Towards a data commons: Imagery and derived data from autonomous and remotely piloted aerial vehicles

Alice Fremand – UK Polar Data Centre, British Antarctic Survey

This report is an output of Work Package 4 of the Environmental Data Service (EDS) UKRI DRI Phase 1b grant.

November 2023

Table of Contents

Abstract 2

Introduction 2

The project 2

Data commons approach definition, methodology and scope 2

1. Towards a commons that supports the UAV community within NERC 3

Definition and scope of the UAV community 3

Identifying the needs, challenges, and vision of the UAV community 4

Results of the survey 4

The vision 6

2. Harmonising UAV data and metadata 7

Review of existing best practices and recommendations 7

EDS: an expertise for data stewardship that supports the UAV community within the commons 7

International recommendations and references 8

Harmonising metadata within the commons 9

UK Gemini metadata 9

Common UAV metadata 9

Use cases and recommendations per discipline 15

Best practices and recommendations 15

3. Skills, training, tools and other considerations to support the management of UAV data 15

Skills and training 16

NetZero, data storage and imagery 16

Controlled vocabularies and persistent identifiers 16

Software 17

Computational infrastructure 17

Privacy, licensing security 17

Summary and Conclusion 17

Recommendations and actions that will help support UAV management within NERC 18

Outputs 18

List of activities 19

Acknowledgements 19

References 20

Appendix – use case examples 20

Abstract

Scientists are increasingly using autonomous and remotely piloted aerial vehicles (commonly referenced here as UAVs) to collect large volume of data in a range of different scientific domains. These data are key to providing environmental information, such as terrestrial or marine monitoring including species detection or collecting atmospheric information. However, the volume of data collected, and the lack of standard workflows make their sharing difficult.

The aim of this project is to enable improvement of collection and management of data derived from UAVs by establishing best practices and metadata recommendations through a data commons approach within NERC Environmental Data Service (EDS).

Following a data commons approach, the project first defines who the main users and stakeholders are, here referenced as the UAV community and identifies their needs, challenges and vision. Then, we review existing UAV data management practices and recommend a new data workflow to promote Findable, Accessible, Interoperable and Re-usable (FAIR, Wilkinson, 2016) data collection and management of UAV data. Finally, we propose a list of topics that needs to be further investigated to achieve the UAV community vision in the future.

We hope that the promotion of these recommendations will improve the reuse of data from UAVs and foster collaborations for new research studies using these platforms and the proposed data commons approach will help create new synergies within the EDS.

Introduction

The use of autonomous and remotely piloted aerial vehicles (commonly referenced here as UAVs) for scientific purposes has significantly increased in the last 10 years with the development of miniaturised sensors and the availability of performant autopilot hardware and software amongst others. It is now a cost-effective way to collect environmental data which has opened the doors to new scientific discoveries. However, the lack of UAV data management best practices, the volume of data collected, and the lack of open hardware and software tools (Irujo, 2023) limit the full potential reuse of these data.

The goal of this study is to take a data commons approach to establish best practices for the collection, analysis, use and re-use of data acquired from UAVs.

The project

This study has been carried out as part of the Natural Environment research Council (NERC) Environment Data Service (EDS) Digital Research Infrastructure (DRI) Phase 1b Working Package (WP) 4 entitled, ‘Communications, Use Cases and Project Management’. The goal is to propose an example of a data commons approach applied to the environmental data collected from UAVs that has the potential to utilise tools being explored as part of the WP2 entitled “EDS Integration Experiments”. The data commons approach is better defined as part of the WP1: “Building Interoperability – a NERC Data Commons RoadMap” where the data commons concept is generalised under the term “asset commons” while WP2 proposes general recommendations for metadata and semantic interoperability.

Data commons approach definition, methodology and scope

The concept of data commons can be defined as “a cloud-based data platform with a governance structure that allows a community to manage, analyse and share its data” (Grossman, 2023). But other definitions and explanations are useful to better understand the concept. In his blogpost, Grossman also defines a data commons as a way to, “organise data for a field or a discipline. More formally, a data *commons* brings together (or co-locates) data with cloud computing infrastructure and commonly used software services, tools and applications for managing, analysing and sharing data to create an interoperable resource for a research community” (Grossman et al., 2016; Grossman, 2018). As part of WP1, Gordon Blair defines an asset commons as “a common place supporting asset discovery, access, interoperability and asset re-use, tailored for a community and managed by that community for the common good”.

To build this data commons approach in the context of the management of UAV data, we followed the 10 lessons identified by Grossman, 2023 with a focus on:

1. **Users**: the commons is built for a specific community (the UAV community) with a specific set of research challenges (lesson 1). Through different workshops and a survey, we identified the needs, challenges and vision of environmental scientists in managing UAV data.
2. **Data and metadata**: Following the second and third lessons of Grossman, 2023, we recommend best practices to curate and harmonise the metadata so the data can be reused to generate new research discoveries.
3. Investigating the **skills, tools and other considerations** that are needed to support the community with an emphasis on the data stewardship of UAV data. This section aims at reducing barriers to access data, increase usage (lesson 4) and touch on a few elements of lessons 5 to 10.

This work focus on the management of digital data and proposes recommendations on how best to manage data derived from UAV through recommendation on harmonisation of metadata and best practices for the collection to publication of UAV data. Although a data commons is defined as “a cloud-based data platform with a governance structure”, no data infrastructure or governance framework have been set up as part of this work. Lesson 4 to 10 with a focus on APIs, access control or infrastructure are only briefly described in section 3. Strategy and additional information regarding how the governance framework of the data commons and tools will be implemented is detailed as part of WP1, WP2 and WP3.

1. Towards a commons that supports the UAV community within NERC

Definition and scope of the UAV community

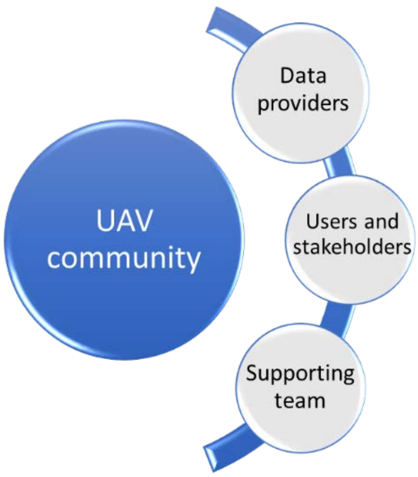

Figure 1 The UAV community

For this project, the UAV community within the commons is defined as any person, group or organisation with an interest in scientific UAV data. The community is composed of:

- The **data providers**: scientists, engineers or mapping staff who collect, analyse, or process UAV data and share them to the UAV community of the commons.
- The **data users and stakeholders**: any person, group, or organisation who use UAV data from the commons. This could include scientists who reuse the data, as well as weather services, general public, FCDO, or UK local management (councils, town planning) for instance.
- A **supporting team** who provides advice and tools to implement the commons.

Identifying the needs, challenges, and vision of the UAV community

One key element to providing a commons is to understand the needs, challenges and vision of the community it serves. As part of this project, we participated in different workshops organised by the NERC NetZero Airborne Capability group on uncrewed aerial systems (NZArC), carried out interviews with scientists from different domains and ran a survey to better understand the UAV community within NERC (see list of activity for additional details). As part of these activities, we wanted to identify:

- The primary usage of UAVs within NERC – Are UAVs used primarily for scientific, media outreach or operational purposes?
- The type of platform used – A large variety of UAV platforms exists. Do scientists principally use quadcopter or fixed-wing platforms?
- The scientific domains – What are the main applications of UAV research surveys?
- The type of data collected – In this report, we distinguish imagery from non-imaging sensor data. Imagery or image data refers here to imaging sensor data that can be processed to produce an image of an area while non-imaging sensor data corresponds to all other sensor data, mostly single value data or "point" data.
- The challenges and opportunities that exist with regards to the data stewardship aspects of UAV.
- The vision of the end-users regarding data collection and reuse.

Results of the survey

The survey was undertaken from 29/06/2023 to 10/09/2023. We publicised the survey through the NERC research news bulletin and at workshops and conferences including the NZArC workshops and the NERC Digital gathering conference held in Cambridge on 10-11 July 2023. In total, we received responses from 21 participants.

Participants are mostly funded by NERC and working at universities, NERC institutes or environment and government agencies as shown by the figure below.

