

Astronomy Dataverse: *enabling scientist data publishing*



An Open-Source Application for
Publishing, Citing and Discovering Research Data



**SEAMLESS
ASTRONOMY**
Linking scientific data, publications, and communities



where should data live?

- Refined data sets are published by astronomers in long lived repositories;
- Published data appear in ADS & are “searchable”
- Published data are **reused** and **cited**, giving astronomers credit for that work.



Hmm, that sounds like a **goal** of the Virtual Observatory... what is the **reality** of data publishing today?

data publishing is driven by the literature.

The screenshot shows a web browser window displaying the abstract page for the paper "Hubble Space Telescope Proper Motions and Stellar Dynamics in the Core of the Globular Cluster 47 Tucanae" by Dean E. McLaughlin et al. The page is from The Astrophysical Journal Supplement Series, 166:249-297, 2006 September. The abstract text describes the use of HST imaging to derive proper motions and U- and V-band magnitudes for 14,366 stars in the core of the globular cluster 47 Tucanae. The page includes a table of contents on the left, a list of online materials (FITS file, master frame reference image, machine-readable table, Table 4, tar file, proper-motion catalog) on the left, and a list of related links at the bottom. A blue arrow points from the "Online Material" section to a callout box on the right.

Online Material
FITS file

▶ [Master frame reference image](#)

Machine-readable table

▶ [Table 4](#)

Tar file

▶ [Proper-motion catalog](#)

References: McLaughlin et al. 2006; <http://adsabs.harvard.edu/abs/2006ApJS..166..249M>

Tables, Tables in tar file

Code in tar file

FITS Files

```

1 #
2 # MACRO pmdat REQUIRES ONE COMMAND-LINE ARGUMENT ...
3 #
4 # $1 = ID number of a star in file '...datfile'
5 #   in either of two formats: M11111 = an exact ID label in the file
6 #   : 11111 = integer part [>0] of a known ID
7 # -- OR --
8 #
9 # $1=0 or $1=[any string not starting with 'M']
10 #   will choose a star at random from the file '...datfile'
11 #
12 #
13 # MACRO pmdat WILL OPTIONALLY TAKE A SECOND ARGUMENT...
14 #
15 # $2 = PRINT [optional]
16 #
17 #   if second argument exists and is PRINT
18 #   then star data are printed to file with extension '.DATA'
19 #
20 #   if second argument is anything else, or does not exist at all,
21 #   then star data are echoed to screen instead
22 #
23 #
24 #
25 # Usage... at the sm prompt, type
  
```

References: McLaughlin et al. 2006; <http://adsabs.harvard.edu/abs/2006ApJS..166..249M>



VizieR

Tables, Tables in tar file

FITS Files

Code in tar file

```

1 #
2 # MACRO pmdat REQUIRES ONE COMMAND-LINE ARGUMENT ...
3 #
4 # $1 = ID number of a star in file "...datafile"
5 #   in either of two formats: M:llll = an exact ID label in the file
6 #                           : llll = integer part (>=0) of a known ID
7 # -- 0 ---
8 #
9 # $! = 0 if $1=[any string not starting with 'M']
10 #   will choose a star at random from the file "...datafile"
11 #
12 #
13 # MACRO pmdat [ID] [OPTIONALLY TAKE A SECOND ARGUMENT...]
14 #
15 # $2 = PRINT [optional]
16 #
17 #   if second argument exists and is PRINT
18 #   then star data are printed to file with extension ".DATA"
19 #
20 #   if second argument is anything else, or does not exist at all,
21 #   then star data are echoed to screen instead
22 #
23 #
24 #
25 # Usage... at the sm prompt, type

```

References: McLaughlin et al. 2006; <http://adsabs.harvard.edu/abs/2006ApJS..166..249M>



