BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 8: e134931
doi: 10.3897/biss.8.134931

OPEN ACCESS

Conference Abstract

Automatically Generated Texts for Fauna and Flora from Structured Data Based on a Controlled Vocabulary

Adeline Kerner‡, Elie M. Saliba§, Nicolas Bailly|, Thierry Bourgoin§, Régine Vignes Lebbe§

‡ UMR 7207 – Centre de Recherche en Paléontologie - Paris CNRS – Sorbonne Université - MNHN, Paris, France
§ Institut de Systématique, Evolution, Biodiversité (ISYEB), MNHN, CNRS, Sorbonne Université, EPHE, Université des Antilles, Paris, France
| University of British Columbia / Beaty Biodiversity Museum, Vancouver, Canada

Corresponding author: Adeline Kerner (kerner@mnhn.fr)

Received: 17 Aug 2024 | Published: 19 Aug 2024

Citation: Kerner A, Saliba EM, Bailly N, Bourgoin T, Vignes Lebbe R (2024) Automatically Generated Texts for Fauna and Flora from Structured Data Based on a Controlled Vocabulary. Biodiversity Information Science and Standards 8: e134931. https://doi.org/10.3897/biss.8.134931

Abstract

Information systems like Xper3 and Fulgoromorpha Lists On the Web (FLOW) play a crucial role in managing and using biological data. These platforms store extensive collections of normalized data and structured taxonomic descriptions. By using controlled terminologies, they standardize the vocabulary, significantly enhancing the processes of identification, description, and comparison of various taxa. The massive assemblages of data hosted in these repositories could be reused to generate texts in natural languages automatically. The most immediate goal is to produce, from these information systems, more accessible and user-friendly displays in the form of taxon summary pages.

This automated production of textual outputs is a great addition that can be continuously updated as the databases evolve. Can structured data be reused to provide better species pages and to ensure updating if the data evolves? Will AI assist in this process, or will specific computing be needed? To address these questions, multiple possible approaches have been identified.

The most basic level of this process involves converting a single line from a taxon-by-character matrix into text that resembles natural language, similar to the descriptions

found in botanical or zoological guides. The primary goal is to move beyond the rigid format of matrix lines or lists of characteristics (such as characters and states) of a species, and instead generate a coherent, easy-to-read paragraph intended for human eyes.

To achieve this objective, a first solution is given by *Descrxp*, a tool currently developed in conjunction with Xper. The user specifies the desired key outline for the output, and *Descrxp* fills this canvas using the database contents. While this approach is highly reliable and adaptable to various contexts, its drawback is being labor-intensive, and requiring significant human input and oversight (Fig. 1).

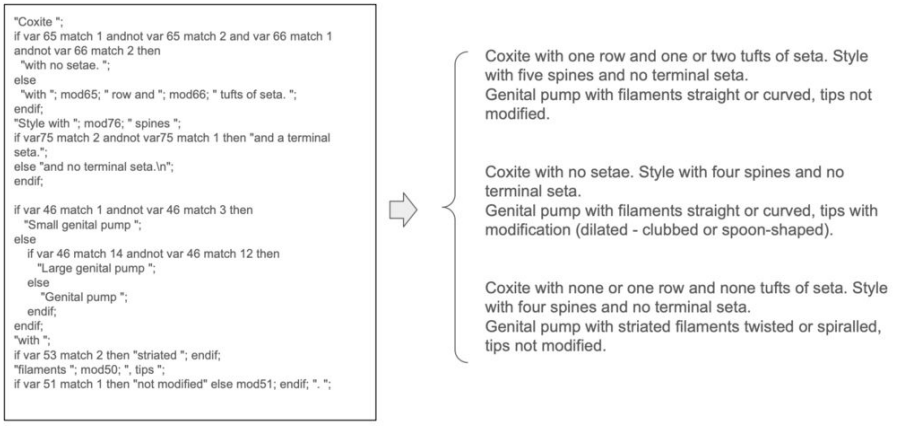

Figure 1.