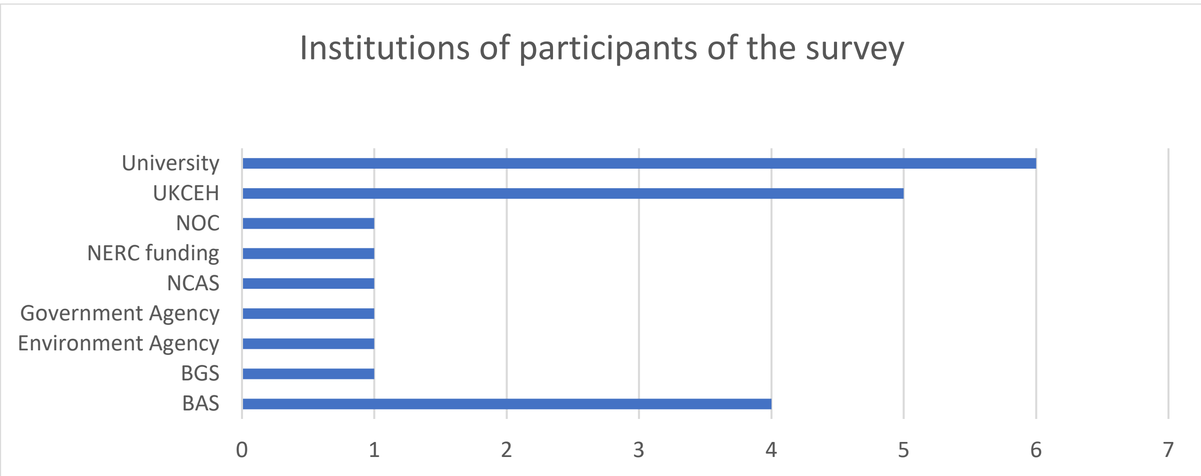

Figure 2 Institution of the participants of the survey

- Type of platforms

The type of platforms used by the participants of the survey consists mainly of quadcopter or fixed-wing remotely piloted aerial systems. Other platforms include vertical take-off and landing (VTOL) UAVs or hexacopters.

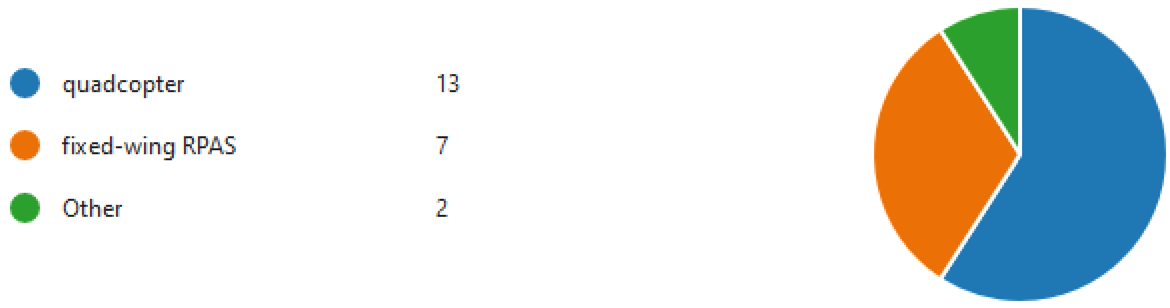

Figure 3 Type of platform used by the participants of the survey

- Usage

Among the participants of the survey, UAVs are mainly used for scientific purposes (70%) but they are also used for operational/support (19%), media outreach (7%) or teaching (4%).

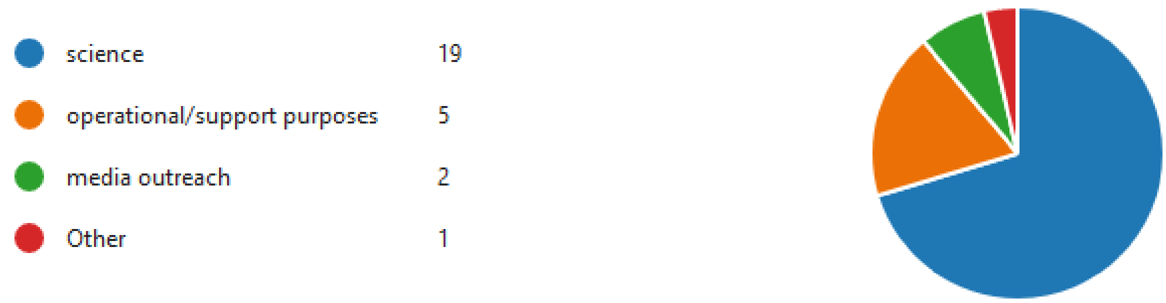

Figure 4 Usage of UAVs by the participants of the survey

- Discipline

As shown in the chart below, the disciplines represented are very diverse and well spread amongst the hydrosphere, atmosphere, geosphere, biosphere and other (mainly cryosphere) subject of study.

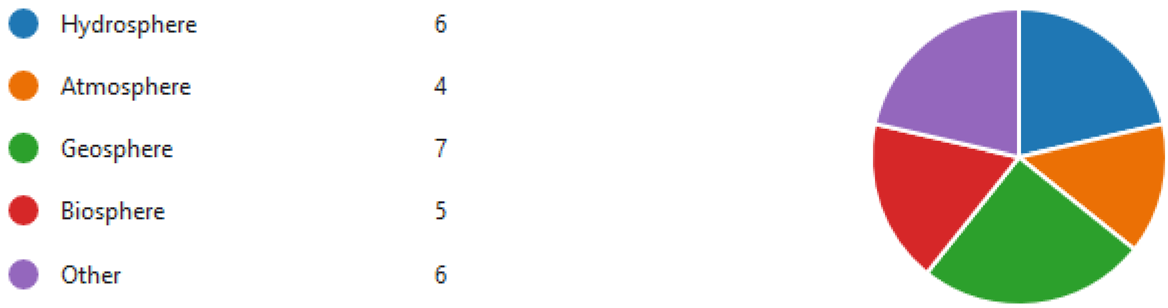

Figure 5 Disciplines of the participants

- Type of data

The data collected from UAVs are primarily images and non-imaging sensor data. Other types of imagery data include thermal and LIDAR imagery.

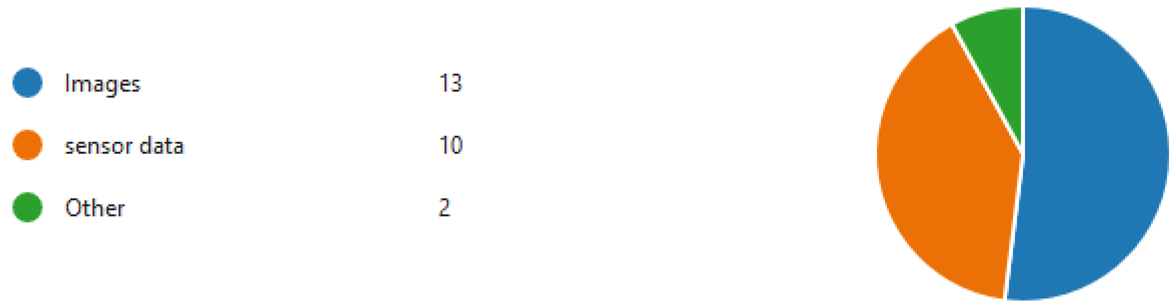

Figure 6 Type of data collected by UAVs

- Data published

Only 25% of the participants have already published their data in a repository.

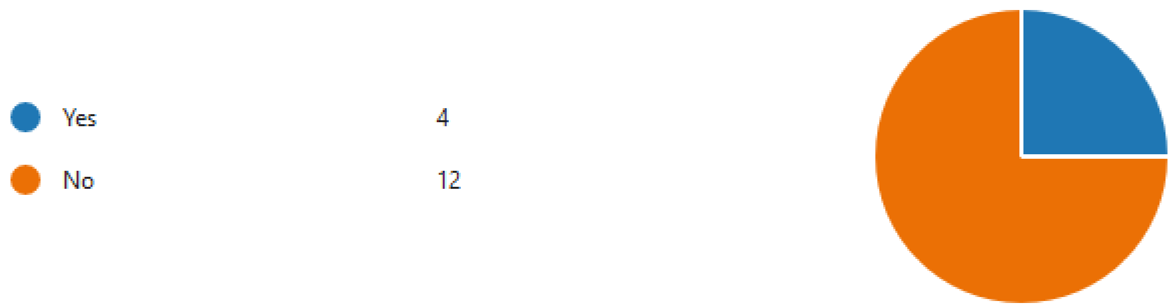

Figure 7 Number of participants who published their data in a repository

- Summary

The results of the survey and discussions with the UAV community highlighted the importance of the data as the main output of any UAV survey. UAV data consists primarily of imagery (>50%) and non-imaging sensor (42%) data that gives unique information about the environment for a specific date and space. Although UAVs can be used for different usages (see Figure 4), this report focuses on scientific purposes and the collection of digital data (physical samples will not be the focus of this report). UAV data are an asset that we need to preserve for the long term as with any other environmental data collected within NERC. However, to date, very few UAV datasets have been published. From the 21 respondents only 4 stated that they had published UAV data.

