Open source digital data collection for field sciences.

Isaac I. Ullah, PhD
San Diego State University

Why Open Science?

"Often described as **'open science'**, these new norms include **data stewardship** instead of data ownership, **transparency in the analysis process** instead of secrecy, and **public involvement** instead of exclusion. ... We believe that there is much to be gained, both for individual researchers and for the discipline, from broader application of open science practices. ... [W]e have identified three elements of open science that cross-cut [these] themes: **open access**, **open data**, and **open methods.**" (emphasis added)

- Ben Marwick + 48 coauthors, "Open Science in Archaeology." *SAA Archaeological Record,* September, 2017.

SAA Open Science Interest Group.
Wiki: **https://osf.io/2dfhz/**

Digital Data Collection as an Open Method

The benefits of paperless technology for the field sciences are obvious:

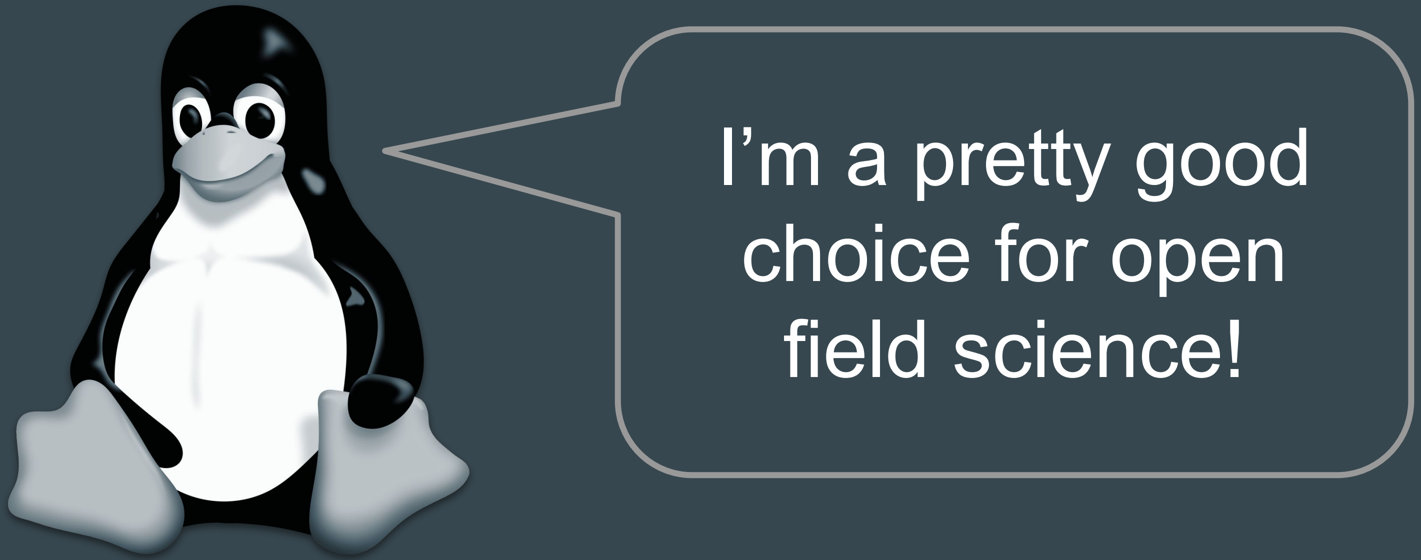
1. Faster in-field collection, little to no need for “post production” of data.
2. Better data standardization, fewer recording and transcription errors.
3. Ability to “course correct” based on real-time data analysis.
4. Near-instant upload, backup, and real-time long-distance collaboration.

However, with the adoption of this new technology, we have an important opportunity choice. Will we support the idea of Open Science by ensuring our paperless data collection workflows are:

1. As transparent as possible so that potential errors can be assessed?
2. Reproducible by others, right down to the hardware components?
3. Freely scalable and changeable so that they can responsively grow along with our disciplinary needs?
4. Available to all, regardless of income, location, or institutional support.

Open Science, Digital Data, and Linux

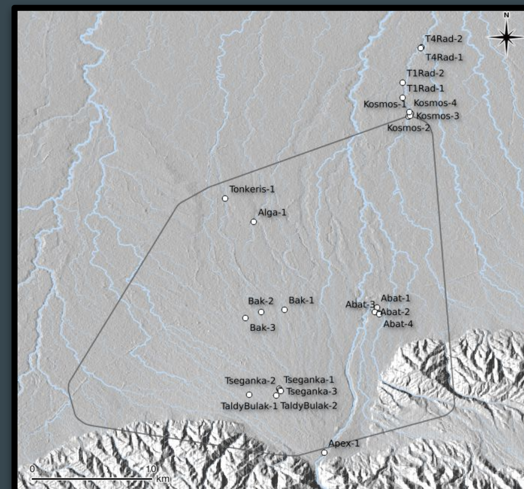
If we are really serious about meeting these four objectives, *everything* about our data needs to be **open**. That includes methods for gathering, storing, manipulating, analyzing, and disseminating those data, right down to the source code of the software(s) that was used to do everything. **Our choice of operating system is an essential part of this chain, but one that is perhaps not frequently considered by *field scientists*.**

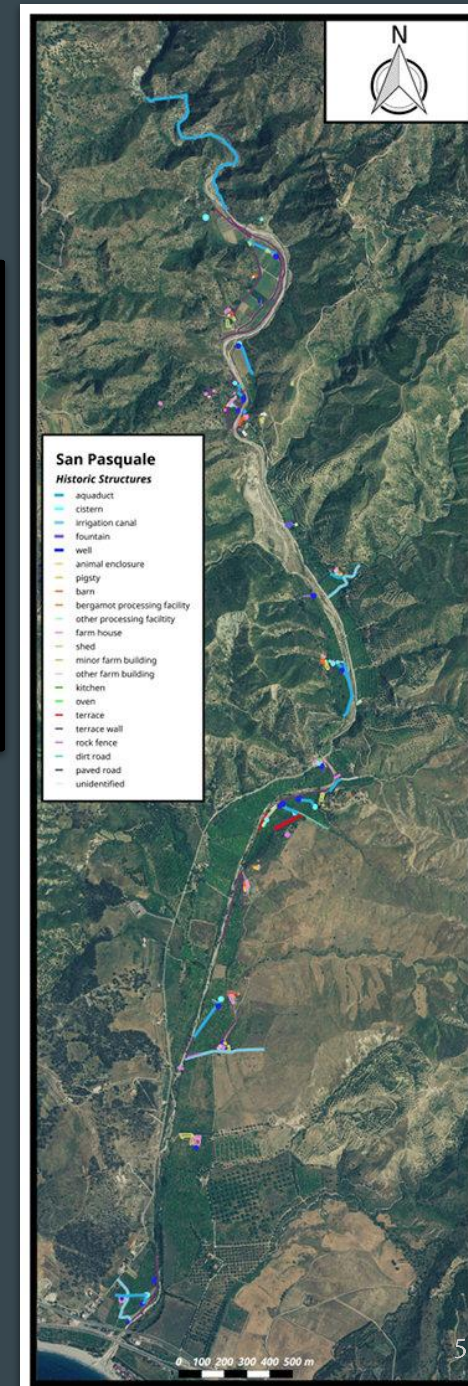


The BMAP and KAAE projects.

