A Flexible, Open, Decentralized System for Digital Pathology Networks

Robert SCHULER[a,1] and David E. SMITH[a] and Gowri KUMARAGURUPARAN[a] and Ann CHERVENAK[a] and Anne D. LEWIS[b] and Dallas M. HYDE[c] and Carl KESSELMAN[a]

[a] *Information Sciences Institute, University of Southern California*
[b] *Oregon National Primate Research Center, Oregon Health & Science University*
[c] *California National Primate Research Center, University of California, Davis*

Abstract. High-resolution digital imaging is enabling digital archiving and sharing of digitized microscopy slides and new methods for digital pathology. Collaborative research centers, outsourced medical services, and multi-site organizations stand to benefit from sharing pathology data in a digital pathology network. Yet significant technological challenges remain due to the large size and volume of digitized whole slide images. While information systems do exist for managing local pathology laboratories, they tend to be oriented toward narrow clinical use cases or offer closed ecosystems around proprietary formats. Few solutions exist for networking digital pathology operations. Here we present a system architecture and implementation of a digital pathology network and share results from a production system that federates major research centers.

Keywords. Digital pathology, digital pathology network, whole slide image, virtual microscopy, DICOM, data federation, data integration

Introduction

Digital pathology is an emerging technology enabled by high fidelity digital imaging and is transforming many sciences, including veterinary and human anatomical pathology, neuroanatomy, oncology, immunology, pharmaceutical research and other clinical and research pursuits. Also known as virtual microscopy, it encompasses *real-time* virtual microscopy, where a microscope is operated remotely, and *static* virtual microscopy, where glass slides are digitized into virtual slides. High-resolution slide scanners digitize the slides at magnifications from 20x to over 100x [1]. Even at 40x magnification, file sizes can exceed 20 Gigabytes (GBs) for a typical whole slide image (WSI), putting stress on the storage and computing resources of pathology departments. Despite the inherent challenges of managing large volumes of virtual slides, digital pathology holds great potential for clinical workflows, collaborative science, educational and training resources, and multi-site organizations.

In addition to standalone virtual microscopy systems, researchers and clinicians are now seeking to develop *digital pathology networks* in order to share virtual slides and facilitate the activities of collaborators. In this paper, we describe a distributed collaborative system developed for pathologists who must manage and share large

[1] Corresponding Author.

numbers of virtual slides as well as metadata associated with virtual slides, such as details about the subject demographics, the specimen, organs and organ systems, disease and etiology, and scanner device settings. This distributed pathology system was developed by researchers and engineers from the Biomedical Informatics Research Network (BIRN) [2] in close collaboration with the National Nonhuman Primate Research Consortium (NHPRC). Key contributions of this system include: 1) flexible data model and ontology support for pathology data and metadata; 2) a decentralized software infrastructure that supports federated query mediation across a digital pathology network while retaining local administrative control of each site; 3) an image processing service that converts images and annotations from proprietary scanner formats into open formats to facilitate data sharing; and 4) a testbed implementation operated by two of the National Primate Research Centers (NPRCs) with a data model for pathology data and metadata developed in cooperation with primate pathologists.

In the next section, we discuss the major obstacles for implementing digital pathology networks. In Section 2, we present the system architecture and implementation of a solution for digital pathology. In Section 3, we share early results from the deployment of the system in a production environment for sharing virtual slides and related pathology data for education and training purposes between two large nonhuman primate research centers. Section 4 describes related work on pathology and medical imaging systems. We conclude with a discussion of our ongoing and future work.

1. Challenges for Digital Pathology Networks

Given the technology considerations for managing large digital pathology collections, it can be a significant challenge to implement a digital pathology network for sharing whole slide images and accompanying information.