The NZArC workshop showcased the challenges and opportunities that exist with regards to the data stewardship aspects of UAV. The challenges to manage UAV data are manyfold: they can generate high volume of data, some parameters such as the payload and stability of the platforms can have a high influence on the quality of the data and they are used by a very diverse community including but not limited to atmospheric, marine, geoscience, and environmental sciences. In addition, the need for standard and open format data and metadata has been recognised by the community.

The vision

The scientific community wants to have their data easily discoverable through a common data portal and be able to run computational analysis directly in the cloud using computational notebooks or Google-Earth Engine equivalent. Being able to discover what has already been done through a common data portal will help reduce duplication and replication of dataset, which is key in decreasing the carbon footprint and cost of data storage.

As the field is still new, there is a remarkable opportunity to make these data FAIR across NERC using a common data management approach. But to implement this approach, we need to:

- Have complete metadata with information about the platforms, the sensors, the calibration parameters and the project so the community can assess the quality, and understand the potential bias of the data.
- Be in line with the international recommendations (Barbieri et al., 2023; Wyngaard et al. 2019)
- Recommend controlled vocabularies especially for instruments.
- Have data in a standardised and open format that both comply with the UAV and domain community requirements.
- Have transparent workflows and common operating and calibration procedures.
- Develop tools and computational infrastructure to serve the data.
- Develop skills and promote best practices for UAV data management.

In section 2, we propose a set of metadata fields to best describe UAV data and in section 3, we investigate how best to support scientists in managing and publishing their data collected from UAVs.

2. Harmonising UAV data and metadata

Scientists agree that having standardised metadata would help sharing and publishing UAV data. In this section, we first review existing best practices and recommendations for the stewardship of UAV data, then we discuss the common UAV metadata requirements through the result of a survey and finally give some recommendations for domain specific data through three use cases.

Review of existing best practices and recommendations

EDS: an expertise for data stewardship that supports the UAV community within the commons

Figure 8 The data centres part of the Environmental Data Service

To implement a data commons approach, the expertise of NERC Environmental Data Service (EDS) and its data centres is essential to advise on UAV data management. The EDS is a trusted UK facility providing data stewardship services ensuring environmental data of long-term value FAIR. The NERC EDS are composed of 5 data centres, each with a specific thematic:

- BODC – British Oceanographic Data Centre (marine)
- CEDA – Centre for Environmental Data Analysis (Atmospheric, Earth Observation, and Solar and space physics)
- EIDC – Environmental Information Data Centre (hydrology, terrestrial and freshwater data resources)

- NGDC – National Geoscience Data Centre (geoscience)
- UKPDC – UK Polar Data Centre (polar and cryosphere)

The management of the UAV data is highly linked to the discipline the data are collected for. Each NERC data centre can provide its own discipline expertise and experience managing similar data. In particular, we can note CEDA's expertise in managing large volumes of data with specific experience managing airborne datasets including data from the following projects:

- FAAM – Facility for Airborne Atmospheric Measurements
- ARF – Airborne Research Facility
- EUFAR – European Facility for Airborne Research
- The Field Spectroscopy facility

Although the UAV data have some specificities, there are high similarities with airborne datasets. As for other airborne campaigns, UAV data are collected on moving platforms, with a wide range of instruments and a wide range of scientific goals. The same principles established for the management of airborne datasets can be applied to manage the UAV data while acknowledging some of their specificities:

- UAV data can generate high volumes of data compared to airborne surveys (e.g. many overlapping images rather than one large image in optical surveys). This can have some impacts on which data to keep. There might be some cost/carbon footprint considerations.
- Payload and stability information may have a high impact on the quality of the data.
- There are potentially many flights per day compared to only one or two for airborne surveys.
- The concept of multiple UAV operating simultaneously (sometimes referred as swarms) is new. Indeed, there are now more and more surveys operating several UAVs simultaneously to collect data.

In the context of UAV data management, we can also utilise BODC's expertise on Persistent Identifiers (PIDs) for instruments and experience publishing data from autonomous marine vehicles.

International recommendations and references

In addition to the expertise of the NERC EDS, a number of initiatives and projects provide recommendations that can help define the best practices for UAV data stewardship. The following groups from the Research Data Alliance (RDA) and the Earth Science Information Partners (ESIP) are of particular interest:

- RDA – Small Unmanned Aircraft Systems' Data Interest Group
- ESIP – Drone Cluster

For ontologies, the following works have been identified as relevant:

- **LANDRS – Linked And Networked DroneS** – Sloan Foundation funded project to build open source tools for managing scientific data on drones through the use of web standards and linked data technologies. https://github.com/landrs-toolkit
- **RDA – InteroperAble Descriptions of Observable Property Terminology WG (I-ADOPT WG)** https://www.rd-alliance.org/groups/interoperable-descriptions-observable-property-terminology-wg-i-adopt-wg
- **SensorML –** https://www.ogc.org/standard/sensorml/

For metadata, we identified the work from Barbieri et al., 2023 and Wyngaard et al., 2019 of particular interest.

Harmonising metadata within the commons

As highlighted previously, UAV data are collected for different applications and topics. To harmonise UAV metadata, we need to acknowledge this specificity when providing metadata and data recommendations. Indeed, the data needs to follow the metadata requirements of the UAV community but also the domain specific recommendations (See Figure 9).

