This workflow was developed at an iDigBio workshop in January 2015. The most recent version is available at https://github.com/iDigBioWorkflows/FlatSheetsDigitizationWorkflows and https://www.idigbio.org/content/workflow-modules-and-task-lists.

Appendix S10. Module 10: Selecting a Database

Below we give a brief overview of some of the things to be considered when selecting a database to store your specimen occurrence records. For another excellent treatment of this topic, see Chapter 6, “Deciding on a particular database solution” in Frazier et al. (2008).

Task ID	Task Description	Explanations and Comments	Resources
T1	Assess institutional need, goals, and policies.	This task is paramount. The subsequent tasks can only be properly undertaken after a thorough understanding of your institutional needs, goals, and policies is achieved. This assessment should take into account future needs, in addition to present/short-term needs. Subsequent tasks provide guidance for acquiring the kinds of knowledge needed to make further assessments.	See Frazier et al. (2008). http://www.gbif.org/resource/80574
T2	Assess institutional IT support, including local computer resources.	If IT support is limited or lacking, options may be limited, to include inexpensive commercial packages or open-source solutions that can be implemented with little need for customization and configuration. Web-based solutions that are maintained by the community (e.g., a Symbiota node within a regional portal) are another increasingly popular possibility. Excel spreadsheets, Access, and Filemaker Pro databases, etc. are relatively inexpensive and easy to use, but still require design and configuration and may be prone to corruption of data. With ample IT resources and time, customization of existing open-source solutions or creation of custom solutions is possible. See Morris (2005) for a discussion of some of the issues and considerations.	Symbiota: http://symbiota.org/. Specify: http://specifyx.specifysoftware.org/. See: Morris (2005). http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.113.2339&rep=rep1&type=pdf . or: http://systbio.org/files/phyloinformatics/7.pdf Commercial or other biodiversity database systems for cost include: Arctos EMu PastPerfect

			Reviews for these and others are at https://www.idigbio.org/content/biological-collections-databases.
T3	Assess collection size.	If digitizing a small collection (100s or a few 1000s of records), it may not be worth the time and money to implement a complex and highly customized database. For a very large collection, some solutions may not be able to hold all records or provide an effective means of querying data.	
T4	Assess the existence and format of legacy data.	Legacy data may need to be imported into a new database solution. The ease of mapping data between the old and new system will vary depending on the structure of the old data and the new database model. An understanding of the Darwin Core terms can facilitate this process. Also consider the formats in which data can be exported from the old database and the import formats recognized by the new database. If using a commercial solution, then the vendor may be able to assist you in the process of database migration.	See: Darwin Core standard: http://rs.tdwg.org/dwc/terms/index.htm#.
T5	Assess the kinds of data that will be managed in the database.	Determine if the institution simply needs a place to store specimen occurrence data, or something that can handle accessioning, loans, multimedia, etc. If digitization includes imaging, ensure the selected database supports linkages to images and other multimedia objects. Simple, flat solutions (e.g., spreadsheets) can be easy to implement and use; however, there are great risks associated with such solutions (e.g., accidental deletion of records, sorting mistakes, etc.).	

		Whether considering an open-source or commercial solution, it is highly recommended to evaluate test versions of the applications to ensure they meet institutional needs.	
T6	Assess the kind and number of staff that will interact with the database.	A wide variety of individuals may interact with an institutional database, including students, curator(s), collection manager(s), volunteers, etc. Some platforms allow for local or remote data entry by multiple simultaneous users, whereas others (e.g., Access, Excel, FileMaker Pro) may offer single-user licenses only. Another consideration is the comfort level of users with computers and the extent of data editing an institution might allow them to perform. For example, some databases have rich functionality for managing the level of access and editing capability individual users are assigned, providing more control for the institution over what types of data might be edited and to what extent. This can be an important consideration for institutions that expect a wide variety of users to be accessing the database. Regardless of the user base, quality control processes are an important consideration. These can range from form fields with controlled input based on underlying authority tables to automated back-end processes that monitor for possible errors.	
T8	Assess whether data from the collection will coexist with data from other collections at the institution.	At some institutions, data from different disciplines will coexist in the same database. Some databases are better at handling such scenarios than others. Also consider future changes that may affect database structure.	
T9	Assess database access needs.	Consider the kinds of users that will access the database and their location. Are internal users centrally located or	

		dispersed across the institution? Do volunteers play a big role in transcribing and managing? Some platforms have web-based interfaces and rich user administration functionality.	
T10	Assess ease of data export.	Determine how easy it is to export data from the database and potentially share it outside of the institution, i.e., with data aggregators. Some platforms facilitate the publishing of data to the internet. For example, some web-based platforms provide instant public access to data. Keep the institution's data sharing policy in mind when considering options that provide instant, public access. The ease with which data for sensitive species can be withheld from public view is a consideration, as well. Local databases can provide greater control over data and their distribution but may require greater IT support (see T3) when data are shared with external users, aggregators, and web-based presentations. Regardless of the database solution, it is important to be able to map specimen data to the terms provided in the Darwin Core standard. A common and increasingly popular means of sharing data over the web is via a Darwin Core Archive. The GBIF Integrated Publishing Toolkit (IPT) is a tool that enables the publishing of such archives.	See: Darwin Core standard terms: http://rs.tdwg.org/dwc/terms/index.htm#. Robertson et al. (2014). doi:10.1371/journal.pone.0102623.
T11	Assess back-up functionality.	Over the short- to medium-term, you will want to back up the data in your database. See the DATAOne website for a more in-depth consideration of this topic. For long-term preservation of digital	See: DATAOne Best Practices for backing-up data: https://www.dataone.org/best-practices/backup-your-data.

		data, it is important to adopt a true digital preservation environment. See Module 9: Image Archiving and Corrado and Moulaison (2014) for more details about this very big topic.	Corrado and Moulaison (2014).

Literature Cited

Corrado, E. M., and H. L. Moulaison. 2014. Digital preservation for libraries, archives, and museums. Rowman & Littlefield, Lanham, Maryland, USA.

DataONE. Backup your data. https://www.dataone.org/best-practices/backup-your-data Accessed 1 May 2015.

Frazier, C. K., J. Wall, and S. Grant. 2008. Initiating a Natural History Collection Digitisation Project, version 1.0. Global Biodiversity Information Facility, Copenhagen, Denmark. http://www.gbif.org/resource/80574

Morris, P. 2005. Relational database design and implementation for biodiversity informatics. *PhyloInformatics* 7: 1–66. http://systbio.org/files/phyloinformatics/7.pdf or http://systbio.org/files/phyloinformatics/7.pdf

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