UTLS-OZONE Data Scoping Study

J. A. Kettleborough, L. J. Gray, and S. R. Williams. British Atmospheric Data Centre, Rutherford Appleton Laboratory.

Summary

- UTLS-OZONE funds a diverse set of projects encompassing field observations, data analysis, model studies and chemical kinetics studies. The data requirements of the different types of project are clearly different.
- Field observation projects will measure a large range of chemical and meteorological parameters from a range of platforms such as the MRF C-130 aircraft, the Egrett aircraft, balloon based observations, and ground based LIDAR. They are the most data intensive projects requiring data dissemination and long term archive facilities. The adopted file format for data exchange and archive will be NASA-Ames format.
- Model and data analysis projects potentially use and produce large volumes of data. Continuous model development and improvement means it is not appropriate to keep this data in a long-term archive. There is a case, however, for storing some chemical fields to act as reference fields. The adopted file format for model data will be NetCDF.
- Chemical kinetics studies produce small amounts of data that is readily disseminated via published papers. This data is most usefully archived along with other chemical kinetic data rather than in isolation.
- There are significant demands for 3rd party data sets to aid the planning and analysis of field observation campaigns, to establish the climatology and variability of the UTLS, and to evaluate the performance of numerical models. These data sets include ECMWF analyses, UKMO Unified Model output, and value-added products such as trajectories.
- Many of the UTLS-OZONE projects involve collaboration with EU and US funded programs. This presents some challenges for dissemination and archiving the data: how much of the data should be stored at the BADC, and what are the most effective ways to maintain references between data centres?

Purpose

The British Atmospheric Data Centre (BADC) has been designated by NERC as the data centre for archiving data collected or produced as part of the UTLS-OZONE thematic program. All projects funded by NERC under the UTLS-OZONE thematic program have a commitment to archive the data they produce as part of the program at the BADC, if appropriate. The BADC will then be responsible for disseminating the data and provide a long-term archive.

The general objective of the scoping study was to determine the data requirements of the program. Having determined these requirements the study can be used to identify and prioritise the tasks needed to ensure that these requirements are met efficiently and effectively.
The specific objectives of the scoping study were to:

1. Determine the type of data to be collected.
2. Determine the 3rd party data sets needed by the projects.
3. Determine a timetable for data collection.
4. Determine the most useful file formats for the data.
5. Determine the structure of the Data Archive.
6. Determine the most appropriate keywords for the data set catalogue.
7. Highlight any areas of data management that are uncertain at present.

Sources

Most of the information used in compiling this study was acquired using a data questionnaire sent to all round 1 and round 2 funded projects. A copy of the data questionnaire used in the data scoping study can be found in appendix A. Copies of the individual returns to the questionnaire can be obtained from the BADC on request. The returns to the questionnaire have been supplemented by e-mail contact with researchers in the individual projects.

Project Types

One of the challenges of data management for the UTLS project is the diversity of the data sets produced as part of the project. For the purposes of data management the projects can be characterised as one of three types.

1. Field observation (FO): using instruments employed to observe geophysical parameters.
2. Model studies (MS) and data analysis (DA): using mainly computational and numerical techniques.
3. Chemical kinetics (CK): using laboratory based instruments.

Clearly some projects encompass more than one type. For instance field observation projects always include an element of data analysis. The division of projects into the different types is not intended to completely define the aims of any one project: it is intended to act as a simple catagorisation to infer where the emphasis for data management tasks will be.
The data management demands for the three types of project are clearly different. Field observations, satellite data apart, tend to produce small data sets confined in time and space. The field observations need both to be disseminated rapidly and to be archived for the long term. Model studies and data analysis often use and produce large data sets. For these projects the 3rd party data requirements, such as meteorological analyses, are often of a higher priority than long term archiving. Long

term archiving of the data sets produced by MS and DA projects is of low priority since they have a limited life span as the models evolve and improve. The results of chemical kinetics experiments are very small data sets that can be disseminated either simply by the originating investigator or in published papers. The archiving of chemical kinetics results is obviously important, but is usually achieved by results appearing in published papers.