1.1. Diverse Applications

The applications of digital pathology vary widely. For example, consider the differences between immunology research involving animal subjects and clinical oncology involving human patients. Since the applications are diverse, the data captured by digital pathology systems vary greatly from one instance to the next. Each application has its own data schema constrained by numerous ontologies for diseases, clinical terms, species, and other data elements. In addition to different data types, the system must expose user interfaces and data curation workflows that fit the unique data models and ontologies of the application. Since there is no single standardized data schema or ontology to support all possible applications in pathology, it is challenging to design a system that is specific enough to complement pathologists' existing procedures while being reusable in different domains or even at different sites within a single domain. Additionally, the requirements for data schema and workflows change overtime as pathologists refine their data processes and gain experience using digital pathology methods. This requires digital pathology systems to be flexible enough to adapt to these changes without continually reengineering the system.

1.2. Proprietary Systems

Pathologists have limited options for infrastructure support for implementing open digital pathology networks. Most of the attention of information technology vendors is placed on standalone clinical pathology systems, while hardware vendors of digital scanners tend to offer closed ecosystems based on proprietary imaging and annotation formats, proprietary application programming interfaces (APIs), and proprietary image viewers. Closed ecosystems prevent sharing of data between hospitals and laboratories, as each site makes independent procurement decisions and acquires slide scanners from different vendors, which produce their own proprietary virtual slide formats. Recently, the Digital Imaging and Communications in Medicine (DICOM) Working Group 26 (Pathology) published DICOM Supplement 145 for Whole-Slide Imaging [3]. While this is a significant step toward open data formats for digital pathology, it may take some time for interoperability to be fully realized in practice. At present, pathologists must contend with proprietary images produced by previously acquired digital scanners and with existing digital archives of proprietary images that predate the new DICOM standard. Further exacerbating the challenge, we have found that pathologists naturally wish to exploit the full potential of their hardware investment by using the proprietary formats supported by the vendor, rather than be limited to the lowest common denominator of features supported by the standardized formats. Pathologists may prefer to use proprietary formats within their own labs in order to exploit fully their chosen technology and then convert a subset of their image collections to standard formats for the purpose of sharing within the digital pathology network.

1.3. Security Concerns

Digital pathology involves highly sensitive data sets. When human patients or research subjects are involved, the data consists of protected health information (PHI) and its usage and sharing must satisfy Health Insurance Portability and Accountability Act (HIPAA) constraints in clinical settings or Institutional Review Board (IRB) policies in research settings. In other domains where HIPAA and IRB rules do not apply, data security and privacy may still be required. These security issues and other factors drive the need for local administrative control. Hospitals and laboratories are reluctant to give up control over their data by submitting it to a third party data warehouse to facilitate the networking of digital pathology. In addition, hospitals and laboratories are equally restrictive when providing access to internal medical, laboratory, and pathology information systems because of the potential damages that could be caused by malicious access and other security breaches. As such, systems for digital pathology networks must be decentralized and allow local administrative control.

Next we present a system that addresses the challenges to digital pathology as outlined above.

2. System Architecture and Implementation for Digital Pathology Networks

In this section, we describe a flexible, open, and decentralized system architecture and implementation for digital pathology networks.