VizieR

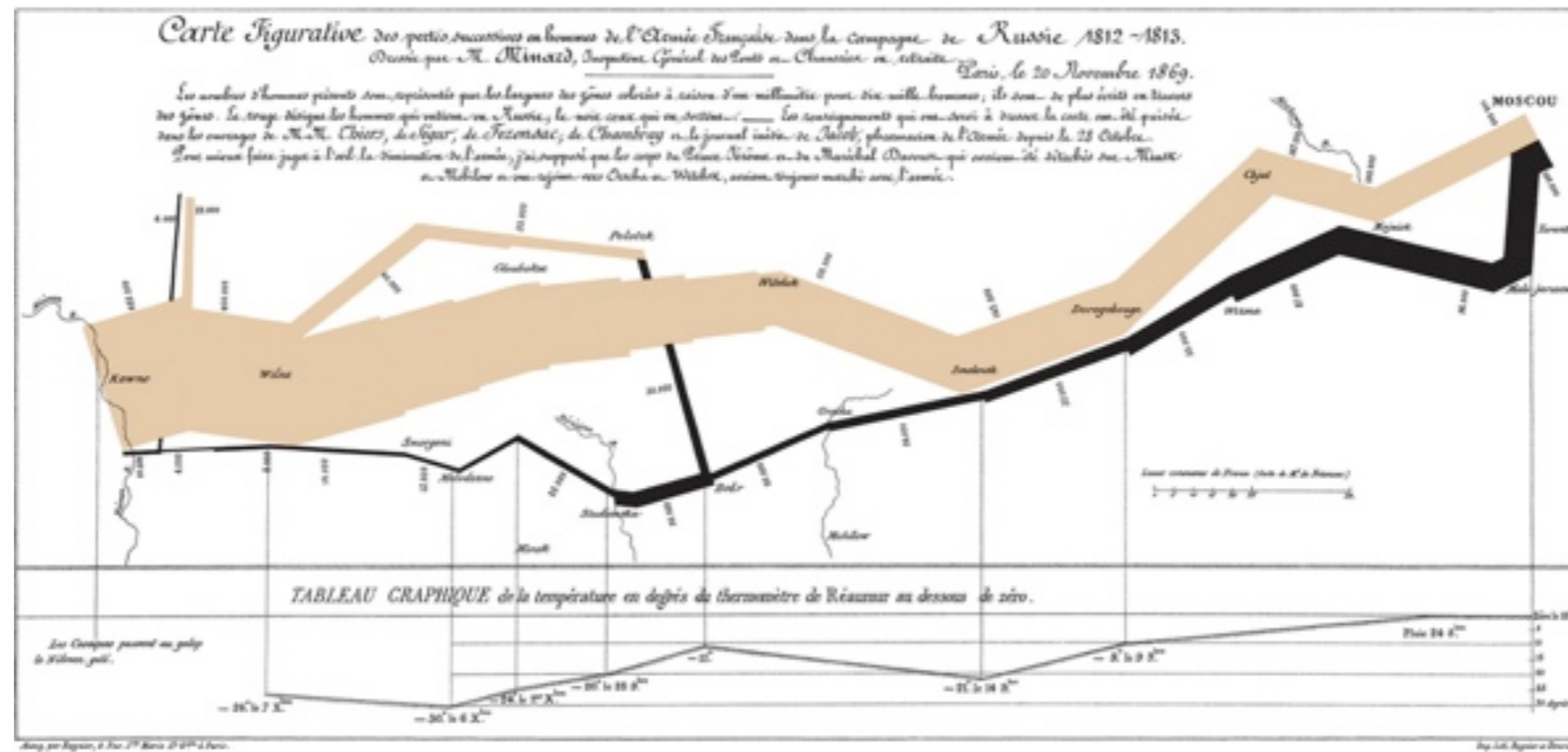
Tables, Tables in tar file

FITS Files

Code in tar file

Unlinked Data is LOST data!

References: McLaughlin et al. 2006; <http://adsabs.harvard.edu/abs/2006ApJS..166..249M>



And now for a remix...

Consider Minard's “infographic” charting the demise of Napoleon’s army on its roundtrip to Moscow...except instead of losing soldiers, we ask about **losing data behind or in a paper...**

References: Charles Minard (1781-1870) (see upload log) [Public domain], via Wikimedia Commons


```

1 # MICRO PRINTS REQUIRES ONE COMMAND-LINE ARGUMENT ...
2 #
3 # S1 = 33 number of a star in file "_datafile"
4 # in either of two formats: M11111 = an exact ID label in the file
5 # ; 11111 = longer path (4) of a known ID
6 # -- OR --
7 # S1=0 or S1=Only string not starting with "W"
8 # Will choose a star at random from the file "_datafile"
9 #
10 # S2 = PRINT [optional]
11 #
12 # MICRO PRINTS WILL OPTIONALLY TAKE A SECOND ARGUMENT ...
13 #
14 #
15 # If second argument exists and is PRINT
16 # then star data are printed to file with extension ".data"
17 #
18 # If second argument is anything else, or does not exist at all,
19 # then star data are echoed to screen instead
20 #
21 # Usage: at the prompt, type