I will use these two case studies to exemplify an open-source approach to:

1) The field data collection workflow
2) Post field-work "data hygiene"
3) Data curation and dissemination

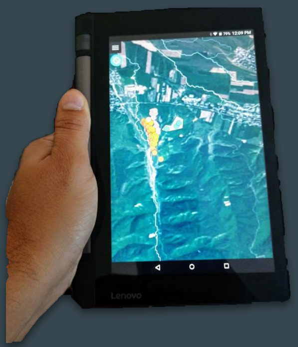




Field data collection workflow.

The hardware used in this workflow is under $4500 USD

8” Android Tablet - Lenovo Yoga Tab ($150 x 4)



- Android offers more Open Source choices. These tablets are cheap, readily available, and easy to replace.

Bluetooth GPS - Bad Elf GNSS Surveyor ($600 x 4)

- These provide ~1m accuracy (or better). The Bad Elf company makes their hardware Open Source, and provides an API. More affordable than comparable Trimble™ products.

Quadcopter - DJI Mavic Pro + 2 batteries ($1200)

- DJI drones are ubiquitous, high quality, and affordable. They offer an API for third-party software controller options. The Mavic Pro includes a high-resolution, stabilized camera, FPV video, long flight time, and compact folding design.

Mobile Data Collection

Open Data Kit and the fork, “GeoODK,” allow easy creation of custom forms that are deployable on multiple mobile devices via the ODK Collect app.

1) Build Form in spreadsheet

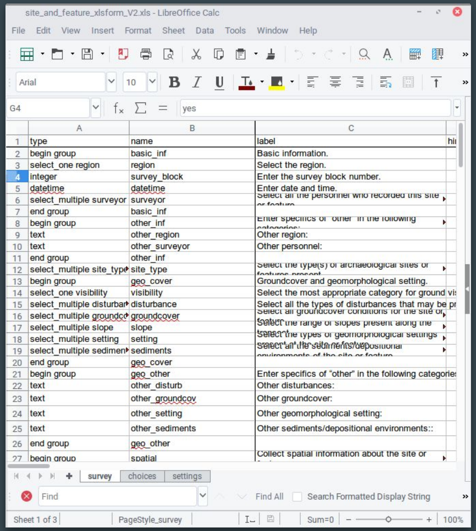
Libreoffice Calc

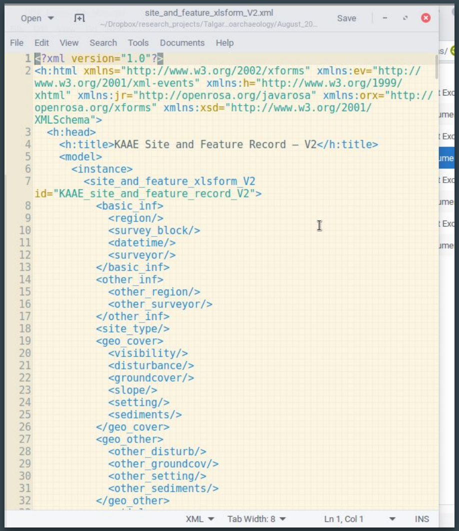
2) Convert to ODK XML format

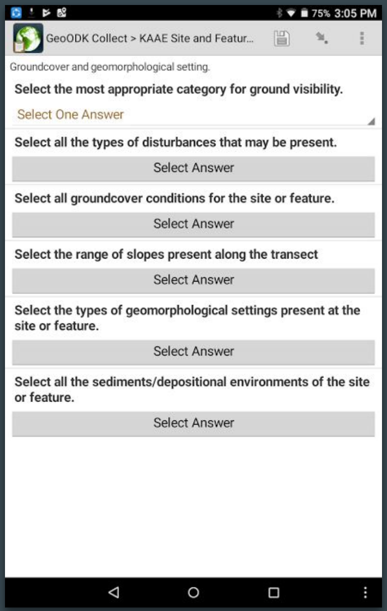
XLS-Form (online or Python)

3) Distribute to devices

Geo/ODK-Collect







Mobile Data Collection

Collected data is aggregated into a central database, and can be exported to common tabular data and GIS formats. This is done with ODK Briefcase or ODK Aggregate.

1) Save completed forms

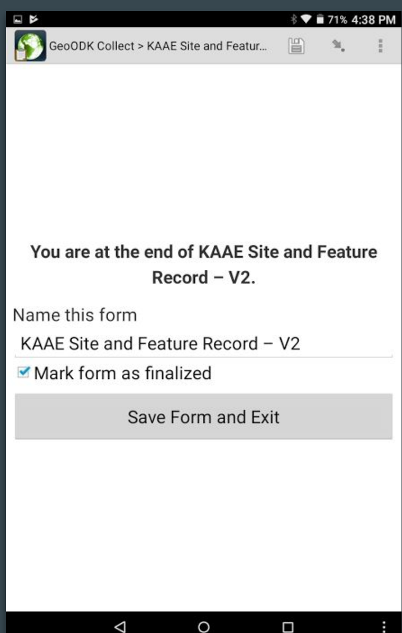
Geo/ODK-Collect

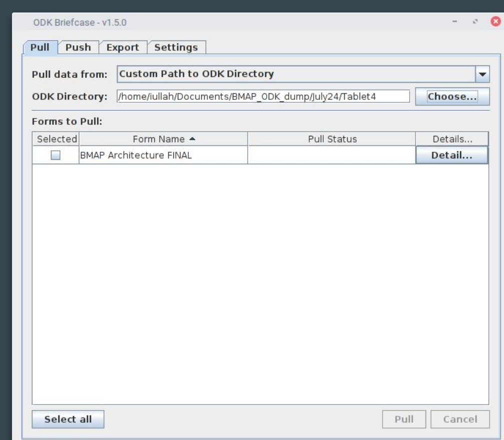
2) Aggregate form data

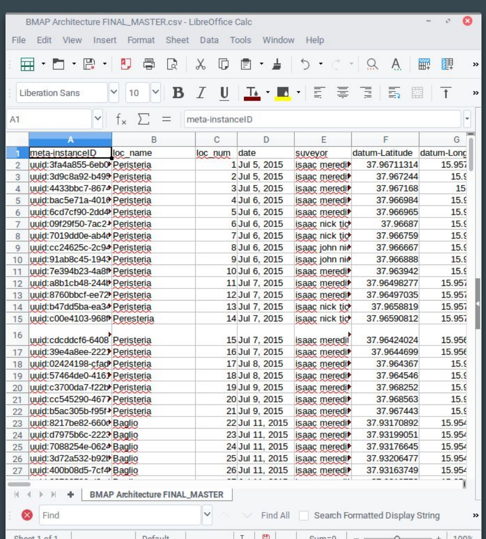
ODK Briefcase or ODK Aggregate

3) Export database

CSV, SHP, etc.





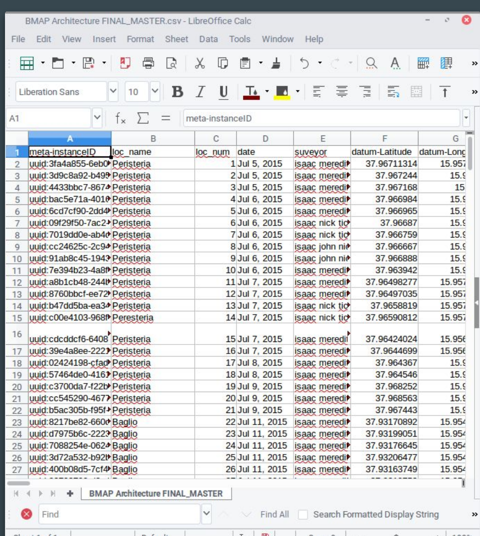


Mobile GIS

At the end of each field day, the form data is aggregated, exported, and a centralized GIS project is updated. The connection of QGIS and the Q-Field app allows every tablet to have a queryable, editable, up-to-date version of the GIS database in the field.

