Evaluating the First 1000 preprints on EcoEvoRxiv

The goal of this Hackathon is to quantify various attributes of first 1000 preprints to better understand preprint practices and publication pathways in Ecology and Evolution. There are currently well >1100 preprints on the new *EcoEvoRxiv*, which was established in 2018.

The aims of this project are to determine: 1) how long it takes for a preprint to become published, and how many preprints remain unpublished; 2) what countries make use of preprints; 3) how career stage and gender impact preprint use; 4) whether data and code are more likely to be shared in preprints; and 5) the extent to which authors make use of preprint servers for registered reports and community driven peer review (i.e., Peer Community In, PCI). 6) citation rates of preprints and published articles. Possibly also looking at social media. 7) how much community discussion occurs along side preprints. 8) versioning of preprints and whether titles change between preprint and published article.

The details outlined in this README have been pre-registered on OSF. We have also captured this pre-study research plan as a release.

Hackathon Participants

Anyone registered for the SORTEE conference is welcome to join the Hackathon which will take place October, 17th 2023 (8:30-10:30 PM AEST). The Hackathon will be held virtually via Zoom and run by Daniel Noble, Shinichi Nakagawa and Malgorzata Lagisz as part of the SORTEE conference.

Participants will need to provide details of their involvement in our delegate data sheet so that we can allocate papers, awknowledge contributions and maintain contact with participants.

Communication During the Hackathon

We will be using Zoom chat for conversations during the Hackthon. Anything after the Hackathon please use the GitHub Discussion Forum.

Preparation for the Hackathon

We will be using GitHub and we would recommend everyone signup and create an account.

Hackathon Schedule

1) Introductions and Formalities (5-10 min)
2) Introduction to the Hackathon and Where to get Material (15 min)

   - GitHub
   - Google Sheets
   - Google Form
   - Communication channels

3) Training session with one paper and Google form (Losia) (10 min)
4) Same assigned paper training everyone extracts data (10 min)
5) Questions and discussions (10 min)
6) Short Break (5 min)
7) Data collection (60 min)

We will probably not finish data collection during the Hackathon and it will be likely that those participating will need to collect additional data outside of the Hackathon in order to complete their 25 papers. As such, we're hoping to stick to the tentative dates outlined below to complete the data collection and a first draft of the manuscript.

1) 6th November, 2023: First round of data collection complete for each co-athor's 25 papers
2) 13th November, 2023: Second round of data collection complete for any additional papers
3) 20th November, 2023: Cross checking of data complete.
4) 27th November, 2023: Data analysis complete.
5) 13th December, 2023: First draft of manuscript complete and sent to all co-authors for feedback.

What will we do?

Participants that are part of the Hackathon will help collect data on a subset of the preprints from *EcoEvoRxiv*. A list of some of the information we would like to collect can be found below.

The first part of the Hackathon will be a training session where we will go through the data collection process and answer any questions. The second part of the Hackathon will be the data collection itself. We will use a Google Sheet to allocate preprints to participants. We will also use a Google Sheet to collate the data collected by participants.

What we intend to produce?

We intend to produce a manuscript that will be submitted to *Nature Ecology and Evolution* as a brief communication. We will also make the data collected available to the public.

The format of the paper as described by the journal is as follows: - Brief unreferenced abstract – 3 sentences, up to 100 words. - Title – up to 10 words (or 90 characters). - Main text – 1,000-1,500 words, including abstract, references and figure legends, and contains no headings. - Display items – up to 2 items, although this may be flexible at the discretion of the editor, provided the page limit is observed. - Extended Data – up to 10 items (figures and/or tables, linked from the main text in the html version of the paper). - Online Methods section should be included. - References – as a guideline, we typically recommend up to 20. Article titles are omitted from the reference list. - Brief Communications should include received/accepted dates. - Brief Communications may be accompanied by supplementary information. - Brief Communications are peer reviewed.

Co-authorship

If you would like to be listed as a co-author on the resulting manuscript: 1) You will need to have filled out the delegate data sheet so that we can contact you. 2) You will need to complete data collection for at least 25 preprints if you're in attendance. 3) If you cannot attend, you must either contribute to the data collection for at least 25 preprints or checking of 50 preprints outside of the Hackathon. 4) Contribute to any discussion and feedback on the resulting manuscript. 5) Reply promptly to any emails regarding the manuscript (within 2 weeks).

