Chapter 42. Information, knowledge and agricultural biodiversity

Dag Endresen

GBIF Norway, UiO Natural History Museum, University of Oslo, Norway

Introduction

Plant genetic resources for food and agriculture include an estimated 7.4 million *ex situ* accessions conserved in genebank collections. An estimated 40% of these accessions are both electronically documented and freely available from online genebank data platforms such as *Genesys* (Alercia and Mackay, 2013; www.genesys-pgr.org/) and *EURISCO* (FAO, 2010; Dias et al, 2011; http://eurisco.ipk-gatersleben.de/). Approximately 21% of the world's flora is classified as a crop wild relative and as such a potential gene donor for crops (Maxted and Kell, 2009). The Global Biodiversity Information Facility (GBIF) (Telenius, 2011) integrates and provides extensive occurrence information about collection data, including genebank accessions *ex situ* and wild plants *in situ,* including many crop wild relatives. However, neither GBIF nor the genebank data portals focus on providing data on the molecular genetic diversity or conservation status of the collections. Some *ex situ* genebank accessions do provide associated measurement data from characterization and evaluation trials. However, the lack of easy access to experimental trait information online continues to be reported as a major limitation to the efficient use of plant genetic resources (FAO, 2010). Data on *ex situ* genebank collections, crop wild relative *in situ* populations, genetic data, and trait measurements are generally created and made available by different sub-groups of practitioners, and each sub-group has showed a tendency to develop its own documentation practices and data standards. This chapter will describe how knowledge organization principles can be used to create a more unified data landscape for agricultural biodiversity. It is concluded that the introduction of persistent and globally unique digital identifiers, resolvable to machine-readable information, and based on a standardized and formally declared data domain model is one of the fundamental first steps for an effective integration of agricultural biodiversity information (FAO, 2014).

What is genetic diversity?

The food crops and farm animals we depend upon for our livelihood are exposed to many stresses such as evolving plant diseases, pests and climate change. Plant breeding programs require access to novel genetic diversity to maintain and improve food crops to keep pace with these challenges (see Ortiz, Chapter 17 of this volume). To select the appropriate genetic diversity to do this, plant breeders and crop scientists need access to relevant information from a wide range of distributed information sources. Heterogeneous data formats such as non-standardized table header labels and database access interfaces makes data integration challenging. In contrast, standardization of data formats and communication protocols makes data acquisition and processing easier. Data interoperability principles describe how heterogeneous systems can exchange data with each other and interpret that data in a manner that is meaningful to the end-user. To make optimal use of the emerging large amounts of information about genetic diversity, we will also need novel approaches and tools for data analysis.

The loss of suitable habitat for the wild relatives of the cultivated plants has raised concerns regarding their survival *in situ* (Iriondo et al, 2008). There are also indications of a gradual replacement of more genetically diverse landraces and traditional cultivars with genetically more uniform modern cultivars (Tanksley and McCouch, 1997; Zeven, 1998). *In situ* and on-farm conservation of plant genetic resources in their natural habitat is generally considered the optimal strategy for maintaining of valuable genetic diversity (Maxted et al, 1997; Brush, 2000). When maintained in their natural environment, plant populations are better able to adapt and respond to changing conditions (Dulloo et al, Chapter 36 of this volume). However, *ex situ* backup for *in situ* populations of crop wild relatives and landraces maintained on farm will also be needed, in particular for two reasons: (1) Many of the *in situ* populations and on farm landraces are threatened and can be lost forever without safety-backup *ex situ*; and (2) the systems established by genebanks for distribution of cultivated genetic diversity for utilization in breeding and research would be an efficient approach for increased mobilization of wild genetic resources also.

The genetic diversity available in the breeders' material has already been intensively explored and there is an emerging need to start looking into traditional landraces (farmers varieties) and crop wild relatives as new sources of genetic materials (Tanksley and McCouch, 1997; Porch et al, 2013). The wider, more general biodiversity community manages much of the relevant information on crop wild relatives. Much of the information on plant biodiversity, including crop wild relatives, is made open and freely available on the Internet by networks such as the Global Biodiversity Information Facility (GBIF; www.gbif.org). The data models and data exchange solutions standardized by the Biodiversity Information Standards (TDWG; www.tdwg.org) society for the Natural History Museums and Botanical Gardens are largely directly compatible with the corresponding solutions for documentation of genetic diversity and genebank collections. Collaborative development of knowledge organization principles and solutions between natural history museums and genebank institutions will not only pool efforts, but also

ensure improved and easier access to a larger amount of relevant and useful information within both communities.

Big Data solutions

The amount of data being produced and made available in the modern world has already reached enormous volumes, and the rate of new data creation does not seem to be slowing down (Marz and Warren, 2015). Data on biological diversity and genetic resources is no exception. In particular, access to unprecedented volumes of molecular genetic data is increasing rapidly. The wide variation of data formats and data models in use makes for substantial challenges when attempting to integrate and analyze these voluminous and dispersed data reservoirs. Attempts to centralize datasets into large central data portals have a tendency to create duplicate sets of data when the links to the source data are broken (Belbin et al, 2013; Mesibov, 2013). The 'tidal wave' of large volumes of data requires completely new strategies for approaching data analysis. The proposed approach is to seek solutions to allow new ways for data to be shared, found and combined with other data to be reused by people or services without the originator ever needing to directly interact with them. These approaches should include technologies to document the context and meaning of data, with resolvable identifiers always provided, and should develop models and serializations to allow different sets of data to be combined more easily.

Linked open data

Tim Berners-Lee is globally famous as the inventor of the Internet (Berners-Lee and Fischetti, 2000). He is also a founding member and administrative director of the World Wide Web Consortium (W3C) where he has contributed to the description of a deployment scheme and best practice guide for publishing linked open data (LOD) on the Internet with the goal to establish a 'Web of Data' (Berners-Lee, 2006; Bizer et al, 2009). According to the best practice guidelines (http://5stardata.info/en/), one star (*) is awarded for simply publishing data on the Web under an open license. If data is published as structured data, e.g. as a MS Excel spreadsheet table, two stars (**) are awarded. Using a non-propriety and open format such as tab-delimited text, comma-separated text (CSV), extensible markup language (XML) or the increasingly popular JavaScript object notation (JSON), will give one more star, to a total of three stars (***). Four stars (****) require the use of URIs (uniform resource identifiers; Berners-Lee et al, 2005) to denote things so that other people can point to them not only by linking to your complete dataset, but by linking all the way to the actual data entity or set of information inside the dataset. The top, five star ranking (*****, 5-star) are awarded for adding links to external data. The best practice for 5-star linked data is to use a structured data model such as RDF (resource descriptions framework).

Semantic web technologies have generated positive expectations in the biodiversity informatics community (Hardisty et al, 2013). However, widely adapted examples of using these technologies for biodiversity data are yet to be completed. The EU INSPIRE directive provides guidelines and examples of environmental and

biodiversity data expressed using RDF and Linked Open Data principles (Tarasova et al, 2015). Baskauf et al (2015, 2016) provide guidelines and examples of how to express and publish biodiversity data using Darwin Core and RDF. For example, the identification of locations where a specimen was collected (dwc:locationID) with persistent identifiers from GeoNames (www.geonames.org) allows the user to discover additional related information about this location by following and resolving a chain of linked identifiers. Persistent identifiers can be used in this manner to build a so-called 'Knowledge Graph' (Singhal, 2012), with decentralized information structured around identified real-world entities based on a relationship graph. The 5-star schema is a useful guideline for moving towards a 'Web of Data' (www.w3.org/2013/data/) with a vision of allowing the user to interact directly with information distributed across the Internet without a central coordinator, in a similar manner as if the data was stored in a local database system. **Box 42.1** *Semantic web technologies* provides an introduction to some of these semantic web technologies (Allemang and Hendler, 2011; Wood et al, 2014).

