J Pathol Inform

Editor-in-Chief:
Anil V. Parwani , Pittsburgh, PA, USA
Liron Pantanowitz, Pittsburgh, PA, USA

OPEN ACCESS
HTML format

For entire Editorial Board visit : www.jpathinformatics.org/editorialboard.asp

Editorial

The need for the pathology community to sponsor a whole slide imaging repository with technical guidance from the pathology informatics community

Jason D. Hipp, Jeffrey Sica, Barbara McKenna, James Monaco[1], Anant Madabhushi[1], Jerome Cheng, Ulysses J. Balis

University of Michigan, Department of Pathology, M4233A Medical Science I, 1301 Catherine, Ann Arbor, MI 48109-0602, [1]Rutgers The State University of New Jersey, Department of Biomedical Engineering, 599 Taylor Road, Piscataway, NJ, USA

E-mail: *Jason D. Hipp - jason.hipp@gmail.com
*Corresponding author

Received: 10 June 11 Accepted: 10 June 11 Published: 26 July 11

This article may be cited as:
Hipp JD, Sica J, McKenna B, Monaco J, Madabhushi A, Cheng J, *et al.* The need for the pathology community to sponsor a whole slide imaging repository with technical guidance from the pathology informatics community. J Pathol Inform 2011;2:31.

Available FREE in open access from: http://www.jpathinformatics.org/text.asp?2011/2/1/31/83191

In our recently published Editorial in the *American Journal of Surgical Pathology* entitled "Digital Slide Repositories for Publications – Lessons Learned from the Microarray Community", we describe a number of potential benefits of a Whole Slide Imaging (WSI) repository for the pathology and biomedical communities, including: 1) an increased level of rigor for peer review of scientific studies submitted for publication, 2) open confirmation and validation of the entire collection of slides from published studies, rather than the contemporary reality where there is only the availability of limited fields of view as selected by authors, 3) the potential that multiple perspectives generated from the review of primary WSI data will identify alternate interpretations and/or conclusions, 4) improved educational opportunities, and 5) availability of reference data sets for inter-observer concordance studies and image analysis algorithm development.[1] While such repositories of scientific data are usually designed by information and computer scientists, we will discuss an alternative solution whereby pathology informaticists, in combination with computer scientists, would design a repository based on the needs and abilities of the end-user in question – in this case, the pathologist. Furthermore, the recent announcement of the closure of the proteomics repository and the Sequencing Read Archive (SRA)[2] highlights the need for the pathology community to sponsor/support such a repository such that it is immune to fluctuations in the federal budget and other external factors associated with the oversight being stewarded by non-pathologists. The purpose of this communication is to alert the pathology community to the need for sponsorship of their own WSI repository and to alert the pathology informatics community of the need for their help in the design and implementation of such a repository.

The proteomics repository and the SRA, which has collected data since the 1990s, was recently inactivated due to budgetary issues (announced in February 2011).[2] Instead, funding is now being reallocated to support newer technologies that support data derived from next-generation sequencers.[3] However, other databases similar to SRA, housed in Europe (EMBL-EBI) and Japan (DDBJ), will continue to remain operational.[2]

These closures should not have been considered as an unanticipated development, with this reality underscored by the observation that the bioinformatics community is not at all distressed, and rather, fully understands why the long-established SRA database was terminated (as discussed by the bioinformatics blog community).[2]

Access this article online

Quick Response Code:

Website: www.jpathinformatics.org

DOI: 10.4103/2153-3539.83191

Based on commentary of scientists at the Phylogenomics blog site,[4] the drawbacks of the SRA were the result of improper design and implementation. Specifically, the SRA protocols – meaning the software used to download, search and to retrieve procedures, etc. – were established by the SRA itself. This resulted, for example, in the use of the file transfer protocol (FTP) program (an old file transfer protocol that is notorious for failing to complete the transmission of large files) and an antiquated search utility developed by the National Center for Biotechnology Information (NCBI). Contrast this with the most successful public resources such as Wikipedia and the sharing of very large videos, where only the minimum basic set of operational rules are mandated; the user community then implicitly enacts a Darwinian-like evolution toward an optimal set of services and operational models. In these instances, searches are performed by Google and file transfers are effected by the use of the far more contemporary and dependable BitTorrent peer-to-peer file sharing protocol. Other bloggers, such as Nick Loman (on Pathogens: Genes and Genomes blog), hold true that a decentralized database, where each sequencing center is responsible for making their own data available and useful to others, would be better.[2]

An important lesson learned from the closure of the proteomics repository and the SRA, and the ensuing response from the bioinformatics blog community, is the need for a simple, graphical user interface that is analogous to a Laboratory Information System (LIS), since most pathologists have little to no programming experience. Of critical importance for the adoption of such a repository will be the ease of use for the pathologist and the ability to search the meta-data of WSI content (i.e. diagnosis, tissue site, synonyms, etc.). Pathology informaticists (working in tandem with computer scientists) are in a unique position to bridge the two worlds of pathology and distributed web repositories, and should be the "drivers" of designing such a repository. As learned from the blogs, it will be important that there be close curation and maintenance of the repository (possibly with the assistance of the pathology informatics community), that is responsive to changing and updating the repository as needed to ensure that it continues to meet the needs of its users, the pathologists.

