1384
doi:10.1017/S1431927619007657
Microsc. Microanal. 25 (Suppl 2), 2019


# Data Fabric Infrastructure for Heterogeneous Cell Biology Image Data

John Henry J. Scott[1*]

[1.] Office of Data and Informatics, Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, USA.
* Corresponding author: johnhenry.scott@nist.gov

Automated data acquisition technology has vastly improved our capacity to generate raw and lightly-processed datasets in a wide variety of scientific and engineering disciplines, including cell biology [1]. In many instances the bottleneck is no longer the measurement, simulation, or modeling step; now our ability to convert image data into scientific conclusions is limited by our data curation, post-processing, and information extraction technologies. These limitations are felt acutely by researchers struggling to exploit large image datasets and publish novel results in the scientific literature or distribute standalone data publications. However, there are more insidious consequences of the data deluge: it has become harder to discover and reuse datasets we have not acquired ourselves, it is more difficult than ever to share our work with others, and the transparency (and in some cases the reproducibility) of our research has been degraded [2-4].

Many sub-disciplines of science have responded to this challenge by developing new policies, standards, and technological infrastructure such as repositories, registries, portals, and data commons. This is certainly a positive trend and has begun to lower barriers to discovery, access, and interoperability, but unfortunately many of these sub-disciplines have undertaken this development as relatively insulated communities of practice. As a consequence, while the new tooling and infrastructure meets the needs of the community itself (e.g. cell biology), interoperability with other disciplines remains elusive and these systems frequently do not play well in distributed, federated architectures such as the emerging scientific data fabric [5]. There are several practical steps the cell biology community can take to improve this situation and avoid the pitfalls that have hampered other disciplines.

To counter the tendency toward isolation, data-savvy cell biologists can join with existing cross-domain research data initiatives, adopt international best practices, and participate in the process of refining and extending a common, shared set of standards and recommendations. In the area of policy and governance the GO-FAIR Initiative is rapidly gaining traction [6,7]. Distinct from the concept of open data, the FAIR principles strive to make research data Findable, Accessible, Interoperable, and Reusable by advocating: globally-unique, resolvable persistent identifiers (PIDs); registration of metadata in searchable, indexed registries; data vocabularies and other semantic assets that are harmonized across domains; clear assertion of licenses explaining terms and conditions for reuse; detailed provenance information for the datasets; and much more. A detailed explanation of the FAIR principles is beyond the scope of this paper, but Figure 1 provides a flavour of the power inherent in some of these changes. The explosive growth and success of the internet was fuelled in part by the design of the Internet Protocol (IP), a key architectural concept that links a wide variety of low-level networking details with higher level applications such as the word wide web, email, and end-user scientific applications (Figure 1, left diagram). Similarly, PIDs will play a similar role in the data fabric for scientific and engineering research data, linking a wide variety of instruments, modelling tools, computer simulations, and reference data with higher level data objects such as plots and visualizations, and analysis software packages (Figure 1, right diagram). Currently the most popular forum for this work is the Research Data

Alliance (RDA), a community-driven initiative supported by the European Commission, the NSF, NIST, and the Australian Government's Department of Innovation [8]. Although it began just a few years ago, the RDA has grown dramatically to 7700 members from 137 countries and conducts its work via 66 interest groups and 35 working groups. The first RDA sessions to explicitly focus on microscopy and microanalysis research data began at the 11th RDA Plenary meeting in Berlin in March 2018 [9].

References:

[1] Editorial, "The data deluge", Nature Cell Biology **14** (2012), p. 775.
[2] DR Baer and IS Gilmore, Journal of Vacuum Science & Technology A **36** (2018), p. 068502.
[3] AL Plant et al., Nature Methods **11** (2014), p. 895.
[4] NIST-NPL led workshop on "Improving Reproducibility in Research: The Role of Measurement Science" http://www.npl.co.uk/upload/pdf/JointNIST-NPLReproducibilityWorkshop_ DraftExecSummKeyRecommendations.pdf (accessed Feb 14, 2019).
[5] Data Fabric IG, https://www.rd-alliance.org/group/data-fabric-ig.html (accessed Feb 14, 2019).
[6] GO FAIR: a bottom-up international approach, https://www.go-fair.org/ (accessed Feb 14, 2019)
[7] Mark D. Wilkinson et al., Nature Scientific Data **3** (2016), Article number 160018.
[8] Research Data Alliance, https://rd-alliance.org/ (accessed Feb 14, 2019).
[9] Microscopy and Microanalysis Data Management, https://rd-alliance.org/microscopy-and-microanalysis-data-management-rda-11th-plenary-bof-meeting (accessed Feb 14, 2019).
[10] Left figure adapted from https://commons.wikimedia.org/wiki/User:OlivierMehani
[11] The Internet of FAIR Data & Services, https://www.go-fair.org/fields-of-action/internet-fair-data-services/ (accessed Feb 14, 2019).

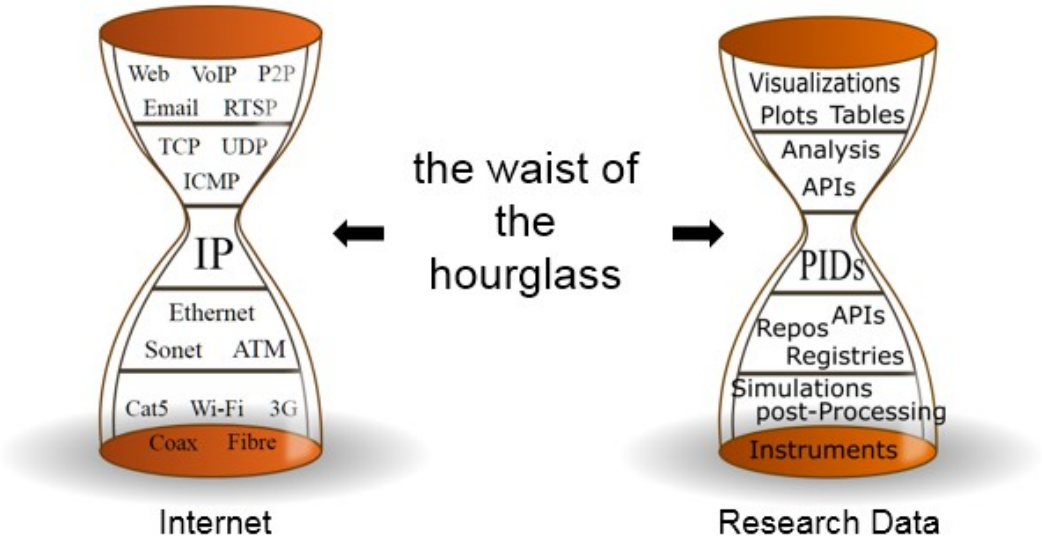


**Figure 1.** Comparison of the role of the Internet Protocol (IP) in the success of the Internet and the role of resolvable, persistent identifiers (PIDs) in the emerging research data fabric. In both cases, information flow in a distributed, federated architecture is enabled by the "waist of the hourglass" [10-11].

https://doi.org/10.1017/S1431927619007657 Published online by Cambridge University Press