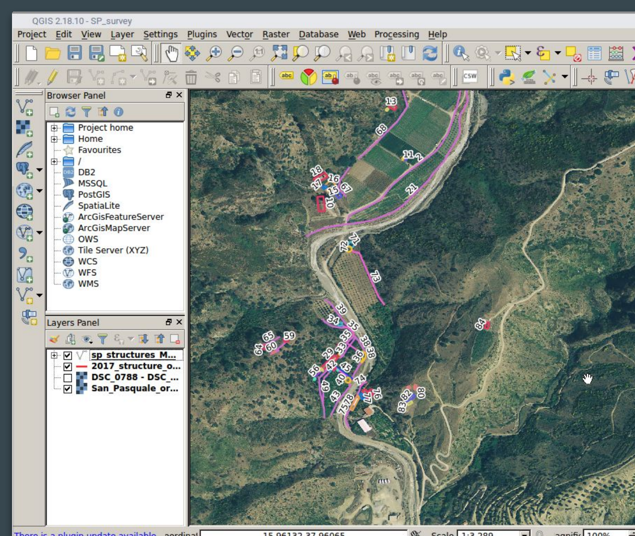
1) Daily export of ODK database

A daily CSV file with coordinates



2) Centralized QGIS project

Styles, labels, layers, and a QGIS project file



3) GIS data in Q-Field

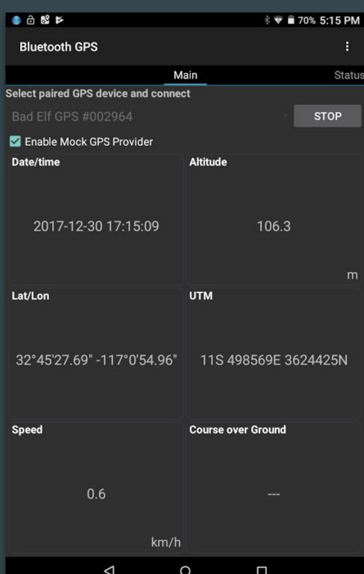
Layers can be hidden/shown

High Precision GPS Tracking

The free (but not open) Bluetooth GPS* app allows the GPS coordinates from the GNSS surveyor to wirelessly replace those from the internal GPS of the device. The GPS Logger app is a flexible solution to record your location in real time. For example, actual survey transect pathways can be recorded, and sweep widths calculated.

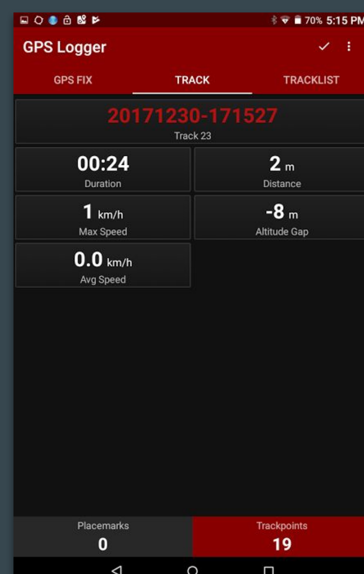
1) Bluetooth GPS connection

High-precision GPS data



2) Real-time GPS logging

Each surveyor logs transect data



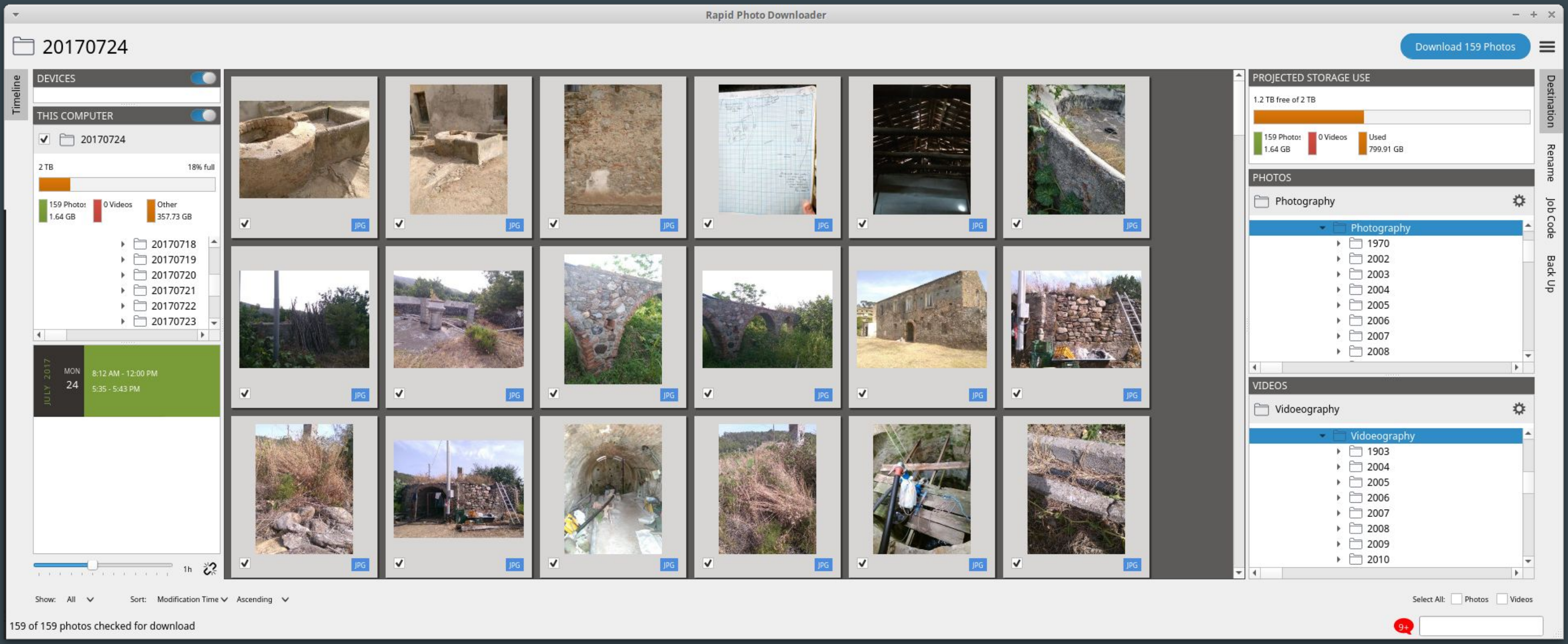
3) GPS track in Q-GIS

Actual walked transect + sweep width

*An Open-Source alternative exists on Sourceforge, but is currently abandoned.

Managing Field Photos

Rapid Photo Downloader greatly smooths the process of downloading images from multiple cameras, and making backups to an external disk. It automatically organizes the images in a user-definable file tree and renaming options, including EXIF tag information. Definable “job codes” can help further differentiate projects.

