bioRxiv preprint doi: https://doi.org/10.1101/110825; this version posted February 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

22 Feb 2017

TOWARDS A GLOBAL SUPPORT OF CORE DATA RESOURCES FOR THE LIFE SCIENCES

W. Anderson[1], R. Apweiler[2], A. Bateman[2], G.A. Bauer[1], H. Berman[3], J.A. Blake[9], N. Blomberg[4], S.K. Burley[5], G. Cochrane[2], V. Di Francesco[6], T. Donohue[21], C. Durinx[10], A. Game[23], E. Green[6], T. Gojobori[14], P. Goodhand[15], A. Hamosh[16], H. Hermjakob[2], M. Kanehisa[22], R. Kiley[17], J. McEntyre[2], R. McKibbin[18], S. Miyano[19], B. Pauly[1], N. Perrimon[12], M.A. Ragan[13], G. Richards[1], Y-Y. Teo[20], M. Westerfield[11], E. Westhof[7], P.F. Lasko[8]

1: The International Human Frontier Science Program Organization, 12 Quai Saint-Jean, 67080 Strasbourg, France
2: European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, CB10 1SD, UK
3: Rutgers University, Center for Integrative Proteomics Research (CIPR), SAS - Chemistry & Chemical Biology, 174 Frelinghuysen Rd, Piscataway, NJ 08854-8076, USA
4: ELIXIR, Wellcome Genome Campus, Hinxton, CB10 1SD, UK
5: Rutgers University, Center for Integrative Proteomics Research, 174 Frelinghuysen Road, Piscataway, New Jersey 088jbl54-8076, USA
6: National Human Genome Research Institute, National Institutes of Health, 31 Center Dr., Bethesda, MD 20892-2152, USA
7: Institut de Biologie Moléculaire et Cellulaire, Université de Strasbourg, 15, rue Descartes, 67084 Strasbourg Cedex, France
8: Department of Biology, McGill University, Bellini Life Sciences Complex, 3649 Sir William Osler, Montreal, Quebec, Canada, H3G 1B1
9: Jackson Laboratory, 600 Main St., Bar Harbor, ME, USA
10: Swiss Institute of Bioinformatics, University of Lausanne, Bâtiment Génopode, 1015 Lausanne, Switzerland
11: Institute of Neuroscience, University of Oregon, Eugene, OR 97403-1254, USA
12: Department of Genetics, Harvard Medical School, 77 Avenue Louis Pasteur, Boston, MA 02115, USA
13: Institute for Molecular Bioscience, The University of Queensland, Brisbane 4069, Australia
14: Computational Bioscience Research Center, 4700 King Abdullah University of Science and Technology, Thuwal 23955-6900, Kingdom of Saudi Arabia
15: Global Alliance for Genomics and Health, MaRS Centre, 661 University Avenue, Suite 510, Toronto, Ontario, Canada, M5G 0A3
16: McKusick-Nathans Institute of Genetic Medicine (IGM), Johns Hopkins University, Blalock 1007 600 N. Wolfe St, Baltimore, MD 21287-4922, USA
17: Wellcome Library, 183 Euston Road, London NW1 2BE, UK
18: Biotechnology & Biological Sciences Research Council (BBSRC), Polaris House, North Star Avenue, Swindon, SN2 1UH, UK
19: Human Genome Center, the Institute of Medical Science, the University of Tokyo, 4-6-1 Shirokanedai, Minatoku, Tokyo 108-8639, Japan
20: Centre for Infectious Disease Epidemiology and Research (CIDER), Saw Swee Hock School of Public Health, National University of Singapore, 12 Science Drive 2, #10-01, Singapore 117549
21: Department of Bacteriology, Great Lakes Bioenergy Research Center, & Wisconsin Energy Institute, University of Wisconsin-Madison, 1552 University Avenue, Madison, WI 53726, USA
22: Institute for Chemical Research, Kyoto University, Uji, Kyoto 611-0011, Japan
23: 3 Poulton Crescent, Marlborough, SN8 1BH, UK

bioRxiv preprint doi: https://doi.org/10.1101/110825; this version posted February 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

On November 18-19, 2016, the Human Frontier Science Program Organization (HFSPO) hosted a meeting of senior managers of key data resources and leaders of several major funding organizations to discuss the challenges associated with sustaining biological and biomedical (i.e., life sciences) data resources and associated infrastructure. A strong consensus emerged from the group that core data resources for the life sciences should be supported through coordinated international efforts that better ensure long-term sustainability and that appropriately align funding with scientific impact. Ideally, funding for such data resources should allow for access at no charge, as is presently the usual (and preferred) mechanism. Designing and implementing a plan for long-term accessibility is the vision. Below, the rationale for this vision is described, and some important considerations for developing a new international funding model to support core data resources for the life sciences are presented.