Box 42.1 *Semantic web technologies*

The Resource Description Framework (RDF; www.w3.org/RDF/) is a standard data model recommended from the W3C for interchange of data on the Web designed to allow for easier data merging even if data is documented using different schema. The RDF data model describe all data as 'triples': subject, predicate and object, or entity-attribute-value, e.g. 'this germplasm' 'is part of' 'this collection'. RDF Schema (RDFS; www.w3.org/TR/rdf-schema/), Web Ontology Language (OWL; www.w3.org/TR/owl-primer) and Simple Knowledge Organization System (SKOS; www.w3.org/TR/skos-primer) are based on RDF and provide languages to express RDF vocabularies. RDFS is a more basic and general-purpose language. OWL is designed to be more expressive with support for more formalized and computable semantics when building ontologies. SKOS is designed for representation of structured controlled vocabularies. SKOS can be used for building a bridge between different types of knowledge representation systems and can also be a suitable and user-friendly technology for exposing and promote the reuse of terminology formally declared using other vocabulary and ontology languages.

Biodiversity information architecture

The Biodiversity Information Standards (TDWG) (formerly the Taxonomic Databases Working Group) is a non-profit association for development and promotion of biodiversity informatics standards. The technical architecture group (TDWG-TAG) of the Biodiversity Information Standards (TDWG) has described three fundamental principles of biodiversity informatics using the metaphor of a three-legged stool (**Figure 42.1** *Biodiversity information architecture illustrated by a three-legged stool*; TDWG, 2006). The first leg (1) represents the ontologies and standardized vocabularies that provide a semantic layer to ensure a common understanding of the data types, data attributes and controlled data values used to describe them. The second leg (2) illustrates the collaborative development of data exchange

technology and standardized protocols for how to publish information. The third leg (3) represents persistent identification technology. Information about physical entities such as biological specimens, events such as material collecting and measurement experiments, ontology and vocabulary terminology, etc. are generally described and documented across distributed data sources. Persistent identifiers will allow data users to connect distributed information about the same things when data publishers reuse the very same identifier names. Identifiers can alternatively be explicitly linked together (e.g. using relationship properties; see **Box 42.2** *Definitions of some relationship properties*). Development of new and cross-dataset, multiple purpose annotation services could also be a good tool for linking identifiers.

Box 42.2 *Definitions of some relationship properties*

owl:sameAs, The property that determines that two given individuals are equal.

rdfs:seeAlso, Further information about the subject resource.

skos:closeMatch, skos:closeMatch is used to link two concepts that are sufficiently similar to be able to be used interchangeably in some information retrieval applications. In order to avoid the possibility of 'compound errors' when combining mappings across more than two concept schemes, skos:closeMatch is not declared to be a transitive property.

skos:exactMatch, skos:exactMatch is used to link two concepts, indicating a high degree of confidence that the concepts can be used interchangeably across a wide range of information retrieval applications. skos:exactMatch is a transitive property, and is a sub-property of skos:closeMatch.

skos:broadMatch, skos:broadMatch is used to state a hierarchical mapping link between two conceptual resources in different concept schemes.

skos:narrowMatch, skos:narrowMatch is used to state a hierarchical mapping link between two conceptual resources in different concept schemes.

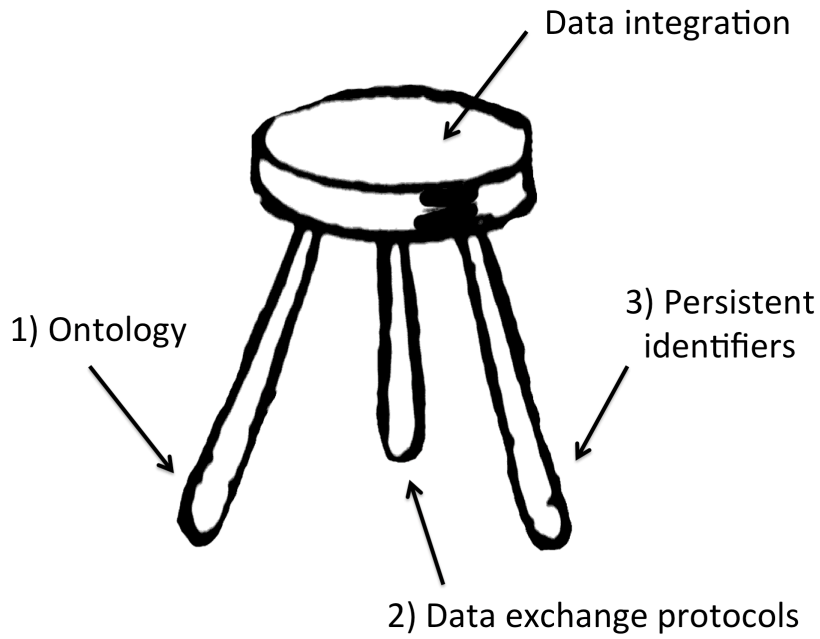


Figure 42.1 *Biodiversity information architecture illustrated by a three-legged stool (TDWG, 2006). Image source: www.clipartbest.com/clipart-MTLM9q6Ta (clipart).*

Ontologies and controlled vocabularies

Ontologies and controlled vocabularies are tools for the development of a shared and agreed standard terminology to organize information and support knowledge representation and management. Ontologies can be developed to organize and categorize physical things such as specimens, genebank accessions, genetic properties, alleles, institutions and people; or events such as the collecting of seed material, trait measurements, and seed distribution (see also **Table 42.1** *Mapping between MCPD, Darwin Core and ABCD 2.06*). Such things and events are here classified as 'classes' (rdfs:Class or owl:Class) in a formal ontology. Ontologies can also include formal descriptions of information attributes such as the scientific name, catalog- or accession-number, or who collected a specimen or genebank accession. When using RDF and ontologies, such information attributes are classified as 'properties' (rdf:Property). When, for example, presenting information in a spreadsheet, the actual physical things or events represented by the information in a record or in a horizontal line is the 'class' and the column headers are 'properties'. Properties are organized into two main types. So-called 'object properties' (owl:ObjectProperty) link individuals to other individuals using URIs, and 'datatype properties' (owl:DataTypeProperty) that link individuals to data values given as literals (www.w3.org/TR/owl-ref/#Property).

The genetic resources and crop genebank communities have achieved wide agreement on using the Multi-Crop Passport descriptor list (MCPD) (Alercia et al, 2015) as a preferred standard data exchange format. The first version of the MCPD was initially introduced in 1997 (Hazekamp et al, 1997) and released in 2001 (Alercia et al, 2001) by the Food and Agriculture Organization of the United Nations (hereafter FAO) and Bioversity International (formerly IBPGR 1974-1991; IPGRI 1991-

2006). The MCPD established a standard set of minimum descriptors for genebank accessions (specimens) of any agricultural crop based on the prior and crop-specific descriptors developed and released by Bioversity International (Bioversity International, 2007; Gotor et al, 2008; Faberova, 2010).

