



DataSeer

DataSeer

Getting beyond the guidelines

Dr. Tim Vines

tim@dataseer.ai



ALFRED P. SLOAN
FOUNDATION



ORIGIN



Stratos



UC3
UC Curation Center

Four problems:

1. Most articles are published without their underlying data
2. Shared datasets are incomplete & in useless formats
3. Metadata about shared data are inadequate
4. Citing datasets is very patchy

Our response:

1. Most articles are published without their underlying data

[Journal] requires, as a condition for publication, that data supporting the results in the paper should be archived in an appropriate public archive, such as [list of approved archives here]. Data are important products of the scientific enterprise, and they should be preserved and usable for decades in the future. Authors may elect to have the data publicly available at time of publication, or, if the technology of the archive allows, may opt to embargo access to the data for a period up to a year after publication. Exceptions may be granted at the discretion of the editor, especially for sensitive information such as human subject data or the location of endangered species.

Our response:

2. Shared datasets are incomplete & in useless formats:

For example, authors should submit the following data:

- › The values behind the means, standard deviations and other measures reported;
- › The values used to build graphs;
- › The points extracted from images for analysis.

Authors do not need to submit their entire data set if only a portion of the data was used in the reported study. Also, authors do not need to submit the raw data collected during an investigation if the standard in the field is to share data that have been processed.

PLOS does not permit references to “data not shown.” Authors should deposit relevant data in a public data repository or provide the data in the manuscript.

<https://journals.plos.org/plosone/s/data-availability>

Our response:

2. Shared datasets are incomplete & in useless formats:

TYPE OF DATA	PREFERRED FILE FORMATS FOR SHARING, RE-USE AND PRESERVATION	Other Acceptable formats
<p>Quantitative tabular data with extensive metadata</p> <ul style="list-style-type: none">• a dataset with variable labels, code labels, and defined missing values, in addition to the matrix of data	<ul style="list-style-type: none">• SPSS portable format (.por)• delimited text and command MS Access (.mdb/.accdb) ('setup') file• (SPSS, Stata, SAS, etc.) containing metadata information• structured text or mark-up file containing metadata information, e.g. DDI XML file	

Our response:

3. Metadata about shared data are inadequate

General overview	Title	Name of the dataset
	Creator	Names and addresses of the organisations or people who created the data
	Identifier	A unique number used to identify the data
	Date	Key dates associated with the data, including project start and end date, time period and other important dates associated with the data. Preferred format is yyyy-mm-dd
Content description	Subjects	Subjects include keywords or phrases describing the subject or content of the data. It can also include Field of Research codes and Socio-economic objective codes as defined by ANZSRC
Technical description	File inventory	All files forming part of the dataset, including extensions (e.g. 'photo1023.jpeg', 'participant12.pdf')
Access	Rights	Any known intellectual property rights, statutory rights, licenses, or restrictions on use of the data

Our response:

4. Citing datasets is very patchy

JOINT DECLARATION OF DATA CITATION PRINCIPLES - FINAL



When citing please use: Data Citation Synthesis Group: Joint Declaration of Data Citation Principles. Martone M. (ed.) San Diego CA: FORCE11; 2014
<https://doi.org/10.25490/a97f-egyk>

Translations

Japanese - https://doi.org/10.11502/rduf_rdc_jddcp_ja (added 31.01.2020).

>>> Endorsement List

PREAMBLE

Sound, reproducible scholarship rests upon a foundation of robust, accessible data. For this to be so in practice as well as theory, data must be accorded due importance in the practice of scholarship and in the enduring scholarly record. In other words, data should be considered legitimate, citable products of research. Data

The logo for Data Citation Principles (DC1) features the letters 'DC' in a large, blue, serif font, with a superscripted '1' to the right.

