DELSA WORKSHOP IV: LAUNCHING THE QUANTIFIED HUMAN INITIATIVE

Elizabeth Stewart,[1,2] *Todd Smith,*[2,3] *Andrea De Souza,*[2,4] *Jack Faris,*[2,5] *Lennart Martens,*[2,6,7] *Sophie Mohin,*[2,8] *Vural Ozdemir,*[2,9] *Courtney MacNealy-Koch,*[1,2] *and Eugene Kolker*[1,2,10,11]

Overview

The mission of the Data-Enabled Life Sciences Alliance (DELSA Global) is to "Accelerate the impact of data-enabled life science research on the pressing needs of the global society." In its first 18 months, DELSA has catalyzed connections and interactions for more effective and sustainable science by bringing stakeholders together through physical or virtual proximity to share ideas, discuss new insights, and form novel collaborations.

During our most recent annual Washington, DC, meeting (May 16–17, 2013), DELSA brought together life and computer scientists, data analysts, research funding agency representatives, and many others to discuss and formulate plans for furthering the initiative of 21st-century collective innovation. In an exciting day of lightning talks and brainstorming, participants discussed the management and analysis of emerging datasets that hold such immense promise for understanding and improving the human condition and our relationship with the worlds around us and within us.

A focus of this meeting was on the Quantified Human (QH) Initiative. QH takes our natural curiosity about self and combines multi-omics and clinical data to draw conclusions about our physical condition both current and future. Measures such as height, weight, and blood pressure have been used throughout medical history; however, it is now possible to track many other measures such as caloric/nutritional intake and output, blood components, and sleep patterns. These data can be viewed in the context of our body as an ecosystem by including measures of the commensal microorganisms, collectively referred to as the microbiome. All of these results, taken together and over a period of time, can lead to a detailed picture of our overall health and open up a whole new level of understanding about the microenvironment that exists inside us. However, the resulting datasets are complex and immense. While the potential exists to use these data to explore the depth and breadth of ourselves in new and unimagined ways, we need new paradigms and policies for organizing, managing, and sharing the data, combined with new publishing and citation models.

The QH initiative issues were divided into four categories:

- Data and Meta-Data,
- Metrics and Evaluation,
- Research Outcomes, and
- Outcomes for the General Public

[1]*Bioinformatics and High-throughput Analysis Laboratory, Seattle Children's Research Institute, Seattle, Washington.*
[2]*Data-Enabled Life Sciences Alliance (DELSA Global), Seattle, Washington.*
[3]*PerkinElmer, Waltham, Massachusetts.*
[4]*Chemical Biology Platform, Broad Institute, Cambridge, Massachusetts.*
[5]*The Fearey Group, Seattle, Washington.*
[6]*Department of Medical Protein Research, Vlaams Instituut voor Biotechnologie (VIB), Ghent, Belgium.*
[7]*Department of Chemistry, Faculty of Medicine and Health Sciences, Ghent University, Ghent, Belgium.*
[8]*Mary Ann Liebert Publishers, New Rochelle, New York.*
[9]*Faculty of Medicine and Desautels Faculty of Management, McGill University, Montreal, Canada.*
[10]*Predictive Analytics, Seattle Children's, Seattle, Washington.*
[11]*Departments of Biomedical Informatics & Medical Education and Pediatrics, University of Washington, Seattle, Washington.*

Data and Meta-Data

The QH initiative is new and builds upon open science and rapidly emerging data-intensive technologies. There are currently two landmark datasets available for exploration, the Dr. Larry Smarr personal multi-omics data[1] and the integrative personal multi-omics profile.[2,3] We were extremely fortunate to have Dr. Larry Smarr as our keynote speaker. To have the person responsible for one of the QH landmark datasets discuss the impact this science has on his health was a powerful call to action.

The Smarr datasets were discussed at the workshop as an example of the opportunities and challenges that could come from such studies. As the Smarr datasets were gathered for more than a decade, the parameters of sample gathering, file formats, and the means of analysis have changed during those years. These changes must be addressed in order to share the data in a format that will allow reproducible analysis. In addition, clinical data are often quite difficult to access, although with the advent of electronic medical records (EMR), it is likely that access will improve. With EMR data, to facilitate meaningful use of the datasets, appropriate standards for data management will need to be developed and implemented in conjunction with the supporting application programmable interfaces (APIs) to facilitate data integration between systems and across datasets.

