

The State of the Curation Nation

Scott Edmunds 

Published April 3, 2012

Citation

Edmunds, S. (n.d.). In *GigaBlog*. GigaBlog. <https://doi.org/10.59350/6rh88-esf96>

Keywords

Publishing, Technology, Biocuration, Conferences, Curation
Feature Image

Copyright

Copyright © Scott Edmunds 2012. Distributed under the terms of the [Creative Commons Attribution 4.0 International License](#), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Of the of the many issues needing addressing in this era of the so-called "data deluge" (apologies [genomics bingo](#)), on top of the well documented difficulties in computing power, bandwidth and storage keeping pace with data production, less attention has been paid on the efforts required to present and package this biological information to users. The key people managing and integrating this data are Biocurators, and this week is the [International Society of Biocuration](#)'s annual get together at the [Biocuration 2012](#) meeting in Washington DC. With growing [challenges](#) in data volumes and heterogeneity – particularly from sequencing technologies and with the promise of [nanopore](#) looming on the horizon, the meeting is a good opportunity to discuss some of the downstream consequences of these rapid developments amongst the people really harnessing the "data-tsunami".

With our publisher [BioMed Central](#) as one of the [sponsors](#) of the meeting, and with its relevance to our big-data scope and associated [GigaDB](#) database, [GigaScience](#) has been pleased to be present at the Georgetown University venue. Our new Biocurator Tam Sneddon has been representing the database side of *GigaScience*, and our Editor-in-Chief Laurie Goodman has also been there on behalf of the journal. The [first day](#) has covered many topics essential to keep on top of these large data-volumes such as community annotation, and workflows and tools to aid and automate tasks for data curators, producers and users.

Having been involved in the [crowdsourcing](#) of the genome of the deadly 2011 outbreak *E. coli* 0104:H4 strain, community annotation is a subject close to our hearts, and it was fantastic to see similar moves to open up and share the burden of annotation and analyses for species as diverse as Skates and Rays, with [Cathy Wu](#) presenting on [SkateBase](#). Wiki's are the obvious platform to handle these types of tasks, and Andrew Su presented on one of the most successful examples of these with [GeneWiki](#). Whilst we have written about this in a [previous meeting report](#), the user base continues to grow, and Andrew's most recent slides are available [here](#).

On top of the distributed "many-eyes"/"many-hands" approach, better automation of curation tasks is essential, and the workflows and tools session provided insight into where the state of the art of curation management currently is, with excellent examples on show in particular from [Reactome](#) and [PRIDE](#). The benefits of this were clearly shown by Attila Csordas (of [personal proteomics](#) fame) from the EBI, who showed that the PRIDE proteomics databases semi-automated pipeline and [tools](#) reduces curation time to 1/6th.

Being both a journal and database, [GigaScience](#) was well placed to take part in the "Databases & Journals – How to have a sustainable long term plan for journals and databases?" panel co-organized by our [editorial board](#) member [Francis Ouellette](#) and [Mike Cherry](#) (Stanford). Being quite a partisan pro-open-access audience and panel, Laurie joined other Editors-in-Chief including Thomas Lemberger from [Molecular Systems Biology](#), and David Landsman from [DATABASE](#), and Michael Galperin representing [the NAR Database issue](#), and all were equally open-data – pushing the need for all supporting data in a paper to be available to aid reproducibility, usability and prevent fraud. Discussion also turned to altmetrics and data citation, with Laurie in particular plugging our [work](#) with [DataCite](#) to give datasets

independently citable DOIs. Bringing a curation perspective to the discussion was the final panelist Pascale Gaudet (chairperson of the ISB), who discussed [BioDBcore](#), the community-defined checklist of the core attributes of biological databases that allows users to fully evaluate the scope and relevance of available resources.

A large proportion of the talks presented databases built on the open-source [GMOD](#) (generic model organism database) platform, and the conference is followed by the satellite [GMOD meeting](#), which also includes a [Galaxy](#) workshop. Laurie and Tam will be on hand all week to answer your questions about *GigaScience*, so feel free to grab them or contact us at editorial@gigasciencejournal.com. Many of the talks have been published in a [virtual Biocuration issue](#) of DATABASE, and you can also follow the action over the rest of the week on twitter at [#isb2012](#).

References

1. Howe D *et al.*, (2008). Big data: The future of biocuration Nature, 455 (7209), 47-50 DOI: [10.1038/455047a](#)
2. Sanderson K (2011). Bioinformatics: Curation generation Nature, 470 (7333), 295-296 DOI: [10.1038/nj7333-295a](#)
3. Burge S *et al.*, (2012). Biocurators and Biocuration: surveying the 21st century challenges Database, 2012 DOI: [10.1093/database/bar059](#)
4. Csordas A *et al.*, (2012). PRIDE: Quality control in a proteomics data repository Database, 2012 DOI: [10.1093/database/bas004](#)

The post [The State of the Curation Nation](#) appeared first on [GigaBlog](#).