Biodiversity Specimen Data

Before considering in detail how specimen data are recorded in a Symbiota network such as OpenHerbarium, it is useful to consider three general concepts: data, metadata, and standards.

Data

Data are isolated bits of information. The height of a plant; the weight of an egg; the number of animals seen. The problem is that different people may report the same data in different ways. For example, they might use different units or count different objects. All these affect how data can be used.

Metadata

Metadata describes data. It may include, for example, information on the units of measurement to be used, how a height is to be measured, which animals are to be counted, and how the data are to be reported. Providing the metadata about for data field is essential when asking people to provide information; conforming to the metadata standards set for an existing data system is also essential.

Standards

Standards, in the context of biodiversity data sharing, are agreements on the metadata that need to be met for effective data sharing. Reaching agreement on the existing standards has not been easy, nor is the task complete. One problem is that different disciplines, as well as different individuals within a discipline, differ in what data to record and how it should be recorded, The reasons for disagreeing are not trivial. They arise from the different habits, morphologies, and life styles of different organisms as well as the interests of the humans studying them and the technologies available. Even 40 years ago, asking collectors to record or determine the latitude and longitude of their collection site was excessive; today it is not unreasonable.

The Global Biodiversity Information Facility

Establishment of the Global Biodiversity Information Facility, a facility that makes biodiversity data freely available, was preceded by about 40 years of discussion of the kinds of data associated with natural history collections, identifying core aspects, items recorded for all kinds of natural history collections, and then how to record each kind of data. Once the core fields had been established, attention was turned to other kinds of data that. Today, the number of fields for which data can be stored is substantial. Which are used varies with field and the purpose. Zoologists often rely on images, tracks, fecal matter, tissue samples, and blood samples for information rather than specimens. There are data fields and metadata associated with each of these kinds of information because, for effective sharing, different data providers must provide the same kind of information in the same way. In other words, for effective biodiversity data sharing, biodiversity scientists must adhere to the standards set through international discussion among people studying different kinds of organismal biodiversity and related disciplines such as ecology, geography, photograph, and genetics.

Despite everyone's best efforts, there are several standards for sharing biodiversity data but the two most widely used are the Darwin Core (DwC) and ABCD (Access to Biological Collections Data) standards. The two standards have many fields in common but, at least initially, ABCD enabled more detailed recording of the location and use of individual specimens.

Since initial adoption of the DwC standards, DarwinCore "Extensions" have been added to the DwC standard. These are additional fields that allow sharing additional information such as fields to accommodate audio and video records. Symbiota, and hence OpenHerbarium, adheres to the DwC standard and its extensions but does not use all the fields recognized. For example, it does not (as yet) have fields for audio and video files, nor for all the EXIF data associated with images. It will shortly be adding fields for recording GENBANK records because it now includes tools for linking to such records.

There are two sets of six documents on how to enter data into OpenHerbarium. Both sets address the basic data entry form (Fig. 1) and contain a separate document for each of its sections. There are two sets because one concerns direct data entry, the other use of the Symbiota Workbook (see the next section). The basic data entry form, or subsets of it, are used to contribute almost all the data used by Symbiota so it important to be familiar with its fields and the way data should be entered into them. The figures are embedded in the relevant documents but are also available as a separate "Figures" document.

Alternative Data Entry Approaches

Symbiota is designed for "Direct Data Entry". This means entering data directly into the networks database via a browser. There are several advantages to this approach: One is that it enables used of Symbiota's tools checking the correctness of certain fields (e.g., scientific name, province) and accelerating data entry (e.g. via duplicate discovery). Another advantage is that it means the database is managed centrally; there is no need for an individual herbarium to have an IT person on staff. A third is the records beccome available immediatly. A major disadvantage of this approach is that it is enormously frustrating if internet access is intermittent, slow, or costly, a situation that faces many herbaria.

The alternative to direct data entry is an indirect approach in which data are entered offline into a local database or spreadsheet and then uploaded to the central database in batches. The data are not available to others until they are uploaded. The advantage of this approach is that it is does not require good internet access. The disadvantage is that the tools for checking data quality and accelerating data capture are not available offline. There are some for tools for checking data quality (data cleaning) on submitted records but it is always better to submit high quality data in the first place.

There are basically two approaches for offline data capture: using a data base system or a spreadsheet. Database systems can be very elaborate with tools for preventing errors in data entry, ensuring consistency, and label generation but the more elaborate they are, the more difficult they are to maintain. The advantage of database systems over spreadsheets comes from their greater ability to prevent errors in data entry. For example, they can be designed to send an error message if a plant is said to have been collected at 22,000 m elevation or if a scientific name has been spelled incorrectly. Unfortunately, there is not yet an appropriate database available.