The Biodiversity Information Standards (TDWG) has ratified and released two different controlled vocabularies for the description of specimens and species occurrences with a similar coverage as the MCPD has for genebank accessions. The Access to Biological Collections Data (ABCD) standard was ratified by TDWG in September 2005 (the current version 2.06 was released in 2007) (TDWG, 2007; Holetschek et al, 2012). The Darwin Core standard (DwC) was ratified in October 2009 and is regularly updated with proposed minor and major revisions to individual terms after consensus is reached during an open peer review period of 30 days announced within the TDWG community (TDWG, 2009; Wieczorek et al, 2012).

Both the ABCD and the Darwin Core standards have been mapped to the MCPD (**Table 42.1** *Mapping between MCPD, Darwin Core and ABCD 2.06*) and extended with all the missing unmapped MCPD descriptors (Berendsohn and Knüpffer, 2006; Knüpffer et al, 2007; Endresen and Knüpffer, 2012). The mapping of terms and descriptors between the ABCD, Darwin Core and MCPD is important not least to achieve integration between agrobiodiversity data (including genebank collections) and other large sources of biodiversity information such as the museum and biodiversity monitoring datasets published within the Global Biodiversity Information Facility (GBIF). The GBIF portal is based on the Darwin Core standard and provides a particularly important source for information on crop wild relatives, where expertise and species occurrence data are often found outside of the agrobiodiversity community and agrobiodiversity data portals such as Genesys and EURISCO.

Table 42.1 *Mapping between MCPD, Darwin Core and ABCD 2.06*

Term	MCPD (2015)	Darwin Core (dwc) and Darwin Core germplasm extension (g)	ABCD 2.06 *
NA	(not applicable)	dwc:datasetID	DataSet/DatasetGUID
0	PUID	dwc:occurrenceID	Unit/UnitGUID
1	INSTCODE	dwc:institutionCode	Unit/SourceInstitutionID
2	ACCENUMB	dwc:catalogNumber	Unit/UnitID
3	COLLNUMB	dwc:recordNumber	Unit/CollectorsFieldNumber
4	COLLCODE	g:collectingInstituteID	Unit/Gathering/Agent/Organisation/Name/Abbreviation
4.1	COLLNAME	dwc:recordedBy	Unit/Gathering//GatheringAgent/AgentText
4.1.1	COLLINSTADDRESS	dwc:recordedBy	Unit/Gathering/Agent/Organisation/Name/Text
4.2	COLLMISSID	dwc:collectionCode	Unit/Gathering/Project/ProjectTitle
5	GENUS	dwc:genus	ScientificName/NameAtomised/Botanical/GenusOrMonomial
6	SPECIES	dwc:specificEpithet	ScientificName/NameAtomised/Botanical/FirstEpithet
7	SPAUTHOR	dwc:scientificNameAuthorship (if SUBTAXA is empty)	ScientificName/NameAtomised/Botanical/AuthorTeamParenthesis + ScientificName/NameAtomised/Botanical/AuthorTeam (if SUBTAXA is empty)
8	SUBTAXA	dwc:infraspecificEpithet	ScientificName/NameAtomised/Botanical/Rank + ScientificName/NameAtomised/Botanical/SecondEpithet
9	SUBTAUTHOR	dwc:scientificNameAuthorship	ScientificName/NameAtomised/Botanical/AuthorTeamParenthesi

			s + ScientificName/NameAtomised/Botanical/AuthorTeam
10	CROPNAME	dwc:vernacularName	TaxonIdentified/InformalNameString
11	ACCENAME	g:breedingIdentifier	ScientificName/NameAtomised/Botanical/CultivarGroupName + ';' + ScientificName/NameAtomised/Botanical/CultivarName + ';' + ScientificName/NameAtomised/Botanical/TradeDesignationName (s)
12	ACQDATE	g:acquisitionDate	Unit/SpecimenUnit/Acquisition/AcquisitionDate
13	ORIGCTY	dwc:countryCode	Unit/Gathering/Country/ISO3166Code
14	COLLSITE	dwc:locality	Unit/Gathering/LocalityText
15.1	DECLATITUDE	dwc:decimalLatitude	Unit/Gathering/SiteCoordinateSets/SiteCoordinates/CoordinatesLatLon/LatitudeDecimal
15.2	LATITUDE	dwc:verbatimLatitude	Unit/Gathering/SiteCoordinateSets/SiteCoordinates/CoordinatesLatLon/VerbatimLatitude
15.3	DECLONGITUDE	dwc:decimalLongitude	Unit/Gathering/SiteCoordinateSets/SiteCoordinates/CoordinatesLatLon/LongitudeDecimal
15.4	LONGITUDE	dwc:verbatimLongitude	Unit/Gathering/SiteCoordinateSets/SiteCoordinates/CoordinatesLatLon/VerbatimLongitude
15.5	COORDUNCERT	dwc:coordinateUncertaintyInMeters	Unit/Gathering/SiteCoordinateSets/SiteCoordinates/CoordinatesLatLon/CoordinateErrorDistanceInMeters
15.6	COORDDATUM	dwc:geodetic.Datum	Unit/Gathering/SiteCoordinateSets/SiteCoordinates/CoordinatesLatLon/SpatialDatum
15.7	GEOREFMETH	dwc:georeferenceSources	Unit/Gathering/SiteCoordinateSets/SiteCoordinates/GeoreferenceSources
16	ELEVATION	dwc:minimumElevationInMeters	Unit/Gathering/Altitude/MeasurementAtomised/MeasurementLowerValue + Unit/Gathering/Altitude/MeasurementAtomised/MeasurementScale set to 'm'
17	COLLDATE	dwc:eventDate	Unit/Gathering/DateTime/ISODateTimeBegin
18	BREDCODE	g:breedingInstituteID	Unit/PlantGeneticResourcesUnit/BreedingInstitutionCode
18.1	BREDNAME	g:breedingInstitute	Unit/PlantGeneticResourcesUnit/DecodedBreedingInstitute
19	SAMPSTAT	g:biologicalStatus	Unit/PlantGeneticResourcesUnit/BiologicalStatus
20	ANCEST	g:ancestralData, g:purdyPedigree	Unit/PlantGeneticResourcesUnit/AncestralData
21	COLLSRC	g:acquisitionSource	Unit/PlantGeneticResourcesUnit/CollectingAcquisitionSource
22	DONORCODE	g:donorInstituteID	Unit/SpecimenUnit/History/PreviousUnit(s)/PreviousSourceInstitutionID
22.1	DONORNAME	g:donorInstitute	Unit/PlantGeneticResourcesUnit/DecodedDonorInstitute
23	DONORNUMB	g:donorsIdentifier	Unit/SpecimenUnit/History/PreviousUnit(s)/PreviousUnitID
24	OTHERNUMB	dwc:otherCatalogNumbers	Unit/PlantGeneticResourcesUnit/OtherIdentification
25	DUPLSITE	g:safetyDuplicationInstituteID	Unit/PlantGeneticResourcesUnit/LocationSafetyDuplicates
25.1	DUPLINSTNAME	g:safetyDuplicationInstitute	Unit/PlantGeneticResourcesUnit/DecodedSafetyDuplicationLocation
26	STORAGE	g:storageCondition	Unit/PlantGeneticResourcesUnit/TypeGermplasmStorage
27	MLSSTAT	g:mlsStatus	(missing)
28	REMARKS	dwc:occurrenceRemarks	Unit/Notes

* 'Unit/' = 'Datasets/Dataset/Units/Unit/'; 'ScientificName/' = 'Unit/Identifications/Identification/Result/TaxonIdentified/ScientificName/'. Table generated by the author based on mapping between MCPD and ABCD made in collaboration with Javier de la Torres (Bioversity International) at the BioCASE Wiki; and mapping between MCPD/EURISCO presented by Walter Berendsohn and Helmut Knüpffer (2006). This mapping is described in Endresen and Knüpffer (2012).

