
2. On seeking a DOI for a sample, the holder should make available at least a set of minimum descriptors, if not all the available information about the sample, by directly uploading data to the GLIS server.
3. Once a DOI is obtained for a sample, the holder would normally use the same DOI for all subsamples and progeny that remain under the holder's management and that are believed to be genetically the same.
4. If the holder deliberately takes subsamples or creates progeny with the intent to make them genetically different from a sample with DOI (for example by growing a single random seed or by selecting a specific variant), and wishes to make these available as well as the original, a new DOI should be obtained for the derived variant.
5. If the holder deliberately takes subsamples or creates progeny with the intent of keeping them genetically the same as the sample with DOI, but suspects they may have unintentionally diverged genetically from the original, the holder should exercise professional judgement, preferably backed up by evidence in the form of DNA of phenotypic tests, to determine whether a new DOI should be obtained for the new material.
6. When a holder provides a sample to a recipient, the provider and recipient are encouraged to use the integration protocol proposed by the Secretariat that will:
    a. Ensure that information errors are minimized;
    b. Ensure that recipient's sample is properly associated to the DOI for the provider's sample and all other relevant passport data;

c. Facilitate the association of a new DOI to the recipient's sample, since all the passport data, other than the recipient's own accession ID or other local identifier, will already be in the central DOI registry;
d. Encourage the recipient to use and reference the newly assigned DOI in publications, online articles and databases, in the same way as described above for all holders;
e. If the transfer was with a Standard Material Transfer Agreement (SMTA), inform the recipient that by referencing the DOI in publications, online articles and databases, the recipient will automatically comply with his/her obligations under article 6.9 of the Treaty.

Managing DOIs

1. Germplasm holders will prepare to use DOIs by adding DOI as an additional field in their database.
    a. It will be an optional field, since not every germplasm record will require a DOI;
    b. It will not be used for labelling samples.
2. If a holder loses a sample for which a DOI has been assigned, the status of the DOI can be changed on the GLIS server to "historic". The status cannot be historic at the moment of assigning the DOI (although this rule may be relaxed in future if a need emerges to use DOIs to track historical samples that are known only through information such as pedigrees).
3. The holder is encouraged to use the DOI in all publications and online articles and databases containing data collected on the germplasm. In a publication or online article, the first reference to the germplasm should include both its DOI and the local identifier normally used by the holder; subsequent references within a single publication would specify only the local identifier.

5. Example use cases

Case 1: genebank accession in the form received

A genebank manager conserves material obtained from elsewhere as an "accession" in the genebank and maintains it under the Multilateral System of Access and Benefit-sharing of the International Treaty on Plant Genetic Resources for Food and Agriculture. He/she wishes to make publicly known that samples of the accession are available to others with a Standard Material Transfer Agreement.

The genebank needs to obtain a DOI for the accession. Essential metadata would include the genebank holding the accession, the accession ID assigned by the genebank to the accession, date, method (which is "Acquisition" in this use case) and its genus or crop name. To achieve the goal of making publicly known the availability of the material, metadata would also include the fact that the material is available with SMTA. The genebank manager would also include available passport data describing the provenance of the sample, ideally by identifying the DOI of the donor's sample as the immediate source of the accession, if available. Associated descriptive data would also be attached by providing DOIs and URLs (called targets in GLIS) where such data are available online, e.g. in Genesys.

Case 2: genebank accession derived from material received

In a twist to the first case, the accession may be intentionally different from the material acquired from elsewhere. The genebank manager may have determined that the original material was a mixture of different types, separated it into its components, and conserve each component as a separate accession. Alternatively the genebank manager may have noted a rare variant in the original material and determined that it should be conserved as a separate accession (rather than discarded as an unwanted off-type). A third possibility is that, with the intent of making the material more easily used, the genebank manager may have subjected the material to one or more generations of single seed descent and self-pollination to make it more a homozygous pure line.

In this case, the associated metadata would be the same as in case 1, except a different method, that the DOI of the immediate source of the material would be the DOI of the variable material received by the genebank, together with data showing that the accession is a derived variant of the material received.

Case 3: a component of a genebank accession

In both the above cases, the genebank manager or other holder of genetic resources makes public the availability of material, at the level of the whole accession. However, where two laboratories share material for the purpose of collaborating in gene discovery, a DOI for the whole accession will often not be adequate. In this case, the holding laboratory will need to expose which specific seed or seed lot of an accession is sent to the receiving laboratory. Thus a DOI will be assigned to the specific material sent.

