Abstract Title: Research Resource IDentifiers for Key Biological Resources
Authors: AE Bandrowski, PhD.[1, 2]; TH Gillespie[1]; ME Martone, Ph.D.[1];
Affiliations: [1] Cetr. for Res. in Biol. Syst. at UCSD / [2] SciCrunch Inc.`
Funding: DA039832[1]

Key Biological Resources, according to the National Institutes of Health, are things that can work differently from lab to lab and protocol to protocol (NOT-OD-16-011[2]). These key resources; specifically: antibodies, transgenic organisms, cell lines and specialty chemicals; are some of the most likely culprits of variability in experiments, therefore contributing to irreproducibility in biological experiments. The National Institutes have asked that investigators include a plan for how these resources will be 'authenticated' to ensure, at a minimum, that they are what they say they are.

The problem with Key Biological Resources is not only that they may be faulty, but in 2013, Vasilevsky and colleagues showed that in 50% of published scientific papers, these resources are not 'findable'. This means that authors do not report sufficient detail to even allow human readers to identify these highly variable resources.

Research Resource IDentifiers, RRIDs, are stable unique identifiers, that can be included in the methods section of research papers. RRIDs are currently used to identify the cell line, antibody, transgenic organism or software tool used in the study (Bandrowski et al., 2015; Bandrowski & Martone, 2016) and they have been adopted by about 100 journals as a matter of publishing best practice, i.e., they are in the instructions to authors, and have appeared in over 500 journals. Including RRIDs in a paper, makes the resources more findable, in fact when Vasilevsky looked at the papers that contained RRIDs, using the same criteria as were used in the 2013 paper, more than 90% of resources were identifiable and therefore findable (Bandrowski et al., 2015). RRIDs were created to introduce a standardized and machine-friendly syntax for reporting on resource use so that it would be easy not only to identify the exact resource used, but to make it possible to find all studies that used a particular research resource.

RRIDs are becoming accepted in the scientific literature, especially in some fields such as neuroscience, where many of the top journals currently enforce the RRID standard (e.g., Neuron[3], Journal of Neuroscience[4]). RRIDs are issued by an authority for a particular type of resource, e.g., stock centers or model organism databases for organisms; Cellosaurus for cell lines, the antibody registry for antibodies and the scicrunch registry for software projects. While not all problems are solved by adding identifiers to a publication, there are three clear benefits of this workflow: 1) reagents are much more unambiguously identified as the records from the database are referred to by authors; 2) We have a means of identifying all papers that used a particular resource and 3) to find the identifiers authors, en masse, have to look at a specific set of web pages.

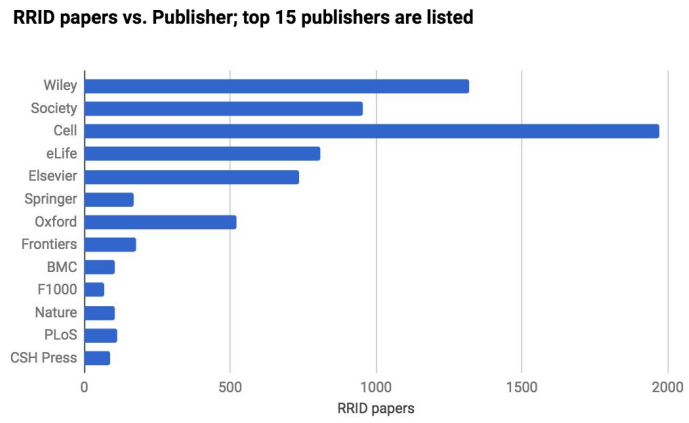
Essentially, RRID's give us a conduit through which information about how resources perform and known issues can be disseminated. Cell lines can become contaminated, may have been mislabeled, or can be overgrown by other cell lines at the stock centers where scientists buy the cell lines. It may be that problems are discovered after the cell lines were sent out and it may very well be that the laboratory that the cell line is used and the purchasing department are disconnected so the company can't let the

[1] https://projectreporter.nih.gov/project_info_details.cfm?aid=9039028&map=y
[2] https://grants.nih.gov/grants/guide/notice-files/not-od-16-011.html
[3] https://www.cell.com/neuron/rrid?code=cell-site
[4] http://www.jneurosci.org/content/preparing-manuscript

laboratory know that the cell line that they had sent out is problematic. There is a group of researchers, the International Cell Line Authentication Committee, ICLAC, who track misidentification problems and put out a report on their website[5], which can be accessed by authors. The ICLAC information can also be accessed on the Cellosaurus[6] website, the naming authority for cell lines. However, if the author is submitting a paper and is asked by the journal editor to include the RRID, the information must be accessed by the author at a central website[7], which contains an up to date list of contaminated and misidentified cell lines from ICLAC & Cellosaurus. Therefore, when looking up RRIDs, the author will come into contact with notes about problems and has time before publication to consider whether the particular problem impacts his or her conclusions.

Because RRIDs are machine readable, or at least detectable via regular expression, we and others have created tools that help our curators find all published papers that used RRIDs in the scientific literature. SciBot (GitHub), is a regular expression-based tool that scans the text of a paper that a curator is reading for RRIDs. It retrieves additional information about each RRID via the SciCrunch RRID resolver, and posts that information and the location of the RRID on the page to the Hypothes.is application programming interface. The information about the entity is brought into the article and anchored inside of the article as an annotation tag by the Hypothes.is client (see 1 minute video). The curator can then easily see the RRID annotations and confirm that they are correct, after which they are uploaded to SciCrunch. The data is made public and then picked up by the CrossRef Event database. Resolving services are provided to the community through the SciCrunch resolver, n2t resolver, and identifiers.org.

Current RRID governance is still coming into focus as a formal advisory board is not yet fully assembled, however the guiding principles for RRIDs are that they are light-weight, adaptable to the needs of the community, and built on current infrastructure. Multiple parties have sought to extend the RRID syntax to other research resource types including physical samples, instruments and chemicals. All of these extensions will need to be evaluated by the advisory board.



Statistics for RRID usage: The graph represents the publishers of periodicals that contain at least one RRID. On June 1, 2018 there were: 7,480 papers published that contained at least one RRID in 527[8] different journals (~9% of journals indexed in PubMed); the number of total RRIDs, noted by authors or curators in these papers, is 101,125 (~13 RRIDs/article). Data has been submitted to Cross Ref's Event database (RRID:SCR_016281; data for individual papers and RRIDs can be found via the Event API example).

[5] http://iclac.org/

[6] https://web.expasy.org/cellosaurus/

[7] http://rrid.site/

[8] https://docs.google.com/spreadsheets/d/1cGtY2oIMfBzg6MKei_S2niCbu0Cici2n17dZNLgBdNw/edit?usp=sharing