The Dublin Core metadata terms were developed by the Dublin Core Metadata Initiative (DCMI; http://dublincore.org/documents/dcmi-terms/) and provide a vocabulary of properties, classes and controlled values for use in resource descriptions (Weibel et al, 1998; Kunze and Baker, 2007). The current version of the Dublin Core terms were revised and released in 2008 as the so-called 'terms namespace' (dct = http://purl.org/dc/terms/). This revision was a refinement of the element terms for the purpose of harmonization with RDF technology. Darwin Core is based on the Dublin Core standard and should be viewed as an extension of the Dublin Core for biodiversity information (Wieczorek et al, 2012). Both Dublin Core and Darwin Core are declared using RDF. Darwin Core is itself designed to accommodate extensions to expand the core set of terms to meet requirements from sub-communities such as the agrobiodiversity community. The missing and un-mapped descriptors from the MCPD were declared as SKOS and released as the Darwin Core germplasm extension for genebanks (Endresen and Knüpffer, 2012; http://terms.tdwg.org/wiki/Germplasm). The germplasm vocabulary thus provides a bridge between the MCPD and Darwin Core.

The Darwin Core vocabulary was initially developed as a pragmatic solution to facilitate standardized biodiversity data exchange using flat text files or XML. To facilitate and promote the use of Darwin Core terms to describe and publish biodiversity collections data as RDF, Baskauf and Webb (2015) have developed a Darwin Core ontology extension, Darwin Core Semantic Web (Darwin-SW or DSW). The DSW influenced some major revisions of the Darwin Core standard and established a more explicit domain model e.g. to separate collection specimens from observations on living organisms in the wild, and the distinction between real-world species occurrences and different types of evidence for species occurrences. This allows the use to, for example, introduce a data model that explicitly allows for more than one set of evidence for the same real-world species occurrence.

Further ontology refinements to the term listing of vocabularies such as the Darwin Core and MCPD have been developed. The Crop Ontology (CO) (Shrestha et al, 2010) includes an RDF representation in the OWL language for the long-standing Bioversity crop descriptor lists and includes OWL representations for the MCPD terms. The CO could be seen as an emerging bridge to the Open Biological and Biomedical Ontologies (OBO) Foundry (Smith et al, 2007; http://www.obofoundry.org/) including ontologies such as the Plant Ontology (PO) (Cooper et al, 2013) and the Plant Trait Ontology (TO) (Jaiswal et al, 2002; Arnaud et al, 2012). The OBO Foundry establishes a set of principles for building semantic interoperability between ontologies. The Biological Collections Ontology (BCO) (Walls et al, 2014) established a bridge between the traditional specimen-based museum collections described using Darwin Core and the genomic information resources described using Gene Ontology (GO) (Ashburner et al, 2000) and the Minimum Information for any (x) Sequence (MIxS) ontology (Yilmaz et al, 2011). The GO and BCO ontologies were also developed within the OBO Foundry system.

FAO had already initiated a multilingual agricultural thesaurus (AGROVOC) in the early 1980s (FAO, 2016). AGROVOC was published online around 2000 and today includes more than 32,000 concepts presented online as Linked Open Data using the SKOS language (Caracciolo et al, 2013).

Data exchange protocols

Even when data attributes and types are standardized using Darwin Core or other vocabularies and ontologies, a user wanting to access and integrate dispersed data published by other people will typically find a number of different access types and file formats. The 5-star Linked Open Data guideline promotes RDF as a data exchange model (Berners-Lee, 2006). Baskauf et al (2015, 2016) have developed best practice guidelines for publishing specimen collection data with the RDF model. However, at the present time very little biodiversity data, including genebank accession data, is published as RDF. A typical solution to facilitate a more homogenous data interface has been to develop and implement standardized data publishing software. When the EURISCO genebank data portal for Europe was developed, the data publishing procedure was dependent on a manual procedure where each national focal point would collect the genebank accession inventory from each genebank in the respective country, harmonize and combine datasets before the national inventory was uploaded to EURISCO (Faberova, 2010). Other agrobiodiversity data portals, including Genesys, have typically followed very similar data publication routines dependent on manual data updates.

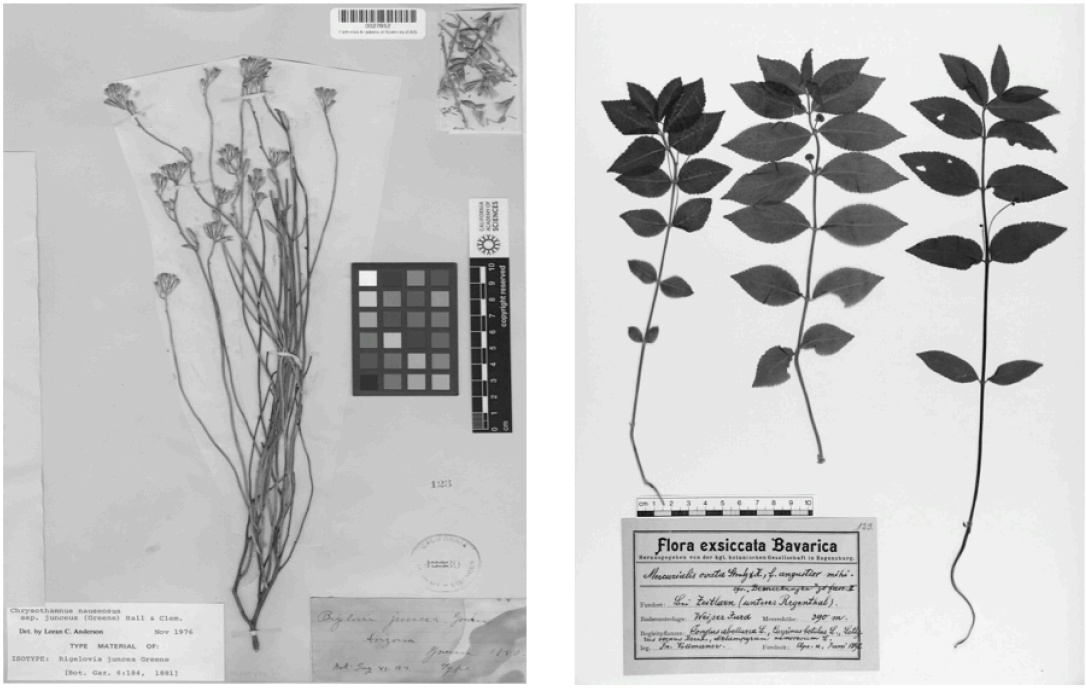
When the GBIF data portal was created and released around 2003, it was based on two of the existing data publishing software packages, DiGIR and BioCASE. The Distributed Generic Information Retrieval (DiGIR) open source software package was initiated at the TDWG conference in Frankfurt in 2000 to publish natural history collections datasets standardized using Darwin Core online as XML (extensible markup language) (Stein and Wieczorek, 2004). The BioCASE (Biological Collections Access Service) (Holetschek et al, 2012) is typically used for biodiversity data, using the ABCD standard online as XML. Typically, this data-publishing approach will require the user of the data service to page through XML responses of e.g. 1000 records at a time until all records have been downloaded. For large datasets and filter conditions matching many data records, retrieving the search results through paging can be time-consuming and might also be vulnerable to page loss. The GBIF Integrated Publishing Toolkit (IPT) (Robertson et al, 2014) therefore introduced the Darwin Core Archive exchange format. The Darwin Core Archive is a compressed zip archive including one or more data files (csv), a single document to describe the relationship between data files (meta.xml) and one document containing metadata that actually describes the dataset (generally expressed using the ecological metadata language, EML). IPT version 1.0 was released in February 2009, with an improved version 2.0 officially released in February 2011. The current version of the IPT will generate globally unique and resolvable Digital Object Identifier (DOI) for each published dataset to support easier data citation and data usage tracking.

