

Open source digital data collection for field sciences.



This work is licensed under a
[Creative Commons Attribution
4.0 International License](https://creativecommons.org/licenses/by/4.0/).

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.

Digital Data Collection as an Open Method

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

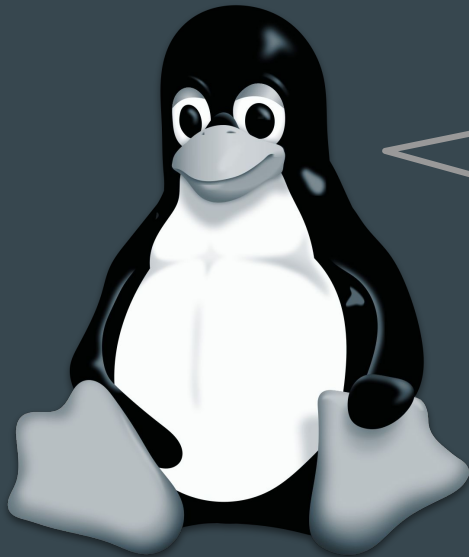
1. Faster in-field collection, little to no need for “post production” of data.
2. Better data standardization, fewer recording and transcription errors.
3. Ability to “course correct” based on real-time data analysis.
4. Near-instant upload, backup, and real-time long-distance collaboration.

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*.**

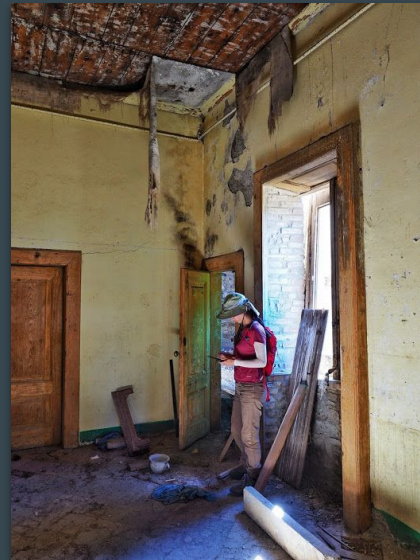
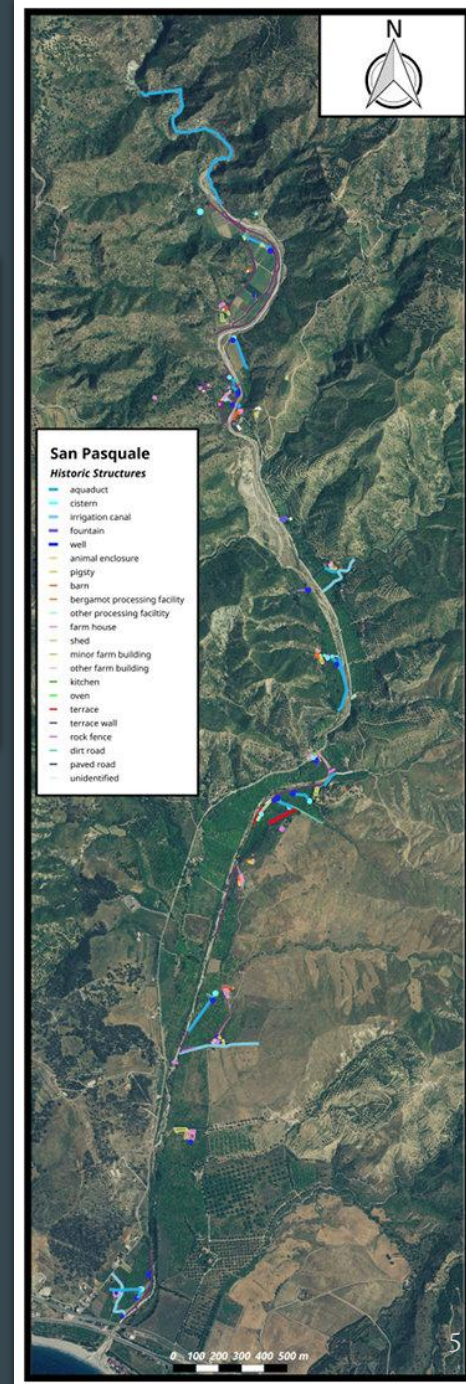
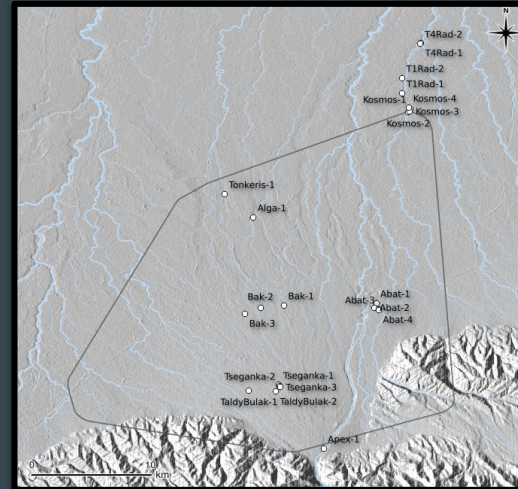


I'm a pretty good choice for open field science!

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

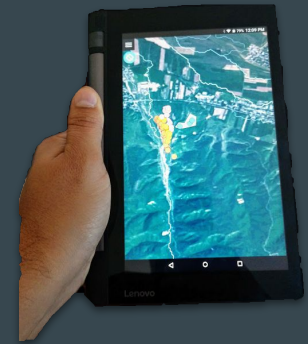


Field data collection workflow.