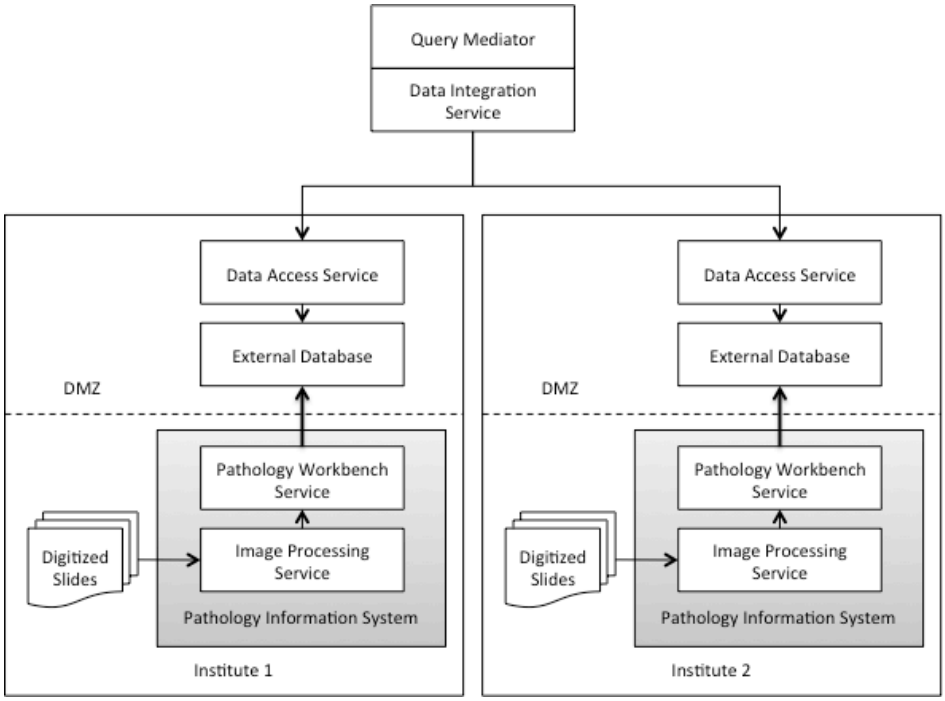

Figure 1. System architecture of a digital pathology network.

As seen in figure 1, the system architecture is decentralized and loosely coupled. Each site operates its own internal pathology information system, which manages local archives of whole slide images and associated pathology information. The local sites are operated within the boundaries of secure networks and protected by institutional firewalls. The local system is responsible for data ingestion from digital scanners and third party systems; data curation workflows for pathologists; search and retrieval over local archives; and data publication of approved subsets for sharing with the network. Next, each site in the network operates a database and data access service. The internal pathology information system publishes approved data by extracting approved subsets of the collection, stripping off fields intended for internal use only, packaging images with metadata records, and pushing the data package out to the data access service. The data access service is operated by each site but hosted outside of the institutional firewall in a network environment commonly known as the demilitarized zone (DMZ). Unlike the local pathology information systems, the data access services accept secure network connections from external clients on the digital pathology network. The data integration service operates from a (potentially) third-party location on the digital pathology network. The data integration service does not centrally manage a database of digital pathology records, as a traditional data warehouse would. Instead, when a client issues a query, the data integration service contacts the distributed data access services in real-time. This loosely coupled integration preserves the local administrative control of each site on the digital pathology network.

The system authenticates users based on username and password credentials, which are managed by third party identity providers (IdPs), such as enterprise class directory services. The distributed Grid services, such as data integration services and data access services, authenticate themselves using the Grid Security Infrastructure

(GSI) [4], a cryptographically-strong certificate based authentication protocol. A Certificate Authority (CA) issues GSI credentials to each of the services and clients in the digital pathology network, thus establishing the "chain of trust" in the system.

2.1. Pathology Information System

Next, we describe the implementation of the pathology information system, which consists of the Pathology Workbench service and the image processing service. These services are responsible for managing local pathology information and preparing data for sharing with the digital pathology network.

2.1.1. Pathology Workbench Service

The BIRN Pathology Workbench is a secure web-based application that provides flexible data schema and ontology management along with open protocols for integrating with third-party systems. It allows pathologists to enter case, specimen, and disease metadata into the system using their own vocabularies. Data entry can be done manually using a workflow-based web user interface, via batch processing using a spreadsheet import utility, or programmatically via a Representational State Transfer (REST) [5] web service interface. The metadata can then be assigned to images on the system and later used when searching for images based on keywords or metadata-specific queries. The Pathology Workbench also allows pathologists to instantly examine their images through the web browser, up to the same high-resolution of the original scan. Pathologists can also view and add graphical annotations to regions of interest on the image. These web-based capabilities allow pathologists to easily examine slides remotely with only minimal client-side software requirements. Finally, pathologists can select which cases to publish to external sites. These cases, along with all relevant metadata and images, can then be exported to other systems for multi-site collaboration or educational use.

