

Draft Technical Manual on Passport Data Quality

Draft Technical Manual on Passport Data Quality Assessment and Improvement

Contents

1 Introduction

- 1.1 General Principles - T.van Hintum & K.U Sattler
- 1.2 International Obligations - S.Harrer
- 2 Generic Methodologies
 - 2.1 Referential Checks - E.Arnaud
 - 2.2 Spelling check- T.Metz
 - 2.3 Cross checks (tabulation) - T.Metz
 - 2.4 Outlier Detection - F.Atieno
- 3 Domain Specific Methodologies
 - 3.1 Taxonomic Data - H. Knüpffer
 - 3.2 Identifiers - S.Gaiji
 - 3.3 Institutional Identification -S.Dias
 - 3.4 Collecting Data - T.Hazekamp
 - 3.5 Geo-spatial Data - A.Jarvis
- 4 Global Resources and References - S.Gaiji
- 5 Documentation Quality Index - T.van Hintum
- 6 References

Introduction

General Principles - T.van Hintum & K.U Sattler

In research the quality of the data used in an analysis is one of the key factors for the success of the project. Garbage in –garbage out. This obviously also hold true for the Generation Challenge Programme (GCP) and all its projects.

The GCP aims at unlocking the genetic diversity in the worlds genebanks by using new technology, based on the increasing knowledge of the genome. Passport data play a crucial role in this attempt. Core collections were selected on the basis of passport data, and subsequently genotyped with molecular markers. Germplasm panels representing specific niches of diversity or cross sections of diversity are created to study physiological contrasts, performing association genetic studies, or creating genotype panels for DArT. Always the quality of the passport data is the major factor for success.

Compared to other types of data produced in the GCP, passport data are relatively simple. We have well established standards for the basic structure of these data: the Multi Crop Passport Descriptor (MCPD) list (FAO/IPGRI 2001). This simple list of passport descriptors proved to be an easy to use structure for data exchange, and is widely used for this purpose. It obviously has its limitations, and with the new information technologies becoming available extension of the list or the application of ontology driven approaches are being explored and applied. But still, the MCPD will act as an anchor, a common set of concepts to which all other approaches will be able to map.

When discussing data quality, we are not alone. Data quality is an extremely important factor in any process that uses data. Companies can lose much money if the data they use in their production, logistics or management is incomplete or unreliable. Data quality becomes an important issue in all data-intensive applications where data do not fulfill the requirements, e.g. because the data are incomplete, inexact or just erroneous. Data quality problems can have many causes. For example problems can occur during manual data entry, or be caused by systematic errors in terms of representation (improper data types or encodings) or inappropriate formatting.

But what exactly do we mean if we refer to ‘data quality’? An often used definition is „fitness for use“. However, this is very general – it does not describe which criteria are satisfied or not. In the information technology literature many different dimensions of data quality were introduced, and several classifications of dimensions have been proposed. However, from 179 dimensions presented by Wang Strong (1996) only a few are typical used. These are

- *completeness* referring to the portion of real-world objects represented in the data set or from a more technical point of view the fraction of non-null values,
- *correctness* describing to which extend data are accurate, where correctness means the nearness of a value to the correct real-world values, and

- *consistency* specifying the fraction of data (records, values) not violating given business rules (e.g. integrity constraints).

Basically, the common idea shared by all definitions of data quality is that data quality is a multidimensional and also a subjective term.

Beside the definition of the relevant dimensions to measure data quality, there are a number of other issues to consider. First, *data quality assessment* is important in order to estimate relevance, significance or generally the value of results of analyzing tasks based on the data. Due to the previously mentioned famous GIGO principle (garbage in – garbage out) results can be only as good as their input. Assessment also provides the indicator for necessary improvements of data quality and allows to evaluate the cost-benefit ratio after improvements. The result of assessment is a set of quality scores which are assigned to the individual dimensions. The second issue is to define an appropriate model for making these scores explicit, i.e. storing them in a database and associate the scores to the data. An explicit *data quality model* allows for further interpretation of quality information as well as ranking or source selection. Finally, if the quality of the data does not fulfill the requirements it has to be improved by applying *data cleaning techniques*. These include domain-independent techniques such as normalization and transformation, outlier detection, identifying and repairing missing values, duplicate detection but also domain-specific approaches. Data cleaning runs often in a feedback loop starting with assessment in order to identify and quantify quality problems followed by the required cleaning steps and a final assessment.