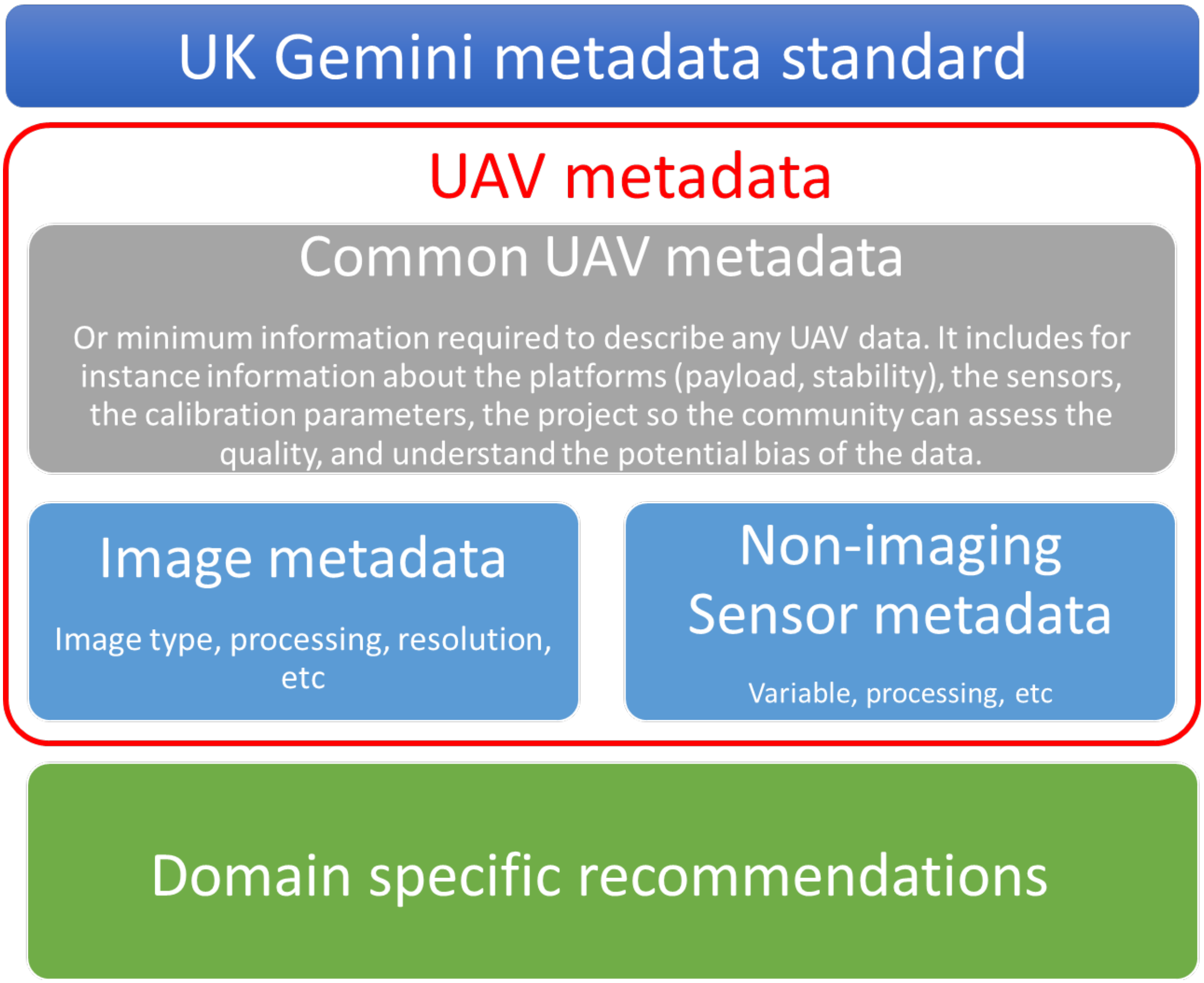

Figure 9 Metadata components for UAV data

UK Gemini metadata

"GEMINI" is the UK geographic metadata standard. It provides guidance on how to publish geographic metadata in a way that conforms to UK government guidelines and the relevant ISO standards. This standard is used across the EDS, and documents how to describe any (geo-)spatial dataset. It comprises a title, an abstract, authors and contributors, lineage, references, dates, license, and spatial information. NERC EDS produces, and regularly updates, good practice guidelines for these fields. It is thus recommended to use the latest recommendations when recording the Gemini metadata.

Common UAV metadata

In addition to the generic Gemini metadata and following the work of Barbieri et al., 2023 and Wyngaard et al., 2019, we have identified a set of additional metadata that are common to describe any UAV data and necessary for their reuse (referred as "Minimum Information Framework" in Barbieri et al., 2023). In this document, we call them the "common UAV metadata". We conducted a survey to better understand the importance of the terms identified in Barbieri et al., 2023 for our specific NERC UAV community.

The survey investigated the need for specific metadata related to:

- The UAV platform and its sensor type (UAV and sensor make and model, sensor information)
- The UAV survey (the extent, dates and sensing target)
- The image data (image type, resolution, and processing information)
- The non-imaging sensor data (sensor data type and processing)
- The sortie (flight log)

For each term, participants were asked to vote on the importance of the term and say whether they:

- Can't use the data without it (MUST)
- Won't use the data without it (SHOULD)
- Can use the data without it but nice to have (COULD)
- Don't need it (WON'T)
- N/A

Result of the survey

In the section below, we present the results of the survey conducted from 15/06/2023 to 10/09/2023. The survey was advertised at the NZArC and the NERC Digital Gathering workshops, as well as through the NERC research community news bulletin. In total, 21 scientists participated in the survey.

- The UAV platform and its sensor type (UAV and sensor make and model, sensor information)

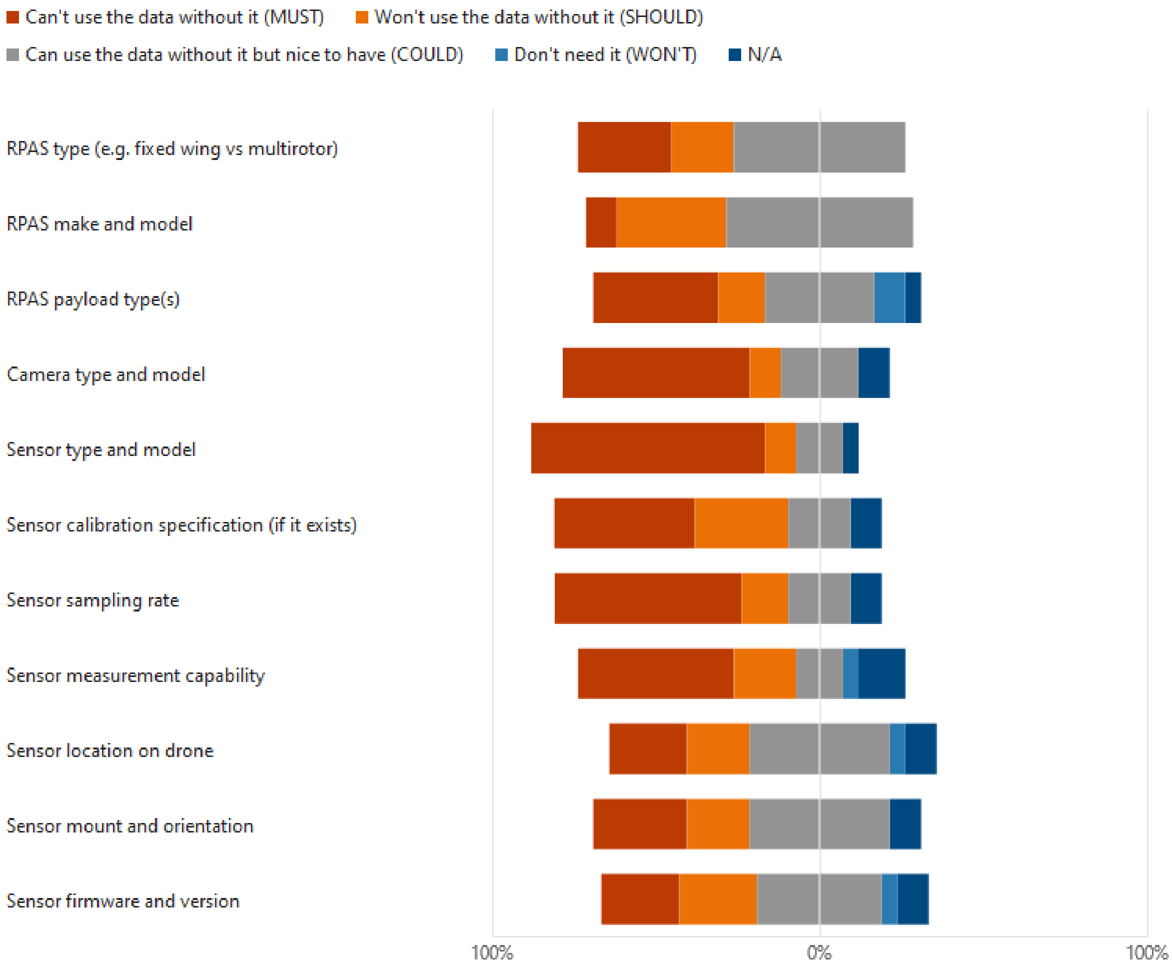

Figure 10 Metadata for the UAV platform and sensor type

- The UAV survey (the extent, dates and sensing target)

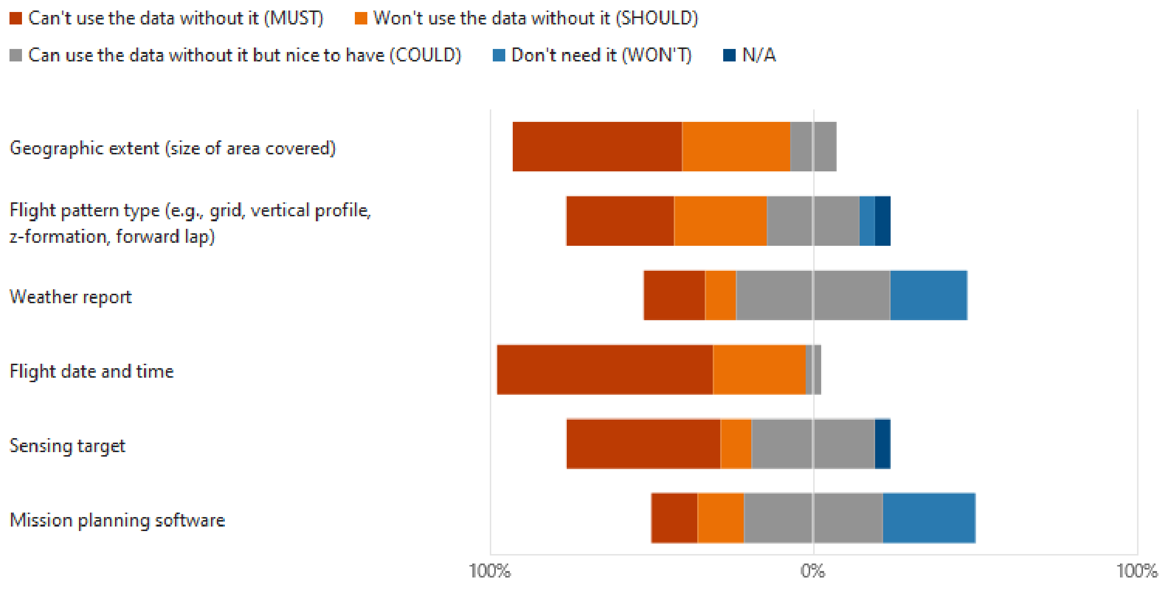

Figure 11 Metadata for the UAV survey

- The image data (image type, resolution, and processing information)

