The documentation of Plant Genetic Resources in Europe

Theo van Hintum, Centre for Genetic Resources, The Netherlands (CGN)

April, 2014

This is a personal view as input for the discussions at the workshop of the Documentation and Information Working Group 'Tailoring the Documentation of Plant Genetic Resources in Europe to the Needs of the User' to be held 20-22 May '14 in Prague, Czech Republic. It provides a conceptual background to the issues to be discussed and formulates a number of draft resolution.

Focus

The current *ex situ* PGR documentation landscape consists of very many 'data sources', i.e., documentation systems of germplasm collections. The passport information from these systems is to a large extent collected by National Focal Points (NFPs) to create National Inventories (NIs). The information in NIs is expected to be regularly uploaded to EURISCO that thus should always provide an overview of the content of these NIs and of the genetic resources in Europe. Parallel to this data-flow and depository, Central Crop Data Bases (CCDBs) have been created since the early days of the ECPGR Crop Working Groups, to collect passport and sometimes additional data on a crop specific basis, however many of these databases do not appear to be up to date or to provide information and features that are not already present in EURISCO.

On a global level, the GeneSys initiative tries to create an entry point to data on all PGR maintained in the world. Data providers and database managers operate in an environment with rapidly evolving technologies and policies. It is therefore expected that developments in information technology, sequencing technology and policies on access and benefit sharing (ABS) will have a large impact on PGR documentation.

The combined data from EURISCO and 46 accessible Central Crop Data Bases originate from 506 data sources in 43 countries. The largest data source, according to EURISCO is IPK in Germany with 128k accessions, followed by the Vavilov Institute in Russia with 123k accessions[1]. The number of accessions currently documented in EURISCO is 1065766, and the total number of accessions in Europe is expected to be around 1.3 million.

In this document some important issues related to the current situation and developments regarding the documentation of PGR in Europe will be discussed, and resolutions will be formulated. These resolutions are aimed at either the EURISCO Management (currently, since beginning of 2014 at IPK) that coordinates the network of NFPs and runs the EURISCO database and web interface, or the ECPGR Doc/Info Working Group, that acts as the steering committee for EURISCO and oversees the documentation activities of ECPGR.

The issues that are to be discussed include: quality and coverage of the passport data in EURISCO, characterisation and evaluation data in EURISCO, the future of CCDBs in relation to EURISCO, PGR Portals, the relation EURISCO – GeneSys, the relation

EURISCO – AEGIS, IT-PGRFA and MLS, the documentation of *in situ* and on-farm PGR, molecular markers and X-omics data, GRIN-Global and a general discussion of some other issues.

Quality and coverage of the passport data in EURISCO

Any search in EURISCO will quickly show that there is room for improvement of the quality and coverage of the EURISCO data. Elements to consider in this regards are:

- *Frequency of the update of NIs.* Although the average age of an accession record in EURISCO is 1.16 years, the 'oldest' 25% of the accessions have an average age of 3.17 years, and the oldest 10% of 4.96 years[2]. Low frequency of updates of the NIs might be an indicator of inefficiency of the respective national information systems. Therefore we should encourage a target frequency of updates. NFPs will be contacted to inquire about the reasons for low frequency of updates. The age of the data sets should also be made visible on the EURISCO web site.
- *Coverage of passport data.* Early December 2013, there were 1066k accessions documented in EURISCO. Following from a detailed comparison with the data in the CCDBs, it could be concluded that about 250k accessions are still missing in EURISCO[2]. This will be remedied by interacting directly with the appropriate NFPs to verify what are the bottlenecks to the inclusion of missing collections in the NIs. The data flow of passport data via the NFPs will not be changed.

Data quality. A superficial analysis of the data in EURISCO shows that on a number of levels the quality, in terms of formatting, consistency and reliability should be improved. This is a responsibility of the data sources, however EURISCO Management could (1) avoid the user seeing the low quality at the interface level, (2) continue to promote and support quality improvement at the data source.