```

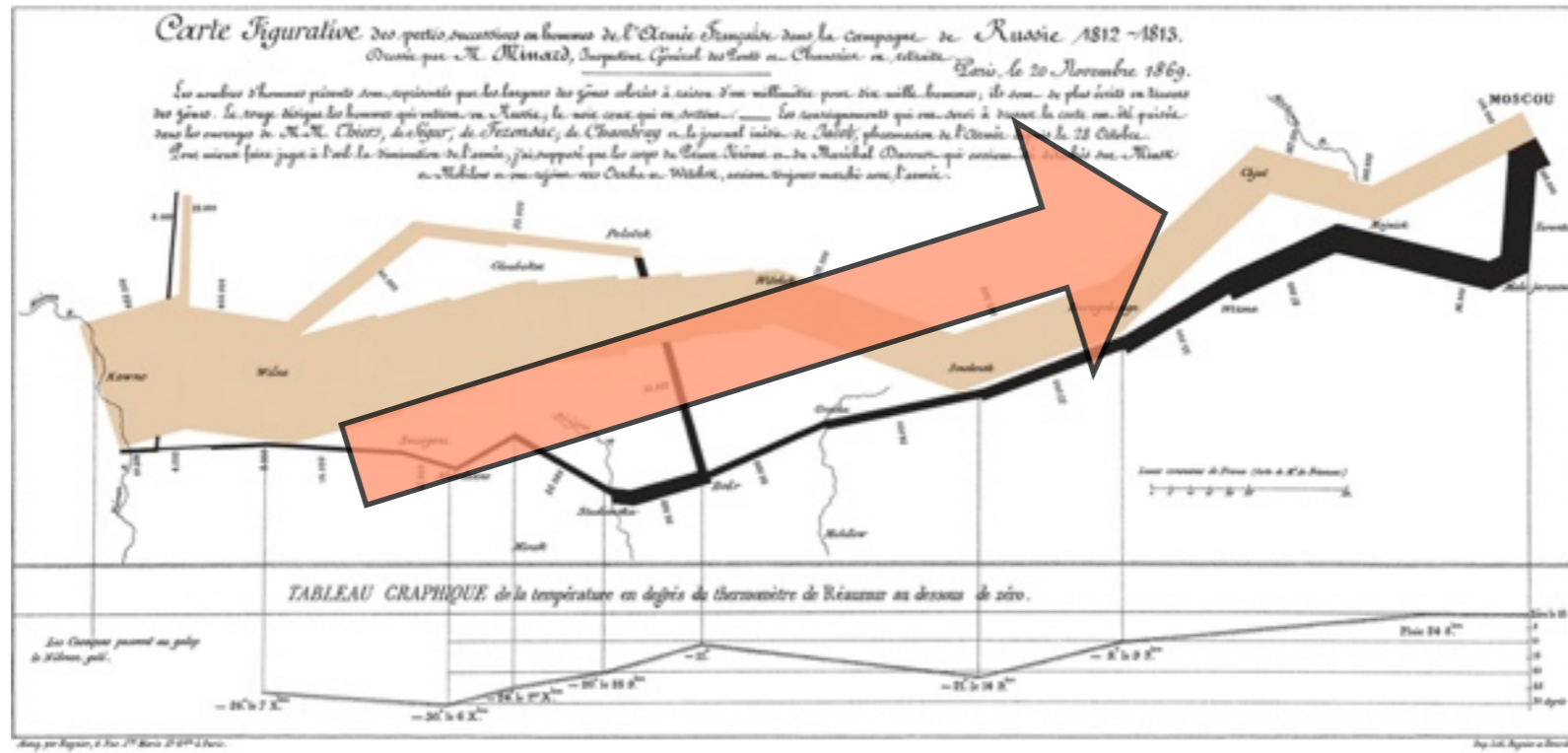
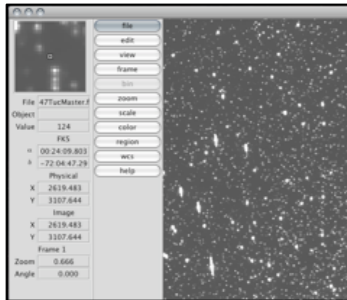


