Proposal for the development of the
‘PGR Duplicate Finder’, a software package to assist in the identification of putative duplicates in germplasm databases

1. Problem statement
In the selection of the Most Appropriate Accessions (MAAs), all Crop Working Groups (CWG) are confronted with the laborious search for putative duplicates. This activity has common elements for all crops, which can be formalized and automated. Software will be developed that will assist the ones responsible for the proposal of MAAs in the identification of probable duplicates. To allow for easy processing of the results this software will be implemented in an Excel environment with easy to use macros.

2. Justification and rationale
A critical step in the creation of AEGIS European Collections will be the identification of the MAAs. Given its labor-intensiveness, all attempts should be made to facilitate this process, allowing the CWG to concentrate on the choice of the MAAs rather than spend time on searching for duplicates.

3. Background
Ideas about the identification of duplicates in germplasm collections have been circulating and applied for a long time (cf. references in section 13). CGN, the main participant of this proposal, has already in its attempts to select MAAs for a number of crops, created some preliminary macros to support the activity and it was clear that a more targeted development of robust tools would be very useful. In the EUROGENEBANK proposal, the Leibniz Institute of Plant Genetics and Crop Plant Research (IPK), Gatersleben, Germany, would have devel-
oped similar tools (however with a much higher budget than the one requested for this project). The Julius Kühn Institute (JKI), Quedlinburg, Germany, also has some experience with the semi-automatic identification of probable duplicates in the framework of the ECPGR Avena database.

Most of the scientific papers about definition and identification of genebank duplicates are (co-)written by the participants in this project (see section 13).

4. Main objective and specific objectives

To develop easy to use software called ‘PGR Duplicate Finder’ for the preliminary identification of putative duplicates on the basis of a list of passport data in the EURISCO Multi-Crop Descriptor (MCPD) format.

5. Materials and methods

Based on a discussion of scientists involved in the identification of duplicates from CGN, IPK and JKI, algorithms will be defined and/or collected. On this basis CGN will develop a prototype of the Duplicate Finder that will be tested by CGN, IPK, JKI and possibly others, and refined based on the feedback of the testers.

The PGR Duplicate Finder will be based on Excel, and can be used by version Excel 2003 and later. It will be a set of macros written in Visual Basic for Applications (VBA). It will use publicly available algorithms and code for routines such as the matching on the basis of sound (SoundEx), and the application of the “keyword in context” (KWIC index) approach.

The data that the PGR Duplicate Finder will analyze has to be copied by the user in one of the spreadsheets using the EURISCO MCPD format. Not all columns have to be present, and additional columns will be allowed. The column headers should comply with the fieldnames in the MCPD.
The passport data will be compared between accessions using transparent rules: identical long accession names will point to probable duplicates, small differences in accession or other names to less likely duplication, accessions matching accession number and holding institute to other accessions’ donor and donor number will be probable duplicates, etc. Furthermore, information such as the geographical and institutional origin of the accession will also be taken into account.

6. **Expected outputs**

The product of this project will be the PGR Duplicate Finder, a set of spreadsheets, macros and menu options allowing the user to easily do a preliminary identification of duplicates in the data.

There are two menu options: one for batch identification of probable duplicate groups, the other of identifying probable duplicates of a selected accession

- When the option for batch processing is chosen, an additional column is created in front of the sheet, with numbers to identify each duplicate group. By sorting these numbers, the probable duplicates will be grouped together.

- When the option for duplicate identification is chosen, all probable duplicates for the currently selected accession are displayed in order of likelihood of duplication, i.e., the accessions that are most likely to be duplicate are displayed first, followed by those that are less likely, etc.

Using these options, the user is able to see the probable duplicates and select the MAA and/or add a pointer to the MAA of probable duplicates, etc.

7. **Benefits and impact**

It can be expected that nearly each ECPGR CWG confronted with the task to identify MAAs will use the software as a starting point in the search for probable duplicates. As such it can
be expected to save very much time and make completion of this enormous task more feasible. Since the identification of MAAs is an essential step in the AEGIS strategy, this will benefit the entire ECPGR community.

8. Innovation

No generally applicable software, nor formalized methodology, for the identification of probable duplicates on the basis of passport data exists. The knowledge generated in the development and application of the software/methodology will be the nucleus on which approaches and ideas can be added in the future.

The product will be freely available for use, and also freely available for further development under the GNU Public License.

9. Application of results

The software and manual developed in this project will be made available to all ECPGR CWG and the rest of the PGR community via the website of AEGIS. It can be expected that it will be applied in (nearly) all CWG, and possibly also by other members of the PGR community to identify probable duplicates within and between PGR collections.

10. Workplan
Remark: Possibly intermediate products will be released on request allowing a rapid deployment by CWGs needing it urgently.

11. Budget

<table>
<thead>
<tr>
<th></th>
<th>Project</th>
<th>In Kind</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>CGN</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Staff time</td>
<td>€ 9 730</td>
<td>€ 7 130</td>
<td>€ 16 860</td>
</tr>
<tr>
<td>Meetings</td>
<td>€ 250</td>
<td></td>
<td>€ 250</td>
</tr>
<tr>
<td>IPK</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Staff time</td>
<td>€ 1 000</td>
<td></td>
<td>€ 1 000</td>
</tr>
<tr>
<td>Meetings</td>
<td>€ 150</td>
<td></td>
<td>€ 150</td>
</tr>
<tr>
<td>Travel</td>
<td>€ 150</td>
<td></td>
<td>€ 150</td>
</tr>
<tr>
<td>JKI</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Staff time</td>
<td>€ 1 000</td>
<td></td>
<td>€ 1 000</td>
</tr>
<tr>
<td>Meetings</td>
<td>€ 150</td>
<td></td>
<td>€ 150</td>
</tr>
<tr>
<td>Travel</td>
<td>€ 150</td>
<td></td>
<td>€ 150</td>
</tr>
<tr>
<td>TOTAL</td>
<td>€ 10 330</td>
<td>€ 9 380</td>
<td>€ 19 710</td>
</tr>
</tbody>
</table>

12. Contributions offered by applicant

CGN will contribute a significant part (42%) of the staff time required to organize the meeting and develop and test the software, and will cover the costs associated with hosting the meeting.

IPK and JKI will cover the staff costs involved in attending the meeting and testing the software.

13. Bibliography