"TO HAVE THE PERSON RESPONSIBLE FOR ONE OF THE QH LANDMARK DATASETS DISCUSS THE IMPACT THIS SCIENCE HAS ON HIS HEALTH WAS A POWERFUL CALL TO ACTION."

These issues are not limited to personal omics and clinical datasets alone but rather part of a bigger issue of data quality and accessibility. The workshop participants agreed that the recent *Nature* checklist recommendation[4] is an appropriate first step by the community as it addresses accessibility and veracity of data as well as the reliability and reproducibility of research—crucial issues as we transform data to knowledge, action, and outcomes. The need for such assessments was supported by DELSA's Letter of Endorsement for Nature's checklist.[5] The development of checklists tailored to different aspects of the life sciences, including QH, was proposed.

Data Management Guidelines

Along with the proposal of checklists came a discussion of the need for data management policies. The benefit of high-quality resources is improved quality and reproducibility of research—anything less than careful attention to the 5V's of big data (volume, velocity, variety, veracity, and value)[6] will lead to waste. However, equally wasteful is a resource that is poorly managed and thus underutilized. Detailed and specific data management guidelines will increase accessibility through tailored usage policies and give user communities clear expectations of what is acceptable use of a collective resource. It was proposed that DELSA formulate such guidelines for general review and adoption, thus aiding the community as it works toward data democratization and effectiveness.

Metrics and Evaluation

As the community works through the requirements and challenges of this new project, there will be the need to establish metrics for data integrity, analysis methods, and perhaps even publication formats. As discussed above, checklists are an excellent first step for framing the data and research integrity requirements of a project. In conjunction, DELSA will work to establish a prototype process and infrastructure for future, larger-scale studies. Studies of one or two individuals give a tantalizing glimpse of what could be possible, but more subjects are needed for a clearer picture.

Publications Types and Impact

Of particular interest to many were the discussions around publications. Many facets of communication were examined. Along with the discussions of the *Nature* checklist were also discussions around other nontraditional publication systems such as Nanopubs.[7] Nanopubs seek to capture and share the smallest unit of publishable information, thus allowing datasets, for example, to be published and tracked for future credit. Nanopubs, and other yet-to-be discovered publication processes, raise questions about the use of publications and citation indexes to evaluate a scientist for career opportunities such as tenure.

A recent suggestion is to do away with the h-index as it leads to biases in choices about scientific questions. Researchers are less likely to chart new territory because their articles will not be cited by others if their choice of work leads to very few colleagues. The impact factor is also under pressure as it has been suggested that researchers preferentially publish in the "top" journals regardless of suitability, and similarly researchers are less likely to work in a little known avenue of research as it would be harder to get a publication or be cited.

Research Outcomes

Along with ensuring the scientific validity of the datasets comes the need to encourage and evaluate their innovative use. The QH datasets concept is so new that there are no

tried-and-true methods for gathering, managing, and analyzing the data. An initial effort by DELSA will be to provide project information to the scientific community so that a globally distributed and highly diverse group of user communities can collectively explore the data and determine both what questions to ask and how to answer them. Such research outcomes generated through collective data analysis, interpretation, and valorization, even possibly through crowdsourcing, would be made available through a quantified human portal powered with end user–friendly tools, a FAQ, and perhaps as part of a publicly available massive open online course (MOOC) unit. In addition, DELSA hopes to host a QH global biochallenge around the above unique datasets to encourage both questions and answers.

"ALONG WITH ENSURING THE SCIENTIFIC VALIDITY OF THE DATASETS COMES THE NEED TO ENCOURAGE AND EVALUATE THEIR INNOVATIVE USE."

Outcomes for the General Public

Datasets about people and patients hold amazing promise to advance our self-knowledge and health, but the same data can bring up issues of privacy, medical insurance, and the ethical use of information. These are issues that belong to all of us. A top priority of DELSA's is the dissemination of information about life sciences to the public, not limited to information only about QH datasets. It is important that society as a whole has accurate and thorough knowledge about data-enabled science in order to support it and encourage it through funding decisions and citizen participation. A Quantified Human FAQ to the public will be used to disseminate information about the project's opportunities and challenges.