The development of the Pathology Workbench depended on an open source Object Relational Mapping (ORM) and Model View Controller (MVC) framework. We extended the framework to dramatically enhance its flexibility for supporting changes to the data schema and ontology, such that changes to the ORM layer are automatically reflected in the full text search facility, the spreadsheet import mechanism, and the REST web service interface.

The development of the application was also driven by security requirements. The web-interface, image content, and REST web service are protected by strong authentication methods, which can be administered internally or in an institutional Lightweight Directory Access Protocol (LDAP) identity provider - giving sites the option of sharing user accounts across multiple sites. Access control authorization is enforced for users and groups at the application level for administration and data curation access. We also introduced fine grained access control at the image-level to determine which users or groups can view or write annotations on an image. Pathologists can share their annotations or keep them private. Shared annotations can be edited and merged by authorized users.

2.1.2. Image Processing Service

The image processing service is responsible for extracting thumbnails and acquisition metadata from images, performing image conversions from proprietary formats to either DICOM or a format that is supported by a common viewer, translating annotations from proprietary formats to an open format, and importing images into the Pathology Workbench.

2.1.2.1. Image and Metadata Processing

The image processing service schedules and automates the workflow to process images and import them from a designated location into the Pathology Workbench. The workflow utilizes the open source Bio-Formats library [6] to read image data and metadata from many proprietary microscopy formats. Information from the original image, such as a thumbnail image and acquisition metadata, is extracted and cataloged with the original image in the Pathology Workbench. Then the images are converted into a pyramidal image structure that can be viewed in a Zoomify web-based image viewer (http://www.zoomify.com) using progressive rendering such that the viewer only retrieves the visible portion of the image. Progressive rendering is critical given the large size of the images. Optionally, lower-resolution pixel data can also be exported to a DICOM file.

2.1.2.2. Annotation Processing

Support for annotations is an absolute necessity for digital pathology solutions. During the image conversion workflow, annotations are translated from proprietary formats to an eXstensible Markup Language (XML) format supported by the Zoomify viewer. Essential parts of an image annotation are point or region of interest (ROI), annotation shape (i.e., arrows, circles, boxes etc.) and annotated text. Proprietary annotation formats specify coordinates using different coordinate systems and scales. The image processing service converts an annotation from a proprietary format by aligning these essential components. This step involves mapping points of interests by transforming coordinates from one scale to another, identifying closest matching shape in Zoomify that can represent the same information, and extracting the annotated text. This logic is then coded in XML Query (XQuery) to transform XML annotations from proprietary formats to the Zoomify-compatible format.

2.2. Query Mediation and Data Integration Services

The BIRN Mediator [7] plays a critical role in creating a decentralized digital pathology network by federating data from autonomous digital pathology systems. Unlike data warehouses, which store copies of data centrally, the mediator integrates data in real-time in response to query requests. Individual sites in the network share only the data that they are willing to expose, yet they maintain control over the source data at all times. This approach also allows pathology departments to define their own distinct data schema and apply their chosen ontologies rather than adopting an external data model. The BIRN Mediator facilitates real-time data integration based on a Global-As-View (GAV) data federation methodology implemented within the OGSA-DAI [8] Distributed Query Processing (DQP) [9] engine. Primary data remains at the sources, and thus the autonomy and local control of individual sites is preserved.

A federated query issued over the global domain schema is rewritten as a set of source schema sub-queries. The DQP engine evaluates the queries, creates an execution

plan and submits sub-queries to the databases. The query responses are joined or merged and returned to the client. The mediator also supports data-value mapping between sources and the global domain, allowing data integration of sites with different vocabularies or ontologies.

3. Results from a Digital Pathology Network Testbed

The digital pathology network has been deployed at two NRPC sites with an independently operated data integration service. The Oregon NPRC, the first deployment of the system, currently has over 27,000 images (~1.5TB) tiled and imported into their system, with the largest image having dimensions of 148480 x 293888 pixels and 15 GB in size. Pathologists have successfully imported all subject information into the database and are currently in the process of entering case data for these images and exporting published case data to external sites. The California NPRC has also deployed the system and is processing images and importing data from other laboratory information systems using the spreadsheet import interface.

