

DATA ARCHIVING AND DISTRIBUTION OF LiDAR AND DERIVED DATASETS IN THE PHILIPPINES

M. E. A. Tupas ^{a*}, S. C. Lat ^b, R. A. Magturo ^b

^a University of the Philippines, Diliman, Quezon City, 1001, Metro Manila, Philippines - matupas@up.edu.ph

^b Phil-LiDAR 1 Program, University of the Philippines Training Center for Applied Geodesy and Photogrammetry (UP TCAGP),
Department of Science and Technology (DOST) - University of the Philippines, Diliman, Quezon City, 1001, Metro Manila,
Philippines,

Commission IV, WG IV/4

KEY WORDS: LiDAR, Geospatial Data Infrastructure, Data Archiving and Distribution, Geospatial Database, Web Standards

ABSTRACT:

LiDAR programs in the Philippines have been generating valuable resource and hazard information for most of the country at a substantial rate since 2012. Significant progress have been made due to the programs design of engaging 16 Universities and research institutions spatially distributed across the country. Because of this, data has been accumulating at a brisk rate which poses significant technical and logistic issues. While a central node, the University of the Philippines, Diliman, handles data acquisition, pre-processing, and quality checking, processing and ground validation are devolved to the various nodes. For this setup to be successful, an efficient data access and distribution system should be in place.

In this paper, we discuss the spatial data infrastructure and data access protocols implemented by the program. At the center of the data access and distribution operations is LiPAD or our LiDAR portal for archiving and distribution. LiPAD is built on open source technologies, established web standards, and protocols. At its back-end a central data archive has been established using state of the art Object Storage technology to store both raw, processed Lidar and derived data sets. Catalog of available data sets ranging from data acquisition foot prints, to DEM coverages, to derived products such as flood hazard, and crop suitability are viewable and accessible on the main site based on the popular GeoNode application. Data exchange is performed using varying protocols to address various logistical problems. Given the various challenges the program is successful in distributing data sets not just to partner processing nodes but to other stakeholders where main requesters include national agencies and general research and academic institutions.

1. INTRODUCTION

The Philippines, an archipelagic nation abundant in both terrestrial and marine natural resources, is one of the few countries situated within both the Pacific Ring of Fire and Pacific Typhoon Belt. With its geographic location and physical environment, it is highly susceptible to various natural hazards. Hence, detailed high resolution mapping of its terrain and topography for resource and hazard assessment is critical.

Light Detection and Ranging (LiDAR) makes it possible to collect billions of individual point measurements of the earth's surface (Crosby et al., 2011). This point cloud measurements enable us to map out topography and characterize features on the ground. These high-accuracy datasets can then be applied to various applications, including floodplain mapping, hydrology, geomorphology, forest inventory, urban planning, and landscape ecology (Chen, 2007). For these reasons it was deemed necessary to use LiDAR to map the Philippines.

In 2012, DOST and University of the Philippines, Diliman (UPD) started the Disaster Risk and Exposure Assessment for Mitigation (DREAM) LiDAR program. From the DREAM program high value products such as high resolution LiDAR datasets, Orthophotographs, Hyperspectral images, and Digital Elevation Models (DEMs), were used to generate Flood Hazard Maps and has successfully covered 1/3 of the total area of Philippine river systems, equivalent to 100,000 sq.km., or 18 major river basins. With DREAM's initial success, DOST

expanded on its goals and coverage by introducing two concurrent programs to succeed it, namely Phil-LiDAR 1 to continue flood hazard mapping and Phil-LiDAR 2 for various resource assessments.



Figure 1. Implementing Institutions and Area Coverages

The two Phil-LiDAR programs aims to cover the remaining 2/3 of the total area of the Philippines in a three year span. To be able to accomplish this, sixteen other spatially distributed and autonomous Higher Education Institutions (HEIs) and State

Universities and Colleges (SUCs) (shown in Figure 1.) have also been engaged. This was done with the goal of distributing the processing workload and leverage local expertise for validation and calibration by devolving these tasks to the HEIs assigned to various geographical areas. This is supported by data acquisition, pre-processing, and quality checking which remains centrally operated by UPD.

However, LiDAR datasets or these billions of XYZ coordinates, 3D points, or point clouds results in massive data spatial datasets (Ackermann, 1999). In LiDAR acquisition and processing operations terabytes to petabytes of disk storage used is the norm, we accounting for raw, intermediate, and processed data. With this in mind an efficient storage and retrieval medium is needed to support processing and distribution operations.

