

# Data Publishing: where are we?

Alberto Accomazzi

DC&P Session  
IVOA Interop Meeting  
Waikoloa, Hawaii  
27 September 2013

# The Thing about Publishing Data

- Scientists are now hearing about data publishing and data sharing, and seem receptive to the idea
- They now have multi-terabyte datasets plus code that they are willing to “put up” for others to use
- They would love to see proper attribution of their dataset and analysis work
- But they don't really want to worry about: Archival, Preservation, Nomenclature, Persistence, Discovery

# Heard on the streets

- “I’m done with this project, want to free up my hard drive and move on”
- “I’m maintaining a dataset which I keep adding data to and want to make it available to others”
- “I want to have my multi-TB data collection published with my paper”
- “I want to have my data published first so I can properly cite them in future publications”

# GHVC G001.8-10.9-173<sup>?</sup>

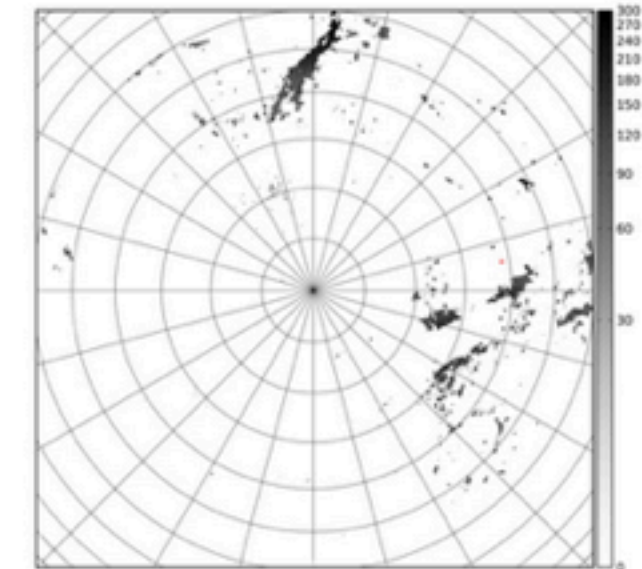
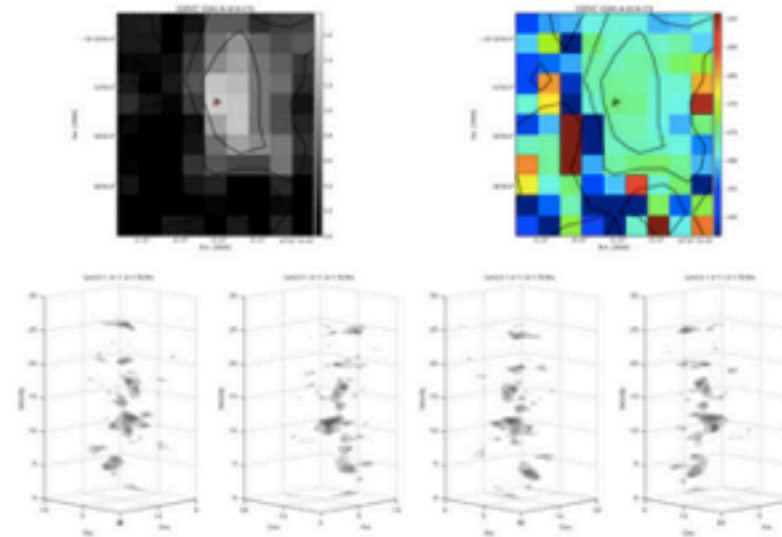
[\(Click here to download FITS cube\)](#)

## DATA

Red values indicate refit parameters.

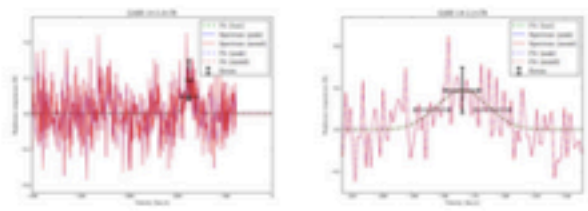
Name	(l+b+v)	GHVC G001.8-10.9-173
RA	(hrms)	18:34:43.90
Dec	(dms)	-32:30:23
VLSR	(km/s)	-173.6
VLSRerr	(km/s)	5.7
VGSR	(km/s)	-166.5
Vdev	(km/s)	-126.1
FWHM	(km/s)	19.7
FWHMerr	(km/s)	11.5
fitTb	(K)	0.09
NH	(cm <sup>-2</sup> )	2.6e+18
NHerr	(cm <sup>-2</sup> )	2.0e+18
Area	(deg <sup>2</sup> )	0.2
dx	(deg)	0.4
dy	(deg)	0.6
Flags	(-)	-
HIPASS_id	(-)	-
WW91_id	(-)	GCN_GC,N

