Joint Declaration of Data Citation Principles

Synthesis Group

By Mercè Crosas, IQSS, Harvard University

@mercecrosas

Data in scholarly publications: Current State

PUBLICATION

in IC 348 (HD 281159) is confirmed to reside in the Perseus cloud, but there might be [illegible] other [illegible] st[illegible]s that in-teract with the cloud (throu[illegible] wind[illegible] a[illegible] UV radiat[illegible]) even though they were not [illegible]ecessar[illegible] f[illegible] cloud com-plex (see, e.g., Walawender et al. 2004; Ridge et al. 2006a; Kirk et al. 2006; Rebull et al. 2007). There is also a large number of nebulous objects associated with outflow shocks (i.e., HH objects and H_2 knots) that have been identified in the cloud complex (Bally et al. 1996b, 1997; Yan et al. 1998; Walawender et al. 2005b; Davis et al. 2008).

The whole Perseus region was first surveyed in ^{12}CO by Sargent (1979), and since then has been mapped in CO at different angular resolutions (all with beams $> 1'$) by a number of other authors (e.g., Bachiller & Cernicharo 1986; Ungerechts & Thaddeus 1987; Padoan et al. 1999; Sun et al. 2006). These maps show a clear velocity gradient in the Perseus molecular cloud complex where the central cloud (LSR) velocity increases from about 4.5 km s^{-1} at the western edge of the cloud to about 10 km s^{-1} at the eastern end. The large velocity gradient in the gas across the entire complex and the fact that different parts of the Perseus cloud appear to have different distances (see above) could possibly indicate that the complex is made up of a superposition of different entities. Recently, the Perseus molecular cloud complex was also observed (and studied) in its entirety in the mid- and far-infrared as part of the "From Molecular Cores to Planet-forming Disks" (aka c2d) *Spitzer* Legacy Project (Jørgensen et al. 2006; Rebull et al. 2007; Evans et al. 2009).

2. DATA

In this paper, we use the ^{12}CO(1–0) and ^{13}CO(1–0) data collected for Perseus as part of the COordinated Molecular Probe Line Extinction Thermal Emission (COMPLETE) Survey of Star Forming Regions,[6] described in detail by Ridge et al. (2006b). The ^{12}CO and ^{13}CO molecular line maps were observed between 2002 and 2005 using the 14 m Five College Radio Astronomy Observatory (FCRAO) telescope with the SEQUOIA 32-element focal plane array. The receiver was used with a digital correlator providing a total bandwidth of 25 MHz over 1024 channels. The ^{12}CO $J = 1$–0 (115.271 GHz) and the ^{13}CO $J = 1$–0 (110.201 GHz) transitions were observed simultaneously using an on-the-fly (OTF) mapping technique. The beam telescope at these frequencies is about 46″. Both maps of ^{12}CO and ^{13}CO are essential for a thorough study of the outflow and cloud properties. The ^{12}CO(1–0) is a good tracer of the cool and massive molecular outflows and provides the information needed to study the impact of these energetic phenomena on the cloud. The ^{13}CO(1–0) provides an estimate of the optical depth of the ^{12}CO(1–0) line and can be used to probe the cloud structure and kinematics.

Observations were made in $10' \times 10'$ maps with an effective velocity resolution of 0.07 km s^{-1}. These small maps were then patched together to form the final large map of Perseus, which is about $6.°25 \times 3°$. Calibration was done via the chopper-wheel technique (Kutner & Ulich 1981), yielding spectra with units of T_A^*. We removed noisy pixels that were more than 3 times the average rms noise of the data cube, the entire map was then resampled to a 46″ grid, and the spectral axis was Hanning smoothed[7] (necessary to keep the cubes to a size manageable by the three-dimensional visualization code, see below). During the obs[illegible]ations [illegible] Pe[illegible] clo[illegible] different OFF positions were us[illegible] depend[illegible]g o[illegible] loca[illegible] was being mapped. Some of these OFF pos[illegible] fa[illegible] significant, emission which resulted in an artificial absorption feature in the final spectra. Gaussians were fitted to the negative feature in regions with no gas emission, and the fits were then used to correct for the contaminating spectral component. The resulting mean 3σ rms per channel in the ^{12}CO and ^{13}CO maps are 0.25 and 0.20 K, respectively, in the T_A^* scale. Spectra were corrected for the main beam efficiencies of the telescope (0.49 and 0.45 at 110 and 115 GHz, respectively), obtained from measurements of Jupiter.

3. COMPUTATIONAL MOTIVATION AND THREE-DIMENSIONAL VISUALIZATION

This study allows for a test of the effectiveness of three-dimensional visualization of molecular line data of molecular clouds in R.A.–decl.–velocity (p–p–v) space as a way to identify velocity features, such as outflows, in large maps.[8] The primary program used for three-dimensional visualization is 3D Slicer[9] which was developed originally at the MIT Artificial Intelligence Laboratory and the Surgical Planning Lab at Brigham and Women's Hospital. It was designed to help surgeons in image-guided surgery, to assist in pre-surgical preparation, to be used as a diagnostic tool, and to help in the field of brain research and visualization (Gering 1999). The 3D Slicer was first used with astronomical data by Borkin et al. (2005) to study the hierarchical structure of star-forming cores and velocity structure of IC 348 with ^{13}CO(1–0) and C^{18}O(1–0) data.