FOs are the most demanding of the types of project for data management. The BADC has gained experience of dealing with field observations through the ACSOE project. Although it has to be noted that not all of the field observation projects funded under UTLS will fit the ACSOE model. Many of the FO projects are funded under UTLS, but form part of larger field campaigns funded by other bodies such as the EU and NASA. These projects may have commitments to other funding bodies: these might include a commitment to archive data somewhere other than the BADC. The issues raised by the collaborations will be covered later in this document.

Projects

Table 1 lists the projects, with a classification of the project type, and any collaborations and/or prior data commitments. Of the thirteen FO projects nine involve collaborations with external projects, or have prior data commitments. Of the eight MS/DA projects three have the specific task of providing modeling support for non-UTLS projects.

Table 1 UTLS round 1 and 2 projects

Project	P.I.	Type	Collaboration
Airborne Measurements of Atmospheric Tracers in the UTLS for studies of atmospheric chemistry and transport	Jones, Pyle, and Gardiner	FO	
Atmospheric Chemistry and transport of Ozone in the UTLS (ACTO)	Penkett *et al.*	FO MS DA	
Campaign participation and modeling studies for APE-GAIA	Chipperfield, Roscoe	FO MS	APE-GAIA
Characterisation of volatile organic compounds in the UTLS	Pilling, Lewis, Bartle	FO	MAXOX
Development of scientific instrumentation for commercial aircraft	Jones	FO	
Extension of THESEO balloon-borne measurements of atmospheric tracers and chemically active gases in the mid-latitude lower stratosphere for test of atmospheric transport	Jones, Pyle, Woods	FO	THESEO
GCM measurements of halogenated source gases in the UTLS region in air samples	Penkett, Oram, Sturges	FO	CARIBIC

from CARIBIC flight program			
Improved upper air forecasting and analysis for APE-THESEO	MacKenzie	FO MS	APE THESEO
Improvements to calculations of lower stratosphere exchange between southern mid-latitudes and Antarctica by radiosonde launches during the Airborne Polar Experiment	Roscoe, Shanklin	FO	APE Antarctic Archive
Ozone LIDAR investigations of the subtropical jetstream and of subtropical intrusions into mid-latitudes	Vaughan	FO	METRO/ TRACAS
Support and Analysis for Far-Infrared emission remote sensing	Hamilton, Ade	FO	SAFIRE/IBEX
The Aberystwyth-Egrett Experiment	Whiteway, Vaughan	FO	
The role of frontal zones in determining upper troposphere chemical distributions	Browning *et al.*	FO MS	MAXOX
A General Circulation Model study of Ozone/Temperature interactions	Shine, Fish	MS	
Development of a microphysical and chemical model of cirrus clouds	Choularton	MS	
Evaluation of the Ozone and Water vapour data sets of the 40 year European Re-Analysis of the Global Atmosphere	Lahoz, O'Neill, Hoskins	DA	ECMWF
Forecast and Analysis of polar stratospheric clouds and cirrus for the NASA SOLVE arctic ozone campaign	Carslaw	MS	SOLVE
Gas Phase and aerosol composition of air entering the upper troposphere through convection	Parker, Carslaw	MS	
Studies of the tropopause region using version 5 data from MLS	Harwood	DA	UARS/MLS
The response of lower stratosphere ozone to solar variability and its impact on radiative forcing and climate	Haigh, Austin	MS	
Three Dimensional model studies for THESEO	Chipperfield	MS	THESEO
Laboratory studies of OH production and removal rates	Heard, Pilling	CK	

for the upper troposphere			
Laboratory studies of the heterogeneous interaction of pollutants from aircraft – of HNO_3, H_2O and soot aerosols	Cox	CK	
Laboratory, theoretical, and modeling studies of Gas phase peroxy radical reactions affecting the UTLS HO_x budget	Rowley, Cox, Clary	CK	

Campaign Times

Table 2 shows the main times when data will be collected. For projects funded in rounds one and two, most of the data will be collected during the first 3 years of funding. This implies that the initial emphasis for data management should be development and population of the archive. Development of value added products will be delayed until later in the program.