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

type	name	label
begin group	basic_inf	Basic information.
select_one region	region	Select the region.
integer	survey_block	Enter the survey block number.
datetime	datetime	Enter date and time.
select_multiple surveyor	surveyor	Select all the personnel who recorded this site or feature.
end group	basic_inf	
begin group	other_inf	Enter specifics of "other" in the following categories:
text	other_region	Other region.
text	other_surveyor	Other personnel.
end group	other_inf	
select_multiple site_type	site_type	Select the type(s) of archaeological sites or features recorded.
begin group	geo_cover	Groundcover and geomorphological setting.
select_one visibility	visibility	Select the most appropriate category for ground visibility.
select_multiple disturb	disturbance	Select all the types of disturbances that may be present.
select_multiple groundcov	groundcover	Select all groundcover conditions for the site or feature.
select_multiple slope	slope	Select the range of slopes present along the transect.
select_multiple setting	setting	Select the types of geomorphological settings present at the site or feature.
select_multiple sediment	sediments	Select all the sediments/depositional environments recorded at the site or feature.
end group	geo_cover	
begin group	geo_other	Enter specifics of "other" in the following categories:
text	other_disturb	Other disturbances.
text	other_groundcov	Other groundcover.
text	other_setting	Other geomorphological setting.
text	other_sediments	Other sediments/depositional environments.
end group	geo_other	
begin group	soatial	Collect spatial information about the site or feature.



2) Convert to ODK XML format

XLS-Form (online or Python)

```
<?xml version="1.0"?>
<html xmlns="http://www.w3.org/2002/xforms" xmlns:ev="http://www.w3.org/2001/xml-events" xmlns:h="http://www.w3.org/1999/xhtml" xmlns:j="http://openrosa.org/javarosa" xmlns:orx="http://openrosa.org/xforms" xmlns:xsd="http://www.w3.org/2001/XMLSchema">
  <head>
    <title>KAAE Site and Feature Record - V2</title>
  </head>
  <model>
    <instance id="KAAE site and feature record V2" data-j="xlsform:form" data-orx="xlsform:form" data-xsd="xlsform:form">
      <basic_inf>
        <region/>
        <survey_block/>
        <datetime/>
        <surveyor/>
      </basic_inf>
      <other_inf>
        <other_region/>
        <other_surveyor/>
      </other_inf>
      <site_type/>
      <geo_cover>
        <visibility/>
        <disturbance/>
        <groundcover/>
        <slope/>
        <setting/>
        <sediments/>
      </geo_cover>
      <geo_other>
        <other_disturb/>
        <other_groundcov/>
        <other_setting/>
        <other_sediments/>
      </geo_other>
    </instance>
  </model>
</html>
```



3) Distribute to devices

Geo/ODK-Collect

Groundcover and geomorphological setting.

Select the most appropriate category for ground visibility.

Select One Answer

Select all the types of disturbances that may be present.

Select Answer

Select all groundcover conditions for the site or feature.

Select Answer

Select the range of slopes present along the transect

Select Answer

Select the types of geomorphological settings present at the site or feature.

Select Answer

Select all the sediments/depositional environments of the site or feature.

Select Answer

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.

GeoODK Collect > KAAE Site and Featur...

You are at the end of KAAE Site and Feature Record – V2.

Name this form
KAAE Site and Feature Record – V2

Mark form as finalized

Save Form and Exit



ODK Briefcase - v1.5.0

Pull Push Export Settings

Pull data from: Custom Path to ODK Directory

ODK Directory: /home/iullah/Documents/BMAP_ODK_dump/july24/Tablet4

Forms to Pull:

Selected	Form Name	Pull Status	Details...
<input type="checkbox"/>	BMAP Architecture FINAL		Detail...

Select all Pull Cancel



BMAP Architecture FINAL_MASTER.csv - LibreOffice Calc

A1	A	B	C	D	E	F	G		
	meta-instanceID	loc_name	loc_run	date	surveyor	datum	Latitude	datum	Long
2	Luidd39a8a55-6a87	Peristeria		1 Jul 5, 2015	isaac meredi		37.9671314		15.95
3	Luidd3dfc8a92-b499	Peristeria		2 Jul 5, 2015	isaac meredi		37.967244		15.9
4	Luidd4433bdc7-8679	Peristeria		3 Jul 5, 2015	isaac meredi		37.967168		15.9
5	Luidd0ac9e71a-4019	Peristeria		4 Jul 6, 2015	isaac meredi		37.966984		15.9
6	Luidd6c07c78c-29a9	Peristeria		5 Jul 6, 2015	isaac meredi		37.966965		15.9
7	Luidd09f29f50-7ac2	Peristeria		6 Jul 6, 2015	isaac nick tip		37.96687		15.9
8	Luidd7019a0de-ab49	Peristeria		7 Jul 6, 2015	isaac nick tip		37.966759		15.9
9	Luiddcc24625c-9c39	Peristeria		8 Jul 6, 2015	isaac john nap		37.966667		15.9
10	Luidd91ab8c45-1943	Peristeria		9 Jul 6, 2015	isaac john nap		37.966888		15.9
11	Luidd7e394b23-4a89	Peristeria		10 Jul 6, 2015	isaac meredi		37.963942		15.9
12	Luidda891c048-2449	Peristeria		11 Jul 7, 2015	isaac meredi		37.96498277		15.95
13	Luidd8780b0c1-ee729	Peristeria		12 Jul 7, 2015	isaac meredi		37.96497035		15.95
14	Luidd8470d5ba-esa3	Peristeria		13 Jul 7, 2015	isaac nick tip		37.9658819		15.95
15	Luiddc00e4103-9e89	Peristeria		14 Jul 7, 2015	isaac nick tip		37.96590812		15.95
16	Luiddcddcd1c-6408	Peristeria		15 Jul 7, 2015	isaac meredi		37.96424024		15.954
17	Luidd39e4a8ee-2223	Peristeria		16 Jul 7, 2015	isaac meredi		37.9644699		15.954
18	Luidd02424198-c7a9	Peristeria		17 Jul 8, 2015	isaac meredi		37.964367		15.9
19	Luidd57464a4d-4109	Peristeria		18 Jul 8, 2015	isaac meredi		37.964546		15.9
20	Luiddc3700da7-c229	Peristeria		19 Jul 9, 2015	isaac meredi		37.968252		15.9
21	Luiddcc545290-4677	Peristeria		20 Jul 9, 2015	isaac meredi		37.968563		15.9
22	Luiddb5ac305b-1994	Peristeria		21 Jul 9, 2015	isaac meredi		37.967443		15.9
23	Luidd8217be82-6609	Baglio		22 Jul 11, 2015	isaac meredi		37.93178992		15.954
24	Luidd4797596c-2223	Baglio		23 Jul 11, 2015	isaac meredi		37.93190051		15.954
25	Luidd7088254e-0e24	Baglio		24 Jul 11, 2015	isaac meredi		37.93176645		15.954
26	Luidd3d72a532-b929	Baglio		25 Jul 11, 2015	isaac meredi		37.93206477		15.954
27	Luidd4003b0a5-7c19	Baglio		26 Jul 11, 2015	isaac meredi		37.93163749		15.954

