DeBarry, JD, et al. 2020. Practical Recommendations for Supporting a Systems Biology Cyberinfrastructure. *Data Science Journal,* 19: 24, pp. 1–12. DOI: https://doi.org/10.5334/dsj-2020-024

PRACTICE PAPER

Practical Recommendations for Supporting a Systems Biology Cyberinfrastructure

Jeremy D. DeBarry[1], Jessica C. Kissinger[1,2,3], Mustafa V. Nural[1,4], Suman B. Pakala[1], Jay C. Humphrey[1], Esmeralda V. S. Meyer[5], Regina Joice Cordy[5,6], Monica Cabrera-Mora[5,7], Elizabeth D. Trippe[1], Jacob B. Aguilar[8], Ebru Karpuzoglu[5], Yi H. Yan[1], Jessica A. Brady[9], Allison N. Hankus[5], Nicolas Lackman[1], Alan R. Gingle[1], Vishal Nayak[1,10], Alberto Moreno[5,11], Chester J. Joyner[5], Juan B. Gutierrez[1,8,12], Mary R. Galinski[5,11] and the MaHPIC Consortium[13]

[1] Institute of Bioinformatics, University of Georgia, Athens, US

[2] Department of Genetics, University of Georgia, Athens, US

[3] Center for Tropical and Emerging Global Diseases, University of Georgia, Athens, US

[4] Department of Computer Science, University of Georgia, Athens, US

[5] International Center for Malaria Research, Education and Development, Emory Vaccine Center, Yerkes National Primate Research Center, Emory University, Atlanta, US

[6] Environmental Health and Safety Office, Emory University, Atlanta, US

[7] Department of Biology, Wake Forest University, Winston-Salem, US

[8] Department of Mathematics, University of Georgia, Athens, US

[9] School of Chemical, Materials, and Biomedical Engineering, University of GA, Athens, US

[10] CSRA Inc., 2 Corporate Blvd Suite 100, Atlanta, US

[11] Division of Infectious Diseases, Department of Medicine, Emory University, Atlanta, US

[12] Department of Mathematics, University of Texas, San Antonio, US

[13] Membership of the MaHPIC Consortium is provided in the Acknowledgments

Corresponding author: Jessica C. Kissinger (jkissing@uga.edu)

Projects in the life sciences continue to increase in complexity as they scale to answer deeper and more diverse questions. They employ technologies that generate increasingly large 'omic' datasets and research teams regularly include experts ranging from animal care technicians, veterinarians, human health clinicians, geneticists, immunologists, and biochemists to computer scientists, mathematical modelers, and data scientists, often located at different institutions. Providing the cyberinfrastructure support framework (IT, data management, communication, documentation, and aspects of project management related to these areas) for these projects requires a diverse set of technical tools and soft skills. These skills must be able to meet both the broad needs of data generators and consumers within the project and the needs of the larger scientific community. Here we describe recommendations for cyberinfrastructure support teams responsible for systems biology research programs. Recommendations are based on lessons learned while establishing and leading a complex, transdisciplinary, host-pathogen malaria systems biology consortium involving many institutions, a variety of disciplines, animal infectious disease models, and clinical studies. While some technical suggestions are included, the primary foci are situational and sociological challenges and tips for handling them.

Keywords: Transdisciplinary; Cyberinfrastructure; Team science; Team communication; Systems biology; MaHPIC

Introduction

Creating and supporting the cyberinfrastructure (IT, data management, communication, documentation, and related project management needs) for a systems biology project is as much about facilitating communication and education as technology, data management, and data integration. Systems biology is, by nature, 'transdisciplinary', bringing together diverse teams of researchers with equally rich expertise. This can create a 'Tower of Babel' scenario where researchers don't speak exactly the same (Zook, et al., 2017) technical language. What is a 'gene'? It may translate as a sequence, a function, a protein, or be taken for jargon. What is a 'database'? Is it an Excel spreadsheet, a file archive, or a relational architecture? Agreement to be the 'informatics' support team for such a project is challenging technologically and socially. Because infrastructure and analysis technologies change rapidly and are often project specific, we focus on situational and sociological 'Tips' to help a cyberinfrastructure support team and project leadership. Sociological tools and strategies are helpful to educate key project members on the critical importance of a cyberinfrastructure, especially when many other competing vital tasks must be completed within the project. The data management and cyberinfrastructure team (DMCIT) should generate a 'Rosetta Stone' to guide the effort and provide a unifying framework for data sharing and data provenance, engaging all project members from the start.

These Tips are organized around project phases (**Figure 1** and Tip 2). The first four concern effort, policies, and best practices to be defined early with guidance from the DMCIT and input from the broader project. Tip 5 is longer and organized into sub-sections that describe the roles of team members, and the main services