1. EURISCO Management should promote and support more frequent updating with a target minimum frequency of once a year.
2. EURISCO Management should develop a strategy that will ensure that the coverage of *ex situ* accessions in Europe increases further.
3. EURISCO Management should make an inventory of data quality issues with appropriate actions to improve the situation.
4. The ECPGR Doc/Info Working Group should create a new steering committee for EURISCO that comments and approves the proposals made by the EURISCO management.

Characterisation and evaluation (C&E) data in EURISCO

Users would like to have access to phenotypic (C&E) data, however providing such access is more difficult compared to other data types due to (1) low availability of these data at the data source, (2) low reliability and comparability of these data (3) genotype times environment interaction making it difficult to compare data over data sets and (4) the heterogeneity of data sets making analysis difficult (different descriptor definitions, scales, etc.). A proposal for inclusion of C&E data in EURISCO has been formulated in 2009, discussed and approved by the ECPGR Doc/Info Network Coordinating Group. However, it has never been implemented by the former EURISCO Management.

Availability of C&E data from EURISCO is one of the outputs expected from ECPGR during Phase IX.

5. EURISCO Management should implement the proposal to provide access to C&E data from EURISCO, as formulated by the ECPGR Doc/Info Network Coordinating Group in 2009.

Future of Central Crop Data Bases in relation to EURISCO

The quality of the Central Crop Data Bases (CCDBs) varies strongly, quite a number of CCDBs listed on the ECPGR website[3] simply doesn’t exist, whereas others are very old (more than 15 years) or an old copy of the data in EURISCO. However, there are also a few CCDBs that are very well maintained, available on the web, and sometimes even providing C&E data, pictures, and/or valuable additional information such as contact details of the curator of the collection.

Some CCDBs have a better coverage as compared to EURISCO although often this appears to be due to no longer existing accessions. A recent analysis[2] showed that 29 out of the 46 CCDBs analysed (63%) had a smaller number of accession accessions of the species in the CCDB as compared to EURISCO. In total EURISCO had 30k accessions more than the CCDBs of the species covered by the CCDBs (4% of the total content of the CCDBs). However, especially France, Russia and Israel seemed to be much better represented in the CCDBs as compared to EURISCO, for the taxa covered by the CCDBs they had respectively 37k, 142k and 11k accessions in the CCDBs and only 4k, 123k and 3k accessions in EURISCO. (In this respect Russia was a special case since it appeared that the accessions reported to the CCDBs were partly different from those reported to EURISCO.)

Apart from a few strong CCDBs, with a clear added value to the user, many CCDBs did not bring clear benefits to the user and only seemed to duplicate existing efforts, sometimes generating confusion for the user. It has also become evident, in the process of selecting accessions for AEGIS, that only those accessions that are included in EURISCO are eligible to become part of the European Collection.

While a number of WGs are still considering their CCDB essential and worth investing in, others have come to the conclusion that it is a duplication of effort that could be avoided if EURISCO could accommodate more information, especially C&E data.

It is suggested here that ECPGR should no longer promote investing in CCDBs by the Crop Working Groups as indispensable tools for their activities. However, whenever self-sustainable and well maintained CCDBs are able to provide additional value to the user community, complementing EURISCO, these should be continued (see PGR Portals in the next issue).

6. Upon consideration that EURISCO will host C&E data, the ECPGR Doc/Info Working Group recommends the discontinuation of the CCDBs with the Crop Working Groups, presenting the alternative: PGR Portals, that can include a CCDB provided it has added value.