Table 2 Times for Experimental Campaigns

Project	**Campaign**	**1998**						**1999**						**2000**					
Hamilton et al	SAFIRE/IBEX																		
Penkett et al	CARIBIC																		
Vaughan et al	TRACAS/METRO																		
McKenzie	APE-THESEO																		
Browning et al	MAXOX																		
Jones et al	THESEO																		
Pilling et al	MAXOX																		
Chipperfield	APE-GAIA																		
Carslaw	SOLVE																		
Jones et al	EGRETT																		
Whiteway et al	EGRETT																		
Penkett et al	ACTO																		

Data Sets

The data sets to be collected by UTLS-OZONE projects are listed in Table 3. The list is, at present, incomplete. More details of the various data sets will be added, as the information becomes available.

Table 3 Data Sets produced as part of Field Campaigns

Quantity	**Instrument**	**Platform**	**Time**	**Project**	**Data set Size**
CH_4	GC	EGRETT	9912-0002	Jones et al.	
3 CFCs	GC	EGRETT	9912-0002	Jones et al.	2.5MBytes
reflectance	MST Radar	Ground based	0101-0201	Whiteway et al	
O_3	Chemical Cell	Balloon	0101-0201	Whiteway et al	
O_3	LIDAR	Ground	0101-0201	Whiteway et	

		based		al	
u, v, w, T		EGRETT	0101-0201	Whiteway et al	
u,v,w,T		C-130	00Spring-00Aut.	Penkett ACTO	
O_3		C-130	00Spring-00Aut.	Penkett ACTO	
H_20		C-130	00Spring-00Aut.	Penkett ACTO	
PAN		C-130	00Spring-00Aut.	Penkett ACTO	
>30 NMHCs		C-130	00Spring-00Aut.	Penkett ACTO	
>40 Halocarbons		C-130	00Spring-00Aut.	Penkett ACTO	
DMS		C-130	00Spring-00Aut.	Penkett ACTO	
Acetone		C-130	00Spring-00Aut.	Penkett ACTO	
CO		C-130	00Spring-00Aut.	Penkett ACTO	
NO NO_2		C-130	00Spring-00Aut.	Penkett ACTO	
NO_y		C-130	00Spring-00Aut.	Penkett ACTO	
HNO_3		C-130	00Spring-00Aut.	Penkett ACTO	
$J(NO_2)$		C-130	00Spring-00Aut.	Penkett ACTO	
$J(O_3\text{->}O^1D)$		C-130	00Spring-00Aut.	Penkett ACTO	
Peroxides		C-130	00Spring-00Aut.	Penkett ACTO	
Peroxy radicals		C-130	00Spring-00Aut.	Penkett ACTO	
HCHO		C-130	00Spring-00Aut.	Penkett ACTO	
CH_4		C-130	00Spring-00Aut.	Penkett ACTO	
N_2O		C-130	00Spring-00Aut.	Penkett ACTO	
Aerosols		C-130	00Spring-00Aut.	Penkett ACTO	400Mbytes
Benzene, Toluene	GC	C-130	9903-9904, 9908	Pilling et al	
H_2O	SAW	C-130		Jones	
CH_4, H_2O, CFCs	GC	Balloon	9901-9912	Jones et al	1.2Mbytes
Halocarbons	GC	Airborne	9801-2010	Penkett	3Mbyte

				CARIBIC	
u,T	Radio Sonde		9909-9910	Roscoe	
O_3	Chemical Cell	Ozone Sonde	9904-0104	Vaughan et al	2.5Mbyte
O_3	LIDAR	Groud Based	9904-0104	Vaughan et al	1Mbyte
Reflectance	MST Radar	Ground Based	9904-0104	Vaughan et al	
OH,NO_2,HOCl ,HO_2, O_3, H_2O, HBr, HOBr, $BrONO_2$, HNO_3	FIR	Various	9807, 981115-981215, 990915-991015	Hamilton	
Reflecatance etc	Chilbolton Radar	Ground based	9901-9905	Browning et al	600Mbytes
NOx, NOy		C-130	9901-9905	Browning et al	
HCHO		C-130	9901-9905	Browning et al	
H_2O_2, RO_2		C-130	9901-9905	Browning et al	
CO		C-130	9901-9905	Browning et al	
Aerosols		C-130	9901-9905	Browning et al	100Mbyte

Collaborating Projects

As noted earlier, many of the UTLS-OZONE projects are part of collaborations with NASA and EU funded projects. The collaborations raise several questions for the dissemination and archiving of the data collected as part of these projects.

