Request for community partnership in data resource licensing planning

May 4, 2017

Dear NIH Scientific Data Council,

We write to initiate a dialog regarding NIH decisions on data use agreements and licenses. We are members of NIH-funded research groups that collect and/or integrate biomedical data from diverse sources for the purpose of advancing diagnosis, prognosis, treatment selection, and mechanistic discovery.

Summary:

- The current diversity of data-use agreements and licenses significantly hampers the ability to reuse and redistribute data in various informatics contexts.
- We believe that any mandatory data licensing policy must also include a plan for ensuring access, sustainability, and data quality.
- We request community partnership with NIH to develop common licensing and data reuse plans.

As part of the Monarch Initiative, the NCATS Data Translator, the Illuminating the Druggable Genome project (IDG KMC), the KnowEnG and other BD2K Centers, and NIH-funded projects, we are semantically integrating data from a wide variety of well-curated databases to create graphs linking diseases, phenotypes, genes, model organisms, medicines, publications, and other data from across the translational spectrum. Our goals are to accelerate discovery science, translation, and improvements in human health by integrating large amounts of diverse data, which is made possible by both the sharing of data assets and the use of appropriate data standards. We believe this is the future of biomedicine is in combining and using such data sets for integrative analyses. ***We also believe that achieving this future requires new innovative approaches to data sharing and licensing models, which will be especially critical for success of visionary programs such as the Cancer Moonshot, Precision Medicine Initiative, and the BRAIN Initiative.***

Our projects make data publicly available via: 1) open APIs (e.g. https://api.monarchinitiative.org/api/, that enables computational access to integrated human, model organism, and non-model organism databases); 2) Web portals (e.g., https://pharos.nih.gov/idg/index, that provides navigation of drug-gene-disease relations to identify new targets); 3) backend resources together with analysis tools, (e.g., Knowledge Network from UIUC); and 4) data exports in common formats (e.g. RDF, FHIR). For data sources that have restrictions on their use and sharing, we provide a view-only option. We provide strong attribution and provenance, such as license information, version history, etc. We recognize the hard work that it takes to create high quality, trustworthy resources and we encourage users to acknowledge the original data sources whenever possible, e.g. in their publications using standardized, persistent identifier strategies[1]. These principles have guided contributions to the greater community, including a response to the RFI on repository metrics[2], a rubric (essentially a maturity model) for evaluating open science projects[3], and we have written on the topic of open data sharing [4–6]. As part of all of these efforts, we have identified one of the ***most significant burdens in data science today: data licensing and use agreements.***

Complex licensing issues hinder most publicly-funded biomedical data from being put to their best use[5,7–10]. Such issues include missing licenses, non-standard licenses, as well as restrictive provisions. The sheer diversity of data-use agreements and licenses required by

different sources (many of which are NIH-funded and aim for openness) are particularly thorny for those that aim to redistribute data [8]. Two such examples include:

1. Currently, NCBI does not place restrictions on the use or distribution of the data contained therein. As a result, some submitters of the original data may still claim patent, copyright, or other intellectual property rights for all or a portion of the data and NCBI does not assess the validity of such claims. Since submitters do not transfer rights to NCBI, NCBI does not have the rights to transfer to a third party, and therefore cannot provide unrestricted permission concerning the use, copying, or distribution of the information contained in the molecular databases.
2. Data sources like STRING have different licenses or data use agreements based on the data set used, even down to the individual annotations in the case of LOVD databases: "The contents...are the intellectual property of the respective curator(s). Any unauthorised use, copying, storage or distribution of this material without written permission from the curator(s) will lead to copyright infringement with possible ensuing litigation."

Such conditions require users to directly contact each subsource to obtain permissions and repeat this process as new versions of data sources are released. We note that for individual investigators to navigate the licensing landscape and select an appropriate licensing mechanism it is both complex and outside their area of expertise in most cases. The legal interpretation and compliance of each source license/data use agreement has therefore become a significant scientific community burden and expense for many fields of science [11]. For example, one of our projects is located at numerous institutions; thus, to redistribute data from an integrated public data source requires a complex set of legal negotiations (Figure 1). We have therefore created a web portal (available soon at http://reusabledata.org) that facilitates community review and navigation of public database licensing and data use terms to make legal reuse and redistribution of data easier and less expensive for everyone[12]. The goal of the reusabledata platform and of this communique itself is to inform our collective thinking by examining the landscape.