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

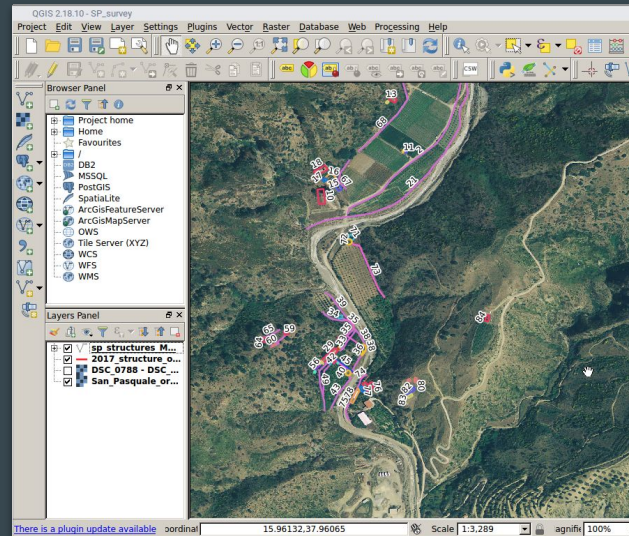
2) Centralized QGIS project

Styles, labels, layers, and a QGIS project file

3) GIS data in Q-Field

Layers can be hidden/shown

meta-instanceID	loc_name	loc_rnum	date	surveyor	datum	Latitude	datum-Long
1	314a4855-64b*	Peristena	1 Jul 5, 2015	isaac_mered*		37.96711314	15.951
2	3d9c8a82-b48*	Peristena	2 Jul 5, 2015	isaac_mered*		37.967244	15.951
3	4433bbc7-867*	Peristena	3 Jul 5, 2015	isaac_mered*		37.967188	15.951
4	ba6c5e71a-401*	Peristena	4 Jul 6, 2015	isaac_mered*		37.966984	15.951
5	cc7c179b-2a8*	Peristena	5 Jul 6, 2015	isaac_mered*		37.966965	15.951
6	09f29f50-7ac2*	Peristena	6 Jul 6, 2015	isaac_nick to*		37.966887	15.951
7	70196d0e-ab4*	Peristena	7 Jul 6, 2015	isaac_nick to*		37.966759	15.951
8	cc24629c-2c9*	Peristena	8 Jul 6, 2015	isaac_nick to*		37.966667	15.951
9	91ab8c4c-184*	Peristena	9 Jul 6, 2015	isaac_nick to*		37.966588	15.951
10	7e394b23-4a8*	Peristena	10 Jul 6, 2015	isaac_mered*		37.963942	15.951
11	a81c1b48-244*	Peristena	11 Jul 7, 2015	isaac_mered*		37.96498277	15.951
12	6760ba14-e72*	Peristena	12 Jul 7, 2015	isaac_mered*		37.96497035	15.951
13	847d45ba-ea3*	Peristena	13 Jul 7, 2015	isaac_nick to*		37.9665819	15.951
14	cc0e4103-9e8*	Peristena	14 Jul 7, 2015	isaac_nick to*		37.96590812	15.951
15	ccddc4f6-64b*	Peristena	15 Jul 7, 2015	isaac_mered*		37.96424024	15.951
16	39e4a8ee-222*	Peristena	16 Jul 7, 2015	isaac_mered*		37.9644699	15.951
17	02424198-cta*	Peristena	17 Jul 8, 2015	isaac_mered*		37.964367	15.951
18	57464a6b-410*	Peristena	18 Jul 8, 2015	isaac_mered*		37.964546	15.951
19	c370da17-f22*	Peristena	19 Jul 9, 2015	isaac_mered*		37.968252	15.951
20	cc545290-467*	Peristena	20 Jul 9, 2015	isaac_mered*		37.968563	15.951
21	b5ac305b-498*	Peristena	21 Jul 9, 2015	isaac_mered*		37.967443	15.951
22	6217a6b2-660*	Baglio	22 Jul 11, 2015	isaac_mered*		37.93170892	15.951
23	47975b6c-222*	Baglio	23 Jul 11, 2015	isaac_mered*		37.93199051	15.951
24	7088254e-062*	Baglio	24 Jul 11, 2015	isaac_mered*		37.93176645	15.951
25	3d72a532-492*	Baglio	25 Jul 11, 2015	isaac_mered*		37.93206477	15.951
26	400b085-7c*	Baglio	26 Jul 11, 2015	isaac_mered*		37.93163749	15.951
27							



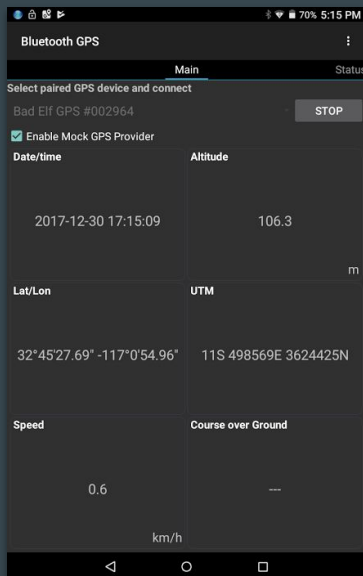


High Precision GPS Tracking

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.