Formats such as the JSON-LD (JavaScript Object Notation for Linked Data) (Sporny et al, 2014) for publishing RDF data on the Web is currently gaining popularity and supporting software implementations. RDF data models using serializations such as the JSON-LD could become the preferred data exchange format also for agrobiodiversity datasets. The *Web of Data* using RDF or similar models will depend

on the successful implementation of globally unique and resolvable identifiers to be described further in the next section.

Globally unique identifiers

Accession numbers are human-readable identifiers that have long-standing and proven utility. However, when combining many genebank collections in large integrated information systems it is very soon discovered that the same alphanumerical string might often be used as the accession number identifier for more than one genebank, often to represent entirely different species. The Genesys portal provides a total of 56 different genebank accessions (from 39 different species and 40 different genebank institutes), all of which have been assigned an identical accession number '123' (www.genesys-pgr.org/explore?filter={%22acceNumb%22:[%22123%22]}). When combining genebank accession information with information sources from the larger biodiversity community, such as occurs in GBIF, the problem of non-unique accession number alphanumeric name strings is even further amplified. The GBIF portal provides a total of 1362 occurrence records with catalog number '123' (GBIF, 2016; www.gbif.org/occurrence/search?CATALOG_NUMBER=123; **Figure 42.2** *Two of the 1362 specimens published in GBIF with catalog Number = 123*).



a) *Bigelowia juncea* b) *Mercurialis ovata*

Figure 42.2 *Two of the 1362 specimens published in GBIF with catalog Number = 123.* (**a**) Bigelowia juncea *Greene, catalogNumber = 123, occurrenceID = urn:catalog:CAS:BOT:123, data publisher = Department of Botany, California Academy of Sciences, accessed at [www.gbif.org/occurrence/543392241] (CAS Botany, 2016; doi:10.15468/7gudyo).* (**b**) Mercurialis ovata *Sternb. & Hoppe, catalog*

number = 123, data publisher = Herbarium der Regensburgischen Botanischen Gesellschaft (© H. Gigglberger), Universität Regensburg, Germany, accessed at [www.gbif.org/occurrence/283363] (Herbarium der Regensburgischen Botanischen Gesellschaft, 2016; doi:10.15468/dnmpiw).

Globally unique and resolvable *persistent identifiers* for genebank accessions would enable information about the same physical accession to be published without central coordination on an open platform such as the World Wide Web and to be linked together through the principles of Linked Open Data (http://linkeddata.org/). Persistent identifier names must be (1) globally unique, (2) resolvable (machine actionable on the World Wide Web), and (3) demonstrate a long-term commitment on providing or enabling persistent access to data and associated metadata. Many different approaches and identifier syntax-schema have already been developed, for an overview of selected identifier schemes (FAO, 2014). Guidelines for deployment in biodiversity informatics and specimen databases are in progress (Page, 2008, 2009; Richards et al, 2011; Hagedorn et al, 2013; Guralnick et al, 2015). The good news is that starting to use almost any of these identifier technologies will immediately provide major benefits and new opportunities for data management and data exchange. Different types of identifier names and resolution services could be mapped, new resolution protocols and formats can be added later to be discovered and used e.g. through (HTTP-) content negotiation, and initial resolution services could be redirected to other emerging resolution services.

The W3C promotes the use of HTTP (hypertext transfer protocol)-URIs (Universal Resource Identifiers) as a general principle for identifier technologies. HTTP-URIs are a good and pragmatic solution and can easily be resolved directly using the Internet. However, embedding the method for identifier resolution directly inside the identifier name-string might lead to undesired limitations during the expected lifetime of the identifiers. Persistent identifier solutions must be designed to last for a very long time. Even long after the physical genebank accessions themselves might be lost, information about them and derived genetic resources and cultivars might reference previous accessions. A good strategy would be to think of the HTTP-URI identifier name string as being composed of two parts, a http-resolver-authority and another part that is persistent and globally unique by itself, irrespective of the prefixed resolver part (http://resolver-authority + globally-unique-persistent-identifier).

The Life Science Identifier (LSID) scheme (Clark et al, 2004) was specifically designed to identify biological specimens and enabled the reuse of locally unique names such as accession numbers and specimen catalog number as the 'objectID' of the identifier name string (urn:lsid:[authority]:[namespace]:[objectID]). LSIDs are a special form of URNs and thus compatible with RDF data models. However, the LSID resolution system is not compatible with the HTTP-URI recommended by the W3C (because LSID is proposed as a new Internet transfer protocol at the same hierarchical level as the HTTP protocol).

The Digital Object Identifier (DOI) system (DOI Foundation, 2016; www.doi.org) is based on the Handle system (http://handle.net/) and was originally developed and

introduced in 2000 by the publishing industry for digital content on the Internet. The centralized DOI system guarantees that DOI name strings are globally unique within the context of the DOI system. Official DOI registration agencies such as DataCite or CrossRef operate services to request and register new DOI names together with descriptive metadata on the object. As of the time of writing (January, 2016) more then 120 million DOI names have been registered and the annual growth rate is 18%. The DOI system operates a global DOI resolver service at 'http://doi.org'. DOI names '<doi>' are generally presented either with a simple prefix: 'doi:<doi>', or as the HTTP-URI form by prefixing with the resolver address: 'http://doi.org/<doi>'. DOI names have been suggested (FAO, 2014) as the preferred object identifier system for the development of the Global Information System (GLIS) on plant genetic resources for food and agriculture (PGRFA) referred to in Article 17 of the International Treaty (FAO, 2009).

A Universally unique identifier (UUID) (Leach et al, 2005) is a 128-bit (16 byte) number typically displayed using the canonical format of 32 hexadecimal digits displayed in five character groups separated by four hyphens. UUIDs can be generated by anybody in a distributed network without central coordination with a very high probability that the number generated is globally unique and will not be unintentionally created again. UUIDs are widely used and tools to generate them are very common across different computer platforms. 'Uniform Resource Names (URNs) are a type or subset of Uniform Resource Identifiers (URIs) that use the 'urn' scheme and '*are intended to serve as persistent, location-independent, resource identifiers*' (Moats, 1997). UUID is formally registered as an URN namespace (Leach et al, 2005) and already widely used as object identifiers across many different domains including biodiversity informatics (Hagedorn et al, 2013). The Genesys portal introduced in April 2015 (Genesys, 2015) PURL prefixed UUIDs to persistently identify all genebank accessions included in the portal. Note that the PURL prefixed UUID only creates a complementary machine-readable identifier for the genebank accession and that the UUID is not intended to *replace* the genebank accession number as the preferred human-readable identifier.