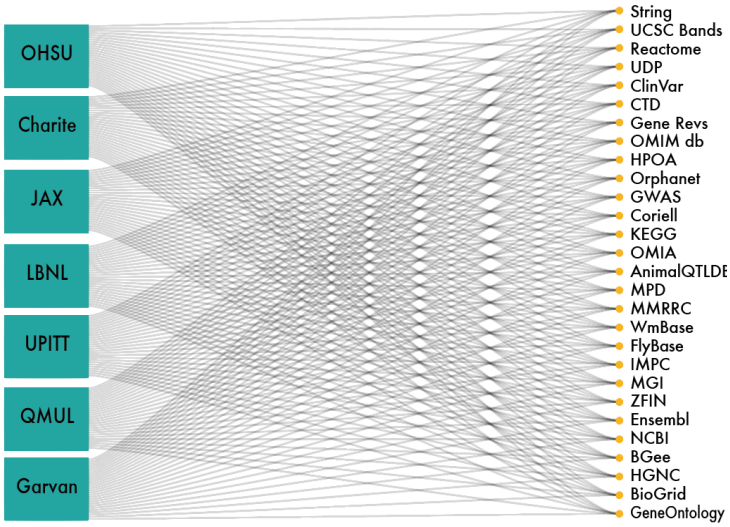

Figure 1. The burden of complex pairwise licensing agreements.

For each of seven institutions in the Monarch consortium (left), data licensing negotiation is needed for each of 28 data sources (right), for a total of 196 separate agreements required for a single project. This number is expected to increase as the number of sources and partners grows as the project matures. Moreover, even if these 196 agreements can be made, it is unlikely that they will all be identically worded, or even necessarily compatible.

Given these challenges we have faced, we support the NIH taking a greater role in standardizing and simplifying data licensing for NIH-funded projects. We also believe that any standardization must offer a plan for access and sustainability. The absence of such a plan may lead to serious unintended impact on data availability. Ultimately, a policy that does not address sustainability, quality assurance, or evaluation will be no better than a mandate that all cities be connected by asphalt with no tax dollars to build such roads and no signposts. Therefore, our goal is to realize better and more standard licensing models that a) make it trivial (e.g. requiring

no legal intervention) to reuse/redistribute data while b) also providing a clear path for entrepreneurship and commercialization in the public interest, including potential flow-through of revenue back to original data sources when presented in aggregate/integrated contexts. This is not dissimilar to how the open source software community functions[13–15]. Commercial or philanthropic revenue is especially important for the sustainability of underfunded or volunteer efforts (one example is the Human Phenotype Ontology[16] but there are many resources in a similar predicament). Numerous commercial entities have expressed enthusiasm for this model, but to date there has been no mechanism to broadly realize these goals. Other disciplines have grappled with similar problems and are addressing them using innovative models[17]. In some disciplines, it has been shown that ~80% of open data is not useful[18,19]. Further, there is indication that openness, in the absence of true utility, sometimes drives inequity rather than addressing it[18]. We believe that experts in these other open communities could be a vital source of inspiration and advice. Our ultimate goal is to effect change on the whole of the biomedical data landscape for public benefit, e.g. improved healthcare and biomedical discovery.

We respectfully request that the NIH initiate a dialog with key stakeholders -- including resource developers/data providers, application developers, open science advocates, data integration experts, repository managers, and economic, policy, ethics, and legal experts -- to develop one or a few common data licenses and an overarching licensing plan. We believe that standardization of data licenses for NIH-funded resources will greatly reduce the burden placed on each downstream user of any given database or data provider, because everyone will be working with the same terms. We recognize that this is no simple request. We take a holistic view regarding improving the whole data ecosystem (e.g. open science, quality informatics, sustainability, and public benefit[7]). Given the uncertainties about any specific perturbation, it may be appropriate to perform pilot projects to prospectively evaluate what will provision for the greatest ecosystem improvement. Thoughtful decisions on this matter will fundamentally affect science for decades to come and we thank you in advance for your consideration over this important matter.