Table 2 WFC3 and ACS Observations of 47 Tucanae

Data Set	Program ID	N _{ex}	Filter	Date
MEYLANG...	7842	15	F330W	1995 Oct 25 = 1995.82
MEYLANG...	6482	16	F300W	1997 Nov 3 = 1997.84
GILLIUS1...	6261	28	F330W	1999 Jul 5 = 1999.51
MEYLANG...	2863	16	F300W	1999 Oct 28 = 1999.82
GILLIUS2...	5206	11	F330W	2001 Jul 15 = 2001.53
WFC-MEUR...	9623	20	F437W	2002 Apr 5 = 2002.26
HEC-MEUR...	9623	40	F437W	2002 Apr 5 = 2002.26
HEC-BORG...	9623	10	F437W	2002 Apr 13 = 2002.28
WFC-KING...	7443	6	F437W	2002 Jul 7 = 2002.52
HEC-KING...	7443	20	F437W	2002 Jul 24 = 2002.56

3.3.3 Astrometric Calibration

We now have a position for the cluster center in the reference frame, which is based on the distance corrected and rotated frame of the first image of 47 Tuc. In order to transform the cluster frame positions into absolute right ascension and declination, we used the image header information from several WFC3 images (GILLIUS1, GILLIUS2, GILLIUS3, and GILLIUS4) to obtain absolute positions for seven stars—five stars at the center and two stars in the outskirts. These four images were taken at different parallaxes and orientations, so they should all use different parallax rates and give independent estimates of the absolute coordinates.

Losses from Data to Literature

- Raw data:
 - ➡ might already be in a telescope archive
 - ➡ linkage partially fixed by post-pub curation
- Theoretical data;
- Analysis codes and logs;
- Processed data:
 - ➡ Reduced data; mosaics;

References: Charles Minard (1781-1870) (see upload log) [Public domain], via Wikimedia Commons