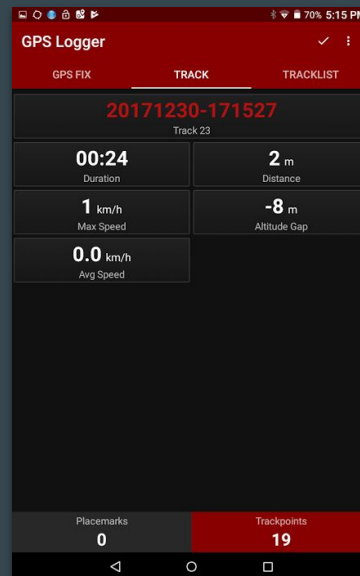
1) Bluetooth GPS connection

High-precision GPS data



2) Real-time GPS logging

Each surveyor logs transect data



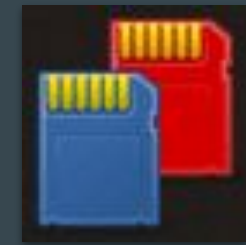
3) GPS track in Q-GIS

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.

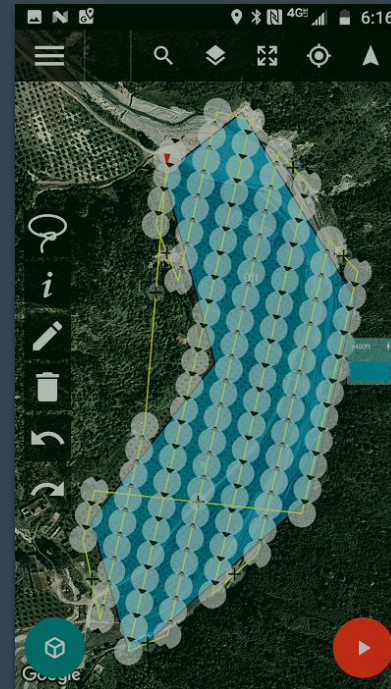
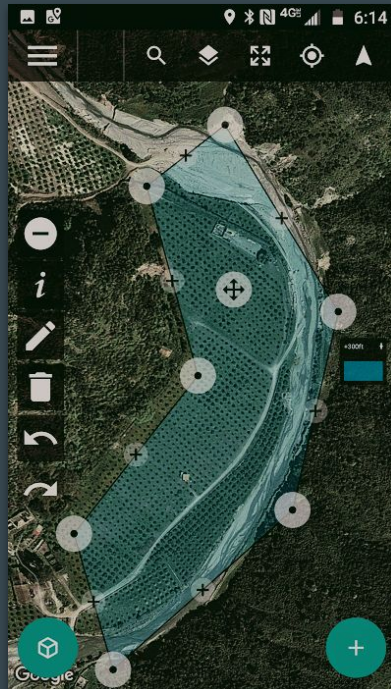
The screenshot shows the Rapid Photo Downloader software interface. The main window displays a grid of 18 photo thumbnails, each with a checkmark and a 'JPG' label. On the left, there's a 'DEVICES' sidebar showing 'THIS COMPUTER' with a folder '20170724' selected. Below it, a storage usage bar shows '159 Photos' (1.64 GB) and '0 Videos' (0 GB). A 'Timeline' sidebar on the far left shows the date 'MON 24 JULY 2017'. On the right, a 'PROJECTED STORAGE USE' sidebar shows '1.2 TB free of 2 TB' and a file tree for 'PHOTOS' and 'VIDEOS'. The 'PHOTOS' tree shows folders for years 1970-2008, and the 'VIDEOS' tree shows folders for years 1903-2010. At the bottom, there's a status bar showing '159 of 159 photos checked for download' and a 'Download 159 Photos' button.

Flight Planning for DJI drones

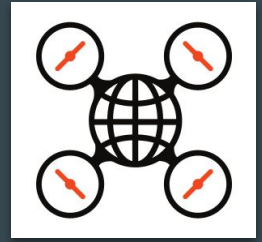


DJI provides a SDK and a well-documented API. Dronepan is *libre* flight planner app available for DJI drones, but not yet on Android. There are, however, several *gratis* Android apps, including DJI's own GO 4, DroneHarmony, and Aerobotics Flight Planner. I have created a *libre* spreadsheet calculator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

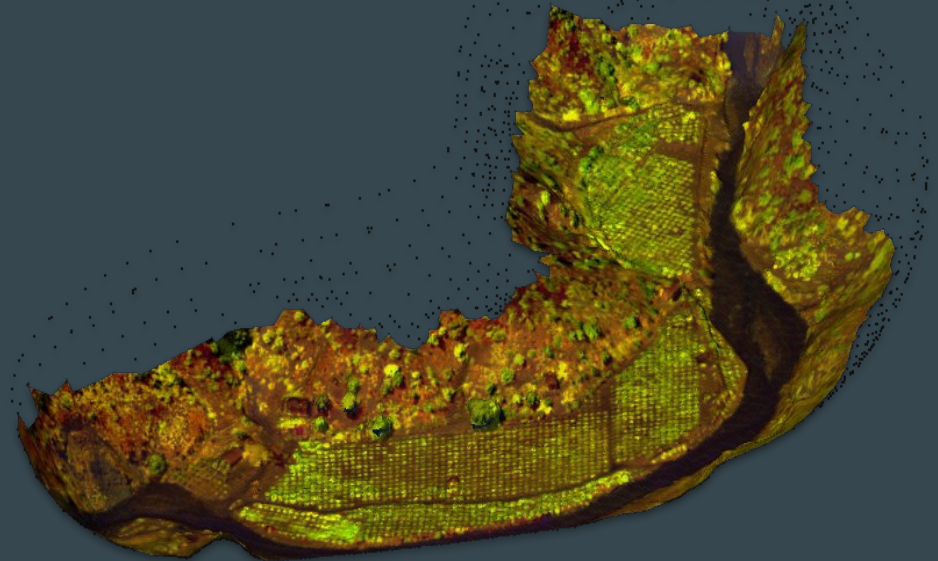
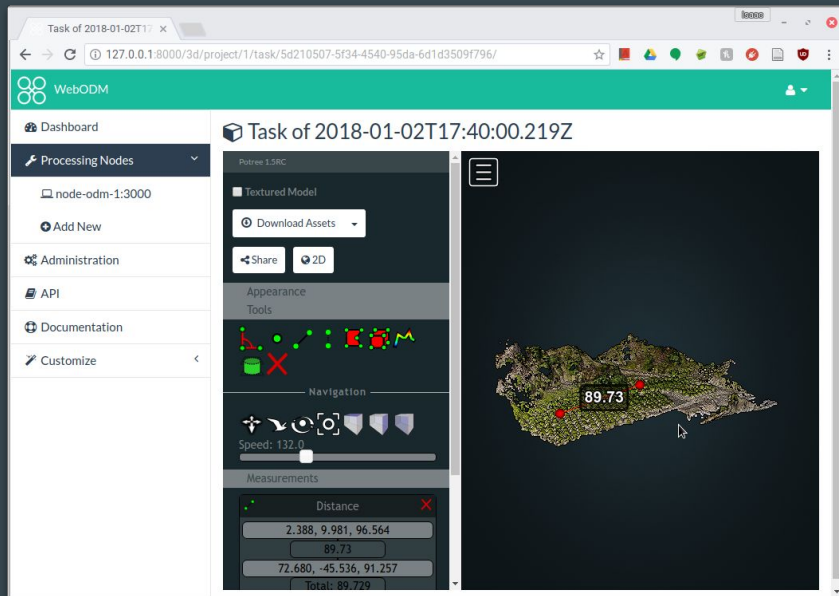
	A	B	C
1	Input Data		
2	Camera Specs:		
3	Drone Altitude (m):	120.00	
4	Sensor X (mm):	6.17	
5	Sensor Y (mm):	4.55	
6	Focal Length (mm):	4.00	
7	Image width (px):	4000	
8	Image height (px):	3000	
9	Photo interval (sec):	4.00	
10	Mission Details:		
11	Side overlap (%):	0.80	
12	Front overlap (%):	0.80	
13	Survey area width (m):	500.00	
14	Survey area length (m):	400.00	
15	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	$\text{Footprint} = \frac{\text{Image width} \times \text{Drone altitude}}{\text{Focal length}}$		
29	The formula for calculating flight time is:		
30	$\text{Flight time} = \frac{\text{Survey area length}}{\text{Flight speed}}$		
31	The formula for calculating resolution is:		
32	$\text{Resolution} = \frac{\text{Ground distance}}{\text{Image width}}$		
	DOI	10.5281/zenodo.1134686	



