Published in final edited form as:

Nat Immunol. 2009 December ; 10(12): 1225–1227. doi:10.1038/ni1209-1225.

Data Management: It Starts at the Bench

Damien Chaussabel, Hideki Ueno, Jacques Banchereau, and Charles Quinn

Baylor Institute for Immunology Research and Baylor Research Institute, Dallas, TX 75204, USA

Abstract

Data management has been neglected but should be made an integral activity in all research laboratories. Chaussabel and colleagues discuss how to implement this at the bench.

Immunology research has transformed over the past decade into a data intensive field. Technological breakthroughs in genomics, proteomics, but also polychromatic flow cytometry and imaging, account for this accelerating trend. The first logical response to this avalanche of data has been to develop novel analytical tools and approaches. However, comparatively little has been done to address growing data management needs. While the current debate centers on data sharing and archiving in public repositories, we argue here the critical importance of making information management an integral part of the activities of the research laboratory.

Data management is critical because it insures that information once collected is and remains secure, interpretable and exploitable. It encompasses plans, policies, programs and practices aiming to control, protect, deliver and enhance the value of data and information assets. Currently, most of the information generated within research laboratories is fragmented between hard drives, CDs, printouts and laboratory notebooks. As a result, the useful lifespan of a dataset often does not extend beyond the publication of the results. The development of public repositories such as the NCBI gene expression omnibus (GEO) does provide means to preserve and share data. Yet, in most cases data is deposited years after it has been generated and is often only accompanied by minimal supporting information. Furthermore such repositories tend to only focus on banking a single data type (e.g. sequence data in GenBank or gene expression profiling data in GEO). This leaves out many parameters, such as flow cytometry results, chemokine/cytokine abundance measurements or detailed clinical phenotype, which are essential to immunological studies. Pioneering initiatives in the immunology field, with for instance the establishment of the ImmPort, immune epitope and T1DBase databases, point us in the right direction^{1–3}. However repositories such as these are clearly not designed to satisfy the day-to-day needs of research laboratories, which is where the data should be managed in the first place.

This commentary describes the tasks involved in managing the data generated everyday “at the bench”. We will also discuss challenges and opportunities the implementation of data management solutions brings for the research investigator and for the research enterprise.

Correspondence to: Damien Chaussabel; damienc@baylorhealth.edu.

Competing interests statement:

The authors have no competing financial interests to declare.

Data management happens at the bench

First data must be captured. Data sources include instrument output but also sample tracking information (e.g. barcode number, location), quality control (e.g. RNA integrity) and other variables (e.g. yield, concentration). Such parameters are captured by Laboratory Information Management Systems (LIMS), a category of software used specifically for the management of laboratory workflows. In the context of clinical studies large amounts of information need also to be captured at the bedside. Finally, information describing the study and experiments performed also need to be recorded for the data to be interpretable. This information is often referred to as “metadata”.

Storage is another important aspect of data management. Instrument output files can be organized on a file server. The data can also be loaded in a database, in which case it will be readily available for query and retrieval. Another consideration is data safety including managing access to the data and maintaining integrity (e.g. redundant storage, backup strategies).

Having the ability to integrate data from multiple sources is becoming critical. This is a difficult task to perform without the appropriate bioinformatics infrastructure in place, yet it is viewed as a key for discoveries that will be made through systems approaches. Traditionally, data from different sources would be organized and linked within a single relational data management system. However, the effort required for developing and maintaining such a system means that it might only be practical to use in large-scale projects. More recent web technologies can be employed to develop applications that will aggregate data from multiple data management systems, in real-time (Fig. 1). This means that data stored in different application databases, different types of databases (e.g. MySQL, MS Access, FileMaker, Oracle etc...), and in different geographic (physical) locations can be queried and retrieved using a single application. This type of approach affords more flexibility and is therefore better suited for research environments and smaller scale projects, as well as collaborative multi-centre projects.

Storing organized data is not sufficient. It must be readily available to bioinformaticians who will carry out downstream analyses; and also to immunologists who can gain considerable insight simply by querying and browsing the data, provided that sufficient information is available for its interpretation.