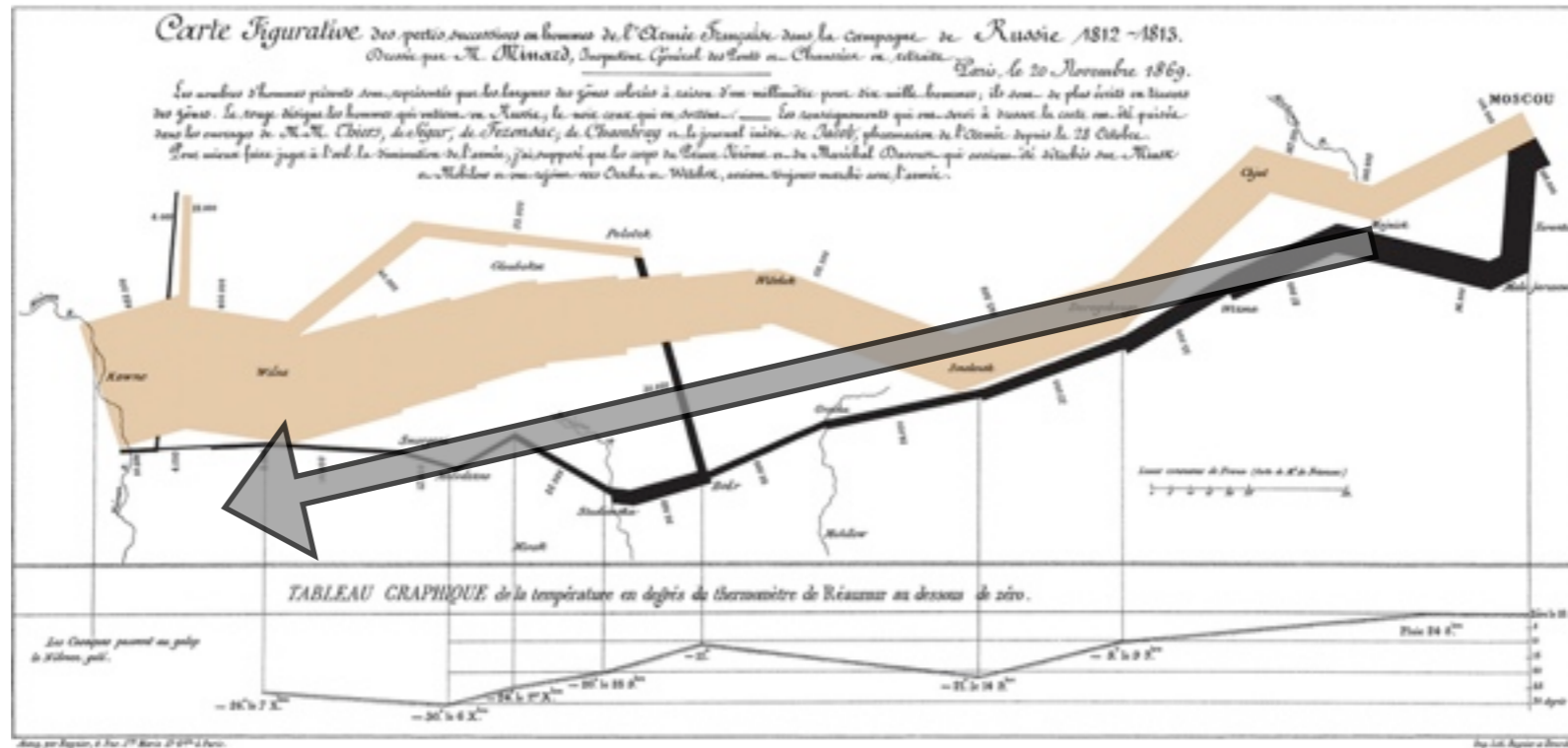


Table 2

WFC3 and ACS Observations of 47 Tucanae

Data Set	Program ID	N _{ex}	Filter	Date
MEYLANG...	7942	15	F330W	1995 Oct 25 + 1995.82
MEYLANG...	6482	16	F300W	1997 Nov 3 + 1997.84
GILLIET...	6261	28	F330W	1999 Jul 5 + 1999.51
MEYLANG...	2863	16	F300W	1999 Oct 28 + 1999.82
GILLIET...	5206	11	F330W	2001 Jul 15 + 2001.53
WFC-MBER...	9623	20	F437W	2002 Apr 3 + 2002.26
HRC-MBER...	9623	40	F437W	2002 Apr 3 + 2002.26
HRC-BORG...	9623	10	F437W	2002 Apr 11 + 2002.26
WFC-KING...	9443	6	F437W	2002 Jul 7 + 2002.52
HRC-KING...	9443	20	F437W	2002 Jul 24 + 2002.56

3.3.3 Astrometric Calibration

We now have a position for the cluster center in the reference frame, which is based on the distance corrected and rotated frame of the first image of SDSS-WCS. In order to transform the cluster frame positions into absolute right ascension and declination, we used the image header information from several WFC3 images ([GILLIET2001](#), [GILLIET2002](#), [GILLIET2003](#), and [GILLIET2004](#)) to obtain absolute positions for seven stars—five stars at the center and two stars in the outskirts. These four images were taken at different parallaxes and orientations, so they should all use different parallax rates and give independent estimates of the absolute coordinates.

Losses *and* Gains from Literature to Archives

- Post-publication curation creates or captures:
 - ➡ *SIMBAD* objects; big archive data references;
 - ➡ large machined tables captured by CDS;
- Data still leaks:
 - ➡ data products that are not machined tables;
 - ➡ data in tar files.
 - ➡ data from external websites (linked as footnote URLs)

References: Charles Minard (1781-1870) (see upload log) [Public domain], via Wikimedia Commons