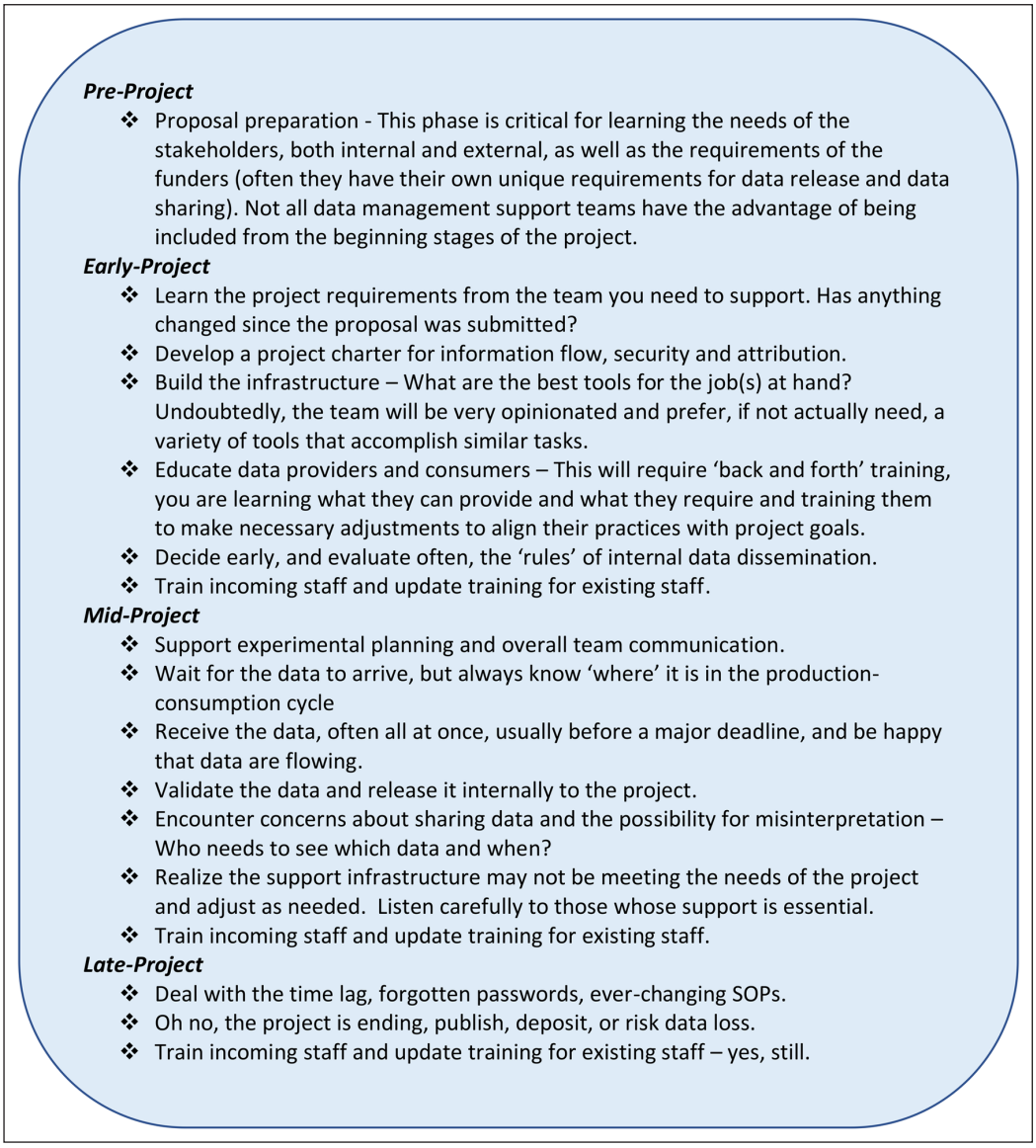

Figure 1: The phases of complex systems projects.

and resources the team will provide on a day-to-day basis. The last four Tips focus on activities that become more relevant as the project matures, though they should be planned for as early as possible.

This manuscript results from experience. In this case, a five-year National Institute of Allergy and Infectious Diseases (NIAID/NIH) contract awarded in 2012 to a large, multi-institutional, transdisciplinary group of researchers to establish a systems biology program to study host-pathogen interactions and the development of pathogenesis in malaria; the Malaria Host-Pathogen Interaction Center – MaHPIC (MaHPIC, 2019). The project involved multiple *Plasmodium* parasite species, their mosquito vectors, and nonhuman primate (NHP) hosts. Multiple data types (genomics, transcriptomics, proteomics, metabolomics, lipidomics, immunomics, clinical and diet) were generated in longitudinal infection experiments spanning 5 years. Extensive data management and bioinformatics processing facilitated data analysis and subsequent mathematical modeling. The MaHPIC was designed from inception to have its data available to, and re-usable by, the larger malaria and systems biology communities, concepts now known as 'FAIR' (Findable, Accessible, Interoperable, Reusable) (Wilkinson, et al., 2016). To this end, it catalogued extensive metadata, including experimental logistics, standard operating protocols (SOPs), equipment details, animal activities, and veterinary clinical care, as well as clinical and demographic data involving humans in malaria-endemic countries. At its peak, 26 terabytes of data residing in nearly one million files generated in a dozen project performance sites needed to be managed, including institutional project approvals and partnering agreements. The DMCIT ensured secure data transfer within and across institutions, internal data validation and analyses and ultimately data publication and data deposition via public repositories and websites. **Figure 2** shows a generalized view of key cyberinfrastructure components based on MaHPIC's experience, and the possible roles and responsibilities of the DMCIT. **Figure 2** is intended as a guide to readers as they consider how they will relate to and serve their own systems biology project.

It is difficult to overstate the consistent support the MaHPIC DMCIT received from project leadership and team members. Of the challenges encountered, many were technical, but many resulted from the ever-present need to prioritize time, effort, and other resources to meet experimental goals balanced with meeting the needs described here. The recognition by all that scientific success was often as dependent on a sound cyberinfrastructure as on sound experimental design and data analyses was key in our overall success.

Tip 1: Define your user community both inside and outside of the project

An assessment of the user community should begin during proposal preparation and continue throughout. You are agreeing to nothing short of creating, adapting, and potentially scaling a set of 'end-to-end' solutions that will require constant support. While the initial focus will be 'inward', this will soon shift externally as data and other deliverables are released. Decisions made early in project planning will impact the ease of release and the ultimate long-term value of project data to the larger community.

