Technical\_Considerations

The technical challenges of using electronic databases to query sample information may be an additional hurdle in the process of requesting samples from museums. [Wieczorek et al.](#ref-wieczorek2012) ([2012](#ref-wieczorek2012)) state that for biodiversity information to be used in broad-scale analyses it must be in digital form, accessible, discoverable, and integrated. The last two of these terms are often given insufficient attention. Even if museums have digitized their collections and made them publicly available, other challenges including an overabundance of search platforms, inconsistent use of taxonomy, and confusing database structure increases the difficulty of use. However, significant gains are being made in the field of biodiversity [informatics@krishtalka2000](mailto:informatics@krishtalka2000). For example, the increasing occurrence of converting information to a machine-readable format is facilitating the identification and transfer of data ([Wilkinson et al. 2016](#ref-wilkinson2016)).

There can be a tension between easy to use, idiosyncratic, local efforts versus rigorous, standardized, reproducible alternatives. Necessary steps to bridge this divide include defining clear and consistent data standards, developing widely accessible tools to reproducibly document and query collections records, and use of standardized file formats and open-access software. Additionally, familiarity with some core data management and transfer concepts provides users with valuable skills for working with collections databases.

### How easily can users find sample information? (Findability)

#### API usage

Application Programming Interfaces, known as APIs, serve as an intermediary for data transfers between clients and databases. Web access to records and associated metadata from digital collections is often provided via remote requests to web APIs, which provide the ability to dynamically query databases while providing script-based automation and reproducibility ([Reichman et al. 2011](#ref-reichman2011); [Varela et al. 2014](#ref-varela2014)). APIs provide a powerful alternative to graphical user interfaces (GUIs), which let users make download requests based on many ‘point and click’ fields on a web forms but don’t provide viable options to document search terms and parameters used. APIs use a standardized set of methods that allow great flexibility in programming language used for requests, and also efficiently use bandwidth, making them a popular choice for data transfer over the internet ([Gadelha Jr et al. 2021](#ref-gadelhajr2021)). Most APIs for museum databases use HTTP (Hypertext Transfer Protocol) requests that can output data records in multiple file formats (e.g. JSON, comma-separated text) that can be incorporated into further text filtering workflows or data visualizations ([Saran et al. 2022](#ref-saran2022)).

Many online NHC and aggregation databases offer APIs for the querying and downloading records. Two databases with extensive documentation are iDigBio and GBIF, both of which offer APIs specific to the R, python and Ruby programming languages ([Chamberlain and Boettiger 2017](#ref-chamberlain2017)). Other API tools focus on more domain-specific querying tasks such as working with RDF-compatible data products ([Boettiger 2018](#ref-boettiger2018)) and using the SPARQL query language ([prud’hommeaux and Seaborne 2013](#ref-prudhommeaux2013)).

### How compatible are different sources? (Interoperability and Reusability)

#### Data standards

Data about specimens in natural history collections (NHCs) are increasingly available online, though their compilation and compatibility is only possible via widely accepted metadata standards. Fortunately, taxonomic and organizational data standards and associated client-server exchange protocols are becoming widely adopted by the natural history collections community. The two most common examples include 1) Darwin Core (primarily North American) and the Distributed Generic Information Retrieval (DiGIR) protocol, a relatively newer set of standards that use the RDF format to consolidate terms collectively shared throughout NHCs without relational structure (flat files), and 2) Access to Biological Collection Data (ABCD; primarily European) and the BioCASE protocol, an older set of standards that are highly structured and seek to capture the wide variety of biodiversity data and associated schemas ([Graham et al. 2004](#ref-graham2004)).

#### Obscure record queries

Integrating data records that are heterogenous in concept and varying in file format can be a challenge. The wide variety of content in NHCs necessitate a consistent but also flexible framework for data abstraction. A working solution to this challenge is semantic annotation, which typically uses a Resource Description Framework (RDF), an XML based language that provides semantic interoperability, to attach machine-readable metadata (as opposed to classic text annotation) to text documents or other unstructured content ([Lo Turco et al. 2019](#ref-loturco2019)). The Semantic Web, widely associated with Web 3.0, allows for relationships between documents and also among data elements within those documents ([Pennington et al. 2019](#ref-pennington2019)). The SPARQL query language directly supports information in RDF form, and allows for the option to query semantic annotation fields ([Skevakis et al. 2014](#ref-skevakis2014); [Stork et al. 2019](#ref-stork2019)) .