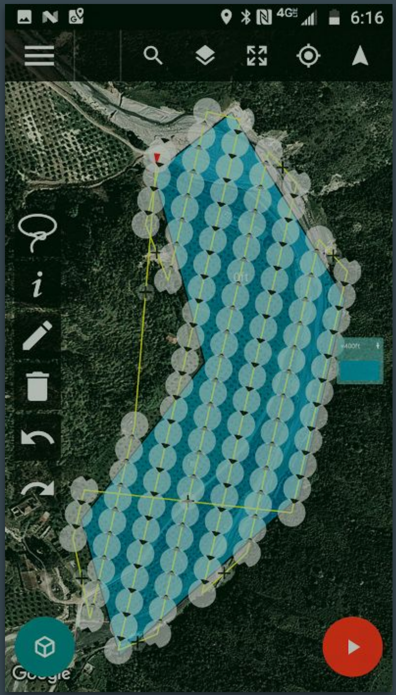


Flight Planning for DJI drones

DJI provides a SDK and a well-documented API. Dronepan is *libre* flight planner app available for DJI drones, but not yet on Android. There are, however, several *gratis* Android apps, including DJI's own GO 4, DroneHarmony, and Aerobotics Flight Planner. I have created a *libre* spreadsheet calculator to help plan flights, but using an automated app-based planner is much more convenient.

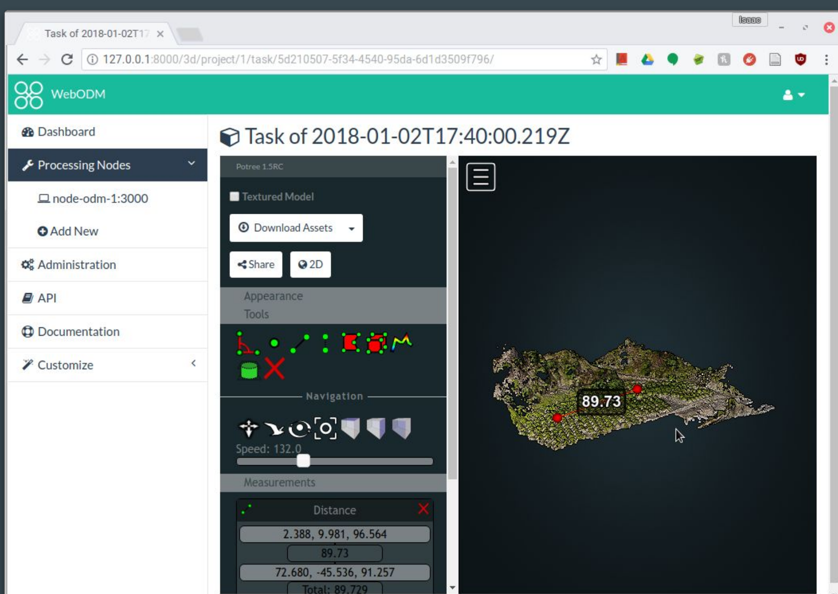
	A	B	C
1	Input Data		
2	*Camera Specs:*		
3	Drone Altitude (m):	120.00	
4	Sensor X (mm):	6.17	
5	Sensor Y (mm):	4.55	
6	Focal Length (mm):	4.00	
7	Image width (px):	4000	
8	Image height (px):	3000	
9	Photo interval (sec):	4.00	
10	*Mission Details:*		
11	Side overlap (%):	0.80	
12	Front overlap (%):	0.80	
13	Survey area width (m):	500.00	
14	Survey area length (m):	400.00	
15			
16	Output Data		
17	***Photo details:***		
18	**Ground X (m):**	**185.10**	
19	**Ground Y (m):**	**136.50**	
20	**Resolution (cm/px):**	**4.63**	
21	***Flight Details:***		
22	**Transect spacing (m):**	**37.02**	
23	**Flight speed (m/sec):**	**6.83**	
24	**Number of Transects:**	**14**	
25	**Flight Time (min):**	**13.68**	
26	**Notes:**		
27	The formula for calculating image footprint assumes a nadir camera angle (straight down) and is:		
28			
29			
30			
31	Flight time/flying speed is influenced by transect spacing, flying height, and the photo interval.		
32	Flying height results in a trade off in coverage vs resolution, and flying higher means faster flight times		