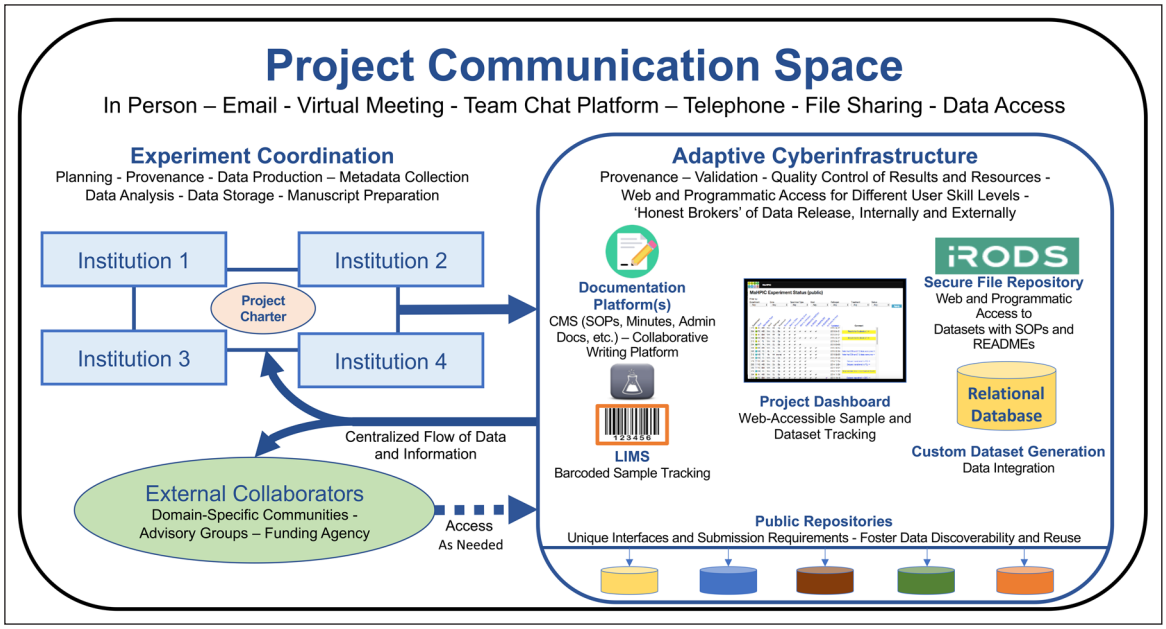

Figure 2: Key components of a systems biology cyberinfrastructure.

It is imperative to determine data formats, construct data dictionaries, and calculate the volume of data that will be generated and consumed during each project phase (**Figure 1**) and to know how these needs will vary within the project. Data producers, often wet lab scientists, may have different needs than data consumers, often mathematical modelers, data scientists, or researchers outside the project's experimental team. It is important to determine what metadata and standard operating procedures (SOPs) will need to be collected and shared within the project and with the community to make them FAIR.

Consider aligning preparations for both internal and external data consumption so that duplication of effort is minimized. Early familiarity with the submission requirements for the archival repositories that will eventually be utilized can avoid last-minute scrambles for details. It is critical that you know who you are serving, how they view the project and its goals, and what they hope to gain from the data. In the case of the MaHPIC, experts from different disciplines might seek knowledge helpful for vaccine development, drug development, treatment of severe illness, or blocking the transmission of infectious parasites to mosquitoes. MaHPIC was fortunate to contain many perspectives on data needs and utilization. Many needs were anticipated, and data and metadata collected accordingly, but there were challenges (Tips 8 and 9).

Tip 2: Understand the phases of the project

Systems biology projects, like all large projects, have distinct phases, (**Figure 1**). From the perspectives of both data and sample management (generation, distribution, use, and storage of both) and team communication (project portals, document sharing, file archives, listservs, online meeting hosting, and/or a real-time experimental dashboard), it is important to be prepared and adaptable. The DMCIT must participate in experimental planning, analysis and publication discussions. Otherwise, unexpected 'new' datasets, data production delays, SOP alterations, or the need for accession numbers or data digital object identifiers (D.O.I's) for a dataset to cite in a paper, can catch the DMCIT by surprise, affect progress, and cause anxiety. The MaHPIC had a leadership and research team that was understanding of the project's data needs and ensured the DMCIT had both an adequate budget and a 'seat at the table' for project-wide discussions and decisions.

Tip 3: Determine what to collect – both data and metadata are essential

It is imperative to invest time to understand the entire project, its members and experiments and sample and data flow. Information related to experimental planning, data production, and data analysis must be captured for the data to be fully interpretable and reusable, both internally and externally. Some may fear that ephemera are collected at the expense of precious project resources, but changes in reagents, equipment, personnel, animal caging and diet, versions of software programs, SOPs, and updates in reference genome sequences may prove essential for data interpretation and re-use. Ideally collection can be optimized and automated to minimize drain on resources and time. Leadership should consider the limits of prescience in making decisions about what will, and will not, be collected, because this is a case of potential 'unknown unknowns'. It is important to consult experts within and beyond the project, especially those with experience integrating data like yours. Gather their dos and don'ts and incorporate them into your decision making. As an example from the MaHPIC, information on the daily routine care of animals (*e.g.*, technician access, diet, medications), and the timing of such, provided understanding of possible physiological or metabolic observations. Leadership's strategic decision to support collecting this data and the team's support in overcoming the technical hurdles to do so, repeatedly benefited downstream analyses. 'Ephemera' collection and integration eventually became standard practice and part of new staff training.

But, How do you discover and account for as many sources of data and metadata as possible? It is wise to adopt a motto of 'Investigate everywhere, trust nothing, test everything'. An intimate understanding of sample and data production is critical for determination of useful metadata and an understanding of how, and where, they are generated. While some may initially view this as an unnecessary intrusion, if you leverage these activities to promote the automated collection of data and metadata where possible, ideally directly from instrumentation, reducing the burden on data producers, most will soon recognize and embrace the benefits (Tip 5). For example, many instruments automatically create result files and several instruments print a metadata summary. These data streams should be automatically captured and saved if possible. Many laboratory activities, even with high-tech omics instrumentation, have significant manual components that require generating sample barcodes, downloading data, re-formatting or editing certain features and then saving them, often for subsequent processing. Each step takes time and provides an opportunity for error. Tasks that were not burdensome when performed once a week become highly burdensome when scaled for months on end. Time spent automating routine tasks reduces errors, ensures provenance (Kazic, 2015), and frees up resources for tasks that infrastructure cannot solve.