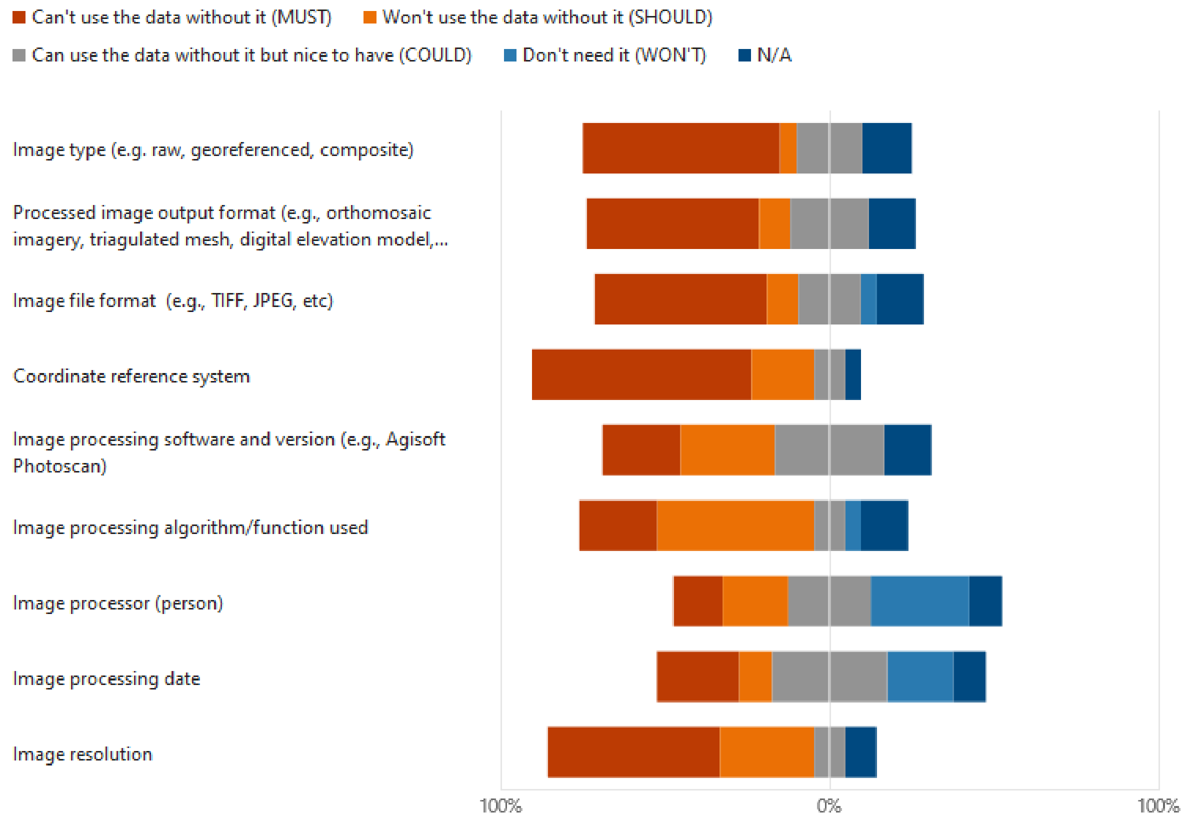

Figure 12 Metadata for image data

- The non-imaging sensor data (sensor data type and processing) and sortie (flight log)

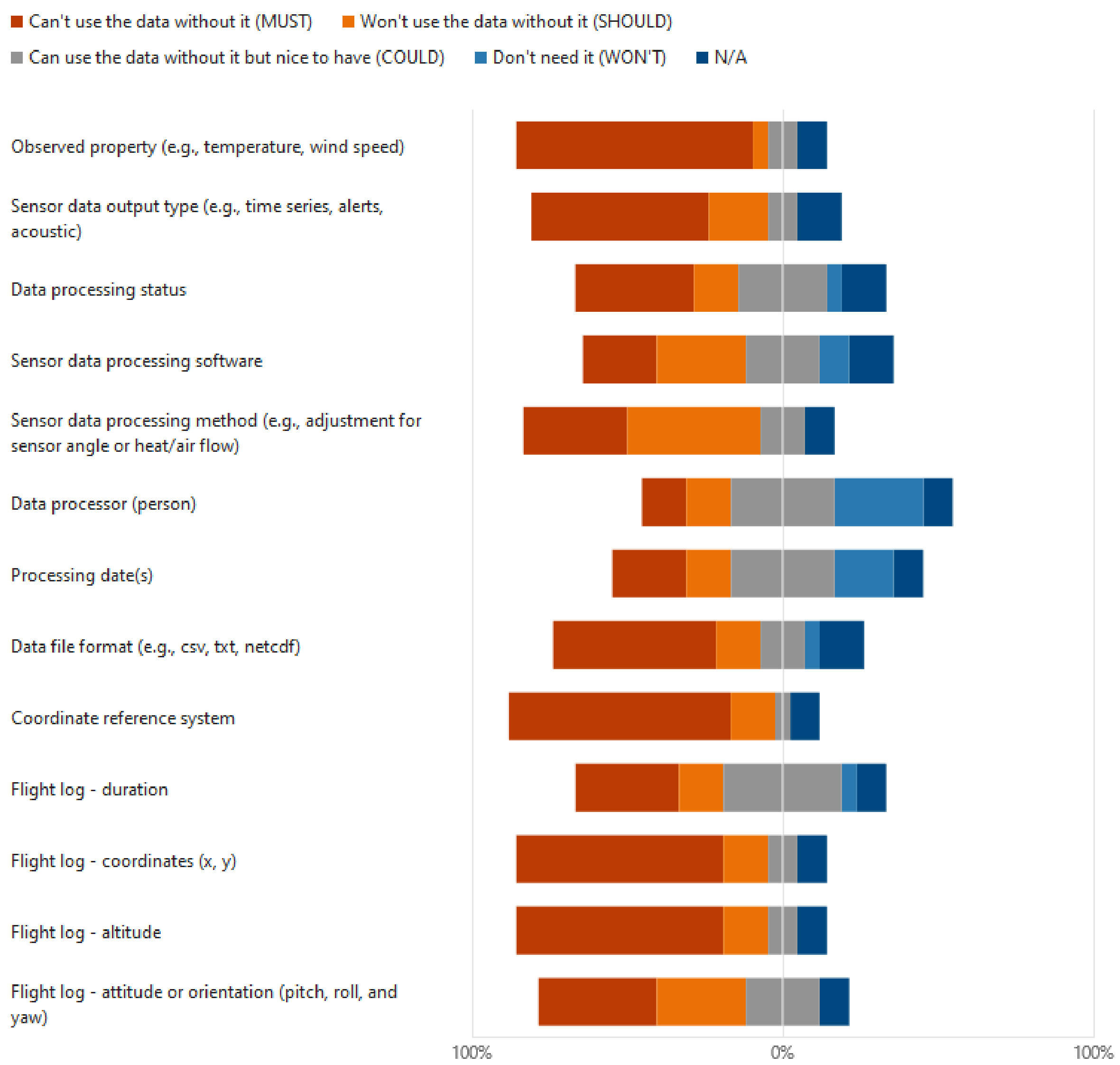

Figure 13 Metadata for sensor data

Recommended common UAV metadata

The table below gives the recommendations on terms used to describe UAV datasets following the results of the survey. The requirements level for each element are either 'M', 'R', 'C', 'O' indicating:

- **Mandatory (M):** the element must be filled in under all circumstances. (≥80% of responders can't use the data or won't use the data without it)
- **Recommended (R):** the element should be used to allow the most potential reuse of the data. (40% - 80% of can't use the data or won't use the data without it)
- **Conditional (C):** the element must be completed for the resource type being described if certain conditions are met e.g. image information are only necessary when a camera is used.
- **Optional (O):** the element may be filled in if desired. Creators are encouraged to populate optional elements if they have the knowledge, as this provides more detailed information for people to search on, allowing better access to, and re-use of data.