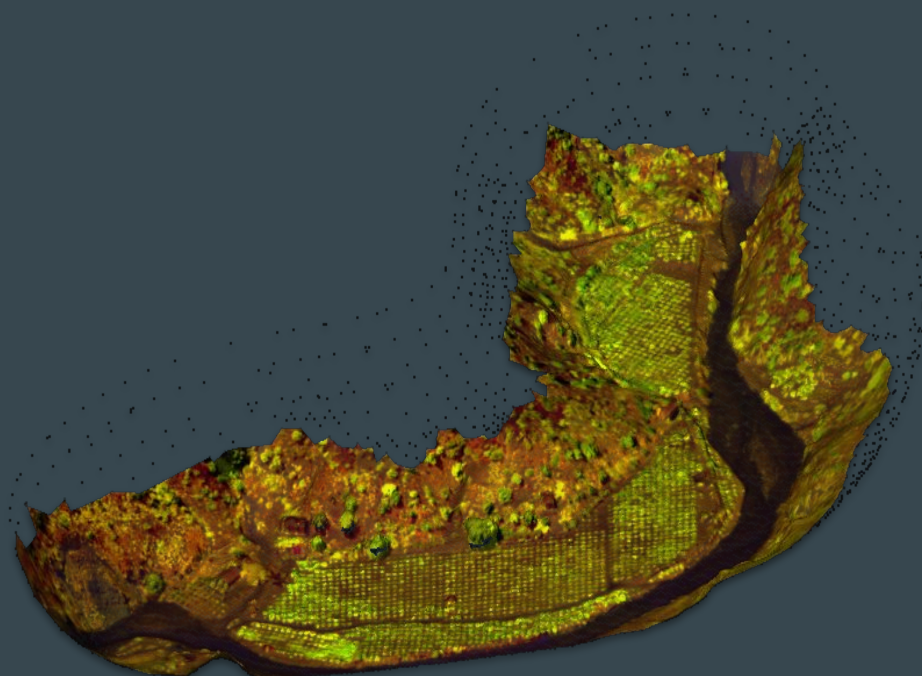
DOI 10.5281/zenodo.1134686



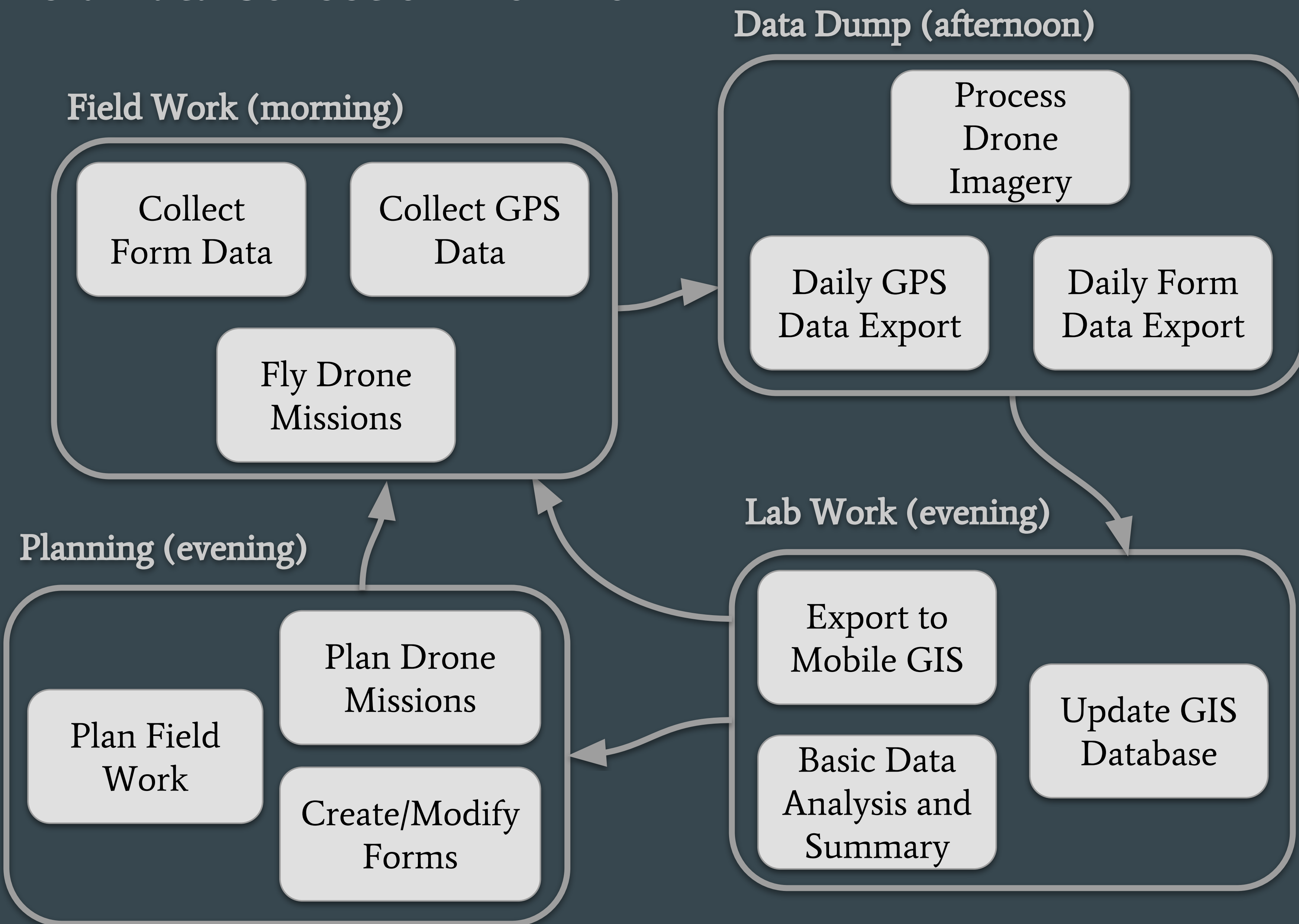
Drone Image Processing

Open Drone Map can create georeferenced orthophotos, pointclouds, 3-D textured mesh, and georeferenced DEMs from unordered drone images with GPS tags. Although it’s best to use a decent computer, it’s really as easy as “load images and hit ‘run’.” In the field, processing time is sped up by choosing to downsize images. Resulting 3D and imagery data are more than good enough for use in the field.

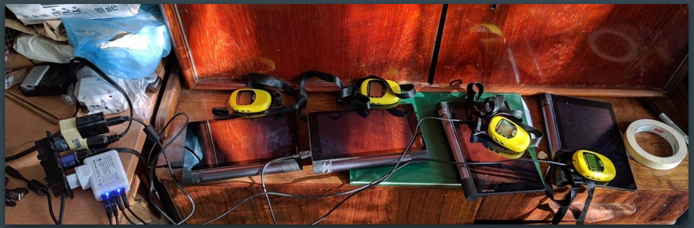




Field Data Collection Workflow



Field Data Collection: Lessons Learned



This last summer marked the third field season in which I employed a version of this paperless workflow, and the sixth field season of paperless data collection altogether. There are several valuable lessons I've learned during this time:

1) **Hardware is the least important part.**
   - Hardware changes fast. Don't worry about "buying the best," just buy what you need.
2) **Form design is paramount to success.**
   - Don't get locked into a suboptimal form! Design your form with potential updates in mind.
3) **Modularity allows for flexibility.**
   - Pick the best software tool for the specific job. Chain them together for a flexible workflow that can adapt to your current needs.
4) **Create "workflow rituals" to prevent mistakes.**
   - Write out the order of operations. Assign specific personnel to specific tasks. Don't duplicate effort, but don't leave something out!
5) **Follow the 3-2-1 backup mantra.**
   - Have at least 3 copies on 2 different forms of media, and store 1 backup offsite (like in the "cloud"). Don't delete anything until you are sure of a backup!

“Data Hygiene” and Post-Processing

Data produced through this workflow are in reasonably good condition. At the end of the field project, a few items of “data hygiene” must be done to correct for any remaining human error.

Once this is done, the final post-processing of the data products can be undertaken. Some of this post-processing can be automated, but it often takes a “human-touch.”

Fixing Form Data

1) Column headers have the “group” name auto-added as prefixes

P	Q	R	S	
external_general-orientation	external_general-height	external_general-levels	external_general-basement	externa
300	10	2	no	
32	11	3	no	
293	3.1	1	no	
296	3.38	1	no	

2) Tracklogs must be connected or converted

K	L	
tracks	tracklog_name	c
yes	GPS 4. 7/10-1	r
yes	GPS 4. 7/10-2	t

43.467675 77.823259 856 3.81;43.467675 77.823261 856 3.81;43.467675 77.82327099999999 856 3.81;43.467673999999995 77.823281 856 3.81;43.467673999999995 77.823291 856 3.81;43.467673999999995 77.823301 856 3.81;43.467676999999995 77.823309 856 3.81;43.467678 77.823308 856 3.81;43.467678 77.82329899999999 856 3.81;43.467678 77.8233 856 3.81;43.46768 77.823309 856 3.81;43.467684999999996 77.823317 856 3.81;43.467689 77.82332799999999 856 3.81;43.467693999999995 77.823341 856 3.81;43.467695 77.823353 856 3.81;43.467698999999996 77.823366 856 3.81;43.467709 77.823371 856 3.81;43.467718 77.823374 856 3.81;43.467727 77.823376 856 3.81;43.467737 77.823379 856 3.81;43.467746999999996 77.823382 856 3.81;43.467757

3) Multiple photos are linked in a subsidiary table, and must be reconnected

A	B	C	
general_image	PARENT_KEY	KEY	SET-OF-images
media/1502252182577.jpg	uuid:f9d490c6-e3c3-44e8-9147-53cbf19d2e2	uuid:f9d490c6-e3c3-44e8-9147-53cbf19d2e2e/other_images-images[	uuid:f9d490c6-e3c3-44e8-
media/1502252197370.jpg	uuid:f9d490c6-e3c3-44e8-9147-53cbf19d2e2	uuid:f9d490c6-e3c3-44e8-9147-53cbf19d2e2e/other_images-images[	uuid:f9d490c6-e3c3-44e8-
media/1502260723293.jpg	uuid:32651e52-af1c-43f9-8045-4513c4bf8c43	uuid:32651e52-af1c-43f9-8045-4513c4bf8c43/other_images-images[1	uuid:32651e52-af1c-43f9-
media/1501830559195.jpg	uuid:41b65c45-2f19-4364-a186-c0fe7721784	uuid:41b65c45-2f19-4364-a186-c0fe77217842/other_images-images[	uuid:41b65c45-2f19-4364-

