Data quality requirements for agrobiodiversity: the case of crop wild relatives

Nora P. Castañeda-Álvarez, Colin K. Khoury, Steven Sotelo, Chrystian C. Sosa, Vivian Bernau, Harold A. Achicanoy

“Biodiversity Data Quality Symposium: developing a common framework to improve fitness for use of biodiversity data”, 8th March 2016, Sao Paulo, Brazil

Helianthus argophyllus (wild relative of sunflower). Credit: Kasia Stepien
CONTENT

• Introduction
  • Agrobiodiversity
  • Crop wild relatives

• Using occurrence data (GBIF mediated data) to assess the *ex situ* conservation of crop wild relatives
  • Species distribution models
  • Quantifying gaps in *ex situ* collections

• Improving the quality of data

• Moving forward: steps for continuous improvement
INTRODUCTION

• What is agrobiodiversity?
INTRODUCTION

Plant genetic resources: the biodiversity of crops
INTRODUCTION: Crop wild relatives

Crop wild relatives: the wild and weedy taxa genetically related to crops, including its ancestors.
INTRODUCTION: Crop wild relatives

INTRODUCTION: crop wild relatives are valuable

Grassy stunt virus resistance from *Oryza nivara*

Salinity tolerance from *Solanum cheesmaniae*

Aluminium tolerance from *Oryza rufipogon*

Resistance to black Sigatoka and Fusarium wilt from *Musa acuminata ssp. burmannica*
To be used, crop wild relatives need to be conserved
To be used, crop wild relatives need to be conserved
Assessing the conservation of crop wild relatives in genebanks
But, to what extent the genetic diversity of crop wild relatives is conserved in genebanks?
Ex situ conservation of crop wild relatives
Identification of “gaps” in genebank collections

1. Gather taxonomic data
2. Gather occurrence data
3. Georeferencing
   - Google Maps Geocoding API
4. Determine gaps in collections
5. Model distributions
6. Make collecting recommendations

**Taxonomic**

**Geographic**

**Ecological**
Identification of “gaps” in genebank collections

1. Gather taxonomic data
2. Gather occurrence data
3. Georeferencing
4. Determine gaps in collections
5. Model distributions
6. Make collecting recommendations

Tools and methods:
- GRIN Taxonomy for Plants
- The Plant List
- Google Maps Geocoding API
- Taxonomic, Geographic, Ecological
Identification of “gaps” in genebank collections

Gather taxonomic data → Gather occurrence data → Georeferencing

Google Maps Geocoding API → GEO Locate

Make collecting recommendations → Determine gaps in collections → Model distributions

Taxonomic
Geographic
Ecological
Identification of “gaps” in genebank collections

1. Gather taxonomic data
2. Gather occurrence data
3. Georeferencing
4. Determine gaps in collections
5. Model distributions
6. Make collecting recommendations

- **GRIN**: Gateway to Genetic Resources
- **THE PLANT LIST**: Taxonomic
- **GENESYS**: Geographic
- **ECOLOGICAL**: Ecological

**Google Maps Geocoding API**
Identification of “gaps” in genebank collections

1. Gather taxonomic data
2. Gather occurrence data
3. Georeferencing
4. Determine gaps in collections
5. Model distributions
6. Make collecting recommendations
Identification of “gaps” in genebank collections

1. Gather taxonomic data
2. Gather occurrence data
3. Georeferencing
4. Make collecting recommendations
5. Determine gaps in collections
6. Model distributions

- GRIN Taxonomy
- The Plant List
- Google Maps Geocoding API
- GEOLocate

Taxonomic
Geographic
Ecological
Identification of “gaps” in genebank collections

Gather taxonomic data → Gather occurrence data → Georeferencing

Make collecting recommendations

Determine gaps in collections

Model distributions

Georeferencing

Google Maps Geocoding API

Taxonomic

Geographic

Ecological
Priorities for conservation

- **High priority taxa for collection**: 71%
- **Mid priority taxa for collection**: 13%
- **Low priority taxa for collection**: 12%
- **No further collection is required**: 5%
Collecting hotspots

Global collecting hotspots for High Priority Taxa, for 76 crop gene pools
Collecting *Avena ventricosa* in Cyprus
Data quality constraints

- Gather taxonomic data
- Gather occurrence data
- Georeferencing
- Make collecting recommendations
- Determine gaps in collections
- Model distributions

Data sources and tools:
- GRIN Taxonomy for Plants
- The Plant List
- GENESYS Gateway to Genetic Resources
- GBIF
- Google Maps Geocoding API

Taxonomic, Geographic, and Ecological data quality constraints.
Data quality constraints

• Taxonomic identity
  • Herbarium specimen
    • Old taxonomical determinations = not aligned with recent taxonomies.
    • No taxonomical determinations at all.
  • Digitization
    • Scientific names misspelled.
    • Need to align with crop wild relatives checklists.

• Geographic localities
  • No geographic coordinates at all.
  • Occurrence records in the sea.
  • Coordinates coinciding with countries centroids (!).
  • Inconsistencies in the country where the record is mapped.
Preparing and improving data

• Taxonomic identity
  • Herbarium specimen
    • Old taxonomical determinations = not aligned with recent taxonomies.
    • No taxonomical determinations at all.
  • Digitization
    • Scientific names misspelled.
    • Need to align with crop wild relatives checklists.

• Geographic localities
  • No geographic coordinates at all.
  • Occurrence records in the sea.
  • Coordinates coinciding with countries centroids (!).
  • Inconsistencies in the country where the record is mapped.

Botanists!
Alignment with cwr checklists
Re-calculate and test coordinates
Preparing data

• Taxonomic identity
  • 89% of our DB is now aligned to GRIN Taxonomy.
  • 5% was aligned using the Taxonomic Name Resolution Service.
  • 5% is aligned to The Plant List (through TaxonStand).
  • <1% JSTOR Global Plants, IPNI, ITIS
Preparing data

• Geographic localities
Preparing data

• Geographic localities
Validation with experts
Obrigada!

This work was undertaken as part of the initiative "Adapting Agriculture to Climate Change: Collecting, Protecting and Preparing Crop Wild Relatives" which is supported by the Government of Norway. The project is managed by the Global Crop Diversity Trust with the Millennium Seed Bank of the Royal Botanic Gardens, Kew UK and implemented in partnership with national and international genebanks and plant breeding institutes around the world. For further information, go to the project website: http://www.cwrdiversity.org/