Conclusion

A key step towards implementing semantic web technologies for genebank data is to establish and use persistent identifiers for your own collections, and to reuse persistent identifiers from external systems as often as possible (Guralnick et al, 2015). Widespread implementation and use of semantic web technologies for biodiversity information have so far been slow to happen. However, the rapidly growing volumes of data produced and made available will demand new data management practices where storing locally cached copies of external data will rapidly become less attractive (Marz and Warren, 2015). Further harmonization and common standardized solutions for data exchange in the larger biodiversity informatics community, including information on genetic resources, is needed because researchers and other users of these data are anticipated to require seamless access to increasingly larger volumes of data maintained and accessed directly from an heterogeneous network of data sources (Hardisty et al, 2013).

References

Allemang, D. and Hendler, J. (2011) *Semantic web for the working ontologist, second edition: Effective modeling in RDFS and OWL*, Morgan Kaufmann, MA, USA

Alercia, A. and Mackay, M. (2013) 'A gateway to plant genetic resources utilization', *Acta Horticulturae,* vol 983, pp25-30

Alercia, A., Diulgheroff, S. and Metz, T. (2001) 'FAO/IPGRI Multi-crop passport descriptors, December 2001', International Plant Genetic Resources Institute (IPGRI) and Food and Agriculture Organization of the United Nations (FAO), Rome, Italy. www.bioversityinternational.org/e-library/publications/detail/faoipgri-multi-crop-passport-descriptors-mcpd/, accessed 8 Jan 2016

Alercia, A., Diulgheroff, S., and Mackay, M. (2015) 'FAO/Bioversity multi-crop passport descriptors V.2.1 [MCPD V.2.1]', FAO/Bioversity International, Rome, Italy. www.bioversityinternational.org/e-library/publications/detail/faobioversity-multi-crop-passport-descriptors-v21-mcpd-v21/, accessed 8 January 2016, doi: 10.13140/RG.2.1.4280.2001

Arnaud, E., Cooper, L., Shrestha, R., Menda, N., Nelson, R. T., Matteis, L., Skofic, M., Bastow, R., Jaiswal, P., Mueller, L. and McLaren, G. (2012) 'Towards a reference plant trait ontology for modeling knowledge of plant traits and phenotypes', in *KEOD 2012 – Proceedings of the International Conference on Knowledge Engineering and Ontology Development*, http://wrap.warwick.ac.uk/id/eprint/59831, accessed 8 January 2016, pp220-225

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. and Sherlock, G. (2000) 'Gene Ontology: tool for the unification of biology', *Nature Genetics*, vol 25, pp25-29

Baskauf, S. J. and Webb, C. O. (2015) 'Darwin-SW: Darwin Core-based terms for expressing biodiversity data as RDF', *Semantic Web*, pre-print, pp1-15, doi: 10.3233/SW-150203

Baskauf, S. J., Wieczorek, J., Deck, J. and Webb, C. O. (2015) 'Lessons learned from adapting the Darwin Core vocabulary standard for use in RDF', *Semantic Web*, pre-print, pp1-11, doi: 10.3233/SW-150199

Baskauf, S., Wieczorek, J., Deck, J., Webb, C., Morris, P. J. and Schildhauer, M. (2016) 'Darwin Core RDF guide', Biodiversity Information Standards (TDWG), http://rs.tdwg.org/dwc/terms/guides/rdf/, accessed 8 January 2016

Belbin, L., Daly, J., Hirsch, T., Hobern, D. and La Salle, J. (2013) 'A specialist's audit of aggregated occurrence records: An 'aggregator's' perspective', *Zookeys*, vol 305, pp67–76

Berendsohn, W. and Knüpffer, H. (2006) 'Draft mapping of Eurisco descriptors to ABCD 2.06', www.bgbm.org/tdwg/codata/Schema/Mappings/EURISCO-2-ABCD.pdf, accessed 8 January 2016

Berners-Lee, T. (2006) 'Linked data', www.w3.org/DesignIssues/LinkedData.html, accessed 8 January 2016

Berners-Lee, T., Fielding, R. and Masinter, L. (2005) 'Uniform resource identifiers (URI): Generic syntax', Internet Engineering Task Force (IETF), Freemont, CA, USA, http://tools.ietf.org/html/rfc3986, accessed 8 January 2016, doi: 10.17487/RFC3986

Berners-Lee, T. and Fischetti, M. (2000) *Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by Its Inventor*, HarperBusiness, New York, USA

Bioversity International (2007) *Development of crop descriptor lists: Guidelines for developers*, Bioversity International, Maccarese, Italy, ISBN: 978-92-9043-792-1, www.bioversityinternational.org/e-library/publications/detail/developing-crop-descriptor-lists/, accessed 8 January 2016

Bizer, C., Heath, T. and Berners-Lee, T. (2009) 'Linked data – the story so far', *International Journal on Semantic Web and Information Systems (IJSWIS)*, vol 5, pp1-22

Brush, S. B. (ed.) (2000) *Genes in the Field: On-farm Conservation of Crop Diversity*, Lewis Publishers, CRC Press, Boca Raton, FL, USA

Caracciolo, C., Stellato, A., Morshed, A., Johannsen, G., Rajbahndari, S., Jaques, Y. and Keizer, J. (2013) 'The AGROVOC linked dataset', *Semantic Web*, vol 4, pp341-348

CAS Botany (2016) California Academy of Sciences: CAS Botany (BOT), CA, USA, http://www.gbif.org/occurrence/543392241, accessed 21 January 2016, doi: 10.15468/7gudyo

Clark, T., Martin, S. and Liefeld, T. (2004) 'Globally distributed object identification for biological knowledgebases', *Briefings in Bioinformatics*, vol 5, pp59-70

Cooper, L., Walls, R. L., Elser, J., Gandolfo, M. A., Stevenson, D. W., Smith, B., Preece, J., Athreya, B., Mungall, C. J., Rensing, S., Hiss, M., Lang, D., Reski, R., Berardini, T. Z., Li, D., Huala, E., Schaeffer, M., Menda, N., Arnaud, E., Shrestha, R., Yamazaki, Y. and Jaiswal, P. (2013) 'The Plant Ontology as a tool for comparative plant anatomy and genomic analyses', *Plant and Cell Physiology*, vol 54, pp1-23

Dias, S., Dulloo, M. E. and Arnaud, E. (2011) 'The role of EURISCO in promoting use of agricultural biodiversity', in N. Maxted, M. E. Dulloo, B. V. Ford-Lloyd, L. Frese, J. Iriondo, and M. A. A. P. de Carvalho (eds) *Agrobiodiversity Conservation: Securing the Diversity of Crop Wild Relatives and Landraces*, CABI, UK

DOI Foundation (2016) '*DOI handbook*', International DOI Foundation, UK, www.doi.org/hb.html, accessed 21 February 2016, doi: 10.1000/182

Endresen, D. T. F. and Knüpffer, H. (2012) 'The Darwin Core extension for genebanks opens up new opportunities for sharing genebank data sets', *Biodiversity Informatics*, vol 8, pp11-29,

Faberova, I. (2010) 'Standard descriptors and EURISCO development', *Czech Journal of Genetics and Plant Breeding*, vol 46, S106-S109