How the management of information is approached will vary based on the scale of the project, the type of data being generated and the laboratory environment (e.g. research lab, core facility). For instance, it may not always be necessary or feasible to rigorously track samples or reagents. However, even small-scale immunology projects will generate electronic files and results that need to be reconciled together with details regarding the experiment in order for the data to remain interpretable.

The Promise

Given the proliferation of high throughput profiling platforms, ramping up efforts to better manage data certainly seems to be “the right thing to do”. Yet, data management can only be a means to an end and it is therefore important to develop a clear rationale for engaging in such efforts.

One of the first goals is to preserve the value of the data and expand its usable lifespan. Well-annotated data once captured will remain interpretable for years to come and by individuals who may not have directly participated in generating it. This notion takes on a particular importance when science becomes driven by data as a source of hypotheses. Clearly any given dataset will yield more hypotheses than can be tested by any given lab. Furthermore, additional

hypotheses may be formulated only when several such datasets are combined and analyzed collectively. Hence, more than ever before it is essential that efforts be made in the biomedical research field to preserve the long-term integrity and interpretability of data. In fact, in a given project how the data is managed may arguably be more critical than how it is analyzed. Indeed, data that is well managed can always be analyzed again, possibly using novel tools or alternate approaches. However, there are no second chances for data that has not been captured or integrated properly from the beginning.

In some situations data management solutions constitute the only means by which a given task can be accomplished. A LIMS system can for instance support a level of throughput in a lab that would otherwise be impossible to sustain. Having the ability to efficiently capture and integrate data is also essential when it comes to mining complex datasets, which may for instance include clinical, genomics, proteomics and flow cytometry data. Furthermore, the reliance on data management solutions increases exponentially when one considers carrying out analyses across several such datasets.

Doing better at managing data translates into considerably improved abilities to share it, whether it is with collaborators, members of research consortia or the scientific community at large. It also provides a unique opportunity to enhance the communication of results in peer-reviewed scientific publications by providing access to raw instrument output and “behind the scenes” experimental details backing a particular finding.

Data Management in Action

To illustrate our point, we are presenting the results of an experiment where peripheral blood mononuclear cells were exposed to influenza, CMV or MART peptides. After 24 hours cells were harvested and processed for gene expression microarray analysis. Abundance of cytokines and chemokines in supernatants were measured using a multiplex protein assay after 48 hours of culture. In addition, flow cytometry analysis was performed after 8 days of culture to measure proliferative responses of CD8⁺ T cells following antigenic exposure. Here we present the gene expression level of the interferon-inducible chemokine CCL7 (also known as monocyte chemoattractant protein 3) as measured by microarrays (see <http://www.biiir.net/gxb/cc17.htm>). This interactive web figure provides the reader with access to several layers of information. Hovering the mouse cursor over the bar of the histogram, representing gene expression levels of CCL7 for each sample, brings up pop up windows that under different modes will display different variables. Sample information provide the characteristics of the donor (demographic information, HLA type), culture conditions (peptides) and sample information (e.g. identifiers, freezer location). Quality Information gives the RNA quality and yield along with quality control parameters generated during the microarray analysis. This information is recorded in our LIMS and simply retrieved as accessory information for this figure. Associated results allows pop up windows to display the corresponding flow cytometry results as well as protein chemokine and cytokine amounts measured in culture supernatants for each sample.

In addition, raw data and associated files can be exported at the click of a button (e.g. output files from flow cytometry and multiplex protein assays, presentation slides describing the gating strategy). Another link provides access to all experimental details necessary to properly interpret and replicate the results.

The level of details provided affords the transparency necessary to replicate this experiment and properly interpret its results. The primary data underlying those results is also made available for re-analysis. Furthermore, presenting integrated results in an interactive format greatly facilitates the interpretation of an experiment where multiple parameters need to be

considered (in this case proliferation of CD4⁺ and CD8⁺ T cells, cytokine production, different antigenic peptides, and donor HLA). Importantly, this data is captured as part of normal laboratory activities where data management is integrated to the workflow. As a result presenting the data in this manner and with this level of details can be done with very little additional effort by the research investigator.