Table 1 Metadata recommendations specific to UAV datasets (M – Mandatory, C – Conditional, R – recommended, O – optional)

The UAV platform and its sensor type	
RPAS type (e.g. fixed wing vs multirotor)	R
RPAS make and model	R
RPAS payload type(s)	R
Camera type and model	C - R
Sensor type and model	C - M
Sensor calibration specification	C - R
Sensor sampling rate	R
Sensor measurement capability	R
Sensor location on drone	R
Sensor mount and orientation	R
Sensor firmware and version	R
The UAV survey	
Geographic extent (size of area covered)	M
Flight pattern type (e.g., grid, vertical profile, z-formation, forward lap)	R
Weather report	O
Flight date and time	M
Sensing target	R
Mission planning software	O
Imagery data	
Image type (e.g. raw, georeferenced, composite)	C - R
Processed image output format (e.g., orthomosaic imagery, triangulated mesh, digital elevation model, dense point cloud)	C - R
Image file format (e.g., TIFF, JPEG, etc)	C - R
Coordinate reference system	C - M
Image processing software and version (e.g., Agisoft Photoscan)	C - R
Image processing algorithm/function used	C - R
Image processor (person)	C - O
Image processing date	C - R
Image resolution	C - M
Non-imaging sensor data	
Observed property (e.g., temperature, wind speed)	C - M
Sensor data output type (e.g., time series, alerts, acoustic)	C - M
Data processing status	C - R
Sensor data processing software	C - R
Sensor data processing method (e.g., adjustment for sensor angle or heat/air flow)	C - M
Data processor (person)	C - O
Processing date(s)	C - R
Data file format (e.g., csv, txt, netcdf)	C -R
Coordinate reference system	C - M
The sortie	
Flight log - duration	R
Flight log – coordinates (x, y)	M
Flight log - altitude	M
Flight log - attitude or orientation (pitch, roll, and yaw)	R

Use cases and recommendations per discipline

To better understand the specificities of each domain, the following four questions have been answered by three volunteers:

- Please could you provide information about the type of research you undertake using autonomous or remotely piloted aerial systems?
- what are the specific characteristics of the data collected in your domain? For instance, is it large datasets, lots of files, or/and proprietary format? Is there challenges that you need to tackle regarding the data you collect to publish them? What would you need to facilitate the publication of the data?
- What are the standards you use (or plan to use) to publish your data? Is there some example of data catalogues/repositories that provide good examples of data publication?
- If there are some recommendations you would like to give to other scientists publishing data of the same type, what would you say?

The answers covered several disciplines with an example in atmospheric, hydrology and geophysics sciences. Full responses of participants can be found in the Appendix – use case examples.

Summary

Answers to the questions showed that there is an evident lack of standard approach to metadata, file format and overall data management practices. Some disciplines are more advanced than others with the development of vocabularies and set of metadata fields specific to disciplines. The National Centre for Atmospheric Science (NCAS) for example worked closely with CEDA to publish atmospheric UAV data, developing new workflows and vocabularies. In atmospheric sciences, the International Society for Atmospheric Research using Remotely-piloted Aircraft (ISARRA) provides a forum for the exchange of knowledge, experience and ideas on the various aspect of atmospheric and related environmental research using UAVs and ACTRIC (Aerosol; Clouds and Trace Gases Research Infrastructure) proposes some data standardisation and calibration including for data derived from UAVs. However, not all disciplines are as advanced, and work still needs to be done to provide case-by-case recommendations, but general rules can apply:

- Work closely with the NERC data centre of your choice
- Liaise with the discipline community for best practice examples
- Use standards and open formats

Best practices and recommendations

To help users with collecting and publishing data derived from UAV, a specific drone data management handbook has been created (Fremand, 2023). From project planning to the publication of the data, the document guides scientists through the different steps for the publication of Findable, Accessible, Interoperable and Reusable (FAIR, Wilkinson, 2016) data. The document presents the general flowchart from planning, collection, processing and publication.

3. Skills, training, tools and other considerations to support the management of UAV data

The goal of this report is to provide a first common guidance on how best to manage UAV data from collection to publication. The document focuses on metadata standardisation and how best to record these. But there is still a long road to achieve the vision drafted by the scientists and many questions still to resolve. In this section, we present a few suggested topics that are relevant to the UAV community that will help achieve the vision.

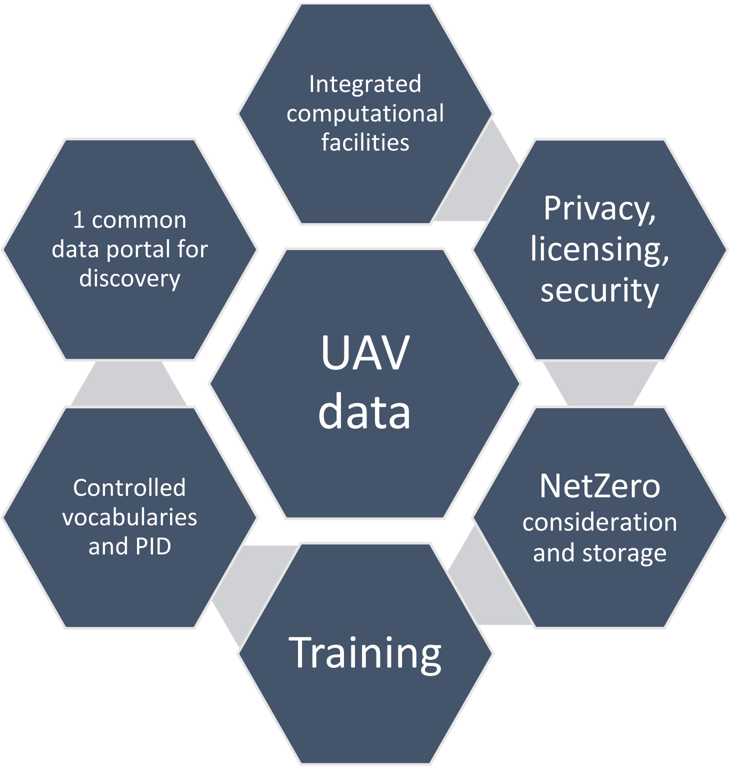

Figure 14 The UAV ecosystem

Skills and training

The different NetZero Airborne capability workshops highlighted the critical lack of understanding within the community of some data management principles that needs to be tackled. Communicating about the NERC data policy, the role of EDS and standardised procedures will be essential to promoting the recommendations for UAV data management.

A first step to advance UAV data management within NERC is the publication of the drone data management handbook as introduced in Section 2. Although the document will need to be updated and properly reviewed by the community, the document provides a first introduction to best manage UAV data in the future.

To promote best practices, it is recommended that an introduction to data management principles is added to any curriculum related to use of UAVs. The principles should be introduced to both scientists and UAV operators as most of the parameters should be recorded at the time of data collection.

NetZero, data storage and imagery

As highlighted in the survey, 50% of the data collected consists of imagery data. Those are often very large size datasets and can easily represents tens of terabytes at the end of a survey. Questions regarding safeguarding all of these data need to be considered for potential reusability over cost and carbon footprint considerations especially when the data of interest often represents a small subset of the data collected. This question is not specific to UAV data but can be broadened to other large size datasets collected as part of NERC funding. Additional guidance for the publication of imagery data would be useful for the community as no common approach is available within the EDS.