FAO (2009) '*International Treaty on Plant Genetic Resources for Food and Agriculture'*, FAO, Rome, Italy, www.fao.org/docrep/011/i0510e/i0510e00.htm, accessed 8 January 2016

FAO (2010) '*The second report on the state of the world's plant genetic resources for food and agriculture'*, Commission on Genetic Resources for Food and Agriculture (CGRFA), FAO, Rome, Italy, ISBN: 978-92-5-106534-1, www.fao.org/docrep/013/i1500e/i1500e00.htm, accessed 8 January 2016

FAO (2014) '*Technical options to facilitate the establishment of data links in the field of plant genetic resources for food and agriculture: Permanent unique identifiers, IT/COGIS-1/15/3, November 2014'*, International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA), FAO, Rome, Italy, www.planttreaty.org/sites/default/files/cogis1w3.pdf, accessed 8 January 2016

FAO (2016) 'AGROVOC Multilingual agricultural thesaurus', Agricultural Information Management Standards (AIMS), FAO, Rome, Italy, http://aims.fao.org/vest-registry/vocabularies/agrovoc-multilingual-agricultural-thesaurus, accessed 21 February 2016

Field Museum of Natural History (2016) *J. F. Macbride's Historical Photographs (1929-1939) of Type Specimens from Berlin (B),* www.gbif.org/occurrence/1211536059, accessed 21 February 2016, doi: 10.15468/c4krdu

GBIF (2016) 'GBIF Occurrence Download: CATALOG_NUMBER=123, Search results', http://doi.org/10.15468/dl.cccmwb, accessed 21 February 2016

Genesys (2015) 'Database upgrade completed', www.genesys-pgr.org/content/news/38/database-upgrade-completed, last updated 16th April 2015, accessed 18th February 2016

Gotor, E., Alercia, A., Rao V. R., Watts, J., and Caracciolo, F. (2008) 'The scientific information activity of Bioversity International: the descriptor lists', *Genetic Resources and Crop Evolution*, vol 55, pp757-772

Guralnick, R. P., Cellinese, N., Deck, J., Pyle, R. L., Kunze, J., Penev, L., Walls, R., Hagedorn, G., Agosti, D., Wieczorek, J., Catapano, T. and Page, R. D. M. (2015) 'Community next steps for making globally unique identifiers work for biocollections data', *ZooKeys*, vol 494, pp133–154

Hagedorn, G., Catapano, T., Güntsch, A., Mietchen, D., Endresen, D., Sierra, S., Groom, Q., Biserkov, J., Glöckler, F. and Morris, R. (2013) 'Best practices for stable URIs', http://wiki.pro-ibiosphere.eu/wiki/Best_practices_for_stable_URIs, accessed 8 January 2016

Hardisty, A., Roberts D. and The Biodiversity Informatics Community (2013) 'A decadal view of biodiversity informatics: challenges and priorities', *BMC Ecology*, vol 13, pp1-23

Hazekamp, T., Serwinski, J. and Alercia, A. (1997) 'Appendix II. Multicrop passport descriptors (final version)', in E. Lipma, M. W. M. Jongen, T. J. L. van Hintum, T. Grass and L. Maggioni (eds) *Central Crop Databases: Tools for Plant Genetic Resources Management*, International Plant Genetic Resources Institute (IPGRI), Rome, Italy and WUR Centre for Genetic Resources (CGN), Wageningen, the Netherlands

Holetschek, J., Dröge, G., Güntsch, A. and Berendsohn, W. G. (2012) 'The ABCD of primary biodiversity data access', *Plant Biosystems,* vol 146, pp771-779

Iriondo, J. M., Maxted, N. and Dulloo, M. E. (2008) *Conserving Plant Genetic Diversity in Protected Areas*, CABI, UK

Jaiswal, P., Ware, D., Ni, J., Chang, K., Zhao, W., Schmidt, S., Pan, X., Clark, K., Teytelman, L., Cartinhour, S., Stein, L. and McCouch, S. (2002) 'Gramene: Development and integration of trait and gene ontologies for rice', *Comparative and Functional Genomics*, vol 3, pp132-136

Knüpffer, H., Endresen, D. T. F., Faberova, I. and Gaiji, S. (2007) 'Integrating Genebanks Into Biodiversity Information Networks', in *Proceedings of the 18th EUCARPIA conference, Genetic resources section, Plant genetic resources and their exploitation in the plant breeding for food and agriculture*, Piešťany, Slovak Republic, ISBN: 9788088872634, doi: 10.13140/2.1.4172.8960, pp34-35

Kunze, J. and Baker, T. (2007) *The Dublin Core Metadata Element Set, RFC 5013*, Internet Engineering Task Force (IETF), Freemont, CA, USA, http://tools.ietf.org/html/rfc5013, accessed 8 January 2016, doi: 10.17487/RFC5013

Leach, P., Mealling, M. and Salz, R. (2005) *A Universally Unique Identifier (UUID) URN Namespace, RFC 4122*, Internet Engineering Task Force (IETF), Freemont, CA, USA, http://tools.ietf.org/html/rfc4122, accessed 8 January 2016, doi: 10.17487/RFC4122

Marz, N. and Warren, J. (2015) *Big Data: Principles and Best Practices of Scalable Real-Time Data Systems*, Manning Publications Co., Shelter Island, NY, USA

Maxted, N., Ford-Lloyd, B. V. and Hawkes, J. G. (eds) (1997) *Plant Genetic Conservation: The In Situ Approach*, Chapman & Hall, London, UK

Maxted, N. and Kell, S. (2009) '*Establishment of a global network for the in situ conservation of crop wild relatives: Status and needs'*, FAO, Rome, Italy, www.fao.org/docrep/013/i1500e/i1500e18a.pdf, accessed 8 January 2016

Mesibov, R. (2013) 'A specialist's audit of aggregated occurrence records', *ZooKeys*, vol 293, pp1–18

Moats, R. (1997) *URN Syntax. RFC 2141*, Internet Engineering Task Force (IETF), Freemont, CA, USA, http://tools.ietf.org/html/rfc2141, accessed 8 January 2016, doi: 10.17487/RFC2141

OAC-BIO Herbarium (2016) 'OAC-Herbarium, Biodiversity Institute of Ontario from University of Guelph', www.gbif.org/occurrence/931031820, accessed 21 February 2016, doi: 10.5886/66f3rsta

Page, R. D. M. (2008) 'Biodiversity informatics: the challenge of linking data and the role of shared identifiers', *Briefings in Bioinformatics*, vol 9, pp345–354

Page, R. D. M. (2009) 'bioGUID: resolving, discovering, and minting identifiers for biodiversity informatics', *BMC Bioinformatics*, vol 10, S5, doi: 10.1186/1471-2105-10-s14-s5

Porch, T. G., Beaver, J. S., Debouck, D. G., Jackson, S. A., Kelly, J. D. and Dempewolf, H. (2013) 'Use of wild relatives and closely related species to adapt common bean to climate change', *Agronomy*, vol 3, pp433-461

Richards, K., White, R., Nicolson, N. and Pyle, R. (2011) '*Beginners' Guide to Persistent Identifiers: Version 1.0*', Global Biodiversity Information Facility (GBIF), Copenhagen, Denmark, www.gbif.org/resource/80575, accessed 8 January 2016, ISBN: 87-92020-14-3