Looking at passport domain, this can be translated in to the following issues:

How complete are the data? Are there ways to increase the number of values – by using other available fields or additional data sources? Going back to information recorded during the collection of the material, or using available collection site information to add longitudes and latitudes could be options.

Are the data correct? Are there ways to check if the accession is actually the species it is recorded to be? Spellcheckers, checklists, but also the actual growing and checking the material are approaches that can be considered here.

Are the data consistent? Do the codes exist, do the species grow where they were recoded as being collected?

These three dimensions of data quality seem to be the most relevant and important ones, and will be explored in some depth in the text you are reading.

A workshop, organized by Bioversity, and funded by the GCP, was held in Rome 3 to 5 July 2007, to discuss passport data quality, and to make an inventory of opportunities to measure and increase the quality of these data. This manual is the outcome of this workshop. In it, we do not want to give an in depth academic analysis of all dimensions of data quality – if you are interested there is plenty of material to read from highly theoretical (such as Wang and Strong 1996) to highly applied (Chapman 2005, check out the GBIF website for other manuals !) In stead, in this manual we will try to provide the gene bank curator and the gene bank database manager with easy to use ideas,

suggestions and tools to check and improve the quality of their passport data. To some readers some of the items will appear to be open doors, and can be ignored, others might be eye openers, or just things that are available but simply not known to some. We trust that it will provide an accessible help in improving the quality of the passport data.

International Obligations - S.Harrer

In 1992 the Convention on Biological Diversity (CBD) set up a first comprehensive international framework for the conservation and sustainable use of all biological diversity. To address more the specific problems of agricultural biodiversity, a second international binding agreement was negotiated under the auspices of the Food and Agricultural Organisation (FAO) of the United Nations. The International Treaty on Plant Genetic Resources for Food and Agriculture (ITPGRFA) came into force in June 2004 and was ratified up to now by 112 countries and the European Union (by July 07). Its objectives are the conservation and sustainable use of plant genetic resources for food and agriculture (PGRFA) and the fair and equitable sharing of benefits derived from their use, in harmony with the CBD. Through the ITPGRFA, countries agree inter alia to establish an efficient, effective and transparent Multilateral System to facilitate access to their PGRFA, and to share the benefits arising from the use of plant genetic resources in a fair and equitable way. The Multilateral System applies to over 60 major crops and forages.

One central element to facilitate the access to the PGRFA in the Multilateral System is the obligation of the Contracting Parties of the ITPGRFA to make “all available passport data” (article 12.3(c)) accessible by means of “catalogues and inventories” (article 13.2(a)) amongst others through the Global Information System (article 17) of the ITPGRFA. The terms of the material transfer are regulated by a Standard Material Transfer Agreement (SMTA) in which the provider of the material commits himself also to make “all available passport data and, subject to applicable law, any other associated available non-confidential descriptive information” (article 5b) of the SMTA) available to the recipient together with the PGRFA provided. In this respect the quality of passport data becomes crucial.

The Global Information System of the ITPGRFA will be based on existing national (e.g. institutional information systems, National Inventories), regional (e.g. EURISCO, GRIN) and international (e.g. SINGER) information systems. Besides international standards for data exchange like the Multicrop Passport Descriptors, data quality, i.e. *completeness*, *correctness* and *consistency*, will be one of the most important prerequisite for its satisfactory implementation and functioning.

Generic Methodologies

Referential Checks - E.Arnaud

A data set checking against referential data sets is needed for most of the Passport data fields, mainly those related to: taxonomy, botanical name, author names, institution codes and name, country names, coordinates, location/place names, collecting environment. Some of these fields in the Passport data are mandatory because they ensure that the recorded Passport data is meaningful. Consequently, as those fields cannot remain empty and has to be properly filled, a referential checking is necessary. The mandatory field are often identified with SQL rules imposing to choose a value in a reference picklist and preventing the record validation if the field is empty or mistyped. Referentials to which this feature applies at best are the Taxonomy lists with Genus, species, accession names because it reduces a lot the misspelling errors.

Referential checks of data sets will use the authority lists or authority data sets that can be found on thematic web sites, in guidelines for data quality, in reference books, or embedded as picklists in the database for data entry assistance. To follow the principle of currency and timeliness, the authority source used and its version should be cited with the checked data set or with the picklist built from the authority list.