If the collaborating project has its own designated data centre, then, presumably, that data centre will be responsible for the dissemination of data during field campaigns. The main question is then how much of the data archive will subsequently be mirrored at the BADC. Although there is a commitment to archive all UTLS-OZONE funded data at the BADC it may be that it is simply not appropriate, and possibly meaningless, to archive the UTLS-OZONE project data in isolation from the collaborating project data. In these cases it may be that the BADC could act as a mirror to the primary archive, although this obviously would entail negotiation with the data centre of the collaborating project. Alternatively, and probably more practically, the BADC could maintain references to the collaborating project data centre. This would pass responsibility for the maintenance of some of the UTLS-OZONE data archive to another data centre, which may have its own problems.

If the collaborating project does not have its own designated data centre, then the BADC could act as the data centre for the project, if required. The dissemination of UTLS data to other researchers in the collaborating projects can be facilitated by the BADC. Collaborating researchers would be required to sign a data agreement, in a similar way to ACSOE. It would have to be decided, probably on a case by case basis, how much of the UTLS-OZONE archive the external researchers would have access too. Presumably collaborating project researchers would only need access to the data sets produced by the UTLS-OZONE project with which they are collaborating. If the BADC is to facilitate the dissemination of data to collaborating projects, it will be necessary to set up the relevant mechanism in the archive at an early stage.

3rd Party Data

Table 4 lists the 3rd party data requirements of the various UTLS-OZONE projects. The column labeled status indicates whether the project already has access to the required data set. If the data set is already available the column labeled source indicates the source of the data. It should be noted that some data sets have separate agreements and are not necessarily available to the all UTLS projects. These include the UKMO/ECMWF forecasts used during APE-THESEO (McKenzie), and, at least initially, the new 40-year ECMWF-ERA data (Lahoz et al).

Forecast data is required for mission planning during field campaigns. ECMWF forecast data can be obtained through the BADC, but requires an independent letter of application to go to the ECMWF.

Some of the 3rd Party data requirements are already being met by the BADC. These include the ECMWF analyses and UKMO UARS assimilated data. Plans are already underway to meet the UTLS requirement for access to the UKMO Unified Model data. No doubt other 3rd party data requirements will become evident throughout the period of the UTLS-OZONE program.

There is some demand for value added products such as trajectories, meteorological variables on isentropic surfaces, and meteorological variables along flight tracks. The BADC has developed a WWW interface to a trajectory model to help fulfil the demand for trajectory data by the UTLS-OZONE projects. The BADC will also investigate the possibility of adding tools to the BADC archive to calculate other value-added products.

Table 4 Third Party data Requirements of the UTLS-OZONE projects

Project	Data Required	Status	Source
Vaughan et al	ECMWF Met. Data	Semi – no isentropic data or PV	BADC
Vaughan et al	Meteosat Water Vapour	Not currently available	
Pilling et al	Trajectories	Yes	HYSPLIT Model
Oram et al	Trajectories	Yes	CARIBIC
McKenzie et al	ECMWF/UKMO data	Yes	UKMO/ ECMWF
Browning et al	UKMO Unified model	Yes	UKMO-JCMM
Browning et al	Network Rain Rate	Yes	
Browning et al	ECMWF data	Yes	BADC
Parker et al	Met Data for initialisations		
Jones	T, H_2O from UKMO/ECMWF	Semi – no along flight track data	BADC
Harwood et al	ECMWF/UKMO	Yes	BADC UKMO
Chipperfield	UKMO UARS assimilations	Yes	BADC
Chipperfield	ECMWF Reanalysis	No	
Shine et al	O_3 Trends	Yes	NASA

