Final Report of the Task Group on GBIF Data Fitness for Use in Agrobiodiversity

Final version 1.0 published on 15 February 2016

Authors (in alphabetical order)

Elizabeth Arnaud, Bioversity International, France - Task Group Chair
Nora Patricia Castañeda-Álvarez, CIAT, Colombia and University of Birmingham, UK
Jean Ganglo Cossi, University of Abomey-Calavi, Benin
Dag Endresen, GBIF Norway, University of Oslo, Norway
Ebrahim Jahanshiri, Crops for the Future, Malaysia
Yves Vigouroux, Institut de Recherche pour le Développement (IRD), France

GBIF contact

Dmitry Schigel, GBIF Secretariat (dschigel@gbif.org) - Programme Officer for Content Analysis and Use

Table of Contents

Acknowledgements

Many thanks to Abdallah Bari, Ahimsa Campos Arceiz, Alberto Tanzi, Arthur Chapman, Asha Karunaratne, Aoudji Augustin, Aryo Feldman, Axel Diederichsen, Christian Leclerc, Christoph Germeier, Chrystian Camilo Sosa, Colin Khoury, Daniel Callo-Concha, Dro Daniel Tia, Evert Thomas, Fabrizio Celli, Gueye Mathieu, Hannes Gaisberger, Harold Achicanoy, Helmut Knüpffer, Holly Vincent, Igor Loskutov, Joana Magos Brehm, Jose Iriondo, Koffi Kouao Jean, Koura Kourouma, Lee Belbin, Maarten van Zonneveld, Mame Codou Gueye, Marc Deletre, Marcelo Simon, Matija Obreza, Marie-Angelique Laporte, Mauricio Parra Quijano, Nidhi Nagabhatla, Nigel Maxted, Ola Westengen, Peter Desmet, Priscila Ambrosio Moreira, Raymond Sognon Vodouhe, Razlin Azman, Reinhard Simon, Robin Goffaux, Ruth Bastow, Samy F. Gaiji, Sean Mayes, Severin Pohlreich, and Theo van Hintum for participation in the online survey; Sue Walker kindly helped to distribute the information about the survey. Thanks to Andrea Hahn, Mélianie Raymond, and Tim Robertson for comments on the first draft report. Many thanks to the GBIF secretariat and Bioversity International for setting up and hosting this task group.

Arthur Chapman, Lee Belbin, Joana Magos Brehm, Shelagh Kell, and Mauricio Parra Quijano provided valuable feedback to the first draft report. We wish to highlight acknowledgement to Joana Magos Brehm for providing particular detailed and constructive comments and suggestions.

Special thanks to Dmitry Schigel, Programme Officer for Content Analysis and Use, GBIF, for facilitating the task group work, the survey preparation and meetings.

Document history

Draft version 0.1 released on 2 October 2015
Final version 1.0 published on 15 February 2016

Top recommendations for GBIF Data Fitness for Use for Agrobiodiversity

- The Multi Crop Passport Data (MCPD) is the data exchange standard for describing crop samples held in gene banks. GBIF must index data attributes described with the MCPD terms to stimulate the use of gene bank data and other ABD data published in GBIF. Most of the MCPD terms were mapped to Darwin Core terms (see table 1 on p.14). Therefore, to enable full compatibility between these standards, only a few terms need to be added to the GBIF data profiles, following the model proposed in the existing Darwin Core germplasm extension. This will be achieved by including the Darwin Core germplasm extension into the GBIF data indexing routines. (Recommendation 6.2.1).
- A more formal agrobiodiversity (ABD) community governance policy is needed for the germplasm extension. The Biodiversity Information Standards (TDWG) could be a suitable platform for implementation of a formal agrobiodiversity community governance policy for the Darwin Core germplasm extension. Darwin Core germplasm extension should be maintained by a TDWG task group to reach a stable standard for germplasm accessions conserved in gene banks (*ex situ* conservation), and should be expanded to address needs of data on the *in situ* and on-farm conservation. (Recommendation 6.2.5).
- Authoritative checklists and classification of crop wild relatives, cultivars, landraces and neglected and underutilized crop species, including vernacular names from authoritative lists along with the language and countries where it applies, should be added to GBIF when developed and validated by an international expert group and community. (Recommendation 6.4.1).
- GBIF should seek to support the integration of popular data cleaning tools such as GEOLocate, OpenRefine (formerly Google Refine), and workflow services from BioVeL and other Galaxy or Taverna compliant protocols with data published to the GBIF portal. It is also important to take into account the requirements on use cases that are being developed by a task group of the TWDG/GBIF data quality interest group. (Recommendation 6.8.2).
- GBIF should improve routines for preliminary quality assessment of data records and datasets (aggregated records) giving levels of confidence to individual data record or datasets and highlight issues to data suppliers. **A level of confidence can only be applied within a specific context so a weighting of the scores (possibly 'weighted completeness' and 'weighted issues'*)* should be proposed in the context of use by ABD community.** (Recommendation 6.11.1).
- GBIF should develop or adapt existing tools to: (a) identify quality improvement thresholds based on the decided weighting of scores such as unreliable coordinates; identify issues with taxon names such as completeness of name-strings and up-to-date nomenclature and whether names are backed by publication reference, sequence, or expert; (b) check the completeness of the data (e.g. index of passport data completeness) through possibly two scores: 'weighted completeness' and 'weighted issues'; (c) provide the percentage of records with actual data reported for each attribute (data column), possibly with

two scores: ‘weighted completeness’ and ‘weighted issues’. (Part of Recommendation 6.11.2).

- Expand the data attributes made available for search from the GBIF portal. Include the most important agrobiodiversity terms from the MCPD and the corresponding Darwin Core germplasm extension as searchable information attributes (such as gene pool and taxon group concepts, trait information, characterization and evaluation data, pre-breeding and breeding information) (Recommendation 6.12.3).
- Resources for the ABD community like the global crop wild relative species checklist (http://www.cwrdiversity.org/checklist/) and the Bioversity Collecting Mission database must be published to the GBIF portal registry of checklists and integrate this checklist to the GBIF taxon backbone. To complement this global list, other crop wild relatives (CWR) checklists can be proposed for publishing as a taxon checklist in GBIF and first the crop wild relative species list developed by the Southern African Development Community (SADC)-CWR project which includes a global list of crop genus names that is a useful tool for national species list of crop wild relatives. (Recommendations 6.3.1 and 6.6.1)
- Stimulate the digitization of relevant collections (i.e. herbaria, gene banks, published articles, MSc and PhD theses, national and regional projects) related to ABD, and stimulate the publishing of already digitized collections, by providing small--grants through competitive calls (Recommendation 6.6.2).
- Train the GBIF Nodes on the value of CWRs, and mobilization of data on crop wild relatives and on species traits useful for crop improvement and for landscape restoration (Recommendation 6.3.4).

1. Abstract

Human wellbeing and food security in a changing climate depend on productive and sustainable agriculture. For this, policies based on analyses and research results are vital to establish conservation priorities of natural resources that underpin the enhancement of sustainable food production. Therefore, data from agrobiodiversity and wider biodiversity sources are required to be available and accessible. Currently, there is a risk that agrobiodiversity and the wider biodiversity data communities remain separated with inefficient data aggregation, unless data flow pathways are harmonized. GBIF has a role to play in contributing to the convergence of the two communities. Biodiversity data in particular on wild relatives of the cultivated species will flow easier into agrobiodiversity conservation priority assessments and analysis with agrobiodiversity data integrated in GBIF.

The Task Group on Data Fitness for Use in Agrobiodiversity was established by the GBIF Secretariat and Bioversity International to help improve the fit of data related to agrobiodiversity to the variety of important uses required and requested by the community of research and policy. The task group has been looking at the key actions for creating interoperability of data on *ex situ*, *in situ* and on-farm conservation of agrobiodiversity, with a focus on plants. A survey and interviews of selected experts and ABD data practitioners were conducted to collect feedback on fitness for use and issues with GBIF-mediated data.

The 53 recommendations of the task group cover the whole data flow, from publishing to data use with a focus on agrobiodiversity, also considering the role of nodes in data mobilization and in promotion and training. Some key recommendations are to (i) promote GBIF to the agrobiodiversity community, (ii) integrate the terms from the long-standing Multi Crop Passport Data standard (MCPD) already used for several decades by agricultural gene banks into Darwin Core indexed attributes, (iii) by installing proper governance, the Darwin Core germplasm extension can be maintained as a stable international standard, (iv) develop agrobiodiversity user profiles on GBIF data portal to improve the user experience in accessing data of interest, (v) add infraspecific taxonomy levels to ensure adequate publication of agrobiodiversity data, by means of integrating into the GBIF taxonomic backbone the reference taxonomies used by the community with additional attributes related to the crop wild relative species, landraces and cultivars, (vi) publish existing digitized ABD data collections, such as the Bioversity Collecting Mission database[1] and the Crop Wild Relative Global Occurrence dataset[2], to support capacity building of agrobiodiversity data publishers, (vii) provide quality filtering of the data only using attributes of interest to the agrobiodiversity data users. Additionally, GBIF needs to provide tools and services to discover, mobilize, or link to additional specialized data sources commonly used by the agrobiodiversity community. Integrated access from GBIF to external sources of key agrobiodiversity data would be an added value for the community. (viii) Assign a level of confidence to individual data records, and (ix) channel feedback to data suppliers.

[1] http://bioversity.github.io/geosite/

[2] http://www.cwrdiversity.org/checklist/cwr-occurrences.php

The task group identified increasing the knowledge of the nodes about agrobiodiversity data through training as a key step to enable them to play a more prominent role in the mobilization of locally available information resources on ABD.

A priority setting of these recommendations, with the feedback of the ABD community, the GBIF country parties and the expert knowledge of the GBIF secretariat and nodes, is needed.

2. Scope: what is agrobiodiversity and why it matters?

"Agricultural biodiversity [agrobiodiversity or ABD in this report] is the diversity of crops and their wild relatives, trees, animals, microbes and other species that contribute to agricultural production. This diversity exists at the ecosystem, species, and genetic levels and is the result of interactions among people, biodiversity components, and the environment over thousands of years. The use of agricultural biodiversity can help make agricultural ecosystems more resilient and productive; and can contribute to better nutrition, productivity and livelihoods" (Bioversity International)[3].

Note of the task group: Given that agrobiodiversity covers a large area of research, and that the focus of this task group is 'crop diversity', it is worth acknowledging that a group of livestock experts should be convened to extend the recommendations related to animal diversity.

Agrobiodiversity contributes to farmers' resilience to climatic events and plant pathologies, and provides options for adaptive strategies to environmental and economic changes; it supports the restoration of ecosystem services and provides a genetic reservoir of new traits and species for farming. An estimated total of 35,000 plant species are cultivated by humans for use in gardening, landscaping and agriculture. An estimated total of 7,000 from these plant species are cultivated for use in agriculture (Khoshbakht and Hammer 2008). About 1,000 species of cultivated plants are threatened globally (Khoshbakht and Hammer 2007). Addressing the loss of species and genetic resources is critical for improving crops, coping with pests and diseases, soil health, global freshwater, and pollinators, and adaptation to climate change. Pollinators contribute to the production of over 80% of crops traded on the world market and up to 10–16% of global yearly harvests are lost to plant diseases.

The discovery, access and adequate use of primary biodiversity data is critical to inform decision making to achieve sustainable use of agrobiodiversity resources, to secure their availability in the future, and to address many of the world's key challenges such as feeding a growing human population, and developing more productive and sustainable agriculture under climate change. It is estimated that various agrobiodiversity data portals and institutions (Genesys[4], EURISCO[5], GRIN[6], CIAT, FAO, national and regional gene banks)

[3] http://www.bioversityinternational.org/why-agricultural-biodiversity-matters-foundation-of-agriculture/

[4] https://www.genesys-pgr.org/welcome

[5] http://eurisco.ipk-gatersleben.de/

[6] http://www.ars-grin.gov/npgs/searchgrin.html

collectively house some 7.4 million specimens (FAO, 2010)[7] of cultivated species and cultivated varieties, genetic samples and other important evidence of patterns and trends in global biodiversity. High numbers of species and specimens of crop wild relatives need to become digitally available alongside the data on cultivated species. Only a fraction of this vast databank of species information and genetic material is freely and digitally available.

Cultivated crop species for food and agriculture are generally conserved *ex situ* in gene bank collections. Traditional cultivars or landraces can also be conserved on-farm, in active farming. Species of crop wild relatives (CWRs) are generally conserved *in situ* in their natural habitat. Currently only a few prioritized and important populations of CWRs have been collected and conserved *ex situ* in botanical gardens and gene bank collections.

Definition of Crop Wild Relative species (CWR)

"those wild plant taxa more or less closely related to species of direct socio-economic importance including food, fodder and forage crops, medicinal plants, condiments, ornamental and forestry species, as well as those related to crops used for industrial purposes such as oils and fibres" (Maxted *et al*. 2006). Crop wild relatives are used as a source of genes for plant improvement.

Definition of Landrace

“Landraces have a certain genetic integrity. They are recognizable morphologically; farmers have names for them and different landraces are understood to differ in adaptation to soil type, time of seeding, date of maturity, height, nutritive value, use and other properties. Most important, they are genetically diverse” (Harlan 1975).

"A landrace is a dynamic population(s) of a cultivated plant that has a historical origin, distinct identity and lacks formal crop improvement, as well as often being genetically diverse, locally adapted and associated with traditional farming systems" (Villa *et al*. 2005).

Neglected and Underutilized Species (NUS)

Also called ‘orphan crops’, NUS are plant species and varieties of importance for the rural communities but to which little or no attention is paid by agricultural researchers, plant breeders and policymakers. NUS are not widely traded (Padulosi *et al*. 2013) and are represented by wild, semi-domesticated or local varieties and many non-timber forest species, adapted to local and often marginal areas. According to the combined gene pool concept (Harlan and de Wet 1971) and Taxon Group categorization (Maxted *et al.* 2006), many NUS are also classified as CWR.

[7] http://www.fao.org/docrep/013/i1500e/i1500e00.htm

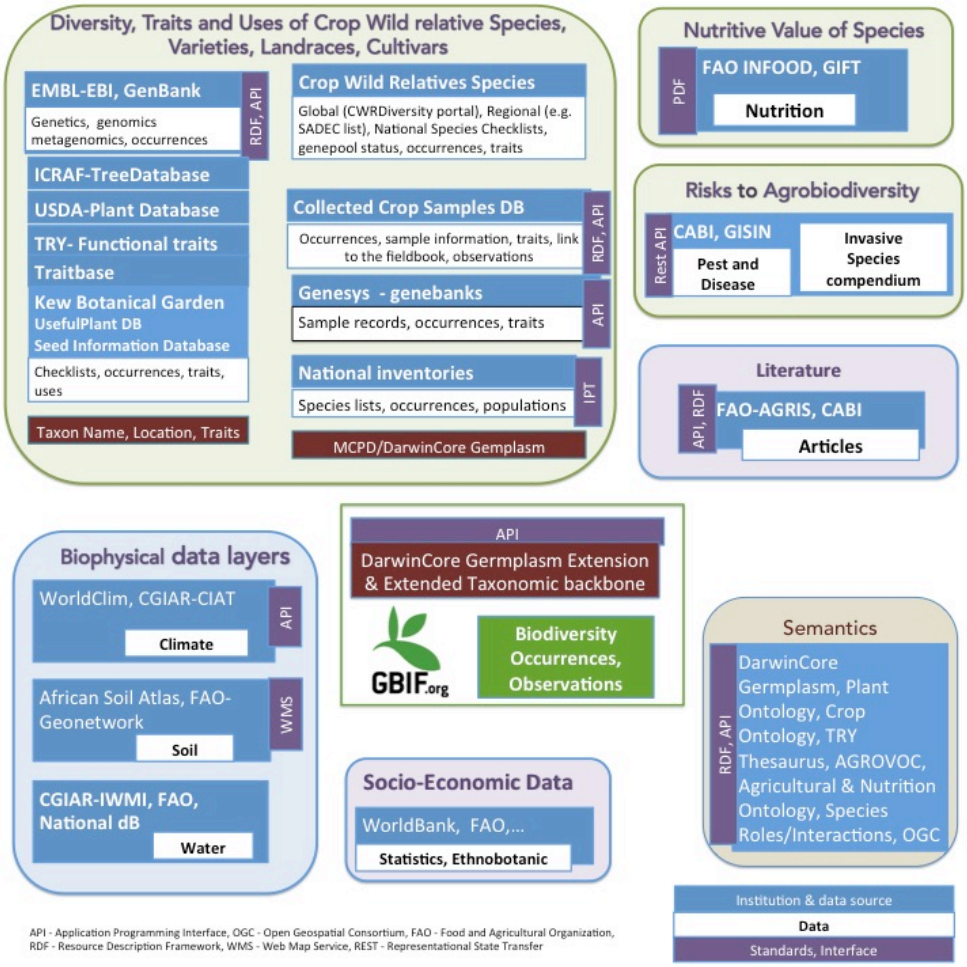
3. Rationale: what can GBIF do for agrobiodiversity data users?

GBIF.org as a global biodiversity data platform can play an important role in the agrobiodiversity landscape by mobilizing and connecting biodiversity datasets that can support research and development for food security and ecosystem services resilience. As part of a broader global strategy on fitness for use of biodiversity data, GBIF and Bioversity International convened a Task Group on Data Fitness for Use in Agrobiodiversity in March 2015. The Task Group identified the need to bridge ecological and agricultural data that are relevant for agrobiodiversity and agroecology uses. In general, the plant *ex situ* conservation data are in a good state with developed data and metadata solutions.

However, data on *in situ* and on-farm conservation, use and management in agrobiodiversity are not yet fully available and often unstructured. Although occurrence data from many of the gene bank collections, including landraces and collected CWR resources (conserved *ex situ*), are already published and made available in the GBIF portal, the agrobiodiversity community does not use GBIF much. The reasons are probably because (a) a number of well-established ABD data portals exist (such as Genesys, EURISCO and a number of crop-specific databases, each serving different subsets of the ABD data) and (b) because research on plant genetic resources diversity generates data at the intraspecies level, mainly unstructured and requiring specific attributes to describe data sets that are not yet available through GBIF.

However, GBIF has great additional potential for the ABD community, providing integrated access from one single portal to ABD-related data, including crop wild relatives (which are generally under-represented in *ex situ* gene bank collections and *in situ* monitoring data in general severely under-represented in the ABD data portals), data for landscape restoration, crop improvement, eco-geographic land characterization and other uses. Such information could be linked to other information outside GBIF such as extinction risk, genotype-level trait data and restoration and molecular genetic data (see figure 1). Agrobiodiversity data users often access these different data from different platforms.

Figure 1: GBIF within the landscape of agrobiodiversity relevant data sources (not exhaustive list))



4. Objectives of the task group

The task group aims to capture the best available experiences, document limitations in existing GBIF services, and suggest improvements in the functionality of GBIF.org for domain-specific needs.

- To make recommendations on improving data availability and use, mobilization, publishing and processing of data / metadata. Also to deliver a vision of the ideal data, data modifications, cleaning steps, analyses and visualization needs of the agrobiodiversity community.
- To document best practices using agrobiodiversity-related data, and to collect the information on repeatable tools and data management solutions.
- To make recommendations on GBIF.org improvements, and to provide guidance in the development of training and outreach materials for data users, to allow the better interconnection of different platforms and to allow different datasets to be combined.

5. Mode of operation of the task group and outputs

A survey of selected agrobiodiversity experts was launched to capture knowledge, experiences and opinions relating to data used in agrobiodiversity research for food security and agroecosystem resilience (see full survey in Appendix I). Fifty-one respondents answered the survey, which ended on 10 September 2015. To complement the survey, results and the ideas and demands captured in the initial phase of its work, the task group subsequently conducted in-depth interviews with experts about data flows and their practices in consulting sources of data. The survey mostly captures the feedback of experts interested in species distribution modelling and genetic diversity analyses of ABD, reflecting the dominant current uses of data accessed through GBIF. The draft report was published online and submitted to community feedback until 15 December through the GBIF Community site and through e-mails. Additional comments received at the end of December were also integrated.

The main deliverable of the task group is a set of practical guidelines and recommendations from the agrobiodiversity community around the issues defined under the Terms of Reference (see Appendix 3), summarized into a short, action-oriented report. An interim summary of demands and an early analysis are presented below. Recommendations will not only be directed at the GBIF Secretariat, but also for GBIF Participant Nodes to target data mobilization activities informed by gap analysis from the task group.

Four Skype meetings of the task group members were held on 28 April, 4 June, 21 July and 29 September 2015. Two face-to-face meetings were held, the first at Bioversity in Montpellier from 10 to 11 July and the second at the GBIF Secretariat in Copenhagen from 10 to 11 September. The meeting notes, draft version of the report and the survey results were shared among task group members through an online shared folder.

Feedback from survey respondents and ABD community stakeholders following the initial release of the draft report, were incorporated into this final version 1.0 of the task group report.

6. Recommendations

Key recommendations were derived from our analysis of the information received through the survey, from interviews and based on our own experience.

6.1 General use of GBIF by the agrobiodiversity community

A survey was sent to agrobiodiversity experts selected based on their previous experience with GBIF and the results provide a good snapshot of their user experience. Respondents are satisfied by the availability of big datasets and the possibility of downloading large data sets (most survey respondents have downloaded data from GBIF.org) but are generally less satisfied with the quality of the coordinates, outdated taxonomic names, and presence of duplicates. 60% of the respondents report problems in accessing the data they need.

However, most of the mentioned issues are related to the lack of access to different types of external auxiliary data, among which the external environmental and external trait data are

the most frequently mentioned. It is worth noting that very few of the respondents have contacted the GBIF helpdesks at the Secretariat and the national nodes to get support. Some nodes provide substantial national helpdesk functions (e.g. France, Spain, Norway), while some other nodes could increase their availability to provide helpdesk services. National pages on GBIF.org should provide the email to the national helpdesk, and/or national mailing lists if nothing else is available at the national level.

There is a clear potential for expanding the use of GBIF in the agrobiodiversity community for publishing and accessing data. The task group recommends that the GBIF Secretariat and GBIF nodes stimulate data mobilization in the agrobiodiversity community by showing the academic and non-academic benefits to potential publishers, with a particular focus on an audience ranging from pure scientists to pure data managers, and on young scientists. Data mobilization approaches include support for crowdsourcing of observation data generation, and making data available for applications on handheld mobile devices. Citizen scientists should be invited to browse through their national records flagged with data quality issues. GBIF and the task group must inform and promote the importance of agrobiodiversity data to partners (IUCN, EOL, BHL, and others). Researchers, students and teachers need encouragement and clear guidelines on how to publish their existing data. The GBIF Secretariat should encourage and support the organization of workshops at international, regional and at national levels through GBIF nodes to build capacity on data types, data mobilization and publication by students, researchers, teachers and node staff members.

Recommendations

6.1.1. A promotional campaign showing that GBIF is a useful resource for agrobiodiversity research (abbreviated ABD) will be required to explain to potential data publishers the academic and non-academic benefits of data sharing through the GBIF portal.

6.1.2. A series of training workshops at international, regional and national levels targeting the ABD research community will be necessary to explain the features of the GBIF portal, including the upload and download of data to/from GBIF and feedback to GBIF. Training materials on the publication and use of biodiversity data should be provided to the agrobiodiversity community. This should be done in conjunction with the training sessions focusing on ABD data, such as best practices and standardized methodology to collect relevant data on crop wild relatives.

6.1.3. Provide training materials and best practices targeting agrobiodiversity users.

6.1.4. The visibility of helpdesk point of contact for the GBIF Nodes needs to be improved. GBIF.org national pages should provide the email to the national helpdesk, and/or national mailing lists when this is available at the national level.

6.1.5. As a result of the above actions, the ABD community should publish existing digitized ABD data collections through GBIF.

6.2 Integration of relevant data standards for agrobiodiversity

The community of *ex situ* gene banks, which predates GBIF, has agreed on a set of core Multi-Crop Passport Descriptors (MCPD). The first version was introduced in 1998 (Hazekamp *et al*. 1998) with the first official version published in 2001 (Alercia *et al*. 2001). An updated version was published in June 2012 (Alercia *et al*. 2012), and the current version including terms for specimen-level persistent identifiers was published in December 2015 (Alercia *et al*. 2015). The MCPD is a long-standing community specific standard. Darwin Core and the germplasm extension dominate in citation by gene bank managers responding to the survey because it integrates the additional data fields that survey respondents suggest should be directly available at GBIF.org. The MCPD covers the essential core terms for an agrobiodiversity specimen data type level backbone aligned one-to-one with the Darwin Core occurrence data type level. Further refinement of the germplasm terminology and the Darwin Core extension is still needed to improve the support for other agrobiodiversity data types such as the crop wild relatives (Thormann *et al*. 2013), pre-breeding and breeding data, and characterization & evaluation trait data.

Most of the MCPD terms already have corresponding terms in the Darwin Core standard[8] (Wieczorek *et al*. 2012, see table 1). Unless all of the current 41 MCPD descriptors are included in and made available in data downloads from the GBIF index, the agrobiodiversity community will need to continue maintaining parallel occurrence-level data flow pathways and independent data indexing solutions. A remedy to this situation will be the addition of the few currently lacking MCPD terms into the Darwin Core set of terms – following the description in the Darwin Core germplasm extension[9] (Endresen and Knüpffer 2012) – that are indexed by GBIF procedures for the occurrence-level backbone. Further work will be required to integrate descriptors for *in situ* conservation of crop wild relatives to the Darwin Core extension (Thormann *et al*. 2013). There is no alternative to full integration.

The integration priorities are as follows:

1. The most important term to add is SAMPSTAT (biological status of sample, g:biologicalStatus) describing the type of germplasm material.
2. Another prioritized set of terms include DONORCODE (g:donorInstituteID), DONORNAME (g:donorInstitute), DONORNUMB (g:donorsIdentifier), ACQDATE (g:acqusitionDate), and COLLSRC (g:acquisitionSource). Germplasm material is living material allowing for living copies to be passed on from one gene bank collection to another. The set of terms for donor institute and the germplasm identifier used by the donor is important to enable tracking regarding the provenance of germplasm. Darwin Core germplasm extension also promotes a persistent identifier (g:donorsID) for the germplasm material held by the donor (this term is also proposed for a future MCPD revision).
3. A similar set of prioritized terms include BREDCODE (g:breederInstituteID), BREDNAME (g:breedingInstitute), ACCENAME (g:breedingIdentifier), and ANCEST (g:ancestralData, g:purdyPedigree). The Darwin Core germplasm extension promotes a persistent identifier (g:breedingID) for this type of source material. Germplasm material can be created by a plant breeder through an

[8] http://rs.tdwg.org/dwc/terms/

[9] http://purl.org/germplasm/germplasmTerm#

active and experiment-based crop improvement and research activity. These terms describe the creation of germplasm material in the situations when this material is created through breeding and not collected *in situ* (or on-farm). Terms for collecting events are already very well covered in Darwin Core.

4. One minor issue here is the recommendation of the MCPD to use the degree-minute-second format for geographic coordinates while Darwin Core prescribes the decimal-degree format.
5. A second issue for the collecting event is the need for a term to describe the FAO WIEWS institute code for the collector (COLLCODE, g:collectingInstituteID).
6. Darwin Core should include information on whether a particular population is cultivated, wild, escaped from cultivation, sub-spontaneous or unknown (Note: dwc:establishmentMeans partly covers cultivated, wild, naturalized etc, while MCPD:SAMPSTAT provides richer information) and the 'basisOfRecord' term should enable to distinguish herbarium specimens from gene bank accessions because "preserved specimens" is too ambiguous. The 'dwc:basisOfRecord' has potential for improvement and this is also under discussion at TDWG.
7. Include in Darwin Core several administrative fields for the description of the site of observation or collection, rather than a two-field called "COLLSITE" and "ORIGCTY" (like in the current MCPD) or three-fields "Country", "County" and "Locality" (like in GBIF format). Such inclusion should also take place soon in the MCPD format. The suggestion came from the quality of georeferencing quality assessment tool called GEOQUAL tool[10].

Table 1: Mapping of MCPD (Alercia et al*. 2001, 2012, 2015; Hazekamp* et al*. 1998) to Darwin Core (Wieczorek* et al*. 2012) using the Darwin Core germplasm extension (Endresen and Knüpffer 2012). 25 terms in MCPD match a corresponding term in Darwin Core. 15 terms from MCPD are not matching terms already described in Darwin Core (highlighted in blue) and 2 terms partly matching (highlighted in grey). [Namespaces, dwc = http://rs.tdwg.org/dwc/terms/; g = http://purl.org/germplasm/germplasmTerm#]*

Term	MCPD (2015)	Darwin Core (dwc), germplasm (g)
NA	(not applicable)	dwc:datasetID
0	PUID	dwc:occurrenceID
1	INSTCODE	dwc:institutionCode
2	ACCENUMB	dwc:catalogNumber
3	COLLNUMB	dwc:recordNumber
4	COLLCODE	g:collectingInstituteID

[10] http://www.capfitogen.net/en/tools/geoqual/

4.1	COLLNAME	dwc:recordedBy
4.1.1	COLLINSTADDRESS	(dwc:recordedBy)
4.2	COLLMISSID	dwc:collectionCode
5	GENUS	dwc:genus
6	SPECIES	dwc:specificEpithet
7	SPAUTHOR	dwc:scientificNameAuthorship (if SUBTAXA is empty)
8	SUBTAXA	dwc:infraspecificEpithet
9	SUBTAUTHOR	dwc:scientificNameAuthorship (if SUBTAXA is not empty)
10	CROPNAME	dwc:vernacularName
11	ACCENAME	g:breedingIdentifier
12	ACQDATE	g:acquisitionDate
13	ORIGCTY	dwc:countryCode
14	COLLSITE	dwc:locality
15.1	DECLATITUDE	dwc:decimalLatitude
15.2	LATITUDE	dwc:verbatimLatitude
15.3	DECLONGITUDE	dwc:decimalLongitude
15.4	LONGITUDE	dwc:verbatimLongitude
15.5	COORDUNCERT	dwc:coordinateUncertaintyInMeters
15.6	COORDDATUM	dwc:geodeticDatum
15.7	GEOREFMETH	dwc:georeferenceSources
16	ELEVATION	dwc:minimumElevationInMeters

17	COLLDATE	dwc:eventDate
18	BREDCODE	g:breedingInstituteID
18.1	BREDNAME	g:breedingInstitute
19	SAMPSTAT	g:biologicalStatus
20	ANCEST	g:ancestralData, g:purdyPedigree
21	COLLSRC	g:acquisitionSource
22	DONORCODE	g:donorInstituteID
22.1	DONORNAME	g:donorInstitute
23	DONORNUMB	g:donorsIdentifier
24	OTHERNUMB	dwc:otherCatalogNumbers
25	DUPLSITE	g:safetyDuplicationInstituteID
25.1	DUPLINSTNAME	g:safetyDuplicationInstitute
26	STORAGE	g:storageCondition
27	MLSSTAT	g:mlsStatus
28	REMARKS	(dwc:occurrenceRemarks)

Genesys, the global catalogue of plant germplasm gene bank accessions[11], that uses the MCPD to aggregate gene bank data appeared in many of the responses gathered through the survey. The overall suggestion is to increase the quality of the existing records, reduce duplication, and collaborate with Genesys. Genesys could be invited to form an agrobiodiversity thematic data node in GBIF and to provide an agrobiodiversity data mobilization helpdesk.

The survey revealed some concerns among the respondents regarding modifications and changes applied to the Darwin Core standard. The Darwin Core standard is ratified by the Biodiversity Information Standards (TDWG) and all modifications require ratification by the TDWG community as is described by the Darwin Core namespace policy[12]. The Darwin

[11] https://www.genesys-pgr.org

[12] http://rs.tdwg.org/dwc/terms/namespace/

Core standard was first ratified by TDWG in 2009 and the Darwin Core decision history[13] list all the approved and implemented modifications, while the normative Darwin Core complete historical record[14] lists all the terms including all historical declarations. Some of the concerns raised by respondents to the survey might relate to the previous versions of Darwin Core that existed prior to the TDWG ratification in 2009. A mapping between the current Darwin Core standard and these older obsolete versions of Darwin Core is presented by TDWG[15]. However, most users of Darwin Core will most likely find the quick reference to the current valid Darwin Core terms[16] to be the most useful presentation.