Controlled vocabularies and persistent identifiers

To improve interoperability, the use of common vocabularies is recommended. The work package 2 on EDS integration experiments subsection on semantic interoperability proposes recommendations on vocabularies for descriptions of observations across the EDS. In the context of UAV, further investigation is needed following on the overall work of EDS to facilitate the use of vocabularies and international working groups such as the:

- **LANDRS – Linked And Networked DRoneS** - Sloan Foundation funded project to build open source tools for managing scientific data on drones through the use of web standards and linked data technologies. https://github.com/landrs-toolkit
- **RDA – InteroperAble Descriptions of Observable Property Terminology WG (I-ADOPT WG)** https://www.rd-alliance.org/groups/interoperable-descriptions-observable-property-terminology-wg-i-adopt-wg
- **NCAS AMF controlled vocabularies:** https://github.com/ncasuk/AMF_CVs/
- **SensorML –** https://www.ogc.org/standard/sensorml/

Modern-day research is also increasingly using persistent identifiers (PIDs) to uniquely and permanently identify an entity, an object like an instrument or a platform, a person or an institution to link them to a dataset, a project, or a scientific paper. This work might be useful to easily link UAV metadata and instruments if those happen to be identifiable via a PID from a common vocabulary. But the domain is still emerging with a number of challenges to overcome (de Castro et al. 2023)

Software

When using UAV, scientists are facing software challenges. In the UAV community, most of the software and hardware are proprietary. In a recent literature review of applications in agro-environmental monitoring, Irujo et al., 2023 showed that more than 80% of studies are using fully closed source drones, and more than 90% of studies use proprietary closed-source software for image processing. This has high consequences on cost linked to software subscriptions but also on the format of the exported data which can often only be read by proprietary software reducing their reuse. In addition, each platform has its own software that requires specific data and IT expertise. The software is also rapidly obsolete, with the need of scientists to train regularly on how to use new versions.

Computational infrastructure

To answer the vision of the scientists, the need for standardised data and metadata is only a first step to providing a full working computational infrastructure. The different workshops have highlighted the need for high performance computing and access to near-real time datasets with user-friendly interface and data visualisation tools. Due to high volume of data to process, the need for near-data processing capability is crucial.

Privacy, licensing security

Another important aspect of the use of UAV data is privacy, licensing, and security. As highlighted in previous sections, imagery is one of the principal types of data collected from UAV. Those might contain personal data which are protected by the General Data Protection Regulation (GDPR) and the Data Protection Act 2018. Other data might also be protected via the environmental Protection Act which prevent scientists from disclosing the location of endangered species for instance.

Summary and Conclusion

The development of miniaturised sensors, attractive prices and improvements in the stability of UAVs have seen their use increased in scientific studies in the last few years. The data generated are often huge in size and difficult to manage.

To facilitate UAV data management, the project used a data commons approach focusing on data users and harmonisation of data and metadata. As part of this project, we worked closely with the UAV community to give them tailored recommendations on how best to manage their data. Following the

10 lessons from Grossman, we also looked sharply at the core issues of the data types, metadata, and vocabularies to work towards interoperability.

As part of this project, we carried out a survey and participated in different workshops to identify and define the UAV community, their needs, challenges and vision. Then, we analysed their practices to propose guidelines and recommendations for UAV data management and comply with the FAIR principles.

One key output of this project is the creation of the UAV data management handbook which provides the key information to the UAV community on how best to manage their data. From data collection to publication, the document presents a workflow, log sheets and metadata recommendations to help data collectors record the required information for data reuse.

In addition, we identified additional topics that need to be further investigated or developed to achieve the UAV community vision in the future. This includes:

- the development of training opportunities for scientists to learn more about data management,
- further investigations towards data storage and NetZero,
- recommendations on privacy, licensing and security,
- the use of controlled vocabularies and PIDs,
- considerations to use open software,
- the development of a specific computational infrastructure to allow online data visualisation and/or cloud-processing.

We hope that the promotion of the generated best practices and data workflow will improve the reuse of autonomous and remotely piloted aerial systems and foster collaborations but we acknowledge that this work is only a first step to achieve a data commons vision for the UAV community.

Recommendations and actions that will help support UAV management within NERC

1. Harmonise data and metadata within NERC following the recommendations of the UAV data management handbook.
2. Include recommendations in the digital stewardship wizard (https://ds-wizard.org/) to promote best data management practices for UAV data.
3. Continue the development of the UAV data management handbook by adding recommendations for specific scientific domains.
4. Investigate and promote the use of vocabularies and PIDs to be used for UAV data.
5. Get involved in international groups focusing on UAV data management best practices (such as RDA, ESIP)
6. Develop strategies to best publish imagery and large size datasets and include them in the UAV data management handbook
7. Train scientists to improve their skills in data management
8. Promote the use of open software when collecting and analysing UAV data
9. Integrate data into a computational infrastructure
10. Develop APIs for easy access to the data
11. Propose recommendations on privacy, licensing and security

Outputs

1. UAV data management handbook
2. This report

List of activities

Deliverables and tasks	Original due date	Status	Comments
Attend WP1 workshop	25-26/05/2023	Completed	
Attend 1st workshop on NERC NetZero Aerial Capability on uncrewed aerial vehicles	05/06/2023	Completed	Additional information: here
Attend 2nd workshop on NERC NetZero Aerial Capability on uncrewed aerial vehicles	20/06/2023	Completed	Additional information: here
Presentation at NZArC	20/06/2023	Completed	Link to the presentation: here
Create a survey to better understand the usage of UAVs	10/06/2023	Completed	
Advertise survey through different channels	01/07/2023	Completed	Survey sent out to the participants of the NZArC workshop and the NERC research community news bulletin
Prepare abstract for NERC Digital Gathering workshop	15/06/2023	Completed	
Attend NERC Digital Gathering workshop	12/07/2023	Completed	Link to poster: here
Meeting with scientist on discipline 1	15/09/2023	Completed	
Meeting with scientist on discipline 2	15/09/2023	Completed	
Meeting with scientist on discipline 3	15/09/2023	Completed	
Attend WP2 workshop	21/06/2023	Completed	
Write report	01/11/2023	Completed	
Recommendations for NERC	01/11/2023	Completed	

Acknowledgements

I would like to thank all the scientists who took the time to answer the questions of the survey. In particular, I would like to thank Dr Alvaro Arenas Pingarron, Dr Tom Jordan, Dr Barbara Brooks and Dr Charles George for taking part in the use case study. Their help was instrumental in defining the best practices for the UAV community and I look forward to continuing the work together.

Funding came from the NERC EDS Digital Research Infrastructure phase 1b WP4 project.

References

Barbieri, Lindsay, Jane Wyngaard, Sarah Swanz, and Andrea K. Thomer. 'Making Drone Data FAIR Through a Community-Developed Information Framework'. *Data Science Journal* 22 (2023): 1–1.

de Castro, Pablo, et al. "Persistent identifiers for research instruments and facilities: an emerging PID domain in need of coordination." (2023). https://doi.org/10.5281/zenodo.7330372

Fremand, Alice. 2023, UAV data management handbook. UK Polar Data Centre, British Antarctic Survey. https://nora.nerc.ac.uk/id/eprint/536392/

Grossman, R.L. Ten lessons for data sharing with a data commons. *Sci Data* **10**, 120 (2023). https://doi.org/10.1038/s41597-023-02029-x

Grossman, R.L. 'A Proposed End-To-End Principle for Data Commons'. (2018). https://medium.com/@rgrossman1/a-proposed-end-to-end-principle-for-data-commons-5872f2fa8a47

Grossman, Robert L., Allison Heath, Mark Murphy, Maria Patterson, and Walt Wells, "A Case for Data Commons: Toward Data Science as a Service," Computing in Science & Engineering 18, no. 5 (2016): 10–20.

Irujo, Gustavo Pereyra, Paz Bernaldo, Luciano Velázquez, Antoni Pérez, Celeste Molina Favero, and Alejandrina Egozcue. 'Open Science Drone Toolkit: Open Source Hardware and Software for Aerial Data Capture'. *PLOS ONE* 18, no. 4 (6 April 2023): e0284184. https://doi.org/10.1371/journal.pone.0284184.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'. *Scientific Data* 3, no. 1 (15 March 2016): 160018. https://doi.org/10.1038/sdata.2016.18.

Wyngaard, Jane, Lindsay Barbieri, Andrea Thomer, Josip Adams, Don Sullivan, Christopher Crosby, Cynthia Parr, Jens Klump, Sudhir Raj Shrestha, and Tom Bell. 'Emergent Challenges for Science SUAS Data Management: Fairness through Community Engagement and Best Practices Development'. *Remote Sensing* 11, no. 15 (2019): 1797.