## IMAGES

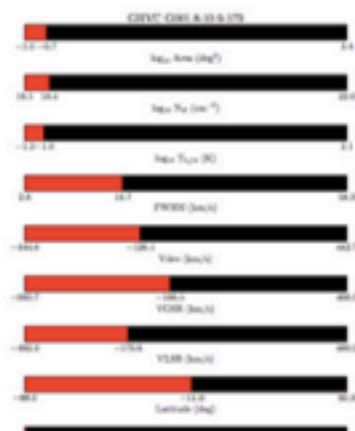


## SPECTRA

[\(Click here to download spectrum\)](#)



## PROPERTIES



## Not Found

The requested URL /vmoss/gasscat/comments.html was not found on this server.

Apache/2.2.15 (Red Hat) Server at s10.physics.usyd.edu.au Port 8080

- [GHVC G000.1-07.3-278](#)
- [GHVC G000.2-11.4-239](#)
- [GHVC G000.6-54.7-091](#)
- [GHVC G000.5-75.7-169](#)
- [GHVC G000.6-21.2-103](#)
- [GHVC G000.9-06.0-258](#)
- [GHVC G001.0-66.4-094](#)
- [GHVC G001.2-15.4-185](#)
- [GHVC G001.2-67.2-122](#)
- [GHVC G001.4-43.0-160](#)
- [GHVC G001.6-04.4-221](#)
- [GHVC G001.6-08.6-166](#)
- [GHVC G001.8-01.3-180](#)
- [GHVC G001.7-15.9-130](#)
- [GHVC G001.8-10.9-173](#)
- [GHVC G001.9-58.4-175](#)
- [GHVC G001.8-08.3-157](#)
- [GHVC G002.1-03.1-200](#)
- [GHVC G002.1-06.2-215](#)
- [GHVC G002.1-21.5-185](#)
- [GHVC G002.1-43.1-126](#)
- [GHVC G002.4-11.6-203](#)
- [GHVC G002.4-07.5-166](#)
- [GHVC G002.5-22.9-134](#)
- [GHVC G002.8-19.9-120](#)
- [GHVC G002.8-23.5-116](#)
- [GHVC G003.1-04.5-160](#)
- [GHVC G003.1-03.9-146](#)
- [GHVC G003.1-05.1-132](#)
- [GHVC G003.2-10.0-114](#)
- [GHVC G003.1-14.1-199](#)
- [GHVC G003.5-08.6-366](#)
- [GHVC G003.6-43.1-140](#)
- [GHVC G003.7-65.5-123](#)
- [GHVC G003.7-09.6-125](#)
- [GHVC G003.9-10.5-109](#)
- [GHVC G003.8-22.7-132](#)
- [GHVC G003.9-36.8-137](#)
- [GHVC G003.8-06.4-154](#)
- [GHVC G004.0-05.2-203](#)
- [GHVC G004.4-01.2-155](#)
- [GHVC G004.5-06.7-214](#)
- [GHVC G004.7-04.7-171](#)
- [GHVC G004.8-50.7-137](#)
- [GHVC G005.0-09.4-139](#)
- [GHVC G005.3-21.4-138](#)
- [GHVC G005.2-33.0-128](#)
- [GHVC G005.3-36.0-342](#)
- [GHVC G005.3-60.7-149](#)
- [GHVC G005.4-24.9-095](#)
- [GHVC G005.5-02.6-474](#)
- [GHVC G005.5-09.8-131](#)
- [GHVC G005.6-00.7-424](#)
- [GHVC G005.6-43.5-115](#)
- [GHVC G005.6-44.8-153](#)
- [GHVC G005.5-63.7-102](#)
- [GHVC G005.8-11.5-123](#)
- [GHVC G005.7-18.6-186](#)
- [GHVC G005.8-16.4-188](#)
- [GHVC G005.9-32.4-110](#)
- [GHVC G006.2-13.4-186](#)
- [GHVC G006.2-09.2-169](#)


# Existing solutions/platforms

- A variety of “data publishing” platforms have appeared in the past few years
- General-purpose repositories: Figshare, Zenodo
- Institutional repositories: CDL, Dataverse, several University repositories
- Discipline-focused initiatives: CfA Astroverse, ScienceDrive (was VObox)
- Astrophysics Archives: Chandra, MAST, CADC, VizieR, ...