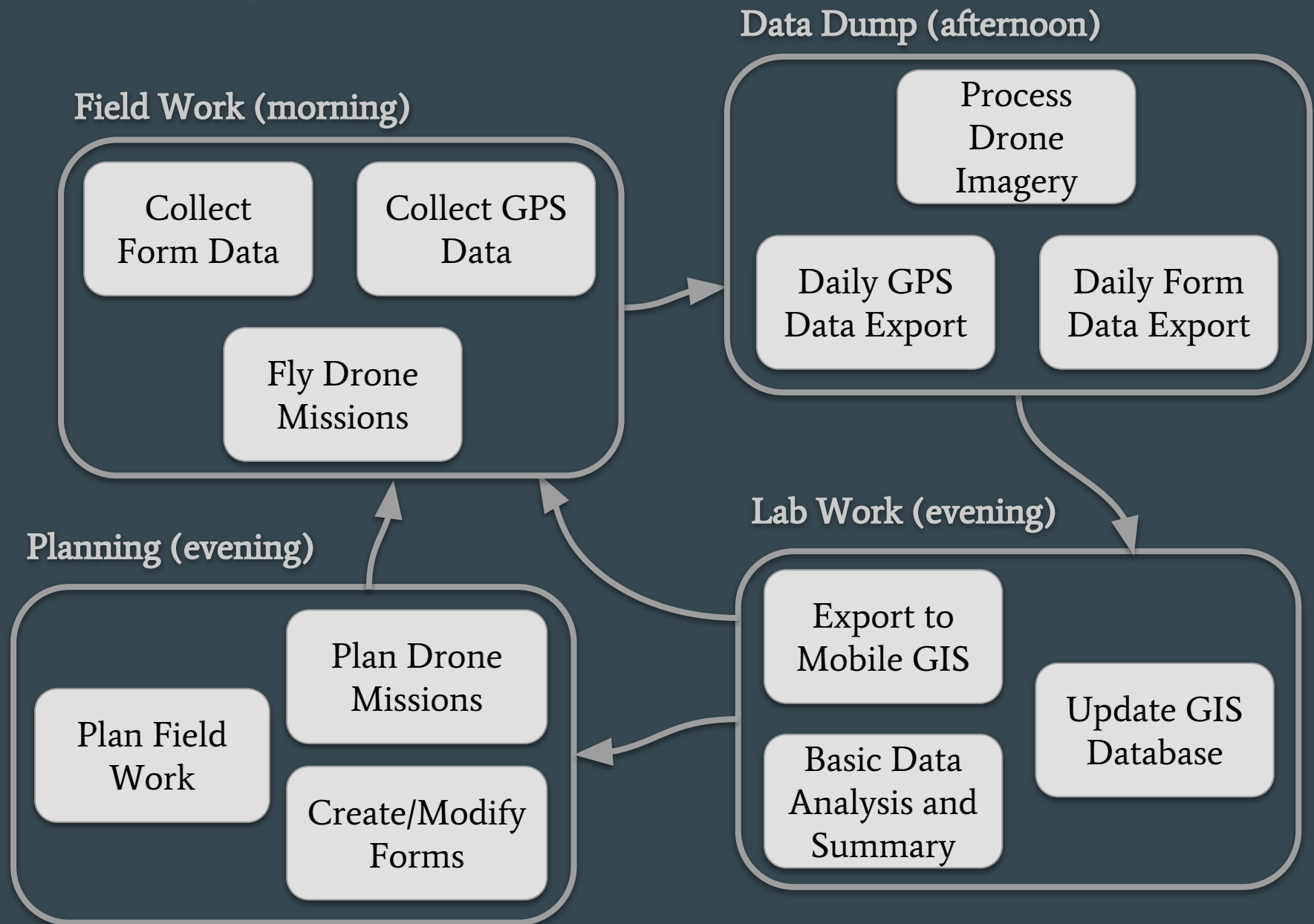
Drone Image Processing



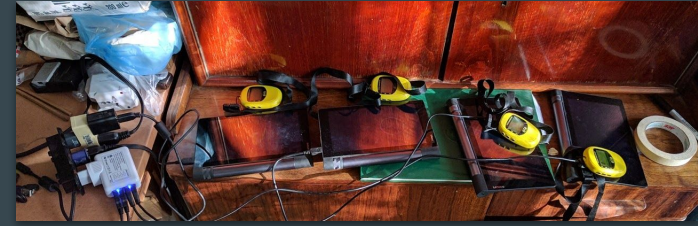
Open Drone Map can create georeferenced orthophotos, pointclouds, 3-D textured mesh, and georeferenced DEMs from unordered drone images with GPS tags. Although it's best to use a decent computer, it's really as easy as “load images and hit ‘run’.” In the field, processing time is sped up by choosing 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	external
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	
yes	GPS 4. 7/10-1	
yes	GPS 4. 7/10-2	

43.467675 77.823259 856 3.81;43.467675 77.823261 856 3.81;43.467675 77.82327099999999 856 3.81;43.4676739999999995
 77.823281 856 3.81;43.4676739999999995 77.823291 856 3.81;43.4676739999999995 77.823301 856
 3.81;43.4676769999999995 77.823309 856 3.81;43.467678 77.823308 856 3.81;43.467678 77.823298999999999 856
 3.81;43.467678 77.8233 856 3.81;43.46768 77.823309 856 3.81;43.4676849999999996 77.823317 856 3.81;43.467689
