Does TDWG Need an API Design Guideline?

Ian Engelbrecht[‡], Hester Maria Steyn[§]

‡ Natural Science Collections Facility, Pretoria, South Africa
§ SANBI, Pretoria, South Africa

Corresponding author: Ian Engelbrecht (ian@nscf.org.za)

Abstract

RESTful APIs (REpresentational State Transfer Application Programming Interfaces) are the most commonly used mechanism for biodiversity informatics databases to provide open access to their content. In its simplest form an API provides an interface based on the HTTP protocol whereby any client can perform an action on a data resource identified by a URL using an HTTP verb (GET, POST, PUT, DELETE) to specify the intended action. For example, a GET request to a particular URL (informally called an endpoint) will return data to the client, typically in JSON format, which the client converts to the format it needs. A client can either be custom written software or commonly used programs for data analysis such as R (programming language), Microsoft Excel (everybody's favorite data management tool), OpenRefine, or business intelligence software. APIs are therefore a valuable mechanism for making biodiversity data FAIR (findable, accessible, interoperable, reusable).

There is currently no standard specifying how RESTful APIs should be designed, resulting in a variety of URL and response data formats for different APIs. This presents a challenge for API users who are not technically proficient or familiar with programming if they have to work with many different and inconsistent data sources. We undertook a brief review of eight existing APIs that provide data about taxa to assess consistency and the extent to which the Darwin Core standard (Wieczorek et al. 2021) for data exchange is applied. We assessed each API based on aspects of URL construction and the format of the response data (Fig. 1).

While only cursory and limited in scope, our survey suggests that consistency across APIs is low. For example, some APIs use nouns for their endpoints (e.g. 'taxon' or 'species'), emphasising their content, whereas others use verbs (e.g. 'search'), emphasising their functionality. Response data seldom use Darwin Core terms (two out of eight examples) and a wide range of terms can be used to represent the same concept (e.g. six different terms are used for dwc:scientificNameAuthorship). Terms that can be considered metadata for a response, such as pagination details, also vary considerably. Interestingly, the public interfaces for the majority of APIs assessed do not provide POST, PUT or DELETE endpoints that modify the database. POST is only used for providing more detailed request

bodies to retrieve data than possible with GET. This indicates the primary use of APIs by biodiversity informatics platforms for data sharing.

An API design guideline is a document that provides a set of rules or recommendations for how APIs should be designed in order to improve their consistency and useability. API design guidelines are typically created by particular organizations to standardize API development within the organization, or as a guideline for programmers using an organization's software to build APIs (e.g., Microsoft and Google). The API Stylebook is an online resource that provides access to a wide range of existing design guidelines, and there is an abundance of other resources available online.

This presentation will cover some of the general concepts of API design, demonstrate some examples of how existing APIs vary, and discuss potential options to encourage standardization. We hope our analysis, the available body of knowledge on API design, and the collective experience of the biodiversity informatics community working with APIs may help answer the question "Does TDWG need an API design guideline?"

Keywords

application programming interface, standards, Darwin Core, JSON,

Presenting author

Ian Engelbrecht

Presented at

TDWG 2021

Acknowledgements

We would like to acknowledge the symposium organisers for the invitation to contribute, and the NSCF funder for financial support. Ben Norton and Elycia Wallis provided comments which greatly improved this abstract.

Author contributions

H. Steyn assessed the APIs considered for this analysis. I. Engelbrecht prepared the abstract and figures.

Conflicts of interest

We have no conflict of interest.

References

- Wieczorek J, Bloom D, Guralnick R, Blum S, Döring M, Giovanni R, Robertson T, Vieglais D (2021) Darwin Core: An Evolving Community-Developed Biodiversity Data Standard. PLOS One 7 (1): e29715- e29715. https://doi.org/10.1371/journal.pone.0029715

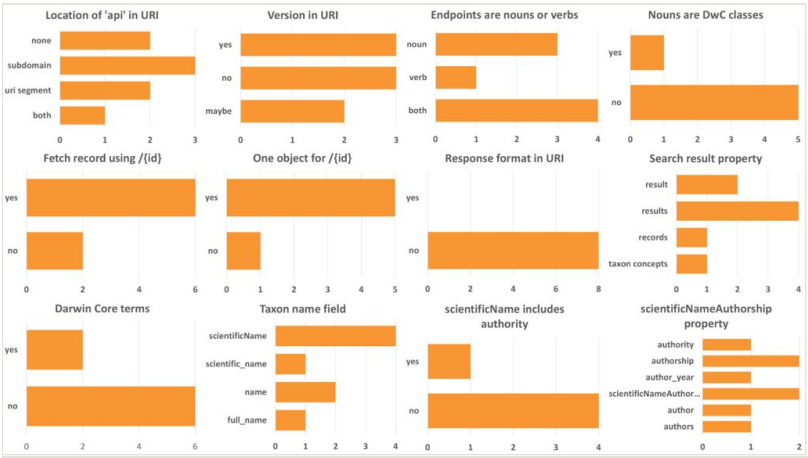

Figure 1.

An assessment of consistency in URI format and response data across eight RESTful APIs providing data for taxa, using criteria commonly used in API design and the Darwin Core standard for data sharing. A: URIs should indicate that they are part of an API either via a subdomain or a URL segment. B: APIs should be versioned as functionality evolves and the version indicated in the URI. C: API endpoints can be nouns or verbs, but nouns are typically recommended. D: If an endpoint is a noun, is it a Darwin Core class (in this case Taxon)? E: APIs typically provide access to individual data records via a URI that includes the record identifier as a URL segment, and not using a query string. F: GET requests for a single data resource should return a single object, and not an array. Conversely, queries should return an array, as in H. G: Some APIs require the response format be included in the URI. This is better achieved using request headers. H: When querying a collection, the response object includes a property for the array of search results. The name of that property is not standardized. I: Darwin Core provides a standard for data sharing between systems that have differing internal data schemas, facilitating use and understanding of data by third parties as well as easier integration and use of APIs. Darwin Core terms are not widely applied in APIs however. J: Related to I, we assessed the terms used for full taxon name in API response data, or the equivalent dwc:scientificName. K: The formal definition of dwc:scientificName is that it must include the taxon authority. It seldom does. L: Darwin Core also provides an atomic term for the taxon authority, dwc:scientificNameAuthorship. A range of different terms are currently used to represent this concept in API response data.