Tip 4: Manage data stewardship, inside and outside the project

The DMCIT may need to train the project in best practices for data stewardship. It is vital to instill and foster a sense of scientific duty related to making data FAIR (Tips 8 and 9). Successful efforts will promote efficient production, validation, and analysis of data. Engagement begins with full immersion in the data generation process, including documentation of the provenance of all samples, files, results, and analyses, as well as the gathering of metadata and SOPs. You may be tasked with quality control of results to ensure the proper format, content, and integrity of results for downstream consumption. Maintaining provenance is essential. Validation of results places the DMCIT on the front lines of ensuring the integrity of the data and potentially as the last line of defense if you are also tasked with deposition to public repositories. The reputation of your project and colleagues will be in your hands. It is critical to define and set 'ground rules' for the flow of project information and results and then provide reliable user-friendly ways for the project to thrive under that paradigm, preserving provenance. This process includes locating and integrating 'back channels' of data and information flow, where members eager to analyze data may be exchanging unvalidated results outside of the agreed-upon framework (*i.e.*, familiar channels that work well in small-scale projects, like shared online folders and email), as well as identification of 'hidden' data producers. For example, technicians, animal handlers, custodians, and 3rd party collaborators or service providers are all valuable sources of project information that ought to be captured. It is important to know when instruments were calibrated or software programs updated. If these points seem overreaching, consider an early 'signal' detected in the MaHPIC project data that was traced back to an instrument repair as opposed to biology, underscoring the importance of recording such analytical metadata. The eventual resolution of this signal reinforced the need to collect and understand the impact of these types of information. In another scenario, two 'versions' of a dataset were being analyzed (same results, but different formal sample naming/IDs) resulting from one group utilizing the official validated project repository and one group utilizing 'back channels.' Such well-intentioned efforts could prove costly, requiring time and resources to identify the problem and ensure that only validated data is being analyzed within and ultimately outside the project.

Investigators may move beyond the original scope of a project or circumvent SOPs in the interest of short-term scientific progress. Sometimes this is needed (Tip 10). However, special exceptions can cause disruptions if hidden. From that point forward a data set is 'outside' of the project tracking and becomes unknown to the official data flow monitoring system. If later this data set needs to be shared officially, provenance is often hard to document, and public deposition can be delayed.

Project data are valuable and must be handled with the utmost care. Attention must be paid to data restrictions imposed by an Institutional Review Board (IRB), patient consent, or collaborator agreements (Zook, et al., 2017). Policies regarding the release of project data, internally and externally are needed. Ideally, a DMCIT will assume the role of 'honest broker', releasing data internally to the project and externally to the community. This requires policies aligned with funding requirements to be created, agreed to, and followed by all. Data sharing, with trusted provenance and high standards of accuracy, like the results they produce, will be recognized and rewarded through re-use and data citations. New programs should consider creating a project charter and set of data sharing principles (Tip 6). In the MaHPIC, group leadership decisions made early in the project were faithfully implemented, and extra attention was then needed to best ensure they were enforced in the face of any transitions of people joining or leaving the project. The DMCIT quickly recognized the need for formal rules and leadership supported the implementation and training.

Tip 5: Build an appropriate infrastructure, it matters

'Cyberinfrastructure' has been generally defined and often invoked, but what is a DMCIT going to actually build? Preceding tips focused on how the DMCIT and a broader project might interact and what each may need. But what are the specific resources that will be needed from the DMCIT?

People

A DMCIT needs the 'right' people. It will need programmers and database administrators, perhaps software engineers. It will need members who can understand the biology and scientific project goals and who can liaise between those building the cyberinfrastructure and those using it. Bioinformaticians, adept at the keyboard and bench, may be available. If not, recruit staff to ensure that all required skills are present and promote project-wide training to fill gaps. If the project does not have an official project manager the DMCIT may need someone with experience to organize and oversee communication and related needs.

The DMCIT should be known for two things: 1) coordination of policies and SOPs to facilitate communication and the flow of data between institutions and larger community and 2) the cyberinfrastructure and tools they provide to implement those policies and needs (**Figure 2**). Consider your own experience with data sharing, collaborative writing, virtual meetings, and documentation resources. How likely are you to revisit a tool that fails or is difficult to use the first time? The second time? You must provide the tools to meet project needs and this requires that you understand those needs. The broader team is likely to be more interested in the proposed 'science' and potential discoveries and less in the data management and cyberinfrastructure that will affect everyone. Additionally, project researchers will come from varied backgrounds with different levels of comfort using the tools and technologies needed, (Boland, et al., 2017), necessitating multiple tools/platforms to accomplish similar tasks.

An effective systems biology cyberinfrastructure support team will: Plan the flow of information including communication and documentation, identify hardware and software components and the needs of all project members, and work with all project stakeholders to create a project charter that will enable attainment of goals. The required components will be complex and require careful integration to support effective communication and the secure movement of data and information. Needs will change as the project matures and the infrastructure should adapt accordingly.

Hardware, cyber connections, and security

Needs will vary but servers and storage will likely be needed. Hardware may need to be redundant to provide enhanced data security (FIPS) and (FISMA) and guard against data analysis and access downtime. Infrastructure to support relational databases as well as fast and slow data storage may be needed. Periodic vulnerability scans and system adjustments to address emerging security issues may be required.

Secure data access agreements between institutions to support encrypted data transfer may be required. Persons involved in data transfer within and between institutions should coordinate to understand these needs early. Within the MaHPIC, human data were de-identified, transferred securely and released in accordance with IRB protocols. Huge files (up to 700 GB) were moved between institutions and daily 'dumps' of a Laboratory Information Management System (LIMS) were transferred, formatted, and provided to animal technicians and veterinarians and compiled in a project database and dashboard. Establishing required agreements takes time and often includes specific hardware and software specifications (dedicated drives or partitions, university port configurations, *etc.*).