 77.823327999999999 856 3.81;43.4676939999999995 77.823341 856 3.81;43.467695 77.823353 856 3.81;43.4676989999999996
 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;43.4677469999999996 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-e3c3-44e8-9147-53cbf19d2e2e	uuid:f9d490c6-e3c3-44e8-9147-53cbf19d2e2e/other_images-images[2]	uuid:f9d490c6-e3c3-44e8-
media/1502252197370.jpg	uuid:f9d490c6-e3c3-44e8-9147-53cbf19d2e2e	uuid:f9d490c6-e3c3-44e8-9147-53cbf19d2e2e/other_images-images[2]	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-c0fe77217842	uuid:41b65c45-2f19-4364-a186-c0fe77217842/other_images-images[2]	uuid:41b65c45-2f19-4364-

4) Correction of typos and “autocorrect” mistakes

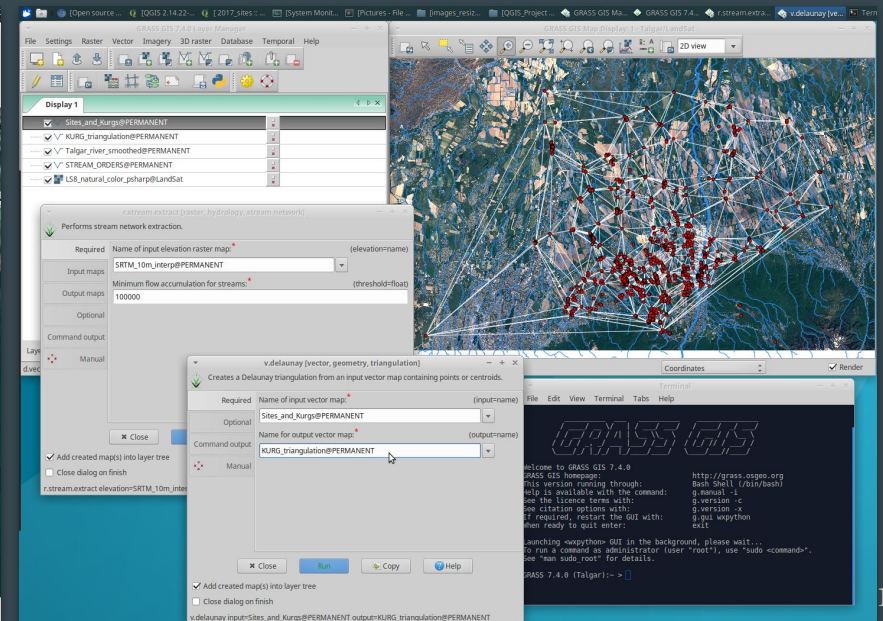
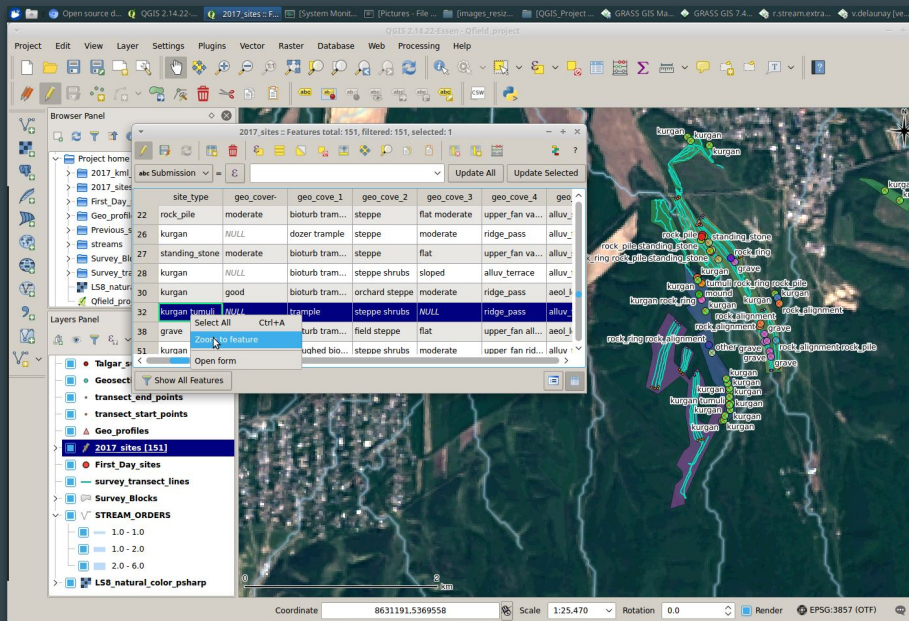
Adjacent to other rock piles to NE, <u>Onan</u> alluvial plain
Large earth mound with depression on east side looted?? Some large rocks on s
Near rock pile recorded separately
Mounds arranged to form rectangular openings
Loose concentration of standing stones.