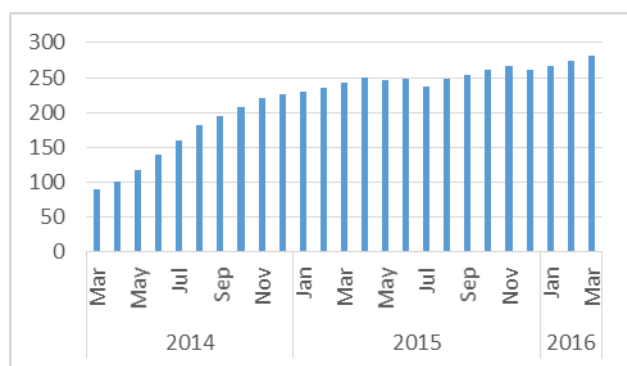


Figure 2. Data Growth Last 2 Years (Tb vs Time)

The data holdings of the three programs currently constitute two hundred ninety terabytes (290 Tb), with the data increasing at roughly eight terabytes per month of operation. Given the operations pipeline is being co-implemented with 16 SUCs/HEIs distributed across an archipelago, this poses significant logistic and technical challenges.

2. RELATED LITERATURE

2.1 LiDAR Data Storage and Archiving

The initial concern is to make sure these datasets are securely stored and can be accessed in a timely manner. A variety of LiDAR storage systems have been demonstrated to have these characteristics, these include Laserdata Information System by Rieg et al. (2013), and Open topography (Crosby et al., 2011, Nandigam et al, 2010) which utilizes a combination of Relational Database Management Systems and organized LiDAR flat files. While others implement Object Relational Database Management Systems with specialized spatial columns such as Lewis et al (2012) that uses Well Known Binary point geometry and Ramsay (2013) who uses a column optimized for storing point cloud clusters.

For faster access and retrieval of the datasets it is imperative to implement means to catalog and spatial index which are done either through 1) built-in DBMS (Lewis et al., 2012; Ramsay, 2013), 2) external DBMS (Crosby et al., 2011) or 3) file folder structured (Chen, 2007; David et al., 2008) When indexing is external to the data, such as when traditional RDBMSs are used, updates incur additional indexing overhead. Making updates computationally intensive and affect performance adversely as data size increases (Fox et al., 2013). This

produces scalability and query time issues (Al-Naami et al., 2014).

2.2 LiDAR Data Distribution

Proper storage and archiving is the initial concern, but with these amount of data, challenges are also encountered in data visualization, data analysis, and rapid data processing (Hungchao & Wang, 2009; Crosby et al., 2011; Lewis et al., 2012), because of these, development of such large spatial databases pose significant technical challenges in terms of the management and as well as web-based distribution (Nandigam et al., 2010; Lewis et al., 2012).

Moreover, for seamless use and processing of geographically referenced data such as those derived from LiDAR, a common understanding of how data is stored and written is needed. This is addressed, atleast for vector and raster datasets, by the use of standards such those provided by ISO TC 211 and OGC (Banks, 2004). And this should be extended to LiDAR as David et al. (2008) points out that for efficient LiDAR processing a standard data format is needed. For LiDAR data most utilized standard file format is LAS (the airborne LiDAR data exchange format) (2010). The data exchange format was devised as an open standard independent of the proprietary format derived from LiDAR data acquisition.

Taking the concept of standards further is the utilization of open web standards for inter-operable GIS data and Spatial Data Infrastructures. SDI interconnects GIS nodes across the World Wide Web to promote information sharing and access (Banks, 2004). SDIs enable spatial data sharing, cataloguing, access, and processing to its stakeholders. Steiniger and Hunter (2009) points out that SDIs can now be implemented with various freely available software.

To operationalize and sustain the use of LiDAR data among SUCs, HEIs, and other stakeholders throughout the country, the capability to reliably store, and efficiently transfer data and its derivatives is of utmost importance for an effective utilization of LiDAR in resource assessment and hazard mitigation.