Software

Use existing software for data management, code versioning, data analysis, communications, *etc.*, (Taschuk and Wilson, 2017) to avoid 're-inventing the wheel'. This will allow you to innovate when needed. When innovation is needed, determine project requirements and build a prototype first, allowing the intended users to review before progressing too far into development. The earlier the DMCIT provides value the sooner they will have the opportunity to innovate.

Within the MaHPIC, software support needs ranged from collaborative writing, team chat, online meetings, data storage and retrieval (including text documents like Institutional Animal Care and Use Committee and IRB approvals and funder progress reports), to SOP documentation and code. Collaborative writing and data storage within subgroups were without specific policies other than ensuring that data security and back-up plans were in place. Larger team writing efforts and data storage were restricted to agreed platforms that could accommodate large numbers of simultaneous writers.

If possible, leverage existing life science cyberinfrastructures. Several mature cloud-based platforms for data storage, management, and even analysis are available. Amazon Web Services is one. CyVerse (Merchant, et al., 2016) is a National Science Foundation-funded national cyberinfrastructure project with freely available resources and training from web-based data management and analysis, to fully customizable APIs. SAGE bionetworks provides an open science platform including Synapse to help researchers collaboratively aggregate, organize and share their data.

File formats and data storage

The nearly universal data and metadata format within the MaHPIC was the tab-delimited file. Notable exceptions included RNA-Seq and proteomics files, flow cytometry, and some blood chemistry. Templates were created for each data type, versioned, and linked to SOPs and data dictionaries. All data files for were validated and shared internally via an iRODS (iRODS-Consortium, 2019) file repository. This open source solution was chosen because it provided data access via dozens of mechanisms, thus suiting a diverse user audience.

Sample tracking

Sample tracking can be challenging, especially for a large multi-institutional project like the MaHPIC. A LIMS will generate a barcode and record sample creation and its relationship to an experimental subject and condition. A barcode reader can log a sample's movements in and out of an institution or laboratory. However, unless all players are linked to the LIMS these movements are only known to the owner of the scanner until the DMCIT is informed of sample movement through other means. This system worked fine for the MaHPIC from the perspective of the data, but it was less than ideal for live status updates with respect to the physical movement of samples and their associated data through the project. The MaHPIC utilized a custom project dashboard to meet this need but this did not provide multi-institutional sample tracking in real time.

Automation

It is critical to automate the flow of data or redundant time-consuming activities wherever possible. Automation (checked often for correct functionality) facilitates focus elsewhere. Work with data producers to eliminate manual data entry wherever possible. The MaHPIC automated transfer of clinical data from an analysis instrument to computer, saving hours of technician time previously used for manual double data entry (error checking). Carefully consider all data being collected and how you can leverage institutional resources to automate.

User support

Make sure you are aware of the potential 'secondary skills' of project members to facilitate infrastructure set up and troubleshooting. This will also help you with an often-hidden trap of team science: you will be responsible for providing the same services and information to all conceivable computational skill levels, ranging from those comfortable at the command line to those who rely solely on a web browser. Note that computational savvy does not correlate with scientific prowess, significant data producers and consumers may need the most training and vice versa. You can identify the training needed early and help ensure the DMCIT will be contacted by users when needed.

Tip 6: Reduce jargon, train, educate, refresh … repeat

Policies and effective cyberinfrastructure use require training, refreshers, reminders, and repeat performances for new project members. Consider the early creation of a project charter, including rules for communication, data-sharing guidelines, and specific goals, created with the entire project's input and have everyone symbolically sign to signify agreement. When onboarding new staff, add their signatures and take advantage of the process to reinforce policies. Foster a universal awareness of the need to evaluate whether everyone is 'speaking the same language' about tasks and goals. Consider monthly 'all-hands' meetings to reinforce the policies and decisions and ensure all are informed and all have their voices heard on data management and other topics across the project. Attendance, in person or remotely, could be an important part of the charter. The MaHPIC established data sharing rules (Tip 4) and other mechanisms to address potential issues (*e.g.,* onboarding SOP described below). The creation of a charter, including guidance similar to these Tips, could potentially save significant time and effort and support the smooth running of large programs.

Coordinate with leadership to create and regularly update an 'onboarding' SOP for all members that covers the contents of the charter, security requirements, and specific survival strategies within the project. What resources are available? Where do you retrieve data, or meeting schedules? Are there shared calendars and listservs to join? As you train newcomers, you will be forced to reevaluate project policies and jargon. If training, education, and refreshers are part of your project culture you will be better able to respond to any staff turnover and be more resilient to perturbations in project goals.

Tip 7: Promote a culture of effective documentation

Documentation will affect every aspect of a project. Who attended that meeting last month? What were the outcomes? Who was tasked with that critical 'thing'? Where can you find an SOP for transcriptomic data production for 'Experiment Alpha' from two years ago? Often critical elements of a single lab's knowledge exists in the 'heads' of its members. Protocols and results may be penned in lab notebooks, with key project-specific details and nuances, and these notes must be identifiable and accessible, an unenviable task if they are not electronically searchable and organized.

Consider what documentation you need, what form it will take, how it will be accessed and by whom, and how often it will be updated. Documentation will mean different things at different stages of the project (**Figure 1**). It will range from data production and analysis SOPs, meeting minutes, data sharing

guidelines, dataset READMEs that detail nuances of the files, team rosters and contact information, MOUs, MTAs, *etc.* You will need a systematic approach to manage your team's documentation infrastructure and it must be accessible across institutions and by team members with different skillsets. This is a vital part of a cyberinfrastructure.