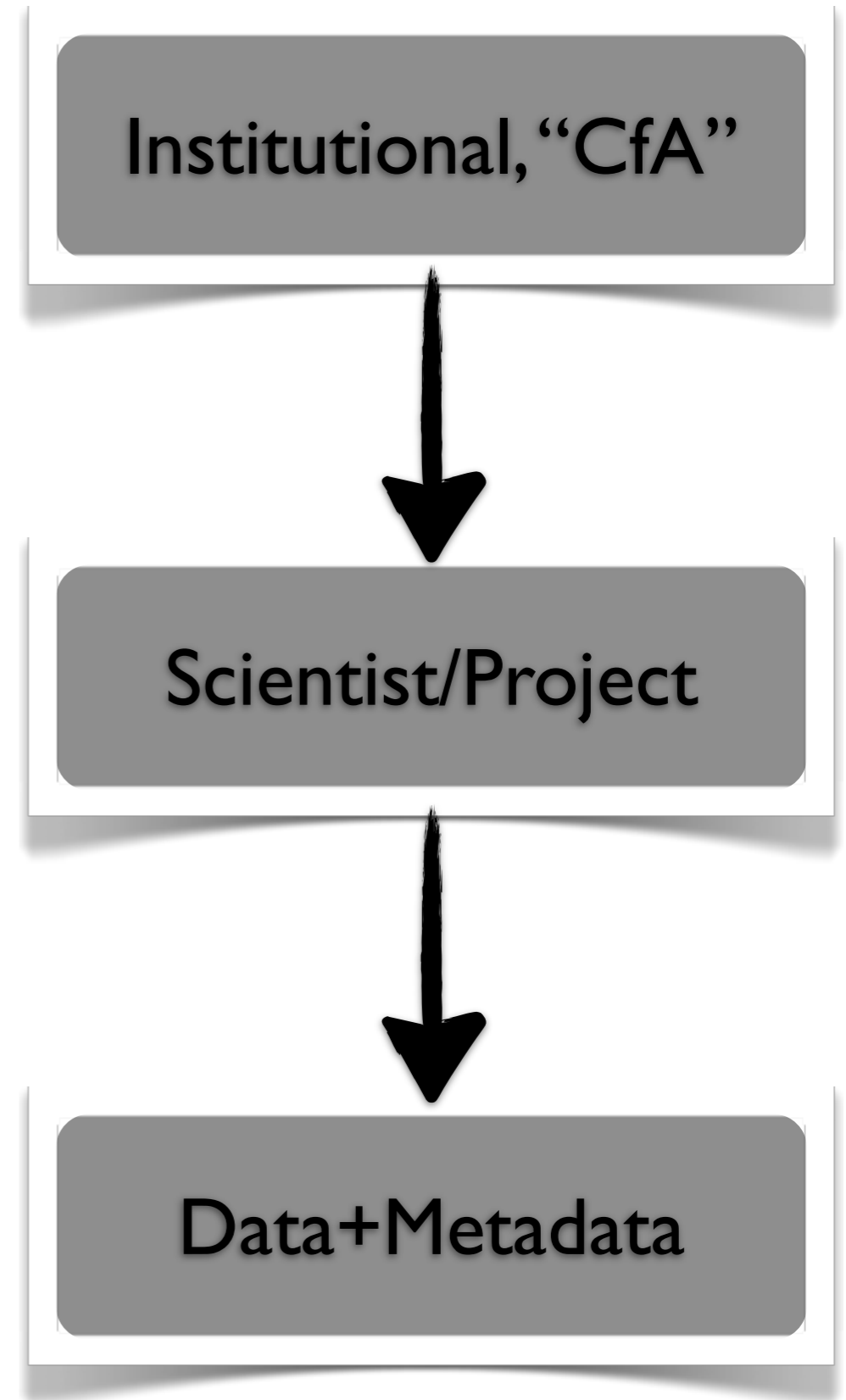