Fixing the GIS Database



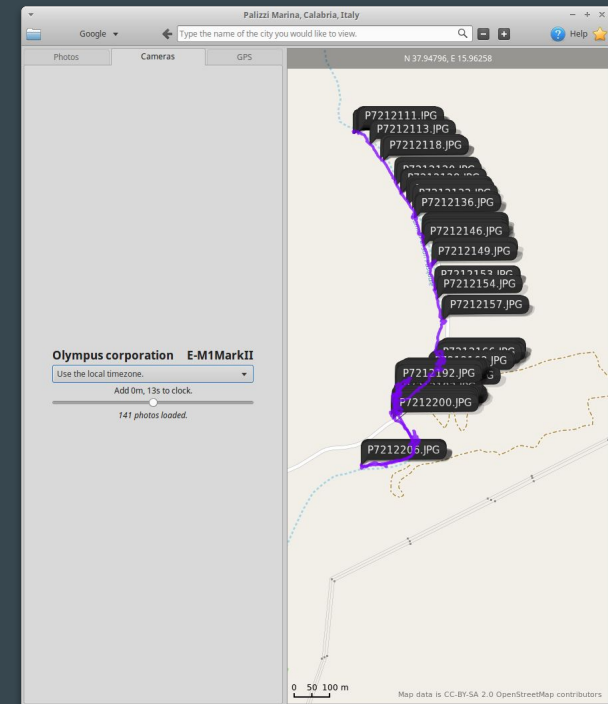
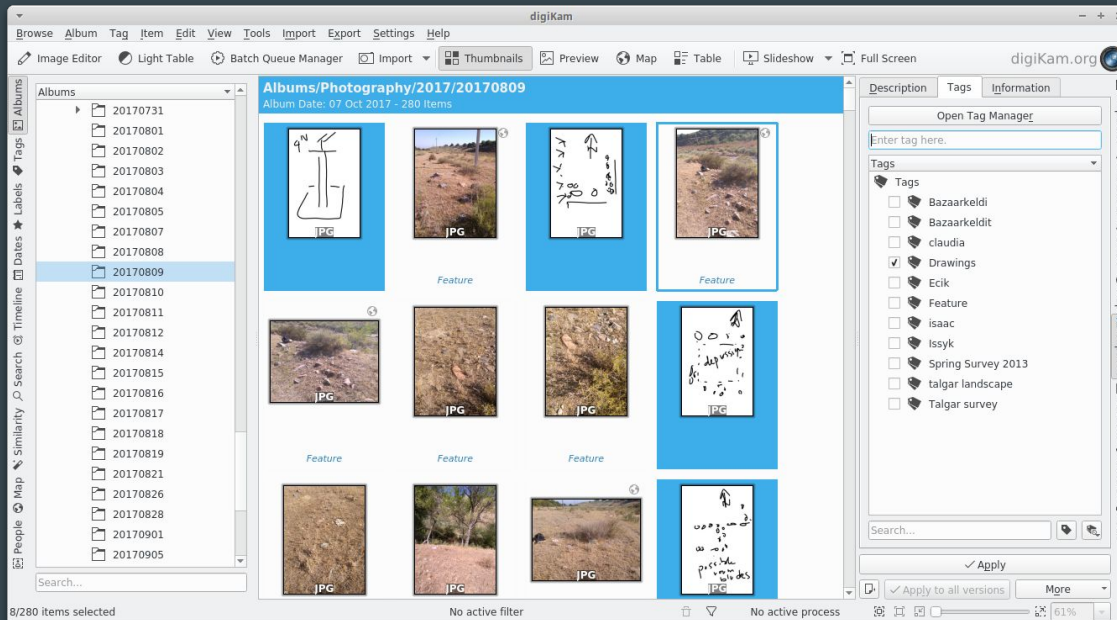
The spatial aspects of the project are best managed in a GIS. GRASS and QGIS are the “dynamic duo” that I recommend. QGIS for daily use in the field and for making nice maps, and GRASS for all follow-up analysis. They play very nicely together.

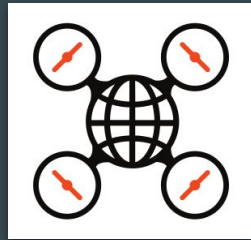
Post-processing includes topology correction and merging of data layers, creation of proper metadata, “project files” with styling and layering, advanced geospatial analyses, statistical analysis, and creation of publication-quality maps.



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

The screenshot displays a Linux desktop environment with several windows open:

- Terminal Window:** Shows the output of a Ceres Solver process, including iteration counts, initial and final costs, and termination status (CONVERGENCE). It also shows bundle setup and teardown commands, and the addition of various image files (e.g., DJI_0178.jpg, DJI_0214.jpg) to the reconstruction process.
- Task Manager Window:** Displays system performance metrics: CPU usage at 64%, 1103 processes running, and memory usage at 3%. It includes a graph of CPU usage over time.
- Task List Window:** Shows a list of running tasks with columns for PID, RSS, and CPU usage. The task `chromium-browser/chromium-browser --type=renderer --enable-pinch --field-trial-handle=11881925186718588514,824680820378283418...` is highlighted, showing it is using 8673 PID, 811.4 MiB of memory, and 0% CPU.
- System Monitor Window:** Provides a detailed view of system resources, including a CPU History graph and a table of CPU usage for each core (CPU1 to CPU52).

