BISS Biodiversity Information Science and Standards

Biodiversity Information Science and Standards 3: e37198
doi: 10.3897/biss.3.37198

OPEN ACCESS

Conference Abstract

# Examining Herbarium Specimen Citation: Developing a literature-based institutional impact measure

Nicky Nicolson[‡,§], Alan Paton[‡], Sarah Phillips[‡], Allan Tucker[§,‡]

‡ Royal Botanic Gardens, Kew, Richmond, United Kingdom
§ Department of Computer Science, Brunel University, London, United Kingdom

Corresponding author: Nicky Nicolson ( nicky.nicolson@brunel.ac.uk)

Received: 12 Jun 2019 | Published: 19 Jun 2019

Citation: Nicolson N, Paton A, Phillips S, Tucker A (2019) Examining Herbarium Specimen Citation: Developing a literature-based institutional impact measure. Biodiversity Information Science and Standards 3: e37198. https://doi.org/10.3897/biss.3.37198

## Abstract

Herbarium specimens are critical components of the research process - providing "what, where, when" evidence for species distributions and through type designation, providing the basis for un-ambiguous, standardised nomenclature facilitating the interpretation of scientific names. Specimen references are embedded within research article texts, by convention usually presented in a relatively formalised fashion. As this is a domain-specific practice, general publishers tend not to provide tools for detecting and tracking specimen references to enable bibliometric-style calculations and navigation to the referenced specimen, as is common practice in literature reference management. This means that it is difficult to measure impact, which affects both the individuals responsible for the collection and determination of herbarium specimens (McDade et al. 2011), and the institutions responsible for their long-term management.

Specimen digitisation - creating searchable data repositories of metadata and/or images - has enabled many new and larger scale uses for herbarium specimens and their associated data, and stimulated interest in quantifying usage and measuring institutional impact. To date, these impact measures have been conducted by examining usage statistics for specimen portals, or by text searching for specimen identifier patterns.

This research uses text mining and document classification techniques to detect article sections likely to contain specimen references, which are then extracted, classified and counted. A dataset of taxonomic publications categorised into paragraph-level units is used to train a text classifier to predict the presence of specimen references within component units of articles (sections or paragraphs). The input to the classifier is a set of features derived from the text contents of paragraphs, which detect content such as latitude/longitude, dates and bracketed lists of herbarium codes. Article units classified as containing specimen references are processed to extract a minimal representation of the specimen reference, including the abbreviated codes for the institutional holder(s) of the specimen material. This allows total and per-institution counts to be calculated, which can be compared to datasets of Global Biodiversity Information Facility data citations, to institutional-level type citations in nomenclatural acts recorded by the International Plant Names Index and to usage statistics recorded by institutional data repositories. As well as counting specimen references, distinct specimen reference styles are detected and quantified, including the use of numeric and persistent identifiers (Güntsch et al. 2017) which can be used to access a standardised metadata record for the specimen.

We will present an assessment of the classification and detection process and initial results, and discuss future work to develop this approach to work with different kinds of literature inputs. These techniques have the potential to allow institutions to make better use of existing information to help assess the use and impact of their specimen and data holdings.

## Keywords

specimen citation, text classification, text mining, citation metrics

## Presenting author

Nicky Nicolson

## Presented at

Biodiversity_Next 2019

## References

- Güntsch A, Hyam R, Hagedorn G, Chagnoux S, Röpert D, Casino A, Droege G, Glöckler F, Gödderz K, Groom Q, Hoffmann J, Holleman A, Kempa M, Koivula H, Marhold K, Nicolson N, Smith V, Triebel D (2017) Actionable, long-term stable and semantic web compatible identifiers for access to biological collection objects. Database 2017 https://doi.org/10.1093/database/bax003

- McDade L, Maddison D, Guralnick R, Piwowar H, Jameson ML, Helgen K, Herendeen P, Hill A, Vis M (2011) Biology needs a modern assessment system for professional productivity. BioScience 61 (8): 619-625. https://doi.org/10.1525/bio.2011.61.8.8