Also, whereas in the first two cases the genebank manager chooses to publicize the availability of material in advance of distributing it, in this case there is generally no need or advantage to publicising the existence of the specific seed lot in advance. Thus the DOI would usually be assigned at the time of sending it.

In a further variant of this case, a genebank manager could, if wished, assign different DOIs for every seed harvest that was used for distribution, at the time of distribution. This might be considered desirable where the genebank manager is concerned about the possible magnitude of unintentional genetic differences between harvests (by drift, selection, pollen contamination, seed contamination, mutation and mislabelling) and wishes to be transparent for all recipients, not only those who require more detailed information.

Given that the DOI system should not replace existing systems, full information tracking this seed lot back to its parental accession should be in the provider's own data management system and would not need to be replicated through DOIs. Thus it would normally be sufficient to specify the DOI of the parental accession as the "immediate source" of the specific seed lot. However, there would be nothing to stop the provider from assigning DOIs for some or all steps in the creation of the sample provided, should the provider wish to do so. This might be desirable, for example, for full transparency of the relationship between a sequenced sample of DNA and a sample of seed sent for phenotyping.

Case 4: novel distinct PGRFA

Recipients often use material received to create novel genotypes, for example through crossing with other samples, inducing mutations or genetic modification. In their own data management systems, they keep (or should keep) full pedigree data on how the novel genotypes were created from the original parental materials.

Under the Treaty, when developers of such novel genotypes send their "PGRFA under development" to collaborators with a Standard Material Transfer Agreement, they are required to identify the original materials from which they derived their PGRFA under development. Full pedigrees are not required.

The DOI system would simplify the documentation and processing for this case. The developer would create a DOI for the PGRFA under development; and its source DOI(s) would be the DOI(s) of the material(s) they previously received under an SMTA and used to develop the PGRFA under development.

Again there would be nothing to stop the developer from recording full pedigree information using DOIs, should they have other reasons to do so.

Case 5: transferring material with DOI

With a system of DOIs already in place, new opportunities arise for creating new DOIs with greater quality control over metadata. When a provider sends a sample to a recipient, as indicated above, a new DOI should normally be assigned to the recipient's sample.

If a DOI was previously assigned to the provider's sample, passport data for the recipient's new DOI will already exist in the system (other than the recipient's local ID for the newly received sample). Thus when the recipient seeks a DOI for the sample received, the correct metadata on the provenance of the recipient's sample can be supplied directly by the system, without requiring input from the new holder; in particular the recipient's DOI would be correctly linked back to the provider's DOI as its immediate source. This would reliably overcome what is arguably the most serious problem with transferring material from laboratory to laboratory, namely that data on provenance of the recipient's sample are not reliably recorded in the recipient's database.

References

FAO. 2016. Data required for the assignation of Digital Object Identifiers in the Global Information System – v.1". Rome. http://www.fao.org/3/a-bp767e.pdf

FAO. 2014. Genebank Standards for Plant Genetic Resources for Food and Agriculture. Rev. ed. Rome. http://www.fao.org/3/a-i3704e.pdf

McMurry J, Blomberg N, Burdett T, Conte N, Dumontier M, Fellows DK, Gonzalez-Beltran A, Gormanns P, Hastings J, Haendel MA, Hermjakob H, Hériché J-K, Ison JC, Jimenez RC, Jupp S, Juty N, Laibe C, Le Novère N, Malone J, Martin MJ, McEntyre JR, Morris C, Muilu J, Müller W, Mungall CJ, Rocca-Serra P, Sansone S-A, Sariyar M, Snoep JL,

Stanford NJ, Swainston N, Washington N, Williams AR, Wolstencroft K, Goble C, Parkinson H. 2015. 10 Simple rules for design, provision, and reuse of identifiers for web-based life science data. https://zenodo.org/record/31765

Sackville Hamilton, N.R., J.M.M. Engels, Th. J.L. van Hintum, B. Koo and M. Smale. 2002. Accession management. Combining or splitting accessions as a tool to improve germplasm management efficiency. IPGRI Technical Bulletin No. 5. International Plant Genetic Resources Institute, Rome, Italy.

Soiland-Reyes S, Williams AR. 2016. What exactly happened to LSID? myGrid developer blog, 2016-02-26. http://dev.mygrid.org.uk/blog/2016/02/what-exactly-happened-to-lsid/ doi:10.5281/zenodo.46804