Articulating the problem

The life sciences research enterprise relies extensively upon a set of core resources that archive, curate, integrate, analyse, and enable ready access to data, information, and knowledge generated worldwide by hundreds of thousands of researchers supported by hundreds of millions of dollars of annual research investment. Some such resources are public repositories of primary data (e.g., nucleic acid sequences and protein structures), while others are public knowledgebases that assemble and curate information and insights about a particular scientific domain, organism or the emergent properties of microbiomes (communities of microbial cells). Many of these core data resources developed from modest beginnings, in some cases with histories that span more than 50 years. Some began as printed books that were regularly updated, and then morphed into web resources as the Internet became better established in the 1990s.

Today, these web-based data resources are heavily accessed around the globe by researchers in academia and industry, students and clinicians, and the interested public. They are critical for ensuring the reproducibility and the integrity of research processes [1]. The ability to deposit to and download data from these resources freely and without restrictions facilitates progress in life sciences research. Significant loss of data from these resources, or introduction of barriers to data access could have devastating consequences for science, medicine, and wider society.

Core data resources are funded by a variety of mechanisms - mostly reflecting the history of how each developed over time. Some are funded by single sources and others by several sources; in almost all cases, the funding comes from national or non-profit granting agencies. The use of public funds to support this essential infrastructure ensures a strong return to society on public investments in research, and, furthermore, enables data to be reused, sometimes in unanticipated ways. However, the current funding model is fragile, with many of the data resources subject to vulnerabilities one associates with grant funding, such as changing

bioRxiv preprint doi: https://doi.org/10.1101/110825; this version posted February 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

priorities, processes, and policies. Of particular concern are relatively short funding cycles (e.g., 3-5 years), and the challenges encountered when grant applications for data resource infrastructures have to compete with research proposals.

In addition, many more areas of the world are research-intensive than was the case when these data resources were first developed decades ago. In some cases, these areas are associated with substantial technical expertise that could make important contributions to operating and improving these resources. Moreover, scientists in these geographies are members of the global research community and rely on these resources in the same way as scientists elsewhere. In this regard, all life scientists, irrespective of where they are based, are stakeholders in the sustainability of core data resources.

Defining core data resources

In order to design and implement an international plan for long-term sustainability, it is important to determine which data resources are of fundamental (i.e., core) importance to global life sciences research. This is a challenging undertaking given the scope, heterogeneity, and complexity of both the resources and the data they contain. For example, the online *Nucleic Acids Research* database catalogue lists around 1600 molecular biology data resources [2]. While some are no longer used or maintained, others have operated for decades and form a globally coordinated infrastructure that serves hundreds of thousands of researchers daily [3]. Operation of these long-standing resources requires a robust governance structure, active service management, and community-driven scientific development that are collectively well beyond the scope of a typical research program of an individual investigator. Some of these resources are connected to institutions committed to service provision [4, 5], while others have effectively navigated major management changes [e.g., transition of the Protein Data Bank (PDB) archive from Brookhaven National Laboratory to the Research Collaboratory for Structural Bioinformatics (RCSB) consortium after 27 years of operation [6]].