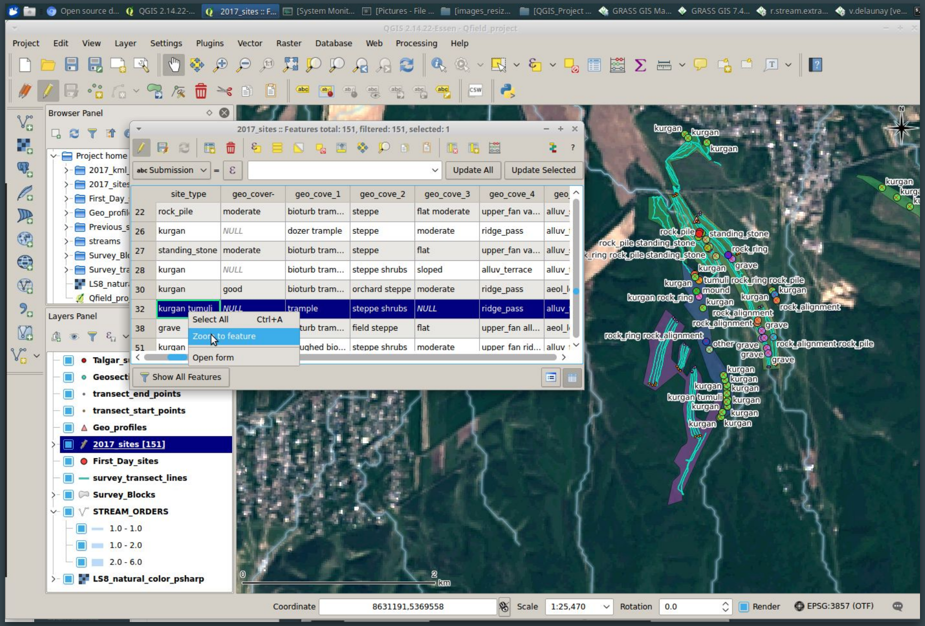
4) Correction of typos and “autocorrect” mistakes

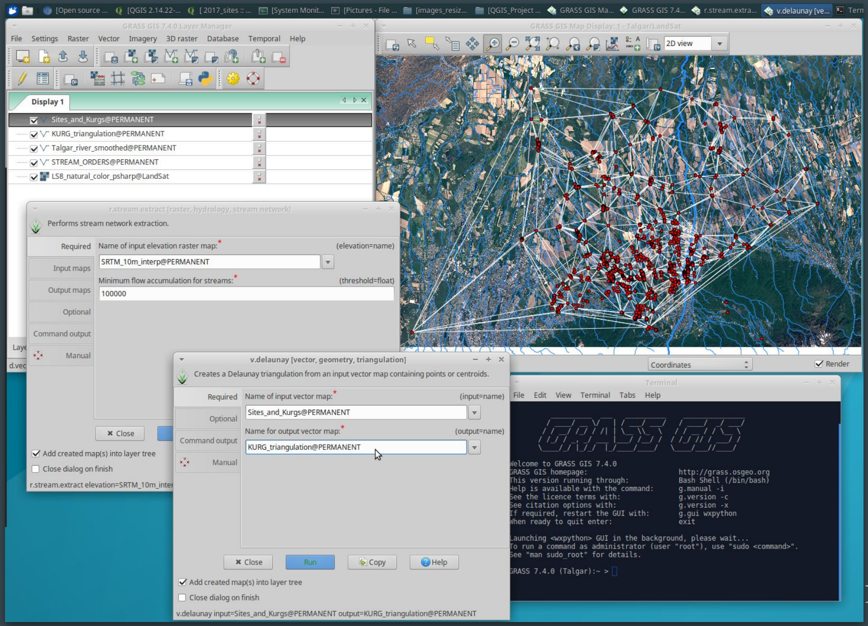
Adjacent to other rock piles to NE, Onan alluvial plain
Large earth mound with depression on east side looted?? Some large rocks on s
Near rock pile recorded separately
Mounds arranged to form rectangular openings
Loose concentration of standing stones.

Fixing the GIS Database

The spatial aspects of the project are best managed in a GIS. GRASS and QGIS are the “dynamic duo” that I recommend. QGIS for daily use in the field and for making nice maps, and GRASS for all follow-up analysis. They play very nicely together.

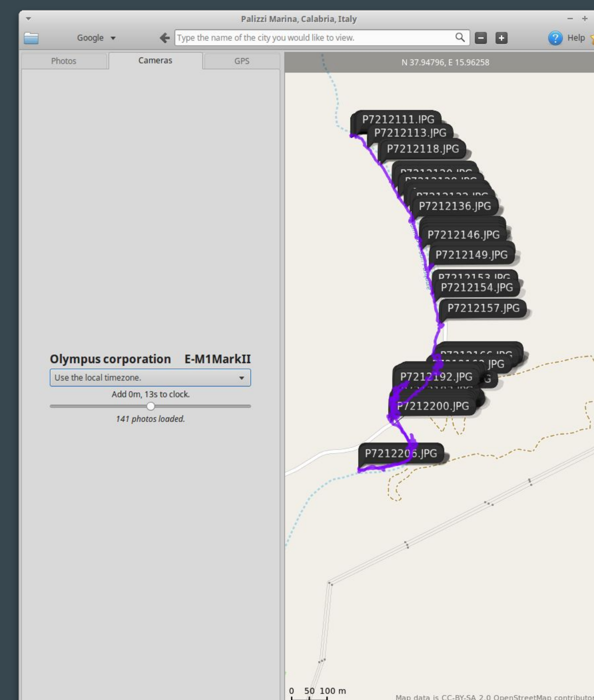
Post-processing includes topology correction and merging of data layers, creation of proper metadata, “project files” with styling and layering, advanced geospatial analyses, statistical analysis, and creation of publication-quality maps.