Project organisers will determine their own authorship position. Co-authorship is otherwise determined by the degree of contribution of a given author (as determined by project organisers) or for equal contribution authorship, authors are listed in alphabetical order.

Organisers retain the right to to deny co-authorship to anyone who does not meet the above criteria and in special cases (e.g., not following code of conduct, not providing authorship details). If you cannot fulfill the authorship criteria listed above or are deemed not to have contributed substantially, then you will be acknowledged in the manuscript as a Hackathon contributor.

We have already started drafting the manuscript (see the `ms/ms.qmd` file). The file can be rendered to a word or pdf document. We will be using the collected data to provide a quantitative analyses of the *EcoEvoRxiv* preprints to answer key questions outlined in the manuscript.

Preprint Data Collection

Data for individual preprints will be collected using our Google Form. The form will be filled out by each Hackathon participant for a subset of preprints. Some journal-level data will be collected later from journal websites and databases.

Data to be manually collected on preprints

- **Extractors first name**: This will be used to identify who collected the data.
- **Extractors last name**: This will be used to identify who collected the data .
- **Preprint DOI**: Copy and paste from the provided list of preprints from the Preprint Meta-data data file. Copy the whole link containing DOI and paste it.
- **Submitting/ corresponding authors firstname**: First name (given name) and second name (or initials, if provided) of the author submitting the preprint - copy and paste from the master list provided. Usually there is one submitting author in the meta-data preprint list. If there are multiple authors then use the first author listed on the preprint.
- **Submitting/ corresponding authors lastname**: Last name (family name) of the author submitting the preprint - copy and paste from the master list provided. If there are multiple authors then use the first author listed on the preprint.
- **Country of the corresponding / submitting author**: Country of affiliation for the author submitting the preprint. You should be able to see it on the first page of the preprint (open the preprint landing page on EcoEvoRxiv using the DOI link from the master list). If multiple affiliations are listed only use the country of first affiliation. To Enter the country name following the standard names ISO 3166 which are in the first column of the table here If you cannot determine the affiliation country enter "NA".
- **Year of first publication for corresponding/submitted author**: Use Google Scholar to collect publication year of journal article of the author submitting the preprint. If submitting author has no Google Scholar profile, enter '0'.
- **How many versions of the preprint exist**: This can be found in the Preprint Meta-data data file.
- **Taxa being studied**: What taxa are focus of the study described in the preprint? (select all that apply). Levels include "Plants", "Animals", "Fungi", "Algi", "Invertabrates", "Vertebrates", "Microorganisms (bacteria, viruses)", "Other". (select all that apply)
- **Discussion on the preprint?**: Use preprint DOI link to look at its EcoEvoRxiv landing page. Have there been any comments made on the preprint in the discussion panel on the preprint landing page?
- **Type of article**: Research Article - is any article-like manuscript intended for publication in research journals with new empirical findings; methods paper, are papers presenting new methodological or computational approaches; Reviews and Meta-analyses, are papers quantitatively or qualitatively synthesising a given topic; opinions, are usually short papers providing new perspectives on a topic; comments, are papers that explicitly comment on an already published research article.
- **Link to Data for Preprint**: Scroll through the first version of the preprint and any related files on EcoEvoRxiv to see if a link (or accession number) to underlying data is provided. From the first preprint, copy and paste link (or accession number) to data repository, if reported. If the preprint is not an empirical or equivalent article based on data and data analyses (so usually for opinions, comments, narrative reviews, theoretical paper, etc.) enter 'NA'. If the preprint is based on data/code but no link

(or accession number) to data repository is reported, enter 'none'. If multiple links are provided, enter the first one.

- **Link to Code for Preprint**: Scroll through the first version of the preprint and any related files on EcoEvoRxiv to see if a link (or accession number) to underlying code is provided. From the first preprint, copy and paste link (or accession number) to data repository, if reported. If the preprint is not an empirical or equivalent article based on data and data analyses (so usually for opinions, comments, narrative reviews, theoretical paper, etc.) enter 'NA'. If the preprint is based on data/code but no link (or accession number) to code is reported, enter 'none'. If multiple links are provided, enter the first one.