Do not rely on email for document storage or a record of project activities. Consider a content management system (CMS) for storage, and a plan for document organization. Access to some files may be restricted to your leadership and administrators, so permissions will be an issue. For SOPs consider a 'wiki' approach but discuss who will be able to create and edit the content. Select a collaborative writing platform that will provide your team with the flexibility to simultaneously edit documents. You will need to budget for these services, which means planning for them before experiments begin. The ease of usability of your documentation platform(s) will directly affect if and how they are used, and this can be a major factor in team efficiency and morale.

Tip 8: Integrate the data to facilitate use, discovery, and public deposition

Individual datasets will likely become part of larger analyses. Within the MaHPIC, a discrete dataset was usually a specific data type (RNA-seq, immune profiles, metabolomics, *etc.*) from a specific experiment designed to investigate specific hypotheses. Each dataset was useful on its own, but when combined for mathematical modeling at the experimental level, including leveraging clinical metadata they provided an unprecedented level of molecular detail about host-pathogen interactions resulting from *Plasmodium* infections (Cordy, et al., 2019, Gardinassi, et al., 2018, Joyner, et al., 2019, Tang, et al., 2018, Tang, et al., 2017). This was possible because of the integration of the datasets within and across experiments made possible by FAIR data practices (Tips 4 and 9). Integration can have varied definitions. It may mean the formatting, organization, and storage of results to facilitate their analysis in parallel. Simply designing folder and file names to allow humans and algorithms to identify where they 'fit' in the grand scheme is a good start. Structured README files and supplementary documents that detail accessory facts about datasets, which versions of software and reference genomes were used, what other datasets are available or being produced, *etc.* will also be helpful. Organized and annotated datasets will facilitate their analysis and rudimentary automated processing of their contents.

Integration can also take the form of a relational database. How data are processed and stored for later consumption will be critical. If similar data types are to be compared across experiments, it will be important to know if normalization, standardization, batch effect correction, *etc.* were conducted to allow 'apples to apples' comparisons. Some consumers will need 'raw' results, preferring to process data themselves. Data producers and consumers must know and agree what 'raw' means. Provenance, validation, and metadata will play key roles. For example, if comparing transcript abundances, protein levels, and metabolites between samples, it is critical that electronic results are associated with the physical samples from which they were derived and that there is sufficient information to distinguish technological artifact from biological reality. The MaHPIC supported data integration through a novel database schema entitled Scientific Knowledge from Data, SKED (Trippe, et al., 2017a), which uses 'data primitives' to enable the programmatic creation of custom datasets and analyses that can be applied over diverse data types (Trippe, et al., 2017b).

Unless you are reading this in the future when data integration is a solved issue, you will have additional challenges and opportunities to consider as you prepare to publicly deposit project data. The integrated datasets you produced will likely have to be separated, like slices of a pie, and deposited in repositories that specialize in specific data types (if they exist). For example, NCBI's SRA (Leinonen, et al., 2011) for RNA-seq, EBI's MetaboLights (Haug, et al., 2013) for Mass Spec, ImmPort (Bhattacharya, et al., 2014) for immunology data, *etc.* Repositories have unique criteria. Some require only minimal free-text metadata forms and will simply host files. Others have optional batch submissions, and programmatic transfer of data and metadata, and require integration of metadata into domain-specific controlled vocabularies and ontologies. READMEs, supporting documentation, and carefully planned and collected metadata are all important for data deposition and re-use.

Many journals require data be public before publication, or even manuscript review. Understanding repository requirements early, and agreement from your funder as to what data will be made public, when, and where, is important. Repositories do not exist for all data types, *e.g.*, MaHPIC's animal clinical data. It may be difficult for future consumers to recreate the larger experimental picture, *i.e.*, the 'whole pie' that produced a 'slice'. Consider providing a persistent website with a list of all public datasets and experimental descriptions, associated publications, and explanations of your project and its products. See MaHPIC's manifest of publicly available datasets (PlasmoDB-MaHPIC).

Ontologies are important for public data deposition and integration, but availability and comprehensiveness vary. In genomics, they are mature, with years of community input, and include a core set of terms to describe a 'study', 'experiment', 'gene', *etc.* Other fields are less developed, the rules are few, and the guideposts shift as communities discuss, test, and refine core concepts. You may have to accommodate needs across this spectrum. This can be a challenge, but can provide an opportunity to take a leadership role, in partnership with ontologists and your data producers, shaping minimum data and metadata standards within a community. Investment up front will pay dividends in the future as the data can be more easily identified, reused, and integrated (Kazic, 2015).

Tip 9: Make a strong and sustained debut by publishing and depositing project data

Project datasets will outlive the project. Sustained presentation of project results from the perspective of the 'whole pie' will foster their discoverability and reuse. Building on Tip 8, make sure you know your funder's repository requirements, work to understand repository standards early, and proactively contact repository staff with questions. Repository staff may be receptive to an alignment of goals. Those supporting emerging data types may want to increase their visibility and you may be able to help. They may be willing to assist your project with training in metadata collection, dataset organization, and the submission process.

Information security restrictions may preclude public sharing of all project results at the same time. Repository availability and requirements may limit what can be shared and in what form. Consider releasing raw instrument output and not just processed files so data can be reanalyzed with new methods in the future. Decide early, and include in your charter, which data will be made public, accounting for requirements for how specific data must be stored as well as what can be accessed and shared within your infrastructure, in papers, and in repositories (Zook, et al., 2017). Constantly work with your team to ensure that all materials and outputs are tracked and released, and that data, SOPs, *etc.*, are internally approved and do not contain sensitive or proprietary information.

Tip 10: Be flexible and grow with the project

Likely you will need to adapt to emerging challenges and opportunities. Staff changes, new technologies, and discoveries may necessitate swift changes. An adage recently expressed by funding officials to describe MaHPIC research and its complexity is, "*Build the car while you are driving it.*"; *i.e.*, keep moving with confidence at a fast pace despite lacking all pieces of the research puzzle. This mindset has been particularly useful for the MaHPIC, as a large fast-paced demanding systems biology consortium. Momentum has been critical, as well as delivering timely results and meeting project milestones, plus providing publicly available data for outside analysis. Build your project culture, communication, documentation, data flow, analysis, *etc.* with these considerations in mind.