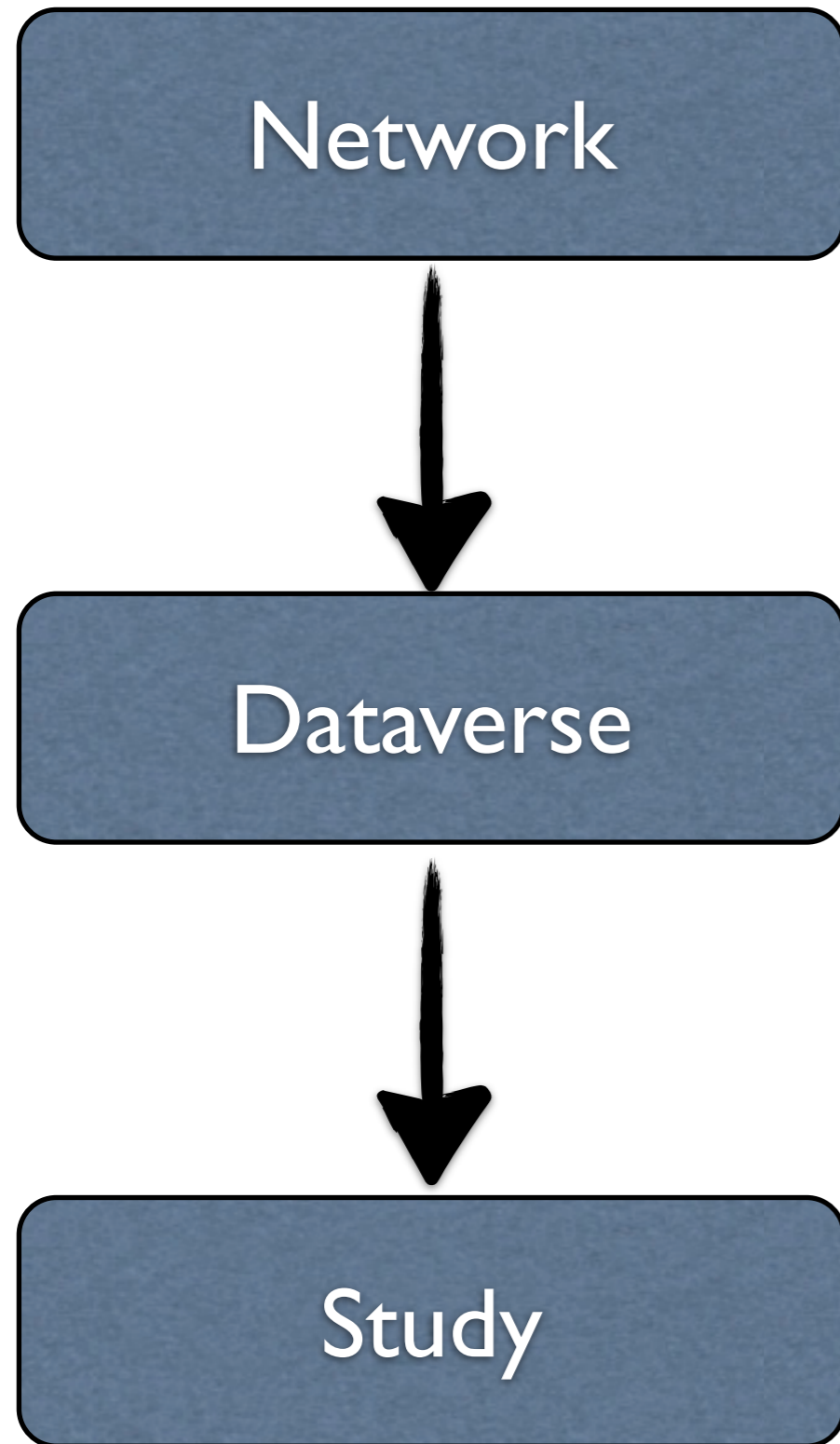