- **Number of citations to preprint**: Collect manually from Google Scholar. If the preprint version is NOT on Google Scholar enter "999". If the preprint version is online and clearly indicated (i.e., EcoEvoRxiv) then take that citation count.

- **PCI recommendation**: Has the preprint been recomended by Peer Community In (PCI)? Two steps to try. 1) PCI recommendations may be associated with the preprint on the preprint landing page. If so, it will clearly indicate this and or provide a link to a recommendation; 2) In a Google Search, add the following search query: "peercommunityin.org recommendation"TITLE OF PREPRINT" ". Replace TITLE OF PREPRINT with the preprint title. See if it's in the first page of the search. If not, then assume it is not recommended.

- **Publication DOI**: If the preprint has been published as a journal article, provide the DOI of that article. Note that if already known to be published there will be a published DOI in the Preprint Meta-data data file. Please copy and paste this DOI. If there is no DOI in the Preprint Meta-data data file then please follow these steps: 1) copy the preprint title; 2) Search in Google for the preprint title and determine on the first page of the search whether it has been published in a peer-reviewed journal. 3) If preprint has been published but it is not yet recorded in the Preprint Meta-data data file then please copy and paste the DOI of the published article. 4) If the preprint has not been published then please indicate 'NA'.

- **Journal name**: If the preprint has been published as a journal article, provide journal name (full name). Please use lower case for all letters. If the preprint has not been published, enter 'NA'.

- **Double-blind Peer review**: Does the published journal have a policy of double blind or blind peer review? Please review the journals policy page. If the journal does not have a policy page or it is unclear then please indicate 'No Blinding'.

- **Journal impact factor/Cite Scores**: Can be collected using the journal name automatically using R packages.

- **Publicaton Date**. If the preprint has been published as a journal article, provide (first) publication date (Month, Day, Year). If the DOI is available in the Preprint Meta-data data file you will still need to visit the journal website to collect the publication date.

The publication date should be listed on the landing page of the published version of the preprint just under author affiliation, but this may vary depending on the journal. If the preprint has not been published, leave empty.

- **Title Change**: If the preprint has been published as a journal article, has the title the changed between the first version of the preprint and the published article? Note that any word change is sufficient for a 'yes'. If the preprint has not been published, leave empty.
- **Number of citations to publication**: Collect manually from Google Scholar. If the preprint has not been published, leave blank.
- **Link to Data for Publication**: If the preprint has not been published enter 'NA'. If the preprint is not and empirical or equivalent article based on data and data analyses (so usually for opinions, comments, narrative reviews, theoretical paper, etc.) enter 'NA'. If the preprint is based on data/code but no link (or accession number) to data repository is reported, enter 'none'. If the preprint has been published as a journal article, and has such link reported, copy and paste this link (or accession number) below.
- **Link to Code for Publication**: If the preprint has not been published enter 'NA'. If the preprint is not and empirical or equivalent article based on data and data analyses (so usually for opinions, comments, narrative reviews, theoretical paper, etc.) enter 'NA'. If the preprint is based on data/code but no link (or accession number) to code is reported, enter 'none'. If the preprint has been published as a journal article, and has such link reported, copy and paste this link (or accession number) below.
- **Gender**: Use R package that determines likely gender of corrsponding author based on first name
- **Preprint comments**: Do YOU have Any additional comments on the preprint. Make note of any relevant comments about the preprint that may be useful for the manuscript.
- **Publication comments**: Do YOU have any comments on the published article related to an EcoEvoRxiv submission. Make note of any relevant comments about the preprint that may be useful for the manuscript.