One of the key contributions of our work was the development of a detailed pathology data model, shown in figure 2. This model was developed over a period of 18 months using an iterative process that included the following steps: 1) discussions between pathologists and system designers identified concepts and relationships that needed to be included in the schema and ontologies; 2) BIRN staff rapidly implemented database prototypes based on the revised data model using the system's flexible framework; and 3) pathologists extensively tested the prototype database and provided detailed feedback to developers, specifying new features needed in the next phase of development.

The core data model was developed by defining relationships among essential concepts, namely, *Subject*, *Case*, *Clinical diagnosis*, *Histological diagnosis*, *Specimens* and *Images*. *Specimens* represent samples taken from a *Subject* during a *Procedure* for a given *Case*. The specimen can be the organs or tissues from which gross images are derived or physical specimens from which histologic slides are generated and then imaged. In the latter case, histologic diagnoses are derived from the evaluation of the specimen. A *Procedure* represents any type of surgery, biopsy or necropsy done on a subject, and *Diagnoses* are performed to identify diseases. Along with the disease, its etiologic agents, disease processes and morphologies are recorded. For each image, metadata are collected, including the image type (gross or histological), the scanner device profile, image annotations and other details about the image. The data model references ontological codes from standard or custom ontologies. Figure 2 illustrates just a subset of the full data model, which spans dozens of entities and relationships.

Currently, we are working to expand the testbed beyond the initial deployment. We are also streamlining the data curation workflows and related user interfaces based on detailed feedback from the pathologists using the system. Finally, we are enhancing the image processing automation with an open source messaging framework.

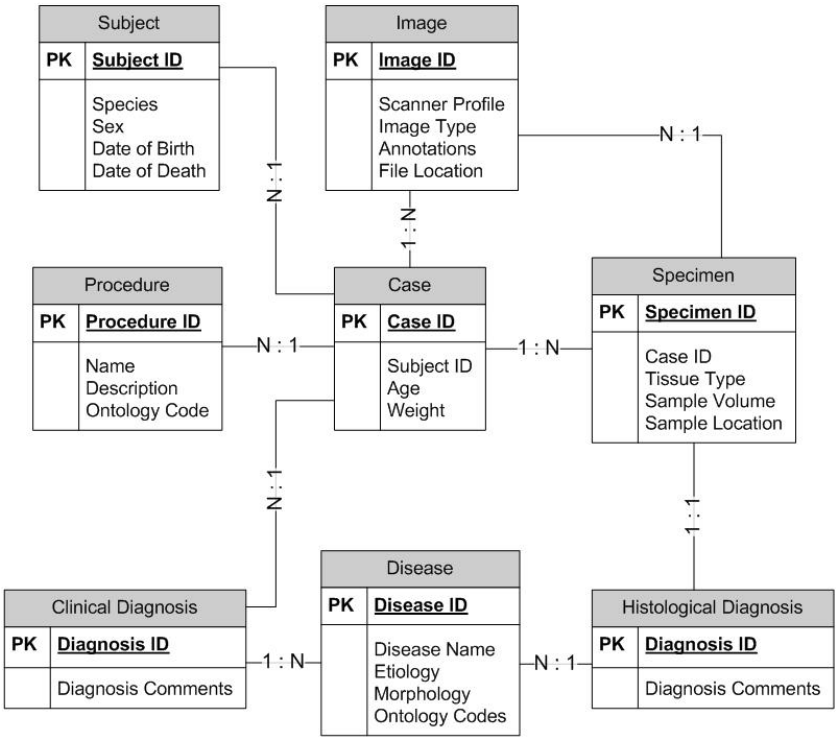

Figure 2. A representative subset of the data schema used in the testbed deployment.