On the left, a key outline defining how to describe the male genitalia (case of an Xper3 database on phlebotomine sandflies, Diptera). On the right, three text descriptions generated from this template, based on data from three different species.

The second solution uses AI. Data on well-known taxa, and on more obscure taxa have been tested with ChatGPT 3.5 or 4.0. ChatGPT succeeds well at generating natural language descriptions for well-known taxa (Fig. 2), providing text that appears accurate and coherent due to the extensive information available online for these groups.

However, when it comes to more obscure taxa, less represented on the internet, the results can be more inconsistent and unpredictable (Fig. 3). This variability can be explained by the quality of the data the model has been trained on, which is significantly richer for well-known species. Consequently, for lesser-known groups, the generated descriptions may lack the precision and reliability seen with more familiar taxa.

However, if more careful attention is given to the output of chatGPT (Fig. 4), it is possible to notice that the AI filled part of the description with inaccurate characters. Due to its inability to recognize specific scientific terminology, and to process hierarchical information, chatGPT ends up producing an over-generalized and redundant description.

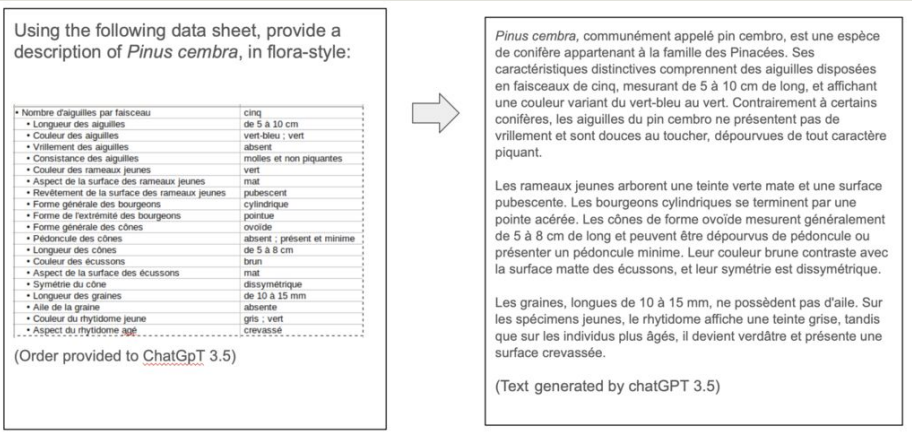

Figure 2.

On data related to common species (e.g., in *Pinus*), ChatGPT produces satisfactory texts (input on the left, output on the right).

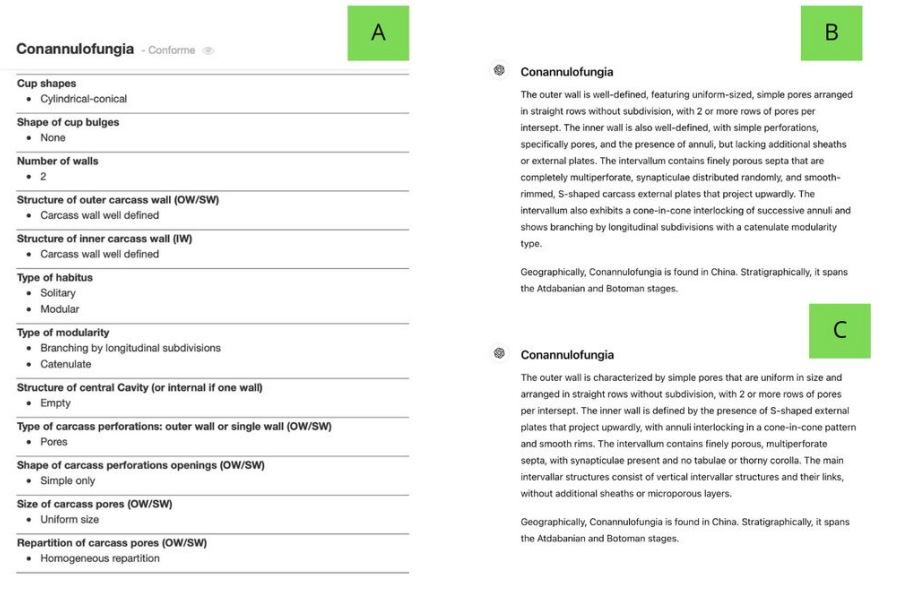

Figure 3.

