SciPost Bundled Datasets

Context - Since its inception in 2016, SciPost has been operating as a community-based Genuine Open Access publisher in the physical sciences. With more than 3300 publications from more than 7000 authors associated to more than 1200 organizations throughout the world, it exemplifies how community-led, open refereed, quality-focused publishing can be successfully deployed as an attractive Diamond alternative for scientists. The bulk of SciPost's activities has up to now been the publishing of scientific articles in a traditional format; there is however by now a demonstrated need for "*beyond-article"* material to be given its due regard within the realm of scientific publishing.

Motivations - When we speak to scientists, mention is often made of the following problems:

- due to their relatively inflexible format, journal articles cannot contain all the relevant and publishable material produced during the research process, or all that is necessary to properly ensure reproducibility;
- journal articles, being static entities, do not enable the use of modern interactive visualization tools;
- the "venue" problem: there is a dearth of high-quality endpoints where "beyond-article" material can be thoroughly peer-reviewed, published, listed and credited; this translates into lack of potential recognition, and thus lack of motivation for authors to invest in such forms of publishing;
- the "interlinking" problem: the points where "beyond-article" material is typically made available (*e.g*. local or institutional repositories) do not systematically, transparently and fully interlink with related material.

With this project, our vision is to improve SciPost's palette of publishing venues by expanding the types of scientific output which can be published through it, adding to it a new fundamental type: Datasets.

Datasets can take on a flexible variety of possible forms: images, measurements, raw instrumentation outputs, filtered outputs with filtering protocols, plots with data and plotting scripts, interactive data visualizations, reruns or reproduction reports, benchmarks, initialization data, control data, test results, (small) databases, etc. Each Dataset is meant to be composed of the data itself, accompanied by the necessary descriptors to ensure intelligibility. Datasets are to be submitted and evaluated similarly to other publications at SciPost: the same thorough editorial procedure will be applied (including open refereeing and College-based decisionmaking). Production of accepted articles (including metadata handling and interlinking) will be performed to the same standards as for other publication types.

This project is designed to provide researchers with a high-quality, meaningful, and creditable outlet for research-level experimental, observational and/or computational data and their descriptors, thereby enhancing the completeness and reproducibility of their published output. It will:

- help develop interoperability standards (by defining strict minimal standards for publishable dataset descriptors, and installing a robust open peer review process on those);
- facilitate authoring, reviewing and publishing of dataset-related material, and its interlinking with the publication corpus (articles, codes etc);
- stimulate wider adoption of beyond-article material as being an integral part of scientific output;
- foster an environment where proper recognition of this form of alternative output can be given.

Related recent developments

As a first step in addressing the problems mentioned in the Motivations above, SciPost has in recent years begun extending the types of scientific output that can be peer reviewed and published on its platform. As a first representative of this "beyond-article" extension, SciPost deployed a journal in 2022, SciPost Physics Codebases, which is specifically tailored for computer codes and algorithms of relevance to research. As part of this, in order to address the "interlinking" problem, we introduced the concept of *Publication Bundles* in order to faithfully enshrine each published element as part of a bigger ensemble. By definition, a **Publication Bundle** is a set of interrelated publications of heterogeneous type, by identical or overlapping sets of authors, detailing different facets of the output of a single coherent research effort. Within SciPost Physics Codebases, for example, publications of type *Userguide* and *Codebase release* (each in themselves full-fledged publications with their own DOI) are grouped together into a Bundle, meaning that readers are naturally directed to the full bundle while looking at any of its elements (*e.g*. PYTHIA 8.3). All elements of the bundle are designed to be individually cited in further literature making use of this material.

The Bundles setup provides an infrastructural backbone which further provides:

- the ability to clearly separate scientific output into distinct but coherent parts, each with their individually appropriate format, each evaluated according to its own appropriate refereeing protocol, and each published with their individual DOI;
- the ability to have these distinct parts viewed and treated (among others in citations) like a coherent whole;
- the flexibility for authors of updating parts while maintaining a clear history of revisions and improvements;
- the ability to highlight different subsets of authors in different ways, depending on which part is concerned.

Once again using Codebases as an example, a Userguide can remain up-to-date while successive Codebase releases are published (each for example representing a new major/minor version of the codebase), each element being authored by author sets which are either identical, reordered, or differentiated.

Project plan

The first implementation step will be to define the scope, requirements and expectations of Datasets, with input from the community of SciPost users deemed likely to author submittable material (in particular large experimental collaborations) and academic/institutional data stewards. Particular attention will be given to detailing expectations on dataset descriptors (how the dataset was generated, formatted and organized; which tools enable its (re)use; maintenance and update plans) and ensure compliance with FAIR principles. This will naturally lead to formulating authoring guidelines, submission templates, practical refereeing guidelines and editorial handling instructions.

Separately from this conceptual and design-level work, a second step in the project (to occur in parallel/close succession to the first) will be to get the platform technically ready to handle the new material and workflows as an extra branch of the Bundles system. SciPost's current infrastructure is built using only open software, and is itself developed as open software available in our repository https://git.scipost.org/scipost , with documentation at https://docs.scipost.org . In line with this policy, the code output of the present project (namely: all the web infrastructure built to empower Datasets) will also be licenced as open software (AGPLv3) and be made openly available in the above-mentioned repositories.

These preparation steps will be followed by a rollout period, during which we will make the community aware of this new addition to our Bundles system, and attract and assist the first wave of contributors. After a certain expected period of adaptation and polishing up, the project will aim for accelerated upscaling of the rate of adoption, in particular by inviting all authors of past publications to "bundle up" datasets not only for their new papers, but also for their past publications at SciPost. Finally, we will work to ensure that these new publishing types are recognized and credited in existing academic evaluation systems.

Resources required

Initiation: 50k euros
Per year expenditures: depending on usage volume - for 100 Dataset publication objects, we estimate 50k euros expenditures.