Quantified Human Initiative

The QH initiative is one exciting part of a larger DELSA vision: Through interdisciplinary research and transdisciplinary engagement, the life sciences community will move from a "single scientist–single project" model to collective innovation. As the life sciences community works to meet the challenges of 21st-century science, it has become apparent that established approaches must be examined and new ways of accomplishing science must be entertained. During the second day of the workshop participants heard from speakers whose mandate was to "Tell us how your efforts can contribute to collective innovation." This statement alone is a departure from the usual way of doing business in the life sciences and demonstrates DELSA's focus. With five-minute lightning talks, the participants covered pertinent subjects such as funding, publications, and outreach to the public and developing countries as well as the need for improved communication and education of scientists and the public. These lightning talks covered subjects that are of interest to a broad range of stakeholders; DELSA leaders frequently exchange ideas and suggestions about these and other subjects with publishers such as Nature Publishing Group, Mary Ann Liebert (specifically *OMICS* and *Big Data*), and PLOS.

Workshop participants also explored opportunities for outreach. Thanks to contributions from a few DELSA founders there are funds available for "microgrants" that will be used in two approaches—funding young researchers to pursue a research avenue they would not have otherwise done, and funding microgrant projects in developing countries where a small amount of funding can have a big impact. Used in a targeted manner, microgrants can serve as force multipliers to accelerate big data R&D through collective innovation.

Education was also an important topic. Recent innovations in educational content delivery have opened up many new avenues of communication. DELSA's working group "Training Data Scientists" is developing a collection of MOOC segments that can be used for young and not-so-young investigators.

The workshop gave the participants a chance to shape the future of an emerging and powerful data-enabled initiative, Quantified Human. From the discussions, DELSA is formulating plans for education efforts around QH and dataset dissemination to the scientific community and the public. These efforts will be followed by DELSA efforts to support research and evaluate results.

In addition, we are continuing the broader efforts of DELSA with our working groups and endorsed projects—addressing issues such as data set accessibility and international outreach.

Next Steps

The workshop concluded with eight actionable items and an open invitation for participation from the community:

1. Launch of microgrant outreach for younger generations
2. Launch of microgrant outreach for developing countries
3. Commitment to progress reports for sponsors, members, and community
4. Data-enabled Life Sciences 101 mini MOOC
5. FAQs for Quantified Human Initiative
6. Quantified Human actionable items for each group by each working group
7. Form planning committee for next DELSA meeting

Conclusions

Our interest in the natural world around us, and in particular within us, is as old as the dawn of thought. The need to know may have started with the basic drive for survival but has grown to encompass the wish to not just survive but to thrive, to improve, and to know for knowledge's sake alone. The life sciences represent this need to know about ourselves and our world and as such are shaped by our expanding abilities to explore both within and without. In particular, the life sciences are being transformed by the ever-increasing availability of datasets that explore the depth and breadth of ourselves and our environment.

DELSA is neither a funding agency nor a consortium but an alliance of individuals and organizations with a common commitment to data-enabled life sciences. Put simply, we plan to contribute to science in ways big and small through 21st-century collective innovation. We invite you to contribute with us.

Disclosure Statement

No competing financial interests exist.

Acknowledgments

We would like to thank our generous sponsors: The Gordon and Betty Moore Foundation, Seattle Children's Research Institute, EMC, and Intel. Our sincere thanks go to Internet2 for recording the workshop.

References

1. Smarr L. Quantifying your body: A how-to guide from a systems biology perspective. Biotechnol J 2012; 7:980–991.
2. Chen R, Mias G, Li-Pook-Than J, et al. Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell 2012; 148:1293–1307.
3. Stanberry L, Mias G, Haynes W, et al. Integrative analysis of longitudinal metabolomics data from a personal multi-omics profile. Metabolites 2013; in press.
4. Announcement: Reducing our irreproducibility. [Editorial] Nature 2013; 496:398.
5. Kolker E, Altintas I, Bourne P, et al. Reproducibility: In praise of open research measures. Nature 2013; 498:170.
6. Higdon R, Haynes W, Stanberry L, et al. Unraveling the complexities of life sciences data. Big Data 2013; 1:42–50.
7. http://nanopub.org

Address correspondence to:

Elizabeth Stewart, PhD
Senior Scientist, Seattle Children's Research Institute
1900 Ninth Avenue
M/S: C9S-9
Seattle, WA 98101

E-mail: elizabeth.stewart@seattlechildrens.org

Courtney MacNealy-Koch
Program Coordinator, DELSA Global
1900 Ninth Avenue
M/S: C9S-9
Seattle, WA 98101

E-mail: courtney.macnealy-koch@delsaglobal.org