The Darwin Core germplasm extension follow the similar design principles as is set by Darwin Core and the overall design guidelines set by the TDWG Vocabulary Management Task Group (TDWG 2013). The terms are declared as RDF (resources description framework) using the SKOS (simple knowledge organization system)[17] language and organized into class[18] or property[19] terms following current W3C (World Wide Web Consortium) recommendations. However, the terms have deliberately very limited semantic declarations. Some of the survey respondents have expressed the requirement of a more formal semantic description for the germplasm terms. The Darwin Core germplasm extension includes a type vocabulary for controlled element values. Recent updates to the Darwin Core standard have transferred the corresponding Darwin Core type terms into the main Darwin Core namespace. A similar approach could be implemented for the germplasm extension. A more formal agrobiodiversity community governance policy is needed for the germplasm extension. The Biodiversity Information Standards (TDWG) could be a suitable platform for implementation of a formal agrobiodiversity community governance policy for the Darwin Core germplasm extension. The Darwin Core germplasm extension is available for collaborative management at the TDWG Term Wiki[20], but is not prepared and submitted for formal community review as a TDWG standard.

Recommendations

6.2.1. The Multi Crop Passport Data (MCPD) is the data exchange standard for describing crop samples held in gene banks. GBIF has to index data attributes described with the MCPD terms to stimulate the use of gene bank data and other ABD data published in GBIF. Most of the MCPD terms were mapped to Darwin Core terms (see table 1). Therefore, to enable full compatibility between these standards, only a few terms need to be added to the GBIF data profiles, following the model proposed in the existing Darwin Core germplasm extension. This will be achieved by including the Darwin Core germplasm extension into the GBIF data indexing routines.

6.2.2. Collaboration with Genesys, the global portal for information on plant genetic resources, is necessary and the task group recommends studying the feasibility of Genesys

[13] http://rs.tdwg.org/dwc/terms/history/decisions/
[14] http://rs.tdwg.org/dwc/terms/history/
[15] http://rs.tdwg.org/dwc/terms/history/versions/
[16] http://rs.tdwg.org/dwc/terms/
[17] https://www.w3.org/TR/skos-reference/
[18] http://www.w3.org/2000/01/rdf-schema#Class
[19] http://www.w3.org/1999/02/22-rdf-syntax-ns#Property
[20] http://terms.tdwg.org/wiki/Germplasm

becoming a thematic data node within GBIF, providing a helpdesk for agrobiodiversity data mobilization.

6.2.3. Further refinement of the germplasm terminology and the Darwin Core extension with additional attributes (terminology) is needed for describing agrobiodiversity species, such as crop wild relatives and pre-breeding and breeding data, and characterization & evaluation of trait data. It should be added as an extension to the GBIF taxon core data profile and be included in the corresponding GBIF indexing routines. The gene pool and taxon group classifications along with traits have the highest priority as extension attributes.

6.2.4. The possibility of accommodating various standards and descriptors used in data sources was mentioned alongside the addition of data generated by predictive characterization using geospatial information. Population data should link to pre-breeding and breeding data.

6.2.5. Indigenous names are needed in addition to vernacular names.

6.2.6. A more formal agrobiodiversity community governance policy is needed for the germplasm extension. The Biodiversity Information Standards (TDWG) could be a suitable platform for the implementation of a formal agrobiodiversity community governance policy for the Darwin Core germplasm extension. Darwin Core germplasm extension should be maintained to reach a stable standard for germplasm accessions conserved in gene banks (*ex situ* conservation), and should be expanded to address the needs of data concerning *in situ* and on-farm conservation.

6.2.7. Technically, in Darwin Core germplasm extension, a clear distinction between classes and properties is required. Darwin Core germplasm extension needs to be revised to align with the very last version of Darwin Core. Using a controlled vocabulary for the value of an element should be considered. Ideally, Darwin Core germplasm extension should be compliant with the DCMI model proposed by Dublin Core[21].

6.2.8. Expand the Darwin Core germplasm extension with standard terminology to describe *in situ* conservation of crop wild relatives (to be based on Thormann *et al*. 2013).

6.2.9. Include in Darwin Core several administrative fields for the description of the site of observation or collection, rather than a two-field called "COLLSITE" and "ORIGCTY" (current MCPD) or three-fields "Country", "County" and "Locality" (GBIF format). Such inclusion should also take place soon in the MCPD format.

6.2.10. GBIF should consult livestock experts to adapt the Darwin Core germplasm extension to livestock in order to better cover agrobiodiversity research.

[21] http://dublincore.org/documents/interoperability-levels/

6.3 Inventories of crop wild relatives

GBIF needs to integrate agrobiodiversity terms and attributes for Crop Wild Relatives (CWR), for *in situ* (Moore *et al*. 2008, Thormann *et al*. 2013) and on farm conservation (FAO 2015).

1. Global, regional and national taxon checklists to identify crop wild relatives should be developed, agreed upon and published via GBIF. A global list of crop genera would be an important tool here.
2. Indigenous names are needed in addition to vernacular names to support the identification of the diversity maintained by local communities.
3. At the taxon-backbone-level a new attribute to identify the priority and conservation status for CWR species would be very helpful.
4. Attributes describing the gene pool category status also need to be added at the taxon/checklist level.
5. Information on interactions between agricultural crop and pest species, at the taxon level, is needed to correlate the respective occurrence data available in the GBIF portal (e.g. using the Darwin Core dwc:ResourceRelationship[22]).

In situ conservation and on-farm management information systems need to combine basic eco-geographic information (climate variables, water availability, soil type, vegetation type, land cover, latitude, longitude, altitude, spatial distribution of pests and diseases, etc). This information is critical to allow users of the information system to locate traits of interest (e.g. drought, disease or salinity tolerance) and also to identify sites with similar conditions where the varieties or landraces could perform well (Thormann *et al*. 2014, 2015).

Classification and identification of crop wild relatives

Harlan and de Wet (1971) propose a classification of crop wild relatives according to the relative crossability between wild and cultivated species as follows:

Gene pools		
GP1A	Primary	Cultivated forms of the crop (cultivars and landraces)
GP1B	Primary	Wild or weedy forms of the crop
GP2	Secondary	Species with which gene transfer is possible but difficult
GP3	Tertiary	Species with which gene transfer is impossible by genetic engineering [1]

[1] With the advance of molecular engineering techniques allowing for complex genetic transfers, Hammer et al. *(2003) suggested adding a fourth gene pool that includes genetic components of artificial origin (transgenes).*

[22] http://rs.tdwg.org/dwc/terms/ResourceRelationship

The main assumption behind the proposed definition is that **taxonomic distance is positively related to genetic distance** (Maxted *et al.* 2006) and thus provides a pragmatic approximation of potential crossability between taxa (when genetic information lacks). Taxon groups are defined as follows:

Taxon groups	
TG1A	Crop (cultivars and landraces)
TG1B	Same species
TG2	Same series or section
TG3	Same subgenus
TG4	Same genus
TG5	Same tribe but different genus

As a rule of thumb, Maxted *et al.* (2006) also suggested the following ranking methodology for establishing conservation priorities:

Degree of CWR relatedness	Gene pool	Taxon group	Conservation priority
Close CWR	GP1B	TG1B, TG2	High priority
Remote CWR	GP2	TG3, TG4	Low priority
Not CWR	GP3	TG5	Excluded

Although it remains difficult to differentiate between “close” and “remote” CWR as often the taxonomy is not given in full and the “series”, “section” and “subgenus” are not informed in most databases, the combined GP/TG definition is convenient and can easily be implemented in an automated request (Delêtre *et al*. 2012).

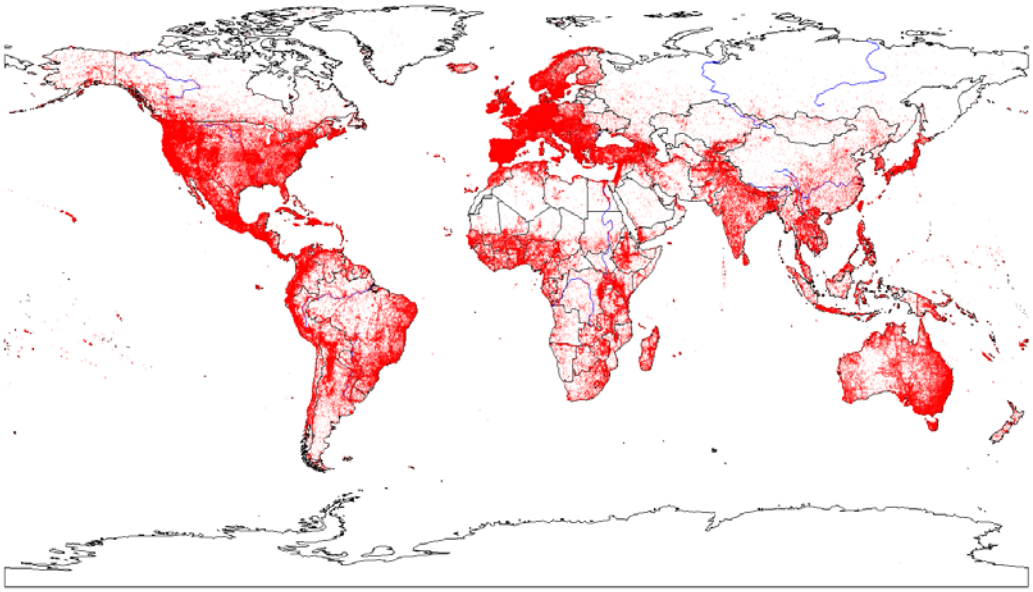
The global checklist of crop wild relatives[23] (Vincent *et al*. 2013) and the GRIN Taxonomy provide important information resources for the classification of taxa as crop wild relatives. Traits of interest for breeding are being added and will complement the species attributes. The CWR classifications from these and similar checklists should be published to the GBIF

[23] http://www.cwrdiversity.org/checklist/

checklist bank and integrated into the GBIF taxon backbone. European Native Seed Conservation Network (ENSCONET) developed a database for crop wild relatives. ABD users should be enabled to filter occurrence data based on the CWR, gene pool and taxon group classifications. National checklists of CWR conservation priority species and their status should also be published into the GBIF checklist bank and made available through the GBIF portal.

Occurrence data from the Global Atlas of crop wild relatives should be published via GBIF (see figure 2). National programmes for the monitoring of CWR species should be promoted by national GBIF nodes and datasets published through GBIF.

Figure 2: Distribution of occurrence data in the global dataset of crop wild relatives (CWRDGC, 2015). This map was prepared with information mobilized via GBIF and other sources (i.e., herbaria, gene bank databases, researchers archives), and includes crop wild relatives occurrence data and cultivated occurrence data, offering a picture of the current availability of ABD occurrence data. Map prepared by Steven Sotelo (CIAT).



Recommendations

6.3.1. The global crop wild relative species checklist (www.cwrdiversity.org/checklist/) has to be published to the GBIF portal registry of checklists and integrated in to the GBIF taxonomy backbone. To complement this global list, other crop wild relatives checklists for publishing as a taxon checklist via GBIF may be proposed, starting with the crop wild relative species list developed by the Southern African Development Community (SADC)-CWR project which includes a global list of crop genus names - a useful tool for national species list of crop wild relatives.

6.3.3. New GBIF indexed attributes has to be added at the taxon-backbone-level to:

1. Identify the conservation priority and status of crop wild relative species at the global level (and at the regional and national levels in a taxon-level extension).
2. Provide information on the relationship of crop wild relatives (CWR) and their associated crops (e.g. gene pool and taxon group).
3. These taxon attributes should be implemented as conditional filters for selecting occurrence data in the GBIF portal.

6.3.4. Train GBIF Nodes on the value of CWRs, and mobilization of data on crop wild relatives and on species traits useful for crop improvement and for landscape restoration.

6.4 Mobilizing data on cultivated plants

In their farms, small-scale farmers maintain a large diversity of cultivated species and recognize many different types ('landraces' *sensu* Harlan 1975 and Villa *et al*. 2008) within each of their crops (Jarvis *et al*. 2008). Over 200,000[24] landraces of rice (*Oryza sativa* L.) are estimated to exist worldwide and about as many varieties of bread wheat (*Triticum aestivum* L. subsp. *aestivum*). There are about 47,000 varieties of sorghum, 30,000 varieties of common bean (*Phaseolus vulgaris* L.), chickpea (*Cicer arietinum* L.), and maize (*Zea mays* L.), about 20,000 varieties of pearl millet, 15,000 varieties of peanut (*Arachis hypogaea* L.), and between 7,000 and 9,000 varieties of manioc (*Manihot esculenta* Crantz) (FAO, 1998, 2010, Deletre *et al*. 2012).

The growing interest in neglected and underutilized crops (NUS) reflects rising concerns over this increasing reliance on a handful of crops to ensure global food security and economic growth (Padulosi *et al*. 1999, Stamp *et al*. 2012). NUS encompass a variety of plant species that are farmed (minor crops), reared (semi-domesticates), or gathered from the wild for a variety of uses and may contribute to nutrition (food, beverage), medicine, cosmetics, fodder, fibres, fuel, or provide material for building. NUS also include some ornamental plants. Although the promotion and conservation of NUS is part of FAO Global Plan of Action for the Conservation and Sustainable Use of Plant Genetic Resources for Food and Agriculture since 1996, NUS have been overlooked by breeders and botanists and data are lacking on their taxonomic/nomenclature, ecology, distribution, genetic diversity, local uses, and nutritional value. Inadequately described or characterized, NUS are at high risk of cultural and genetic erosion (Vietmeyer 1986).

Information required:

1. Taxonomy and checklists of traditional names in relevant languages.
2. Geospatial distribution information in cultivated areas.
3. Morpho-taxonomy, agronomic traits (farmers and breeders), functional traits, local uses, characterization and evaluation data.
4. Use, agronomic practices, cultural practices, seed conservation and exchange

[24] *FAO 2010 actually estimates a total of 35% landraces of rice from a total of 773,948 rice accessions, which amounts to approximately 270,882 rice landrace accessions*

Recommendation

6.4.1. Authoritative checklists and classification of crop wild relatives, cultivars, landraces and neglected and underutilized crop species, including vernacular names from authoritative lists along with the language and countries where it applies, should be added to GBIF when developed and validated by an international expert group and community.

6.5 Interactions between species

ABD scientists study the spatial distribution and species interactions in order to predict, for a given area, the possible gene flows between crops and wild species relative to crops, draw conclusions on the opportunities for the evolution of on-farm diversity and crop adaptation, and identify trade-offs and risks in interventions for conservation, land management or restoration actions. Data about the presence of livestock, of pests, diseases, helpful insects like pollinators, or about the predators of pests are needed. Additionally, the risk of a species turning invasive in a given environment is important information for decision making on landscape restoration, natural resource management, and conservation of threatened species.

Recommendations

6.5.1. Additional attributes are needed about species relation to crops or between species at the taxon name level like pest, predator, pollinator, etc.

6.5.2. Additional attribute ‘pathogen’ with the scientific name of the pathogens and vernacular names of the diseases should be made available.

6.5.3. The GBIF portal should enable the selection and download of crop occurrences along with occurrences of pests, diseases, pollinators, livestock, etc.

6.5.4. Information on the risk of a species becoming invasive has to be made available through a link between the GBIF portal and other databases holding such information, like the CABI database on invasive species.

6.6 Improving the mobilization of new data sources

Existing CGIAR and other gene bank databases are critical sources of information for the ABD community. However, there are other alternative sources of ABD information that remain in the grey literature and need to be digitized. Others exist in digitized formats but are not yet available through GBIF. Approaches for making this information readily available are listed below.

- Stimulate the digitization of relevant collections (i.e. herbaria, gene banks, published articles, MSc and PhD theses, national and regional projects) related to

ABD, by providing small grants in calls for competitive projects, as through the EU Biodiversity for Development (BID) call.

- Keep additional initiatives on the radar, as they progress with the digitization of herbaria specimens important for both the wider biodiversity and agrobiodiversity communities. Examples of such initiatives include JSTOR plants, Beyond the Box digitization competition[25], and data repatriation projects such as the "Capture of primary biodiversity data on West African plants", where images of specimens from large herbaria are being digitized.
- Provide alternative tools that require low technical expertise, including further development of the GBIF spread sheet template for publishing data to GBIF.org. The current IPT service is demanding in terms of informatics skills. For reporting the mandatory dataset metadata, an offline template and/or an online form should be provided. Alternatively a solution based on the Global Registry of Biodiversity Repositories (GRBio)[26] or similar can be explored as a solution for dataset metadata registration.
- Promote the use of citizen data portals such as the EarthSky[27] and iSpot[28] as a means for publishing data through GBIF.org.
- Existing digitized ABD data collections that should be considered to be made available through GBIF as they contain new information:
  1. The Crop Wild Relative Global Occurrence Dataset[29],
  2. The Bioversity Collecting Mission Database[30],
  3. Data, including the CWR genera list from the Southern African Development Community (SADC)-CWR project[31].

Recommendations

6.6.1. Existing digitized ABD data collections, such as the Bioversity Collecting Mission database[32] and the Crop Wild Relative Global Occurrence dataset (see map on page 20, figure 2), should be published through GBIF.

6.6.2. Stimulate the digitization of relevant collections (i.e. herbaria, gene banks, published articles, MSc and PhD theses, national and regional projects) related to ABD and stimulate the publishing of already digitized collections, by providing small grants through competitive calls.

6.6.3. Support the publishing of occurrences on ABD by rendering the upload of data records easy, requiring a very low technical expertise, and by providing an offline template, online

[25] https://beyondthebox.aibs.org/
[26] http://grbio.org/
[27] http://earthsky.org/earth/citizen-scientists-hit-one-million-mark-for-observations-of-nature
[28] http://www.ispotnature.org/communities/global
[29] http://www.cwrdiversity.org/checklist/cwr-occurrences.php
[30] http://bioversity.github.io/geosite/
[31] http://www.cropwildrelatives.org/sadc-cwr-project/
[32] http://bioversity.github.io/geosite/

form or a system (such as e.g. GRBio) for reporting the mandatory metadata describing the data set.

6.6.4. Promote the use of citizen data portals such as the EarthSkySea and iSpot as a mean for publishing data to GBIF.org (as complementary to systems such as the iNaturalist that are already publishing citizen scientist observations through GBIF).

6.7 Data Mobilization targets for Nodes

As indicated in the report '*Agrobiodiversity in perspective*' (Delêtre *et al.* 2012) commissioned by Sud Experts Plantes and Bioversity International, the input of national experts is essential to create national inventories of CWR, NUS and landraces, to refine species checklists, and to identify and document knowledge gaps. Specific objectives should be to:

1. Revise or complete the taxonomy of lesser known plant genera.
2. Evaluate the species' genetic diversity.
3. Gather detailed information on species distribution and ecology.
4. Collect ethnobotanical data on folk knowledge and traditional uses.
5. Assess the nutritional value and potential for commercialization of NUS.

With the help of the national agrobiodiversity research community, nodes are encouraged to contact the national agrobiodiversity data holders to improve ABD data availability through GBIF. Together, they can assess and influence national priorities.

However, Nodes should receive guidance and training on the mobilization and cleaning of agrobiodiversity data, and on CWRs and their importance for human food security. Experts could share their knowledge with Nodes by developing and sharing a global and a national checklist of CWR and neglected species so that CWRs can be identified as priorities for mobilization by the Nodes. A recommendation on this could be formulated by the task group on 'accelerating the discovery of bio-collections data'[33].

Recommendations

6.7.1. GBIF nodes could have a significant role to play if they are properly trained in the identification of data relevant to agrobiodiversity (i.e., crop wild relatives and on-farm diversity). Experts of agrobiodiversity data can provide support and best practices to Nodes to get acquainted with the expected data types (e.g. data collection methodology developed for the Crop Wild Relative Project of Southern African Development Community (SADC)[34]).

6.7.2. A key and simple step is to increase the knowledge of nodes through training so that they can play a more prominent role in the mobilization of locally available information resources on ABD in GBIF.

[33] http://www.gbif.org/governance/task-groups

[34] http://www.cropwildrelatives.org/sadc-cwr-project/

6.8 Services and tools for data processing and cleaning

Starting an agrobiodiversity study often means creating a reliable checklist of species, subspecies and cultivars that will be used for extracting relevant data from various sources. Scientists responding to the survey reported that this labour-intensive work is usually done manually. Therefore, GBIF could promote (and potentially integrate into the GBIF portal) tools selected based on popularity to serve the agrobiodiversity community, in particular for the curation of georeferences and taxonomy. An example is GEOLocate[35] were GBIF might provide a service to package data extracted from the GBIF portal into a format suitable for upload into this tool, or potentially explore possibilities to integrate the GEOLocate tool into the GBIF portal. Modelling pipelines and workflow services such as BioVeL and other Galaxy and Taverna compliant protocols, in conjunction with universal workflow technology, can be used to help with data pre-processing. Additionally, the GBIF helpdesk and tool directory can be improved for supporting data processing with ample help and manuals for the users. Tools for data processing can be hosted in an online environment for workflow processing, such as Galaxy Toolshed[36].

GBIF could become the point of access for the most reliable and up to date taxonomy for agrobiodiversity. Checklist resources such as USDA GRIN Taxonomy[37] (52,577 names of the respective plant agrobiodiversity species) is already integrated into the GBIF backbone taxonomy. Similarly, the Mansfeld´s World Database of Agricultural and Horticultural Crops[38] taxonomy (6,100 names, and the most complete checklist for cultivated species) should be closely integrated with additional information attributes (using an extension to the Taxon Core for additional terms such as gene pool status, etc).

Recommendations

6.8.1. GBIF to become the point of access for the most reliable and up-to-date taxonomy (including cultivars).

6.8.2. GBIF should seek to support the integration of popular data cleaning tools such as GEOLocate, OpenRefine (formerly Google Refine), and workflow services from BioVeL and other Galaxy or Taverna compliant protocols with data published through the GBIF portal. It is also important to take into account the requirements on use cases that are being developed by a task group of TWDG/GBIF data quality the interest group).

6.8.3. GBIF needs to increase visibility of existing taxonomic name reconciliation tools Global Names Architecture (GNA)[39] on GBIF.org, and provide access to the Plant List[40] developed by botanical gardens.

6.8.4. GBIF has to implement or improve tools for cross-checking and validating nomenclature of records published by different collections on the GBIF portal using

[35] http://www.museum.tulane.edu/geolocate/

[36] https://toolshed.g2.bx.psu.edu/, https://wiki.galaxyproject.org/ToolShed

[37] http://www.gbif.org/dataset/66dd0960-2d7d-46ee-a491-87b9adcfe7b1

[38] http://mansfeld.ipk-gatersleben.de

[39] http://globalnames.org/

[40] http://www.theplantlist.org/

taxonomic authorities such as GRIN Taxonomy, Mansfeld's World Database of Agricultural and Horticultural Crops, IPNI, the Plant List and Tropicos to resolve naming issues.

6.9 De-duplication of occurrence-level data records

The agrobiodiversity community currently uses a number of different data flow mechanisms and portals for germplasm records. They need guidance in retrieving data across this broad landscape of data sources, including GBIF.

There are two types of duplicates – duplicate data records about the same accession (published from different datasets) and duplication of accessions originating from the same collecting event (copies of the same living material conserved in different gene banks). In fact, the very same occurrence of originally collected living plant material can be copied across multiple gene bank collections and lead to a physical duplication of the actual physical biological material. All these specimens or accessions are connected to the same original occurrence in nature or in a farmer's field. It is estimated a total of 7.4 million gene bank accessions conserved in *ex situ* collections worldwide originate from a total of almost 2.2 million original collected material samples (FAO 2010). Consequently, in the gene bank world, information about different accessions for each occurrence can be conserved and made available from several different gene bank collections. This “duplication” of occurrence records (across gene bank collections) carries varying richness and very different types of information for the same occurrence record, according to local needs.

The other type of duplication concerns information on the very same accession or specimen included in more than one dataset published via GBIF. Some datasets such as the Genesys gene bank portal or the Global Inventory of CWRs provides a meta-catalog with a mixture of datasets already published in GBIF from the source and other datasets not yet published in GBIF. Other datasets such as trait experiments by crop researchers and plant breeders contribute new information not available from the gene bank collection but linked to the accessions in the gene bank collections.

A useful solution is the new feature of the GBIF API that implements *occurrenceID* as a searchable term for emerging persistent identifiers coming from data holders and data publishers. Having occurrenceID as a searchable term demonstrates to data publishers the value of providing persistent identifiers for their specimens. Occurrence information from multiple data sources (datasets published by different data owners) can more easily be combined/merged with persistent occurrence-level identifiers.

A valuable functionality for the GBIF portal would be an “occurrence-backbone” to merge and combine occurrence information on the same occurrence/specimen provided by different data owners (different datasets but using the same specimen/accession-level persistent identifier) in the same manner in which taxon names from multiple sources are combined to form a GBIF taxon backbone. The same occurrence can be included in different datasets from different data owners for a number of valid reasons such as an interest or focus on different types of attribute information.

Agrobiodiversity specimens (accessions) are routinely the object of different types of experiments conducted by crop scientists or commercial crop breeding companies each with

their own database systems. Following the implementation of an “occurrence-backbone”, an estimated 600,000 occurrences aggregated by Genesys and a few hundred thousand occurrences from CIAT that are not yet published through GBIF could readily be added. A portion of the records included in the above mentioned agrobiodiversity datasets originate from GBIF. The best strategy would be to refresh existing datasets using locally unique (currently not globally unique) record identifiers. De-duplication using resources based on GBIF-mediated data mixed with unique new data, cross-linking towards other datasets published in GBIF would be needed.

Recommendations

6.9.1. Implement as an additional feature to the GBIF portal an ‘occurrence backbone’ by merging and combining information for the same occurrence/specimen linked using persistent identifiers, for datasets provided by different data owners, in the same manner as for taxon names.

6.9.2. Publishing additional occurrences from Genesys, the Crop Wild Relative occurrence dataset, and other ABD datasets containing a mixed set of new records and records already published via GBIF using shared or linked specimen-level identifiers so that information about the same specimen/accession can be linked together.

6.9.3. Using unique and semi-unique identifiers to identify “duplicate” records between different datasets could improve the mechanisms for updating and refreshing datasets in the GBIF portal. Cross-linking and de-duplication would be needed when using resources based on existing GBIF-mediated data records mixed with unique new data.

6.10 Agrobiodiversity user profile access

Users accessing the GBIF portal need to register with a user profile before downloading data. The user registration system should be expanded to allow users to choose from a set of predefined user profiles. An ‘agrobiodiversity user profile’ could offer thematically designed search widgets and tools to support commonly performed operations on the GBIF portal by this type of user. The downloaded files could also be offered within a specific optimized ‘agrobiodiversity-format’. A hierarchy of data profiles will be useful here. The purpose of such user profiles is not to reduce the content by hardcoding a rigid filtering to exclude data records outside the ABD domain, but to put upfront thematically designed search tools to support ABD users to find the information they need with fewer operations. The so-called “reference dataset” approach might prove to be less effective here, because the ABD user generally wants to discover new sources of crop and crop wild relative data records rather than find those records previously identified and included in a “reference dataset”. Algorithms to identify newly available data records, not yet discovered as agrobiodiversity relevant data records, would generally be more useful.

Recommendations

6.10.1. A hierarchy of data-profiles and user-profiles, thematically designed search widgets, and tools would enable thematic users (such as ABD users) to increase efficiency of their use of the GBIF portal to find the range of data they need.

6.10.2. Implementing a hierarchy of thematically designed data/user profiles at the API level would also help with simplified access.

6.11 Improving fitness for use through data quality

Concerns regarding the general quality of the data were frequently mentioned in the survey as a major bottleneck for the reuse of occurrence and taxon data published in GBIF. GBIF should perform preliminary quality filtering of the data resulting in the attribution of a level of confidence to individual data records, at the dataset-level (aggregated records) and provision of clear feedback to the data suppliers. Having data pre-processed by the GBIF portal to flag potential issues will greatly help data scientists identify and select cleaned and final data records to be included in their studies. Part of such mechanisms could be the creation of crop/species expert groups, including taxonomy expertise, to review and endorse the quality of agrobiodiversity data made available in the GBIF portal. In this regard, GBIF could consider providing user-configurable and tuning-enabled online services for checking data quality, identifying outliers, data preparation, manipulation and visualization. Such services together with the option of adding multiple layers of data will boost the definition of a scientific hypothesis, research collaboration and knowledge gap analysis.

The survey reported that only an estimated 50% of GBIF published data records are useful in analysis because of out-dated names, geographic issues, dates, and identification issues. The data publishable with the GBIF Integrated Publishing Toolkit (IPT) (Robertson *et al*. 2014) should bear a quality stamp. Publication of incomplete data sets should not be blocked, but data owners should be able to inform about data quality issues of their data such as the completeness of the dataset. Completeness can be addressed at the national level, and at the regional level. Tools to check the completeness of data, to identify improvement targets, such as flagging incomplete taxonomic names, should be made available at national level. Node managers should be provided with adequate tools and enabled to perform data cleaning (i.e. flagging data quality issues) at national levels before data publication. By screening upfront, publishers should be alerted to improvements needed, such as percentage of completeness. Statistics should be available on reliable coordinates and reliable taxonomy. Complete name string and assessments of how up-to-date the information is should be backed by publication, sequence, or expert. Users would benefit from receiving alerts when datasets that they use are improved with some indication of trust. Actual use of datasets may guide the correction and should be channelled back to the data publishers as a guideline for improvements and republishing. With respect to publication of the so-called “improved-reference-datasets”, “improved datasets” can be indexed in the GBIF portal as yet another source of “evidence” for an interpreted “occurrence-backbone” entity. Following this approach, information pieces provided by different data publishers about the very same occurrence entity should be linked together and combined. Information pieces on the same occurrence from different data publishers should be displayed together, with conflicting information highlighted. GBIF should also find a way of getting feedback from data users on the quality of downloaded data and steps for publishing the cleaned data.

GBIF.org could further improve the routines for preliminary quality assessment of data records, for assigning a level of confidence to individual data record and provide clear

feedback to data suppliers, highlighting potential issues. A level of confidence could only be applied within a specific context. A standard suite of tests should be developed that can be applied at the record level to the fields of Darwin Core and germplasm extension. The tests that return suspect or in-error status could be tabulated and accumulated into a score for the record (multiple tests can be applied to each Darwin Core field), but the weighting given to the tests would be use-dependent. Ideally, the aim should be to publish those weightings (formulae) for each application/use (Belbin, L. communication).

Evaluation at the dataset-level is necessary as some quality assessments can only be inferred from aggregates of data records. This is included in the *Guidelines for listing mechanisms for DQ Validation, Measurement and Improvement*[41] developed by the TDWG interest group on data quality. There are two aspects of dataset-level evaluations — accumulating record-level evaluations into an overall evaluation, and evaluating single records based on dataset context. The former should be easy in theory and would involve accumulating weighted scores for each Darwin Core field. Context-type evaluations such as testing against an environmental envelope are in place, but certainly need clarification and standardization. An important issue is that that annotations themselves need to be standardized and always remain with the record. Darwin Core probably needs to expand to accommodate this (Belbin *et al*., 2013).

Recommendations

6.11.1. GBIF should improve routines for preliminary quality assessment of data records and datasets (aggregated records) giving levels of confidence to individual data records or datasets and highlight issues to data suppliers. **A level of confidence can only be applied within a specific context, and a weighting of the scores (possibly 'weighted completeness' and 'weighted issues'*)* should be proposed in the context of use by the ABD community.**

6.11.2. GBIF should develop or adapt existing tools to:

1. Identify quality improvement thresholds based on the decided weighting of scores e.g. unreliable coordinates; issues with taxon names regarding completeness of name-strings and up-to-date nomenclature; if names are backed by either publication reference, sequence, or expert.
2. Check the completeness of the data (e.g. index of passport data completeness) through possibly two scores: 'weighted completeness' and 'weighted issues'.
3. Provide the percentage of records with actual data reported for each attribute (data column), possibly with two scores: 'weighted completeness' and 'weighted issues'.
4. Highlight attributes in a search result with no actual data reported.
5. Provide statistics on the percentage of completeness and issues that over time can be used to produce a graph about the completeness of data over time and issues.

[41] http://community.gbif.org/pg/pages/view/42614/guideline-for-listing-mechanisms-for-dq-validation-measurement-and-improvement

6.11.3. Data formatted for being published in GBIF should bear a quality stamp provided by the data owner and the publisher. Sometimes the data owner lacks the resources or the relevant expertise to improve the data and address a set of known and documented data quality issues. User-friendly tools, such as the sandbox of the Australian Living Atlas[42] should be proposed.