<https://www.force11.org/datacitationprinciples>

The unifying theme:

- Policies and guidelines are all generally worded
- They appear in diverse locations
- Applying them all to your research is **hard work**

The solution:

- **Get Specific!**
- Tell authors what they need to do *right now* for *this document*
 - No need to locate & interpret multiple guidelines and policies
 - Less inconsistency between articles & between authors

The solution:

- **Get Specific!**
- Tell authors what they need to do *right now* for *this document*
 - No need to locate & interpret multiple guidelines and policies
 - Less inconsistency between articles & between authors
- DataSeer uses this approach to tackle all four problems



DataSeer

- Uses NLP to 'read' articles
- Finds sentences describing data collection
- Works out the type of data

Remobilization of crustal carbon may dominate volcanic arc emissions

Emily Mason, Marie Edmonds,* Alexandra V. Turchyn

RESEARCH | REPORT

Compiled carbon and helium data



and other carbon reservoir estimates over geologic time.

We compiled a global data set for carbon and helium isotopic composition of volcanic gases in arcs and evaluated whether assimilation of carbon from overlying crustal carbonates could dominate the global arc volcanic carbon flux. We identified arcs for which this mechanism may dominate the carbon flux. We found a large effect of crustal carbonates on the carbon budget, which requires reinterpretation of the global carbon isotope mass balance throughout Earth history. This has direct implications for the fractional burial of organic carbon (19, 20) through geological time.

The speed of memory errors shows the influence of misleading information: Testing the diffusion model and discrete-state models

Jeffrey J. Starns^{a,*}, Chad Dubé^b, Matthew E. Frelinger^a

3. Experiment 1

In this experiment, participants studied a list of words and then completed a single-item test followed by a forced-choice test. On the single-item test, participants saw a word for each trial and decided whether or not it was on the study list. On the forced-choice test, participants saw two words for each trial and were asked to select which one was on the earlier study list. The forced-choice words were always ones that had previously appeared on the single-item test and received the same response (either both called “Studied” or both called “Not Studied”), but only one of the words was actually studied. Thus, the forced-choice trials always had one word with a previous correct response and one with a previous error response. The result of primary interest was whether forced-choice accuracy was related to the single-trial RT for the word with a previous error response.

3.1. Participants

We collected data from 122 University of Massachusetts Amherst undergraduates in exchange for extra credit in their psychology courses. We excluded participants with near-chance performance levels on the single-item trials, defined as a difference between the hit rate and false-alarm rate less than 0.1. This resulted in the removal of 12 participants, leaving 110 for the analyses reported below.

Participant data



5.1. Participants

Participants were sampled from the same population as Experiment 1, but no participants contributed data to more than one experiment. For Experiment 2A, we collected data from 66 participants, and one participant was excluded from analyses for near-chance performance on the single-item trials. For Experiment 2B, we collected data from 48 participants and excluded two for near-chance performance on the single-item trials.

Natural selection in a postglacial range expansion: the case of the colour cline in the European barn owl

SYLVAIN ANTONIAZZA,*¶¹ RICARDO KANITZ,*†¹ SAMUEL NEUENSCHWANDER,*†‡
RETO BURRI,§ ARNAUD GAIGHER,* ALEXANDRE ROULIN* and JÉRÔME GOUDET*†

Materials and methods

Sampling and molecular analyses

From 20 locations throughout Europe, a total of 390 barn owls were sampled by collaborators working in survey programmes, recovery centres and museums (Fig. 1). Genomic DNA was extracted from the basal 1 mm of breast feather quills or from blood or muscles stored in 96% ethanol. Extractions were performed either on a BioSprint 96 extraction robot using the BioSprint 96 DNA blood kit or using the DNeasy blood and tissue kit, following the manufacturer's protocols (Qiagen, Hilden, Germany).

der *et al.* 2008a,b). For the observed data, colour measurements (phenotypes) were obtained for each individual by sampling four reflectance spectra with an Ocean optic USB 4000 spectrophotometer (Ocean Optics, Dunedin, FL, USA) on five breast feathers per