These long-standing data resources fall broadly into two categories:

Archival data repositories contain primary experimental data upon which many other databases are built. Typically, these repositories distribute data at no charge and without limitations on use, reflecting the widely held view that these fundamental data constitute a public good. Current best practices in the life sciences call for data producers to deposit primary data and metadata into such repositories prior to manuscript submission (or even sooner), with those data then made publicly accessible upon the manuscript's publication. Archival data repositories include the collection of nucleotide sequence data managed by INSCD, the International Nucleotide Sequence Database Collaboration [3] and the PDB [7], which contains information about the three-dimensional structures of biological macromolecules and is managed by the Worldwide Protein Data Bank (wwPDB) partnership. A more recently established example is the ProteomeXchange collaboration, which brings together four proteomics databases across the U.S., Europe, and Japan [8].

bioRxiv preprint doi: https://doi.org/10.1101/110825; this version posted February 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Knowledgebases add value to primary data by integrating information from multiple sources, often using computational approaches, and typically including expertly curated material. Some have a very broad scope, such as the Universal Protein Resource (UniProt [9]), which covers protein sequences and function, MetaCyc which contains extensive information on metabolic pathways and enzymes from organisms across all domains of life [doi: 10.1093/nar/gkv1164], KBase, a collaborative open environment for systems biology modeling of plants, microbes, microbial communities and microbiomes [doi: 10.1101/096354], and the Kyoto Encyclopedia of Genes and Genomes (KEGG), which focuses on genes and genomes [10]. More specialized knowledgebases, with deep integration of a particular domain, include the Online Mendelian Inheritance in Man (OMIM) database [11], the Arabidopsis Information Resource (TAIR), the *Escherichia coli* database (EcoCyc; doi: 10.1093/nar/gkw1003) and Model Organism Databases (MODs) such as the Mouse Genome Database (MGD), the Saccharomyces Genome Database (SGD), the Rat Genome Database (RGD), the online database of the genetics of *C. elegans* (WormBase), the online database for Drosophila genetics and molecular biology (FlyBase [21]), and the Zebrafish Information Network (ZFIN [12-16]). Note that the latter six knowledgebases plus the Gene Ontology Consortium (GOC [17]) recently formed the Alliance of Genome Resources (AGR) (http://www.alliancegenome.org).

The core data repositories and knowledgebases mentioned above are presented as possible examples, and are not intended as exclusionary.

Assessing life sciences data resources

In determining whether a life sciences data resource merits 'core' designation (and thus shared international support), we recommend the use of a broad set of well-defined and transparent indicators, such as those already being used by the European life science infrastructure ELIXIR [18]. These indicators are both quantitative and qualitative, with some mapping to the FAIR principles to make data Findable, Accessible, Interoperable, and Reusable [19, 20] and others measuring the impact of the resource on the scientific community and its role in accelerating science. Such indicators should also assess scientific focus and quality, the size of the research community served, the quality of the technical services provided, and the presence of a governance structure that supports open science.

While the set of data resources designated as 'core' should account for long-term and international requirements, such a portfolio must be dynamic so as to adapt to changing scientific needs. In this regard, the aforementioned indicators should be used in an ongoing fashion in managing the life cycle of all core data resources – from start-up through maturity and, when appropriate, to termination.

Determining costs and quantifying benefits

Having defined the appropriate set of core data resources for the life sciences, it will then become necessary to determine the fully burdened cost of operating each resource. In the case of archival data repositories, the replacement value of the primary data and metadata must be assessed, so as to establish whether long-term

bioRxiv preprint doi: https://doi.org/10.1101/110825; this version posted February 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

data storage is appropriate (*versus* future data regeneration, on an as needed basis). Furthermore, a reliable set of metrics for tracking the impact and cost/benefit balance of each core data resource, whether archival or knowledgebase, must be established. Finally, it will be essential to understand consequences of terminating the operation of a given resource. Addressing this final issue will require reliable quantitative and qualitative measures of the scientific, educational, and economic impact of each core data resource.

Towards a global solution for supporting core data resources

We propose an international coalition whose mission is to collectively support those core data resources deemed essential to the work of life science researchers, educators, and innovators worldwide. Through this coalition, funders of the life sciences should commit to the long-term shared responsibility to sustain the open access to core data resources because of their value to the global life science community and adhere to the oversight principles outlined above.

The new coalition we propose should be international and would include representatives of major life science research funders from most, ideally all, of the countries that are active in life science research. Initial efforts of this coalition would necessarily address some guiding questions including (1) what are the precise indicators that will be used for establishing a set of core data resources that will be eligible for shared international support; (2) will there be a binding and universal policy of global free access to the content of all designated core data resources that is appropriate and practical (as we recommend); and (3) what fraction of overall research funding from contributing nations should would be dedicated to supporting core data resources (note that informal estimates of 1.5-2% have been proposed, but a more accurate accounting is warranted going forward to guide the efforts of the new coalition).