Frequently Asked Questions (FAQ)

During the tutorial by Losia and the following training session of the Hackathon, a range of questions were raised by the delegates. Most of them have already been accommodated in the updated Google Form and are likely no longer relevant; for sake of clarity and transparency, the questions are listed here:

- **Q: Does the form require us to have access to the full text of the publication?** A: Yes, it would for some questions. If you cannnot access it, move to another preprint and leave this one to someone who does have access. If you cannot access to published articles at all, contact Shinichi, Dan, or Losia.
- **Q: Should we use ORCIDs, if they are available on any of the preprints?** A: Yes, if it is difficult to find the submitting author on Google Scholar, ORCIDs might be

a helpful alternative.

- **Q: What should we fill in for citation number when the preprint is *not* on Google Scholar?** A: The field is restricted to numeric values; please use '999'.
- **Q: Which date should we use for publication date (publication date online or publication date print)?** A: Please use the date of online publication.
- **Q: Are we submitting the form once we've entered all the data in the form?** A: Yes. Once all required fields are filled in (marked with '*'), the form can be submitted.
- **Q: What should we do if a preprint/publication has more than 1 repository in which they stored data/code?** A: Please pick either of the repositories. Key is whether the code and data are available at all, so as long as it is at least in one place, it's good.
- **Q: What should we do if the data/code is submitted as supplementary files?** A: Please use the DOI of the paper to which the supplements belong, and add a comment in `Preprint_comments` or `Publication_comments` at the end of the Google Form.
- **Q: In meta-analyses, the taxa being studied might be multiple. How should we record this?** A: You can tick multiple boxes under `Taxa_being_studied`, so please tick all that apply.
- **Q: What country code should we use for `Submitting_author_country`?** A: Please use the ISO 3166 English short name (i.e., the first column) as listed here: ISO online browsing platform.
- **Q: What should we fill in when the publication date is only month and year (e.g., March 2023)?** A: In these instances, please select the first day of the month; i.e., 01/03/2023.
- **Q: What should we do if the corresponding author has multiple affiliations?** A: Please use the first affilitation to fill in `Submitting_author_country`.
- **Q: If the author's first publication in Google Scholar is a thesis or conference paper do we still use that year? Or do we use the year for their first peer-reviewed publication?** A: Please use the first peer-reviewed publication.
- **Q: Should we use first and middle name (e.g., Israel T) or only first name (Israel) for the corresponding author?** A: Please use the first and middle names as provided.

Meta Data for files in `data/` folder

This is the main dataset that contains the most recent version of each preprint submitted to EcoEcoRxiv along with whether the preprint is published (some might be missing) and the dates preprint was published. It should be noted that the date the most recent publication was submitted should be the date of publication of a published DOI listed.

Column Names and Information for `20231003_EER_preprints_metadata.xlsx`**. Note: These columns are the same descriptors as that found in the** **Preprint Meta-data**

- **Preprint ID**: Janeway's internal identifier for the preprint
- **Preprint Title**: Title of the preprint
- **Preprint DOI**: DOI of the preprint
- **Publisher DOI**: DOI of the postprint/publisher's article, if any
- **Reuse Licence**: Creative Commons reuse licence
- **Submission Date**: Date preprint was submitted to EcoEvoRxiv
- **Accepted Date**: Date preprint was accepted to EER
- **Published Date**: Date preprint was published in EER (may differ from accepted date)
- **Update Date**: Date preprint was last updated by an EER moderator
- **Current Version**: Current version now
- **Version creation date**: Date that version was created/submitted (may differ from update date)
- **Submitting Author**: Name of submitting author
- **Submitting Author Email**: Submitting author's email address
- **Authors List**: List of all authors
- **Total authors**: Total number of authors

Column Names and Information for `20231003_EER_allversions.xlsx`

This data contains version history information for each preprint on EcoEvoRxiv. It is meant to supplement the main dataset `20230824_EER_Preprints_metadata.xlsx`.

- **Preprint ID** Title of the preprint
- **Preprint DOI** DOI of the preprint
- **Publisher DOI** DOI of the postprint/publisher's article, if any
- **Reuse Licence** License selected by author
- **Submitting Author** Name of submitting author (first, last)
- **Submission Date** Date of version submission
- **Accepted Date** Date of acceptance for version
- **Published Date** Date published on EcoEvoRxiv
- **Update Date** Date published preprint was updated
- **Current Version** Current version number of preprint
- **Version date** Date of version