PGR Portals

To change the PGR information on the internet from being supply driven towards demand driven, the PGR Portal concept is being developed. A PGR Portal serves as an entry point for a specific PGR user group, providing the information needed by that group. For example, a portal could be developed for users of lettuce germplasm, for malting barley breeders or for people interested in traditional cultivars suitable for their back yard. It should provide access to all information relevant to this specific user group, such as, depending on the specific user group, information about the main genebanks from which material can be ordered, where and how research populations (RILs, NILs, MAGIC populations, etc.) can be accessed, links to available phenotypic data and links to available genomic resources. In addition, the portal may inform users about ordering procedures, including whether material can be ordered on-line, and if not, who needs to be contacted and how, what kind of material transfer agreement needs to be signed, and whether phytosanitary regulations do apply. It could collect experiences of users ordering material. The portals should evolve according to the needs of the user community, and hence they should develop into interactive platforms, where social media may be used to share experiences and ideas. Currently, pilots of such crop portals are being developed at CGN for lettuce and potato (e.g. http://pgrportal.nl/lettuce).

7. Based on the experiences with the prototypes of the PGR Portals, the ECPGR Doc/Info Working Group should stimulate and support the ECPGR Crop Working Groups in developing these information resources.

Relation EURISCO – GeneSys

Based on initial funding of the Global Crop Diversity Trust, Bioversity International and the International Treaty on PGRFA, a global database and website has been developed: GeneSys. At present it is coordinated by the Global Crop Diversity Trust. According to its site, GeneSys is a gateway from which germplasm from genebanks around the world can be easily found and ordered. It currently includes the data from EURISCO, GRIN and the CGIAR Genebank system, including many of the C&E data of the latter two. It is currently very much oriented towards the crop listed on Annex I of the IT-PGRFA.

If this system develops according to the expectations, it could be able to replace the web-interface of EURISCO, as from the user perspective there is no reason to restrict the domain of queries to Europe. The function of EURISCO would be reduced to a node to aggregate European information, and even this function might become redundant. It can be expected that data sources, in the near future, will offer their information either via web-services or via systems like the Integrated Publishing Toolkit as used by GBIF. Any authorised system, be it EURISCO or GeneSys, will, in that situation be able to use the data for their purposes. Making an intermediate aggregation, such as that of EURISCO would become redundant. In this situation EURISCO would become a network of data sources, the importance of which should not be underestimated. For the roles of EURISCO that cannot be taken over by GeneSys, such as the visibility of European material or the registration of AEGIS material, additional small systems using views on GeneSys could be built (similar to the additional functionalities required by the former SINGER community).

It is of high importance to the European PGR community to follow and influence the developments of GeneSys, so that it will support the functions of EURISCO. Given the fact that the chair of the ECPGR Doc/Info Working group is, on personal title, member of the GeneSys Oversight Committee, no further action is needed for the time being. On the longer run it would be desirable to have a more formal representation of ECPGR in the body guiding or steering the development and management of GeneSys.

A request from GeneSys to allow on-line ordering of material documented in EURISCO via automatically generated email messages to the curators of the material was turned down by the ECPGR Doc/Info Network. This was done for two very different reasons. First of all, providing passport information about material to EURISCO doesn't, so far, imply that this material is available for distribution, and a descriptor indicating availability was once considered unacceptable by the data source community (this might be reconsidered). Allowing on-line ordering would thus lead to frustration of both the user, who doesn't receive the requested material, and the genebank, that can't answer the requests. And secondly, some genebanks use on-line ordering procedures, with click-wrap SMTAs. Emails requesting material would be a major step back and create an undesired additional work load. However the latter problem could be avoided by accessing these local ordering systems via GeneSys. A new alternative that could be considered in this context has arisen that involves the infrastructure created by the IT-PGRFA for on-line generation of SMTAs. In any case, on-line ordering is an issue that might have to be reconsidered.

8. The ECPGR Doc/Info Working Group stays, via a formal channel such as a steering committee or advisory board, in touch with GeneSys to assure that the respective developments are planned in a harmonized, synergistic way.

Relation EURISCO – AEGIS, IT-PGRFA and MLS

EURISCO is currently used as a platform on which AEGIS indicates which accessions form part of the European Collection, if and where accessions are being safety-duplicated and similarly which are part of the Multi-Lateral System under the IT-PGRFA. At some stage it also has been considered to use the EURISCO infrastructure to report about SMTAs transactions to the IT-PGRFA, but that was shown not to be a good idea because of legal and confidentiality reasons.