- The Dataverse Network (DVN) Project was built originally for Social Science Data;
- Collaboration between the Harvard/CfA “Seamless Astronomy” team and the DVN team to reuse this framework for Astronomy Data.
 - Conducting Data “Interviews” with Astronomers to deduce their needs;
 - Metadata mapping between the Data Documentation Initiative (DDI) standard used by DVN and Astronomy’s VO standards;
 - Technical training for astronomers to use platform.
- Institutional support from Harvard Library to support the infrastructure and training for Astronomy.



- Gives **ownership and recognition** to data owner
- Generates a **persistent data citation**
- Converts data sets to a **preservable** and verifiable format
- Distributes data to the **public**, but also supports **restricted** access
- Indexes all metadata for quick data **discovery**
- Supports **subsetting and analysis** for (some) data files
- Can be branded as your web site.
- Inter-operates** with other systems using **standards**

Dataverse “Overview”



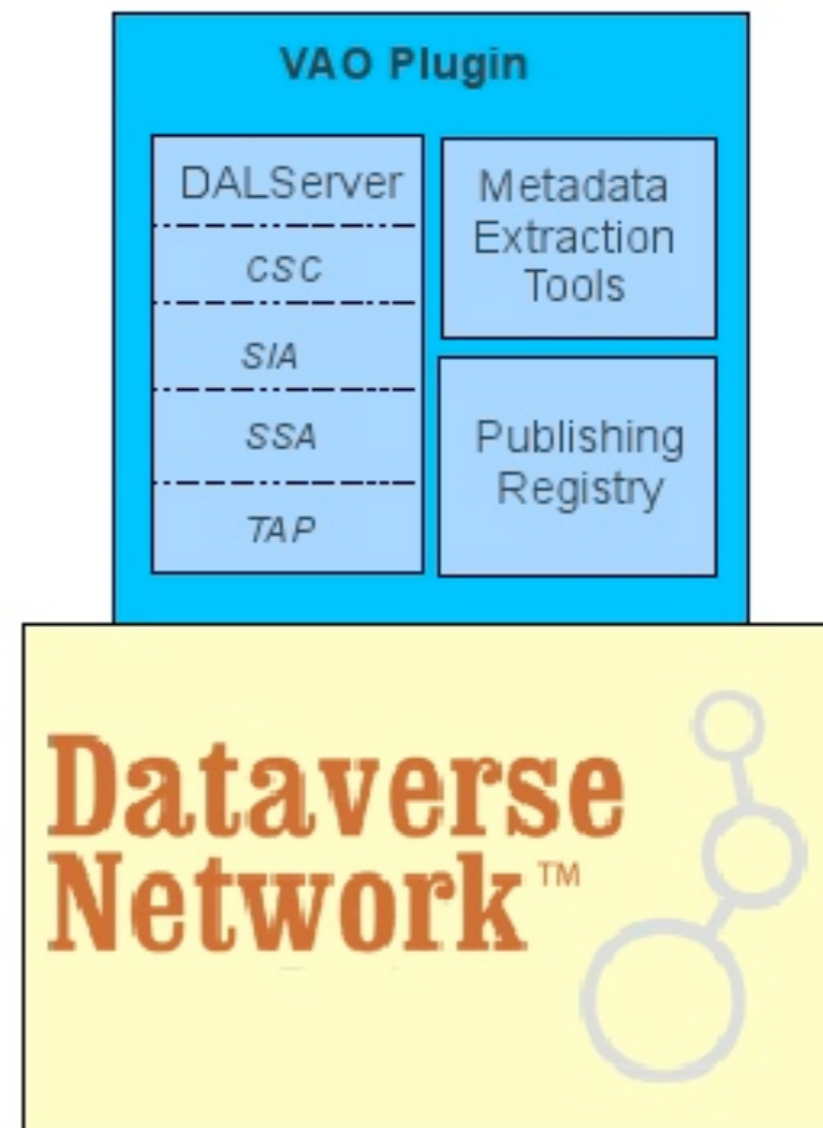


Studies, Bags, & Blobs

References: Ton Zijlstra; <http://www.flickr.com/photos/tonz/2463875144/>

VAO Plugin to DVN

- Index individual “datatypes” in a published data study;
- Expose services for datatypes;
- Manage publication registration to VO.



theastrodata.org

How do Astronomers Use a Dataverse?

or what do you need to publish your data.

data interviews

- We interviewed 10 astronomy research groups (or individuals) about their data;
- Followed a “data interview” format;
- Coded and extracted information about typical science stories in astronomy.

we asked a bunch of you some questions

“Mostly FITS”

“thousands of lines, hundreds of columns. hundreds of MBs at most.”

“Terabite-ish.”

“Currently KB, MB
(reduced)”

“General public”

“No. No Licensing; No obligations.”

we asked a bunch of you some questions

“I don't have a website where I store these data. Most of it is in various stages of mess.”

“if we were rich and organized, we would be like Sloan...”

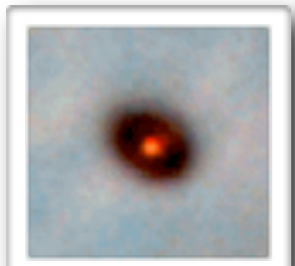
“Visibility from ADS, Vizier, arXiv... Interface: 1. ability to retrieve the data, 2. simple visualization, 3. VO-interoperability”

“We don't anticipate any fancy interactive data browsing capability. You just download the data and you do anything you like with it.”

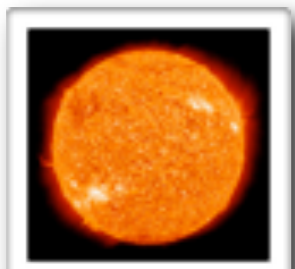
Archetypes in a Dataverse



Asteroid You have small, data sets you'd like to see stay in reliable orbits.



Protostar You're young and eager to become a full-grown star, so you want to share all the data you can, and embed links to it in your publications.



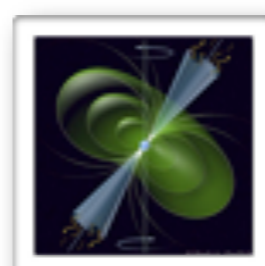
Main-sequence Star You've been at this for a while, so you have long data history and a good future. You'd like to upload important data to go with "old" papers now, and more in the future.



Cluster You collect things in catalogs and lists, and you want to group the catalogs for the greater good.



Supernova Your disks are **EXPLODING** with data, and you don't know what to do with it. You want to permalink vast data sets directly to papers, and more...



Pulsar You really like it when things *change*. Time-domain astronomy is your thing, and you want online identifiers that understand time.



Galaxy You love everything, but you're organized. You make and collect Surveys you don't want to lose, and you want people to find them from far away.



Quasar Your energy is nearly unlimited, so you suck up (mine) and spit out as much data as you can find. And you like to share in showy ways.



Black Hole You suck down any and all data, with unbridled appetite. Dataverse is *NOT* for you.

