A.14. Information system and data management

A.14.1. Overview

Why is data critical to crop wild relative conservation and use?

It is widely accepted within the PGRFA conservation and user community that one major factor hindering effective conservation and use is the lack of easy access to data, as well as obstacles to information exchange due to the many different approaches in managing data. If we are to inventory and build a national management plan for CWR conservation, then consistent data collation and management is required.

To conserve CWR efficiently there is necessarily a significant requirement for data and associated information, that data needs to be sourced, managed and analysed to help ensure the most appropriate conservation actions are implemented. This process is likely to involve taxonomic, ecogeographic occurrences and temporal distribution, threats and conservation status and genetic structure data, as well as the ability to track using time-series data and predicted demographic and genetic changes within a species in relation to land management and environmental factors. The data sources are often not readily available and for CWR are particularly disperse because of the broad taxonomic range of species and the fact that much data are held by those outside of the PGR community. Accessing such information is not only time-consuming, but comparing data sets is often difficult due to the diversity of information management models used. If CWR are to be conserved and sustainably utilized, a means of bringing together this information into an accessible and standard format is required.

To help manage this data both CWR descriptors and information management tools have been developed. The first attempt to produce a set of CWR descriptors was made by the EC funded PGR Forum project\(^{104}\) and these were developed further within the GEF funded ‘In situ conservation of crop wild relatives through enhanced information management and field application’\(^{105}\) and are now being further developed with the EC funded PGR Secure project\(^{106}\); the current version of the CWR descriptors is available at PGR Secure helpdesk (http://pgrsecure.org/). Within PGR Forum a stand-alone information systems was developed to help make available CWR data for Europe and the Mediterranean to the user community, the Crop Wild Relative Information System (CWRIS) (see Box 59) and

\(^{104}\) See Moore et al. (2006) and http://www.pgrforum.org/cwris/cwris.asp
\(^{105}\) See http://www.cropwildrelatives.org/
\(^{106}\) See http://pgrsecure.org/
this was extended in the EC funded AEGRO project\(^\text{107}\) (see Box 60). Although there are currently no plans to develop CWRIS further, it is functional and can be used in the creation of national checklists for Europe and the Mediterranean countries, CWRIS users gain access to the checklist data by selecting the country or geographical units of interest and then downloading the dataset. These data can then be cross-checked against local floras, databases and other documentation, verified and edited as necessary to ensure it meets the national requirement.

**Box 59. CWRIS**

The Crop Wild Relative Information System (CWRIS - http://www.pgrforum.org/cwris.htm) was the first information management system specifically designed to facilitate CWR conservation and use. CWRIS has two main dimensions: taxon information breadth is provided by the PGR Forum CWR Catalogue for Europe and the Mediterranean, while the CWR descriptors for conservation and use for individual CWR taxa provide taxon information depth. The CWR descriptors provide a comprehensive set of data standards that can be used to effectively manage genetic conservation of CWR taxa and their component populations. The descriptors provide the structure within which existing data can be accessed or mapped onto the data model, and novel data can be provided. CWRIS was designed to facilitate access to CWR data for a diverse range of user communities, including plant breeders, protected area managers, policy-makers, conservationists, taxonomists and the wider public. CWRIS also provides access to ancillary information on the taxa contained in the Catalogue via links to external online resources, such as Mansfeld’s World Database of Agricultural and Horticultural Crops, GRIN Taxonomy, European Nature Information System (EUNIS), the IUCN Red List, Electronic Plant Information Centre (EPIC) and key publication search engines. CWRIS comprises:

- A searchable database of crop species and their associated wild relatives that occur in Europe and the Mediterranean region. The taxonomic back-bone to CWRIS was provided by Euro+Med PlantBase (http://www.emplantbase.org/home.html) version August 2005. CWRIS provides occurrence records according to geographic boundaries, not political boundaries.

- Information on the taxa contained in the database via external web links.

- A data model for the management of CWR information, with an emphasis on site and population data, which is required for the effective genetic conservation of *in situ* CWR populations. The data model is illustrated with a number of CWR case studies.