In conclusion, we believe that it is time to reshape the approach for funding core data resources in the life sciences and we propose the launching of a coordinated, international effort to harness global expertise and to create a sustainable and egalitarian data infrastructure that will support scientific endeavors well into the future.

References

1. P.E. Bourne, J.R. Lorsch and E.D. Green, *Sustaining the big-data ecosystem*. Nature, Vol 527, S16-S17 (5 November 2015)
2. Rigden, D.J., Fernandez-Suarez, X.M., and Galperin, M.Y., *The 2016 database issue of Nucleic Acids Research and an updated molecular biology database collection.* Nucleic Acids Res, 2016. **44**(D1): p. D1-6. doi:10.1093/nar/gkv1356.
3. Cochrane, G., Karsch-Mizrachi, I., Takagi, T., and International Nucleotide Sequence Database Collaboration, *The International Nucleotide Sequence Database Collaboration.* Nucleic Acids Res, 2016. **44**(D1): p. D48-50. doi:10.1093/nar/gkv1323.

bioRxiv preprint doi: https://doi.org/10.1101/110825; this version posted February 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

4. NCBI Resource Coordinators, *Database resources of the National Center for Biotechnology Information.* Nucleic Acids Res, 2016. **44**(D1): p. D7-19. doi:10.1093/nar/gkv1290.
5. Cook, C.E., Bergman, M.T., Finn, R.D., Cochrane, G., Birney, E., and Apweiler, R., *The European Bioinformatics Institute in 2016: Data growth and integration.* Nucleic Acids Res, 2016. **44**(D1): p. D20-6. doi:10.1093/nar/gkv1352.
6. Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E., *The Protein Data Bank.* Nucleic Acids Res, 2000. **28**(1): p. 235-42. doi:10.1093/nar/28.1.235.
7. Berman, H., Henrick, K., Nakamura, H., and Markley, J.L., *The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data.* Nucleic Acids Res, 2007. **35**(Database issue): p. D301-3. doi:10.1093/nar/gkl971.
8. Vizcaino, J.A., Deutsch, E.W., Wang, R., Csordas, A., Reisinger, F., Rios, D., Dianes, J.A., Sun, Z., Farrah, T., Bandeira, N., Binz, P.A., Xenarios, I., Eisenacher, M., Mayer, G., Gatto, L., Campos, A., Chalkley, R.J., Kraus, H.J., Albar, J.P., Martinez-Bartolome, S., Apweiler, R., Omenn, G.S., Martens, L., Jones, A.R., and Hermjakob, H., *ProteomeXchange provides globally coordinated proteomics data submission and dissemination.* Nat Biotechnol, 2014. **32**(3): p. 223-6. doi:10.1038/nbt.2839.
9. The UniProt Consortium, *UniProt: the universal protein knowledgebase.* Nucleic Acids Res, 2016. doi:10.1093/nar/gkw1099.
10. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K., *KEGG: new perspectives on genomes, pathways, diseases and drugs.* Nucleic Acids Res, 2017. doi:10.1093/nar/gkw1092.
11. Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A., *OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders.* Nucleic Acids Res, 2015. **43**(Database issue): p. D789-98. doi:10.1093/nar/gku1205.
12. Blake, J.A., Bult, C.J., Eppig, J.T., Kadin, J.A., Richardson, J.E., and Mouse Genome Database, G., *The Mouse Genome Database: integration of and access to knowledge about the laboratory mouse.* Nucleic Acids Res, 2014. **42**(Database issue): p. D810-7. doi:10.1093/nar/gkt1225.
13. Engel, S.R. and Cherry, J.M., *The new modern era of yeast genomics: community sequencing and the resulting annotation of multiple Saccharomyces cerevisiae strains at the Saccharomyces Genome Database.* Database (Oxford), 2013. **2013**: p. bat012. doi:10.1093/database/bat012.
14. Laulederkind, S.J., Hayman, G.T., Wang, S.J., Smith, J.R., Lowry, T.F., Nigam, R., Petri, V., de Pons, J., Dwinell, M.R., Shimoyama, M., Munzenmaier, D.H., Worthey, E.A., and Jacob, H.J., *The Rat Genome Database 2013--data, tools and users.* Brief Bioinform, 2013. **14**(4): p. 520-6. doi:10.1093/bib/bbt007.
15. Harris, T.W., Baran, J., Bieri, T., Cabunoc, A., Chan, J., Chen, W.J., Davis, P., Done, J., Grove, C., Howe, K., Kishore, R., Lee, R., Li, Y., Muller, H.M., Nakamura, C., Ozersky, P., Paulini, M., Raciti, D., Schindelman, G., Tuli, M.A., Van Auken, K., Wang, D., Wang, X., Williams, G., Wong, J.D., Yook, K., Schedl, T., Hodgkin, J., Berriman, M., Kersey, P., Spieth, J., Stein, L., and Sternberg, P.W., *WormBase 2014: new views of curated biology.* Nucleic Acids Res, 2014. **42**(Database issue): p. D789-93. doi:10.1093/nar/gkt1063.
16. Howe, D.G., Bradford, Y.M., Eagle, A., Fashena, D., Frazer, K., Kalita, P., Mani, P., Martin, R., Moxon, S.T., Paddock, H., Pich, C., Ramachandran, S., Ruzicka, L., Schaper, K., Shao, X., Singer, A., Toro, S., Van Slyke, C., and Westerfield, M., *The Zebrafish Model Organism Database: new support for human disease models, mutation details, gene expression phenotypes and searching.* Nucleic Acids Res, 2016. doi:10.1093/nar/gkw1116.
17. Gene Ontology Consortium, *Gene Ontology Consortium: going forward.* Nucleic Acids Res, 2015. **43**(Database issue): p. D1049-56. doi:10.1093/nar/gku1179.
18. Durinx, C., McEntyre, J., Appel, R., Apweiler, R., Barlow, M., Blomberg, N., Cook, C., Gasteiger, E., Kim, J.H., Lopez, R., Redaschi, N., Stockinger, H., Teixeira, D., and Valencia, A., *Identifying ELIXIR Core Data Resources.* F1000Res, 2016. **5**. doi:10.12688/f1000research.9656.1.
19. *FAIR principles for data stewardship.* Nat Genet, 2016. **48**(4): p. 343. doi:10.1038/ng.3544.