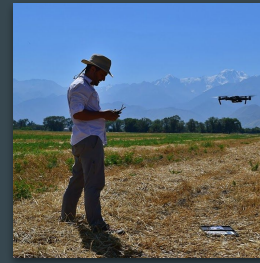
CPU	Usage	CPU	Usage	CPU	Usage	CPU	Usage
CPU1	69.3%	CPU17	73.3%	CPU33	70.6%	CPU49	80.8%
CPU2	70.9%	CPU18	70.7%	CPU34	69.7%	CPU50	65.3%
CPU3	68.3%	CPU19	68.0%	CPU35	71.0%	CPU51	69.6%
CPU4	69.6%	CPU20	69.2%	CPU36	67.0%	CPU52	66.7%
CPU5	68.0%	CPU21	70.9%	CPU37	68.7%		
CPU6	71.3%	CPU22	71.0%	CPU38	69.7%		
CPU7	69.0%	CPU23	70.0%	CPU39	71.0%		
CPU8	67.6%	CPU24	70.0%	CPU40	70.6%		
CPU9	67.6%	CPU25	70.3%	CPU41	72.0%		
CPU10	74.3%	CPU26	72.4%	CPU42	69.7%		
CPU11	69.0%	CPU27	70.4%	CPU43	71.0%		
CPU12	68.4%	CPU28	70.1%	CPU44	72.5%		
CPU13	69.2%	CPU29	69.1%	CPU45	69.3%		
CPU14	70.0%	CPU30	70.0%	CPU46	67.3%		
CPU15	71.6%	CPU31	70.1%	CPU47	68.7%		
CPU16	72.0%	CPU32	71.3%	CPU48	65.7%		

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

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.

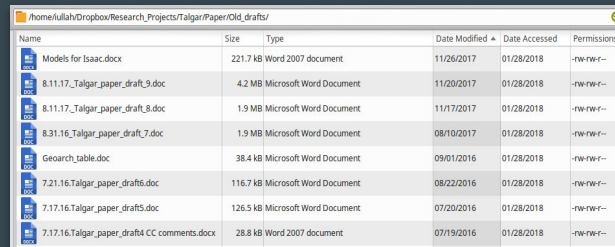
2. **Better:** The Cloud.

- 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.”

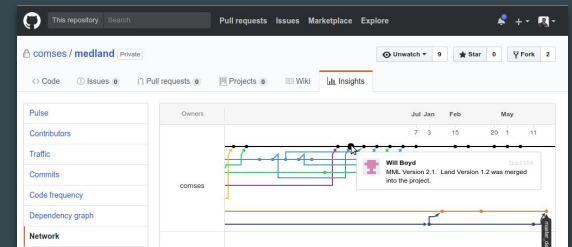
3. **Best:** Open Online Repositories.

- 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



Name	Size	Type	Date Modified	Date Accessed	Permissions
Models for Isaac.docx	221.7 KB	Word 2007 document	11/26/2017	01/28/2018	-rw-rw-r--
8.11.17_Talgar_paper_draft_9.doc	4.2 MB	Microsoft Word Document	11/20/2017	01/28/2018	-rw-rw-r--
8.11.17_Talgar_paper_draft_8.doc	1.9 MB	Microsoft Word Document	11/17/2017	01/28/2018	-rw-rw-r--
8.31.16_Talgar_paper_draft_7.doc	1.9 MB	Microsoft Word Document	08/10/2017	01/28/2018	-rw-rw-r--
Geoarch_table.doc	38.4 KB	Microsoft Word Document	09/01/2016	01/28/2018	-rw-rw-r--
7.21.16.Talgar_paper_draft6.doc	116.7 KB	Microsoft Word Document	08/22/2016	01/28/2018	-rw-rw-r--
7.17.16.Talgar_paper_draft5.doc	126.5 KB	Microsoft Word Document	07/20/2016	01/28/2018	-rw-rw-r--
7.17.16.Talgar_paper_draft4 CC comments.docx	28.8 KB	Word 2007 document	07/19/2016	01/28/2018	-rw-rw-r--



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

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”).

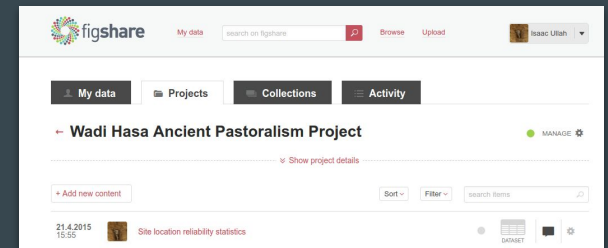
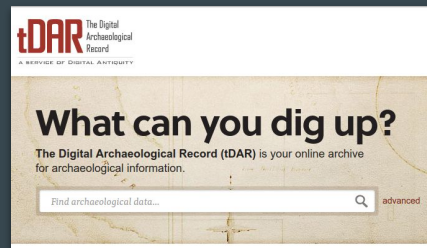
3. Static repositories (FigShare, Xenodo, OSF, Library Repositories):

- **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 data 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.

C-MAPLE
www.cmaple.org

C-MAPLE Home · Computational Models · Archaeological Data · Community ▾

COMMUNITY FOR MODELING AGRO-PASTORAL LIFEWAYS IN EURASIA

ARCHAEOLOGICAL DATA

Open datasets relating to agro-pastoral subsistence and culture

Pastureo Dataset
by Bryan K. Hanks
Human-Animal Relationships in the Eurasian Steppe Iron Age.

Talgar Alkhalil Fan Dataset
by Claudia Chang and Perry A. Tourtellotte
Settlement study of the Talgar Alkhalil Fan, Almaty Province, Kazakhstan.

Cross-cultural data for multivariate analysis of subsistence strategies
By Isaac L. Ulsh
These are datasets of human subsistence, mobility, demographic, and environmental variables for the 184 cultures of the Standard Cross Cultural Sample.

COMPUTATIONAL MODELS

Open-Source Computational Models of Agro-Pastoralism

M

The Mediterranean Landscape Dynamics Project's Modeling Laboratory. Read the documentation available at [Qeios](#) for information about this model.
added by Isaac L. Ulsh

Musical Chairs and Nice Musical Chairs models. Read the documentation available at [Qeios](#) for information about these models.
added by Andros Angelourakis

COMMUNITY WORKS

Community-generated white-papers, posters, conference presentations, and pre-prints

Modeling Agro-Pastoralism in Eurasia
"Lightning-Round" Forum
82nd Annual Meeting of the Society for American Archaeology
2017, Vancouver BC

Page: 1 / 37

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