Best regards,

Melissa Haendel (OHSU, Monarch Initiative, NCATS Data Translator, Force11, ISB, OBO Foundry)
Chris Mungall (LBNL, Monarch Initiative, NCATS Data Translator, Gene Ontology, OBO Foundry)
Andrew Su (Scripps Research Institute, NCATS Data Translator, Wikidata, BioThings)
Peter Robinson (JAX, Monarch Initiative, NCATS Data Translator)
Christopher Chute (JHU, NCATS Data Translator)
Russ B. Altman (Stanford University, Co-PI Pharmacogenomics Knowledgebase, PharmGKB.org)
Philip R.O. Payne (Washington University in St. Louis School of Medicine)
Mark Lawler (Queen's University Belfast, UK)
Tudor I. Oprea (UNM)
John Wilbanks (Sage Bionetworks)
Subha Srinivasan (UIUC, KnowEnG BD2K center)
Lawrence Hunter (U. Colorado)
Ida Sim (UCSF, Mobile Sensor Data-to-Knowledge BD2K Center, Open mHealth, Vivli)
Sean McDonald (Stanford University, FrontlineSMS)
Sean Mooney (Chief Research Information Officer, U Washington School of Medicine)
Damian Smedley (Queen Mary University of London, UK & Genomics England, UK)
Emma Ganley (PLOS)
Amye Kenall (Springer Nature)
Tim Clark (Harvard Medical School, Massachusetts General Hospital)
Carole Goble (The University of Manchester, UK, ELIXIR UK Head of Node, FAIRDOM)
Michel Dumontier (Maastricht University, NCATS Data Translator, CEDAR, BioThings)
Kristi Holmes (Northwestern University)

Mark Diekhans, University of California, Santa Cruz
Adrienne Zell, Oregon Health and Science University
Casey L. Overby (Johns Hopkins University, NCATS Data Translator)
Gustavo Glusman (Institute for Systems Biology, NCATS Data Translator)
Leigh Carmody (Jackson Laboratory for Genomic Medicine, Scientific Curator)
Guoqian Jiang, Mayo Clinic
Monica Munoz-Torres (LBNL, Apollo, Gene Ontology, International Society for Biocuration)
Maureen Hoatlin (OHSU, NCATS Data Translator)
Jeremy Goecks (OHSU, Galaxy Project)
Victor Jongeneel (University of Illinois and Swiss Institute of Bioinformatics)
Joshua Bittker (Broad Institute)
Jean-Philippe Gourdine (Monarch Initiative, Undiagnosed Diseases Network, OHSU)
Matthew H. Brush (OHSU, Monarch Initiative, NCATS Data Translator)
Richard L. Zhu (Johns Hopkins University, NCATS Data Translator)
Lara Mangravite (Sage Bionetworks)
Brett Tyler (Oregon State University Center for Genome Research and Biocomputing)
Mark D. WIlkinson (Ctr for Plant Biotechnology and Genomics, Universidad Politécnica de Madrid)
Michael R. Crusoe (Common Workflow Language project)
Raja Mazumder (George Washington University)
Nicholas P. Tatonetti (Columbia University, NCATS Data Translator)
Peter D'Eustachio (NYU School of Medicine, Co-PI Reactome Knowledgebase)
Nicole Vasilevsky (OHSU, Monarch Initiative, Force11, International Society for Biocuration)
Julie McMurry (OHSU, Monarch Initiative, NCATS Data Translator)
Robin Champieux (OHSU Library, Ontology Development Group)

Acronyms

IDG KMC: Illuminating the Druggable Genome Knowledge Management Center
BD2K: Big Data to Knowledge
API: Application programming interface
RDF: Resource Description Framework
FHIR: Fast Healthcare Interoperability Resources
RFI: Request For Information
LOVD: Leiden Open (source) Variation Database
HPO: Human Phenotype Ontology

References

1. McMurry J, Juty N, Blomberg N, Burdett A, Conlin T, Conte N, et al. Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data [Internet]. bioRxiv. 2017. p. 117812. doi:10.1101/117812 (in press, PLoS Biology).