Sampling information

Population genetic statistics were estimated from genotypes obtained for 22 polymorphic microsatellite loci [Ta-202, Ta-204, Ta-206, Ta-210, Ta-214, Ta-214, Ta-215, Ta-216, Ta-218, Ta-220, Ta-305, Ta-306, Ta-310, Ta-402, Ta-408 and Ta-413 from Burri *et al.* (2008) and 54f2, Calex-05, FEPO42, Oe053, GgaRBG18 and Tgu06 from Klein *et al.* (2009)]. Polymerase chain reactions (PCR) were performed in five multiplexes using the QIAGEN Multiplex PCR Kit (Qiagen) following the protocol: initial step of denaturation for 15 min at 95 °C, 34 cycles of 30 s denaturation at 94 °C, annealing for 1.5 min at 57 °C and elongation at 72 °C for 1 min. Final elongation for 30 min was conducted at 60 °C. The primer concentration and multiplexes composition is given in Table S1 (Supporting information). Fragment analyses were run on an ABI 3100 sequencer with a ROX 500 size standard, and allele lengths were assigned using GENEMAPPER 4.0 (Applied Biosystems, Foster City, CA, USA). After verifying that no null alleles were present (MICRO-CHECKER 2.2.3, Van Oosterhout *et al.* 2004) and that populations were not showing departure from Hardy–Weinberg equilibrium (Goudet 1995), the data