Appendix – use case examples

Geophysics

Example of usage of fixed-wing platform for geophysics data collection

Answered given by Alvaro Arenas Pingarron, radioglaciologist at BAS

- **Please could you provide information about the type of research you undertake using UAV?**

The research field is glaciology in Antarctica, using the UAV platform to transport an active radar (with integrated transmitter and receiver) for along-flight englacial observations of shallow and deep ice. Among the applications are the estimations of: the ice thickness by detecting the subglacial base (bedrock or sea water) down to 5 km, the profiling of the internal layers, the shape of subglacial landforms, and the crystal orientation of englacial bulk ice.

- **What are the specificities of the data you collect?**

The main particularity of the data collection is that the path travelled by the radar signals is also part of the target of interest, and hence not fully known. In other radar surveys, the targets of interest are found *after* the path, but in our application they are *along* the path. Often, to interpret the deeper ice layers and the subglacial base, we need a proper interpretation of the surface and shallow ice.

The radar data (and derived products) contain a great number of ambiguities that hinder the data processing and the glaciological interpretation. There are two main sources: 1) the angular radiation pattern of the radar antenna system, and 2) morphology of the surveyed area. To recognise the ambiguities these two aspects, if possible, should be informed to the data processor and interpreters in form of metadata.

1) The radiation pattern of the antenna system needs to be measured, to identify possible ambiguities in the radar image interpretation. In our system the antennas point vertically downwards, and then the glaciers are illuminated approximately within a single vertical cone. The shape of this ideal cone depends on the type of antennas and their arrangement. However, unexpected irregularities in the antennas or even in the airborne platform can modify the ideal illumination pattern.
2) Along the flight, the glacier morphology can affect the radar interpretation. For example, if a surface is highly crevassed, then the more likely the radar signals reflected back from the englacial ice (vertically to the radar) can be masked by signals reflected off-vertically from the surface. This problem can be more acute if the angular radiation pattern is not measured.

The data collected involves a number of transmitters and receivers, independently paired, so each dataset is a group of 60 pairs of transmitter-receiver. These 60 pairs could be integrated into a one large file by adding two dimensions for the number of transmitters and receivers, but maybe this would increase the volume of unwanted data if for example only interested in one particular transmitter-receiver pair. Although there is specialised software to only read chunks of data, the files should be prepared by compatible software, which in my case is a difficulty. The metadata is not a difficulty, as only variations of the transmitter and receiver descriptions should be added.

The main problem I found is regarding the processed data from the raw data, because the processed output could involve several input files and intermediate steps. I usually feel that besides providing the processed output data, I should provide the intermediate inputs and raw data, which is not feasible. The solution would be to include the software and processing steps, to go from the input to the final output, but again this is not easy in my case.

- **What is the standard you use (or planning to use) to publish your data? Is there some example of data catalogues/repositories that provide good examples of data publication?**

I use netCDF, and I plan also to use extended csv format. As examples for data publication, I have not used a particular file or repository, but the guidelines from several examples to define the netCDF variables and attributes in https://cfconventions.org/Data/cf-conventions/cf-conventions-1.10/cf-conventions.html, and a list of global and variable attributes in https://wiki.esipfed.org/Attribute_Convention_for_Data_Discovery_1-3. The most important directions were from conversations with the UK Polar Data Centre, to better recognise the scope of the attribute and variables.

- **If there are some recommendations you would like to give to other scientists publishing data of the same type, what would you say?**

I would recommend putting themselves in the position of someone who, being not familiar with the data, tries to identify if their data set is useful. This way the metadata will definitely improve regarding the case of only considering potential users experts on the field, because there might be conventions that in other fields are opposite.

Although admitting that the metadata standards evolve with the time, it is also very important to think about data management before collecting or processing the data, to save time later, or before realising that some metadata are not possible to recover.

For netCDF files, I would recommend visualising the content with the software Panoply (https://www.giss.nasa.gov/tools/panoply/), a free portable stand-alone application.

Imagery

Answered given by Charles George at CEH

- **Please could you provide information about the type of research you undertake using autonomous or remotely piloted aerial systems?**

- Hyperspectral imagery
  - vegetation diversity of grasslands
  - quality of grassland forage
- RGB SfM
  - Forage volume
  - Hydrology of Colombian Paramo's
- LiDAR – landscape biodiversity metrics
  - River flow estimates
  - River habitat responses to drought

- **What are the specificities of the data you collect?**

The data tend to be large datasets, but of course this depends upon what resolution you need or how high you fly and what overlap between images you need. As an example for a 1.3 km area for RGB flying at 80m height we end up with 4,400 images totalling 81.5GB. If it were hyperspectral it would be a much larger dataset. The same area for Lidar flown at 60m is 114GB. The RGB are JPEG images. The lidar in this case are proprietary DJI files which as far as I am aware can only be efficiently pre-processed into more portable formats using DJI software.

- **What are the standard you use (or planning to use) to publish your data? Is there some example of data catalogues/repositories that provide good examples of data publication?**

We haven't got as far as publishing yet but this is a good question and something we need to think about as a community. For publication of the actual data, we really need a specialised data centre that is quick and easy to use. We already put some of our data in the EIDC, but it's not really made for.

- **If there is some recommendations you would like to give to other scientists publishing data of the same type, what would you say?**

Not applicable.

Atmospheric science
Answered given by Barbara Brooks at NCAS

- **Please could you provide information about the type of research you undertake using autonomous or remotely piloted aerial systems?**

- Hot spot GHG monitoring
  - development of analysis techniques providing regulatory compliant estimates of emission fluxes.
  - Lifting of a teflon sampling tube connected to a ground based sensor
  - Aeries technologies MIRA Stratos LDS: Laser absorption spectrometer for GHG detection. 2 kg with battery.
  - Flylogix work - community accepted best practices for determining GHG emissions flux. Calibration against ICOS network instruments.
- Boundary layer profiling
  - WESTCON and use of external contractor
    - Wind Speed: Range of approaches taken including, 5 and 7 hole probes, hot wire\film sensors, sonic anemometer, GPS & DGPS, triangulation from platform rotors.
    - PTU sensors: emphasis on stability and housing to eliminate dynamic and radiative effects.

- Radar calibration
    - carrying target and holding it steady in beam
- Insect sample collection
    - drag nets, sticky on leading edges - radar - aerobiology
- Volcano response
    - Sensors but no platform - making package platform independent, calibration
    - Sniffer 4D: Wide range SO_2, high resolution SO_2. 0.3 kg
    - Sniffer 4D: Wide range H_2S, wide range C_xH_y/CH_4/LEL. 0.3 kg
    - Alphasense: OPC series 30 - 100 g depending on model

- **What are the specificities of the data you collect?**

We produce lots of little files: one per flight per data product.
We use netcdf 4 - classic and most of us work with python 3 or higher - we are stating to think about moving to netdf4 or HDF5
We don't have instrument PIDs at the moment but are looking to move that way. At the moment, we have a defined vocabulary for what will be accepted by the compliance checker as entries in metadata fields and how file names are structured.
Each data product is defined in terms of its specific set of variables. We work on the principle that when a user opens the file they may find fewer variable but they will never find more. Each product is supported by documentation and we also supply an example files of both real data and artificial - the latter is so we can show what a fully populated file looks like.
The biggest challenge was getting our providers onboard - all they could see was a burden. We've been supporting them and now they are beginning to see the benefits through more efficient work streams and code development support. Users are also now seeing benefits as data is being delivered in a traceable consistent fashion, DOIs are more easily acquired and analysis code developed.
The next big challenge is to introduce traceable uncertainty into the data files that is CF compliant and develop tools that bring multiple streams together for comparison.

- **What are the standards you use (or planning to use) to publish your data? Is there some example of data catalogues/repositories that provide good examples of data publication?**

Data will follow the protocols and standards we have developed through the NCAS Data Activity - this work has defined the data products, metadata, and file formats we use for all our data. We've worked very closely with Wendy Garland and Graham Parton over at CEDA to set this all up. Graham has recently taken over from me as the lead for this activity.
We are moving out of the development phase for these data standards and are entering the phase where we are getting them embedding in the way NCAS scientists provide data. The next phase of development is around uncertainty and how we put them in our files.
We are making use of GitHub as a repository for processing code and CEDA as the node we archive all our data to.

- **If there are some recommendations you would like to give to other scientists publishing data of the same type, what would you say?**

Say what you do, do what you say and ensure everything is documented and grounded in citable literature. Provide tools and support at all times and try to lead rather than use a big stick. As early as possible start a dialogue between the archive, IT support, users and providers so you define a common vocabulary - we might all recognise the words but how they are interpreted is unique to the individual.