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*, vol 9:e102623,

Shrestha, R., Arnaud, E., Mauleon, R., Senger, M., Davenport, G. F., Hancock, D., Morrison, N., Bruskiewich, R. and McLaren, G. (2010) 'Multifunctional crop trait ontology for breeders' data: field book, annotation, data discovery and semantic enrichment of the literature', *AoB PLANTS*, vol 2010, plq008

Singhal A. (2012) 'Introducing the knowledge graph: things, not strings', GoogleBlog, Google Inc., Mountain View, CA, USA, http://googleblog.blogspot.co.uk/2012/05/introducing-knowledge-graph-things-not.html, accessed 20 February 2016

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C., The OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenber, A., Sansone, S. A., Scheuermann, R. H., Shah, N., Whetzel, P. L. and Lewis, S. (2007) 'The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration', *Nature Biotechnology*, vol 25, pp1251-1255

Sporny, M., Longley, D., Kellogg, G., Lanthaler, M. and Lindström, N. (2014) *JSON-LD 1.0: A JSON-based Serialization for Linked Data*, World Wide Web Consortium (W3C), Cambridge, MA, USA, www.w3.org/TR/json-ld/, accessed 8 January 2016

Stein, B. R. and Wieczorek, J. (2004) 'Mammals of the world: MaNIS as an example of data integration in a distributed network environment', *Biodiversity Informatics*, vol 1, pp14-22

Tanksley, S. D. and McCouch, S. R. (1997) 'Seed banks and molecular maps: unlocking genetic potential from the wild', *Science*, vol 277, pp1063-1066

Tarasova, T., Mynarz, J. and Archer, P. (2015) *SmOD INSPIRE Vocabularies*, World Wide Web Consortium (W3C), Cambridge, MA, USA, www.w3.org/2015/03/inspire/, accessed 8 January 2016

TDWG (2006) *Technical Roadmap 2006*, Technical Architecture Group (TAG), Biodiversity Information Standards (TDWG), www.tdwg.org/activities/tag/documents/, accessed 8 January 2016

TDWG (2007) *Access to Biological Collection Data (ABCD)*, *Version 2.06*, Access to Biological Collection Data task group, Biodiversity Information Standards (TDWG), www.tdwg.org/standards/115, accessed 21 February 2016

TDWG (2009) *Darwin Core*, Darwin Core task group, Biodiversity Information Standards (TDWG), www.tdwg.org/standards/450, http://rs.tdwg.org/dwc/, accessed 21 February 2016

Telenius, A. (2011) 'Biodiversity information goes public: GBIF at your service', *Nordic Journal of Botany*, vol 29, pp378-381

Walls, R., Deck, J., Guralnick, R., Baskauf, S., Beaman, R., Blum, S., Bowers, S., Buttigieg, P. L., Davies, N., Endresen, D., Gandolfo, M. A., Hanner, R., Janning, A., Krishtalka, L., Matsunaga, A., Midford, P., Morrison, N., O Tuama, E., Schildhauer, M., Smith, B., Stucky, B. J., Thomer, A., Wieczorek, J., Whitacre, J. and Wooley, J. (2014) 'Semantics in support of biodiversity knowledge discovery: an introduction to the biological collections ontology and related ontologies', *PLoS ONE*, vol 9:e89606

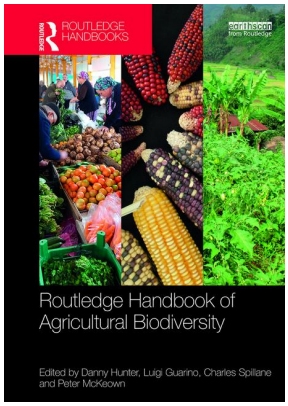
Wieczorek, J., Bloom, D., Guralnick, R., Blum, S., Döring, M., Giovanni, R., Robertson, T. and Vieglais, D. (2012) 'Darwin Core: an evolving community-developed biodiversity data standard', *PLoS ONE*, vol 7:e29715

Wood, D., Zaidman, M., Ruth, L. and Hausenblas, M. (2014) *Linked Data*, Manning Publications, New York, NY, USA

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J.R., Amaral-Zettler, L., Gilbert, J. A., Karsch-Mizrachi, I., Johnston, A., Cochrane, G., Vaughan, R., Hunter, C., Park, J., Morrison, N., Rocca-Serra, P., Sterk, P., Arumugam, M., Bailey, M., Baumgartner, L., Birren, B. W., Blaser, M. J., Bonazzi, V., Booth, T., Bork, P., Bushman, F. D., Buttigieg, P.nL., Chain, P. S. G., Charlson, E., Costello, E. K., Huot-Creasy, H., Dawyndt, P., DeSantis, T., Fierer, N., Fuhrman, J. A., Gallery, R. E., Gevers, D., Gibbs, R. A., Gil, I. S., Gonzalez, A., Gordon, J. I., Guralnick, R., Hankeln, W., Highlander, S., Hugenholtz, P., Jansson, J., Kau, A. L., Kelley, S. T., Kennedy, J., Knights, D., Koren, O., Kuczynski, J., Kyrpides, N., Larsen, R., Lauber, C. L., Legg, T., Ley, R. E., Lozupone, C.A., Ludwig, W., Lyons, D., Maguire, E., Methé, B. A., Meyer, F., Muegge, B., Nakielny, S., Nelson, K. E., Nemergut, D., Neufeld, J. D., Newbold, L. K., Oliver, A. E., Pace, N. R., Palanisamy, G., Peplies, J., Petrosino, J., Proctor, L., Pruesse, E., Quast, C., Raes, J., Ratnasingham, S., Ravel, J., Relman, D. A., Assunta-Sansone, S., Schloss, P. D., Schriml, L., Sinha, R., Smith, M. I., Sodergren, E., Spor, A., Stombaugh, J., Tiedje, J. M., Ward, D .V., Weinstock, G. M., Wendel, D., White, O., Whiteley, A., Wilke, A., Wortman, J. R., Yatsunenko, T. and Glöckner, F. O. (2011) 'Minimum information about a marker gene sequence (MIMARKS) and minimum

information about any (x) sequence (MIxS) specifications', *Nature Biotechnology*, vol 29, pp415-420

Zeven, A. C. (1998) 'Landraces: A review of definitions and classifications', *Euphytica*, vol 104, pp127-139



Routledge Handbook of Agricultural Biodiversity

Edited by: Danny Hunter, Luigi Guarino, Charles Spillane, Peter C. McKeown

Print publication date: 03 October 2017

Online publication date: 03 October 2017

Print ISBN: 9780415746922

eBook ISBN: 9781315797359

Adobe ISBN: 10.4324/9781317753285

URL (handbook): https://www.routledge.com/Routledge-Handbook-of-Agricultural-Biodiversity/Hunter-Guarino-Spillane-McKeown/p/book/9780415746922

Cite chapter 42 as: Endresen D (2017) Information, knowledge and agricultural biodiversity. Chapter 42, pp. 646-661 *in*: Hunter D, Guarino L, Spillane C, and McKeown (eds.) Routledge Handbook of Agricultural Biodiversity. Routledge Handbooks. Abingdon: Routledge. Published 09 Oct 2017. ISBN: 9780415746922, "doi:10.4324/9781317753285-43". Accessed from https://www.duo.uio.no/ (date and year).

URL (chapter 42): https://www.routledgehandbooks.com/doi/10.4324/9781317753285-43

https://doi.org/10.4324/9781317753285-43