



EDI Profile for EML

Margaret O'Brien, M. Gastil-Buhl *University of California, Santa Barbara*
John Porter *University of Virginia*, Mary Martin *University of New Hampshire*
Corinna Gries, Colin Smith *University of Wisconsin, Madison*
Mark Servilla, Duane Costa, Kristin Vanderbilt *University of New Mexico*



Why a Profile for EML?

- EML documents can be highly customized, with 1000's of XPaths possible because of EML's reuse of XML Types. Constraining practice to a subset of XPaths enables
- Code development to be more easily scoped
- Completeness of EML-generating tools to be measured

How is a Profile expressed?

- List of XPaths
- Rule-based language (e.g., Schematron)
- Text description (document)

What does it mean to “match a Profile”?

Variation among dataset EML documents is expected and necessary - both within a given research site, and between sites. Also, each data management system tends to follow a pattern of usage.

The profile is a concept of the aggregate, and not expected to completely match any one EML document.

Some XPaths will be present in every EML document.
Some XPaths will be present in at least some EML documents from every contributor.

Rarely used XPaths may not warrant inclusion in a Profile because the effort to code for them does not repay resources proportionately. Exclusion from a profile does not prohibit use; it just may not be supported by all tools.

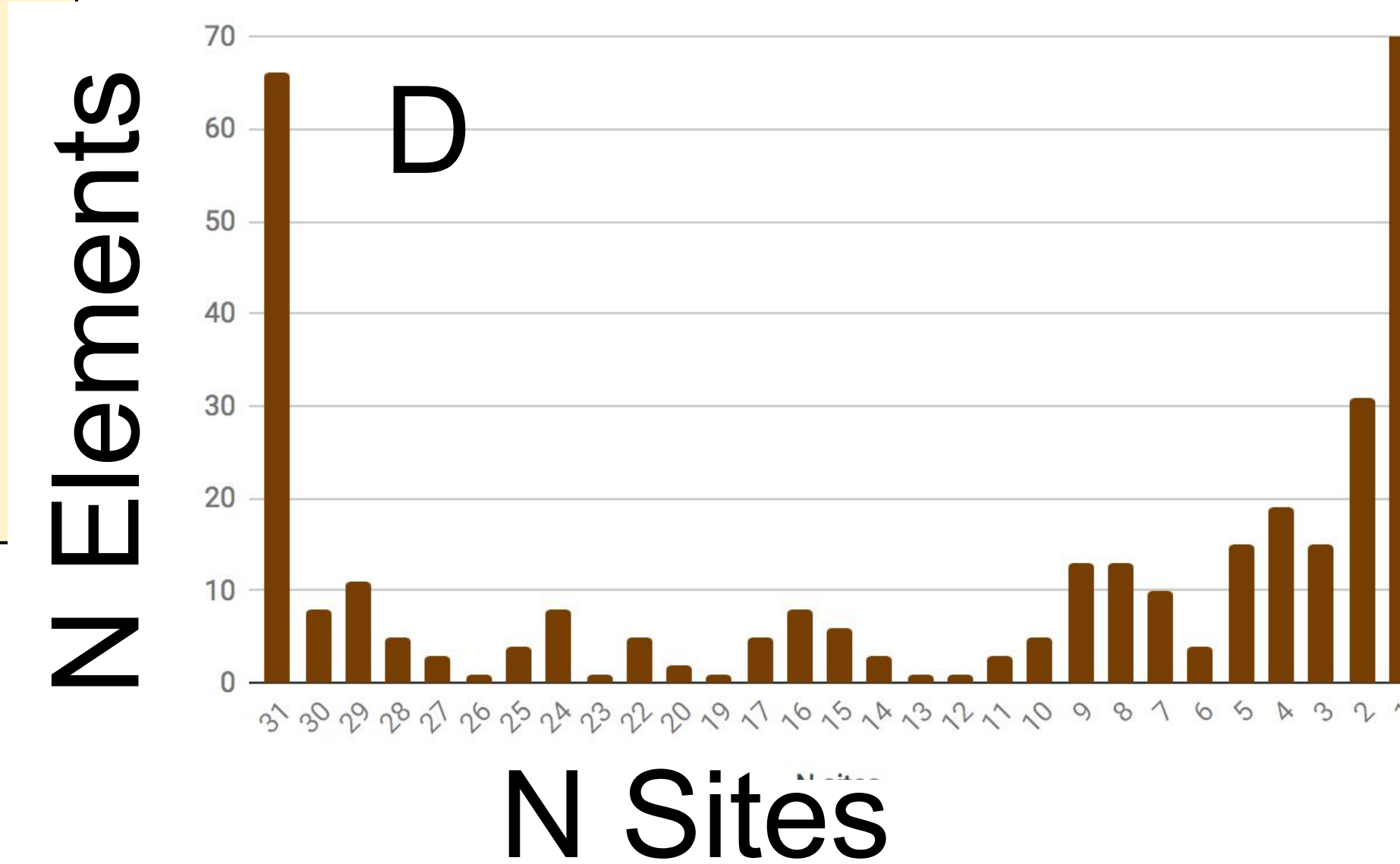
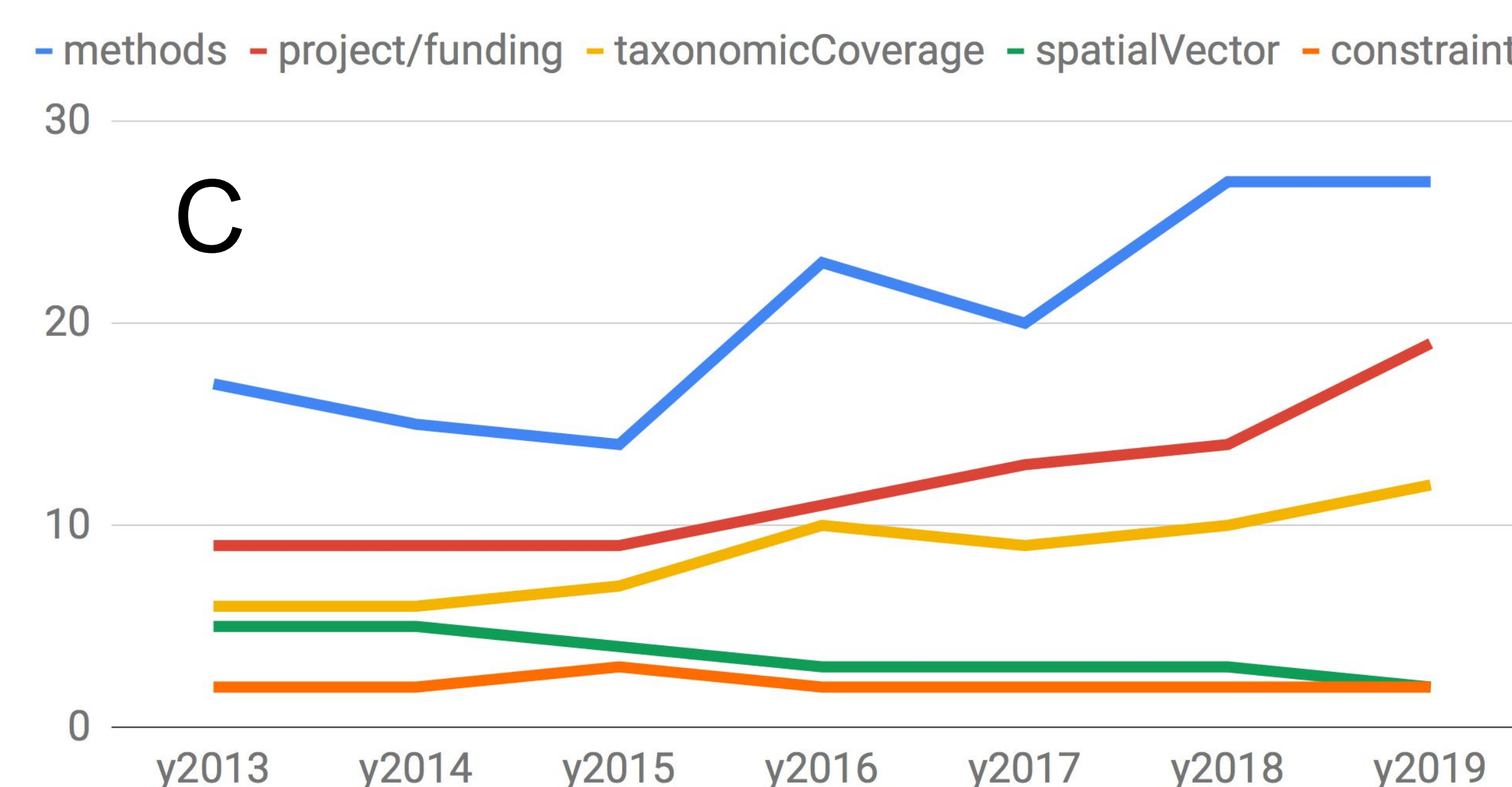
Profile to be informed by

- Common Patterns of Usage (CPU)
- Best Practice
- Indexing requirements

Preliminary Results

CPU data mined from 6838 EML docs in the EDI data repository, contributed by

- 30 LTER sites (past and current)
- 100s independent researchers (via EDI, included as a “site”)



A: Element presence by site, sorted by abundance (>1 site)

B: Abundance by year, sorted by abundance (>2 sites)

C: 5 Example elements' abundance.

D: Histogram

analysis by J. Porter, M. Servilla, M. Gastil-Buhl, M. Martin

<https://lternet.edu>, <http://environmentaldatainitiative.org>

This work is supported by the US National Science Foundation, grants ABI-1565103 and DEB-1629233