2. NOT-OD-16-133: Request for Information (RFI): Metrics to Assess Value of Biomedical Digital Repositories [Internet]. [cited 24 Apr 2017]. Available: https://grants.nih.gov/grants/guide/notice-files/NOT-OD-16-133.html

3. Haendel M, Su A, McMurry J. Metrics To Assess Value Of Biomedical Digital Repositories: Response To Rfi Not-Od-16-133 [Internet]. Zenodo; 2016. doi:10.5281/zenodo.203295

4. Vasilevsky NA, Minnier J, Haendel MA, Champieux RE. Reproducible and reusable research: are journal data sharing policies meeting the mark? PeerJ. PeerJ Inc.; 2017;5: e3208. doi:10.7717/peerj.3208

5. Wilbanks J, Friend SH. First, design for data sharing. Nat Biotechnol. 2016;34: 377–379. doi:10.1038/nbt.3516

6. Siu LL, Lawler M, Haussler D, Knoppers BM, Lewin J, Vis DJ, et al. Facilitating a culture of responsible and effective sharing of cancer genome data. Nat Med. 2016;22: 464–471. doi:10.1038/nm.4089

7. Cancer Moonshot. Enhanced Data Sharing Working Group Recommendation: The Cancer Data Ecosystem [Internet]. Zenodo; 2016. doi:10.5281/zenodo.193064

8. Oxenham S. Legal confusion threatens to slow data science. Nature. 2016;536: 16–17. doi:10.1038/536016a

9. Gilbert N. Legal tussle delays launch of huge toxicity database. Nature News. 2016; doi:10.1038/nature.2016.19365

10. Balasegaram M, Kolb P, McKew J, Menon J, Olliaro P, Sablinski T, et al. An open source pharma roadmap. PLoS Med. 2017;14: e1002276. doi:10.1371/journal.pmed.1002276

11. Desmet P. Analyzing the licenses of all 11,000+ GBIF registered datasets [Internet]. NOVEMBER 22, 2013 [cited 20 Apr 2017]. Available: http://peterdesmet.com/posts/analyzing-gbif-data-licenses.html

12. How to License Research Data | Digital Curation Centre [Internet]. [cited 1 May 2017]. Available: http://www.dcc.ac.uk/resources/how-guides/license-research-data

13. Matt Jacobs |, Counsel G. Software Licensing Decisions: Consider Dual Licensing. In: Black Duck Software Blog [Internet]. [cited 28 Apr 2017]. Available: http://blog.blackducksoftware.com/software-licensing-decisions-consider-dual-licensing

14. Choosing an open-source licence | Software Sustainability Institute [Internet]. [cited 28 Apr 2017]. Available: https://www.software.ac.uk/resources/guides/adopting-open-source-licence

15. Top Ten Licensing Types – Where is the market going? In: The ITAM Review [Internet]. 8 Dec 2014 [cited 28 Apr 2017]. Available: https://www.itassetmanagement.net/2014/12/08/top-ten-licensing-types/

16. Köhler S, Vasilevsky N. The Human Phenotype Ontology in 2017. Nucleic Acids. Oxford Univ Press; 2016; Available: http://nar.oxfordjournals.org/content/early/2016/11/28/nar.gkw1039.abstract

17. Bishop T. Steve Ballmer launches USAFacts, using business principles for an unprecedented government report. In: GeekWire [Internet]. 18 Apr 2017 [cited 24 Apr 2017]. Available: http://www.geekwire.com/2017/steve-ballmer-launches-usafacts-using-business-principles-for-unprecedented-government-analysis/

18. McDonald S. The Open (Data) Market – Sean McDonald – Medium. In: Medium [Internet]. 1 Feb 2016 [cited 4 May 2017]. Available: https://medium.com/@McDapper/the-open-data-market-92f9557fd63d

19. Center for Open Data Enterprise [Internet]. [cited 4 May 2017]. Available: http://www.opendataenterprise.org/what-we-do