We divided the Perseus cloud into six areas (with similar cloud central LSR velocities) for easier visualization and outflow search in 3D Slicer (see below). The borders of these areas are similar to those named by Pineda et al. (2008), who also based their division mainly on the cloud's central LSR velocity. The regions, whose outlines are shown in Figure 1, overlap between 1 and 3 arcmin to guarantee complete analysis. This overlap was checked to be sufficient based on the fact that new and known outflows which crossed regions were successfully double-identified.

For each area, an isosurface (constant intensity level) model was generated in 3D Slicer, using the ^{12}CO(1–0) map. The threshold emission intensity level chosen for each isosurface model was the lowest level of emission above the rms noise level for that particular region. This creates a three-dimensional model representing all of the detected emission. The high-velocity gas in this three-dimensional space can be identified in the form of spikes, as shown for the B5 region in Figure 2, which visually stick out from the general distribution of the gas. These sharp protrusions occur since one is looking at the radial velocity component of the gas along the line of sight, thus causing spikes wherever there is gas at distinct velocities far away from the main cloud velocity. Instead of having to go through each region and carefully examine each channel map, or randomly scroll through the spectra by hand, this visualization allows one to instantly see where the high-velocity points are located (see also Borkin et al. 2007, 2008).

6 See http://www.cfa.harvard.edu/COMPLETE.

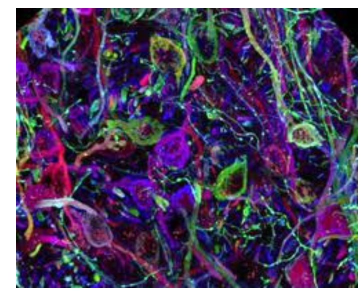
7 See http://www.cfa.harvard.edu/COMPLETE/projects/outflows.html for a link to the molecular line maps.

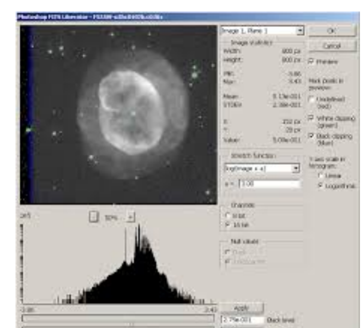
8 This work is done as part of the Astronomical Medicine project (http://am.iic.harvard.edu) at the Initiative in Innovative Computing at Harvard (http://iic.harvard.edu). The goal of the project is to address common research challenges to both the fields of medical imaging and astronomy including visualization, image analysis, and accessibility of large varying kinds of data.

9 http://www.slicer.org/

Most Publications:

1. **No link to data**
2. **Or** footnote with a link to data, but **no persistent identifier**
3. **Or** persistent citation to data, but **no common machine-actionable metadata**





Are we ready to make **data citable products** of research?

A Brief History of Citing Data

Standards in Scholarly Citation:
author/creator, title, dates, publisher
or distributor of the work

1906
Chicago Manual of Style

~1960
First scientific digital repositories
(e.g., ICPSR,
World Data Center)

1979
ASBR (“Data File” type)
MARC (machine readable catalog)
Domain Repositories (e.g., GenBank)

1999 - Now
Data Repositories
(e.g., NESSTAR, Dataverse,
Dryad, Figshare)
DOI services (DataCite)

Altman M., Crosas M., 2014, “The Evolution of Data Citation: From Principles to Implementation” IASSIST Quarterly, *In Press*

The Making of the Principles

- Decades of research and practices in data citation
- Consolidated to a single set of Principles
- By a synthesis group representing 25+ organizations
- Driven by the premise that:

"sound, reproducible scholarship rests upon a foundation of robust, accessible data”

"data should be considered legitimate, citable products of research”

Synthesis Group

CODATA Principles

Amsterdam Manifesto

DataCite Principles

Joint Declaration of Data Citation Principles

DC1

Data Citation Principles

+ contribution from 25 organizations (DCC, ORCID, Publishers, Repositories, …)

The Principles

1 Importance
2 Credit and Attribution
3 Evidence
4 Unique Identification
5 Access
6 Persistence
7 Specificity and Verifiability
8 Interoperability and flexibility

https://www.force11.org/datacitation

Example

A data citation generated by the **Dataverse Repository** (dataverse.org)

Resolves to landing page with access to metadata, docs, and data

Authors, Year, Dataset Title, DOI, Data Repository, UNF, version

Principle 2:
Credit and Attribution

Principle 4, 5, 6:
Unique Id Access Persistence

Principle 7:
Specificity and Verifiability

Principle 8: Interoperability and flexibility:
Repository exports citation metadata in XML, JSON formats

See Details at: Altman and King, 2007, A Proposed Standard for the Scholarly Citation of Quantitative Data; Altman, Crosas, 2013, The Evolution of Data Citation