Example with a database on Archaeocyatha (A) Genus sheet from Xper3 (B) Description generated by chatGPT 4.0 based on a knowledge base whose data are structured with several levels of dependencies (C) Description generated by chatGPT 4.0 based on a knowledge base whose data have a minimal hierarchical structure.

To go further, it would be possible to use a comparison between taxa to produce a text that highlights the remarkable features of a taxon. Moreover, it could be interesting to try

and synthesize several lines of the initial matrix to extract generalized descriptions of groups of taxa, such as the description of a genus based on its included species.

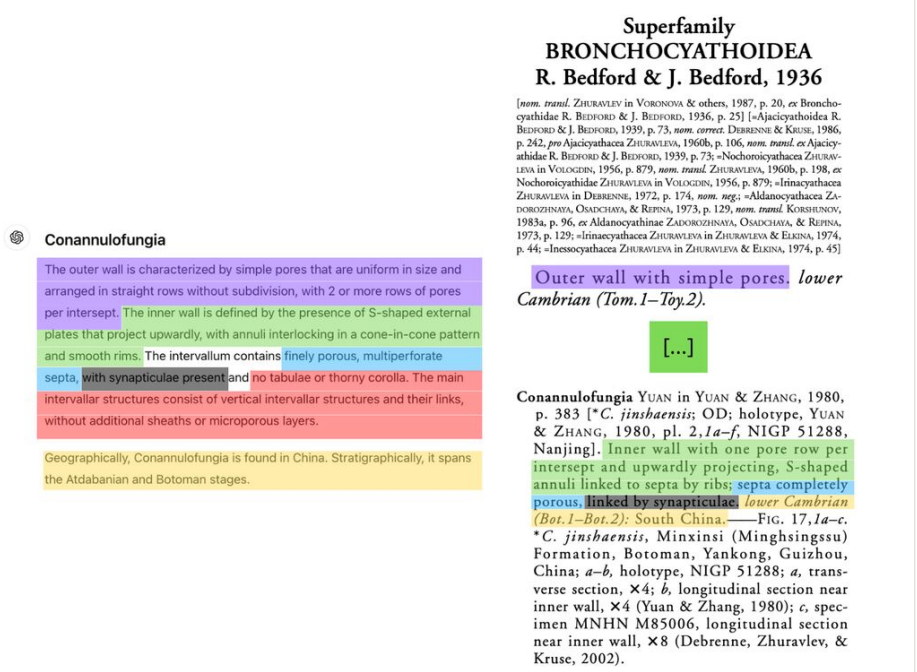

Figure 4.

On the left: same description generated with chatGPT 4.0 as Fig. 3C. On the right the description from the reference treatise (Debrenne et al. 2012). Equivalent structures are denoted by the same colors. The red part is a random add-on.

Using data from FLOW, Fishbase, and Xper3, the French ACDC (Counterfactual Learning for Controlled Data-to-text) project is creating a variety of text outputs. These range from "fill-in-the-blank" canvas to fully AI-generated content, utilizing data matrices derived from these pilot databases.

Keywords

AI, data-to-text, Xper3

Presenting author

Adeline Kerner

Presented at

SPNHC-TDWG 2024

Funding program

ACDC (Counterfactual Learning for Controlled Data-to-text) project, ANR-21-CE23-0007

Conflicts of interest

The authors have declared that no competing interests exist.

References

- Debrenne F, Zhuravlev AY, Kruse PD (2012) Treatise Online no. 50: Part E, Revised, Volume 4, Chapter 19: Systematic descriptions: Archaeocyatha. Treatise Online https://doi.org/10.17161/to.v0i0.4335