Source: Kell *et al.* (2008)

---

\(^{107}\) See http://aegro.jki.bund.de/aegro/index.php?id=95
Box 60. CWRIS PLIS

The Crop Wild Relative Information System (CWRIS - http://www.pgrforum.org/cwris.htm) was also extended to provide information at the species level within the EC funded AEGRO project, using four independent modules collectively called "Population Level Information System" for *Avena, Beta, Brassica* and *Prunus* European species population level occurrence data. The population level information system was designed to facilitate CWR conservation management and monitoring via:

a. Data exploration
   - Search for occurrences by taxonomic criteria (hierarchical search through taxonomic ranks including synonyms according to different taxonomic views)
   - Search for occurrences by geographic information (hierarchical search through levels of administrative units or within protected areas)
   - Combined search by taxonomic and geographic criteria
b. Data acquisition
   - Downloading results and displaying them on a map
c. Data contribution
   - Editing taxonomic and geographic data for atomization, harmonization and geo-referencing
   - Acquisition of population data in the field with portable data assistants and uploading these data to a central database.

The data exploration and data acquisition use cases have been fully implemented in CWRIS-PLIS (http://aegro.jki.bund.de/index.php?id=168), while the data contribution use cases have been only partly implemented.

Germeier *et al.* (2012)
Rubus plicatus Weihe & Nees, a CWR, in Lithuania (photo: Juozas Labokas).

A.14.2. Methodology

Information on CWR is available from a wide range of sources, but retrieving it presents a number of challenges. Firstly, in existing databases, such as those managed by plant gene banks, CWR accessions are not identified as CWR; this issue is not helped by the fact that in the current FAO/IPGRI Multi-crop Passport Descriptors V. 2\textsuperscript{108} the SAMPSTAT descriptor allows for designation of wild species samples but does not make a distinction between CWR and non-CWR wild species. Secondly, although information on CWR per se is possible only of specific interest to the PGR conservation and use community because CWR are ‘normal’ wild species they are also collected, conserved and studied by a broad community of taxonomists, ecologists, geneticists, physiologists, etc. and so when collating CWR information these other communities need to be consulted. Further these non-PGR communities often have significantly larger data sets than the PGR community itself. These challenges are not insurmountable but they do demand a carefully considered and tested approach (particularly with regard to obtaining information from non-PGR communities) and a considerable amount of time. However, like all data mining activities the more background data available the more predictive the analysis results in formulating effective conservation plans.

Information at the CWR at the taxon level is primarily gathered from the relevant literature: monographs, revisions, field guides, floras, gazetteers, articles, papers,

soil, vegetation and climatic maps, atlases, etc., while at the accession level it is
gathered from herbarium and germplasm collections of the target taxon from the
target area, and the latter will often involve visiting the herbarium or gene bank to
collect the data. However in recent years there has been exponential growth of
web-enabled ecogeographic datasets, most notably the Global Biodiversity
Information Facility (GBIF) established in 2001 (http://data.gbif.org), which
provides extensive access to global taxon nomenclature, taxon and accession
distribution, conservation and environmental data.

Table 5. Internet resources for CWR

<table>
<thead>
<tr>
<th>Data set</th>
<th>Description</th>
<th>URL</th>
</tr>
</thead>
<tbody>
<tr>
<td>European Native Seed Conservation Network</td>
<td>European database of major <em>ex situ</em> botanic garden gene bank holdings</td>
<td><a href="http://enscobase.maich.gr/">http://enscobase.maich.gr/</a></td>
</tr>
<tr>
<td>FAOSTAT</td>
<td>Agricultural statistics and data</td>
<td><a href="http://www.fao.org/">http://www.fao.org/</a></td>
</tr>
<tr>
<td>Gap Analysis Project</td>
<td><em>Ex situ</em> gap analysis results of 13 crop gene pools</td>
<td>gisweb.ciat.cgiar.org/gapanalysis/</td>
</tr>
<tr>
<td>GBIF</td>
<td>Global Biodiversity data</td>
<td><a href="http://data.gbif.org/">http://data.gbif.org/</a></td>
</tr>
<tr>
<td>GENESYS</td>
<td>Global database of major <em>ex situ</em> gene bank holdings</td>
<td><a href="http://www.genesys-pgr.org/">http://www.genesys-pgr.org/</a></td>
</tr>
</tbody>
</table>

