Integrating Data and Publications

how to make things better

Integration of Research Data and Publications

Eefke Smit
International Association of STM Publishers
Director, Standards and Technology

Amsterdam, Prepared Workshop at IDCC 2013,
17 January 2012
A famous paper in Nature: DNA structure - 1953

- 1 page
- 2 authors
- 1 figure
- no data
Nature in 2001: The human genome issue

- 62 pages, 49 figures, 27 tables

Source: V. Kiermer, Nature Publishing Group, 2011
The human genome at 10 – 2010
Nature now in an iPad edition:

Source: V. Kiermer, Nature Publishing Group, 2011
A thousand genomes – 2010

http://www.nature.com/nature/journal/v467/n7319/full/nature09534.html

Raw data: 12,145 SRA run ids submitted to Short Read Archive

Source: V. Kiermer, Nature Publishing Group, 2011
new publishing models
author information
doi
live updates
article-level metrics
Tool box to print,
download reference,
share: email, social media,
bookmark
Related content
Figure previewer
Collapsible sections

Source: V. Kiermer, Nature Publishing Group, 2011
Data Repositories and Journals

PANGAEA
Data Publisher for Earth & Environmental Science

Citation: Indermühle, Andreas; Monnin, Eric; Stauffer, Bernhard; Stocker, Thomas F.; Wahlen, Martin (2000): A high-resolution record of the atmospheric CO2 concentration from 60-20 kyr BP from the Taylor Dome ice core, Antarctica. doi:10.1594/PANGAEA.710905

Reference(s): Indermühle, Andreas; Monnin, Eric; Stauffer, Bernhard; Stocker, Thomas F.; Wahlen, Martin (2000): Atmospheric CO2 concentration from 60 to 20 kyr BP from the Taylor Dome ice cores, Antarctica. Geophysical Research Letters, 27(8), 735-738, doi:10.1029/1999GL010660

Abstract: A high-resolution record of the atmospheric CO2 concentration from 60 to 20 thousand years before present (kyr BP) based on measurements on the ice core of Taylor Dome, Antarctica is presented. This record shows four distinct peaks of 20 parts per million by volume (ppmv) on a millennial time scale. Good correlation of the CO2 record with temperature reconstructions based on stable isotope measurements on the Vostok ice core (Antarctica) is found.

Project(s): European Project for Ice Coring in Antarctica (EPICA)

Coverage: Latitude: -77.783300 * Longitude: 158.718670
Minimum DEPTH, ice/snow: 380 m
Event(s): Taylor_Dome (TD) * Latitude: -77.783300
Date/Time End: 1994-12-31 10:00
TAYLOR_DOME * Basis: Same
Comment: Depth is top of sample interval (m)

Parameter(s):
- Name
  - DEPTH, ice/snow
  - Gas age
  - Carbon dioxide
  - Carbon dioxide, standard deviation

GRL Geophysical Research Letters

Abstract

Atmospheric CO2 concentration from 60 to 20 kyr BP from the Taylor Dome Ice Core, Antarctica

Andreas Indermühle
Climate and Environmental Physics, Physics Institute, University of Bern, Bern, Switzerland
Eric Monnin
Climate and Environmental Physics, Physics Institute, University of Bern, Bern, Switzerland
Bernhard Stauffer
Climate and Environmental Physics, Physics Institute, University of Bern, Bern, Switzerland
Thomas F. Stocker
Climate and Environmental Physics, Physics Institute, University of Bern, Bern, Switzerland
Martin Wahlen
Sciences Institute of Oceanography, University of California San Diego, La Jolla, California, USA

A high-resolution record of the atmospheric CO2 concentration from 60 to 20 thousand years before present (kyr BP) based on measurements on the ice core of Taylor Dome, Antarctica is presented. This record shows four distinct peaks of 20 parts per million by volume (ppmv) on a millennial time scale. Good correlation of the CO2 record with temperature reconstructions based on stable isotope measurements on the Vostok ice core (Antarctica) is found.
From the Journal Article to the Data Repository

Earth and Planetary Science Letters
Volume 290, Issues 3-4, 20 February 2010, Pages 319–330

Obscuring of long eccentricity cyclicity in Pleistocene oceanic carbon isotope records
Pindian Wanga, Jun Tiana, Lucas J. Lourensbb
a State Key Laboratory of Marine Geology, Tongji University, Shanghai, China
b Faculty of Geosciences, Department of Earth Sciences, Utrecht University, Utrecht, Netherlands

http://dx.doi.org/10.1016/j.epsl.2009.12.028

PANGAEA® – Related Data
Stable isotope record of Mediterranean Sea sediments

Data Description

doi:10.1029/99PA01125


Minimum Age: 1212.000 ka BP * Maximum Age: 5330.000 ka BP


Parameter(s)