Lahoz et al	ERA O_3 and H_2O	Yes	ECMWF
Haigh et al	SBUV O_3	Yes	
Haigh et al	SOLSTICE/SUSIM	Yes	
Whiteway	ECMWF Forecasts		BADC/ECMWF
Whiteway	Forecasts		
Penkett et al ACTO	Trajectories	Yes	U. Reading/ BADC

Data File Formats

The adoption of common, standard file formats both eases the dissemination of results and helps ensure the long-term integrity of a data set. Standard formats reduce any ambiguity in a data set and simplify the maintenance of reading and writing software and associated documentation. A good standard data format should be:

1. Portable across computer platforms.
2. Self-describing: there is enough information in the file to determine the contents of that file.

As well as being portable across platforms a good standard format will have potential for use in many analysis packages. In deciding the standard file format to be adopted it is important to consider users previous experience and the resources available for analysis.

Figure 1 indicates the experience of researchers funded by UTLS with various established data formats. NASA-Ames Format for Data Exchange is the format that has had most previous usage. This is an ASCII based format ideal for exchange of FO data. Each NASA Ames file consists of a header, which describes the contents of the file, followed by the data.

A binary format is often more useful for storage of large gridded data sets. Although researchers have about equal experience of GRIB and NetCDF format some thought that NetCDF was easier to use. NetCDF is a portable binary format, with access routines in FORTRAN, C, C++, perl, java, and IDL.

Figure 2 shows the analysis/display methods used by the researchers. IDL is the most widely used, although clearly a significant number use Excel and MATLAB.

Given the previous experience of data formats and analysis/plotting packages used UTLS Ozone will adopt NASA-Ames as the standard data format. Large gridded data sets will use NetCDF. NASA-Ames files are readable by IDL, Excel, and MATLAB. NetCDF files are supported by IDL, and some freely available software packages for plotting NetCDF files, such as FERRET. LiveAcess server can also be used to give WWW access to NetCDF files.