3. METHODOLOGY

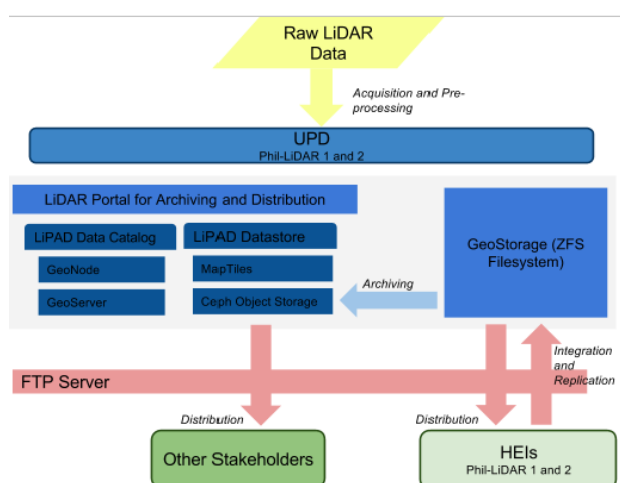


Figure 3. Conceptual Framework of Data Distribution and Replication

The end goal is the establishment of an efficient SDI for the Philippine LiDAR programs which can be extended for the whole country. For this the Data Archiving and Distribution (DAD) component of the program based in UPD currently performs the necessary tasks leading to this, these includes providing 1) spatial data storage or archiving, 2) spatial data cataloguing, and 3) spatial data access or distribution to enable efficient processing and validation between the various nodes of the program. At the center of SDI is our LiDAR Portal for Archiving and Distribution, fancifully called LiPAD (meaning flight in the Tagalog language). LiPAD integrates the above capabilities using proven scalable open-source technologies.

3.1 Data Storage and Archiving

The first task of DAD is to properly store and archive acquired, and processed, LiDAR datasets as well as other ancillary datasets. For data processing operations of raw LiDAR datasets tiled flat file organization based on acquisition hierarchies have been utilized. To centrally manage this, ZFS Samba and Windows file sharing was implemented with datasets grouped logically by river basins. However, due to performance limitations, a more advanced archiving solution has been implemented. While at the program level, to further ensure the security of data acquired an inherent redundancy is in place between the main node in UPD and partner HEIs as shown in Figure 3. The technical solution for archiving adopted by the program is a novel Object storage system.

Main considerations are Object storage's built-in features for redundancy, clustering, remote backup and remote data location (Mesneir et al., 2003). In an Object storage data can be re-organized indefinitely with minimal impact on data search. This allows the system to scale towards larger datasets with minimal impact on maintaining a database of all available data for distribution. For its implementation, DAD utilized Ceph (Inktank Storage, Inc., 2015), Ceph is a stand-alone object-storage solution that can be deployed without any dependency on any cloud-based technology. Ceph was chosen due to being open source, its compatibility with widely used cloud services and the broad spectrum of programming languages it supports. It runs on commodity hardware, and is designed to be self-healing and self-managing. Therefore Ceph is more usable for smaller, more customized deployments which is the case for Phil-LiDAR which do not have the luxury of large data centers.



Figure 4. Semi-automated Archiving and Metadata updating

Processes have been optimized for river basin and flight missions, to be consistent with end products such as flood hazard maps. To address storage and distribution needs, our implementation utilizes semi-automated workflows such as tiling (shown in Figure 4), caching, naming, re-projecting, among others. The functions are accessible on the LiPAD management interface.

3.2 LiDAR Data Catalog and Management

What is lacking from the Ceph implementation is the means to catalog and manage geospatial metadata. To provide

management and front end for the data uploaded into the object storage, a geospatial content management solution was utilized. For this purpose, GeoNode (GeoNode Development Team, 2013) is an open-source web-based application and platform for geospatial information systems (GIS) content management system written in Django. It is able to store and share rasters and vectors and uses GeoExplorer to display the managed data sets. GeoNode not only provides a web interface for upload, display, and download of vectors and rasters but also allows sharing of these data sets using Open Geospatial Consortium (OGC) standard web services. Additional features can be developed on any open-source platforms it uses.

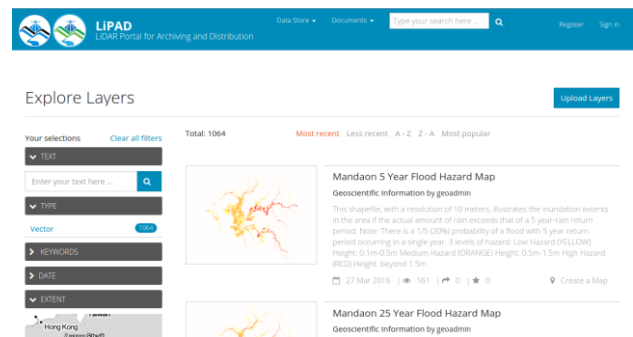


Figure 5. Flood Hazard Layers available in LiPAD

For end products such as resource layers and hazard layers, LiPAD leverages GeoNode's built-in capability to visualize, catalog and distribute both vector and raster layers either as flat file download or through OGC services.