<table>
<thead>
<tr>
<th>#</th>
<th>Name</th>
<th>Short Name</th>
<th>Unit</th>
<th>Principal Investigator</th>
<th>Method</th>
<th>Comment</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Event label</td>
<td>Event</td>
<td></td>
<td></td>
<td>Metadata</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>AGE #</td>
<td>Age</td>
<td>ka BP</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Globigerinoides</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Globigerinoides</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

License: CC BY Creative Commons Attribution 3.0 Unported
Size: 8238 data points
Data from: Synchrotron X-ray tomographic microscopy of fossil embryos

When using this data, please cite the original article:


Additionally, please cite the Dryad data package:


Abstract

Fossilized embryos from the late Neoproterozoic and earliest Phanerozoic have caused much excitement because they preserve the earliest stages of embryology of animals that represent the initial diversification of metazoans. However, the potential of this material has not been fully realized because of reliance on traditional, non-destructive methods that allow an inspection of embryonic structures, but do not permit the extraction of detailed information about the internal structures of the embryos. Here, we present a methodology that allows an inspection of the internal structures of the embryos, and we provide a freely available data set that can be used to study the development of the earliest metazoans.
Puerto Rican parrot

These data represent the first assembly of a genome sequence for a critically endangered parrot (Amazona vitata) endemic to the United States, and also the first genome of a species from the diverse and ecologically important genus Amazona native to South America and the Caribbean. One sample has been selected from the non-reproductive female at Rio Abajo Breeding Facility in Puerto Rico (IACUC#201109 1), and sequenced on Illumina HiSeq platform with both fragment and paired-end sequencing approaches, resulting in a total of 42,476,949,706 bases. We predicted a total coverage depth of 28.89X of the parrot’s genome: 17.09X coverage for the short fragment reads, and 9.8X coverage for the mate pairs. The sequencing was initiated with the construction of two genome libraries: a short fragment library (~300 bp inserts) for sequencing the majority of the genome, and a long fragment library (~2.5 Kb inserts) to generate scaffolds to be used to order and assemble contigs derived from the short fragment library. The Illumina paired-end and mate-pairs reads were assembled together with Ray (http://denovocassembler.sourceforge.net), with the k-mer defined iteratively. In total, given that the genome size is predicted to be 1.58 Gb, with the total scaffold length of 1,184,594,388 bp, the overall coverage of the genome is around 76%, a value that might be slightly overestimated given that some of the scaffolds may be overlapping but could not be assembled. Filtering followed by assembly resulted in 259,423 contigs.

In accordance with our terms of use, please cite this dataset as:
Olinsky, TK; Gubier, W; Pombort, JF; Valentin, R; Martinez-Cruzado, JC (2012): Genomic data of the Puerto Rican Parrot (Amazona vitata) from a locally funded project. GigaScience.
http://dx.doi.org/10.5524/100039

Related manuscript available at doi:10.1186/2047-217X-1-14
From The BioChemical Journal, Portland Press:
Every wanted to inspect data referenced in articles? Utopia Documents allows you to interact directly with curated database entries. Play with molecular structures; edit sequence and alignment data; even plot curated tabular data yourself.

http://www.biochemj.org/bj/semantic_faq.htm
Elsevier offers data-viewers from within the article, to data stored elsewhere:
The Graph depicts the average size of a Journal of Neuroscience article and supplemental material in megabytes.

As a consequence, the Journal no longer accepts supplementary files to manuscripts, soon the supplementary material would outgrow the article volume. The burden on the peer review process became simply too large.

Editors suspect researchers to treat supplements as data dumping grounds (Emily Marcus, Cell)

Publishers cannot guarantee proper preservation and future accessibility of supp files.
Publications with data

Processed Data and Data Representations

Data Collections and Structured Databases

Raw Data and Data Sets

(1) Data contained and explained within the article

(2) Further data explanations in any kind of supplementary files to articles

(3) Data referenced from the article and held in data centers and repositories

(4) Data publications, describing available datasets

(5) Data in drawers and on disks at the institute
The Pyramid’s likely short term reality:

1. Top of the pyramid is stable but small
2. Risk that supplements to articles turn into Data Dumping places
3. Too many disciplines lack a community endorsed data archive
4. Estimates are that at least 75% of research data is never made openly available

Data on Disks and in Drawers
(1) More integration of text and data, viewers and seamless links to interactive datasets

(2) Only if data cannot be integrated in article, and only relevant extra explanations

(3) Seamless links (bi-directional) between publications and data, interactive viewers within the articles

(4) More Data Journals that describe datasets, data mgmt plans and data methods
ODE recommendations: How can publishers help to make things better

• Clearer editorial policies on the availability of underlying data
• Recommend reliable and trustworthy Data Archives to authors
• Enhance articles for better integration of underlying data
• Endorse guidelines for proper citation of data
• Launch and sponsor Data Journals
• Ensure persistent identifiers and bi-directional linking
• Partner with reliable Data Archives for further integration of Data and Publications, including interactivity for re-use.
Questions?

Eefke Smit
International Association of STM Publishers
Director, Standards and Technology
smit@stm-assoc.org

stm