Reflectance spectra data

microsatellite genotype data



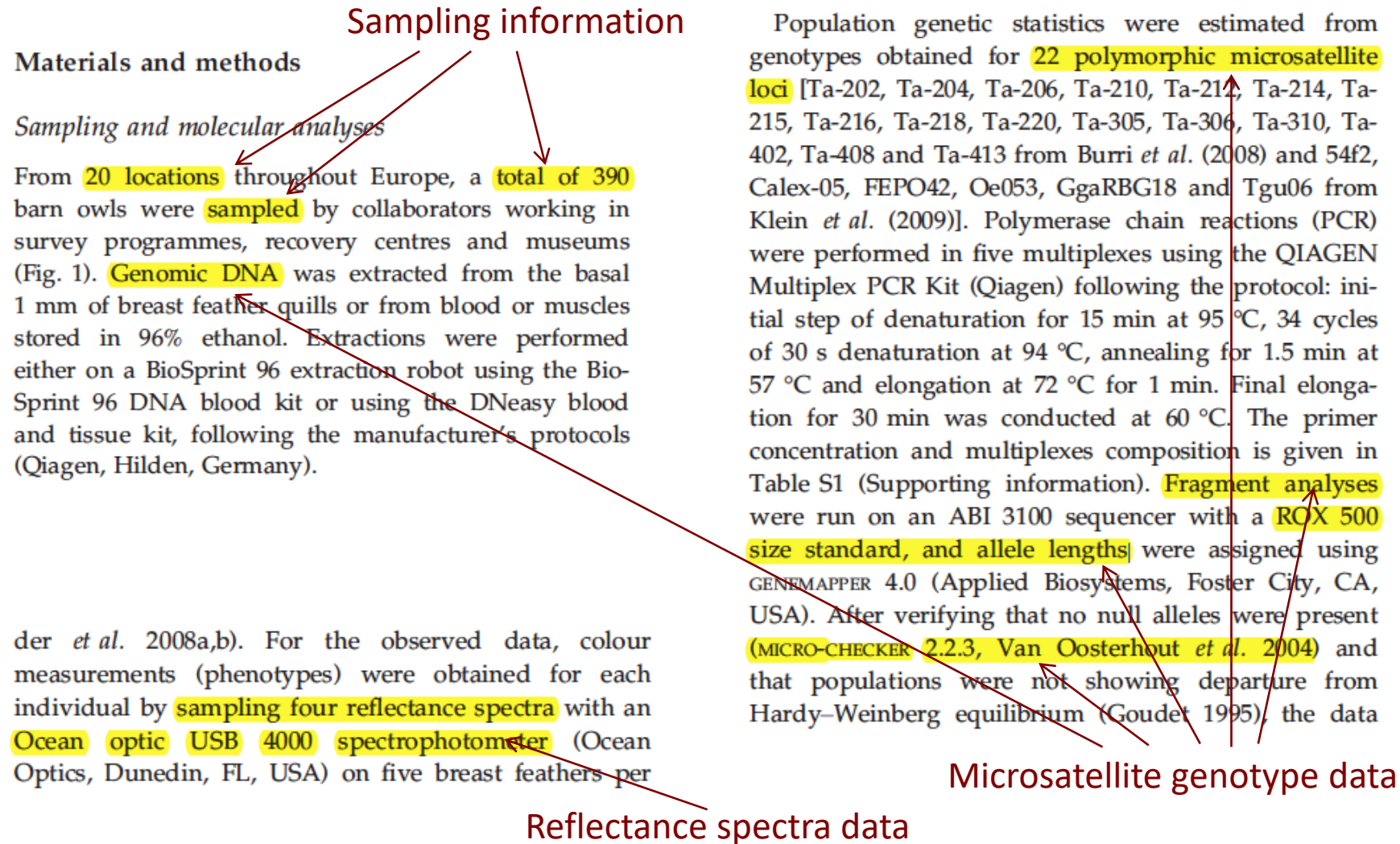
DataSeer

- Uses NLP to 'read' articles
- Finds sentences describing data collection
- Works out the type of data

- Then gives best practice sharing advice &
- Recommends a repository

Natural selection in a postglacial range expansion: the case of the colour cline in the European barn owl

SYLVAIN ANTONIAZZA,*¶¹ RICARDO KANITZ,*†¹ SAMUEL NEUENSCHWANDER,*†‡
RETO BURRI,§ ARNAUD GAIGHER,* ALEXANDRE ROULIN* and JÉRÔME GOUDET*†



Natural selection in a postglacial range expansion: the case of the colour cline in the European barn owl

SYLVAIN ANTONIAZZA,*¶¹ RICARDO KANITZ,*†¹ SAMUEL NEUENSCHWANDER,*†‡
RETO BURRI,§ ARNAUD GAIGHER,* ALEXANDRE ROULIN* and JÉRÔME GOUDET*†

Sampling information →

Sample Table (Tabular Data)

MeSH ID: n/a

Description:

Tabular data consist of a table with rows and columns. Each cell of the table contains numbers or text. For a sample table, the rows contain information about individual samples, and the columns contain the measured variables and other information. Sample tables must contain detailed information about exactly when and where each sample was obtained, and columns indicating any subgroup each sample belongs to (e.g. 'sand dunes', or 'control group').

Best practice for sharing this type of data:

Tabular data should be saved as a .txt or .csv file. The first row(s) should contain information about the dataset, such as the data file name, author, today's date, when the data within the file were last modified, and companion file names. Please also state which symbol has been used to denote missing data (NA is preferred). Column headings should describe the content of each column and contain only numbers, letters, and underscores - no spaces or special characters. Lowercase letters are preferred. Row names should be consistent with those used in the article and in other related datasets.

Most suitable repositories:

Many repositories are suitable for tabular data. Your institution, journal, or funder may recommend specific repositories, otherwise [Dryad](#), [Zenodo](#), or [FigShare](#) are good choices.

Natural selection in a postglacial range expansion: the case of the colour cline in the European barn owl

SYLVAIN ANTONIAZZA,*¶¹ RICARDO KANITZ,*†¹ SAMUEL NEUENSCHWANDER,*†‡
RETO BURRI,§ ARNAUD GAIGHER,* ALEXANDRE ROULIN* and JÉRÔME GOUDET*†

Reflectance spectra data 

Tabular data

MeSH ID: n/a




Description:

Tabular data consist of a table with rows and columns. Each cell of the table contains numbers or text.