6.11.4. Node managers should be capacitated and provided with adequate tools to perform data cleaning (by highlighting data quality issues) at national levels before data publication. The GBIF portal could display such data quality issues as annotations to the respective occurrence or taxon entity. Data quality annotations could also be made available in data downloads from the GBIF portal as extra columns.

6.11.5. Reference datasets, with improvements suggested by nodes or users, should be published to the GBIF portal following the procedures for standard dataset publication and indexing. Such datasets could be flagged as reference-datasets in the IPT. Information elements from the reference-dataset could be displayed alongside information provided by the data owner so that conflicts with external/additional annotation information are identified. A browser extension, easy to install, that provides at-a-glance insights about datasets available through GBIF.org was developed by GBIF Belgium and should be tested and promoted for quality improvement[43].

6.12 GBIF portal improvements

Search

GBIF.org portal search functionality will need to be improved. The current search functionality is not intuitive and a search for a valid taxon name in the primary search box does not return any relevant results. The search box on the front page is limited to news and information pages at http://www.gbif.org, while a much broader search functionality here, including taxon and occurrence information, is expected by many users. Search for a taxon name, dataset, taxon, data originator pages would ideally have faceted tab stats. GBIF would ideally include solutions for searching using all of the core data properties and thus become a richer data infrastructure.

Download

Many agrobiodiversity users find the download procedure rather complicated, most likely because of the large amount of data to be filtered and the number of steps, including registration, needed before the user can download search results.

Tagging and visualization of ABD data

It is desirable that at GBIF.org agrobiodiversity-related data are identified and visualized. Tagging or indication of agrobiodiversity-related sources, without reducing the agrobiodiversity user's access to the full content, would be very useful.

[42] http://sandbox.ala.org.au

[43] http://devpost.com/software/gbif-dataset-metrics-xfvzns

Attribution

Better ways of crediting data publishers and data originators, for e.g. by attaching names to datasets in a more visible way, and including lists of contributors to-thank and to-invite-as-co-authors together with the citation and DOI information for downloads.

Access to features

It is desirable that the most relevant functionalities of the portal are placed upfront and promoted to our community, following solutions enabled by the implementation of an agrobiodiversity user profile. Few people currently use the temporal axis feature of data, but those who do seem to be satisfied with it. This feature needs to be promoted as it is relevant to monitoring ABD, and to studies of genetic evolution or erosion.

The GRIN Taxonomy is already embedded in the taxonomy backbone but this is not well known by ABD users. Additionally, the Global Names Architecture[44] (GNA), Global Names Index[45] (GNI) and the Taxonomic Name Resolution Service[46] (TNRS) provide solutions for cleaning taxon names, including helping to extract all synonyms used by GBIF when enabling access to occurrence data, but this functionality is not well known in the ABD community. As an example, a survey respondent used the 'Plant List' to manually reconcile the names for the Neglected and Underutilized species checklist. This situation can also be resolved through training modules on topics such as the creation of taxon name checklists.

Application Programme Interface (API)

A little over 10% of the respondents use the GBIF portal API but some are major data sources for ABD (e.g. FAO).

Recommendations

6.12.1. Expand or enable the search functionality on the front page of GBIF.org to include a search for taxon and occurrence information. Place the Darwin Core descriptors with their definition in a more accessible position so that users can easily check the meaning of each field when they download data.

6.12.2. Inform the ABD community about the current availability of GRIN Taxonomy, developed by US Department of Agriculture and used by gene banks, in the GBIF taxonomy backbone.

6.12.3. Expand the data attributes made available for search. Include the most important agrobiodiversity terms from the MCPD and the corresponding Darwin Core germplasm extension as searchable information attributes (such as gene pool and taxon group concepts, trait information, characterization and evaluation data, pre-breeding and breeding information).

6.12.4. If profiles for agrobiodiversity users are enabled, they should offer users an alternative data download format including the most important terms from the MCPD and the Darwin Core germplasm extension, utilizing some of the currently provided Darwin Core terms that are relevant for use in agrobiodiversity research.

---

[44] http://globalnames.org/

[45] http://gni.globalnames.org/

[46] http://tnrs.iplantcollaborative.org/

6.12.5. Develop simple guidelines on the steps for downloading GBIF data. These guidelines should be placed prominently on GBIF.org.

6.12.6. Lack of use of the temporal axis in searches is a consequence of ignorance among users regarding this feature. We suggest that GBIF post a teaser similar to the 'Google Earth history timeline' that shows the spread of diseases across months.

6.13 Mashup agrobiodiversity data sources into a single access point

Providing contextual information to ABD data is an added value for the ABD community and is likely to increase the use of data published through GBIF.org. The lack of availability of external environmental and trait data are most frequently mentioned in the survey.

Such contextual information can be proposed through the user profile by calling existing open access repositories through their API or web services from relevant sources like:

- Genetic related data: NCBI genbank, DRYAD
- Taxonomy: USDA GRIN taxonomy, Mansfeld taxonomy, Plant List
- Climate and environmental data: Worldclim, CRU, ISRIC
- Scientific literature and other related references: Elsevier API, PLOS API
- Trait databases: TRY, Crop Ontology, Biopop, LEDAtrainbase
- Agriculture statistics: FAOSTATS
- Satellite derived images, remote sensing: NASA Geoexplorer, ESA scientific data hubs.
- Socio-cultural data: archaeology, linguistic and ethnographic data [https://www.ethnologue.com/]

This contextual information should be downloaded into a single meaningful dataset. The FAO AGRIS website is a mashup site for the large FAO database of bibliographic references which provides contextual information from external sources by using web services (see figure 3). Aside from bibliographic references, AGRIS displays data from the Biodiversity Collecting Mission Database, using the country and the genus of the crop as key searchable terms, as well as statistics from the World Bank database.

Figure 3: FAO-AGRIS web site (http://agris.fao.org/agris-search/index.do) showing bibliographic references displayed with contextual data from external sources.

Recommendations

6.13.1. Communicating and advertising new links and procedures more widely so that publishers can include more information on these in their websites.

6.13.2. Making GBIF API more interconnected and accessible to the agricultural information providers such as AGRIS, FAOSTATS, CABI.

6.13.3. GBIF and partners should explore and test connecting GBIF-mediated data with Linked Open Data (LOD). GBIF could provide access to occurrence data in adequate formats for semantic web services to use.

6.14 Combining GBIF-mediated data with external data sources

Many survey respondents reported that combining GBIF-published data with other types of external data, including environmental spatial data, occurrence/genotype-level trait data and molecular genetics diversity data, in their studies. GBIF is not expected to mobilize or channel all of these data types, but helping enable smooth linking would be very useful. According to the survey, occurrence-level data are most often linked to external data sources using geographic coordinates, followed by taxon name. However, in most cases, persistent identifiers are NOT used to build such links. This user pattern is possibly influenced by dominance of distribution modelling among respondents' use cases.

There are “external” data attributes important for the agrobiodiversity community, such as identification of cultivated plants and overlay with poverty maps, ethnographic, archaeological, social, linguistic data, agronomy, morphology, physiology, environmental data (sunshine hours, temperature, light intensity, rainfall, soils observations and profiles) and threat categories from the IUCN Red List. Databases relating the diversity to local problems and other useful ‘combined-with-GBIF’ data, such as the database under construction for neglected and underutilized species of Crop For the Future, could be connected to GBIF. Interaction data will be of better use if found not only in verbatim view, which is also retrievable using the API, but also potentially more integrated in the portal and searchable.

In addition to GBIF providing access to CWR resources from herbaria and museum collections, the main sources of ABD occurrence data are local (gene bank) databases, Genesys[47], EURISCO[48]. Non-occurrence sources are FAOSTAT[49], WorldClim[50] and USDA GRIN Taxonomy[51]. Some datasets used by agrobiodiversity users have more detailed information on taxonomy, place and time of collection, photos, and plant uses. For example, there are a number of records of wild species collected in research field stations that some users would assume to be “native” or within the natural range of occurrence of the species. Therefore it would be useful to have the cultivation status information linked to agrobiodiversity records and information on international treaty legislation governing the access and regime for agrobiodiversity species and populations.

Recommendations

6.14.1. Occurrence-level data are most often linked to external data sources using geographic coordinates, followed next by taxon name, and taxon ID. Consequently quality of these data is paramount (see recommendation 6.9.x and 6.11.x).

6.14.2. There is a growing interest in the ABD community for GBIF to expand currently supported data types (taxon, occurrence and event) to include experimental data such as crop trait information and characterization and evaluation data (using for e.g. Darwin Core MeasurementOrFact terms).

6.14.3. The most relevant experimental data in ABD include field trials on crop disease resistance, greenhouse trials and other economically useful trait measurements subsequently required in crop improvement activities and commercial plant breeding.

7. Use Cases

Several use cases are described using a Use Case Template to illustrate the real uses of the data. A use case enables the understanding of data flow, the user experience and the implications for GBIF in terms of data accessibility, quality, and use by the ABD community.

[47] https://www.genesys-pgr.org/welcome
[48] http://eurisco.ipk-gatersleben.de/
[49] http://faostat.fao.org/
[50] http://worldclim.org/
[51] http://www.gbif.org/dataset/66dd0960-2d7d-46ee-a491-87b9adcfe7b1

Please note that the TDWG/GBIF biodiversity data quality interest group[52],[53] (TDWG/GBIF DQIG 2015a) has initiated a collection of use cases with the aim of building tools to convert use cases from plain text to a formal framework (TDWG/GBIF DQIG 2015b)[54]. The Use Cases requirement spreadsheet developed by this group should be included in the ABD Use cases.

See full description of used cases in Appendix 2.

Template proposed by the task group for developing use cases

1. Describe the objective	
2. Who are the actors ?	
3. Data/information products to be produced	
4. Data sources the most used	
5. Tools the most used	
6. Describe the data flow into steps ("to be able to...") indicating who is doing the indicated step (a scientist, a developer, the system?)	
To be able to …	
7. Identify GBIF role and improvements	
i. status of data (quality, coverage)	

[52] http://www.tdwg.org/activities/biodiversity-data-quality/interest-group-charter/
[53] http://community.gbif.org/pg/groups/21292/
[54] http://community.gbif.org/pg/pages/view/47749/

ii. Additional attributes in demand	
iii. Additional sources - > what are the connectors	
iv. What data mobilization is needed? By whom?	
→ Is this a priority? if not implemented how will it block the progress?	
Link to recommendations #	

8. Bibliography and sources

Alercia A, Diulgheroff S, and Metz T (2001). FAO/IPGRI Multi-crop passport descriptors, December 2001. Food and Agriculture Organization of the United Nations (FAO), and International Plant Genetic Resources Institute (IPGRI), Rome, Italy.

Alercia A, Diulgheroff S, and Mackay M (2012). FAO/Bioversity Multi-crop passport descriptors, v.2. Food and Agriculture Organization of the United Nations (FAO), and Bioversity International, Rome, Italy. Available online at http://www.bioversityinternational.org/index.php?id=19&user_bioversitypublications_pi1%5BshowUid%5D=6901.

Alercia A, Diulgheroff S, and Mackay M (2015). FAO/Bioversity multi-crop passport descriptors, v.2.1. Food and Agriculture Organization of the United Nations (FAO), and Bioversity International, Rome, Italy. Available at http://www.bioversityinternational.org/e-library/publications/detail/faobioversity-multi-crop-passport-descriptors-v21-mcpd-v21/ doi:10.13140/RG.2.1.4280.2001

Belbin L, Daly J, Hirsch T, Hobern D, and LaSalle J (2013). A specialist's audit of aggregated occurrence records: An 'aggregator's' perspective. *ZooKeys* 305: 67-76. doi:10.3897/zookeys.305.5438

CWRDGC (The Crop Wild Relative Occurrence Data Global Consortium) (2015). A global occurrence dataset for crop wild relatives. *In prep.*

Delêtre M, Gaisberger H, and Arnaud E (2012). Agrobiodiversity in perspective: A review of questions, tools, concepts and methodologies in preparation of Sud Experts Plantes for Sustainable Development Programme. *unpublished report*. Bioversity International and Sud Experts Plantes. Available at http://doi.org/10.13140/RG.2.1.1467.5680

Endresen DTF, and Knüpffer H (2012). The Darwin Core extension for genebanks opens up new opportunities for sharing genebank data sets. *Biodiversity Informatics* 8:11-29. doi:10.17161/bi.v8i1.4095

FAO (1998) The state of the world's plant genetic resources for food and agriculture. FAO, Rome, Italy. Available at ftp://ftp.fao.org/docrep/fao/meeting/015/w7324e.pdf (verified 28 January 2016).

FAO (2010). The second report on the state of the world's plant genetic resources for food and agriculture. Food and Agriculture Organisation of the United Nations (FAO), Rome, Italy. ISBN: 978-92-5-106534-1. Available online at http://www.fao.org/agriculture/crops/thematic-sitemap/theme/seeds-pgr/sow/sow2/en/

FAO (2015). Input paper from Bioversity International, CIAT, CIP and GBIF: Global information system for *in situ* conservation and on-farm management of plant genetic resources for food and agriculture. The International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA), Food and Agriculture Organisation of the United Nations (FAO), Rome, Italy. IT/ACSU-2/15/Inf.3. Available online at http://www.planttreaty.org/sites/default/files/acsu2i3.pdf

Hazekamp T, Serwinski J, and Alercia A (1998). Appendix II. Multicrop passport descriptors. p. 97-90 *in*: Lipman E, Jongen MWM, van Hintum TJL, Grass T, and Maggioni L (eds). Central crop databases: Tools for plant genetic resources management. International Plant Genetic Resources Institute (IPGRI), Rome, Italy, and CGN, Wageningen, Netherlands. ISBN 92-9043-320-5.

Hammer K, Arrowsmith N, and Gladis T (2003). Agrobiodiversity with emphasis on plant genetic resources. *Naturwissenschaften* 90: 241–250. doi:10.1007/s00114-003-0433-4
Harlan JR (1975). Our vanishing genetic resources. *Science* 188: 618–621. doi:10.1126/science.188.4188.617

Harlan JR, and de Wet JMJ (1971). Toward a rational classification of cultivated plants. *Taxon* 20(4): 509-517. doi:10.2307/1218252

Jarvis DI, Brown AHD Cuong PH, Collado-Panduro L, Latournerie-Moreno L, Gyawali S, Tanto T, Sawadogo M, Mar I, Sadiki M, Hue NT-N, Arias-Reyes L, Balma D, Bajracharya J, Castillo F, Rijal D, Belqadi L, Rana R, Saidi S, Ouedraogo J, Zangre R, Rhrib K, Chavez JL, Schoen D, Sthapit B, De Santis P, Fadda C, and Hodgkin T (2008). A global perspective of the richness and evenness of traditional crop-variety diversity maintained by farming communities. PNAS 2008 105 (14) 5326-5331. doi:10.1073/pnas.0800607105

Khoshbakht K and Hammer K (2007). Threatened and rare ornamental plants. *Journal of Agriculture and Rural Development in the Tropics and Subtropics* 108: 19-39.
Khoshbakht K, and Hammer K (2008). How many plant species are cultivated? *Genetic Resources and Crop Evolution* 55: 925-928. doi:10.1007/s10722-008-9368-0

Maxted N, Ford-Lloyd BV, Jury SL, Kell SP, and Scholten MA (2006). Towards a definition of a crop wild relative. *Biodiversity and Conservation* 15: 2673–2685. doi:10.1007/s10531-005-5409-6

Moore JD, Kell SP, Iriondo JM, Ford-Lloyd BV, and Maxted N (2008). CWRML: representing crop wild relative conservation and use data in XML. *BMC Bioinformatics* 9: 116. doi:10.1186/1471-2105-9-116

Padulosi S, Eyzaquirre P, and Hodgkin T (1999). Challenges and strategies in promoting conservation and use of neglected and underutilized crop species. pp. 140-145 *in*: Janick J (Ed.) Perspectives on New Crops and New Uses. ASHS Press, Alexandria, USA.

Padulosi S, Thompson J, and Rudebjer P (2013). Fighting poverty, hunger and malnutrition with neglected and underutilized species (NUS): needs, challenges and the way forward. Bioversity International, Rome. ISBN: 978-92-9043-941-7.

Robertson T, Döring M, Guralnick R, Bloom D, Braak K, Otegui J, Russell L, and Desmet P (2014). The GBIF Integrated Publishing Toolkit: Facilitating the efficient publishing of biodiversity data on the Internet. *PLoS ONE* 9(8): e102623. doi:10.1371/journal.pone.0102623

Stamp P, Messmer R, and Walter A (2012). Competitive underutilized crops will depend on the state funding of breeding programmes: an opinion on the example of Europe. *Plant Breeding* 131(4): 461-464. doi:10.1111/j.1439-0523.2012.01990.x

Thormann I, Gaisberger H, Mattei F, Snook L, and Arnaud E (2012). Digitization and online availability of original collecting mission data to improve data quality and enhance the conservation and use of plant genetic resources. *Genetic Resources and Crop Evolution* 59(5): 635-644. doi:10.1007/s10722-012-9804-z

Thormann I, Alercia A, and Dulloo ME (2013). Core descriptors for *in situ* conservation of crop wild relatives v.1. Bioversity International, Rome, Italy. ISBN: 978-92-9043-935-6.

Thormann I, Parra-Quijano M, Endresen DTF, Rubio-Teso ML, Iriondo MJ, and Maxted N (2014). Predictive characterization of crop wild relatives and landraces. Technical guidelines version 1. Bioversity International, Rome, Italy. ISBN: 978-92-9255-004-2.

Thormann I, Parra Quijano M, Iriondo JM, Rubio Teso ML, Endresen DT, Maxted N, Dias S, and Dulloo ME (2015). Two predictive characterization approaches to search for target traits in crop wild relatives and landraces. *Crop Wild Relative* 10: 16-18.
TDWG (2013). Report of the TDWG Vocabulary Management Task Group (VoMaG). Baskauf S, Ó Tuama É, Endresen D, and Hagedorn G (eds). 25 pp. Available online at http://www.gbif.org/resource/80862

Villa TCC, Maxted N, Scholten M, and Ford-Lloyd B (2005). Defining and identifying crop landraces. *Plant Genetic Resources: Characterization and Utilization* 3(3): 373-384. doi:10.1079/PGR200591

Vincent H, Wiersema J, Kell S, Fielder H, Dobbie S, Castañeda-Álvareza NP, Guarino L, Eastwood R, León B, and Maxted N (2013). A prioritized crop wild relative inventory to help underpin global food security. *Biological Conservation* 167: 265-275. doi:10.1016/j.biocon.2013.08.011

Vietmeyer ND (1986). Lesser-known plants of potential use in agriculture and forestry. *Science* 232: 1379–1384. doi:10.1126/science.232.4756.1379

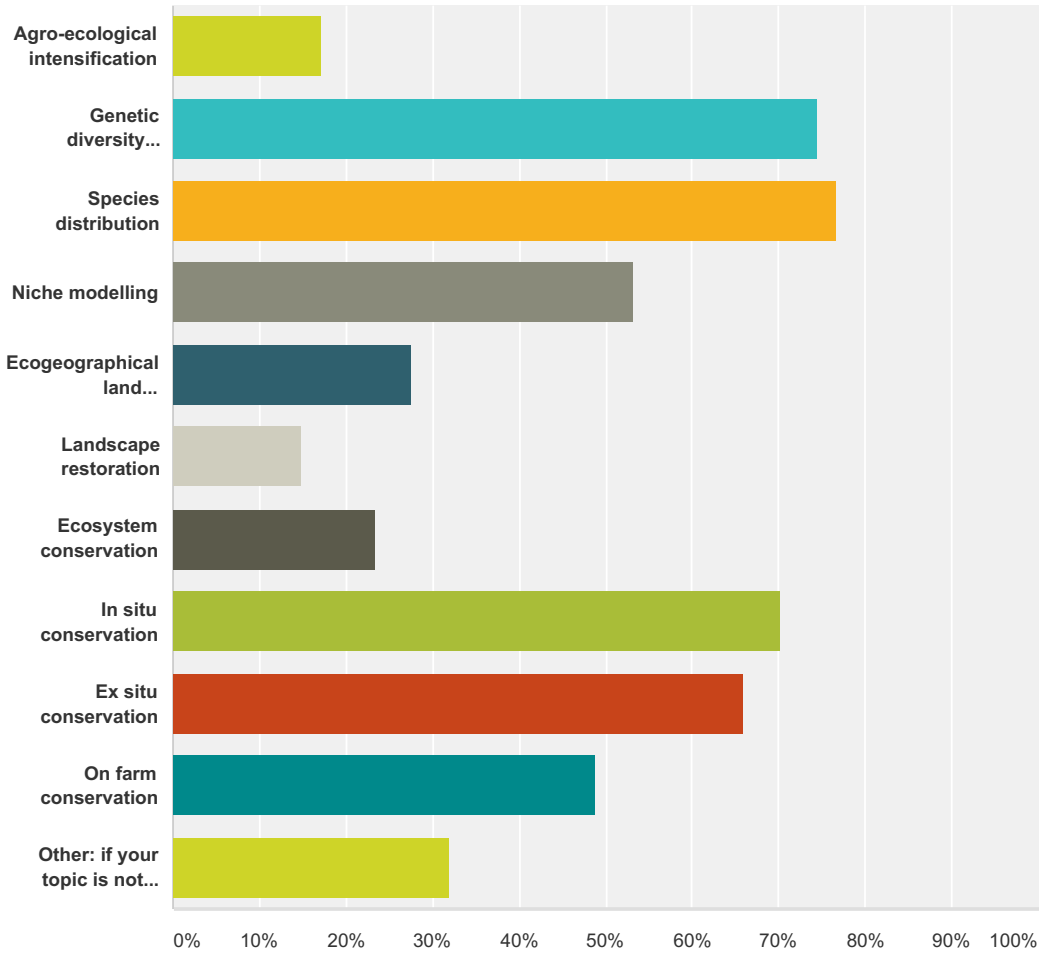
Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, and Vieglais D (2012). Darwin Core: An evolving community-developed biodiversity data standard. *PLoS ONE* 7(1): e29715. doi:10.1371/journal.pone.0029715

Appendix 1

Summary of survey on data fitness for use for agrobiodiversity

Q6 What is your interest in accessing/using Agrobiodiversity Data? Multiple choice question

Answered: 47 Skipped: 5

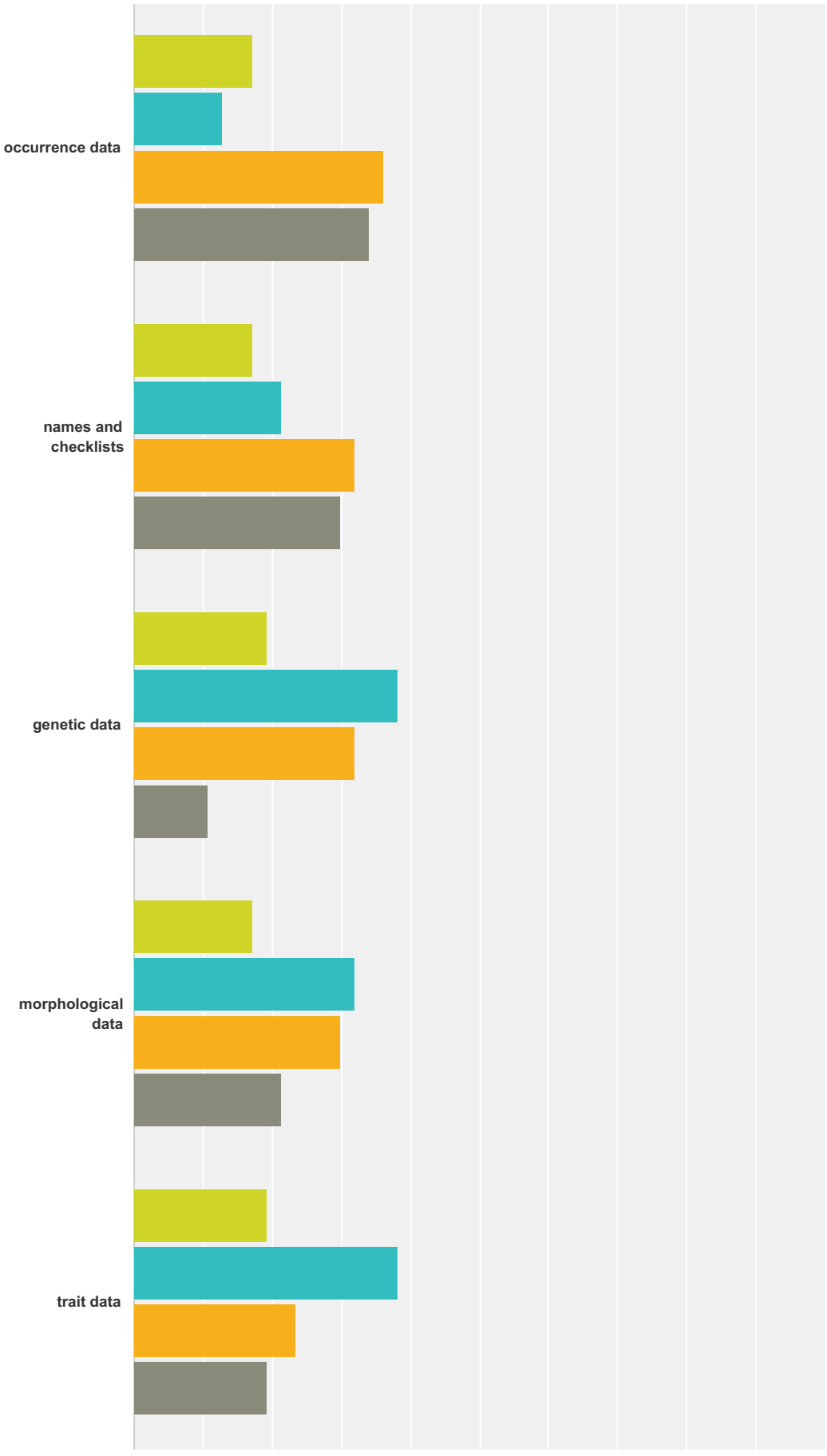


Answer Choices	Responses	
Agro-ecological intensification	**17.02%**	8
Genetic diversity analysis	**74.47%**	35
Species distribution	**76.60%**	36
Niche modelling	**53.19%**	25
Ecogeographical land characterization	**27.66%**	13
Landscape restoration	**14.89%**	7
Ecosystem conservation	**23.40%**	11
In situ conservation	**70.21%**	33
Ex situ conservation	**65.96%**	31
On farm conservation	**48.94%**	23
Other: if your topic is not listed, please specify, we are interested to learn about it, please specify (max. 200 characters)	**31.91%**	15
Total Respondents: 47		

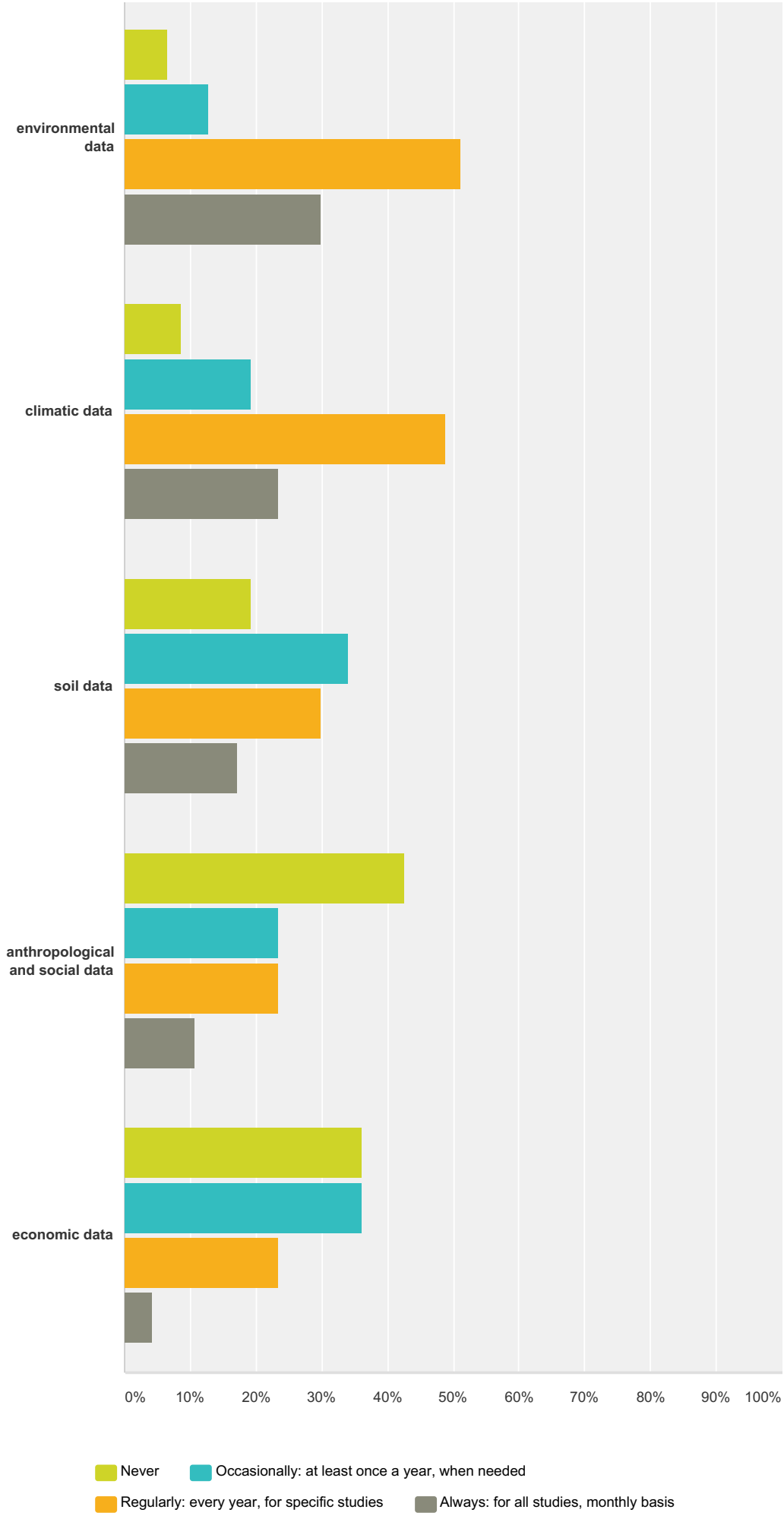
#	Other: if your topic is not listed, please specify, we are interested to learn about it, please specify (max. 200 characters)	Date
1	Ecogeographical overviews, taxonomic databases, genetic resources documentation	7/20/2015 10:42 AM
2	Genesys	7/14/2015 2:31 PM
3	predictive characterization (focused identification of germplasm strategy); nomenclature of crop wild relatives (checklist data)	6/30/2015 12:32 PM
4	breeding	6/30/2015 11:16 AM
5	Making better use of genetic variation for plant breedung	6/28/2015 7:45 PM
6	Utilization of genetic resources	6/26/2015 3:21 PM
7	Landscape ecology	6/22/2015 9:15 PM
8	Sustainable (broadly) production	6/18/2015 4:44 PM
9	Agrobiodiversity and Climate Change	6/16/2015 10:06 PM
10	Actually we don't really use the data. Our work is to make research data more visible, allowing its re-use by other researchers. For genetic resources issues, we will have interactions with ministry of agriculture for addressing FAO state of the world and plan of action informations and data requirements.	6/16/2015 11:09 AM
11	inventory of traditional knowledge and practices Nutrtional value	6/15/2015 1:37 PM
12	Spatial analysis	6/15/2015 12:57 PM
13	Eco-system services restoration	6/13/2015 9:15 AM
14	Germplasm utilisation, plant breeding	6/12/2015 6:28 PM
15	Policy	6/12/2015 5:50 PM

Q7 What are the type of data that you mostly use? Multiple choice question

Answered: 47 Skipped: 5



GBIF and Bioversity International: Data Fitness for Use in Agrobiodiversity 2015

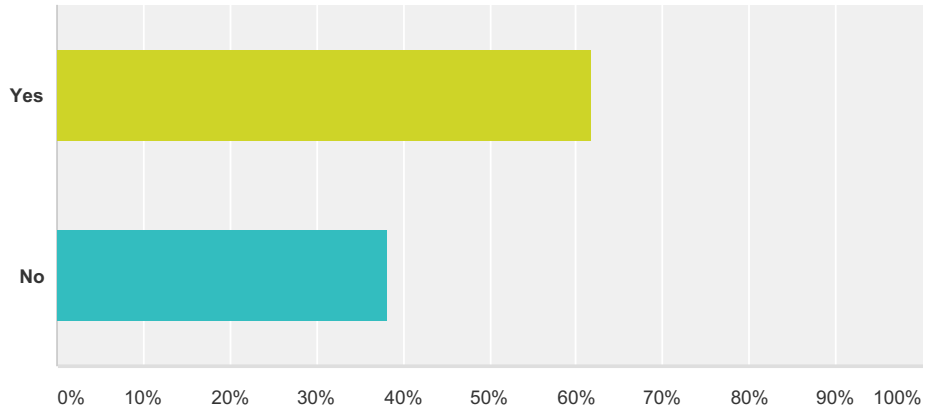


	Never	Occasionally: at least once a year, when needed	Regularly: every year, for specific studies	Always: for all studies, monthly basis	Total
occurrence data	**17.02%** 8	**12.77%** 6	**36.17%** 17	**34.04%** 16	47
names and checklists	**17.02%** 8	**21.28%** 10	**31.91%** 15	**29.79%** 14	47
genetic data	**19.15%** 9	**38.30%** 18	**31.91%** 15	**10.64%** 5	47
morphological data	**17.02%** 8	**31.91%** 15	**29.79%** 14	**21.28%** 10	47
trait data	**19.15%** 9	**38.30%** 18	**23.40%** 11	**19.15%** 9	47
environmental data	**6.38%** 3	**12.77%** 6	**51.06%** 24	**29.79%** 14	47
climatic data	**8.51%** 4	**19.15%** 9	**48.94%** 23	**23.40%** 11	47
soil data	**19.15%** 9	**34.04%** 16	**29.79%** 14	**17.02%** 8	47
anthropological and social data	**42.55%** 20	**23.40%** 11	**23.40%** 11	**10.64%** 5	47
economic data	**36.17%** 17	**36.17%** 17	**23.40%** 11	**4.26%** 2	47

#	Other (please specify)	Date
1	Ontologies	6/28/2015 7:45 PM
2	Actually we don't use the data. We only bring information about the data through metadata.	6/16/2015 11:09 AM
3	agronomlc traits, funtional traits	6/13/2015 9:15 AM
4	historical data; agronomic evaluation data (evaluations); other phenotypic data (chemical etc.)	6/12/2015 6:28 PM

Q8 Do you have problems accessing the types of data you need?