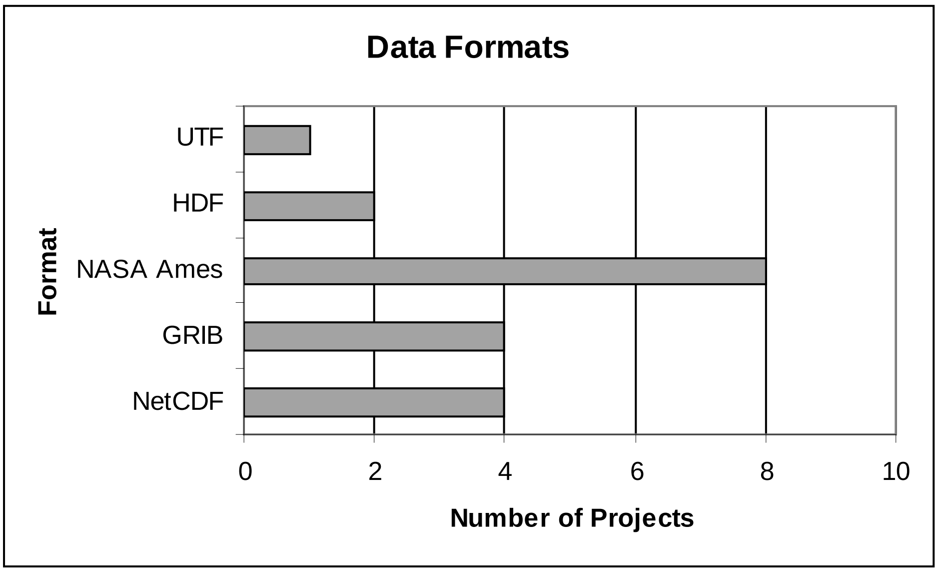

Figure 1 Data format experience of UTLS-OZONE projects

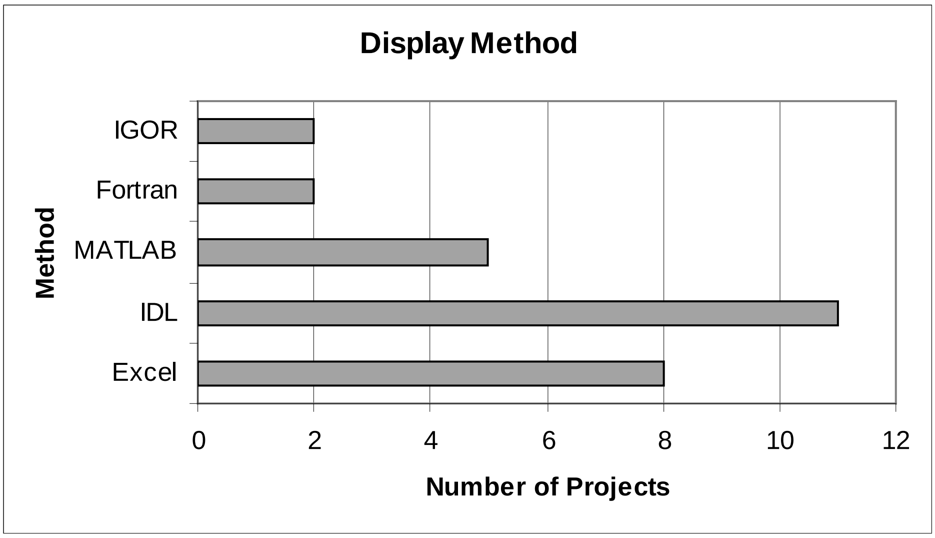

Figure 2 Display/Analysis Tools used by UTLS-OZONE projects

Appendix A UTLS-OZONE Data Questionnaire

UTLS-Ozone Data Questionnaire

Project Details

Title of Project:

PI's Name:

Your Name (if not the PI):

Your Address:

E-mail:

Data sets that you will produce

2.1 What data sets are you likely to produce as part of your UTLS-Ozone funded work? Please give some indication of the size (to an approximate order of magnitude) of each of these data sets.

2.2 Which of these data sets do you consider might be worth archiving in the long-term?

2.3 What will be the primary times of your data collection and/or model result production?

Start time	Finish time

2.4 Do you think that other projects within UTLS-Ozone would find it useful to have access to your data?

☐ Yes ☐ No

2.5 Will you **need** to share data with other institutes?

☐ Within your UTLS project ☐ With other UTLS projects
☐ Outside UTLS (*e.g.* with EU partners)

If so, please give details:

Do you already have plans for doing this, or would you like assistance from the BADC?

☐ Distribute data yourselves ☐ Use the BADC

2.6 When do you anticipate submitting data to the BADC archive?

☐ As soon as possible to allow you to use the BADC to share and validate data with other UTLS projects.
☐ When you are happy that the data is in a suitably validated form for other people to use.
☐ At the end of the project.
☐ Not Applicable

Third Party Data Requirements

3.1 If you anticipate using any third party data (*e.g.* Met. data or trajectories), please tell us what they are.

3.2 Do you already have access to these data sets? If not how do you anticipate getting them?

3.3 Are there any issues of intellectual property rights that will restrict the availability of these data to other UTLS projects or that will prevent them from being added to the "public" UTLS data sets at the end of the UTLS project?

4 Computing Issues

4.1 Which of these standard data formats do you have experience of working with?

- ☐ netCDF
- ☐ NASA Ames
- ☐ GRIB
- ☐ HDF

Please list any other formats that you might like to use:

Would you be prepared to write your data to standard data format(s), if assistance was available from the BADC to help with this job?

4.2 What computer hardware are you likely to use to work on your data? Please tick any appropriate boxes:

- ☐ Supercomputer
- ☐ Sun Unix workstation
- ☐ HP Unix workstation
- ☐ DEC Unix workstation
- ☐ Linux workstation
- ☐ Windows 95/98/NT PC
- ☐ Macintosh PC

Other - please specify:

4.3 If you intend to use commercial software to manipulate your data, for example, IDL, PV-Wave or Excel, please list them in the space below:

Anything Else?

Are there any other UTLS-Ozone data issues with which the BADC could help you? Do you have any other comments about the UTLS-Ozone data management? Are you happy with the UTLS-Ozone data protocol as it stands at present?