The proper authority list has to be chosen following recommendations, guidelines. Reference web sites recommend the use of referential for the Passport data _ multi crop sites e. g. GBIF, SINGER, EURISCO, GRIN and also Crop specific sites. The referential data sets can also be quality data from other databases compiling the same types of data and identified as being an authority by domain experts. Several referential adapt sets are cited in the following chapters according to the type of data set to be checked for the MCPD along with the methodology on when and how using those referential data sets.

Checking data sets against referentials allows verifying that the recommended standards is properly applied but can also highlight various problems: • A wrong codification, wrong IDs • misspelling • gaps • Erroneous or inaccurate data.

It also allows identifying new data that is not yet included in the authority list. In this case, the a specific temporary code or name may need to be found and explained to be easily retrieve and be substituted by the official one once it is produced.

Additional information on the level of quality control or certainty of identification can inform users of the data status and their fitness for use (Chapman, 2005): • data has been fully checked against referentials/authority lists • data is temporarily attributed waiting for validation • data has not been checked.

An authority list or referential data sets allows checking and integrating various standards. For the MCPD: • An official codification: e.g. ISO country codes, ISO languages codes, • An official ID: e.g. FAO institution codes, Unique accession identifier, alternate identifier • An official botanical name: Reference taxonomical checklists (Taxonomy checker) • Appropriate coordinates or location name: GIS, Geomancer, Atlases, gazetteer. • Environments: GIS, maps.

Spelling check- T.Metz

Values in free-text fields may be duplicates that differ only by minor variations in spelling, e.g. capitalization, punctuation, typos, white space. Examples of such free-text fields are bibliographic references, addresses, names of persons, places, institutions, etc.

The commonly available string comparison functions[1] can only make a binary distinction, whether text strings are equal or not. They do not allow fuzzy comparison[2] of text strings. An easily available and generally applicable tool[3] that uses the similarity of text strings to quickly find possible fuzzy duplicates in a list of strings is described here. A more complete description of different string similarity metrics can be found here[4].

Fuzzy duplicates may lead to various data quality problems:

- o The number of distinct entities is inflated, e.g. the number of distinct recipients of germplasm as counted from a contacts database is too large if there are fuzzy duplicates of recipient names/addresses.
- o Multiple records for the same entity lead to data redundancy and data inconsistency and to query results that vary according to the spelling of the input.
- o The number of distinct states is inflated, e.g. the description of the habitat at a collection site contains fuzzy duplicates that only orthographic variants rather than descriptions of different habitats.
- o Linkages within the database or to external databases may be missing or inconsistent as there are usually exact matches required. This may affect linkages by taxonomy, bibliography, place name, etc.

The String Similarity Checker (<http://cropwiki.irri.org/spellcheck/>) is a web-based tool that allows users to check long lists of strings (names, addresses, bibliographic references, etc.) for possible fuzzy duplicates. Comparisons of all strings in the list will yield groups of similar strings that are above a user-defined similarity threshold.

An additional functionality is available to check a list of strings against a user-supplied dictionary of strings in order to find possible fuzzy duplicates. This is useful if users have a clean dictionary of distinct strings (e.g. names, addresses, etc.) and want to check whether the addition of new strings would create fuzzy duplicates.

A typical output from a similarity check of a list of names is shown below.

ALAN CARPENTER

ALAN J. CARPENTER [0.90]

ALAN L. CARPENTER [0.90]

B. H. SIWI

B.H. SIWI [0.95]

CHRISTOPHER TSEU

CHRISTOPHER TSUE [0.94]

The similarity measure is a value between 0 (completely different – not a single character in common) and 1 (exactly identical – character by character). The measure is based on the edit distance as well as the length of the strings. A more detailed explanation of the functionalities and the definition of the similarity measure are available in the user guide (<http://cropwiki.irri.org/spellcheck/userguide.html>).

Limitations: Spelling check based on fuzzy string comparison only addresses syntactic similarity (e.g. Philippines vs. Phillippines), but not semantic similarity (e.g. UAE vs. United Arab Emirates). If naming systems are based on small syntactic differences, e.g. sequential numbering of accessions where the smallest possible difference is also the true difference between genuinely different names, a fuzzy string comparison approach may yield too many false positives to be useful. The tool quickly highlights possible fuzzy duplicates but users need to make an independent decision whether an observed similarity between two strings is a fuzzy duplicate or a genuine difference.