Answered: 47 Skipped: 5



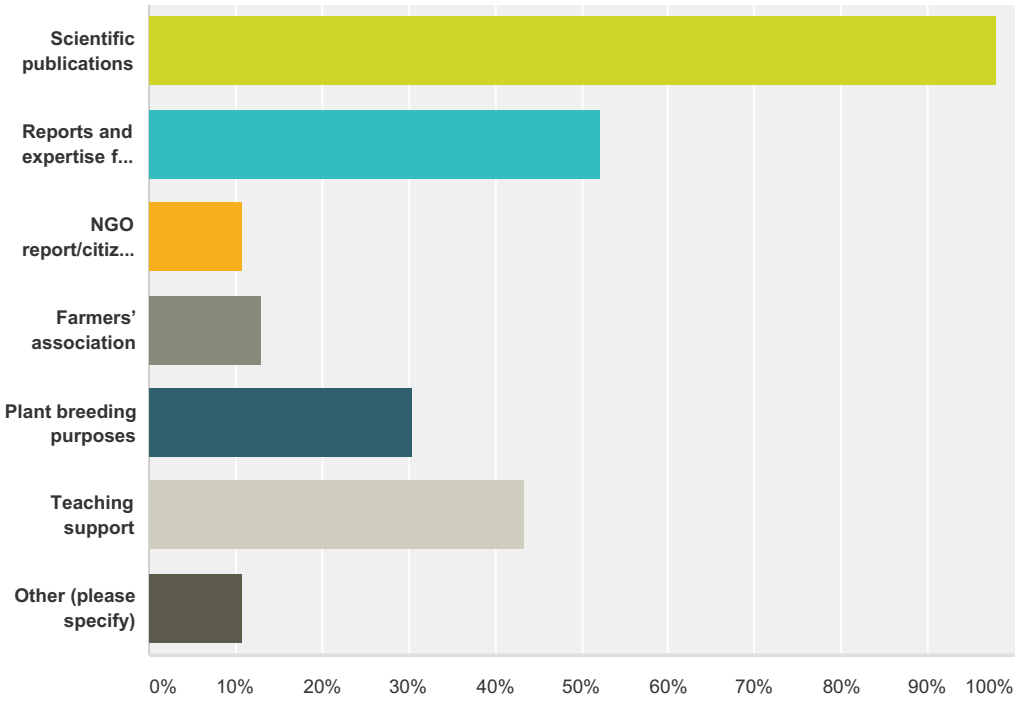
Answer Choices	Responses	
Yes	**61.70%**	29
No	**38.30%**	18
Total		**47**

#	If yes, please explain which type of data, which coverage (temporal, spatial), and why you cannot access it?	Date
1	In particular for the morphological, trait, and genetic data, they lack or they are not clearly summarised in a way that could be possible to compare between different occurrences	8/30/2015 4:11 PM
2	Data is difficult to obtain due to legal and practical constraints.	7/14/2015 2:31 PM
3	duplicata of herbarium should be excluded to avoid false estimation of abundance in a local occurrence.	7/6/2015 11:27 AM
4	lack of information on distribution and phenology of wild plants	7/4/2015 7:10 AM
5	Occurrence, genetic, environmental, climatic, soil and economic data for some countries (e.g. African countries) and regions (e.g. SADC region, Middle East),	7/2/2015 4:45 PM
6	Data need to be geolocalized and it not always the case	6/30/2015 12:26 PM
7	Data is often not assigned by genotype, so for many wild and landrace species, there is no genotype-line trait data link	6/30/2015 11:16 AM
8	Trait data together with environmental data (GxE) are not available for many smaller crops	6/30/2015 9:31 AM
9	Inconsisten sources with no metadata	6/30/2015 7:02 AM
10	Little data on intra-specific diversity available	6/29/2015 5:38 PM
11	local weather data, genetic diversity	6/29/2015 3:18 AM
12	Acessing the germplasm associated with the data in publication	6/28/2015 7:45 PM
13	Occurrence data - very often not online, and when online, difficult to query and access easily	6/26/2015 9:51 PM
14	Large size	6/26/2015 3:21 PM
15	climatic, soil and environmental data is difficult to come by, especially when looking for site-specific data	6/23/2015 4:24 AM
16	Characterization data are not available for many plant species, especially for neglected and underutilised species.	6/22/2015 7:18 PM
17	Problem of internet connection	6/18/2015 8:58 PM
18	the data formats do not match	6/16/2015 10:06 PM
19	As we don't access data, we don't face this problem. But we know we'll have problems with data linked to commercially used genetic resources because of industrial secrecy.	6/16/2015 11:09 AM

20	Global and national trait and genotypic data access are very restricted, global solar radiation data are not available, some genebank occurrence data are not accesible at national scales	6/15/2015 12:57 PM
21	low quality of data	6/15/2015 11:00 AM
22	Problems with accessing ocurrence data with detailed MCPD	6/15/2015 9:55 AM
23	the phenotypic data ar difficult to access; infraspecific taxonomy and diversity	6/13/2015 9:15 AM
24	Inconsistent taxonomy; poor documentatio of characterization data	6/12/2015 6:28 PM
25	The passport data we have at CIP are the most complete ones in terms of potato diversity, however they are relatively old and distribution data, indicating the frequency/abundance of potato landrace diversity showing the current status in situ is not available. We are working at community/field level to generate our own baseline.	6/12/2015 6:00 PM
26	Not a big problem, but often depend on colleagues' assistance to access spatial data needed (beacuse I'm not an IT expert)	6/12/2015 5:54 PM
27	Global digitised passport data is often lacking	6/12/2015 5:50 PM

Q9 What kind of information product do you produce and for whom ?

Answered: 46 Skipped: 6

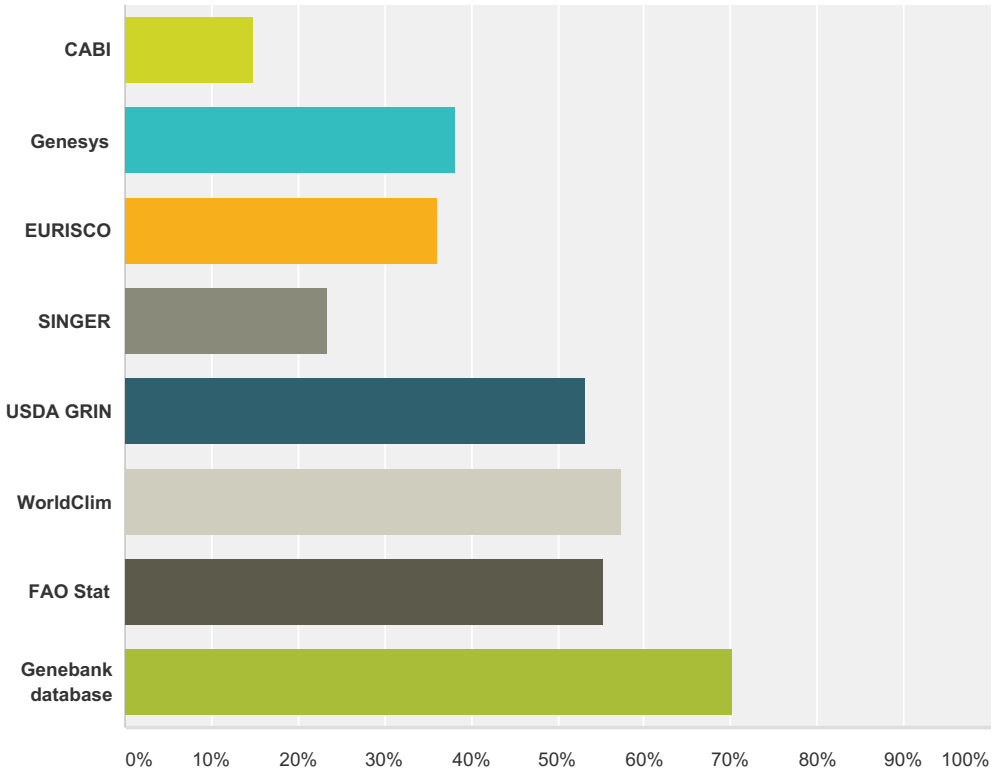


Answer Choices	Responses	
Scientific publications	**97.83%**	45
Reports and expertise for policy makers	**52.17%**	24
NGO report/citizen association	**10.87%**	5
Farmers' association	**13.04%**	6
Plant breeding purposes	**30.43%**	14
Teaching support	**43.48%**	20
Other (please specify)	**10.87%**	5
Total Respondents: 46		

#	Other (please specify)	Date
1	Genesys	7/14/2015 2:31 PM
2	National strategic action plans for the conservation of plant genetic resources	7/2/2015 4:45 PM
3	Plant genetic resources evaluation and characterisation	6/30/2015 9:31 AM
4	Technology and practical applications	6/15/2015 12:57 PM
5	Genebank database data	6/12/2015 6:28 PM

Q10 What are your main sources of data (databases, portal, catalogs / inventories)?

Answered: 47 Skipped: 5



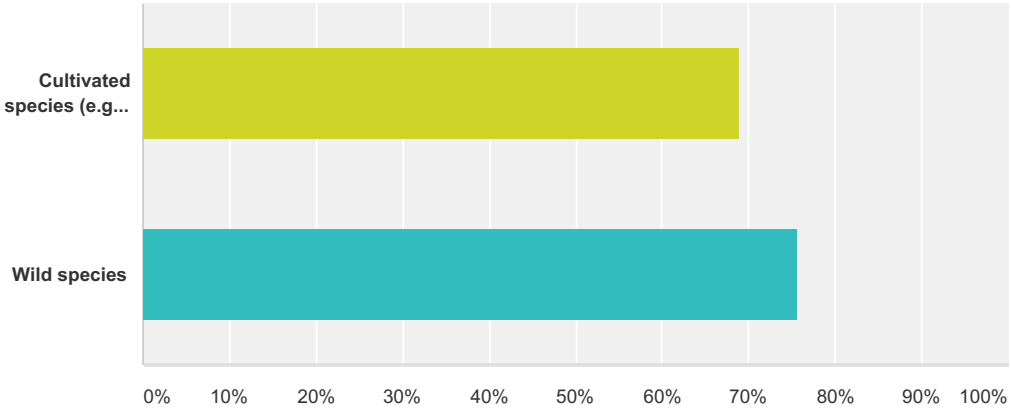
Answer Choices	Responses	
CABI	**14.89%**	7
Genesys	**38.30%**	18
EURISCO	**36.17%**	17
SINGER	**23.40%**	11
USDA GRIN	**53.19%**	25
WorldClim	**57.45%**	27
FAO Stat	**55.32%**	26
Genebank database	**70.21%**	33
Total Respondents: 47		

#	Other (please specify) and examples	Date
1	FAO WIEWS	7/20/2015 10:42 AM
2	GBIF	7/19/2015 9:23 AM
3	CRIA	7/6/2015 11:27 AM
4	Web of Science and scientific publications repositories	7/4/2015 7:10 AM
5	GBIF; Mansfeld database of cultivated plants	6/30/2015 12:32 PM
6	GBIF for distribution and occurrence data	6/30/2015 9:31 AM
7	Plant Ontology, Crop Ontology, iPlant, EBI	6/28/2015 7:45 PM
8	World Checklist of Species ILDIS Tropicos IPNI PlantList Harlan and de Wet Inventory	6/26/2015 6:01 PM

9	Science Direct, Google Scholar,	6/23/2015 4:24 AM
10	GBIF	6/18/2015 8:58 PM
11	Own gathering	6/18/2015 4:44 PM
12	Regional and National level data sources	6/16/2015 10:06 PM
13	We only use metadata. Some come from genebanks, some by other research or private intiatives structured as 'research observatories'.	6/16/2015 11:09 AM
14	The Plant List Species link Tropicos (Missouri) GBIF	6/15/2015 1:40 PM
15	National inventories	6/15/2015 12:57 PM
16	National inventories; Bioversity collecting mission database; GBIF, AfSys (soil data); Crop ontology; Plant ontology	6/13/2015 9:15 AM
17	WIEWS (!!); GRIN-CA; Mansfeld Encyclopedia and IPK genebank information system; SESTO	6/12/2015 6:28 PM
18	Field surveys and distribution maps and historical registries in publications. This is also important for timeline establishments. Further, expert opinion and survey.	6/12/2015 6:00 PM
19	GBIF	6/12/2015 5:50 PM

Q11 What type of plant agrobiodiversity data have you obtained from GBIF during the last year? Multiple option question

Answered: 45 Skipped: 7

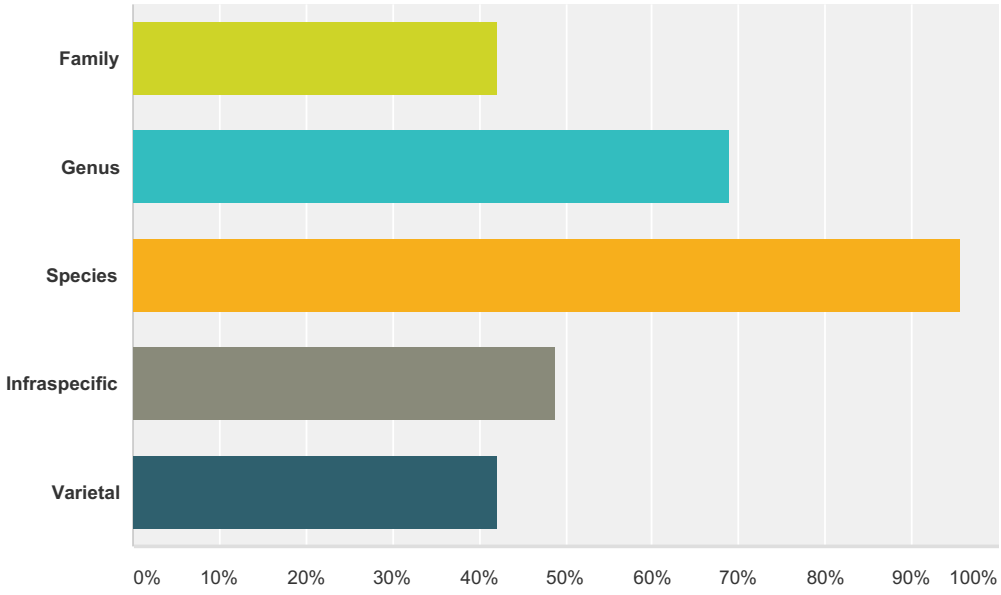


Answer Choices	Responses	
Cultivated species (e.g. varieties, landraces)	**68.89%**	31
Wild species	**75.56%**	34
Total Respondents: 45		

#	Other (please specify)	Date
1	none. The species I was looking for are not available	8/30/2015 4:17 PM
2	None	6/30/2015 11:19 AM
3	None	6/30/2015 7:04 AM
4	None	6/16/2015 11:13 AM
5	access of occurences per country as basleine data	6/13/2015 11:22 AM
6	Nothing from GBIF during last year.	6/12/2015 6:32 PM
7	we use first our own data to identify diversity hotspots of potato	6/12/2015 6:00 PM

Q12 Please select the data precision requirements that apply to your research in terms of taxonomic determination. Multiple choice question

Answered: 45 Skipped: 7

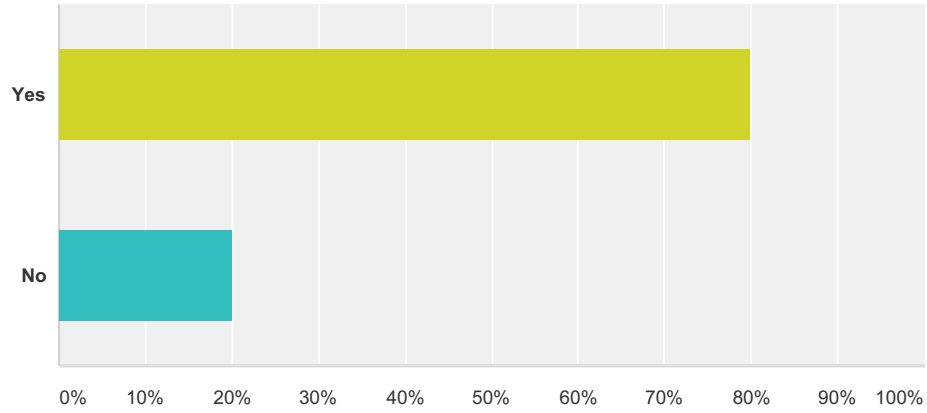


Answer Choices	Responses	
Family	**42.22%**	19
Genus	**68.89%**	31
Species	**95.56%**	43
Infraspecific	**48.89%**	22
Varietal	**42.22%**	19
Total Respondents: 45		

#	Other (please specify)	Date
1	Genetically identified individuals are needed for breeding and assessment work	6/30/2015 11:19 AM
2	year of collection, herbarium and person who determined the sample	6/29/2015 5:40 PM
3	We don't do research	6/16/2015 11:13 AM
4	Varietal is part of infraspecific	6/12/2015 6:32 PM
5	in terms of potato species it is not yet really clear which taxonomic philosophy is the right one	6/12/2015 6:00 PM

Q13 Is the taxonomic determination linked to an international nomenclature system (i.e. the Catalog of Life) relevant for your research?

Answered: 45 Skipped: 7

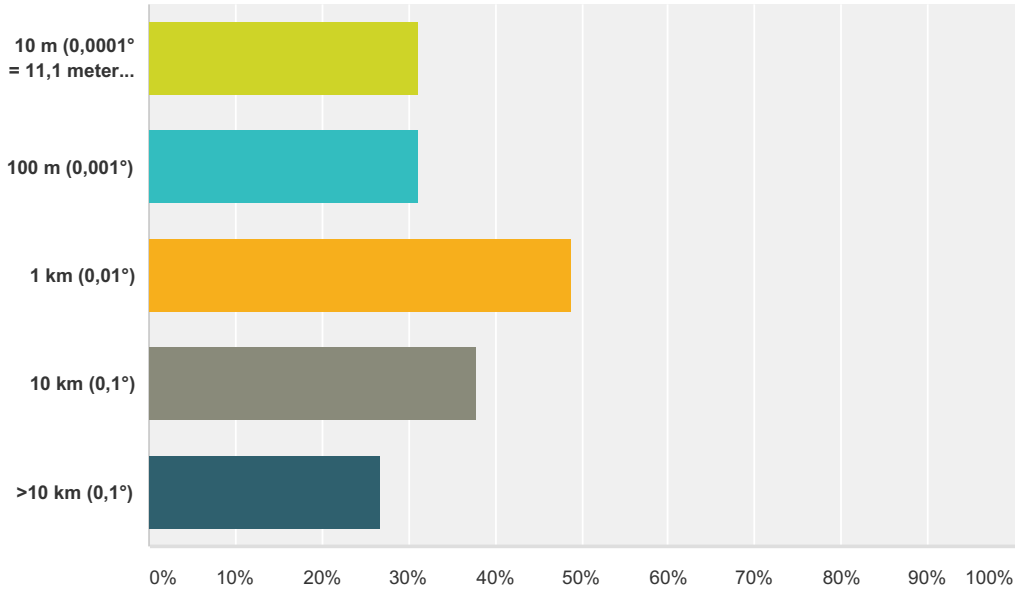


Answer Choices	Responses	
Yes	**80.00%**	36
No	**20.00%**	9
Total		**45**

#	If no, please explain and provide details of the nomenclature system relevant for your research	Date
1	local names, ethnovarieties	7/6/2015 11:29 AM
2	It can be important but in some countries we just have to use the nomenclature system that is used there	7/2/2015 4:48 PM
3	some species classification has to be used with caution	6/30/2015 12:28 PM
4	Beyond an agreed taxon name, I would especially consider the original determination by the observer / collector of importance and the taxonomic system used for this	6/30/2015 9:37 AM
5	GRIN or the Plant List	6/29/2015 5:40 PM
6	Our metadata will allow to precise which nomenclature system is used by our partners	6/16/2015 11:13 AM
7	Not all. We use infraspecific taxonomy, which sometime has no international nomenclature systems	6/15/2015 8:25 AM
8	GRIN-USDA is a reference for crops	6/13/2015 11:22 AM
9	GRIN Taxonomy is the standrad we use for practical reasons.	6/12/2015 6:32 PM

Q14 Please select the data precision requirements that apply to your research in terms of geographic coordinates. Multiple choice question.

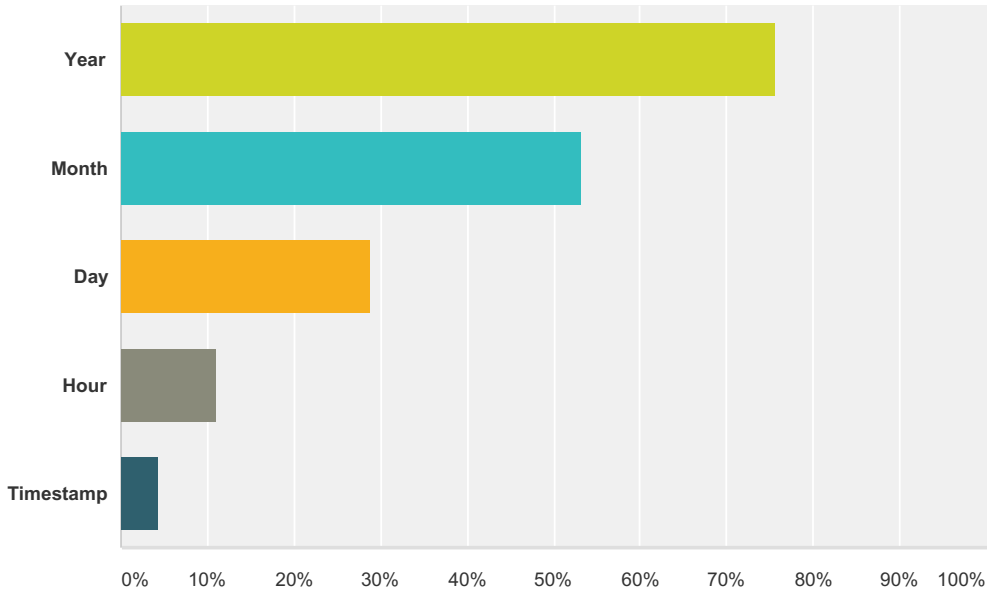
Answered: 45 Skipped: 7



Answer Choices	Responses	
10 m (0,0001° = 11,1 meter at equator)	**31.11%**	14
100 m (0,001°)	**31.11%**	14
1 km (0,01°)	**48.89%**	22
10 km (0,1°)	**37.78%**	17
>10 km (0,1°)	**26.67%**	12
Total Respondents: 45		

Q15 Please select the data precision requirements that apply to your research in terms of time. Multiple choice question

Answered: 45 Skipped: 7

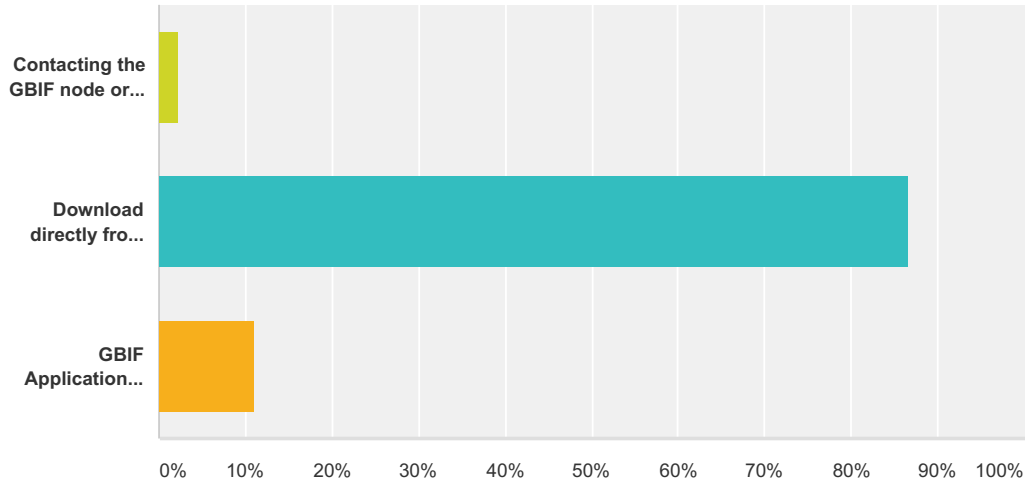


Answer Choices	Responses	
Year	**75.56%**	34
Month	**53.33%**	24
Day	**28.89%**	13
Hour	**11.11%**	5
Timestamp	**4.44%**	2
Total Respondents: 45		

#	Other (please specify)	Date
1	it depends on which data. for local climatic data it can be per hour for crops.	6/13/2015 11:22 AM

Q16 When using GBIF for obtaining data records, what option do you use the most?

Answered: 45 Skipped: 7

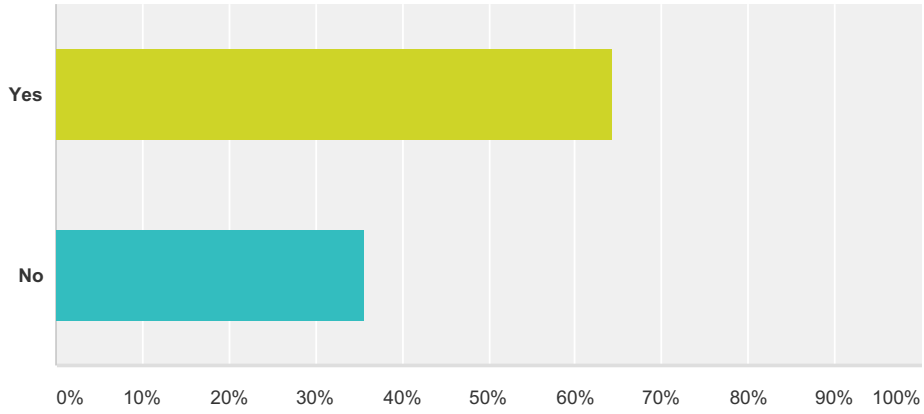


Answer Choices	Responses	
Contacting the GBIF node or secretariat	**2.22%**	1
Download directly from GBIF web portal (www.gbif.org)	**86.67%**	39
GBIF Application Programme Interface (API) webservices (http://api.gbif.org/v1/)	**11.11%**	5
Total		**45**

#	Other apps or software, please specify	Date
1	none	8/30/2015 4:17 PM
2	R (rgbif)	6/30/2015 12:35 PM
3	R package	6/30/2015 12:28 PM
4	N/A so far	6/30/2015 11:19 AM
5	I do not requested GBIF data (so it is an hypothetical response)	6/18/2015 4:47 PM
6	We don't use data.	6/16/2015 11:13 AM
7	in R, dismo package, gbif function	6/15/2015 1:00 PM
8	also used the GBIF helpdesk	6/13/2015 11:22 AM
9	GBIF was not used by me.	6/12/2015 6:32 PM
10	genesys, worldclim	6/12/2015 5:57 PM
11	rgbif R package	6/12/2015 5:14 PM

Q17 Do you combine occurrence or taxon records retrieved from GBIF to data from other data sources to make a relevant data set? If yes, select one or more options below, if no, skip next question

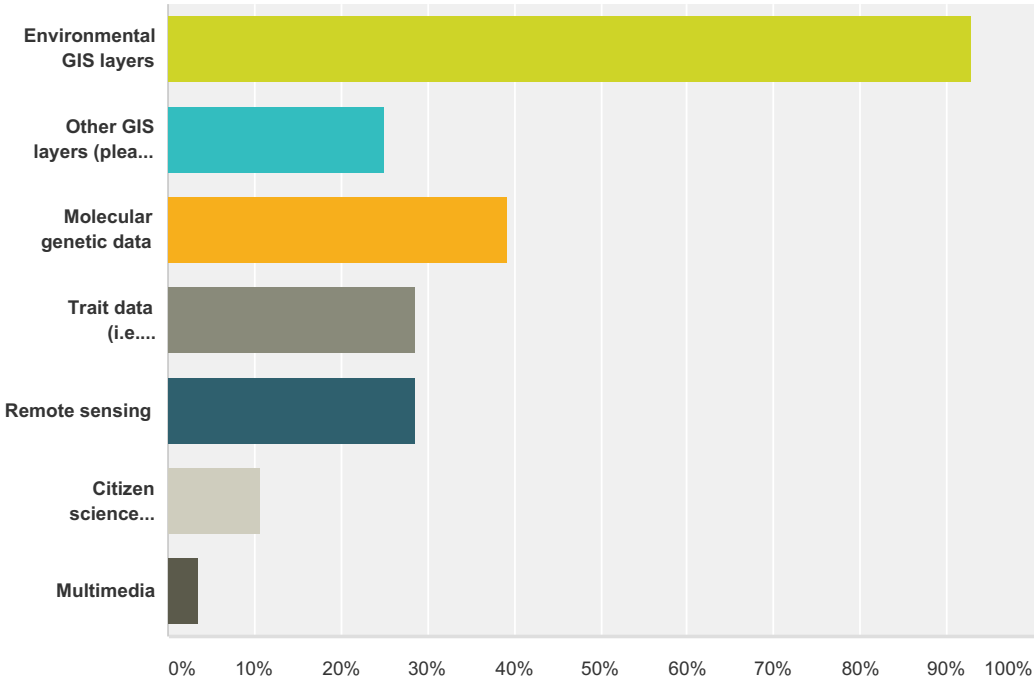
Answered: 42 Skipped: 10



Answer Choices	Responses	
Yes	**64.29%**	27
No	**35.71%**	15
Total		**42**

Q18 You responded that you combine occurrence or taxon records retrieved from GBIF with data from other data sources to make a relevant data set.Please select those that apply to your case

Answered: 28 Skipped: 24

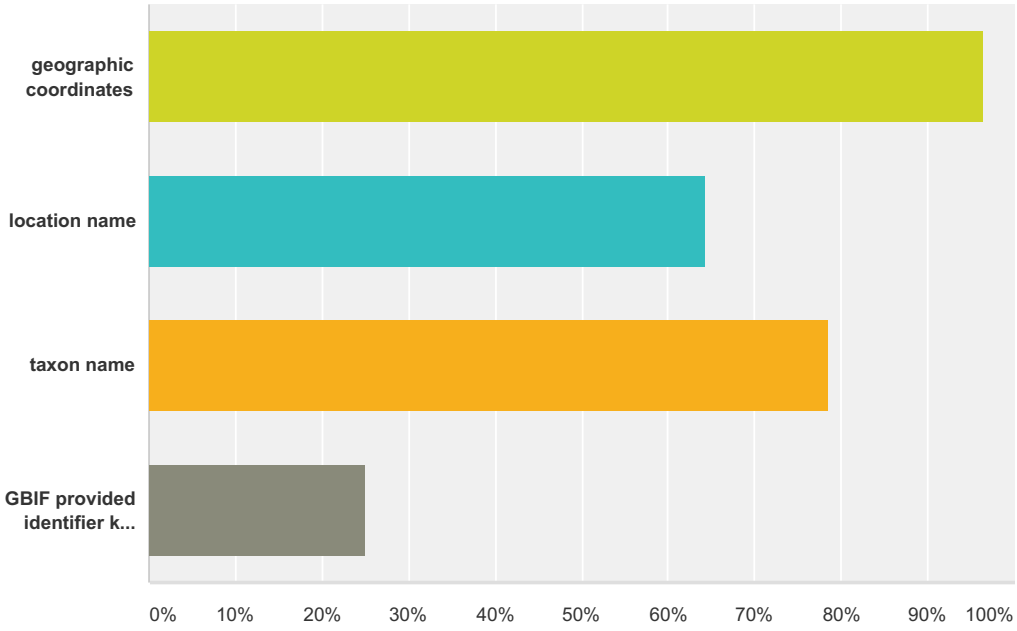


Answer Choices	Responses	
Environmental GIS layers	**92.86%**	26
Other GIS layers (please specify, i.e: census)	**25.00%**	7
Molecular genetic data	**39.29%**	11
Trait data (i.e. characterization and evaluation data)	**28.57%**	8
Remote sensing	**28.57%**	8
Citizen science observations	**10.71%**	3
Multimedia	**3.57%**	1
Total Respondents: 28		

#	Other (please specify)	Date
1	archeologic data, ethnografic data, linguistic data	7/6/2015 11:31 AM
2	Occurrence data from other sources not covered by GBIF (genebanks, herbaria, bibliographic references, personnal communications, field observations)	7/2/2015 4:51 PM
3	administrative geographical units (NUTS, LAU)	6/30/2015 9:40 AM
4	occurrence data from other sources	6/26/2015 9:54 PM
5	threat layers, other private occurrence records from various organisations/people	6/26/2015 6:04 PM
6	Demographic, land use, economic data	6/15/2015 1:02 PM
7	Other occurrence data, GPS taken	6/15/2015 9:57 AM

Q19 What element of information is necessary to connect occurrence or taxon records retrieved from GBIF to data listed above?