In all decisions, it should remain clear that EURISCO serves to aggregate information on *ex situ* maintained PGRFA in Europe. Additional objectives or tasks should support and not hinder that objective. Additional functionalities of EURISCO, such as monitoring the change of AEGIS status in newly uploaded datasets, should be considered in that light, but always keeping in mind that EURISCO and AEGIS serve the same objective: increasing the efficiency of the European PGR system.

9. If AEGIS requests services from EURISCO these should be considered positively by the EURISCO Management. If these services require substantial investments, EURISCO Management should raise the issue with the ECPGR Doc/Info Working Group.

Documentation of in situ and on-farm PGR

The documentation of *ex situ* PGR in Europe is fairly well organised, data standards are in common use, and the flow of information to EURISCO goes smoothly. The situation in the *in situ* (crop wild relatives (CWR) in nature) and on-farm (landraces in cultivation) communities is not equally advanced. The communities, due to their heterogeneity in terms of actors, are much less coordinated and despite the creation of several tools and portals, there is still no agreed infrastructure for information exchange or aggregation.

The on-farm community can be primarily considered to be users of the *ex situ* PGR conservation efforts, since these *ex situ* efforts provide the long term (back-up) conservation function and to some extend the access to conserved germplasm for on-farm users. This implies little need of centralised documentation at the accession level of the on-farm community. Due to the dynamic nature of this domain, both in terms of actors and in terms of material, it can be questioned if such accession-based documentation makes sense at all. Only where it concerns a true conservation task, such as for fruit trees, a centralised information activity is feasible and has added value. This could follow the *ex situ* approaches and be incorporated in the current (*ex situ*) National Inventories. Minor modifications in the uploading format would have to be considered to that end. Systems documenting European landraces and traditional cultivars on the variety level would have added value for both the on-farm as the *ex situ* community and the development of such a system should be supported by the ECPGR Doc/Info Working Group.

In contrast, the CWR *in situ* conservation community should establish a close collaboration with the *ex situ* community to allow a mechanism for efficient prioritising of collecting missions where crop wild relatives are under threat. This requires an inventory and monitoring system for CWR across Europe. It should indicate where a CWR grows, its conservation status and under what conditions it can be accessed. Attempts to create such a system have not been very successful so far, due to several reasons. Although several initiatives have been made or are ongoing (PGR-Forum, AEGRO and PGR Secure tools, Bioversity's crop wild relatives portal and descriptors, the Trust's CWR initiative and inventory and several national CWR sites, inventories, portals, etc.) and a documentation system has been designed (the Crop Wild Relative Information System, CWRIS), only very few countries have actually made an inventory of CWR, and even fewer countries have a monitoring system in place. However, if such a system would be established its users would be the conservationists, not the PGR users, as the use of CWR should preferably go via genebanks that would collect the material in nature to make it available to the user community. Provided that PGR users should not be stimulated to go out and collect the material themselves *in situ*, integration of such a CWR inventory system into EURISCO would thus not make much sense, as EURISCO is primarily aimed at PGR users and genebank managers. However, PGR users might benefit from information about the occurrence of CWR conserved *in situ,* and the best approach to get access to those CWR.

10. The ECPGR Doc/Info Working Group should support, when called upon, the on-farm community in the creation of a register of European landraces and traditional varieties.
11. The ECPGR Doc/Info Working Group should support, when called upon, the *in situ* community in the creation of an inventory and monitoring system for European CWR *in situ*.

Molecular markers and X-omics data

The PGR community has not embraced the possibilities offered by the molecular marker technology, the number of generated data sets is limited, and the availability of these data has never become a real issue. It is quite probable that this will be very different for the genomics data and especially the sequence information that is already being generated. This information might prove so powerful in its ability to predict traits and identify valuable alleles, that the genebank community either finds ways and means to deal with these developments or become irrelevant to plant breeding and crop science in the medium term, and will have to fall back on only its conservation role.