4. Related Work

Numerous online digital pathology resources exist. The European mutant mouse pathology database (Pathbase) [10] provides an online resource for histopathology images derived from mutant or genetically manipulated mice. The University of Connecticut virtual Pathology Museum (PathWeb) [11] provides an online database of low power virtual slides with clinical descriptions. The Zebrafish Atlas [12] provides anatomical reference slides of the zebrafish online for research and education. Brainmaps.org [13] is a large online multi-species neuroanatomical resource for reference and educational purposes. The Pathology Education Instructional Resource (PEIR) Digital Library [14] and the Internet Pathology Laboratory for Medical Education (WebPath) [15] offer instructional resources including gross and microscopic pathology for health science education. While these services represent important online resources for digital pathology, they do not claim to offer reusable infrastructure for implementing digital pathology networks.

The Cancer Biomedical Informatics Grid (caBIG) provides the caTissue Suite [16] for bio-specimen inventory management, tracking, and annotation. A large part of caTissue's functionality is related to inventory management, tracking specimens across storage repositories and keeping track of their storage details like containers, concentration and quantity. Specimen management plays a critical role in the internal pathology information systems of an organization; however, caTissue does not offer

facilities for creating digital pathology networks with whole slide image and annotation sharing.

The Open Microscopy Environment (OME) [17] develops software and standards for virtual microscopy. The OME Remote Objects (OMERO) Platform [6] is a client-server tool for managing microscopy images. With OMERO, users can import images from major microscopy formats, organize image collections by tags and other metadata, analyze images using Python scripts, view images using standalone and web-based viewers, and export images and data for external usage. OMERO is designed as a standalone virtual slide server and as such does not directly provide facilities necessary to create digital pathology networks.

Neuroimaging systems have been developed to address the needs of neuroscientists, particularly in Functional Magnetic Resonance Imaging (fMRI) and related modalities. The eXtensible Neuroimaging Archive Toolkit (XNAT) [18] and the Human Imaging Database (HID) [19], [20], both developed in association with the BIRN project, address many challenges in neuroimaging projects. XNAT was built on an open source Picture Archiving and Communication System (PACS) and supports DICOM medical imaging formats and protocols for medical imaging but not yet for whole slide images required by microscopy. The HID was built for multi-site neuroimaging applications and, like XNAT, supports low resolution image modalities.

Virtual slide scanner vendors offer web-based (e.g., Aperio ViewPort Control, Olympus WebSlide Browser) and client-side (e.g., Aperio ImageScope, Olympus OlyVIA, Hamamatsu NDP.view) software tools that can progressively view and annotate full-resolution whole slide images. Generally, these tools are limited in the web browsers and operating systems they support, and they tend to favor their own propriety image formats as discussed in earlier sections. Aperio has strongly advocated open standards, participated in the DICOM standards working group, and offered a centralized architecture for multi-site digital pathology [21]. In contrast, our system architecture is loosely coupled, decentralized and preserves local administrative control.

5. Conclusion

We have presented a systems architecture and implementation for supporting digital pathology networks that address a general set of needs shared by research pathologists, clinical pathologists, and scientists in other domains who rely on digital pathology methods. The system satisfies the requirements for high-resolution whole slide imaging, image annotations, complex pathology metadata and data management workflows, and decentralized deployment scenarios where security is paramount. We have deployed the system and tested it in production at two large-scale research centers with active use by pathologists and related research and support staff.

6. Acknowledgments

This work was supported in part by the NIH through the NCRR grant Biomedical Informatics Research Network (NIH U24 RR025736-01), the BIRN Community Service Award (NIH U24 RR026057), the BIRN-CC supplemental award (NIH 3U24RR025736-01S1), and “Support for National Primate Research Center” (NIH SP51 RR000163).

References

[1] S. Mikula, I. Trotts, J. M. Stone, and E. G. Jones, "Internet-enabled high-resolution brain mapping and virtual microscopy," *NeuroImage*, vol. 35, no. 1, pp. 9-15, 2007.

[2] K. G. Helmer et al., "Enabling collaborative research using the Biomedical Informatics Research Network (BIRN).," *Journal of the American Medical Informatics Association : JAMIA*, vol. 18, no. 4, pp. 416-422, Jul. 2011.