Due to limited resources for processing and data preparation, the program adopts a delivery on demand approach on larger datasets. To which data is prepared for access and distribution only after a complete request is lodged, this is done to minimize data transfers due to the ever changing spatial data holdings. This system can only operate efficiently if a data catalog is available. The current data catalog of the program consists of metadata information and a series of coverage vector files. Coverage files are made available for acquired raw LiDAR, processed DEMs, and orthophotos. These files are regularly updated and shared with co-implementers for them to determine which datasets are available for their assigned areas. Figure 6 shows DEM coverages available in LiPAD.

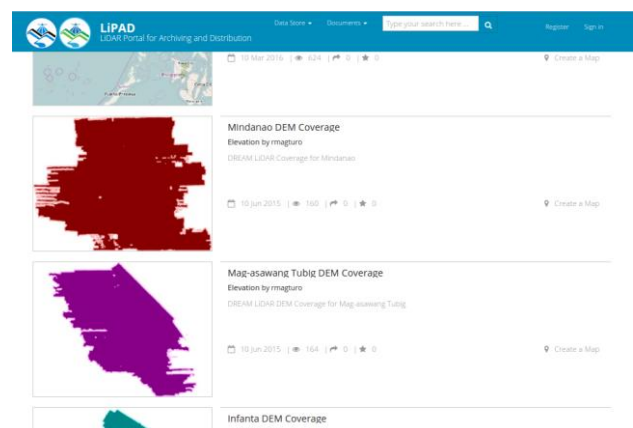


Figure 6. Vector Based DEM Data Coverage in LiPAD

LiPAD houses meta-data and coverage information. Various stakeholders are given credentials where limitations are set such as spatial coverage and data type restrictions can be enforced e.g. regional users can only view data within their assigned region. Also part of LiPAD is the end user facing interface for data distribution. This interface is similar to other LiDAR distribution websites utilizing a tiled approach; 1 x 1 km tiles are indexed and displayed on the portal. Figure 7 shows the tile-based selection interface which includes searching, querying and selection functionalities.

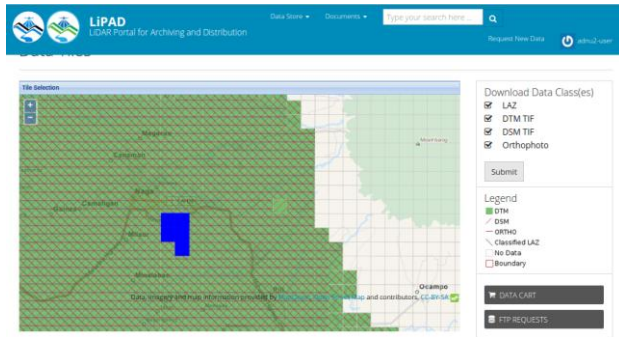


Figure 7. Tile Based DEM Data Coverages in LiPAD

These tiled datasets however are not available as http downloads but rather available using secure file transfer protocol (SFTP). This approach was used due to non-ideal internet connectivity in the country. This was done to be able to secure the transfers and enable resume for slow or dropping connections.

4. DISCUSSIONS

Given the various challenges, the program through the Data Archiving and Distribution Component is successful in distributing data sets not just to partner processing nodes but to other stakeholders where main requesters include national agencies and general research and academic institutions. From February 2013 to February 2016 the program has achieved 90.1% (273/303) successful distribution rate with the 9.9% of the requests are those not yet covered, not yet processed, and outside of jurisdiction. However, this initial rate was mainly from manually distribution using physical mediums such as DVDs and Hard drives.

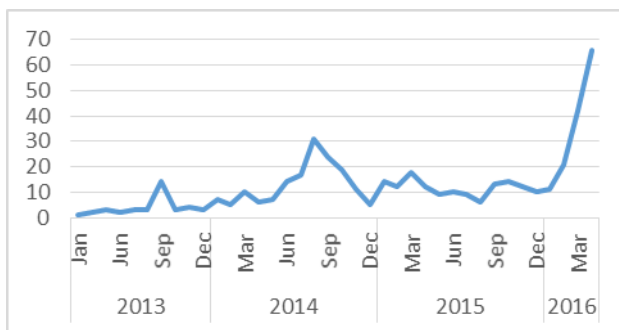


Figure 8. Data Requests per Month

The introduction of LiPAD is an initial step towards an SDI for LiDAR in the Philippines. LiPAD provides catalog services for the various datasets generated from the Phil-LiDAR programs. For end products the default capability of GeoNode for which

LiPAD is an extension of, is used for searching, querying, visualizing and downloading. However, for raw and intermediate datasets such as DEMs, and Point cloud datasets, a tiled catalog is provided. To provide the various datasets, varying protocols and standards are being utilized from OGC services to secure file transfer protocol.