KA Tel My AD: f i ju Ricl 27 Goc t Bar: M DCI API - G Wel httj wal AD: Lon bre Fly: x Twi

figshare.com/articles/W5\_CO\_3\_2\_Data\_Cubes/808583


Google Bookmarks Google Bookmark Bookmark on Delicio Import to Mendeley Altmetric it Other Bookmarks

  [Browse](#) [Upload](#) [Sign up](#) [Login](#)

## W5 CO 3-2 Data Cubes

 W5S.fits	<a href="#">download</a>
 W5Ridge.fits	<a href="#">download</a>
 W5N.fits	<a href="#">download</a>
 W5SE.fits	<a href="#">download</a>
 W5S201.fits	<a href="#">download</a>
 w5outflows_ellipses.reg	<a href="#">preview</a>   <a href="#">download</a>

[Download all](#)

26 views 0 shares 

Published on 26 Sep 2013 - 02:50 (GMT)  
Filesize in total is 1.68 GB

### Categories

- Astrophysics
- Galactic Astronomy

### Authors

Adam Ginsburg  
Jonathan Williams  
john bally

### Tags

- jcmt
- carbon monoxide
- data cube
- herp

Feedback?

Share this: [f Share](#) 0 [Tweet](#) 0 [+1](#) 0 [Embed\\*](#)

Cite this: W5 CO 3-2 Data Cubes. Adam Ginsburg, Jonathan Williams, john bally. figshare.  
http://dx.doi.org/10.6084/m9.figshare.808583  
Retrieved 08:34, Sep 27, 2013 (GMT)

DOI

# Zenodo (CERN)

- Upload from desktop or Dropbox
- Create & curate content via communities  
(Institution, research group, conference, workshop)
- Obtain DOI, cite & share
- Alternative metrics integration (Twitter, Facebook)
- Data stored in CERN cloud infrastructure
- Open source & open access (all types of material)
- Reporting to funding agencies



KA Tel My AD: f i ju Ricl 27 Goc t Bar: M DCI API - G Wel htt: wai AD: Lon & bre: Fly: x Twi

← → ↻ <https://zenodo.org/collection/user-cfa-sci-ed> ☆ 📧 📧 📧 📧 📧

Google Bookmarks Google Bookmark Bookmark on Delicio Import to Mendeley Altmetric it Other Bookmarks

**zenodo** Research. Shared.


Search Communities Browse Upload Get started Email Password Sign in

Home / Communities / Harvard-Smithsonian Center for Astrophysics Science Education Department

Search 12 records for:  [Q Search](#)

# Harvard-Smithsonian Center for Astrophysics Science Education Department

## Recent Uploads



**03 September 2013** **Book** **Open access**

### Challenges in Physical Science: Windmills (Workbook)

Coyle, Harold P. ; Hines, John L. ; Rasmussen, Kerry J. ; Sadler, Philip M.

A Supplemental Curriculum for Middle School Physical Science. [...]

Uploaded by [CfA Library](#) on 04 September 2013.

[View](#)

Community collection

### Harvard-Smithsonian Center for Astrophysics Science Education Department

The [Science Education Department \(SED\)](#) of the [Harvard-Smithsonian Center for Astrophysics](#) develops curricula and materials that reflect current scientific and educational philosophy. SED staff includes education researchers, scientists, teachers, media specialists (see the [Science Media Group's](#)

<https://zenodo.org>

DOI



oac.cdlib.org/findaid/ark:/13030/c8r78fzq/

**OAC**  
Online Archive of California

Search OAC  go

Home Browse Institutions Browse Collections Browse Map About OAC Help **What is OAC?**

> Home > UC San Diego > Research Data Curation Program [Share / Save](#)

**Collection Guide** <http://www.oac.cdlib.org/findaid/ark:/13030/c8r78fzq>

Collection Title: The guide to the Santa Fe Light Cone Simulation research project files RCIDC.0001

Collection Number: RCIDC.0001

Get Items: [Online items available](#)  
[Contact UC San Diego: Research Data Curation Program](#)

**View entire collection guide** [?](#)  
PDF (144.75 Kb) HTML

**Search this collection**  
 go  
 Entire Collection Guide  Online Items

**Table of contents** [?](#)

**Collection Overview**

**Description** The project files consists of data in three broad categories: the simulation data ("Data at Redshift" components); analysis tools and example scripts (Data Processing Tools) for processing the data; and project administration and background documents (Historical Documents) related to the project. All these materials were created between 2005 and 2012, beginning with a proposal for the LUSciD Project, continuing on to the simulation data, and ending with the recent analysis tools. The historical documents are proposals and progress reports that were part of grants or requests for computational resources supporting the research. The component for analysis tools and example scripts contains the source code to yt (<http://yt-project.org/>), which was used to produce the example data analysis results. The results are a combination of structured text, binary files, and images. The historical documents and analysis tools are described in greater detail in their component descriptions.