Best practice for sharing this type of data:

Tabular data should be saved as a .txt or .csv file. The first row(s) should contain information about the dataset, such as the data file name, author, today's date, when the data within the file were last modified, and companion file names. Please also state which symbol has been used to denote missing data (NA is preferred). Column headings should describe the content of each column and contain only numbers, letters, and underscores - no spaces or special characters. Lowercase letters are preferred. Row names should be consistent with those used in the article and in other related datasets.

Most suitable repositories:

Many repositories are suitable for tabular data. Your institution, journal, or funder may recommend specific repositories, otherwise  [Dryad](#),  [Zenodo](#), or  [FigShare](#) are good choices.

Natural selection in a postglacial range expansion: the case of the colour cline in the European barn owl

SYLVAIN ANTONIAZZA,*¶¹ RICARDO KANITZ,*†¹ SAMUEL NEUENSCHWANDER,*†‡
RETO BURRI,§ ARNAUD GAIGHER,* ALEXANDRE ROULIN* and JÉRÔME GOUDET*†

Microsatellite data



Microsatellite Repeats

MeSH ID:  [D018895](#)




Description:

A variety of simple repeat sequences that are distributed throughout the genome. They are characterized by a short repeat unit of 2-8 basepairs that is repeated up to 100 times. They are also known as short tandem repeats (STRs).

Best practice for sharing this type of data:

When possible the original data output should be shared (e.g. the **Electrophoresis** or sequencer output file). The scored genotypes should be shared as **Tabular data**, ensuring that the names for individual samples are consistent with those used in the article and in other related datasets. Tabular data should be saved as a .txt or .csv file. The first row(s) should contain information about the dataset, such as the data file name, author, today's date, when the data within the file were last modified, and companion file names. Please also state which symbol has been used to denote missing data (NA is preferred). Column headings should describe the content of each column and contain only numbers, letters, and underscores - no spaces or special characters. Lowercase letters are preferred. Row names should be consistent with those used in the article and in other related datasets.

Most suitable repositories:

Tables containing microsatellite genotype data and the associated files can be added to any repository able to host generic file types (e.g.  [Dryad](#),  [Zenodo](#), and  [FigShare](#))

Remobilization of crustal carbon may dominate volcanic arc emissions

Emily Mason, Marie Edmonds,* Alexandra V. Turchyn

RESEARCH | REPORT

Compiled carbon and helium data



and other carbon reservoir estimates over geologic time.

We compiled a global data set for carbon and helium isotopic composition of volcanic gases in arcs and evaluated whether assimilation of carbon from overlying crustal carbonates could dominate the global arc volcanic carbon flux. We identified arcs for which this mechanism may dominate the carbon flux. We found a large effect of crustal carbonates on the carbon budget, which requires reinterpretation of the global carbon isotope mass balance throughout Earth history. This has direct implications for the fractional burial of organic carbon (19, 20) through geological time.

Remobilization of crustal carbon may dominate volcanic arc emissions

Emily Mason, Marie Edmonds,* Alexandra V. Turchyn

Compiled carbon and helium data →


Dataset Re-use

MeSH ID:  D064886

Description:

Works consisting of organized collections of data, which have been stored permanently in a formalized manner suitable for communication, interpretation, or processing.

Best practice for sharing this type of data:

Generally these data are already publicly available or available to appropriately vetted users. For access-controlled data authors should provide a link to instructions for obtaining access (e.g. here is the information page for ADNI (Alzheimer's Disease Neuroimaging Initiative):  <http://adni.loni.usc.edu/data-samples/access-data/>). For public datasets please provide a DOI or other stable identified for the dataset itself *and* include a citation for the dataset in the reference list. Be sure to indicate exactly which data has been re-used. In many cases, this is best achieved by sharing the code used to extract the part of the data that you analyzed.

Most suitable repositories:

Not applicable



- This is how it looks in practice



DataSeer

- Thanks to:
 - Kristen Ratan @ StratOS
 - Danielle Lowenberg @ UC Curation Centre
 - Jason Roberts @ Origin Editorial
 - Josh Greenberg @ the Sloan Foundation