Until an appropriate offline database has been developed for Symbiota, the best alternative is a workbook. A workbook consists of a linked series of spreadsheets. In the Symbiota Workbook, there are seven spreadsheets. The first six are for data recording; one for each section of the basic data entry form (Fig. 1). The seventh is completed automatically and is used to generate the spreadsheet used to upload the data. The workbook documents in this handout are based on use of the Symbiota Workbook modified for use in Pakistan. It us saved as PKWorkbook.

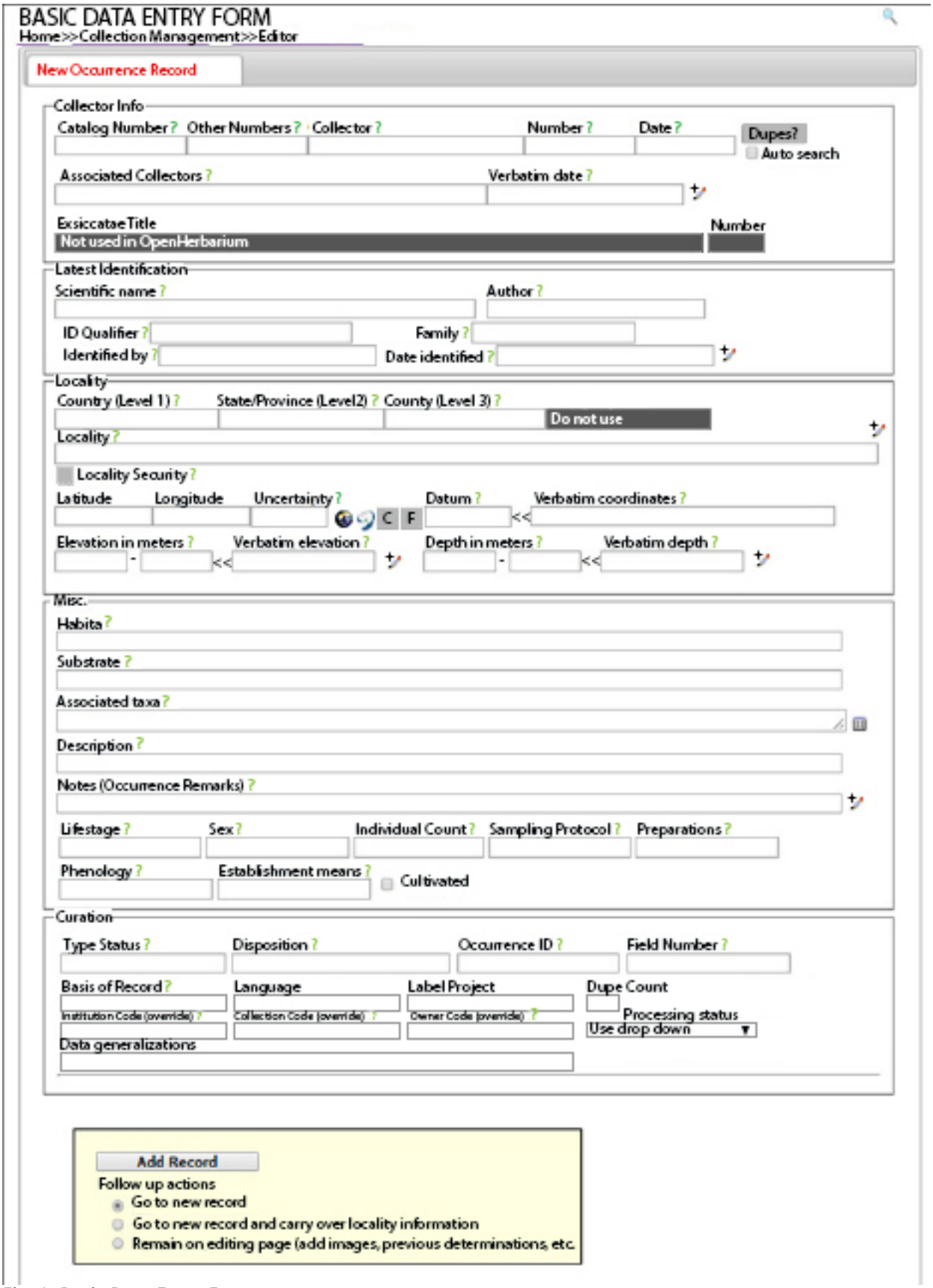

Fig. 1. Basic Data Entry Form