While the front end and other management functions are available from the extended Django interface, the back end is connected to a novel Object Storage implementation using Ceph. Since the beta release of the LiPAD portal in January 2016, data requests from various partners and stakeholders has seen a significant uptick, as shown in Figure 8. This reflects as an initial success for the system with one of its goals to provide better access to the datasets generated by the program.

ACKNOWLEDGEMENTS

This work is supported by the Department of Science and Technology - Philippine Council of Industry, Energy, and Emerging Technology Research and Development (DOST-PCIEERD) under Nationwide Hazard Mapping using LiDAR Data (PHL-LiDAR 1 – Project 5: Data Archiving and Distribution) Program.

REFERENCES

- Ackermann, F., 1999. Airborne laser scanning—present status and future expectations. *ISPRS Journal of Photogrammetry and Remote Sensing* 54 (2-3), 64–67.
- ASPRS, 2010. LAS Specification, Version 1.3-r11. Available online: http://www.asprs.org/a/society/committees/standards/LAS_1_3_r11.pdf
- Banks, E., 2004. Importance of Open Spatial Data Infrastructure in Data Sharing. Pdf file retrieved from: <http://www.isprs.org/proceedings/XXXV/congress/comm4/papers/356.pdf>
- Chen, Q., 2007. Airborne lidar data processing and information extraction. *Photogrammetric Engineering & Remote Sensing*, 73(2), 91-95.
- Crosby, C.J., Arrowsmith, J R., Nandigam, and V., Baru, C., 2011. A Geoinformatics Approach to Online Access and Processing Of LIDAR Topography Data. In R. Keller and C. Baru, Eds., *Geoinformatics: Cyberinfrastructure for the Solid Earth Sciences*, pp. 251-265. London: Cambridge University Press.
- David, N., Mallet, C. & Bretar, F., 2008. Library concept and design for lidar data processing. In *Geographic Object Based Image Analysis (GEOBIA) Conference*, Calgary, Canada.
- Fox, A., Eichelberger, C., Hughes, J., and Lyon, S., 2013. Spatio-temporal indexing in non-relational distributed databases. 2013 IEEE International Conference on Big Data, pp. 291–299.
- GeoNode Development Team, 2013. About GeoNode — GeoNode 2.0 documentation. Retrieved July 2015, from <http://docs.geonode.org/en/master/organizational/about.html#about>

Hongchao, M., & Wang, Z., 2011. Distributed data organization and parallel data retrieval methods for huge laser scanner point clouds, *Computers & Geosciences*, v.37 n.2, p.193-201.

Inktank Storage, Inc., 2015. Welcome to Ceph — Ceph Documentation. Retrieved July 2015, from <http://ceph.com/docs/master/>

Lewis, P., Mc Elhinney, C. P., McCarthy, T., 2012. LiDAR data management pipeline; from spatial data population to web-application visualization. In: *Proceedings of the 3rd International Conference on Computing for Geospatial Research and Applications, COM.Geo 2012 (pp. 10)*, New York, NY, USA

Mesnier, M., Ganger, G. R., and Riedel, E., August 2003. Object-Based Storage. *IEEE Communications Magazine*, pp. 84–90.

Nandigam, V., Baru, C., Crosby, C. J., 2010. Database Design for High-Resolution LIDAR Topography Data. SSDBM, volume 6187 of *Lecture Notes in Computer Science*, page 151-159.

Ramsey, P., 2013. LIDAR in PostgreSQL with PointCloud. Available online: <http://boundlessgeo.com/wp-content/uploads/2013/10/pgpointcloud-foss4-2013.pdf>

Steiniger, S., and Hunter, A.J.S., 2009. Free and open source GIS software for building a spatial data infrastructure. *Geospatial Free and Open Source Software in the 21st Century: Proceedings of the first Open-Source Geospatial Research Symposium, 2009, LNG&C, Springer, Heidelberg*, pp. 247-261.