The MaHPIC DMCIT was well-supported by the consortium's leadership and anticipated many challenges and opportunities. Still, in hindsight, what would we do differently, or better? Naturally, each new endeavor will bring new unknowns. For sure, we would aim to meet with each data producer early on and ask them to fully describe their experimental and analysis pipelines to ascertain every point where automation could make their life easier and reduce errors (Tip 3). As an example, during such discussions with MaHPIC members, we realized that some investigators were scaling data production and analyses using standardized methods developed for much smaller projects. Specifically, one group was using an instrument that only produced a PDF data file. The data from printed PDFs were being manually entered into spread sheets. The DMCIT wrote a custom program to parse the printstream from the application and populate the database, removing considerable manual effort and preventing potential error. As 'creatures of habit', if left alone, we can remain unaware of potential solutions with potentially transformative effects. Exploring ways to increase efficiency, reduce error, and increase provenance is clearly useful early and often. Time spent assessing and transforming manual data steps may reward an entire project by reducing errors, increasing provenance, and creating more time for the science.

Finally, the time lapse between the planning and execution of a wet-lab experiment, the generation of *in silico* results, the writing of a paper and dissemination of the data, can be long. From the beginning, think about and enable the future! In an idealized scenario, project data would be curated in perpetuity and all SOPs, READMEs, and project metadata would be machine accessible, semantically annotated and self-describing to maximize their future utility. The potential exists even though this is not the current reality. It is a challenge to even describe a full experiment in such a way that all its components and documentation can be found in data archives that are restricted by data type. Also, documentation and tracking will be

critical to be sure all data has been secured and remain beneficial for the duration of the project, and possibly beyond. Forensic validation of results, *e.g.*, poring over emails or written records is not how you want to expend precious project resources (or, worse, your time after a project's funding period has ended). With strong documentation procedures in place, your cyberinfrastucture team can focus on adapting, innovating, and realizing the full potential of the project and delivery of its accomplishments to the world.

Concluding remarks

Several themes tie these tips together: 1) Just as the analyses, data, and metadata will be integrated at different times and in different ways, so too, the parts of a systems biology project must be elevated from individual components to part of a larger whole. How well they fit will depend on the specifics of the project and the culture within it; 2) You need a cyberinfrastructure plan and that plan needs to be flexible enough to allow you to 'grow as you go' and innovate 'on the move'; 3) Much, if not all, of your project's deliverables will be public at some point and in some form; plan accordingly; 4) Investigate everywhere and everything, trust nothing, test everything; 5) To provide effective support, your team should be in the vanguard whenever and wherever possible.

A strong cyberinfrastructure can become the tie that binds, facilitating the growth of specific projects as well as the transdisciplinary scientific culture that has evolved in today's post-genomic era. Systems biology projects are helping to mature the overall phenomenon of team science in the life sciences and the perspective of a DMCIT is an excellent starting point to understand the challenges and opportunities that will affect overall success.

Data Accessibility Statement

Data referred to in this manuscript are organized and accessible at the National Institutes of Health, Bioinformatics Resource Center, PlasmoDB (Aurrecoechea, et al., 2009) that is part of EuPathDB (Aurrecoechea, et al., 2017) via a permanent central organizing site (PlasmoDB-MaHPIC) that contains the datasets with DOIs or links to the data in recognized archival databases, *e.g.* NCBI.

Acknowledgements

We give our heartfelt thanks to the MaHPIC consortium for growing with us during this project and for sharing their needs and gratitude over the years. Membership in the MaHPIC Consortium (MaHPIC, 2019) over the 2012–2017 five-year period of NIAID support has included: Dave C Anderson, Dalia Arafat, Sophia Banton, John W Barnwell, Steven E Bosinger, Patrick Breen, Robert Bridger, Cristiana Brito, Jung-Ting Chien, Julia Crutchfield, Megan Bass DeBarry, Christopher Duncan, Valerie Flint, Luis L Fonseca, Patricio Gallardo, AnaPatricia Garcia, Luiz G Gardinassi, Swetha Garimalla, Diego Giraldo, Anuj Gupta, Trenton Hoffman, Chris Ibegbu, Jianlin Jiang, Xuntian Jiang, Dean P Jones, Manoj Khadka, Amy Kong, Rachel Kutner, Miriam Lachs, Tracey J Lamb, Stacey A Lapp, Kevin J Lee, Frances Eun-Hyung Lee, Noah Legall, Shuzhao Li, Loukia Lili-Williams, Tim Morris, Douglas P Nace, Gregg Orloff, Dan Ory, Mariko S Peterson, Jan Pohl, Sarah T Pruett, Zhen Qi, Matthew Reed, Jorge L Salinas, Celia L Saney, Ignacio Sanz, Maren R Smith, Stephanie Soderberg, Andrew Spruill, Hannah Stealey, Amy K Stout, Olga Stuchlik, Mark P Styczynski, JoAnn Sullivan, Yan Tang, Gregory K Tharp, Rabindra Tirouvanziam, ViLinh Tran, Christopher Tseng, Karan Uppal, Eberhard O Voit, Douglas Walker, Susanne Warrenfeltz, Brent Weatherly, Lance Wells, Tyrone Williams, Zerotti Woods, Weiwei Yin, as well as the authors of this manuscript. We are also greatly thankful to institutional IT staff that have supported the project: Paul Brunk, Curtis Combs Jr., and Guy Cormier at the University of Georgia; Andre Bosman, Chang-Kwei Lin, Patrick Maloney, Sharon Mason, Shailish Nair, and Aaron Olson at Emory University; Jimmy Lummis at Georgia Institute of Technology.

