Conference Abstract

Integrating Collector and Author Roles Across Specimen and Publication Datasets

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

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

Abstract

This work builds on the outputs of a collector data-mining exercise applied to GBIF mobilised herbarium specimen metadata, which uses unsupervised learning (clustering) to identify collectors from minimal metadata associated with field collected specimens (the DarwinCore terms *recordedBy*, *eventDate* and *recordNumber*). Here, we outline methods to integrate these data-mined collector entities (large scale dataset, aggregated from multiple sources, created programatically) with a dataset of author entities from the International Plant Names Index (smaller scale, single source dataset, created via editorial management). The integration process asserts a generic "scientist" entity with activities in different stages of the species description process: collecting and name publication. We present techniques to investigate specialisations including content - taxa of study - and activity stages: examining if individuals focus on collecting and/or name publication. Finally, we discuss generalisations of this initially herbarium-focussed data mining and record linkage process to enable applications in a wider context, particularly in zoological datasets.

Presenting author

Nicky Nicolson