Answered: 28 Skipped: 24

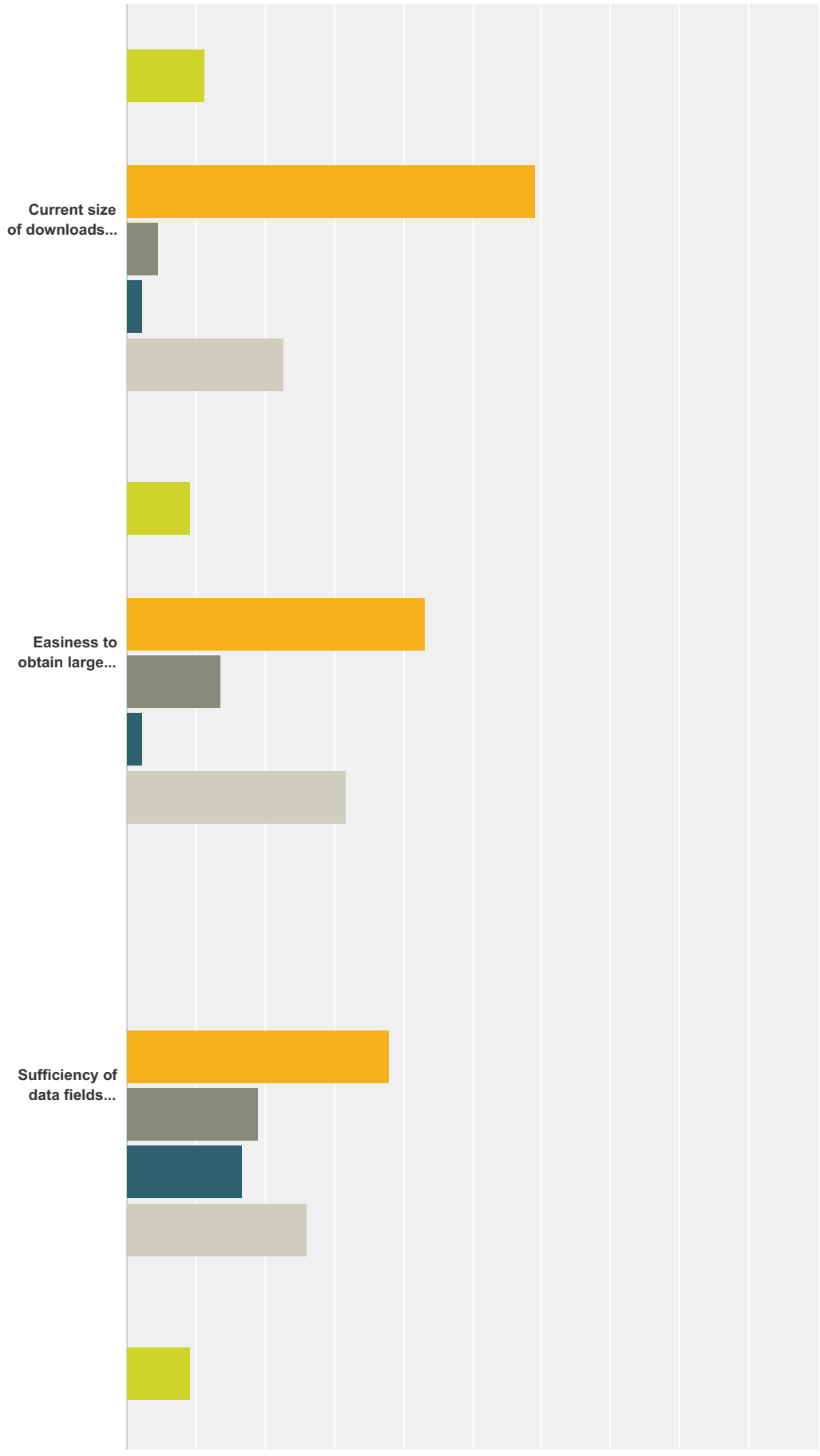


Answer Choices	Responses	
geographic coordinates	**96.43%**	27
location name	**64.29%**	18
taxon name	**78.57%**	22
GBIF provided identifier keys (taxonKey, occurrenceID, etc)	**25.00%**	7
Total Respondents: 28		

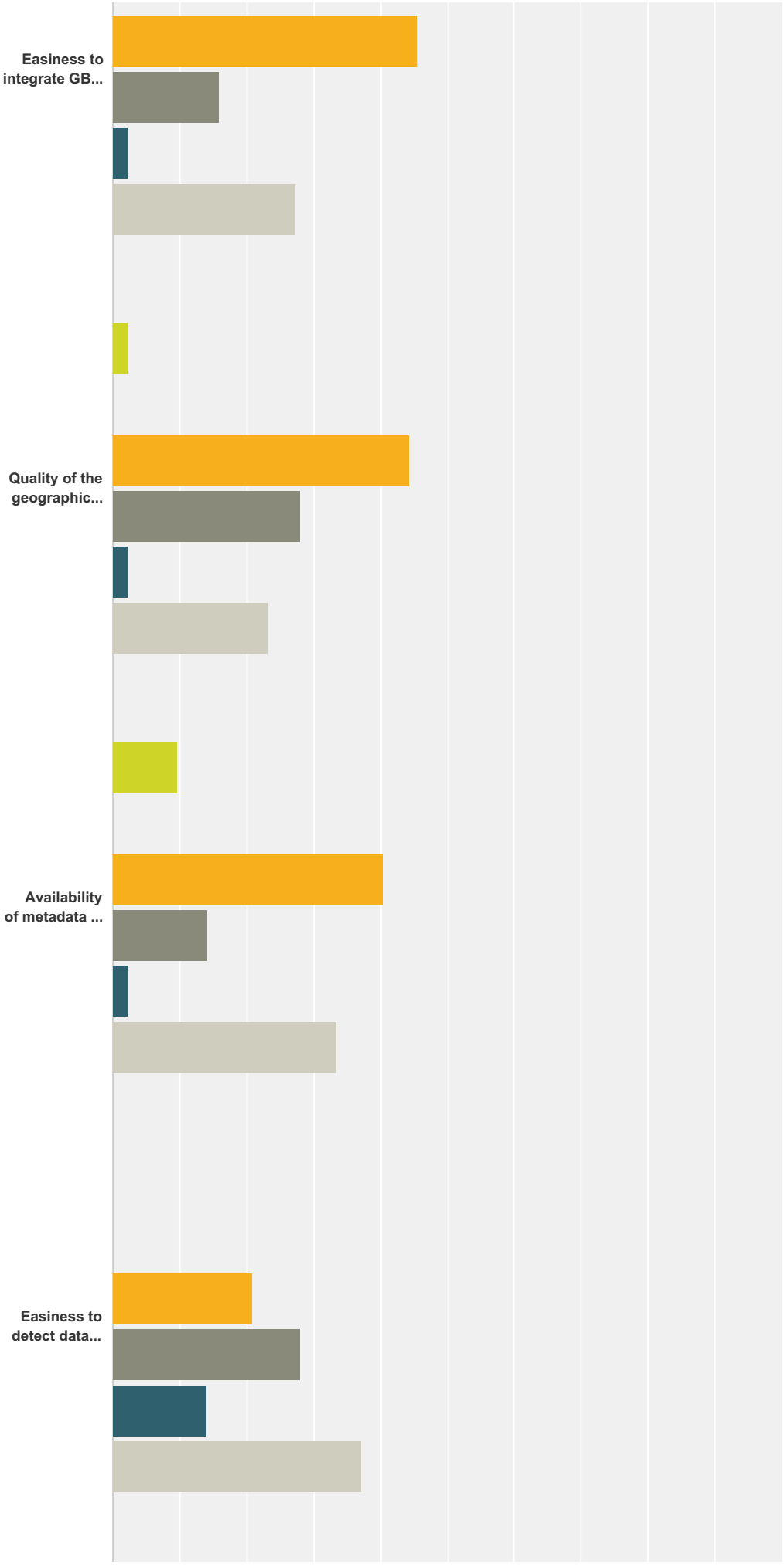
#	Other (please specify)	Date
1	village	7/6/2015 11:31 AM
2	And date of collecting so we detect duplicate records	7/2/2015 4:51 PM
3	Genebank accession identifiers, collecting numbers	6/30/2015 9:40 AM
4	Other MCPD like sample status and coordinate uncertainty	6/15/2015 9:57 AM

Q20 Rate your satisfaction with accessing and using GBIF mediated data

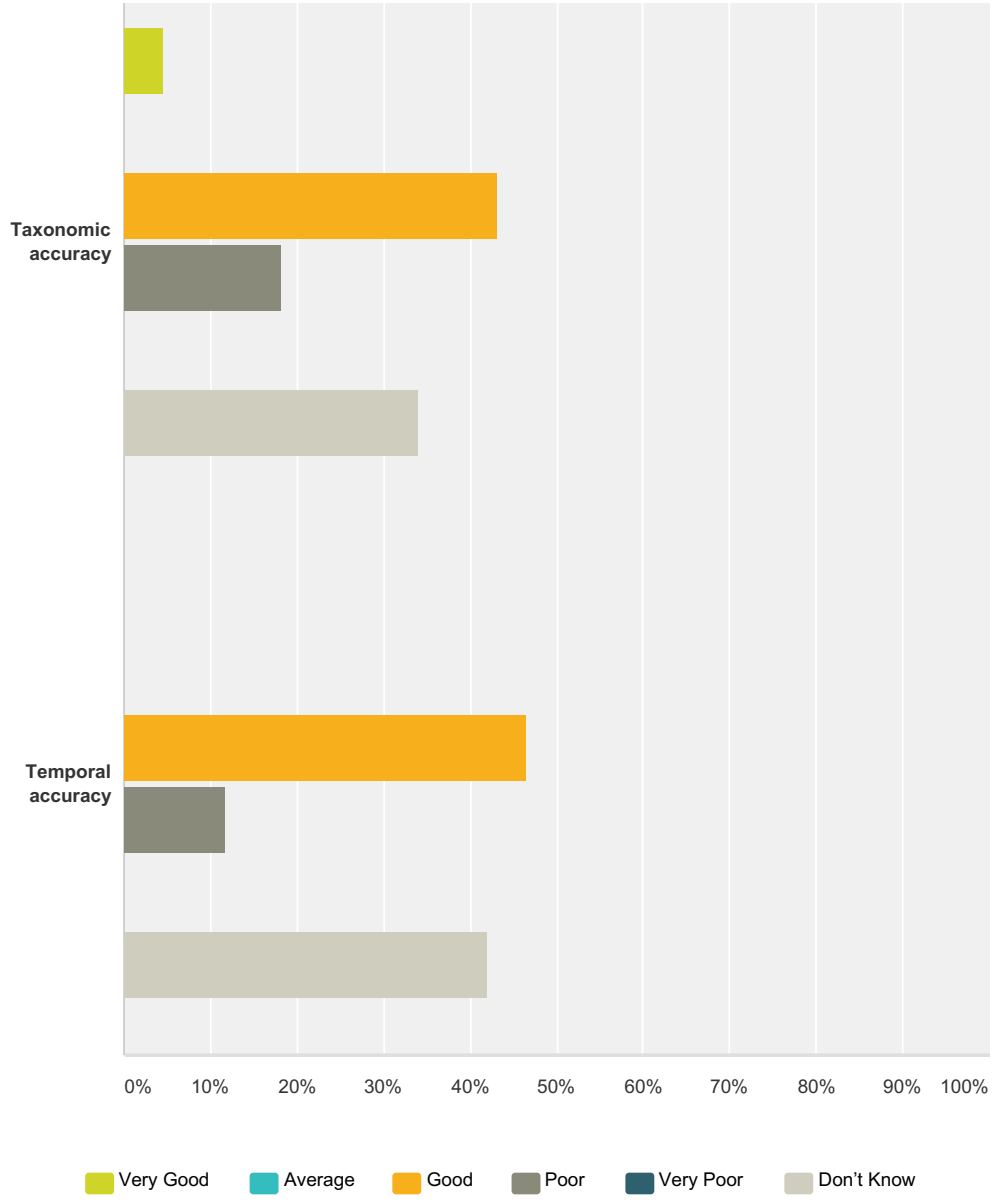
Answered: 44 Skipped: 8



GBIF and Bioversity International: Data Fitness for Use in Agrobiodiversity 2015



GBIF and Bioversity International: Data Fitness for Use in Agrobiodiversity 2015



	Very Good	Average	Good	Poor	Very Poor	Don't Know	Total
Current size of downloads enabled through the portal	**11.36%** 5	**0.00%** 0	**59.09%** 26	**4.55%** 2	**2.27%** 1	**22.73%** 10	44
Easiness to obtain large amounts of occurrence records (> 100.000 records)	**9.09%** 4	**0.00%** 0	**43.18%** 19	**13.64%** 6	**2.27%** 1	**31.82%** 14	44
Sufficiency of data fields available for agrobiodiversity related data	**0.00%** 0	**0.00%** 0	**38.10%** 16	**19.05%** 8	**16.67%** 7	**26.19%** 11	42
Easiness to integrate GBIF data with information from other different databases	**9.09%** 4	**0.00%** 0	**45.45%** 20	**15.91%** 7	**2.27%** 1	**27.27%** 12	44
Quality of the geographic coordinates provided with the occurrence records	**2.33%** 1	**0.00%** 0	**44.19%** 19	**27.91%** 12	**2.33%** 1	**23.26%** 10	43
Availability of metadata for citing original data sources provided through GBIF	**9.52%** 4	**0.00%** 0	**40.48%** 17	**14.29%** 6	**2.38%** 1	**33.33%** 14	42
Easiness to detect data duplicates	**0.00%** 0	**0.00%** 0	**20.93%** 9	**27.91%** 12	**13.95%** 6	**37.21%** 16	43
Taxonomic accuracy	**4.55%** 2	**0.00%** 0	**43.18%** 19	**18.18%** 8	**0.00%** 0	**34.09%** 15	44

Temporal accuracy	**0.00%** 0	**0.00%** 0	**46.51%** 20	**11.63%** 5	**0.00%** 0	**41.86%** 18	43

#	Comments or remarks	Date
1	My remarks are mostly based on my experience working with GBIF data in 2012 and thus may not reflect the current quality of the data provided. Mostly I remarked problems link to taxonomic resolution and the difficulty, therefore, to download exhaustive data for a genus or species for which the taxonomic data was not updated. Other problems I have come across are the lack of geographic coordinates for several accessions.	9/10/2015 3:13 PM
2	Have not used GBIF data downloads for some time; cannot provide rates	7/20/2015 10:46 AM
3	There are many specimens that do not have coordinates at all	7/2/2015 5:05 PM
4	More data properties could of course be indexed by GBIF, however, this will obviously never result in a complete set - and approaches to link between data sources with additional information on the specimens (or taxonomic entities) provides more effective and realistic solutions. When established persistent identifiers for the entities of interest to GBIF are lacking, support or at least strong pressure from GBIF to establish such identifiers is useful. Metadata for citing original data sources may be lacking due to a backlog at the nodes for providing these information. Relatively recent developments have made stronger focus on EML metadata and more precise classification of the roles of institutes and people, however, most of the data sets published through our national node regrettably lack complete metadata.	6/30/2015 12:48 PM
5	When I last used it, there were several issues with duplicate genebank data and wrong or confounded genebank accession numbers. Correct linking to genebank accessions (genebank identifier - accession number) I would consider of high importance. Duplicate import of accession data (e.g. from EURISCO + genebanks) should be avoided.	6/30/2015 9:49 AM
6	Geographic location data by default requieres quality check. Taxonomic accuracy of less common plants can be estimated knowing the person who determined the sample and the herbarium where it is stored. Some species and genera are more easy to determine than others. so maybey GBIF can provide an index of taxonomic accuracy on the basis how difficult a genus or species-complex is for botanical identification	6/29/2015 5:44 PM
7	Very difficult to know how good quality the co-ordinate data is. Or how the co-ordinates were calculated.	6/26/2015 6:07 PM
8	Keep up the good work! More useful information on underutilised crop species would be appreciated.	6/23/2015 4:35 AM
9	Multiple species query is difficult and time consuming	6/15/2015 9:59 AM
10	Data are easy to integarte once you have cleaned the sets for taxonomy and georeferences.	6/13/2015 11:26 AM
11	Based on earlier experience, it was very useful to inspect classical local floras.	6/12/2015 6:34 PM
12	as we are working on a holistic baseline for long-term monitoring of potato diversity in its center of origin. We first need real time data from farmers fields. This is now done in identified diversity hotspots. Data from GBIF have not been included yet but comparisons might be carried out in future.	6/12/2015 6:00 PM
13	I have not used GBIF for research. Just played with it...	6/12/2015 5:58 PM

Q21 Please mention the software, services and/or tools that you use to check the taxonomy of the data you obtain from GBIF. Max 200 characters

Answered: 29 Skipped: 23

#	Responses	Date
1	I have toyed with many online taxonomic reconciliation services and I use now mostly The Plant List (www.theplantlist.org/) which I found to be the most user friendly and the most complete.	9/10/2015 3:18 PM
2	Maxent software, R statistical software	7/14/2015 6:49 PM
3	Usually local floras	7/2/2015 5:08 PM
4	http://gbif.no/datasets Artsnavnebasen (the Norwegian taxon checklist) R	6/30/2015 12:53 PM
5	We back to collection report and check each accession. Then manually. So, we need to know all information concerning herbarium, collector and so on	6/30/2015 12:34 PM
6	N/A so far	6/30/2015 11:20 AM
7	Taxonomic databases like GRIN Tax, genebank databases, Euro+Med Plantbase	6/30/2015 9:54 AM
8	By hand with GRIN	6/29/2015 5:45 PM
9	apis to GRIN, TNRS, and R package TaxStand	6/26/2015 9:56 PM
10	GRIN PlantList	6/26/2015 6:08 PM
11	GRIN, TaxonStand, ThePlantList	6/26/2015 3:30 PM
12	Expert opinion	6/26/2015 3:26 PM
13	N/A - have yet to fully explore the datasets from GBIF	6/23/2015 4:37 AM
14	The International Plant Names Index, The Plant List.	6/22/2015 10:23 PM
15	I use functions of Excell; I also use Catalogue of Life to verify the valid names of species	6/18/2015 9:15 PM
16	the plant list	6/16/2015 12:15 AM
17	Flora of Benin	6/15/2015 5:55 PM
18	Tropicos website The Plant List website IPNI website Lista da Flora do Brasil website	6/15/2015 2:12 PM
19	CJB Tropicos IPNI	6/15/2015 1:52 PM
20	GRIN taxonomy	6/15/2015 1:07 PM
21	custom build	6/15/2015 11:06 AM
22	USDA GRIN Taxonomy for Plants	6/15/2015 10:01 AM
23	Genetic Resources Handbook	6/13/2015 1:22 PM
24	Taxonomic reconciliation withTRNS on iPlantCollaborative adn the Plant List (http://www.theplantlist.org/); http://tnrs.iplantcollaborative.org/; GRIN Taxonomy fromUSDA-ARS	6/13/2015 11:32 AM
25	GRIN, Mansfeld Database	6/12/2015 6:36 PM
26	n.a.	6/12/2015 6:00 PM
27	We check the national floras when appropriate (e.g., Flora Iberica) GRIN Taxonomy	6/12/2015 5:57 PM
28	Plant List GRIN Taxonomy Flora Europeae	6/12/2015 5:56 PM
29	For taxonomy, we use taxstand, tnrs and GRIN	6/12/2015 5:34 PM

Q22 Please mention the software, services and/or tools that you use to assess the accuracy of geographic coordinates of the data you obtain from GBIF (max 200 characters)

Answered: 28 Skipped: 24

#	Responses	Date
1	I work with both ArcMap and DIVA-GIS. I have also been using Google Maps extensively to try to locate accessions for which no coordinates was available but only a mention of a locality.	9/10/2015 3:18 PM
2	Google earth, QGIS, ArcGIS	7/14/2015 6:49 PM
3	Leve of geographic precision based on what has been suggested in Maxted N, Magos Brehm J and Kell S (2013) Resource book for the preparation of national plans for conservation of crop wild relatives and landraces. Rome, FAO. Available from: http://www.fao.org/fileadmin/templates/agphome/documents/PGR/PubPGR/ResourceBook/TEXT_ALL_2511.pdf. Also CAPFITOGEN tools.	7/2/2015 5:08 PM
4	http://gbif.no/datasets	6/30/2015 12:53 PM
5	we use collect informations (when available !) to cross information (town, road, city, region...)	6/30/2015 12:34 PM
6	N/A so far	6/30/2015 11:20 AM
7	Just mapping with Google Maps and comparing with locality strings and looking on satellite image	6/30/2015 9:54 AM
8	R to idneitfy climatic outliers and consistency of administrative units	6/29/2015 5:45 PM
9	google geocoder, geolocate	6/26/2015 9:56 PM
10	Geolocate, Google Map API	6/26/2015 3:30 PM
11	DIVA-GIS, R packages	6/26/2015 3:26 PM
12	N/A - have yet to fully explore the datasets from GBIF	6/23/2015 4:37 AM
13	QGIS	6/22/2015 10:23 PM
14	I use QGIS to display the occurrence data on Benin layer so that to find out wether the occurrence data are within the limits of Benin	6/18/2015 9:15 PM
15	Arc GIS	6/16/2015 10:09 PM
16	R	6/16/2015 12:15 AM
17	Don't check coordinates accuracy	6/15/2015 5:55 PM
18	Locality name of the country National Mapping Service	6/15/2015 1:52 PM
19	GEOQUAL (CAPFITOGEN tools) and Google earth	6/15/2015 1:07 PM
20	DIVA-GIS	6/15/2015 11:06 AM
21	Geolocate, BioGeomancer (till the online tool was working), GeoNames, CAPFITOGEN	6/15/2015 10:01 AM
22	Google Maps	6/13/2015 1:22 PM
23	BioGeoBIF http://biodiversity.colorado.edu/bgb/; GEOLocate http://www.museum.tulane.edu/geolocate/	6/13/2015 11:32 AM
24	Google Earth	6/12/2015 6:36 PM
25	n.a.	6/12/2015 6:00 PM
26	Geoqual see 24	6/12/2015 5:57 PM
27	Mapping software	6/12/2015 5:56 PM
28	We use a tool developed in-house in java for assessing the accuracy of coordinates, and re-calculate coordinates when required.	6/12/2015 5:34 PM

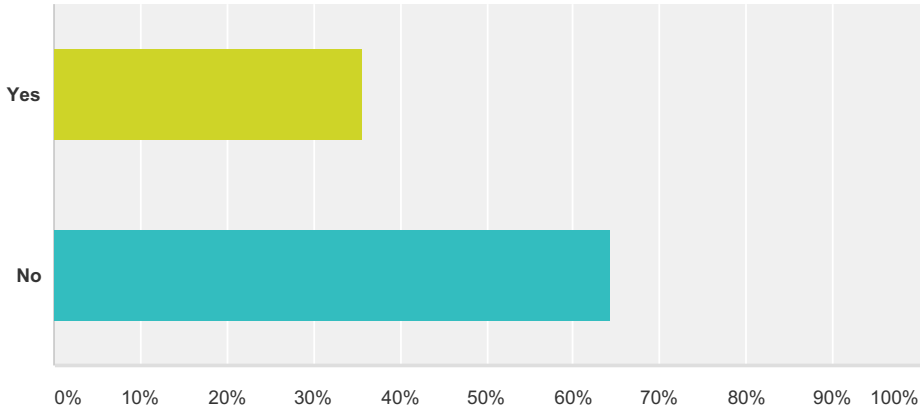
Q23 Please mention the software, services and /or tools that you use to manage your data and associated metadata (max 200 characters)

Answered: 27 Skipped: 25

#	Responses	Date
1	Mostly I use Excel to store and format the data, and R / ArcMap to analyse it.	9/10/2015 3:18 PM
2	SQL and R	7/14/2015 6:49 PM
3	Excel, Access	7/2/2015 5:08 PM
4	http://gbif.no/datasets Global Registry of Biological Repositories (GRBio.org) FAO WIEWS institute database	6/30/2015 12:53 PM
5	R	6/30/2015 12:34 PM
6	N/A so far	6/30/2015 11:20 AM
7	Own databases: e.g. AEGRO, European Avena and Beta Databases	6/30/2015 9:54 AM
8	SQL	6/26/2015 9:56 PM
9	MySql Python R ArcMap QGIS	6/26/2015 6:08 PM
10	R	6/26/2015 3:26 PM
11	N/A - have yet to fully explore the datasets from GBIF	6/23/2015 4:37 AM
12	Excel, Acces	6/22/2015 10:23 PM
13	QGIS,Excel	6/18/2015 9:15 PM
14	Arc GIS	6/16/2015 10:09 PM
15	R	6/16/2015 12:15 AM
16	Manually	6/15/2015 5:55 PM
17	ESRI Mapinfo	6/15/2015 1:52 PM
18	CAPFITOGEN tools	6/15/2015 1:07 PM
19	custom build	6/15/2015 11:06 AM
20	ArcGIS personal geodatabase, Access DB	6/15/2015 10:01 AM
21	Software R,	6/13/2015 1:22 PM
22	for traits,we use Crop Ontology (www.cropontology.org); Plant Ontology	6/13/2015 11:32 AM
23	GRIN, Excel	6/12/2015 6:36 PM
24	n.a.	6/12/2015 6:00 PM
25	Microsoft Excel and Access, ArcGIS, R software environment	6/12/2015 5:57 PM
26	Access	6/12/2015 5:56 PM
27	All occurrence data and metadata associated are stored in a SQL database.	6/12/2015 5:34 PM

Q24 Have you or your team developed tools for using or preprocessing (e.g. check taxonomy, cleaning coordinates, etc.) agrobiodiversity data obtained from GBIF or from other sources?