[1] [http://en.wikipedia.org/wiki/String_functions_\(programming\)](http://en.wikipedia.org/wiki/String_functions_(programming))

[2] http://en.wikipedia.org/wiki/Fuzzy_string_searching

[3] <http://cropwiki.irri.org/spellcheck/>

[4] <http://www.dcs.shef.ac.uk/~sam/stringmetrics.html>

Cross checks (tabulation) - T.Metz

Databases provide easy-to-use query functionality (SQL) for tabulating or counting the unique values of categorical fields. Text books on SQL usually provide only basic examples where this functionality is used for reporting, while implicitly assuming that the underlying data are complete and correct. The same syntax that is used for reporting can also be used for checking on data quality. The following MS Access SQL examples show different applications of the basic SQL tabulation functionality for data quality checking. The techniques used in different examples may be combined for more complex problems.

```
SELECT soil_texture, count(*) AS cnt
  FROM collection
  GROUP BY soil_texture;
```

This is the basic query that generates a frequency tabulation of the different values that a categorical variable (soil_texture) can have. Undefined values, missing values, spelling mistakes, data entry errors, or other anomalies and/or inconsistencies in the domain of a categorical variable can easily be spotted from such a frequency tabulation.

```
SELECT medium, type_test, count(*) AS cnt
  FROM germination
  GROUP BY medium, type_test;
```

This query generates a frequency tabulation from a combination of two categorical

variables. If particular combinations of values from the two categorical variables are not possible, they can be spotted from such a frequency tabulation.

```
SELECT iif(lat = "", "N","Y") AS lat_pres,  
       iif(lon = "", "N","Y") AS lon_pres,  
       count(*) AS cnt  
FROM passport  
GROUP BY iif(lat = "", "N","Y"),  
         iif(lon = "", "N","Y");
```

In this query, continuous variables (latitude, longitude) are recoded into absence or presence states and then a frequency tabulation of the combination of latitude presence and longitude presence is produced. This is a general approach to check multiple variables on their presence/absence patterns and find those patterns that are considered a data completeness or data consistency problem.

```
SELECT year(collect_date) AS year, count(*) AS cnt  
FROM herbarium  
GROUP BY year(collect_date);
```

In this query, a date function is used to convert a continuous variable (collect_date) into a categorical variable (year) and then the new categorical variable is tabulated. This may reveal unexpected pattern, outliers, data entry errors, etc. There are several other date functions (e.g. month, week, dayofweek, dayofmonth) that can be used in a similar way.

```
SELECT weekday(test_date) AS weekday, count(*) AS cnt  
FROM germination  
GROUP BY weekday(test_date);
```

In this query, a date function is used to convert a continuous variable (test_date) into a categorical variable (weekday) and then the new categorical variable is tabulated. If germination tests are not performed during weekends (1=Sunday, 7=Saturday), this tabulation will quickly reveal such records.

```
SELECT date_test, count(*) AS cnt  
FROM germination  
GROUP BY date_test  
ORDER BY count(*);
```

This query simply counts the number of tests performed on each distinct date and then sorts by this count. Looking at the top of the list and the bottom of the list may reveal test dates with unexpected low numbers or high numbers of tests, which may indicate data quality problems, as tests are usually conducted in similar-size batches that are manageable.

```
SELECT accession_no, count(*) AS cnt  
FROM passport  
GROUP BY accession_no  
HAVING count(*) > 1;
```

This query assumes that accession_no is not defined as a unique field in table passport and through that prevented from containing duplicate values. If all accession numbers appear only once, the query will not return any results. This type of query is generally useful to tabulate categorical values that are out of an expected frequency range.

```
SELECT (weight*1000) MOD 10 AS val_lastdigit, count(*) AS cnt  
FROM inventory
```

```
GROUP BY (weight*1000) MOD 10
ORDER BY (weight*1000) MOD 10;
```

In this query a numeric function (modulus) is used to extract the rightmost decimal digit, representing grams in this example, and then the frequency of this new categorical variable is tabulated. Such a tabulation reveals information about the actual precision of measurements. E.g. if the large majority of cases are the digits 0 and 5, then the actual precision is probably 5 grams and not 1 gram, as the numerical precision of 3 decimal places may suggest.

```
SELECT round((avpcgerm)/10)*10 AS pcgermclass, count(*) as cnt
FROM germination
GROUP BY round((avpcgerm)/10)*10;
```

In this query a numeric function (round) is used to categorize a continuous variable, and then the frequency of this new categorical variable is tabulated. Such a tabulation may reveal deviations from an expected distribution.

```
SELECT len(trim(pretreat)) AS numtreat, count(*) as cnt
FROM germination
GROUP BY len(trim(pretreat));
```

In this query character functions are used to extract the length of a string variable and then tabulate these length values. Depending on the length pattern of the string values involved, certain anomalies may be detected, e.g. unexpectedly short or long strings, unexpected lengths.

