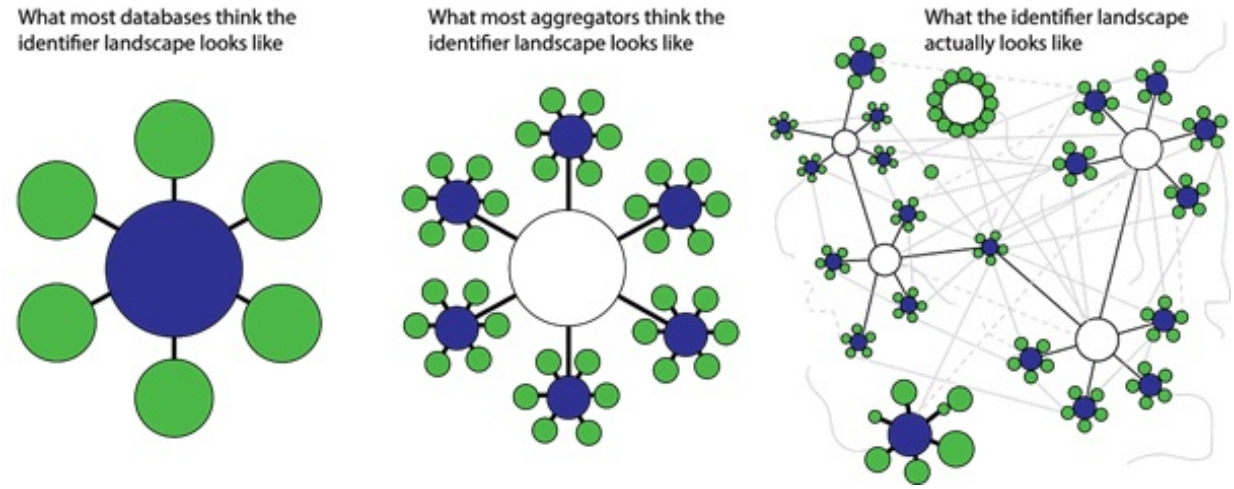
Future-proofing 'big data' biological research depends on good digital identifiers

June 29 2017



In the life sciences, if individual communities think about identifiers at all, it is usually in the context of a single database 'hub' and a variety of cross-referenced 'spokes', or an aggregation of these; however, the real complexity of the inter-relationships is often overlooked -- and with it, the importance of persistent identifiers to hold everything together. Identifier issues such as broken links undermine the flow and integrity of data for data providers and consumers alike. Credit: Julie McMurry and Lilly Winfree from the Monarch Initiative.

"Big data" research runs the risk of being undermined by the poor design of the digital identifiers that tag data. A group of worldwide researchers, led by Julie McMurry, at Oregon Health & Science University, has assembled a set of pragmatic guidelines to create, reference and

maintain web-based identifiers to improve reproducibility, attribution, and scientific discovery. The guidance, publishing June 29 in the open access journal *PLOS Biology* helps address the frequent problems associated with persistent identifiers linked to scientific data.

Over the past decade, the life sciences have drastically changed as data continues to evolve to be larger, more interdependent and natively web-based. In this landscape, the broader scientific research community has struggled to engineer this data for the web so that it is persistently accessible, reusable and attributable.

Depending on the individual database involved, identifiers can signify a gene, a genome, a chemical, an organism, a set of experimental data, or even a published article. The usefulness of all these items depends on the robustness and uniqueness of their respective identifiers, enabling them to be linked and discovered in perpetuity. The authors point out that the organic way in which most identifiers have arisen threatens that usefulness, and recognise that it is difficult to create and sustain persistent identifiers or web addresses that won't break and that are used consistently.

This work calls on professionals to do a better job of identifier engineering - according to emerging community-developed conventions - so that data can be utilized more effectively for scientific discovery. It also calls on users to be aware enough of these conventions, and of available tooling, to not get burned by broken links and missed connections.

"As with plumbing fixtures, the question of how identifiers work should only need to be understood by those that build and maintain them. However, everyone needs to know how identifiers should be used, and this is where convention is important," said McMurry. "Through this work, we hope to encourage all participants in the scholarly ecosystem -

including authors, data creators, data integrators, publishers, software developers, and resolvers - to adhere to best practice in order to maximize the utility and impact of life science data."

More information: McMurry JA, Juty N, Blomberg N, Burdett T, Conlin T, Conte N, et al. (2017) Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data. *PLoS Biol* 15(6): e2001414. doi.org/10.1371/journal.pbio.2001414