bioRxiv preprint doi: https://doi.org/10.1101/110825; this version posted February 23, 2017. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

20. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., Bouwman, J., Brookes, A.J., Clark, T., Crosas, M., Dillo, I., Dumon, O., Edmunds, S., Evelo, C.T., Finkers, R., Gonzalez-Beltran, A., Gray, A.J., Groth, P., Goble, C., Grethe, J.S., Heringa, J., t Hoen, P.A., Hooft, R., Kuhn, T., Kok, R., Kok, J., Lusher, S.J., Martone, M.E., Mons, A., Packer, A.L., Persson, B., Rocca-Serra, P., Roos, M., van Schaik, R., Sansone, S.A., Schultes, E., Sengstag, T., Slater, T., Strawn, G., Swertz, M.A., Thompson, M., van der Lei, J., van Mulligen, E., Velterop, J., Waagmeester, A., Wittenburg, P., Wolstencroft, K., Zhao, J., and Mons, B., *The FAIR Guiding Principles for scientific data management and stewardship.* Sci Data, 2016. **3**: p. 160018. doi:10.1038/sdata.2016.18.
21. L. Sian Gramates, Steven J. Marygold, Gilberto dos Santos, Jose-Maria Urbano, Giulia Antonazzo, Beverley B. Matthews, Alix J. Rey, Christopher J. Tabone, Madeline A. Crosby, David B. Emmert, Kathleen Falls, Joshua L. Goodman, Yanhui Hu, Laura Ponting, Andrew J. Schroeder, Victor B. Strelets, Jim Thurmond, Pinglei Zhou, the FlyBase Consortium; FlyBase at 25: looking to the future. *Nucl Acids Res* 2016; 45 (D1): D663-D671. doi: 10.1093/nar/gkw1016