#### What to do with query output

Many natural history museum databases include records that are georeferenced, allowing data visualization of query results ([Janicki et al. 2016](#ref-janicki2016)). The mapview package in program R provide user-friendly functions to interactively map record coordinates with any combination of user-defined auxiliary data annotated to each point ([Appelhans et al. 2022](#ref-appelhans2022)). Mapview functions as a data-driven API for the more extensive leaflet package, which provides additional functionality and customization options ([Lovelace et al. 2019](#ref-lovelace2019)).

Appelhans T, Detsch F, Reudenbach C, Woellauer S (2022) Mapview: Interactive Viewing of Spatial Data in R. Version R package version 2.11.0.9000URL <https://github.com/r-spatial/mapview>

Boettiger C (2018) RDF Storage and Retrieval with rdflib

Chamberlain SA, Boettiger C (2017) R Python, and Ruby clients for GBIF species occurrence data

Gadelha Jr LMR, de Siracusa PC, Dalcin EC, et al (2021) A survey of biodiversity informatics: Concepts, practices, and challenges. WIREs Data Mining and Knowledge Discovery 11:e1394. <https://doi.org/10.1002/widm.1394>

Graham CH, Ferrier S, Huettman F, et al (2004) New developments in museum-based informatics and applications in biodiversity analysis. Trends in Ecology & Evolution 19:497–503. <https://doi.org/10.1016/j.tree.2004.07.006>

Janicki J, Narula N, Ziegler M, et al (2016) Visualizing and interacting with large-volume biodiversity data using client–server web-mapping applications: The design and implementation of antmaps.org. Ecological Informatics 32:185–193. <https://doi.org/10.1016/j.ecoinf.2016.02.006>

Lo Turco M, Calvano M, Giovannini EC (2019) DATA MODELING FOR MUSEUM COLLECTIONS. Int Arch Photogramm Remote Sens Spatial Inf Sci XLII-2/W9:433–440. <https://doi.org/10.5194/isprs-archives-XLII-2-W9-433-2019>

Lovelace R, Nowosad J, Muenchow J (2019) Geocomputation with R. Chapman and Hall/CRC, New York

Pennington DR, Link to external site this link will open in a new window, Cagnazzo L (2019) Connecting the silos: Implementations and perceptions of linked data across European libraries. Journal of Documentation 75:643–666. https://doi.org/<http://dx.doi.org/10.1108/JD-07-2018-0117>

prud’hommeaux E, Seaborne A (2013) SPARQL 1.1 Query Language

Reichman OJ, Jones MB, Schildhauer MP (2011) Challenges and Opportunities of Open Data in Ecology. Science 331:703–705. <https://doi.org/10.1126/science.1197962>

Saran S, Chaudhary SK, Singh P, et al (2022) A comprehensive review on biodiversity information portals. Biodivers Conserv. <https://doi.org/10.1007/s10531-022-02420-x>

Skevakis G, Makris K, Kalokyri V, et al (2014) Metadata management, interoperability and Linked Data publishing support for Natural History Museums. Int J Digit Libr 14:127–140. <https://doi.org/10.1007/s00799-014-0114-2>

Stork L, Weber A, Gassó Miracle E, et al (2019) Semantic annotation of natural history collections. Journal of Web Semantics 59:100462. <https://doi.org/10.1016/j.websem.2018.06.002>

Varela S, González-Hernández J, Casabella E, Barrientos R (2014) rAvis: An R-Package for Downloading Information Stored in Proyecto AVIS, a Citizen Science Bird Project. PLOS ONE 9:e91650. <https://doi.org/10.1371/journal.pone.0091650>

Wieczorek J, Bloom D, Guralnick R, et al (2012) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLOS ONE 7:e29715. <https://doi.org/10.1371/journal.pone.0029715>

Wilkinson MD, Dumontier M, Aalbersberg IJJ, et al (2016) The FAIR Guiding Principles for scientific data management and stewardship. Sci Data 3:160018. <https://doi.org/10.1038/sdata.2016.18>