Considerations Moving Forward

Given the volume of information currently generated and the resources engaged generating it, making the case for data management is relatively straightforward. However, the barriers that must be overcome before data management becomes a reality are significant. There are a few points that will have to be considered while seeking solutions to our growing data management needs. Firstly, managing data is a long-term project. The amount of efforts required by such an endeavor is far from negligible. The development of data management solutions stretches well beyond the initial implementation phase. Indeed, it basically never ends since the instrumentation and laboratory workflow evolve continuously, thus making data management an ever-moving target.

Secondly, managing data is not exciting. As critical as the implementation of data management solutions in the laboratory are, it is not as interesting and instantaneously rewarding as, for example, data analysis. It is important to stay focused and on track, keeping in mind the downstream, and sometimes long-term, benefits of managing information at the bench. Another barrier is to acknowledge there is more to it than just managing omics data, especially in the case of immunology where a wide variety of results can be generated from a single experiment. Today even small projects can generate considerable amounts of data. All information generated about an experiment should be captured electronically, essentially superseding the role of the laboratory notebook.

Data must also be managed proactively. Gathering information retroactively, months or sometimes years after the data has been generated takes a considerable amount of efforts, and is often met with hardship. The challenge is often more cultural than technological. Developing the infrastructure for managing data is only one part of the problem, getting people to use it is another matter. Indeed, besides changes in workflows that it may require, managing data is a time consuming task that often adds to an already busy workload. Also, adequate staffing and training is necessary. It is also important for individual investigators to understand that data management brings benefits for the group as well as for themselves.

Finally, data sharing and data management are related yet distinct issues. While preserving and sharing data in public repositories is critical, and the current debate necessary^{4, 5}, it should not distract from the data management needs at the bench. For instance, while funding agencies have started mandating data sharing plans in grant applications, it is surprising that at the same time data management should be left out of the evaluation process. Making data management in research laboratories a reality should become an immediate priority. It comes with its own set of goals and challenges that are distinct from those associated with the sharing of data in large public repositories.

Conclusion

In conclusion, with established platforms such as gene expression microarrays now more robust and affordable than ever and the recent introduction of breakthrough technologies, such as deep sequencing, the trend towards ever expanding data acquisition capabilities shows no signs of abating.

One of the factors currently limiting our ability to take full advantage of these advances is the lack of adequate solutions for managing data. While the problem is widely recognized it is also a difficult one to address. It will have to be confronted nonetheless as meeting the data management challenge in biomedical research will truly prove transforming.

Acknowledgments

We would like to acknowledge Durgha Nattamai, Laure Bourdery and Jill Plants for their help generating the data used to illustrate this commentary. David Jutras for his help with the application development. Karolina Palucka for her critical reading of the manuscript.

Supported by the Baylor Health Care System Foundation and the National Institutes of Health (U19 AIO57234-02, U01 AI082110, P01 CA084512).

References

1. <https://www.immport.org>.
2. Hulbert EM, et al. T1DBase: integration and presentation of complex data for type 1 diabetes research. *Nucleic Acids Res* 2007;35:D742–746. [PubMed: 17169983]
3. Peters B, et al. The immune epitope database and analysis resource: from vision to blueprint. *PLoS Biol* 2005;3:e91. [PubMed: 15760272]
4. Nelson B. Data sharing: Empty archives. *Nature* 2009;461:160–163. [PubMed: 19741679]
5. Piwowar HA, Becich MJ, Bilofsky H, Crowley RS. Towards a data sharing culture: recommendations for leadership from academic health centers. *PLoS Med* 2008;5:e183. [PubMed: 18767901]

Figure 1.

Pulling it all together. Information generated during the course of an experiment is captured by a Laboratory Information Management System (LIMS) and other *ad hoc* databases designed to store microarray, flow cytometry or multiplex protein assays results. A web application brings together information from multiple sources in order to afford the user a comprehensive, integrated view of the results. This concept is illustrated in an interactive figure accessible online at: <http://www.biir.net/gxb/ccl7.htm>