109 Castañeda Álvarez *et al.* (2011)
<table>
<thead>
<tr>
<th>Resource</th>
<th>Description</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glob cover</td>
<td>European Space Agency Global Land Cover map, latest version = 2009</td>
<td><a href="http://ionia1.esrin.esa.int/">http://ionia1.esrin.esa.int/</a></td>
</tr>
<tr>
<td>IUCN Red List</td>
<td>Database of red list (extinction threat) assessments</td>
<td><a href="http://www.iucnredlist.org/">http://www.iucnredlist.org/</a></td>
</tr>
<tr>
<td>JSTOR herbaria</td>
<td>Herbaria resources</td>
<td><a href="http://plants.jstor.org/">http://plants.jstor.org/</a></td>
</tr>
<tr>
<td>Plant list</td>
<td>Working list of all known plant species</td>
<td><a href="http://www.theplantlist.org/">http://www.theplantlist.org/</a></td>
</tr>
<tr>
<td>Tropicos (Missouri Botanical Gardens, USA)</td>
<td>Herbaria resources</td>
<td><a href="http://www.tropicos.org">http://www.tropicos.org</a></td>
</tr>
<tr>
<td>UNEP WCMC World Database of Protected Areas</td>
<td>World Database on Protected Areas (polygons)</td>
<td><a href="http://www.protectedplanet.net/">http://www.protectedplanet.net/</a></td>
</tr>
</tbody>
</table>

**National program accession datasets**

<table>
<thead>
<tr>
<th>Country</th>
<th>Database</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>Russia</td>
<td>AgroAtlas</td>
<td><a href="http://www.agroatlas.ru">www.agroatlas.ru</a></td>
</tr>
<tr>
<td>Brazil</td>
<td>CRIA</td>
<td><a href="http://www.cria.org.br">www.cria.org.br</a></td>
</tr>
<tr>
<td>Japan</td>
<td>NIAS</td>
<td><a href="http://www.gene.affrc.go.jp/databases_en.php">www.gene.affrc.go.jp/databases_en.php</a></td>
</tr>
<tr>
<td>Mexico</td>
<td></td>
<td><a href="http://www.biodiversidad.gob.mx/genes/proyectoMaices.html">www.biodiversidad.gob.mx/genes/proyectoMaices.html</a></td>
</tr>
</tbody>
</table>

**Other accession datasets**

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td>CWRIS PLIS</td>
<td><a href="http://aegro.jki.bund.de/index.php?id=168">http://aegro.jki.bund.de/index.php?id=168</a></td>
</tr>
<tr>
<td>Harold and Adele Lieberman Germplasm Bank (cereals)</td>
<td><a href="http://www.tau.ac.il/lifesci/units/ICCI/genebank1.html">www.tau.ac.il/lifesci/units/ICCI/genebank1.html</a></td>
</tr>
<tr>
<td>Manchester Museum</td>
<td><a href="http://emu.man.ac.uk/mmcustom/BotQuery.php">http://emu.man.ac.uk/mmcustom/BotQuery.php</a></td>
</tr>
<tr>
<td>Natural History Museum, UK</td>
<td><a href="http://www.nhm.ac.uk/research-curation/collections/departmental-collections/botany-collections/search/index.php">www.nhm.ac.uk/research-curation/collections/departmental-collections/botany-collections/search/index.php</a></td>
</tr>
<tr>
<td>Royal Botanic Gardens Kew</td>
<td><a href="http://apps.kew.org/herbcat/navigator.do">http://apps.kew.org/herbcat/navigator.do</a></td>
</tr>
<tr>
<td>--------------------------</td>
<td>----------------------------------------</td>
</tr>
<tr>
<td>Royal Botanical Garden of Edinburgh</td>
<td><a href="http://www.rbge.org.uk/databases">www.rbge.org.uk/databases</a></td>
</tr>
<tr>
<td>SolanaceaeSource</td>
<td><a href="http://www.nhm.ac.uk/research-curation/research/projects/solanaceaesource">www.nhm.ac.uk/research-curation/research/projects/solanaceaesource</a></td>
</tr>
<tr>
<td>United States Virtual Herbarium</td>
<td><a href="http://usvirtualherbarium.org">http://usvirtualherbarium.org</a></td>
</tr>
</tbody>
</table>