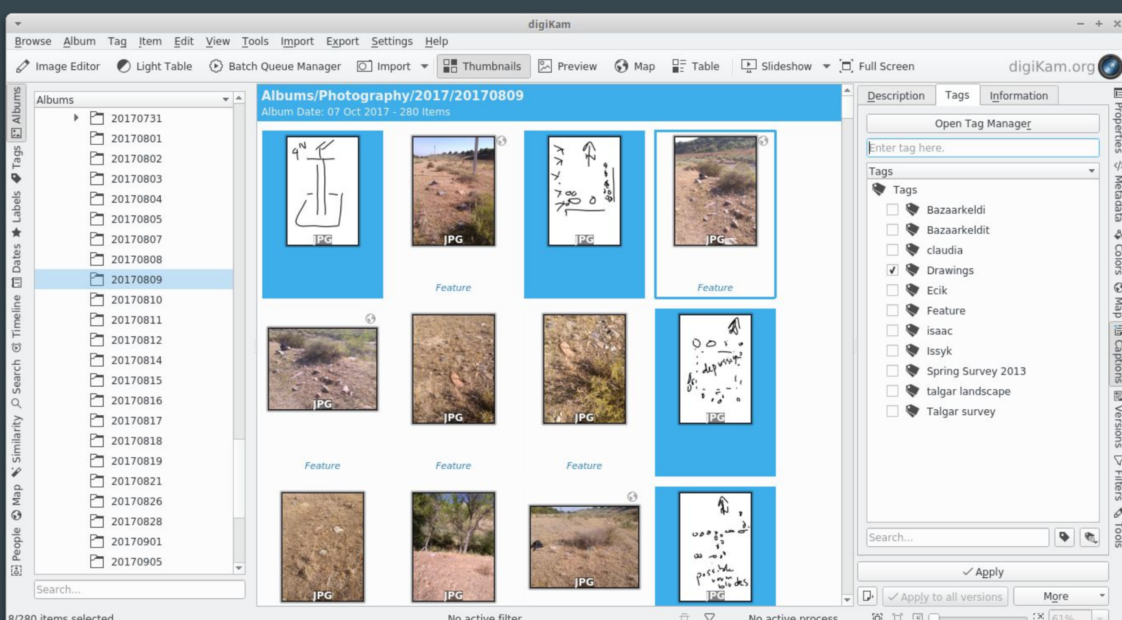




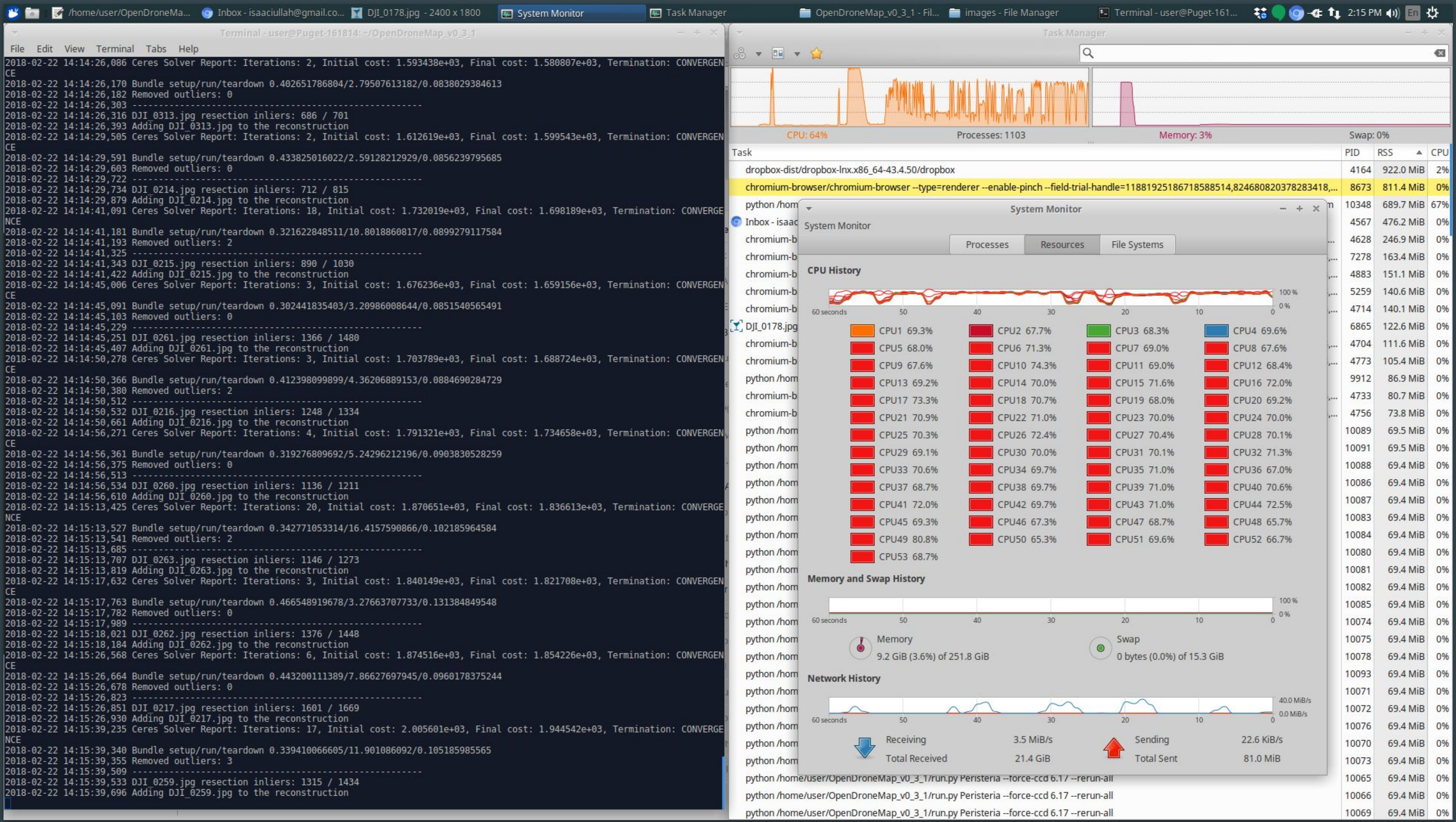
Digital Asset Management (D.A.M.) and Imagery

Images need to be tagged and added to a searchable database. DigiKam makes this easy! It works well with the file structure produced by Rapid Photo Downloader. “Geotags” can be added from GPS tracks easily with GottenGeography. All tags can be stored as EXIF or XMP tags, so they travel with your images.





Final Hi-Res Drone Image Processing



Running Open Drone Map in parallel == NERDVANA!

Using hugin-tools and ImageMagick to Make a Better Airphoto Mosaic

Post-Processing: Lessons Learned

This year, I've had three students working on "data hygiene" and post-processing for these two projects.

1) **Do not underestimate the effort needed for data hygiene.**
   - Despite the fact that digital field collection means that there is no "data entry" or A/D conversion needed, digital data still needs a lot of correction and alteration.
2) **Focus more upfront effort on creating correctly-designed forms.**
   - Many of the most painstaking corrections can be mitigated by using automation (pull-downs, checkboxes, auto-calc fields) in ODK.
3) **Use a D.A.M., and do it early!**
   - Especially for photographs, programs like Rapid Photo Downloader, DigiKam, and GottenGeography are lifesavers to keep them organized and searchable
   - Keyword tagging and geotagging should be begun in the field, not after return.
4) **Computing power may be necessary**
   - High-resolution post-processing of drone images with ODM is computationally intensive. A fast multi-cpu computer (or GPU enabled installation) is very handy!

Data Curation, Versioning, and Dissemination

CENTER FOR OPEN SCIENCE

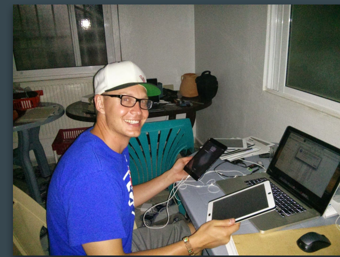
The goal of open, reproducible science requires any project that generates data to **curate**, **version**, and **disseminate** that data. There are several tools available to facilitate this. Which one you choose will depend on the project's goals and time/money budget, but some things to consider are:

1) How much data will my project **generate**, including all secondary data products?
2) How will I keep track of **metadata** about changes, analyses, and secondary data products?
3) For **how long** should I plan to make my data available?

The principles of open, reproducible science are best met through use of **open-source software tools**, employed through **scripted workflows**, that generate **plain-text or open-standard data formats**, with **abundant, informative, metadata**, and released with **permissive licensing**.

Use of a Linux-based operating system offers a seamless and thorough avenue to achieve these goals

Curating Data



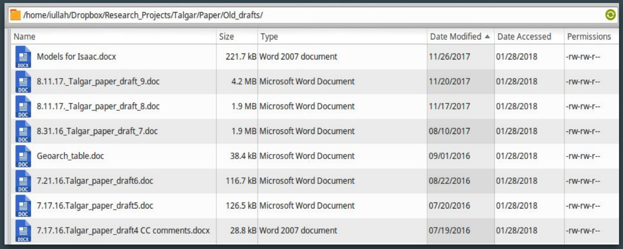
Data **curation** consists of archiving and maintaining long-term copies of the data.

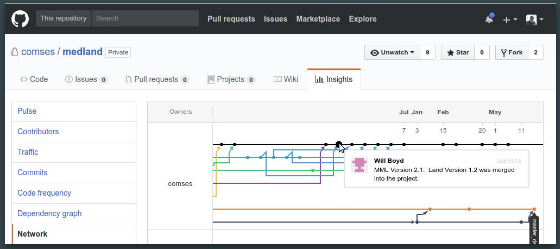
1. **Adequate**: Hard-drives and physical media.
    - This is possibly the most common and certainly the longest running system for data curation. Hopefully you are following the 3-2-1 rule! Sharing consists “burning” a physical copy, or PTP sharing (FTP, etc.). Media have shelf lives, however, and can be lost or destroyed. Arguably the least “open” way to curate data.