Answered: 42 Skipped: 10



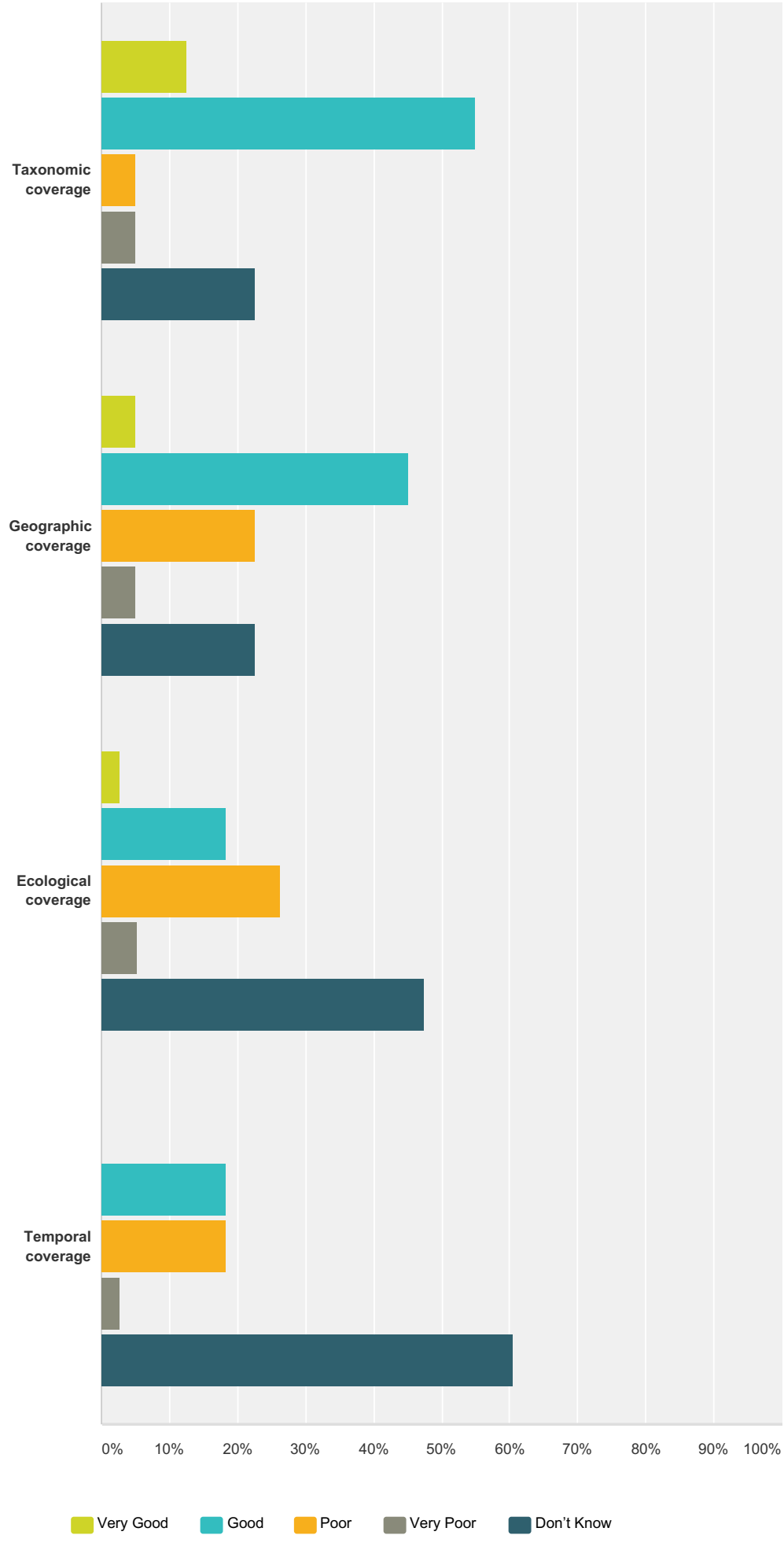
Answer Choices	Responses	
Yes	**35.71%**	15
No	**64.29%**	27
Total		**42**

#	If yes, please specify where information can be found (e.g. URL), programming language used, and eventual licensing or the tool.	Date
1	We used to use the "Taxonomic Nomenclature Checker" http://pgrdoc.bioversity.cgiar.org/taxcheck/grin/index.html, which has not been available for some time; it is said that there are problems in maintaining it by Bioversity	7/20/2015 10:49 AM
2	https://github.com/CIAT-DAPA/cwr_occurrencesvalidation	7/14/2015 6:49 PM
3	http://gbif.no/datasets (GitHub repository in planning)	6/30/2015 12:53 PM
4	AGAP, DDSE team	6/30/2015 12:34 PM
5	For AEGRO we used an MS Access Tool to integrate it with political administrative units (NUTS, LAU)	6/30/2015 9:54 AM
6	R	6/29/2015 5:45 PM
7	at CIAT now; planning to make available in the future	6/26/2015 9:56 PM
8	R package exsic (CRAN)	6/26/2015 3:26 PM
9	http://www.capfitogen.net/en	6/15/2015 1:07 PM
10	local Excel	6/15/2015 11:06 AM
11	Varoius tools were used by datbase manager.	6/12/2015 6:36 PM
12	http://www.capfitogen.net/en/tools/geoqual/	6/12/2015 5:57 PM
13	A tool for validating the taxonomy and coordinates of occurrence records was developed by our data manager (more info here: https://github.com/CIAT-DAPA/cwr_occurrencesvalidation)	6/12/2015 5:34 PM

Q25 Agrobiodiversity data published in GBIF have an uneven coverage. Please evaluate the coverage of GBIF mediated data with respect to your needs in your own research. Multiple choice question

Answered: 40 Skipped: 12

GBIF and Bioversity International: Data Fitness for Use in Agrobiodiversity 2015



GBIF and Bioversity International: Data Fitness for Use in Agrobiodiversity 2015

	Very Good	Good	Poor	Very Poor	Don't Know	Total
Taxonomic coverage	**12.50%** 5	**55.00%** 22	**5.00%** 2	**5.00%** 2	**22.50%** 9	40
Geographic coverage	**5.00%** 2	**45.00%** 18	**22.50%** 9	**5.00%** 2	**22.50%** 9	40
Ecological coverage	**2.63%** 1	**18.42%** 7	**26.32%** 10	**5.26%** 2	**47.37%** 18	38
Temporal coverage	**0.00%** 0	**18.42%** 7	**18.42%** 7	**2.63%** 1	**60.53%** 23	38

#	Please specify taxonomic coverage, or add other remarks or comments	Date
1	Most of my work with relevance to GBIF focusses on neglected and underutilized crop species, for which there is a well-known lack of data. Besides working on some of these NUS species to contribute to fill in these gaps, I have used GBIF precisely to identify which data are missing	9/10/2015 3:25 PM
2	Intensity of use and management. Some species are not cultivated (sensu strictu) but the traditional human use could affect their distribution.	7/6/2015 11:37 AM
3	Many common taxa as well as many crop wild relatives are usually not covered	7/2/2015 5:13 PM
4	Most of the agrobiodiversity occurrence data available from other accessible sources, are also available from GBIF, even if not all of the relevant data properties are indexed in GBIF.	6/30/2015 1:11 PM
5	Main interests Avena, Beta, Legumes. In Avena there are many invalid taxon names, but probably reflecting original determinations. Parallel taxonomic information system would be needed to translate from various taxonomic systems. I would consider this more important than referring to one "agreed" taxonomy, which probably not really exists.	6/30/2015 10:04 AM
6	data from many taxa availabale	6/29/2015 5:46 PM
7	Difficult to answer these questions; not a taxonomic expert so must in many cases take data at face value	6/26/2015 9:58 PM
8	Obviously some areas are much easy to collect in than others, thus shown by an increased number of observations. Places like China have fewer observations of species which we are interested where we should expect more.	6/26/2015 6:12 PM
9	Don't have comment	6/15/2015 6:02 PM
10	Coverages are biased by the original sources interests (botanical garden, herbaria, seed collections, etc), then as GBIF is the most important compilator, its coverage is the available coverage, even biased	6/15/2015 1:19 PM
11	-	6/15/2015 11:08 AM
12	Infraspeicif level is mising; not much occurences of this level outside those from Genesys.	6/13/2015 11:39 AM
13	I found GRIN was not tailored to the community interested in utilisation of plant genetic resources for food and agriculture, so have not used it in a while.	6/12/2015 6:40 PM
14	I will need to try the GBIF system to give an appropriate feedback for this question. In future this might be happen. For now it is not my priority.	6/12/2015 6:00 PM
15	assessment referred to crop wild relatives in Europe	6/12/2015 6:00 PM

Q26 From your experience, what other relevant agrobiodiversity occurrence datasets (such as collections, particular project data, etc) should be made available through GBIF. Please be specific, but max 200 characters

Answered: 27 Skipped: 25

#	Responses	Date
1	Mentioning plant diseases affecting accessions would be a great addition to GBIF as would allow to link host plants with their pathogens in a heuristic database.	9/10/2015 3:25 PM
2	Intensity of use and management. Some species are not cultivated (sensu strictu) but the traditional human use could affect their distribution. Ethnographic, archeologic and linguistic data.	7/6/2015 11:37 AM
3	In an ideal world, everything that has ever been registerd such as in grey literature, other bilbiographic references, experts' knowledge, national databases, personnal collections, etc	7/2/2015 5:13 PM
4	Vaviliv collection in St Petersburg (even if included in EURISCO and GeneSys) GeneSys gateway to genetic resources, this portal includes more or less the very same specimens as are also available in GBIF - but includes many important data properties not indexed by GBIF - and GeneSys includes valuable experimental trait data Crop Wild Relative portals, http://www.cwrdiversity.org/, http://www.cropwildrelatives.org/, ... some occurrence records included in/or originating from GBIF, but includes valuable data properties not indexed or included in GBIF.	6/30/2015 1:11 PM
5	When a dataset was cleaned, it should be very good to joint this dataset to Gbig through a link.	6/30/2015 12:36 PM
6	If they are not available, all GC genebank data should be available	6/30/2015 11:21 AM
7	Collection mission data are important	6/30/2015 10:04 AM
8	Environmental layers, although not related so much can be included in GBIF to provide one stop shop for the biodiversity data.	6/30/2015 7:10 AM
9	intra-specific data	6/29/2015 5:46 PM
10	the aim should be inclusion of all significant datasets from herbaria, genebanks, and researchers	6/26/2015 9:58 PM
11	The global occurrence CWR data set compiled by CIAT with Nora's team.	6/26/2015 6:12 PM
12	Field data (e.g. agronomic, morphological and physiological data). Accurate weather data (e.g. sunshine hours, temperature, light intensity, rainfall). Soil characteristics (physical and chemical). Socioeconomic data (e.g. number of farmers per area, population density, size of the local market).	6/23/2015 4:41 AM
13	Other data should be made available to GBIF	6/22/2015 10:48 PM
14	Neglected species, threaten species...	6/18/2015 9:21 PM
15	Local varieties Indigenous use	6/16/2015 10:11 PM
16	Have no experience on agrobiodiversity	6/15/2015 6:02 PM
17	Germplasm bank databases would be another important source of occurrence data for abrobiodiversity information.	6/15/2015 2:12 PM
18	Name of the project who provid the data Form of consumption	6/15/2015 2:01 PM
19	Occurrence data from national projects and other national sources that have not yet been gathered	6/15/2015 1:19 PM
20	increase quality, reduce duplication, collaborate with GeneSys, don't expand	6/15/2015 11:08 AM
21	Bioversity Collecting Mission Database (http://bioversity.github.io/geosite/)	6/15/2015 10:04 AM
22	Bioversity geospatial database of collected samples; Database on Crop Wild Relatives from the Trust/Birmingham/CIAT project	6/13/2015 11:39 AM
23	Datasets from herbaria and genebanks is still very poorly covered.	6/12/2015 6:40 PM
24	Socio-cultural aspects that have impact on agrobiodiversity.	6/12/2015 6:00 PM
25	No suggestions	6/12/2015 6:00 PM
26	Once the GCDT CWR dataset is available then it is just a case of making more big herbaria collections available, particularly for China	6/12/2015 6:00 PM

27	The global occurrence dataset on crops and their wild relatives, prepared at CIAT	6/12/2015 5:45 PM

Q27 Are you involved or know about an ongoing initiative rescuing and/or producing relevant agrobiodiversity data? Please describe (maximum 200 characters)

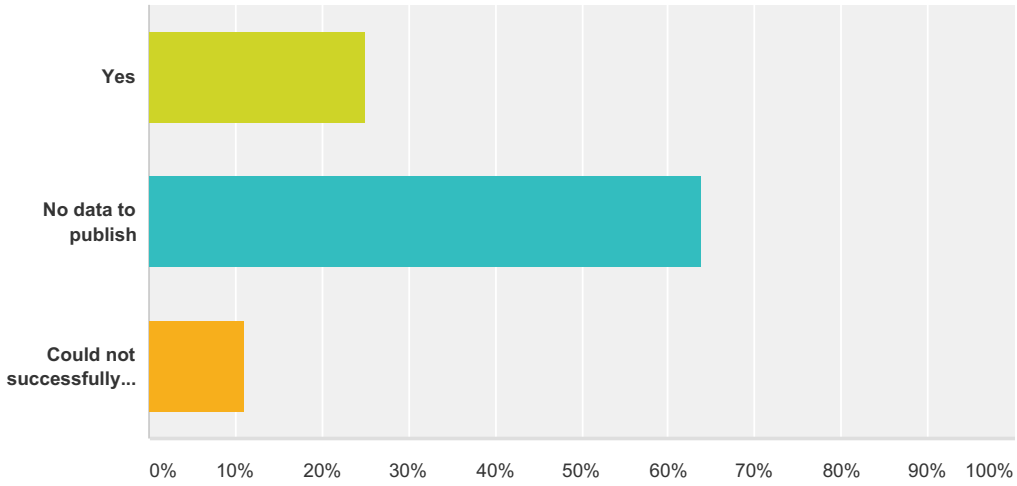
Answered: 27 Skipped: 25

#	Responses	Date
1	I am currently involved in a project on the genus Pachyrhizus, as well as on a project on seed exchange networks and plant epidemiology focussing on African roots and tuber crops	9/10/2015 3:25 PM
2	"Mansfeld's Database of Agricultural and Horticultural Crops" (mansfeld.ipk-gatersleben.de); Database for checklists of cultivated plant species (not online); IPK Genebank Information System (GBIS); EURISCO	7/20/2015 10:53 AM
3	History of domestication of plants and landscapes in Brazillian Amazonia.	7/6/2015 11:37 AM
4	Some of my students recorded data on distribution of edible fruits in Perak	7/4/2015 7:13 AM
5	Any initiative related to plant diversity in general is important because they usually have information on crop wild relatives (FloraOn in Portugal), many genebanks and herbaria in the SADC region, the Royal Botanic Garden in Jordan	7/2/2015 5:13 PM
6	Yes, the Norwegian/Nordic crop wild relative conservation strategy Mater, PhD students, postdocs, and colleagues at the museum	6/30/2015 1:11 PM
7	no	6/30/2015 11:21 AM
8	Several initiatives in genebank world (Bioversity collecting mission database, ECPGR crop databases, USDA crop databases (maize, cotton, soybean, mostly biased to genomic data and large cash crops)	6/30/2015 10:04 AM
9	http://www.biodiversitymapping.org/, https://data.nbn.org.uk/	6/30/2015 7:10 AM
10	yes- Crop Wild Relative global project www.cwrdiversity.org	6/26/2015 9:58 PM
11	N/A	6/23/2015 4:41 AM
12	We are involved in collecting data	6/22/2015 10:48 PM
13	No	6/18/2015 9:21 PM
14	http://www.uni-passau.de/biodiva/startseite/ BioDIVA project	6/16/2015 10:11 PM
15	RGscope is the genetic resources side of our project ECOSCOPE. This metadata portal will bring information about existing french research databases on genetic resources, but ECOSCOPE as a broader range and agrobiodiversity issues would be addressed through agricultural research observatories that will be referenced in ECOSCOPE metadata portal.	6/16/2015 11:19 AM
16	No	6/15/2015 6:02 PM
17	No.	6/15/2015 2:12 PM
18	Computerization of herbarium data Various ethnobotanical study in the university Plant Genetic Resources Programme of ISRA	6/15/2015 2:01 PM
19	The most important initiatives are emerging silently at national scales in several countries interested on improve their agrobiodiversity databases	6/15/2015 1:19 PM
20	involved in EURSICO and GeneSys and in setting up the ITPGRFA GIS	6/15/2015 11:08 AM
21	Bioversity Collecting Mission Database (http://bioversity.github.io/geosite/)	6/15/2015 10:04 AM
22	SADC Crop Wild Relatives project led by Bioversity; Agrobiodiversity Bioversity projects: surveys at the household level; Restoration projects led by Bioversity in Guatemala and Ethiopia; identification of edible species with their nutritional value ; Seeds4Needs project: varietale assessment with farmers communities	6/13/2015 11:39 AM
23	We are steadily producing, managing and communicating such data at the Canadian national genebank and look forward to the implementation of GRIN Global.	6/12/2015 6:40 PM
24	yes, its the Chirapaq ñan initiative. W establish a baseline to monitor potato diversity at the genetic/allele, variety/species, landscape and local/collective knowledge level. The most important impact factors for dynamics in the potato diversity will be monitored as well. This could be considered as the 5th level.	6/12/2015 6:00 PM
25	No	6/12/2015 6:00 PM
26	GCDT CWR dataset	6/12/2015 6:00 PM

27	Yes, during the last 4 years I have been part of a major effort of gathering an preparing the occurrence records of crop wild relatives globally.	6/12/2015 5:45 PM

Q28 Have you published agrobiodiversity related data through the GBIF network?

Answered: 36 Skipped: 16



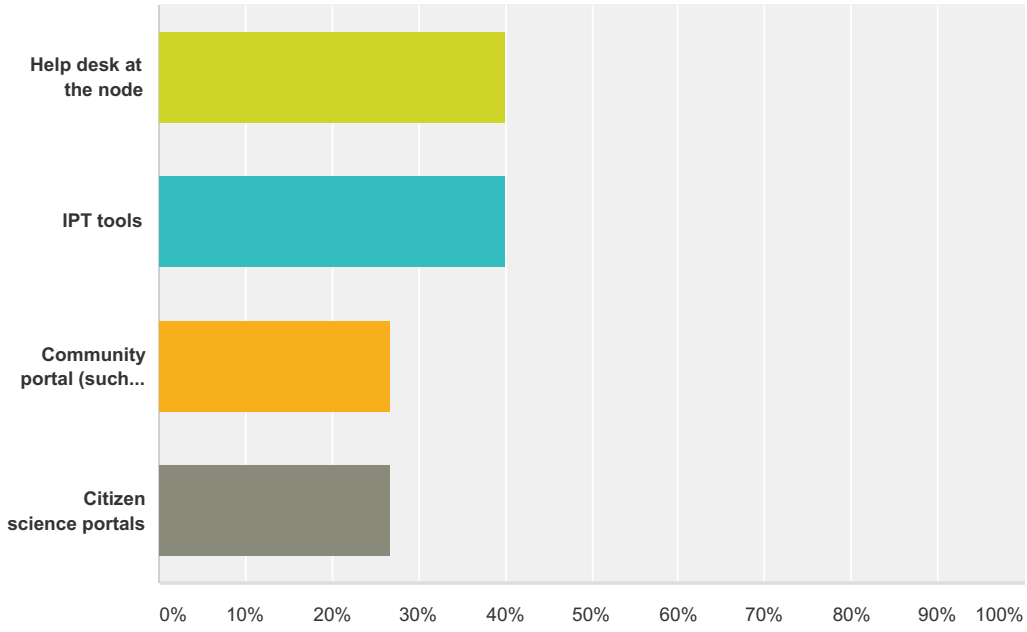
Answer Choices	Responses	
Yes	**25.00%**	9
No data to publish	**63.89%**	23
Could not successfully publish data	**11.11%**	4
Total		**36**

#	If yes, was is it easy? If you failed, please elaborate on your problems	Date
1	As yet I do not have data ready for publication but I am planning to publish all data related to Pachyrhizus once these are ready.	9/10/2015 3:25 PM
2	There were technical problems to connect GBIS to EURISCO (can provide details on request)	7/20/2015 10:53 AM
3	It is underway.	7/19/2015 9:54 AM
4	No. because of the duplicates.	7/6/2015 11:37 AM
5	Previously installing the appropriate Python libraries for BioCASE could cause issues. With IPT the java libraries and the stability of IPT in the given server environment might cause issues. For agrobiodiversity data the Darwin Core and ABCD data standards are incomplete in coverage for the core data properties (such as the Multi-crop passport descriptor standard).	6/30/2015 1:11 PM
6	We are mainly interested in trait data, while we consider GBIF more a botanic occurrence information system. Thus we rather use GBIF for specific purposes mainly in wild plants.	6/30/2015 10:04 AM
7	plan to do it	6/26/2015 9:58 PM
8	It was not so easy but we need to follow up	6/18/2015 9:21 PM
9	Human resource cost for this activity was not factored in the project	6/16/2015 10:11 PM
10	with assistance of NL-BIF	6/15/2015 11:08 AM
11	I didn't succeed to use the GBIF Integrated Publishing Toolkit. Programming skills are required	6/15/2015 10:04 AM
12	No.	6/12/2015 6:40 PM
13	First we need to verify our data then it depends on our data curator whether and how genebank data will be shared with the GBIF framework. For now its more important to share the data among the stakeholders of our in situ conservation network.	6/12/2015 6:00 PM
14	Yes, it was quite easy. Always received great assistance. I have done it two times with data from genebanks	6/12/2015 6:00 PM
15	Yes but I did not do it myself, I worked via collaborators	6/12/2015 6:00 PM

16	We are interested in publishing the crop wild relatives dataset through GBIF, we are still curating and updating some records (finishing by July-August 2015)	6/12/2015 5:45 PM

Q29 How did you publish your data (multiple choice)

Answered: 15 Skipped: 37

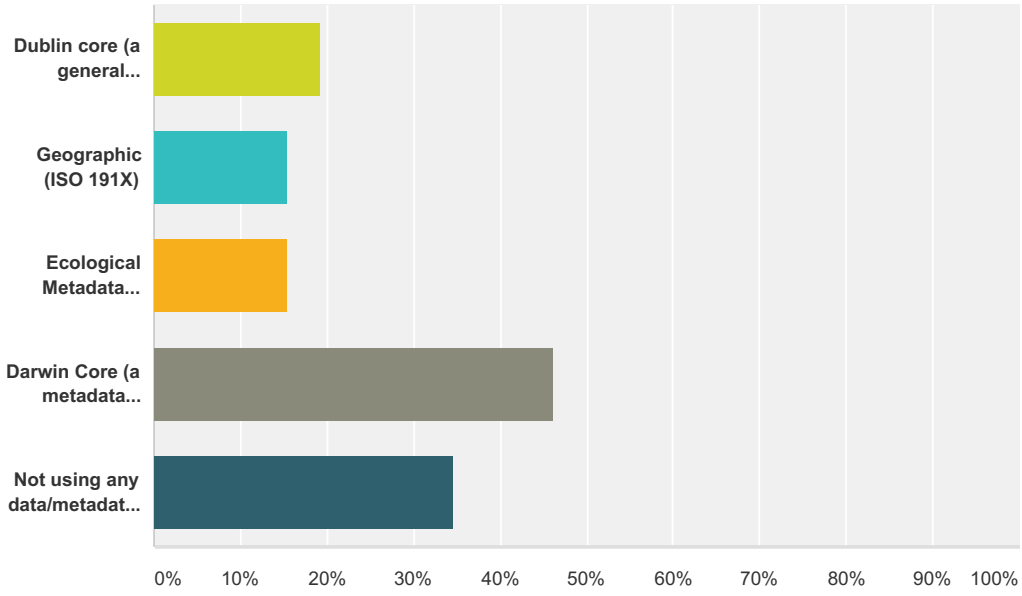


Answer Choices	Responses	
Help desk at the node	**40.00%**	6
IPT tools	**40.00%**	6
Community portal (such as EURISCO)	**26.67%**	4
Citizen science portals	**26.67%**	4
Total Respondents: 15		

#	Other (please specify)	Date
1	I have never published this kind of data (because I do not really generate it)	7/4/2015 7:13 AM
2	publications and project web-sites	6/30/2015 11:22 AM
3	Own Central Crop Databases, Project Information Systems like AVEQ, AEGRO	6/30/2015 10:08 AM
4	N.A	6/23/2015 4:41 AM
5	We don't produce data	6/16/2015 11:19 AM
6	Unpublished data in GBIF as Senegal not	6/15/2015 2:05 PM
7	NA	6/15/2015 1:21 PM
8	Archives	6/13/2015 11:43 AM
9	GRIN-CA and linked databases; scientific publications	6/12/2015 6:41 PM
10	see 26	6/12/2015 6:00 PM

Q30 What type of data/metadata standard are you using?

Answered: 26 Skipped: 26

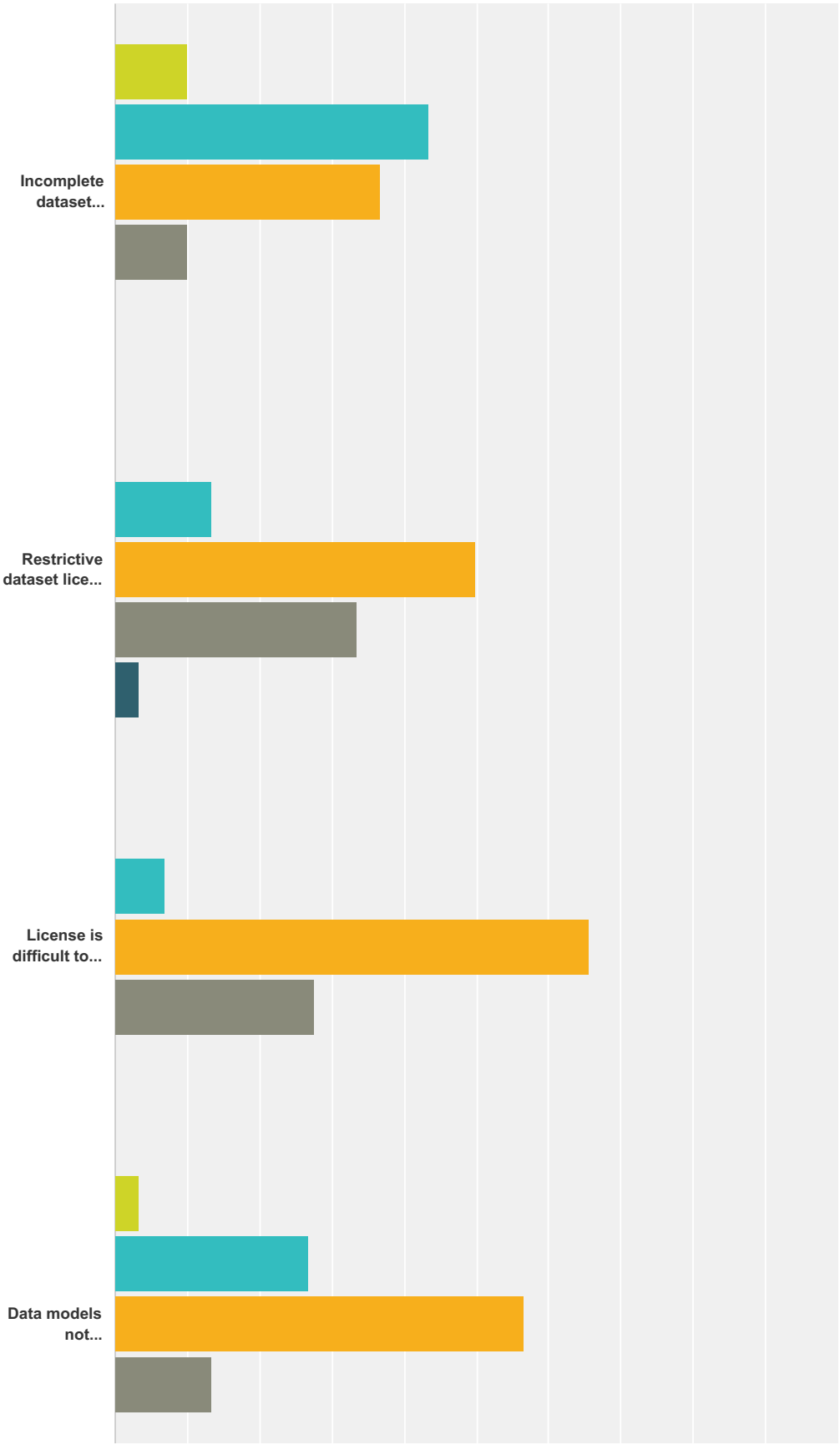


Answer Choices	Responses	
Dublin core (a general metadata template)	**19.23%**	5
Geographic (ISO 191X)	**15.38%**	4
Ecological Metadata Language (EML)	**15.38%**	4
Darwin Core (a metadata specifically for the biodiversity data)	**46.15%**	12
Not using any data/metadata standard to describe or document my data	**34.62%**	9
Total Respondents: 26		

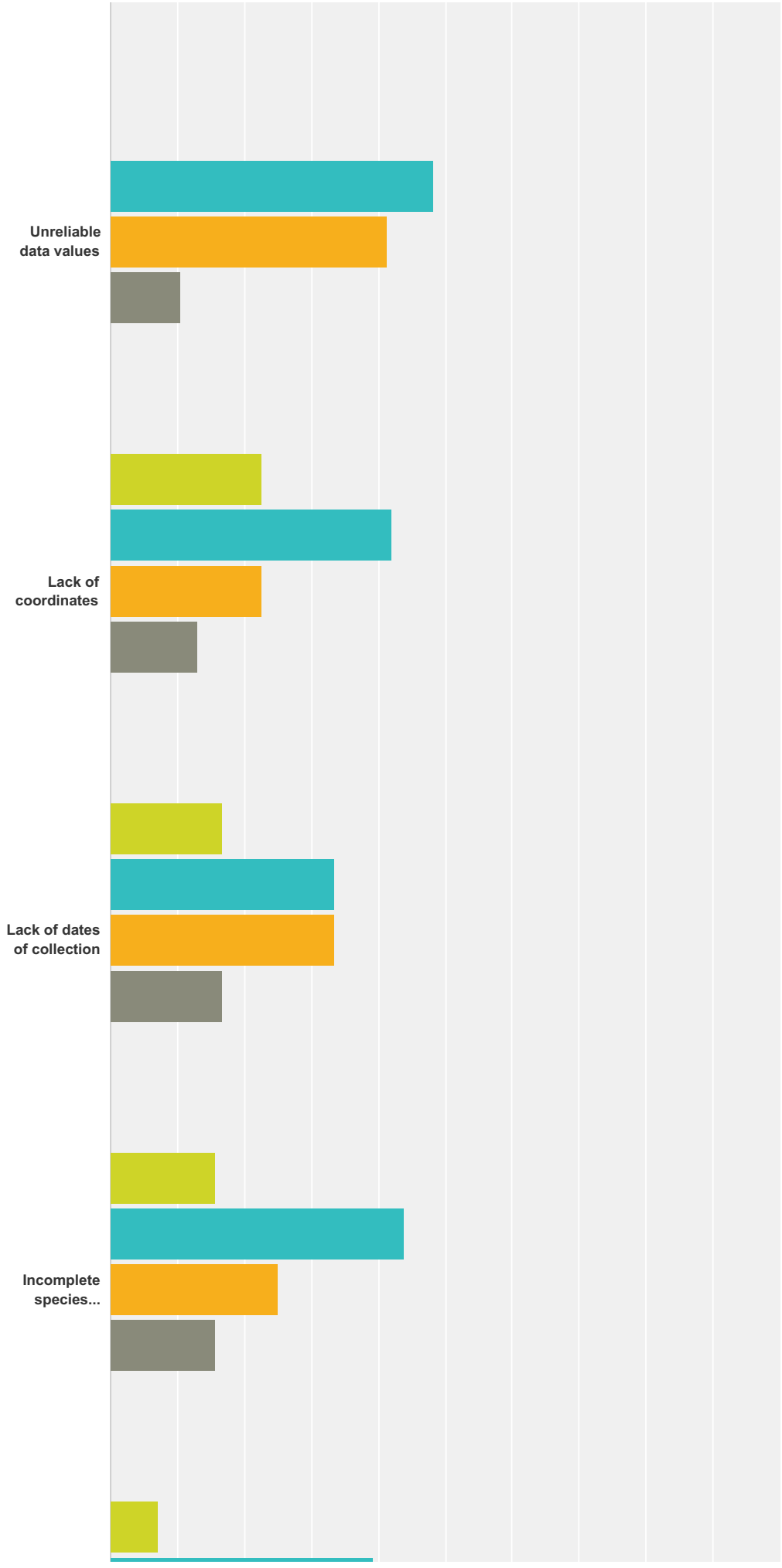
#	Other (please specify)	Date
1	EURISCO MCPD (in EURISCO, GBIS)	7/20/2015 10:54 AM
2	NA	7/4/2015 7:13 AM
3	breeding standard and molecular genetics	6/30/2015 11:22 AM
4	Multi-crop passport descriptors, NUTS, LAU for location name standard, own models for C+E (phenotypic) data	6/30/2015 10:08 AM
5	Multi-crop passport descriptor (MCPD) from FAO-Bioversity 2012	6/15/2015 1:21 PM
6	MCPD	6/15/2015 10:05 AM
7	Multicrop Passport data; trait standards	6/13/2015 11:43 AM
8	Following the standards set by GRIN.	6/12/2015 6:41 PM

Q31 From your experience, which are the main bottlenecks limiting the reuse of occurrence or taxon data retrieved from GBIF?

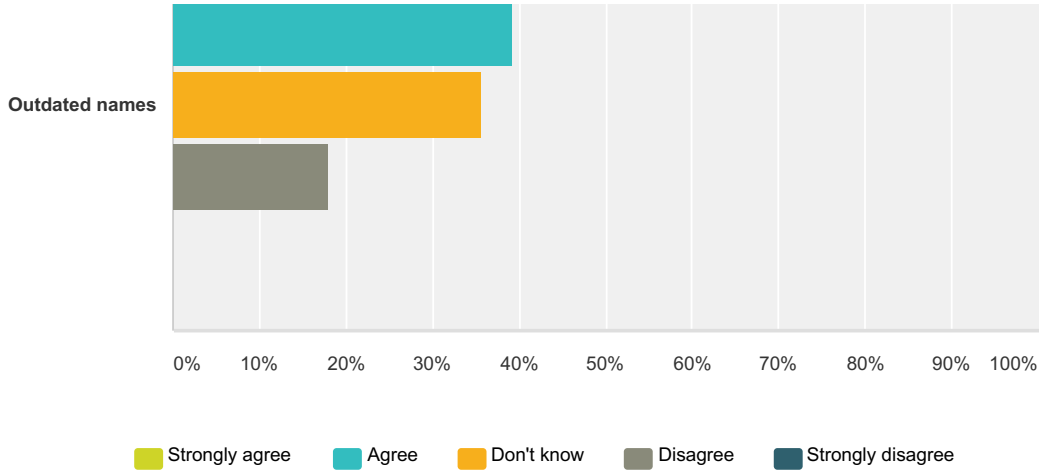
Answered: 33 Skipped: 19



GBIF and Bioversity International: Data Fitness for Use in Agrobiodiversity 2015



GBIF and Bioversity International: Data Fitness for Use in Agrobiodiversity 2015



	Strongly agree	Agree	Don't know	Disagree	Strongly disagree	Total
Incomplete dataset metadata (e.g. missing methodology, dataset descriptions, author, date)	**10.00%** 3	**43.33%** 13	**36.67%** 11	**10.00%** 3	**0.00%** 0	30
Restrictive dataset license (GBIF supports CC0, CC-BY or CC-BY-NC)	**0.00%** 0	**13.33%** 4	**50.00%** 15	**33.33%** 10	**3.33%** 1	30
License is difficult to understand, no machine readable license is provided	**0.00%** 0	**6.90%** 2	**65.52%** 19	**27.59%** 8	**0.00%** 0	29
Data models not satisfactory	**3.33%** 1	**26.67%** 8	**56.67%** 17	**13.33%** 4	**0.00%** 0	30
Unreliable data values	**0.00%** 0	**48.28%** 14	**41.38%** 12	**10.34%** 3	**0.00%** 0	29
Lack of coordinates	**22.58%** 7	**41.94%** 13	**22.58%** 7	**12.90%** 4	**0.00%** 0	31
Lack of dates of collection	**16.67%** 5	**33.33%** 10	**33.33%** 10	**16.67%** 5	**0.00%** 0	30
Incomplete species determination	**15.63%** 5	**43.75%** 14	**25.00%** 8	**15.63%** 5	**0.00%** 0	32
Outdated names	**7.14%** 2	**39.29%** 11	**35.71%** 10	**17.86%** 5	**0.00%** 0	28

#	Other (please specify)	Date
1	Lack of persistent identifiers	6/30/2015 1:31 PM
2	More then lack of coordinates, the low quality of the current available coordinates and the idea to pack the complete locality description in a single field, creating a complex of administrative and non administrative labels, disgregate them in several fields, please! take note of the GADM (http://www.gadm.org) structure	6/15/2015 2:02 PM
3	difficulties to publish phenotypic data	6/15/2015 11:11 AM
4	Incomplete MCPD	6/15/2015 10:09 AM
5	I will need to try the GBIF system to give an appropriate feedback for this question. In future this might be happen. For now it is not my priority.	6/12/2015 6:00 PM

Q32 Do you have any thoughts on how the presentation of GBIF-mediated agrobiodiversity data could be improved to better support your research?