The types of data managed will fall into four basic types, which may be subdivided:

- Ecogeographic data (taxonomic, ecological, geographic and genetic);
  - Taxonomy and nomenclature,
  - Degree of relationship between crop and CWR,
  - CWR uses: historic, current and potential,
  - Other uses: other than as a trait donor,
  - Current, historical and potential distribution, including:
    - Country occurrence/extent of occurrence,
    - Number of populations,
    - Record of extinctions,
    - Mapping function/GIS layers,
  - Genetic diversity and biology,
  - Ecology and habitat,
  - Threat status,
  - Conservation measures, including:
    - Occurrence in named protected areas and genetic reserves,
    - Conservation management techniques,
    - Ex situ holdings in gene banks,
  - References to specific research projects,
  - Contacts,
- Field population data (passport);
  - Precise population location (distributional polygon),
  - Land management regime (protected area, private ownership, common land),
  - Population characteristics,
    - Size,
    - Cover
    - Genetic characterisation,
    - Age structure,
- Obligate associated species (associated keystones, pollinators, seed dispersers)

- Conservation management data (curatorial);
  - In situ criteria
    - Management regime and interventions
    - Monitoring regime
    - Place in national, regional and global CWR networks
    - Place in non-CWR specific conservation networks
    - Local community participation
  - Ex situ criteria
    - Gene bank holding collection,
    - Location of seed in gene bank,
    - Germination and regeneration testing,
    - Access and benefit sharing policy,

- Characterization and evaluation data (descriptive);
  - Taxonomic morphological description
  - Genetic description,
  - Agronomic description
  - Breeder desired characteristic evaluation (disease, pest, drought resistance, etc.)

Although this list of CWR data types is extensive it is not exhaustive, it is indicative of the types of data involved in CWR conservation and use.

Each of these data types are collated using some type of standard descriptor. A descriptor may be defined as “any attribute referring to a population, accession or taxon which the conservationist uses for the purpose of describing, conserving and using this material”. Descriptors are abstract in a general sense, and it is the descriptor states that conservationists actually record and utilise. Standard descriptors for ecogeographic, field and conservation management data are included in the Descriptors for CWR\textsuperscript{110}, while formal characterization and evaluation descriptors are associated with various standardized ‘Crop descriptor lists’ published by FAO, Bioversity, UPOV (see http://www.bioversityinternational.org/publications.html) – these may or may not be suitable for describing the crop's associated CWR. It is important to stress that standard lists of descriptors should be used when they are available. The use of well-defined, tested and rigorously implemented descriptor lists for scoring

\textsuperscript{110} http://www.cropwildrelatives.org/
descriptors considerably simplifies all operations concerned with data recording, such as updating and modifying data, information retrieval, exchange, data analysis and transformation. When data are recorded, they should be classified and interpreted with a pre-defined list of descriptors and descriptor states to consult. This clearly saves a considerable amount of time and effort associated with data entry. The use of lists ensures uniformity, while reducing errors and problems associated with text synonyms.

A.14.3. Examples and applied use

There are a growing number of National management plans for CWR conservation that have been completed in recent years and each involves significant data collation and analysis, and its application to practically conserve the priority CWR taxa. In terms of data management each step in creation and updating the National management plan for CWR conservation (see Figure 25) involves:

a. **CWR National Checklist** – The common first step in production of a National management plan for CWR conservation is to produce a national CWR checklist; this is normally a simple table of the Latin names of the CWR taxa present in the country, as follows for the national CWR checklist of Saudi Arabia:

<table>
<thead>
<tr>
<th>Genus</th>
<th>Species</th>
<th>Species Author</th>
<th>Subspecific Rank</th>
<th>Subspecific Author</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aegilops</td>
<td>kotschyi</td>
<td>Boiss.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aegilops</td>
<td>peregrina</td>
<td>(Hack.) Maire &amp; Weiller</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aegilops</td>
<td>vavilovii</td>
<td>(Zhuk.) Chennav.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aerva</td>
<td>javanica</td>
<td>(Burm.f.) Juss. ex Schult.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Aerva</td>
<td>lanata</td>
<td>(L.) Juss.</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agathophora</td>
<td>alopecuroides</td>
<td>(Moq.) Bunge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Agathophora</td>
<td>alopecuroides</td>
<td>(Moq.) Bunge</td>
<td>var. papillosa</td>
<td>(Maire) Boulos</td>
</tr>
</tbody>
</table>
Figure 10. Summary of data flow in CWR conservation
c. CWR National Inventory – The difference between the checklist and the inventory is that the in an inventory the checklist is annotated; in that it each taxon has a range of ancillary information associated with each CWR taxon. As a result the data structure is now more complex and usually involves a multiple file structure such as Figure 26.

![Figure 26. Entity relationship model for the CWR database](image)

Figure 26. Entity relationship model for the CWR database


d. CWR Gap analysis – The CWR checklist and inventory are primarily taxon based but the gap analysis based largely on data associated with individual accessions that represent those taxa. Normally significant resources will be invested in the collation of large herbarium specimen and gene bank accession data sets. There is no standard format for the database that contains this data, but Annex 5 contains an extended list of data descriptors\(^\text{112}\) that will include those used as a basis for gap analysis.

e. CWR conservation – The data associated with CWR management will vary depending on whether it is associated with in situ or ex situ conservation, but falls

---

\(^{111}\) Vincent (Pers. Comm.)

\(^{112}\) Castañeda Álvarez et al. (2011)
into three basic categories (ecogeographic, field population, conservation management and monitoring) as detailed with examples above.

f. Promotion of use – As stated throughout the book, CWR conservation should be directly linked to utilisation, so once the CWR diversity is conserved it needs to be characterized and evaluated so that the potential users have some basis on which to select the accessions they wish to utilise. The data associated with characterization and evaluation is, as noted in the previous section, often lacking and seldom available to the user community. However, within the context of the European Cooperative Programme for Plant Genetic Resources (ECPGR) several Central Crop Databases were established that hold accession passport data and, to varying degrees, characterization and primary evaluation data of the major crop related collections in Europe, these database are web-enabled. The next conceptual advance in making characterization and evaluation data easily available to the user community has been to develop an internet portal that facilitates access to the existing data. This is currently being developed as the Trait Information Portal (TIP), which is envisaged will provide a unique entry point for access trait-specific information to help direct their research and allow them to effectively exploit CWR diversity.

Box 61. Trait Information Portal
The TIP is planned to have a simple platform architecture accommodating input and output data types, including the following elements: (a) Use a document store database system; (b) Have an upload system with flexible template driven options for data being sent by providers; (c) Include and use the Generation Challenge Programme (GCP) data annotation and trait ontology curation tools developed by the Bioversity team; (d) Be searchable through ontology-driven views; (e) Include information on traits, locations, trial sites, georeference, geographical information; (f) Use web scraping (gather related information/data) to include external data sources, molecular data, bibliography, characterization and evaluation data, images, etc.; (g) Link with external information sources; and (h) Provide data analysis outputs. Additionally, the TIP will include three different entry points (trait information, CWR and LR inventories), allowing users to choose their entry/access point to the information they require, while maintaining the capacity to link or tap into existing online sources of information such as GENESYS, EURISCO and ECCDBs.

This concept has been planned to create a system that primarily serves the data provider so that it can efficiently serve the users. To make the most of this idea the rationale for the TIP framework conceptualization was to use existing developments and resources, focusing the development team’s efforts towards
using and further enhancing existing and evolving resources being developed in other communities of practice. The TIP is being developed in the context of the PGR Secure project (http://pgrsecure.org/) and is expected to be available as a beta test version in 2013.

**TIP platform architecture**

Source: Dias (2012)
A.14.4. List of references used to compile the text

A.14.5. Additional resources and materials


European Cooperative Programme for Plant Genetic Resources (ECPGR) Documentation and Information Network. Available at http://www.ecpgr.cgiar.org/networks/documentation_information.html