2. **Better**: The Cloud.
    - Services like Dropbox, Google Drive, and Spider-Oak make it easy to keep data archives curated in multiple locations. Sharing is easier, but still person-to-person. What happens when your project runs out of money to pay for these services? Only moderately “open.”

3. **Best**: Open Online Repositories.
    - Library and third-party online repositories like GitHub, Figshare, Xenodo, and the OSF facilitate longer-term, more open curation of data. Will these be around 40 years from now?

Versioning Data





<u>Data **versioning** consists of tracking and recording all changes made to data over time.</u>

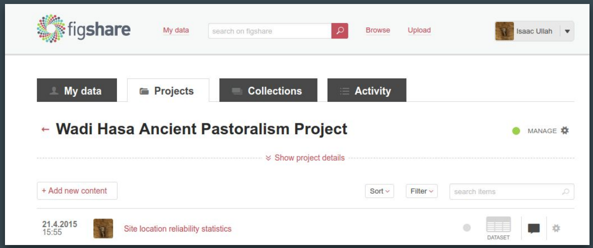
1. Manually:
   - **Pros**: Simple, and “easy.” Anyone can do it.
   - **Cons**: Fallible, hard to share, not particularly transparent.

2. Cloud services (Dropbox, Google Drive, Spider-Oak, etc.)
   - **Pros**: Most services have good to decent integration on Linux (Dropbox daemon, the Grive project, Spider-Oak application). Data files are backed-up to the cloud, auto-synced across computers, and a certain length of file history is maintained. Reasonably easy sharing via links.
   - **Cons**: You have to pay for all this convenience. Versioning information is not public, and is limited. File conflicts are not easy to solve (e.g., “so and so’s conflicted version”).

3. Static repositories (FigShare, Xenodo, OSF, Library Repositories):
   - **Pros**: Really easy: sign up and upload. Can be connected to Cloud/GitHub. Open Access.
   - **Cons**: “Versions” are only static releases. You have to remember to upload.

4. Dynamic repositories (Git, GitHub, GitLab):
   - **Pros**: Native Linux CLI integration (can be easily scripted). Truly excellent versioning, and much better tools to merge conflicts. Unlimited storage in public archives (GitHub, GitLab). Easy, open-access, sharing. Can make static releases. Can connect to static repositories.
   - **Cons**: Git archives can get weird with very large filesets. Everyone on the team needs to know Git.

Disseminating Data

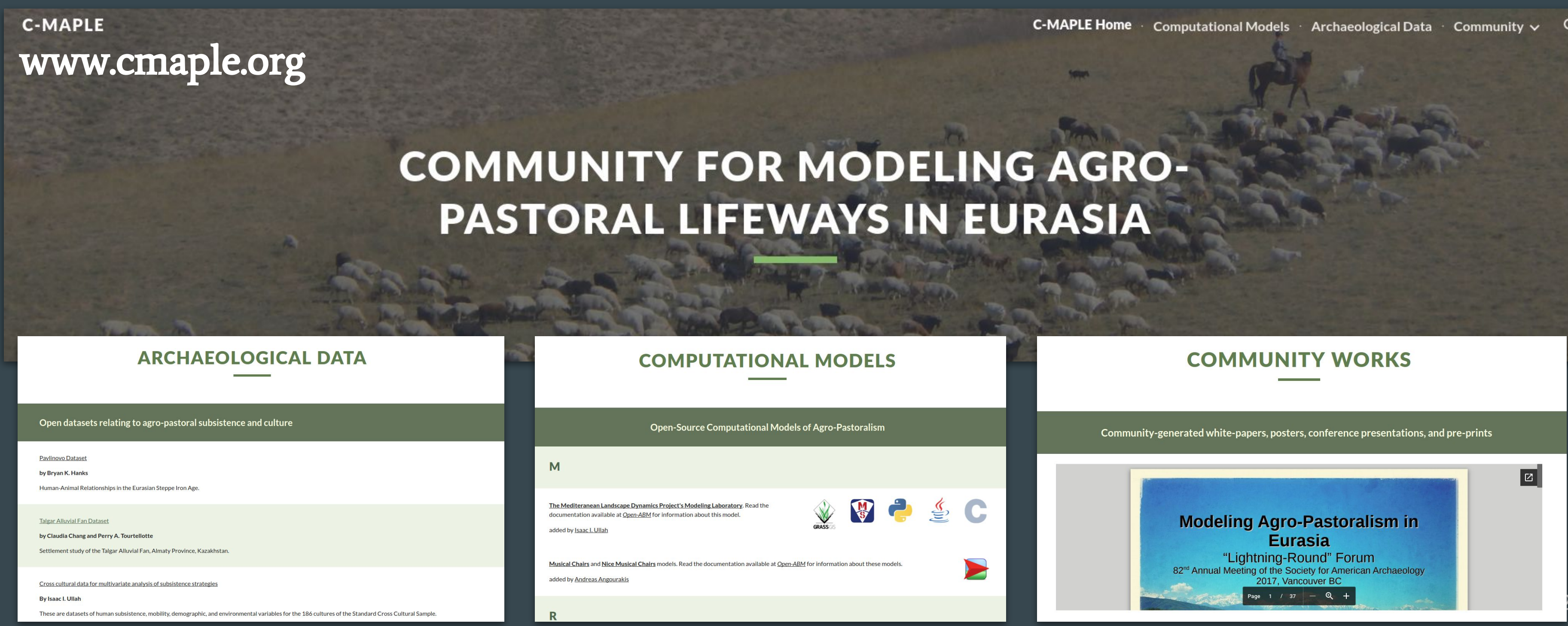


Data **dissemination** consists of making data findable and publicly available.

1. Personal websites and servers:
    - **Pros**: Allow you more control over how the date are presented. Can use any file hosting service.
    - **Cons**: What happens when you forget to pay the domain name or hosting fees? Licensing unclear.
2. Library and institutional archives:
    - **Pros**: Hopefully long-lasting and well organized. Hopefully permissive licensing is encouraged.
    - **Cons**: Access is sometimes restricted, not always an intuitive place to look for datasets (and possibly not indexed by search engines).
3. Third-party repositories (Figshare, Xenodo, GitHub, OSF):
    - **Pros**: Offer a good “middle ground.” They are indexed by search engines, and can be linked to academic works via static DOI numbers. Data collections can be “released.” Licensing must be clear.
    - **Cons**: Uncertainty about longevity of services in a volatile world of “tech startups.” Generic archives of data: how to find related datasets? Standardizations?
4. Domain-specific repositories (tDAR, OpenContext, FAIMS, UK Heritage, ComSES, Systems):
    - **Pros**: Centralizes datasets on specific subjects. Encourages standardization, and permissive licenses.
    - **Cons**: Often comes with considerable up-front costs (not always). Requires a lot of overhead. Competition between similar archives: which one to choose?

Data dissemination: Lessons learned.

There are a plethora of ways to make data available. While a richness of options is a good thing, it makes finding the actual pieces of data you are interested in quite difficult. Rather than force people to use one method, I suggest the creation of **community-driven catalogs**, where links to data on similar topics can be centralized, curated, annotated, and shared by and for the community that wants to use them.



Thank You!

Thanks to the BMAP team, the KAAE team, my students in the Computational Archaeology Lab at SDSU, the ComSES and C-MAPLE communities, all the devel teams of all these wonderful pieces of F/LOSS, and to YOU!

More information, including links and downloads, can be found at my website:

isaacullah.github.io