Another critical lesson we can learn from the closure of the above databases is that such a repository for pathology will need to be immune to external pressures that are beyond the control of the users, such as fluctuations in the federal budget. This may require a large initial financial investment by the pathology community itself, as the benefits of a functioning WSI repository might not be immediately apparent to other biomedical communities that are currently focused on the latest molecular technology rather than the potential for numerical-automated analysis of the anticipated vast library of digital data resulting from scanning of H&E-stained microscope slides. However, once established, such a repository could be a unique growth opportunity for the field of pathology, by reinforcing its central role in biomedical and translational research. For example, if clinician scientists or basic scientists are required to deposit their slide image data into a repository, such effort will encourage collaborations with pathologists to ensure accurate histopathology review.

While the initial benefits of a WSI repository would be for pathologists and pathology informaticists, we envision the benefits would become more broadly evident to the clinical trial research community, as it would enable the benefits associated with increased transparency, open access to the review of primary pathology data, and enhanced longitudinal continuity between past and current studies. Life scientists studying biomarkers or proteins of interest would have access to slide data from related works that may help in the understanding of tissue-level distribution of such proteins, their expression patterns and finally, cellular localization.

However, for all this to happen, the importance of a WSI repository will first need to be recognized by the pathology community (which was the goal of the AJSP editorial). However, as a next step, there must be a strategy of succession, whereby support is guaranteed in perpetuity by the pathology community (the goal of this communication). Who would be better to sponsor/endorse such an archive than the large pathology organizations, perhaps as a joint activity? The combined prestige and membership of the United States and Canadian Academy of Pathology, the College of American Pathologists and the American Society for Clinical Pathology, joining forces with the Association for Pathology Informatics, would bring prominence and permanence to such a repository. Furthermore, with participation of google.org (Google's philanthropic foundation) to sponsor the effort through the use of their servers, or similarly, support for cloud services by Amazon.com, the issue of significant storage space requirements could be mitigated. There would also be a need for such organizations and foundations to provide ongoing funding, possibly in conjunction with sponsorship from the National Institutes of Health/Department of Defense/National Institute of Standards and Technology/Agency for Healthcare Research and Quality, that could and should create Request For Proposals (RFPs) to fund the creation of these databases and their continued stewardship. Database curation is expensive, not so much due to the costs of digitization and storage of data, but rather due to the immense cost and effort associated with the creation of training datasets. For example, image classification libraries and test image sets necessarily require deep annotation of the pathology data, a time consuming and laborious process.

All this cannot happen at once. In an effort to catalyze this process, we have instantiated a fully functioning prototype WSI repository at the University of Michigan, Department of Pathology, Division of Pathology Informatics: www.wsirepository.org. Learning from the multitude of bloggers' comments, we have ensured that our developers have worked hand in hand with pathologists to realize a workflow process for submission and retrieval of digital slides, which is easy and straightforward. Along with ease of use, the system is designed to address high levels of web traffic, while at the same time, allowing for large payloads of WSI data. Our goal in creating this WSI repository is to promote open collaboration and contribution to the total aggregate of available slide data. By empowering collaborators with access to this data, the future of digital pathology and image analysis is hopefully served rather than hindered by current practices.

This effort has been attempted by Gurcan, Rajpoot, and Madabhushi who organized a Histopathology Image Analysis competition in conjunction with the International Conference on Pattern Recognitions meeting in Istanbul, Turkey, in 2010.[5] A small cohort of carefully annotated histopathology data was made available to participants, who ran their algorithms for nuclear and lymphocyte detection. The goals were: (a) to engage the larger pattern recognition community to start working together on problems in digital pathology and (b) to raise awareness of the need for carefully annotated digital pathology repositories.

There are still issues that need to be addressed for successful implementation of a WSI repository, such as copyright ownership, readership misuse and disagreement, regulatory standards, and publishers' participation. Rather than reinventing the wheel, we can build upon the knowledge that has been already gained from the microarray community's experience.