These 'ways to deal with these developments' will have several components. Genebanks will have to become able to supply the material that answers the renewed demands of the user community: when possible homozygous lines, representing a broad diversity with good passport and phenotypic data allowing for genome wide association studies and the selection of allelic panels. On the information side, the community will have to assure that the information in the genebanks' databases can be linked to the information in the genomics community's databases. Initiatives such as DivSeek (previously named SeedSeq and the Digital Seed Bank) are looking at ways to achieve these goals.

12. The ECPGR Doc/Info Working Group has to stay involved in DivSeek initiatives and where possible play an active role in guiding the genebanks to establish the appropriate link with X-omics information.

GRIN-Global

The development of an open source genebank documentation system makes complete sense, since each genebank in the world needs a documentation system, and the required functionality of these systems is quite similar. So, if a kernel of code is developed centrally, the PGR community could pick this up and develop all locally required functionality as appropriate and share that with the community. Furthermore, data sharing mechanisms can easily be implemented, stimulating an easy exchange of PGR data (including SMTA reporting).

Based on funds from the Global Crop Diversity Trust, Bioversity International and the USDA, such open source system has been developed: GRIN-Global. It took much longer to develop than planned, and although the system is open source, no community of developers has been created yet. Since the system still appears to be rather immature, it has only been installed at very few sites, and has not yet been thoroughly evaluated. It is therefore very difficult to predict if GRIN-Global will be able to live up to the expectations. In any case, the ECPGR Doc/Info working group will have to stay informed and if possible involved in these developments. It should act as a channel for the information and experiences between the GRIN-Global community and the ECPGR community. The Working Group could also play a role in actively supporting the installation and further development of GRIN-Global, provided it proves to be a useful system.

13. The ECPGR Doc/Info Working Group has to become/stay involved in the further development of GRIN-Global and where possible play an active role.

Other issues

New tools and methods become available to analyse the data in PGR documentation systems and better serve the user or curator of PGR collections. GIS based software and methods such as Capfitogen for the analysis of germplasm in relation to environmental variables, the Focussed Identification of Germplasm Samples (FIGS) or the MaxEnt species habitat modelling based gap analysis approaches as developed by CIAT, might all become highly relevant to the PGR community. Also methods to select core collections or optimise sampling for germination testing are, to some extent, already available for adoption. However, at the moment they are only being used by very few genebanks.

Where these examples concern approaches that would make PGR conservation and use more efficient, there are also new requirements that will require new functionalities of local PGR documentation systems. A good example is the reporting duty regarding SMTA transactions under the IT-PGRFA.

The ECPGR Doc/Info Working Group would be the appropriate platform to facilitate genebanks to use these tools and methods, and to implement these functionalities. This will require a capacity building programme supported by an activity monitoring developments, testing tools and methods and developing supporting material. However, these activities cannot be developed on the basis of input in kind.

14. Depending on financial resources, the ECPGR Doc/Info Working Group should develop a capacity building programme to support the ECPGR members to effectively use new tools and meet new requirements regarding PGR documentation.

[1] Based on a recent analysis by Theo van Hintum of the EURISCO dataset exported November 19th 2013 with the CCDBs downloaded and received on request between November 13th and December 6th 2013. For matching the data all taxa were first mapped on GRIN-Tax, the 'valid species' as defined by GRIN-Tax was used for further analysis. The institute codes of FAO were used to identify institutes (sometimes renaming was required, such as in the case of old DDR or Yugoslavian institutes still used in some CCDBs). Matches were based on institutes and species.

[2] Based on a recent analysis by Theo van Hintum of the EURISCO dataset exported November 19th 2013. The age of a record is the time between upload of the NI to EURISCO and the day of analysis, it doesn't imply anything about the frequency of upload from the data source to the NI.

[3] The list of CCDBs was checked on the ECPGR website on November 2013, the current website, May 2014, has been updated and the links have been made up to date.