Outlier Detection - F.Atieno

Outlier Detection

Most real world datasets include a certain amount of exceptional values generally termed as “outliers”. Outliers are defined as “an observation (or subset of observations) which appears to be inconsistent with the remainder of the dataset”. A more relevant definition is “values that lie very far from the middle of the distribution in either direction”, referring to the numeric distance.. Causes of outliers could include flawed values resulting from poor data quality i.e a data entry or data conversion error. Likewise physical measurements when performed with malfunctioning equipment/tools may produce certain amounts of distorted values. However, occasionally it is also possible that an outlier may represent a correct, though exceptional information i.e true but extreme, data values. As such due diligence should be exercised when dealing with data values until definite errors are proven.

Why isolate outliers?

The isolation of outliers is important both for improving the quality of original data and for reducing the impacts of outlying values in the process of data analysis. As such the main reason of isolating outliers is associated with data quality assurance. [] asserts that unreliable data represents unconformity between the state of the dataset and the state of the real world. Since the exceptional values have higher probability of being incorrect thus removing or correcting the outliers will improve the quality of stored data and

subsequently positive impact on the results of data analysis and data mining. Therefore, outlier detection is a critical part of data analysis. While often outliers are removed to improve accuracy of the estimators, this practice is not recommendable because sometimes outliers can have very useful information.

Outlier detection and treatment

Most outlier detection methods use some measure of distance to evaluate how far away an observation is from the centre of the data. After identification of outliers and true (extreme or normal) values, the researcher must decide what to do with problematic observations. The options are limited to correcting, deleting, or leaving unchanged. However, there are some general rules for which option to choose. For example, impossible values are never left unchanged, but should be corrected if a correct value/match can be found, otherwise they should be deleted. What should be done with true extreme values and with values that are still suspect after the diagnostic phase? One may wish to further examine the influence of such data points, individually and as a group, on analysis results before deciding whether or not to leave the data unchanged.

Most existing methods of outlier detection are based on manual inspection of datasets. This entails the use of supervised and unsupervised methods categorized into univariate methods, which examine each variable individually, and multivariate methods which take into account associations between variables in the same dataset.

Univariate approach*Italic text*

According to this method, a value is considered an outlier, if it is far away from other values of the same attribute. An example is given where one or few data points stand out clearly apart from the rest. Perhaps the most popular univariate outlier detection technique for survey data is the quartile method. This method creates an allowable range for the data using lower and upper quartiles: data falling outside of the range are outliers. The method is not only robust, but simple and non-parametric.

Example 1:

Example 2:

Multivariate approach*Italic text*

Some definite outliers can be detected only by examining the values of other attributes. For example, where one or few data points stand clearly apart from a bivariate relationship formed by other points. Such an outlier can only be detected by a multivariate method since it is based on a dependency between two variables. This also addresses another special problem with datasets, that of erroneous inliers, i.e., data points generated by error but falling within the expected range. Erroneous inliers will often escape detection. Inliers are mainly discovered to be suspect if viewed in relation to other variables.

Multivariate methodologies available include: using scatter plots or consistency checks.

Example 1:

Example 2:

Conclusion

Data cleaning often leads to insight into the nature and severity of error-generating processes. The researcher can then give methodological feedback to improve data gathering and precision of outcomes. It may be necessary to amend the study protocol, regarding design, observer training, data collection, and quality control procedures. In extreme cases, it may be necessary to restart the study or do the whole exercise all over again.

Domain Specific Methodologies

Taxonomic Data - H. Knüpffer

Identifiers - S.Gaiji

Institutional Identification -S.Dias

Collecting Data - T.Hazekamp

Collecting_data_data_quality_first_draft-Hazekamp.doc

Geo-spatial Data - A.Jarvis

Global Resources and References - S.Gaiji

SINGER

Introduction

The System-wide Information Network for Genetic Resources (SINGER) is the germplasm information exchange network of the Consultative Group on International Agricultural Research (CGIAR) and its partners.