**Background** The Santa Fe Light Cone Simulation project was the result of an ongoing effort by the Laboratory for Computational Astrophysics, beginning with the LUSciD Project in 2005. This led to the development of the ENZO simulation software to the point where it was able to complete a seven-level adaptive mesh refinement (AMR) cosmology simulation.

**Extent** 683.0 Gigabyte(s) 39 digital objects collectively containing 1,797 digital files of various types.

**Collection Overview**  
[Collection Details](#)  
[Project Background](#)  
[Scope and Contents note](#)  
[Use References](#)  
[Arrangement note](#)  
[Immediate Source of Acquisition note](#)  
[Processing Information note](#)  
[Access](#)  
[Rights](#)  
[License](#)  
[Preferred Citation](#)

**Collection Contents**

- [Data at Redshift = 3 \(RD0009\)](#)
- [Data at Redshift=2.75 \(RD0010\)](#)
- [Data at Redshift=2.5 \(RD0011\)](#)
- [Data at Redshift=2.4 \(RD0012\)](#)
- [Data at Redshift=2.3 \(RD0013\)](#)
- [Data at Redshift=2.2 \(RD0014\)](#)
- [Data at Redshift=2.1 \(RD0015\)](#)
- [Data at Redshift=1.9 \(RD0017\)](#)
- [Data at Redshift=1.8 \(RD0018\)](#)
- [Data at Redshift=1.7 \(RD0019\)](#)

DOI

Harvard Dataverse Network > CfA Dataverses >

POWERED BY THE **Dataverse Network** PROJECT V. 3.6

## Greg Snyder Dataverse

REPLICATION DATA FOR: MODELING MID-INFRARED DIAGNOSTICS OF OBSCURED QUASARS AND STARBURSTS

hdl:10904/10188  
Version: 2 - Released: Wed Aug 21 08:32:23 EDT 2013

[Cataloging Information](#) **DATA & ANALYSIS** [Comments \(0\)](#) [Versions](#)

**i** Use the check boxes next to the file name to download multiple files. Data files will be downloaded in their default format. You can also download all the files in a category by checking the box next to the category name. You will be prompted to save a single archive file. Study files that have restricted access will not be downloaded.

Select all files  Total Number of Files: 2 Total Downloads: 5

<input type="checkbox"/> Documentation		
<input type="checkbox"/> README.txt Plain Text - 7 KB - 3 downloads	<a href="#">Download</a>	Description of data file
<input type="checkbox"/> FITS variables		
<input type="checkbox"/> agn_midir_Snyder_et_al_2013.fits application/fits - 19 MB - 2 downloads	<a href="#">Download</a>	Data file This is a FITS file with 8 HDUs total. In addition to the primary HDU of type Image, it contains 7 Image HDU(s); The following recognized metadata keys have been found in the FITS file, and their values will be made searchable in the DVN, once the study has been indexed: EXTNAME;

[Collapse \[-\]](#)

# DOI

# SciDrive

- VAO/JHU solution to data storage + sharing
- REST API, Dropbox API, Openstack architecture, OpenID + OAuth
- 100TB available so far, no limits (so far) to file size and usage
- Automated metadata extraction for FITS, CSV, Excel, etc.
- Provision for sharing data via http links

[EMail contact](#)

# SciDrive

A free, easy-to-use data hosting and sharing service  
for the science community

[Sign up](#)   [Login](#)

VAO OpenID account is required to use the system.

Drag-and-drop files to the browser window and we will not only store your data, but will also recognize the filetype, extract whatever metadata we can from your files.

If we find a tabular data set – we will automatically load it into a database table.

Share your data with the scientific community by simply providing URLs that everyone can browse. Create a writeable share for your workgroup to analyze data together.

Supported by:



# Store What and Where?

- catalogs, tables, plots
- raw data (images, spectra, cubes)
- software (source + executables)
- data reduction workflows, documentation
- the project website itself
- who decides what's useful and worth preserving?
- who will curate the data in the long run?