Answered: 16 Skipped: 36

#	Responses	Date
1	It should be possible to map all accessions of a target species of genus at once, and explore the map to locate the most interesting specimens or identify straightaway geographical areas where sampling is missing	9/10/2015 3:38 PM
2	I find it not straightforward how you get to have a list of plant diversity occurrences for a particular country.	7/2/2015 5:22 PM
3	Improved cross-linking with other data sources including richer data properties and/or more precise data type classification. GBIF/Darwin Core occurrences, specimens and observations provide a collective bag of things that might often be typified more precisely in other systems (eg. cultivation status).	6/30/2015 1:31 PM
4	please, include collecting information as it was in SINGER	6/30/2015 12:39 PM
5	N/A	6/30/2015 11:25 AM
6	Better atomization (e.g. locality, subspecific taxonomy) of information would be always a wish. But it is difficult to achieve	6/30/2015 10:32 AM
7	Data cleaning can be done by volunteers so that the amount of time for pre-processing can be reduced for the analysts.	6/30/2015 7:21 AM
8	No	6/23/2015 4:43 AM
9	Standard data/metadata format at all scales and levels	6/16/2015 10:15 PM
10	That's fine with me for now	6/15/2015 6:11 PM
11	1. Do not pack the complete locality description in a single field, disgregate them in several fields, using GADM (http://www.gadm.org) structure as a model. 2. Creating an easy georeferencing quality evaluator for the occurrence data, using it to evaluate all occurrence data with coordinates and making this evaluator available to the users	6/15/2015 2:02 PM
12	-	6/15/2015 11:11 AM
13	Use complete MCPD. Use georeferencing tools with uncertainty description	6/15/2015 10:09 AM
14	Agrobiodiversity data could be presented as a separated (or special) collection within GBIF.	6/12/2015 6:11 PM
15	Digitising large herbaria collection data, particularly for China	6/12/2015 6:04 PM
16	I will need to try the GBIF system to give an appropriate feedback for this question. In future this might be happen. For now it is not my priority.	6/12/2015 6:00 PM

Q33 Which additional data fields associated to agrobiodiversity data should be directly available at GBIF.org?

Answered: 14 Skipped: 38

#	Responses	Date
1	Taxonomic resolution/reconciliation tools should be directly implemented in GBIF	9/10/2015 3:38 PM
2	Yes	7/19/2015 10:12 AM
3	See 35.	7/2/2015 5:22 PM
4	The missing core data properties from the Multicrop passport descriptor standard; International treaty legislation governing the access and benefit regime relevant for the respective agrobiodiversity specimens and in situ populations. Metadata imported and synchronized from the FAO WIEWS and the GRBio and similar accessible systems.	6/30/2015 1:31 PM
5	N/A	6/30/2015 11:25 AM
6	Genebank identifiers, if data relate to collected biodiversity	6/30/2015 10:32 AM
7	N/A	6/23/2015 4:43 AM
8	Indigenous names	6/16/2015 10:15 PM
9	That's fine with me for now	6/15/2015 6:11 PM
10	The following fields from MCPD (FAO 2012) format are required: LATITUDE and LONGITUDE (if the original source provide coordinates in sexagesimal format), GEOREFMETH, SAMPSTAT and COLLSRC. Additionally a new field to identify CWR species would be very helpfull	6/15/2015 2:02 PM
11	-	6/15/2015 11:11 AM
12	MSPD: Missing sample status, collecting source, complete administrative description like admin1, admin2, admin3 and collecting site. At the moment everything is in one field	6/15/2015 10:09 AM
13	Fields indicating whether the record represents a germplasm accession, or a herbarium specimen.	6/12/2015 6:11 PM
14	I will need to try the GBIF system to give an appropriate feedback for this question. In future this might be happen. For now it is not my priority.	6/12/2015 6:00 PM

Q34 Which additional data fields associated to agrobiodiversity data should be linked to the GBIF.org?

Answered: 12 Skipped: 40

#	Responses	Date
1	As suggested earlier, a link to a database describing pathogens either identified in the accessions or in the locality where the accessions were collected could greatly enrich GBIF data.	9/10/2015 3:38 PM
2	Trait information,	7/2/2015 5:22 PM
3	Experimental trait data following the Bioversity crop descriptors, crop ontology and similar systems; Experimental molecular data for sequence or marker analysis; Environmental data	6/30/2015 1:31 PM
4	N/A	6/30/2015 11:25 AM
5	Characterization and evaluation (phenotyping and genotyping), political administrative units to easy locate reponsibilities for in-situ/on-farm conservation	6/30/2015 10:32 AM
6	A link to the global reliable environmental datasets (like ISCRIC) will be very useful. Other resources for cultivated species will also be useful for the analysts.	6/30/2015 7:21 AM
7	N/A	6/23/2015 4:43 AM
8	Studies with farm scale diversity examples Season infromation	6/16/2015 10:15 PM
9	That's fine with me for now	6/15/2015 6:11 PM
10	The following fields from MCPD (FAO 2012) format are required: LATITUDE and LONGITUDE (if the original source provide coordinates in sexagesimal format), GEOREFMETH, SAMPSTAT and COLLSRC.	6/15/2015 2:02 PM
11	GeneSys; -omics data (ref DivSeek)	6/15/2015 11:11 AM
12	I will need to try the GBIF system to give an appropriate feedback for this question. In future this might be happen. For now it is not my priority.	6/12/2015 6:00 PM

Q35 Indicate your recommendations to GBIF for improving the fitness of the data for your use in agrobiodiversity research.

Answered: 36 Skipped: 16

#	Responses	Date
1	Mostly, taxonomic data must be updated or a taxonomic reconciliation tool should be implemented to allow retrieval of exhaustive datasets related to a target genus/species. It is also of utmost importance that data be provided with accurate geographical coordinates.	9/10/2015 3:38 PM
2	Presently no particular recommendations	7/20/2015 10:57 AM
3	Find solutions to correct the main bottlenecks limiting the reuse of occurrence or taxon data retrieved from GBIF	7/19/2015 10:12 AM
4	It's possible improve the accuracy of geographic coordinates and well documented its metadata.	7/14/2015 6:54 PM
5	Conect occurrence data with management and use plant information. Add archeological, linguistic and ethnographic data to plant species occurrence.	7/6/2015 11:40 AM
6	Sorry, I do not use GBIF	7/4/2015 7:14 AM
7	- Cover more countries and a broader range of sources. - Georeference records without coordinates, assign a level of geographic precision. - Make coordinates more accurate where possible and indicate whether the coordinates come from original data or are georeferenced. - Tag the low quality records. - Integrate other sources of data (EURISCO, ENSCONET, etc…). - More actively engineer data submitted to GBIF to be better quality - working with the donors of data.	7/2/2015 5:22 PM
8	Start with an even stronger push for persistent identifiers, contribute to further development of a data domain model for integration of data types (including cultivation status). Provide services for GeneSys, EURISCO, SINGER, USDA GRIN, ... with respect to integration of source data. Ensure that GBIF provides a complete data infrastructure (eg. includes solutions for all of the core data properties) for the needs of these valuable agrobiodiversity portals and networks.	6/30/2015 1:31 PM
9	Allowed access to cleaned dataset	6/30/2015 12:39 PM
10	it needs to be able to take accession and molecular data related to breeding and research, if there is a real desire to integrate the occurrence and breeding aspects of germplasm use	6/30/2015 11:25 AM
11	The download procedure is rather complicated, probably because of the large amount of data to be filtered. Excel needs special adjustment to display the data. Probably this could be improved in the download procedure.	6/30/2015 10:32 AM
12	Having data pre-processed will help the data scientists greatly. In this regard, GBIF could consider providing online services for data preparation, manipulation and visualisation. These with the option of adding multiple layers of data will boost the hypothesis forming, research collaboration and knowledge gap analysis.	6/30/2015 7:21 AM
13	An index on taxonomic accuracy of a sample. For example thosw without the name of a person who determined a sample are less reliable.And some taxa groups are more diffcicult to assess than others	6/29/2015 5:49 PM
14	futher data processing and detection of errors is always appreciated, although we recognize that primary responsibility rests with data providers	6/26/2015 10:00 PM
15	Better quality co-ordinate data and a column to show to what degree of precision this has been done. Who has defined the co-ordinates - are they original or has someone else geo-reffed them. Taxonomically verified observations - although this may be difficult in practise, does cause some issues as some records can be wrongly identified	6/26/2015 6:16 PM
16	The platform should have a data system to provide different taxonomic names (i.e old names), also a filter to use some period (i.e. 1970-2015)	6/26/2015 3:34 PM
17	.	6/26/2015 3:29 PM
18	Compare datasets against pre-existing ones like FAO stat. Discuss with other dataset portals about fitness of respective data with reliability based on methodology of collecting data.	6/23/2015 4:43 AM
19	we must add photos species and comments on their use	6/22/2015 11:14 PM
20	Update the taxonomic names, ceck formore accurate coordinates...	6/18/2015 9:25 PM
21	Good	6/16/2015 10:15 PM
22	As mentioned before, we don't use data.	6/16/2015 11:20 AM
23	ffkff	6/16/2015 12:17 AM

24	In general lines, GBIF data need to improve taxonomic accuracy (not restricted to agrobiodiversity taxa), since there are a number of names that are misspellings and/not updated names. This can be done by crosschecking records from different collections available through GBIF with taxonomic authorities such as IPNI, The plant list, Tropicos. Also, there are a number of records of wild species collected in research field stations that some users would assume to be “native” or within the natural range occurrence of the species. Therefore is would be useful to have the information native/cultivated linked to agrobiodiversity records.	6/15/2015 9:20 PM
25	That's fine with me for now	6/15/2015 6:11 PM
26	You can include several new fields in your data format (as I mentioned before), develop (for all GBIF data, not only for agrobiodiversity data) a friendly georeferencing quality evaluator/indicator to help to select data with good coordinates, you can create links with taxonomy systems related to agrobiodiversity (GRIN), reduce the steps to download datasets in the GBIF portal, please eliminate the registration requirement to download data (it goes very slow and drives away the users), and finally, disgregate locality description field (called "locality") in several administrative fields according to administrative global structure such as GADM.	6/15/2015 2:02 PM
27	do not duplicate (internally or with other informationservices, but provide effective linking	6/15/2015 11:11 AM
28	Focus more on data quality than on quantity	6/15/2015 10:09 AM
29	Put dataset for argobiodiversity with more complex information for complete taxonomic identification, place and time collection, weather, climatic and soil information this place, if it possible results of some valuable evaluation.	6/15/2015 8:38 AM
30	GBIF should continue to strengthen the capacity of young scientists	6/13/2015 1:44 PM
31	Develop Crop/Species expert group to review the quality of the data! Develop methods/tools to automatically identify outliers. Assign level of confidence to individual data records. Develop better annotation system, feedback from donors.	6/12/2015 10:36 PM
32	Links to data in genebanks might be useful.	6/12/2015 6:43 PM
33	An important portion of agrobiodiversity data is managed by germplasm banks. It is key to engage whether the germplasm banks, or platforms that already centralize this information (i.e. EURISCO, Genesys), to make sure the most updated data is available through GBIF.	6/12/2015 6:11 PM
34	It would be great if GBIF could do some preliminary quality filtering of the data they hold, and provide some feedback to data suppliers so bad quality data are eliminated from the system.	6/12/2015 6:06 PM
35	Digitising large herbaria collection data, particularly for China	6/12/2015 6:04 PM
36	I will need to try the GBIF system to give an appropriate feedback for this question. In future this might be happen. For now it is not my priority.	6/12/2015 6:00 PM

Appendix 2

Proposed Use cases

Use case A: Find gaps in conservation in crop wild relatives - proposed by Nora P. Castañeda-Álvarez
1. Describe the objective
Objectives: Given the previous background, a study to detect the extent of representativeness of crop wild relatives in genebanks, establish the *ex situ* conservation priorities of the crop wild relatives analyzed, and understand the patterns of global richness where future field collections could be conducted.
2. Who are the actors ?
Data managers, modelers, genebank curators, herbaria curators, sample collectors, breeders
3. Data/information products to be produced
In order to be able to achieve these objectives, it was necessary to produce potential distribution models, and to apply a gap analysis methodology to assess the sufficiency of CWR samples in genebanks (Ramirez-Villegas *et al*., 2010).

4. Data sources the most used

The analyses conducted relied heavily on geographic explicit information — occurrence records obtained from GBIF.org, germplasm mobilization platforms [e.g., the CGIAR's System-wide Information Network for Genetic Resources (SINGER), the European Plant Genetic Resources Catalogue (EURISCO), and the United States Department of Agriculture's Genetic Resources Information Network (GRIN), mainly], data provided by researchers, peer-reviewed and gray literature, and herbaria.

5. Tools the most used

- GRIN Taxonomy (http://www.ars-grin.gov/~sbmljw/cgi-bin/taxcrop.pl?language=en-) and the Harlan and De Wet Crop Wild Relative Inventory (http://www.cwrdiversity.org/checklist/) were largely used to identify the relationship between crop wild relatives and their associated crops, confirmed and potential uses in breeding, and the native distributions of crop wild relatives.
- An API to GRIN Taxonomy, the Taxonomic Name Resolution Service (http://tnrs.iplantcollaborative.org/), and TaxonStand (Cayuela *et al*., 2012) were used as references to identify possible misspellings in the taxonomic names of crop wild relatives, and to standardize them.
- GEOLocate and the Google Maps Geocode API were used to georeference records.
- MaxEnt was used as the algorithm to produce potential distribution maps (Phillips *et al*., 2006).
- An in-house tool was designed to process batches of occurrence records in the following order: 1) standardize fields, 2) check and standardize taxonomic names, 3) check and re-calculate geographic coordinates, 4) final standardization. The tool consists of a stand-alone application prepared in java and available here: www.github.com/CIAT-DAPA/cwr_occurrencesvalidation
- In terms of the gap analysis, the code used works with R and is adapted to analyze a single taxon or a complete crop genepool (https://github.com/npcastaneda/gap-analysis-maxent).

6. Describe the data flow into steps ("to be able to...") indicating who is doing the indicated step (a scientist, a developer, the system?)

To be able to …

7. Identify GBIF role and improvements

i. status of data (quality, coverage)

Major gaps in occurrence records are persistent, especially in countries like DRC, China, Russia, Argentina, and regions like the Amazon and central Africa. Improved access to occurrence records of crop wild relatives in such regions will help to detect populations that might not yet be represented in genebanks (and therefore require further

conservation actions).

ii. Additional attributes in demand

links to CWR checklists (e.g.,the Harlan and De Wet Crop Wild Relative Inventory, GRIN Taxonomy, and Mansfeld's World Database of Agricultural and Horticultural Crops -http://mansfeld.ipk-gatersleben.de/apex/f?p=185:3::::::-)

iii. Additional sources - > what are the possible connectors between GBIF and the sources

- Genesys
- European Plant Genetic Resources Catalogue (EURISCO)
- United States Department of Agriculture's Genetic Resources Information Network (GRIN)

iv. What data mobilization is needed? By whom?

- other initiatives (e.g.,GRIN global, Genesys) support the digitization and mobilization of passport data associated with germplasm accessions. It has been estimated that less than a third of the existing genebanks in the world had organized their collections and made their passport data digitally and openly available to the public. Establishing a close collaboration between GBIF and these initiatives (GRIN global and Genesys) will increase the visibility of plant genetic resources occurrence data (including crop wild relatives) via GBIF.
- The project "Adapting agriculture to climate change: collecting, conserving and preparing the crop wild relatives" has produced a CWR occurrence database, including information not yet mobilized through GBIF (i.e., researchers archives, gray literature). This information can also help to improve the completeness of CWR data mobilized by GBIF.

→ Is this a priority? if not implemented how will it block the progress?

Link to recommendations #

6.2.1, 6.2.2, 6.2.3, 6.3.1, 6.3.2, 6.3.3, 6.6.1, 6.6.2, 6.6.3, 6.6.4, 6.6.7, 6.7.2, 6.8.1, 6.8.2, 6.8.3, 6.8.4, 6.9.1, 6.9.2, 6.9.3, 6.10.1, 6.10.2, 6.11.1, 6.11.2, 6.11.3, 6.11.5, 6.12.3, 6.14.1

Bibliographic references

Brummitt, N. A., Bachman, S. P., Griffiths-Lee, J., Lutz, M., Moat, J. F., Farjon, A., … Nic Lughadha, E. M. (2015). Green plants in the red: a baseline global assessment for the IUCN Sampled Red List Index for plants. Plos One, 10(8), e0135152. doi:10.1371/journal.pone.0135152

Cayuela, L., Granzow-de la Cerda, Í., Albuquerque, F. S., & Golicher, D. J. (2012). taxonstand: An R package for species names standardisation in vegetation databases. Methods in Ecology and Evolution, 3(6), 1078–1083. doi:10.1111/j.2041-210X.2012.00232.x

FAO. (2010). The second report on the state of the world's plant genetic resources for food and agriculture. Rome, Italy: Commission on Genetic Resources for Food and Agriculture, Food and Agriculture Organization of the United Nations.

Jarvis, A., Lane, A., & Hijmans, R. J. (2008). The effect of climate change on crop wild relatives. Agriculture, Ecosystems & Environment, 126, 13–23. doi:10.1016/j.agee.2008.01.013

Phillips, S. J., Anderson, R. P., & Schapire, R. E. (2006). Maximum entropy modeling of species geographic distributions. Ecological Modelling, 190(3-4), 231–259. doi:10.1016/j.ecolmodel.2005.03.026

Ramírez-Villegas, J., Khoury, C., Jarvis, A., Debouck, D. G., & Guarino, L. (2010). A gap analysis methodology for collecting crop genepools: a case study with Phaseolus beans. PloS One, 5(10), e13497. doi:10.1371/journal.pone.0013497

Use case B: Conservation plan for Crop Wild Relatives - by Dag Terje Endresen

1. Describe the objective

Background: Genetic diversity from crop wild relatives (CWRs) can be exploited as gene donors to provide needed new properties for food crops emerging from new climate regimes, demand for more food to feed growing world population, to keep up with evolving crop pests, etc. Novel molecular breeding and genetic modification methods allow for genetic diversity and alleles to be transferred between crops of different species and open for more effective use of the CWRs as a genetic resource.

Objective: Develop and implement a conservation plan for CWR (national --> regional --> global). Start with a national CWR conservation plan, collaborate with neighbouring countries to build a regional CWR conservation plan. A global CWR conservation plan based on both a global assessment and available national and regional CWR conservation plans.

2. Who are the actors ?

CWR expert team to develop a national list of priority CWR taxa. National protected area managers to include selected priority CWR taxa in their monitoring plans. This activity will gather the required information to support conservation policy decisions, including assigning designated CWR conservation populations together with the national CWR expert team. Students and researchers to collect molecular evidence on genetic diversity in and between the designated CWR conservation populations. This information will support conservation policy decisions (strengthen when unique genetic diversity, weaken when genetic diversity is similar across designated CWR populations in different locations).

3. Data/information products to be produced

- Global checklist of CWR, what species are classified as CWR (same genus as a crop)
- National checklist of CWR, national conservation status, national conservation priority
- Occurrences for nationally designated conservation populations of CWRs
- Connecting monitoring data on CWRs collected from national activities
- Abundance and quantities to verify that the CWR designated conservation population is ok
- Connecting genetic monitoring data to verify CWR designated population is ok

Definitions of CWR checlists :

CWR Checklists:

"Complete CWR checklist – A list of all CWR taxa found in a certain geographic area comprising a list of taxon names and authorities." (Maxted *et al*. 2015)
"Partial CWR checklist – a partial list of CWR found in a certain geographic area that is the result of a first prioritization, usually on crop genepools, providing the botanical names and authorities."
This CWR checklist (either the complete version or the partial version) can be (further) "prioritized to produce a shorter list for which specific active conservation is considered most necessary. The prioritized national CWR checklist forms the basis of the national CWR inventory."

The national inventory is thus a prioritized list of CWR with anciliary species specific information.

4. Data sources the most used

- Taxon checklists, GRIN Taxonomy, Mansfeld's World Database (for CWR checklist)
- GBIF occurrence data (for Distribution Modelling, SDM)
- Environment data from WorldClim and the Norwegian mapping agency (for SDM)
- External data including spatial extents for protected areas in the country
- A global checklist of crop species (that could have CWRs in the country)
- Data on the economic and other values for society of the crops related to the CWRs

5. Tools the most used

- Species distribution modeling software (Maxent, R)

6. Describe the data flow into steps ("to be able to...") indicating who is doing the indicated step (a scientist, a developer, the system?)

- To be able to develop a national CWR checklist, first the ABD crops and respective species need to be identified. The Mansfeld's world database of crop plants have a comprehensive list of crops.

- Next the CWRs in a country are identified. The general rule of species belonging to the same genus as a crop species can be used.
- A national checklist can be used to extract all species of the same genus as a crop on a global crop species checklist information resource.
- The next step is to identify nationally prioritized CWR species. We need to develop a set of prioritization criteria.
- Data on the economic and other values for society of the crops that are related to CWRs can provide important input to the prioritization of CWR checklist species.
- Direct economic or societal value of the CWR species itself can provide another prioritization criteria.
- With a national prioritized CWR checklist, a conservation strategy can be developed.
- One objective is to ensure *in situ* conservation for designated conservation populations.
- First step is to identify prioritized and stable CWR populations within existing protected areas. Comprehensive occurrence data from GBIF can be used.
- Without comprehensive occurrence data on CWRs species distribution modeling can be used as a predictive tool. Species predicted to occur inside protected areas must of course be verified by an expert to be present before any designated conservation populations are decided.
- If a prioritized CWR species is not conserved within any existing protected areas, then it is possible to start to gather evidence for proposing modification of the protected areas.
- However, more often threatened prioritized CWR species should anyway be collected for *ex situ* conservation to avoid them to be lost.
- When designated CWR conservation populations are decided, an *ex situ* backup copy is warranted both as part of the conservation goal and to provide more rapid and easier access to this genetic resource for use and research purposes.

7. Identify GBIF role and improvements

i. status of data (quality, coverage)

ii. Additional attributes in demand

- Indexing CWR and crop species status into the GBIF portal to allow for these attributes to be used in user-initiated search.
- GBIF portal to implement solutions to identify and link occurrence level information published from different data owners in different datasets.

iii. Additional sources - > what are the possible connectors between GBIF and the sources

iv. What data mobilization is needed? By whom?

- Mobilisation of taxon checklist information resources for CWRs and ABD crops.
- Mobilisation of occurrence data for CWR species, gaps analysis and strategies for filling gaps for these species

Recommendation: Publish in GBIF national, regional, and global CWR checklist with conservation priority assessment status. GBIF nodes to assist national CWR expert to publish checklists and monitoring data for designated CWR populations. Provide and USE persistent identifiers that identify the designated CWR populations (new term needed?) and conservation sites (dwc:locationID)

→ Is this a priority? if not implemented how will it block the progress?

Link to recommendations #

6.2.1 MCPD to be indexed by the GBIF portal
6.2.3 Extension with attributes to describe CWR species
6.2.5 Agrobiodiversity community governance for the Darwin Core germplasm extension, including a process to cover in situ CWR
6.3.1 CWR species checklist will be the starting point for national CWR conservation strategies
6.3.2 Taxon level extensions for CWRs to be indexed by the GBIF portal taxon backbone
6.3.3 Training for nodes on CWRs to participate in CWR data mobilization and use
6.6.1 The global CWR occurrence dataset will guide the national CWR conservation strategy
6.7.1 Training for nodes on CWRs to participate in CWR data mobilization and use
6.8.1 In particular taxon backbone information on CWRs would be useful
6.8.2 Georeferencing data cleaning tools and services embedded in the GBIF portal
6.8.4 Taxon names cleaning tools and ABD checklist resources in the portal
6.9.1 Improved tools to identify occurrence level duplicates will improve modeling prerequisites
6.9.2 Solutions and methods to allow refined versions of CWR occurrences into the portal
6.9.3 Cross-linking duplicated occurrence level information between datasets
6.14.1 Combining occurrence data with other data

Recommendation: An extension to the GBIF Taxon core for CWR attributes/information. Minor additions to the taxon data models implemented by GBIF will enable data exchange of attributes to support CWR conservation by using the existing GBIF infrastructure - replacing the need for a new and parallel data flow mechanism.

- Perhaps simply the taxon extension for description might provide sufficient functionality(?): http://rs.gbif.org/extension/gbif/1.0/description.xml
- The species profile taxon extension also provides similar functionality that could be explored: http://rs.gbif.org/extension/gbif/1.0/speciesprofile.xml

Bibliographic references

Asdal Å, Philips J, and Maxted N (2013). Boost for crop wild relative conservation in Norway. *Crop wild relative* 9:20-21.

Brehm JM, Maxted N, Martins-Loução MA, and Ford-Lloyd BV (2010). New approaches for establishing conservation priorities for socio-economically important plant species. *Biodiversity and Conservation* 19(9): 2715-2740. doi:10.1007/s10531-010-9871-4

Fielder H, Brotherton P, Hosking J, Hopkins JJ, Ford-Lloyd B, and Maxted N (2015). Enhancing the Conservation of Crop Wild Relatives in England. *PLoS ONE* 10(6): e0130804.

doi:10.1371/journal.pone.0130804

Harlan JR, and de Wet JMJ (1971). Towards a rational classification of cultivated plants. *Taxon* 20(4): 509–517. doi:10.2307/1218252

Iriondo JM, Maxted N, and Dulloo ME (eds.) (2008). Conserving Plant Genetic Diversity in Protected Areas: Population Management of Crop Wild Relatives. Wallingford, UK: CAB International. ISBN 9781845932824. doi:10.1079/9781845932824.0000

Maxted N, Ford-Lloyd BV, and Hawkes JG (eds.) (1997). Plant Genetic Conservation. The *in situ* approach. Chapman and Hall, London, UK. ISBN 978-94-009-1437-7. doi:10.1007/978-94-009-1437-7

Maxted N, Scholten M, Codd R, and Ford-Lloyd B (2007). Creation and use of a national inventory of crop wild relatives. *Biological Conservation* 140(1-2): 142-159. doi:10.1016/j.biocon.2007.08.006

Maxted N, Dulloo E, Ford- Lloyd BV, Iriondo JM, and Jarvis A (2008). Gap Analysis: A tool for complementary genetic conservation assessment. *Diversity and Distributions* 14(6): 1018-1030. doi:10.1111/j.1472-4642.2008.00512.x

Maxted N, and Kell S (2009). Establishment of a global network for the in situ conservation of CWR: Status and needs. Background study paper no. 39. FAO Commission on Genetic Resource for Food and Agriculture, Rome, Italy. Available online at http://www.fao.org/docrep/013/i1500e/i1500e18a.pdf

Maxted N, Kell S, Ford- Lloyd B, and Toledo Á (2012). Towards the systematic conservation of global crop wild relative diversity. *Crop Science* 52(2): 1-12. doi:10.2135/cropsci2011.08.0415

Maxted N, Dulloo ME, Ford-Lloyd BV, Frese L, Iriondo JM, and Pinheiro de Carvalho MAA (eds.) (2012). Agrobiodiversity conservation: Securing the diversity of crop wild relatives and landraces. Pp. 72-77. CAB International, Wallingford, UK. ISBN 9781845938512. doi:10.1079/9781845938512.0000

Maxted N, Brehm JM, and Kell S (2013). Resource book for the preparation of national conservation plans for crop wild relatives and landraces. University of Birmingham, UK. Available at http://www.fao.org/fileadmin/templates/agphome/documents/PGR/PubPGR/ResourceBook/TEXT_ALL_2511.pdf

Phillips J, Kyratzis A, Christoudoulou C, Kell S, and Maxted N (2014). Development of a national crop wild relative conservation strategy for Cyprus. *Genetic Resources and Crop Evolution* 61(4): 817-827. doi:10.1007/s10722-013-0076

Vincent H, Wiersema J, Kell S, Fielder H, Dobbie S, Castañeda-Álvarez NP, Guarino L, Eastwood R, Leon B, and Maxted N (2013). A prioritized crop wild relative inventory to help underpin global food security. *Biological Conservation* 167: 265–275.

Use case C: Find agrobiodiversity hotspots and monitor changes for decisions on conservation - proposed by Yves Vigouroux

1. Describe the objective

Humanity relies on a few crops for food supply. Genetic diversity of these crops and their wild relatives is a key asset for adaptation of agriculture to future pest and climate conditions. We still do not know enough about the genetic diversity of our crop and wild relatives, how this diversity was built up over time, how it evolves and changes. Understanding patterns of diversity and evolution of this diversity is key information to develop informed conservation strategy.

The objectives are to identify the hotspots of agrobiodiversity, understand how these hotspots appear and change over time. The final aim is to develop a strategy for the *in situ* conservation of agrobiodiversity.

2. Who are the actors ?

Actors are researchers at the descriptive level and local community with the assistance of national conservation services (government, NGOs, Universities and other research centers) for preservation of specific areas. The analysis of the origin of the hotspots is also an important research question: co-occurrence with wild relatives, historical origin of crop and known diffusion path, human/farmer cultural diversity**.**

3. Data/information products to be produced

- Map of agrobiodiversity across wild relatives and cultivated species and main factors explaining such patterns
- Map of agrobiodiversity changes over time (few or several decades)

4. Data sources the most used

Occurrence data (GBIF, GENESYS, CGIAR genbank, …) and associated fields mainly latitude/longitude and varietal names
Environmental database (environmental data like BIOCLIM current, past and future climate; soil data, ..) and using plant names with statistical database (FAO stats for estimation of cultivated areas, yields with the possibility to use temporal variation of these datasets)
Genetic/genomics depository system (NCBI Genbank, Gramene, DRYAD, ..) or plant specific system (MaizeDB, Coffee Genome hub, Rice databases, …)
Cultural datasets (anthropological, linguistic map, …)

5. Tools the most used

- Genetic diversity analyses tools i.e. structure, estimation of diversity
- Niche modeling (Maxent)
- Maps of hotspots of agrobiodiversity with a high resolution

6. Describe the data flow into steps ("to be able to...") indicating who is doing the indicated step (a scientist, a developer, the system?)

a. To be able to identify the different species occurrences, we need :

a) occurrences with geographical coordinates with validated data for the occurrence location (latitude/longitude and village/country) and date of sampling

b) completeness of the occurrence of the different species, with valid species names

b. To be able to link between database /datasets, accessions names (fields) need to allow connection with

a) Phenotypic variation of the studied species (CGIAR Genebank or other data sources where phenotypic data might be available);

b) Genetic/genomic diversity work (NCBI Genbank, Gramene, DRYAD, .. or plant specific system MaizeDB, Coffee Genome hub, Rice databases, …).