With regards to copyright ownership and readership interpretation, we envision taking an approach similar to NCBI GEO, where such WSI repositories serve as a "holding tank" for digital slides, just as NCBI GEO houses gene expression data. There is language on the NCBI GEO website that indicates that they are not in a position to validate any claims of patent or copyright privileges to the material in the repository, and so freely allow downloading and reproduction of their contents.[6,7] The NCBI GEO site also includes a disclaimer stating that the information presented in their GEO website is based on data independently submitted by the scientific community, making it impossible for the NCBI to independently verify the validity, quality or biological significance of the submitted data.[7] A WSI repository should make use of similar indemnifications and approaches.

To address the regulatory standards for submitted data, the microarray community developed the Minimum Information About a Microarray Experiment (MIAME) standard,[8] which enables the interpretation and reproducibility of the experiment in a manner which is minimally likely to divulge subject identity, operating in consonance with the Bell Ethical Premise. This standard is used by microarray repositories in the US (GEO at NCBI), the UK (EBI), and Japan (CIBEX at DDBJ).

Additionally, one can envision adapting standards such as the one developed by DICOM working group 26 (Supplement 145) for the submission of digital slides to WSI repositories.[9] The scientific merit of submitted data would fall under the jurisdiction of the peer-review process of the scientific journals to which each manuscript would be submitted. In addition, the public availability would allow other investigators to confirm or refute the validity of the data, as we previously opined in our editorial about stem cell microarray data.[1]

The success of such a repository could not have been achieved without the support of both scientific journals and funding agencies. This was briefly addressed in our editorial by Hipp *et al.*[1] In fact, it was through open letters from the scientific community, ultimately published in leading journals, that the goals of a central repository and associated data encoding and submission standards were achieved.[10-15]

We would like to end with a blog comment from *Attractor from the Tree of Life* blog where substituting pathologist for biologist is applicable: "In general, the design of biological databases should be led by biologists [pathologists] rather than programmers. Programmers tend to think of the latest fancy technologies, but leave most biologists [pathologists] in a mess".[4]

To paraphrase the voice heard in the movie *Field of Dreams*, "If we build it, they will come".

REFERENCES

1. Hipp JD, Lucas DR, Emmert-Buck MR, Compton CC, Balis UJ. Digital slide repositories for publications: Lessons learned from the microarray community. Am J Surg Pathol 2011;35:783-6.
2. Callaway E. Unpopular genomic database faces budget axe. Nature Newsblog. 2011 February 15. 2011.
3. NCBI. NCBI To Discontinue Sequence Read Archive and Peptidome. Available from: http://www.ncbi.nlm.nih.gov/About/news/16feb2011. [cited in 2011].
4. Eisen JA. The Tree of Life: Though I generally love NCBI, the Sequence/Short Read Archive (SRA) seems to have issues; what do others think? Available from: http://phylogenomics.blogspot.com/2011/02/though-i-generally-love-ncbi.html. [cited in 2011].
5. Gurcan M, Madabhushi A, Rajpoot N. Pattern recognition in histopathological images: An ICPR 2010 Contest. Workshop in Conjunction with International Conference on Pattern Recognition, 2010 In press.
6. NCBI. Copyright and Disclaimers. Available from: http://www.ncbi.nlm.nih.gov/About/disclaimer.html. [cited in 2009].
7. NCBI-Gene-Expression-Omnibus. GEO Disclaimer. Available from: http://www.ncbi.nlm.nih.gov/geo/info/disclaimer.html. [cited in 2011].

8. FGED-Society. Minimum Information About a Microarray Experiment - MIAME. Available from: http://www.mged.org/Workgroups/MIAME/miame.html. [cited in 2011].
9. DICOM Standards Committee WG, Pathology. Digital Imaging and Communications in Medicine (DICOM), Supplement 145: Whole Slide Microscopic Image IOD and SOP Classes 2010.
10. Edgar R, Domrachev M, Lash AE. Gene expression omnibus: NCBI gene expression and hybridization array data repository. Nucleic Acids Res 2002;30:207-10.
11. Brazma A, Robinson A, Cameron G, Ashburner M. One-stop shop for microarray data. Nature 2000;403:699-700.
12. Bassett DE Jr, Eisen MB, Boguski MS. Gene expression informatics: It's all in your mine. Nat Genet 1999;21(1 Suppl):51-5.
13. Kellam P. Microarray gene expression database: Progress towards an international repository of gene expression data. Genome Biol 2001;2:REPORTS4011.
14. A guide to microarray experiments-an open letter to the scientific journals. Lancet 2002;360:1019 author reply.
15. Ball CA, Brazma A, Causton H, Chervitz S, Edgar R, Hingamp P, *et al*. Submission of microarray data to public repositories. PLoS Biol 2004;2:E317.