[3] DICOM Standards Committee, Working Group 26, "Digital Imaging and Communications in Medicine (DICOM): Supplement 145: Whole Slide Microscopic Image IOD and SOP Classes." NEMA, Rosslyn, Virginia, pp. 1-59, 2010.

[4] V. Welch et al., "Security for Grid services," in *High Performance Distributed Computing, 2003. Proceedings. 12th IEEE International Symposium on*, pp. 48-57.

[5] R. T. Fielding and R. N. Taylor, "Principled design of the modern Web architecture," *ACM Transactions on Internet Technology*, vol. 2, no. 2, pp. 115-150, May 2002.

[6] J. R. Swedlow, I. G. Goldberg, and K. W. Eliceiri, "Bioimage informatics for experimental biology," *Annual review of biophysics*, vol. 38, pp. 327-46, Jan. 2009.

[7] N. Ashish, J. L. Ambite, M. Muslea, and J. A. Turner, "Neuroscience data integration through mediation: an (F)BIRN case study," *Frontiers in Neuroinformatics*, vol. 4, no. 118, 2010.

[8] M. Antonioletti et al., "The design and implementation of Grid database services in OGSA-DAI," *Concurrency and Computation Practice and Experience*, vol. 17, no. 2-4, pp. 357-376, 2005.

[9] M. N. Alpdemir et al., "OGSA-DQP: A Service for Distributed Querying on the Grid," *Proceedings of the First International Conference on Service Oriented Computing ICSOC 2003*, no. 2910, pp. 467-482, 2003.

[10] P. N. Schofield et al., "Pathbase: a database of mutant mouse pathology.," *Nucleic acids research*, vol. 32, no. Database issue, pp. D512-5, Jan. 2004.

[11] L. I. Cheng, "Ode to Pathology Images Online¡," *Toxicologic Pathology*, vol. 35, no. 4, pp. 618-619, Jun. 2007.

[12] N. A. Sabaliauskasa et al., "High-throughput zebrafish histology," *Methods*, vol. 39, no. 3, pp. 246-254, 2006.

[13] E. G. Jones, J. M. Stone, and H. J. Karten, "High-resolution digital brain atlases: a Hubble telescope for the brain.," *Annals of the New York Academy of Sciences*, vol. 1225, pp. E147-59, May 2011.

[14] K. N. Jones, R. Kreisle, R. W. Geiss, J. H. Holliman, P. H. Lill, and P. G. Anderson, "Group for Research in Pathology Education Online Resources to Facilitate Pathology Instruction," Oct. 2009.

[15] J. M. Spak, "WebPath: The Internet Pathology Laboratory for Medical Education on CD-ROM," *Bulletin of the Medical Library Association*, vol. 88, no. 2. Medical Library Association, p. 205, 01-Apr-2000.

[16] A. Brink, D. Mulvihill, L. Jackel, and D. Rashmi, "caTissue Suite v1.2: User's Guide," St. Louis, MO, 2010.

[17] I. Goldberg, "Open Microscopy Environment," in *2005 IEEE Computational Systems Bioinformatics Conference - Workshops (CSBW'05)*, pp. 380-380.

[18] D. S. Marcus, T. R. Olsen, M. Ramaratnam, and R. L. Buckner, "The Extensible Neuroimaging Archive Toolkit: an informatics platform for managing, exploring, and sharing neuroimaging data," *Neuroinformatics*, vol. 5, no. 1, pp. 11-34, 2007.

[19] D. B. Keator, "Management of information in distributed biomedical collaboratories.," *Methods in molecular biology (Clifton, N.J.)*, vol. 569, pp. 1-23, Jan. 2009.

[20] D. B. Keator et al., "A national human neuroimaging collaboratory enabled by the Biomedical Informatics Research Network (BIRN).," *IEEE transactions on information technology in biomedicine : a publication of the IEEE Engineering in Medicine and Biology Society*, vol. 12, no. 2, pp. 162-72, Mar. 2008.

[21] S. J. Potts, "Digital pathology in drug discovery and development: multisite integration.," *Drug discovery today*, vol. 14, no. 19-20, pp. 935-41, Oct. 2009.