Funding Information

This project was funded in part by Federal funds from the US National Institute of Allergy and Infectious Diseases, National Institutes of Health, Department of Health and Human Services under contract # HHSN272201200031C (PI: Mary R. Galinski), which supported the MaHPIC, and National Center for Research Resources [ORIP/OD P51OD011132], as well as resources from the Georgia Advanced Computing Resource Center, a partnership between the University of Georgia's Office of the Vice President for Research and Office of the Vice President for Information Technology.

Competing Interests

The authors have no competing interests to declare.

Author Contributions

Jeremy D. DeBarry and Jessica C. Kissinger contributed equally to this work.

References

Aurrecoechea, C, et al. 2009. PlasmoDB: A functional genomic database for malaria parasites. *Nucleic Acids Research*, 37: D539–D543. DOI: https://doi.org/10.1093/nar/gkn814

Aurrecoechea, C, et al. 2017. EuPathDB: the eukaryotic pathogen genomics database resource. *Nucleic Acids Res*, 45: D581–D591. DOI: https://doi.org/10.1093/nar/gkw1105

Bhattacharya, S, et al. 2014. ImmPort: disseminating data to the public for the future of immunology. *Immunol Res*, 58(2–3): 234–9. DOI: https://doi.org/10.1007/s12026-014-8516-1

Boland, MR, Karczewski, KJ and **Tatonetti, NP.** 2017. Ten Simple Rules to Enable Multi-site Collaborations through Data Sharing. *PLoS Comput Biol*, 13(1): e1005278. DOI: https://doi.org/10.1371/journal.pcbi.1005278

Cordy, RJ, et al. 2019. Distinct amino acid and lipid perturbations characterize acute versus chronic malaria. *JCI Insight*, 4(9): e125156. DOI: https://doi.org/10.1172/jci.insight.125156

FIPS. Federal Information Processing Standards. Available at https://www.nist.gov/itl/current-fips [Last accessed December 2019].

FISMA. Federal Information Security Modernization Act. Available at https://www.dhs.gov/cisa/federal-information-security-modernization-act [Last accessed December 2019].

Gardinassi, LG, et al. 2018. Integrative metabolomics and transcriptomics signatures of clinical tolerance to Plasmodium vivax reveal activation of innate cell immunity and T cell signaling. *Redox Biol*, 17: 158–170. DOI: https://doi.org/10.1016/j.redox.2018.04.011

Haug, K, et al. 2013. MetaboLights–an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res*, 41(Database issue): D781–6. DOI: https://doi.org/10.1093/nar/gks1004

iRODS-Consortium. 2019. iRODS 4.0 Available at https://irods.org [Last accessed 2019].

Joyner, CJ, et al. 2019. Humoral immunity prevents clinical malaria during Plasmodium relapses without eliminating gametocytes. *PLoS Pathog*, 15(9): e1007974. DOI: https://doi.org/10.1371/journal.ppat.1007974

Kazic, T. 2015. Ten Simple Rules for Experiments' Provenance. *PLoS Comput Biol*, 11(10): e1004384. DOI: https://doi.org/10.1371/journal.pcbi.1004384

Leinonen, R, et al. 2011. The sequence read archive. *Nucleic Acids Res*, 39(Database issue): D19–21. DOI: https://doi.org/10.1093/nar/gkq1019

MaHPIC. 2019. Malaria Host-Pathogen Interaction Center. Available at http://systemsbiology.emory.edu [Last accessed December 2019].

Merchant, N, et al. 2016. The iPlant Collaborative: Cyberinfrastructure for Enabling Data to Discovery for the Life Sciences. *PLoS Biol*, 14(1): e1002342. DOI: https://doi.org/10.1371/journal.pbio.1002342

PlasmoDB-MaHPIC. MaHPIC Data Catalog and Access Site. Available at http://plasmodb.org/plasmo/mahpic.jsp [Last accessed December 2019].

Tang, Y, et al. 2017. Integrative analysis associates monocytes with insufficient erythropoiesis during acute Plasmodium cynomolgi malaria in rhesus macaques. *Malar J*, 16(1): 384. DOI: https://doi.org/10.1186/s12936-017-2029-z

Tang, Y, et al. 2018. Metabolic modeling helps interpret transcriptomic changes during malaria. *Biochim Biophys Acta Mol Basis Dis*, 1864(6 Pt B): 2329–2340. DOI: https://doi.org/10.1016/j.bbadis.2017.10.023

Taschuk, M and **Wilson, G.** 2017. Ten simple rules for making research software more robust. *PLoS Comput Biol*, 13(4): e1005412. DOI: https://doi.org/10.1371/journal.pcbi.1005412

Wilkinson, MD, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data*, 3: 160018. DOI: https://doi.org/10.1038/sdata.2016.18

Zook, M, et al. 2017. Ten simple rules for responsible big data research. *PLoS Comput Biol*, 13(3): e1005399. DOI: https://doi.org/10.1371/journal.pcbi.1005399

How to cite this article: DeBarry, JD, Kissinger, JC, Nural, MV, Pakala, SB, Humphrey, JC, Meyer, EVS, Cordy, RJ, Cabrera-Mora, M, Trippe, ED, Aguilar, JB, Karpuzoglu, E, Yan, YH, Brady, JA, Hankus, AN, Lackman, N, Gingle, AR, Nayak, V, Moreno, A, Joyner, CJ, Gutierrez, JB, Galinski, MR and the MaHPIC Consortium. 2020. Practical Recommendations for Supporting a Systems Biology Cyberinfrastructure. *Data Science Journal,* 19: 24, pp. 1–12. DOI: https://doi.org/10.5334/dsj-2020-024

Submitted: 09 December 2019 **Accepted:** 08 May 2020 **Published:** 09 June 2020

]u[*Data Science Journal* is a peer-reviewed open access journal published by Ubiquity Press.

OPEN ACCESS