The connection could be at the level of the accession name/occurrence name (ideal) or could be from closely geographic samples (less optimal as statistical inference is needed based on latitude/longitude).

c. To be able to retrieve environmental datasets (environmental data like BIOCLIM current, past and future climate; soil data...) and with cultural datasets (anthropological, linguistic map), a dataset with accurate geographical coordinates is needed.

d. To be able to monitor the temporal change in agrobiodiversity, supplementary information on crops like varietal names and date of sampling (precision day) is needed. Only an annual update of status is necessary.

7. Identify GBIF role and improvements

i. Status of data (quality, coverage)

1) Quality of geo-referenced datasets is crucial and good coverage is also important
2) Quality in species determination i.e. up-to-date species name and new fields like name of the variety.
3) Ability to link GBIF occurrence with the same occurrence found in other databases (CGIAR Genebank or other data sources where phenotypic data might be available; NCBI Genbank, Gramene, DRYAD, .. or plant specific system MaizeDB, Coffee Genome hub, Rice databases, etc)

ii. Additional attributes in demand

Completeness of the supplementary information associated with crop taxa: varieties names are necessary

iii. Additional sources - > what are the possible connectors between GBIF and the sources

Keep or share unique identifier across databases. Either GBIF keeps the other database's unique identifier (best solution for entry from GBIF on agrobiodiversity data) or a unique identifier is shared across database (necessitate coordination)

iv. What data mobilization is needed? By whom?

- Mobilization of new occurrence data about agrobiodiversity, to be accessible through GBIF, is needed by the agrobiodiversity community
- Centralization and completeness of occurrence at a given entry point (GBIF) by publishing through GBIF occurrence data from other sources

→ Is this a priority? if not implemented how will it block the progress?

- Latitude and longitude validity is very important and access to tools that ensure this quality are necessary. Data of poor quality should not be considered.
- Up-to-date species names are also very important because it allows to discuss a given and precise category

- Link between databases and unique identifiers are of utmost importance since it allows to link very different datasets useful for this objective.

Link to recommendations #

6.1.1, 6.1.2, 6.1.3, 6.1.4, 6.1.5, 6.2.1, 6.2.2, 6.2.3,6.2.4, 6.3.1, 6.4.1, 6..6.1,6.6.2,6.6.3, 6.6.4, 6.9.1, 6.9.2, 6.9.3, 6.11.1, 6.14.1, 6.14.2, 6.14.3

Bibliographic references

Use case D: Crop modelling - proposed by Ebrahim Jahanshiri

1. Describe the objective

Objective: To include/combine crop specific descriptors and vocabularies in addition to the common GBIF protocols/data standards to help the crop performance and modelling community use the most relevant crop cultivars/landrace and their associated information.
Background: Prediction and modelling of the crops is of utmost importance.
Crop specific data are crucial for specific modelling and yield projection exercises. Such data are normally calibrated for specific crop-cultivars and are available per crop model. Variety of standardisation is underway to improve the data flow from trials to modelling.

2. Who are the actors ?

Crop modelling community can converge more on using the standards to describe their field trials in addition to describing the range of environmental conditions. GBIF could consider extending the protocols. International Benchmark Sites Network for Agrotechnology Transfer (IBSNAT), The Agricultural Model Intercomparison Project (AgMIP) and Bioversity are the players in the standardisation descriptors and variables.

3. Data/information products to be produced

GBIF can provide assistance for this data in two ways. First, it can store the crop specific variables such as the plant descriptors and crop ontology terms that are published by Bioversity International in its internal database and relate them to the proper taxonomic information and second, assist with the agriculture diversification through recognising the inter-relationship of alternative crops and crop wild relatives that are currently mediated in relation to all species. GBIF should also assist in data completeness by encouraging gap analysis at national and global levels to assure reliable quality of data with regards to valid attributes of geographic coordinates, time and taxonomy. The users will be able to filter, relate and find information on the new crops for niche modelling, crop potential mapping, etc.

4. Data sources the most used

1-The standards developed by the International Benchmark Sites Network for Agrotechnology Transfer (IBSNAT) project and subsequently revised by the International Consortium for Agricultural Systems Applications (ICASA) were of considerable value for describing experiments (White, *et.al* 2013). The new ICASA v2 can be implemented as extensions. ICASA v2 contains 600 variables and therefore only a core set of it related to experiments can be implemented. Bioversity International is currently mapping Crop Ontology to the ICASA variables for trial management.

2- Also relating to the other types of species that have relationships with the plants/crops of interest for example the parasitic relationship (insect, pest, etc.) with the species of interest, Darwin Core has an extension that can be used to relate the two species and describe their relationships. Implementing the additional information in the Darwin Core for host and parasitic information, is very valuable (Ex. "host: Quercus alba", "parasitoid of:Cyclocephala signaticollis | predator of Apis mellifera" Darwin Core terms, 2015).

3- Another section (or extending the overview section result of the species search) related to physiological and agronomic aspects. As an example the following are the basic information needed for one of the popular crop models (FAO AquaCrop, 2015).

- Number of plants per hectare
- Time from planting to emergence
- Maximum canopy cover
- Time from emergence to start senescence
- Time from emergence to maturity, i.e. length of crop cycle
- Time from emergence to flowering
- Length of the flowering stage
- Maximum effective rooting depth
- Time from sowing to maximum rooting depth
- Reference Harvest Indexes
- Water Productivity

It will be interesting to enrich GBIF data with links to efforts like Crop Ontology that strive to compile and validate standard agricultural concepts and trait dictionaries.

4- GBIF could act as mediator for the linked information on the new food crops through filters and search functions that are designed to accommodate standard agricultural terms related to agronomy. This facility would be very helpful in agricultural diversification projects that are ongoing (for example Crops For the Future mandate on introducing alternative crops). Filter and search mechanism within the GBIF could be extended to include the abovementioned physiological characteristics that can be used for research on the crop modelling using the information on the related crops.

Linking to data standards in agriculture (White, *et al* 2013), the invasive species databases like CABI CropWise database will be useful in distinguishing the invasive species and a link to a database describing pathogens either identified in the accessions or in the locality where the accessions were collected could greatly enrich GBIF data. Interactions between insects and crops: e.g. the sensitive stage is when the crop is at the stage of seedling.

5. Tools the most used

The tools that are already in use by the agricultural community and modellers to facilitate the inter-comparison of the models can be cited in the GBIF section for agrobiodiversity tools (Porter, *et al* 2014). New database like CFF database, under development for underutilised crops, will be able to link easily with GBIF backbone for taxonomy checking and other related information to food crops that can be used for agriculture diversification.

6. Describe the data flow into steps ("to be able to...") indicating who is doing the indicated step (a scientist, a developer, the system?)

To be able to …

7. Identify GBIF role and improvements

i. status of data (quality, coverage)

Rate of adoption of standards in ABD is increasing and this is a very good time to add GBIF data sources to the big picture of ABD research.

ii. Additional attributes in demand

Crop Ontology, ICASA standard variables, IBSNAT, Darwin core extension

iii. Additional sources - > what are the possible connectors between GBIF and the sources

Bioversity International, AgMIP, CFF

iv. What data mobilization is needed? By whom?

GBIF task force on ABD could create a list of possible relevant data and their associated ID to be linked to GBIF data.

→ Is this a priority? if not implemented how will it block the progress?

Link to recommendations #

6.1.1, 6.1.2, 6.1.5, 4.2.4, 6.2.6, 6.4.1, 6.8.2, 6.10.1, 6.10.2, 6.11.1,6.11.2, 6.11.3, 6.12.2, 6.12.3, 6.14.1,6.14.2,6.14.3

Darwin Core Terms: A quick reference guide. (n.d.). Retrieved October 11, 2015, from http://rs.tdwg.org/dwc/terms/#locationindex

FAO AQUACROP. (n.d.). Retrieved October 11, 2015, from http://www.fao.org/nr/water/infores_databases_aquacrop.html

Porter, C. H., Villalobos, C., Holzworth, D., Nelson, R., White, J. W., Athanasiadis, I. N., … Jones, J. W. (2014). Harmonization and translation of crop modeling data to ensure interoperability. *Environmental Modelling & Software*. http://doi.org/10.1016/j.envsoft.2014.09.004.

Raes, D., Steduto, P., Hsiao, T. C., & Fereres, E. (2009). AquaCrop the FAO crop model to simulate yield response to water: II. Main algorithms and software description. *Agronomy Journal*, *101*(3), 438–447.

White, J. W., Hunt, L. A., Boote, K. J., Jones, J. W., Koo, J., Kim, S., … Hoogenboom, G. (2013). Integrated description of agricultural field experiments and production: The ICASA Version 2.0 data standards. *Computers and Electronics in Agriculture*, *96*, 1–12. http://doi.org/10.1016/j.compag.2013.04.003

Use case E: Predict distribution of use of [currently] orphan crops- proposed by Ebrahim Jahanshiri

1. Describe the objective

Objective: To increase known diversity of crop species by predicting the distribution of orphaned crops that will eventually provide enough diversity and choice to improve the traits of available crops. To broaden available crops, provide alternative crops for changing climate, new diseases, etc, providing the suitability index for crops.

Background: To secure food supplies, one solution is to diversify the food crops. Currently there are 7000 species listed as neglected or underutilised crops (Williams, 2002). These crops have the potential to substitute the current food crops or provide additional sustenance. A common hindrance to spread of these crops is the lack of knowledge and distribution of their species and their closed relatives. One important question is that given the basic characteristics of the species and other related information regarding available data on farming communities that grow these crops, where it can be found, so that the diversity that is needed to improve these crops can also be found. Crops For the Future

2. Who are the actors ?

Distribution modellers that specifically work on food crops to use the occurrence data more efficiently. GBIF to provide filters and hierarchy of the relationships with orphaned crops bearing specific tags to recognise the crops and their immediate and wild relatives. Also ABD community providing feedback on the use of GBIF data for their prediction modelling exercise.

3. Data/information products to be produced

The same methodology for mapping the distribution of the species for conservation, landscape restoration can be used specifically to map the distribution of the orphaned species. The focus here however should be on the eco-geography and trait distributions. Ethnobotanic data on the use of the crops will be necessary. Therefore, crop mapping for their traits is a valuable information for the breeders to relate the geography to traits distribution and make inferences on the availability of crop relatives for breeding purposes.

4. Data sources the most used

Inclusion, integration and linking of databases such as Germplasm Resources Information Network (GRIN), National Plant Germplasm System (NPGS), System-wide Information Network for Genetic Resources (SINGER), EURISCO, Genesys, SpeciesLink, JSTOR Plant Sciences, Botanical Research and Herbarium Management System (BRAHMS) and other with emphasis on the crop species is important (Crop Genebank knowledge base, 2015). There are some ethnobotany databases such as international and national ethnobotany databases that could be useful to link to.

5. Tools the most used

GIS and mapping software.

6. Describe the data flow into steps ("to be able to...") indicating who is doing the indicated step (a scientist, a developer, the system?)

- To be able to shortlist the crops and their relatives based on the location (ABD, scientists)
- To be able to retrieve the occurrence crops that are currently orphan and underutilised under at various levels (genera, cultivar/landrace) (GBIF to tag species based on this attribute and ABD scientists (Bioversity Int. and Crops For the Future currently can help with recognising the list of these crops)

7. Identify GBIF role and improvements

i. status of data (quality, coverage)

ii. Additional attributes in demand

Orphaned, underutilized species, checklists of species and landraces, traits, use of the plant, nutritional value

iii. Additional sources - > what are the possible connectors between GBIF and the sources

Bioversity Int., Crops For the Future.

iv. What data mobilization is needed? By whom?

Mediators can be contacted from both of these organisations to further retrieve information on these crops. The mediators will be able to provide the information and start collaboration with GBIF.

→ Is this a priority? if not implemented how will it block the progress?

Link to recommendations #

6.1.1, 6.1.4, 6.1.5, 6.2.4, 6.3.1, 6.4.1, 6.8.1, 6.8.3, 6.14.1

Bibliographic references

Crop Genebank Knowledge Base. (n.d.). Retrieved October 11, 2015, from http://cropgenebank.sgrp.cgiar.org/index.php?option=com_content&view=article&id=662
Williams, J. T. (2002). Global Research on Underutilized Crops. An Assessment of Current Activities and Proposals for Enhanced Activities.

Use case F: Restoration of degraded landscapes and ecosystem services in a given country (Ethiopia) - proposed by Elizabeth Arnaud

1. Describe the objective

'As global population continues to rise, forests and agricultural land must be sustainably managed and more effectively used to satisfy increasing food demands and mitigate carbon emissions (World Resource Institutes Web Site, http://www.wri.org/our-work/project/global-restoration-initiative)'. When land is degraded, many of the benefits that ecosystems provide to local communities and agricultural production are also degraded, food security is compromised and resilience is reduced. Agrobiodiversity can be used to restore important services in agroecosystems e.g. soil quality, nutritional options at landscape level, pest and diseases control, pollination.

The objective is to enable researchers and local authorities to access regionally-relevant data sets on agrobiodiversity, soil, water, pest & diseases, gender-sensitive socio-economy, and policies to enable the identification of a mix of plant species and varieties maximizing the diversity for traits matching the needs of the restoration strategy along with information on seed availability - A major output of the research project will be a **participatory agricultural ecosystem restoration toolkit** supported by information systems. First product will be a Species database storing traits useful for restoration and nutrition-functional diversity in Nile region, available in standards APIs.

2. Who are the actors ?

Data managers, crop modelers, species modelers, restoration experts, local authorities, local communities in charge of the restoration

3. Data/information products to be produced

- Published national species checklist
- Mix of species and Landraces/cultivars lists with traits of interest for restoration of the target area and ecosystems services, and uses by the communities
- Predictive ecogeographic distribution of species
- Identification of data gaps and guidance for targeted inventories/field survey
- Identification of the species dispersal potential/barriers
- Localization of the seed sources

4. Data sources the most used

- Occurrences : GBIF, Collecting mission database, Genesys, national inventories, Kew databases
- Taxon name: GBIF taxonomy backbone, PlantList (Kew), Checklist of CWR, checklist of Neglected and underutilized species, National checklist of endemic species (Ethiopia)
- Traits for restoration : Treedatabase (ICRAF), useful plants (Kew), TRY, FAO databases for Food
- Species distribution, land cover, land use : Geonetwork (FAO)
- National Red list of Threathened species (Ethiopia)

5. Tools the most used

- Taxonomy curation : using Taxonomic Nomenclature Resolution Service v3.0 (TNRS) , The Plant List
- Georeference curation : Geolocate
- Ecological niche modeling : DIVA-GIS, Maxent, Floramap
- Ecogeograhic modelling tools : BioVel, Capfitogen (http://www.planttreaty.org/content/tools-capfitogen International Treaty)
- Ecosystem services : tools of Natural Capital (Invest, Cost)
- Google Earth maps

6. Describe the data flow into steps ("to be able to...") indicating who is doing the indicated step (a scientist, a developer, the system?)

a. To be able to compile a base list of endemic and introduced species, including trees, and a list of landraces/cultivars

1. Download from global, regional and national sources and compile available list of endemic and introduced species lists, cultivars names checklists for the given country where the restoration will take place.
2. Compare with the validated results of community surveys already performed in the area : preferred species/landraces, uses of the plants
3. Resolve taxonomic names using the GBIF taxonomic backbone - Taxonomic reconciliation is a time-consuming but essential step.
4. Classify the species into shrub, tree, grass, succulent, etc
5. Add principal uses: food, fodder, fuelwood, apiculture, intercropping, etc
6. Get traits: ecological, functional, agronomic, inter-species relations
7. Add
    - Crop Wild Relative Status
    - Neglected and Underutilized Species status
    - Invasive species status
    - Nutrition value
    - Threat status (IUCN)

b. **To be able to use this checklist to regularly extract data from GBIF , and other**

prefered sources, and getting occurrences, kml file and any complementary information included in Darwin Core Germplasm

1. Upload the checklists in the GBIF taxonomy Backbone
2. Get the data sets
3. go to a simple pipeline to check the coordinates and improve quality
4. Need to save this checklist in myGBIF space for :
5. reuse on GBIF later and get the updates on data sets
6. for searching additional sources linked to GBIF to get more data (e.g. CWR, NUS, Invasive species, IUCN status, Nutritional value …)
7. Save the search results in myGBIF space

c. To be able to produce predictive mix of species, landraces and cultivars

1. Get quality occurrences for the species distribution in the country, at the site level. Compile occurrence data of key species in the restoration process and link them to environmental data (climate (WORLDCLIM), soil…) to Perform a Species Distribution Model in targeted areas
2. Get time series to show the evolution of the land cover and species distribution from e.g. the last 20 years
3. Run a species eco-geographic model to predict species with traits of interest for the restoration needs and species with an adaptive potential for the targeted restoration.

d. To be able to check the seed availability of the predictive mix of species, landraces and cultivars

1. Locate where seeds are conserved and available
2. Identify if methods for regeneration, seedlings are available

e. To be able to identify the threats and beneficial services in the target regions

1. Get occurences for the target regions for of pest and diseases,
2. Occurrences useful diversity (pollinators, underground diversity)
3. Occurrences of livestocks

7. Identify GBIF role and improvements

GBIF role

Provide an entry point for starting a compilation of species and landraces/cultivars for a target region for restoration. GBIF portal must be a source of quality occurrences and relevant additional attributes (e.g. Darwin Core germplasm) along with a complete taxonomy that includes infraspecies levels and propose taxon name resolvers.
Propose a profile that guides the users seeking for taxon attributes that are relevant for restoration (traits, species status, land cover). Resulting checklist can be uploaded in GBIF and stored in a user's specific space on the portal 'myGBIF'. This user's space should enable storing access to the preferred external sources and running on demand stored queries.

i. status of data (quality, coverage)

- not enough occurrences at national level - Usually a minimum of 20–50 records is needed to produce accurate species distribution models based on 'presence-only' data
- Need a species distribution map per country of higher granularity
- Taxonomy backbone is incomplete: add Mansfield taxonomy, PlantList, etc,

authoritative lists of landraces/cultivar names
- Provide taxonomic name resolvers
- Classification of the species into shrub, tree, grass, succulent, etc currently appears through the load of the wikipedia page and not as a searchable filter

ii. Additional attributes in demand

- National checklists of species, authoritative lists of landraces/cultivars
- Darwin Core germplasm extension with all attributes of the Multi Crop Passport Data
- Crop Wild Relative Status
- Neglected and Underutilized Species status
- Invasive species status
- Nutrition value
- Threat status (IUCN)
- Traits useful for restoration: ecological traits, agronomic traits, nutritional value, usage of the plant
- Seed availability
- Add or link to appropriate controlled vocabularies (traits from the TRY thesaurus, Crop Ontology, Planteome, AGROVOC, CABI, etc)
- add shapefiles for download
- Enable routine to get updates and resolves new names
- Propose cleaning pipelines (e.g. BioVel) and storage of the resulting maps in 'myGBIF'

iii. Additional sources - > what are the connectors

- Genesys – Darwin Core germplasm attributes
- Collecting mission database – taxon name, country, location
- CWR checklist of the Crop wild relatives and climate change Portal – (http://www.cwrdiversity.org/) - taxon name, country
- Treedatabase (ICRAF), Useful Plants database (Kew) – taxon name, country, administrative boundaries, traits
- TRY – taxon names, country, traits
- Geonetwork – country

We need also a discovery system for identifying other relevant sources of data

iv. What data mobilization is needed? By whom?

- Data mobilization on species occurrences at the national level by the GBIF nodes and scientists from herbaria, research institute
- Regional data mobilization by countries sharing borders and similar restoration project
- Publish quality data - Currently, one-third of the names entered into online databases are estimated to be incorrect and about 15% of species names in herbaria specimens are misspelled (Whitfield 2011).
- Collect of relevant maps and remote sensing data by experts
- Collect of community preferences and usage done of the plants and trees by NGOs, extensionists, conservationists engaged in restoration

→ Is this a priority? if not implemented how will it block the progress?	

Link to recommendations #

6.2.1, 6.2.2, 6.2.3, 6.2.4, 6.2.5, 6.3.1, 6.3.2, 6.4.1, 6.5.1, 6.5.2, 6.5.3, 6.5.4, 6.6.1, 6.6.2, 6.6.4, 6.7.1, 6.8.1, 6.8.2, 6.8.3, 6.8.4, 6.10.1, 6.10.2, 6.11.1, 6.11.5, 6.11.6, 6.12.3, 6.13.1, 6.13.2, 6.13.3

Bibliographic references

Vivero J.L., Kelbessa E. and Demissew S. (2005) - The Red List of Endemic Trees & Shrubs of Ethiopia and Eritrea – Fauna & Flora International, Cambridge, UK, booklet, 28 p.

Awas T. - Endemic plants of Ethiopia (2009)- Preliminary working list to contribute to National plant conservation target - Institute of Biodiversity Conservation (IBC) - http://www.ibc.gov.et/wp-content/uploads/downloads/Endemic_plants_of_Ethiopia-Reported.pdf

Climate-Resilient Green Economy (CRGE). (2011). Ethiopia's Climate-Resilient Green Economy, Green Economy Strategy, brochure.

Government of the Federal Democratic Republic of Ethiopia (2013) National Nutrition Programme June 2013-June 2015

Maundu, P.; Bosibori, E.; Kibet, S.; Morimoto, Y.; Odubo, A.; Kapeta, B.; Muiruri, P.; Adeka, R.; Ombonya, J. (2013) UNESCO Safeguarding Intangible Cultural Heritage, 17p

Use case G: Access and benefit sharing (ABS) mechanism (FAO International Treaty) - proposed by Dag Terje Endresen

1. Describe the objective

International agreements and treaties prescribe access and benefit sharing for designated crops (Treaty Annex 1 list of crops) (FAO 2009a). The Convention of Biological Diversity (CBD) establishes biodiversity resources as a national sovereign property. Need to develop mechanisms to assess the use of this genetic resource in commercial activities such as commercial crop improvement breeding companies - so that breeding companies can pay the 1.1% fee on profits (sales gross income) from protected products as prescribed by the multilateral system (MLS) agreement (FAO 2009b).

2. Who are the actors?

- Genebank with germplasm material assigned to be part of the multilateral system (MLS) need to issue a so-called Simple Material Transfer Agreement (SMTA) when distributing MLS material. The SMTA must be registered and shared with the International Treaty (ITPGRFA FAO).
- Germplasm users (such as seed companies, plant breeders and researchers) requesting living germplasm material in the MLS need to sign and comply with the terms and conditions of the SMTA and the ITPGRFA.
- The International Treaty secretariat (ITPGRFA FAO) needs to maintain a global

information system (GLIS) with information about all the material in the MLS and all the SMTAs issued in response to respective seed request.

3. Data/information products to be produced

- A global information system (GLIS) with information about all germplasm material assigned into the multilateral system (MLS), and including in particular the DOI for the genebank germplasm accession and the DOI assigned to the recipient's copy of the living germplasm material.
- SMTA transaction information including information about all (MLS) germplasm material shipped by a genebank in response to a legible seed request from a germplasm user. All germplasm accessions (or similar germplasm material) shipped under SMTA conditions must include the DOI of the genebank accession and the DOI assigned to the recipient's copy of this germplasm material. The recipient will need to keep track of this assigned DOI for reporting use and final products.

4. Data sources the most used

- GeneSys
- EURISCO

5. Tools the most used

- The International Treaty on Plant Genetic Resources for Food and Agriculture (FAO 2009a)
- SMTA guidelines (FAO 2009b)
- Easy SMTA (FAO 2012)

6. Describe the data flow into steps ("to be able to...") indicating who is doing the indicated step (a scientist, a developer, the system?)

To be able to …

7. Identify GBIF role and improvements

i. Status of data (quality, coverage)

ii. Additional attributes in demand

iii. Additional sources - > what are the possible connectors between GBIF and the sources

iv. What data mobilization is needed? By whom?

→ Is this a priority? if not implemented how will it block the progress?

Link to recommendations #

Bibliographic references

FAO (2009a). International Treaty on Plant Genetic Resources for Food and Agriculture. Food and Agriculture Organization of the United Nations, Rome, Italy. Available at http://www.fao.org/docrep/011/i0510e/i0510e00.htm

FAO (2009b). Illustrated version of the Standard Material Transfer Agreement. Food and Agriculture Organization of the United Nations, Rome, Italy. Available at http://www.planttreaty.org/node/1646

FAO (2010). The Second Report on the State of the World's Plant Genetic Resources for Food and Agriculture. Available at: http://www.fao.org/docrep/013/i1500e/i1500e.pdf

FAO (2012). Easy-SMTA: User manual. Food and Agriculture Organization of the United Nations, Rome, Italy. Available at http://www.planttreaty.org/sites/default/files/Easy-SMTA_Manual_en.pdf

Use Case H: Identifying gaps in data for Agrobiodiversity to guide targeted data collect proposed by Jean Cossi Ganglo

1. Describe the objective

The process of data curation reveals important gaps in GBIF mediated data and depending on data uses important data are lost with regards to original data downloaded from GBIF site. It is important to fill the gaps usually identified on data attributes, mainly taxonomic names, geographic coordinates, time collection... Gaps are also noted with regards to spatial / environmental coverage and taxonomic groups.

2. Who are the actors ?

GBIF experts in capacity building, node managers and other data publishers

3. Data/information products to be produced

At national and global levels, increase significantly the percentage of curated data so that at least 80% of GBIF mediated data can be used in Ecological Niche Modelling

4. Data sources the most used

- GBIF data portal
- National inventories
- Checklist of species and landraces/cultivars

5. Tools the most used

1. Google refine
2. Geolocate
3. Taxonomic name resolver

4. Maxent, Floramap, etc

6. Describe the data flow into steps ("to be able to...") indicating who is doing the indicated step (a scientist, a developer, the system?)

To be able to increase the quality data in GBIF

1. Identify with the Agrobiodiversity community the gaps in species occurrences, time series, taxonomic groups, national coverage
2. Capacitate node managers and other data publishers to perform gap analyses on GBIF data at national levels
3. Set priorities on data collections to fill in the gaps identified
4. Train the node managers in collecting the data relevant for Agrobiodiversity

7. Identify GBIF role and improvements

i. status of data (quality, coverage)

ii. Additional attributes in demand

iii. Additional sources - > what are the possible connectors between GBIF and the sources

v. What data mobilization is needed? By whom?

Mobilize occurrences of species of importance for agrobiodiversity and sample-base data, national checklist of species and landraces, trait data

→ Is this a priority? if not implemented how will it block the progress?

Link to recommendations #

6.3.3, 6.4.1, 6.11.1, 6.11.2

Bibliographic references

Appendix 3

Terms of Reference for Task Group on Data Fitness for Use in Agrobiodiversity

The discovery, access and adequate use of primary biodiversity data is critical for informed decision making to achieve sustainable use of agrobiodiversity resources, secure their availability in the future, and address many of the world's key challenges such as feeding a growing human population, and developing a more productive and sustainable agriculture under the scenario of climatic change. It is estimated that various institutions collectively house several billion specimens of more than a million species and cultivated varieties, genetic samples and other important evidence of patterns and trends in global biodiversity. Only a fraction of this vast databank of species information and genetic material is freely and digitally available, and the continuous efforts to use the digitally available information on agrobiodiversity have resulted in a growing – but dispersed – body of scientific tools and literature assisting better understanding of the changing environments and providing the foundation for decision making based on data and scientific evidence.

As part of a broader global strategy on fitness for use of biodiversity data, GBIF is convening a Task Group on Data Fitness for Use in Agrobiodiversity (DFFU-A) to help improve the fit of data related to agrobiodiversity to the variety of important uses required and requested by this community of research and policy. This focus on Agrobiodiversity data results from a long-lasting collaboration with Bioversity International, which is a global research-for-development organization. Bioversity research delivers scientific evidence, management practices and policy options to use and safeguard agricultural and tree biodiversity to attain global food and nutrition security.

DFFU-A will capture the best available experiences, document limitations in existing GBIF services, and suggest improvements in the functionality of GBIF.org for domain-specific needs. Its activities will be informed by ongoing work in GBIF to improve the functionality of its data services by enabling users to filter for relevant data fit for particular purposes, based on pre-defined profiles.

The Task Group is established with three **objectives**:

1. Based on domain specific data use experience, to make recommendations on improving data availability and data use, data mobilization, data and metadata publishing, and data processing. The Task Group is expected to deliver a vision of the ideal data serving the needs of the agrobiodiversity community, as well as past, current and expected data modifications, cleaning steps, analyses and visualization needs,
2. To document best practices from ongoing initiatives using agrobiodiversity related data, and to collect the information on repeatable tools and data management solutions by consulting with ongoing agrobiodiversity and related initiatives, and the broader biodiversity data community, e.g. through GBIF Community Site and TDWG, in order to bring together the different stakeholders, and catalyse activities around use of agrobiodiversity data.

3. Based on information from GBIF Secretariat about current developments relating to quality and fitness for use, to make recommendations on GBIF.org improvements, and to provide guidance in the development of training and outreach materials to help data users to build upon the currently available resources and to share the new developments.

Mandate

1. Develop a schedule and activities for the Task Group
2. Liaise with other experts and define the data use priorities essential for the agrobiodiversity community
3. Liaise with on-going initiatives and projects to document best practices initiatives using agrobiodiversity related data
4. Liaise with potential donor organisations that might be interested in funding data fitness for agrobiodiversity uses
5. Consult with and encourage agrobiodiversity stakeholders to document and share the repeatable tools and data management solutions
6. Provide guidance in the development of training and outreach materials to help data users to build upon the currently available resources and to share the new developments
7. Consult widely and determine key questions that need to be addressed for the community specific data use needs on data availability and data use improvements, data mobilization, data and metadata publishing, and data processing at institutional, national, regional, and global levels
8. Develop the recommendations for the GBIF portal improvements

Outputs

The main deliverable will be a set of practical guidelines and recommendations around the issues defined under the Terms of Reference presented in the form of a report to the Executive Secretary of GBIF, on or before October 1, 2015.

Timeline

The Task Group will operate for a period of six months, and is expected to begin in April 2015 after the members have been confirmed by the Executive Committee. A first report will be presented to GBIF Secretariat by 1 August 2015 and circulated among the attendees of the 22nd GBIF Governing Board meeting. The final report will be delivered to GBIF S by 1 October 2015, after which the Task Group shall be dissolved. The final report will be presented to the 23rd GBIF Governing Board in September/October 2016.

Timeline	*Activity*
March 2015	GBIF EC approves candidate members of the Task Group. Candidate members invited.

April 2015	Establishment of the Task Group. Wireframes for the reports, identification of the priority directions. Initial teleconference. Communication and writing tasks distributed among the Task Group members.
May 2015	Initial inputs from Task Group members collated. Key stakeholders, GBIF, TDWG and agro community invited to contribute: consultation open.
July 2015	Comments / inputs from key stakeholders and broader community collated into the first (intermediate) report. Teleconference meeting of the Task Group to collectively draft first round report
August 1, 2015	First report presented to GBIF S and forwarded to GB 22.
September 2015	Consultation closed. Revise report based on inputs. Face-to-face meeting of the Task Group to write up the final report.
October 1, 2015	Deliver final report

Mode of operation

The Task Group will deliver its recommendations to the GBIF Governing Board through the Secretariat. The Task Group will be coordinated by, and report through, the Programme Officer for Content Analysis and Use at the GBIF Secretariat. The Task Group shall mostly operate remotely through email and associated collaborative tools such as wiki, Skype and conference calls; and hold at least one face-to-face meeting, covering the cost of travel and accommodation and other essential costs using funds available from the GBIF Secretariat as defined in the approved GBIF Work Programme 2014-1016, in particular to consolidate and write its report. The task force may also use other events of interest to hold additional meetings.

Task Group membership

The Task Group shall be comprised of **6 members** reflecting a global representation. The members are invited based on the priority list of candidates offered by the GBIF S and the Chair of the Task Group for the approval by the GBIF Executive Committee. The Task Group is expected to consult widely within the community of agrobiodiversity community and GBIF community, including GBIF stakeholders and the GBIF Secretariat will provide support for this consultation in the form of mailing lists, discussion groups, etc. It is intended to be the nucleus team that will consult widely with other experts, institutions, initiatives & projects. The Task Group is chaired by **Dr Elizabeth Arnaud**, Bioversity International.

GBIF Secretariat contact

Dmitry Schigel, Programme Officer for Content Analysis and Use, e-mail: dschigel@gbif.org