Together, the members of SINGER hold more than half a million samples of crop, forage and tree diversity in their germplasm collections. This diversity is vital for food security and agricultural development; SINGER provides easy access to information about this diversity SINGER is an initiative of the CGIAR System-wide Genetic Resources Programme (SGRP). Basic Search Functionality

SINGER web interface provides search functionality based on all fields available in the passport descriptors. In addition, users can also search seed distribution and cooperators database.

Web link: <http://singer.cgiar.org/>

EURISCO

Introduction

The EURISCO web catalogue automatically receives data from the National Inventories (NI). It effectively provides access to all ex situ PGR information in Europe and thus facilitates locating and accessing PGR. EURISCO is hosted at and maintained by the International Plant Genetic Resources Institute (IPGRI) on behalf of the Secretariat of the European Cooperative Programme for Crop Genetic Resources Networks (ECP/GR).

The central infrastructure of EURISCO has been developed with open source softwares. This strategic choice is intended to allow EURISCO National Focal Points to benefit from the development of EURISCO for their national implementation. The uploading mechanism concept is designed to allow an easy data checking of the information provided in national inventories both on essential descriptors and on a line-per-line checking. The checking and validation procedures assist the national partners in their efforts to improve the accuracy of their information with their data providers at national levels. Basic Search Functionality

Web link: <http://eurisco.ecpgr.org/>

GBIF

Introduction

The GBIF data portal is a service that provides access to millions of scientific data records that are being shared via the GBIF network. These data are generously made available through the GBIF network by a wide range of institutions and organisations from around the world. To see the range of data providers involved, please see the list of data providers and datasets. The two types of data currently being shared through the GBIF Network are: ■ Species occurrence records (based on specimens and observations) - information about the occurrence of species at particular times and places. ■ Names and classifications of organisms - information on the names (both scientific and common) used for species and on the classification of those organisms into taxonomic hierarchies.

Basic Search Functionality Searching for occurrences Access to species occurrence records is at the core of the GBIF Data Portal. The functions provided by the Occurrence Search page provide the ability to perform complex searches in order to explore and locate records of interest.

Many other pages in the portal provide links to this page and set some initial search filter. For example, the "Explore occurrences" link from a species' Overview Page opens the Occurrence Search to find records for that species.

Viewing details for an occurrence record Occurrence records can be found using the Occurrence search (see Searching for occurrences). In the tabular view offered by the Occurrence search, there is a View link shown to the right of each record. This link opens the Occurrence detail view, which in turn offers a link to retrieve the original record directly from the provider's web site. For more information on searching GBIF: <http://data.gbif.org/tutorial/tutorial> Web link: <http://www.gbif.org/>

USDA (United States Department of Agriculture)

Introduction

The Agricultural Research Service (ARS) is the U.S. Department of Agriculture's chief scientific research agency. ARS conducts research to develop and transfer solutions to agricultural problems of high national priority and provide information access and dissemination to:

- ensure high-quality, safe food, and other agricultural products
- assess the nutritional needs of Americans
- sustain a competitive agricultural economy
- enhance the natural resource base and the environment, and
- provide economic opportunities for rural citizens, communities, and society as a whole.

Basic Search Functionality USDA portal supports free text search of different data elements and in particular users can search the following through their data portal:

1. Accession Area Queries
2. Taxonomic Queries
3. Research Crops and Descriptor/Evaluation
4. Data Queries
5. Search Multiple Databases

Web link: <http://www.ars-grin.gov/npgs/searchgrin.html>

Reference:

1. <http://singer.cgiar.org/>
2. <http://eurisco.ecpgr.org/>

3. <http://www.gbif.org/>

4. <http://www.ars.usda.gov/main/main.htm>

Documentation Quality Index - T.van Hintum

Passport_data_quality_-_Measuring_TvHintum.doc

References

Chapman, A.D., 2005. Principles and Methods of Data Cleaning. Report for the Global Biodiversity Information Facility 2005. 75pp. Copenhagen: GBIF. pdf.

FAO/IPGRI, 2001. Multicrop Passport Descriptors. Available from the Bioversity website: pdf.

Wang, R and D. Strong, 1996. Beyond Accuracy: What data quality means to data consumers. Journal on Management of Information Systems 12(4), 1996.