Astronomy Dataverse:

enabling scientist data publishing

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

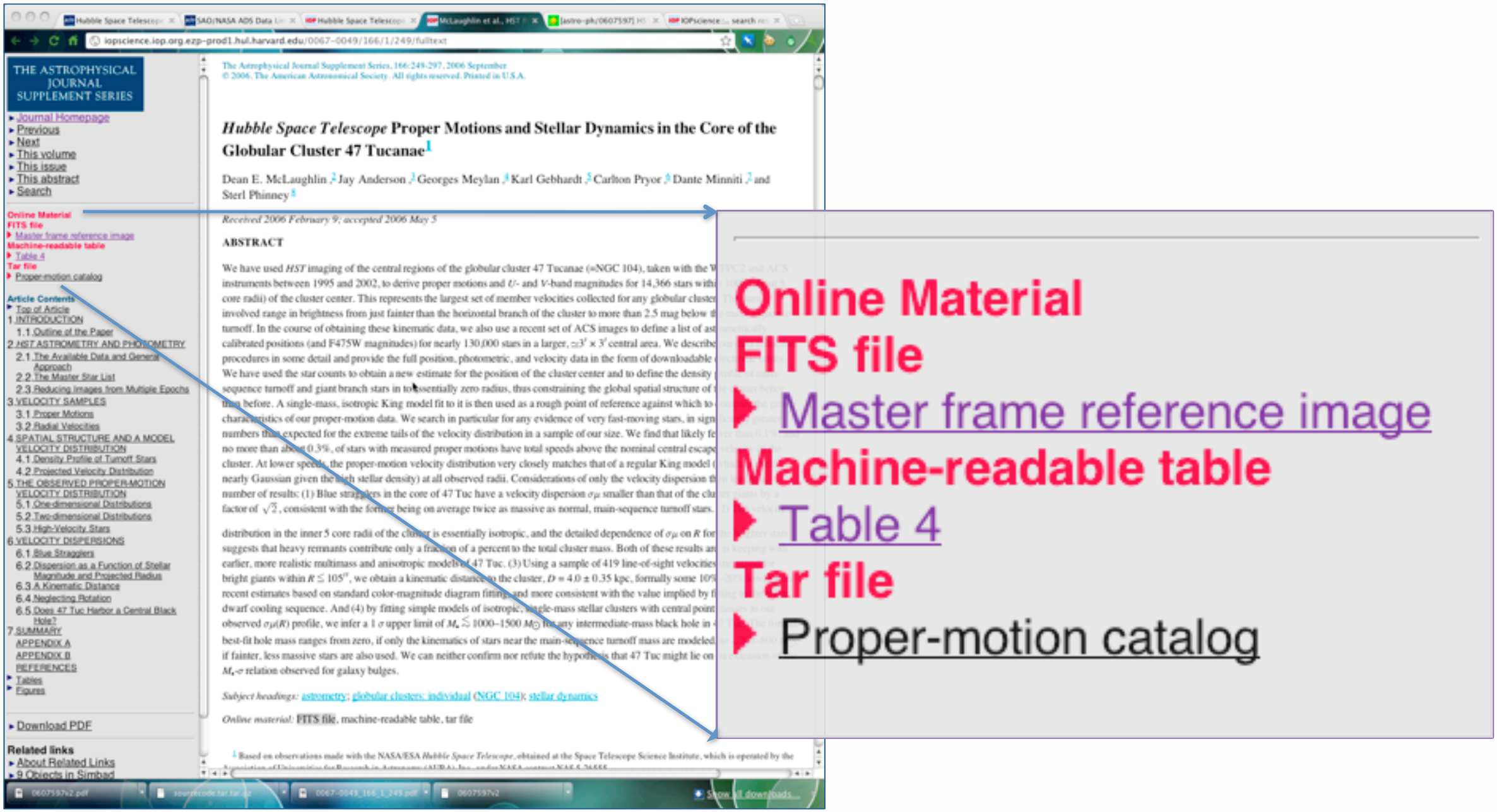
SEAMLESS
ASTRONOMY
Linking scientific data, publications, and communities

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.

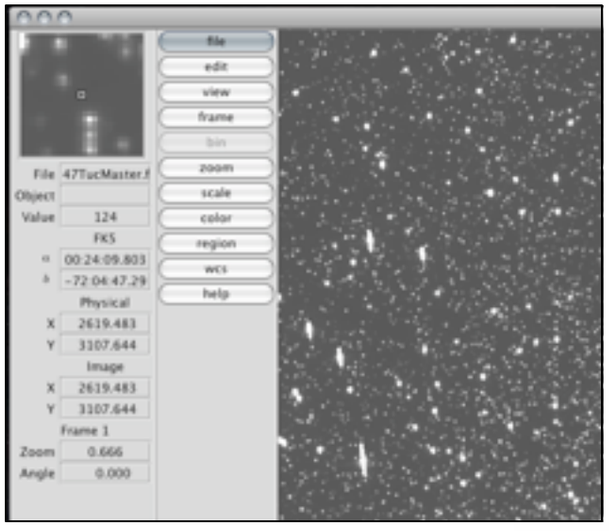


References: McLaughlin et al. 2006; http://adsabs.harvard.edu/abs/2006ApJS..166..249M

Tables, Tables in tar file

Code in tar file

FITS Files

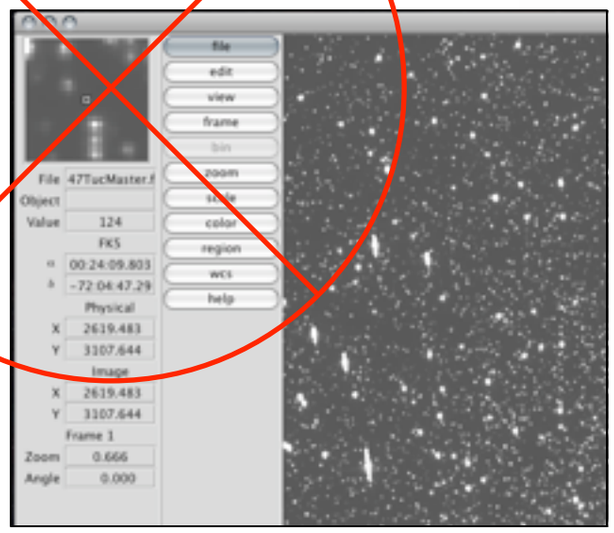


References: McLaughlin et al. 2006; http://adsabs.harvard.edu/abs/2006ApJS..166..249M

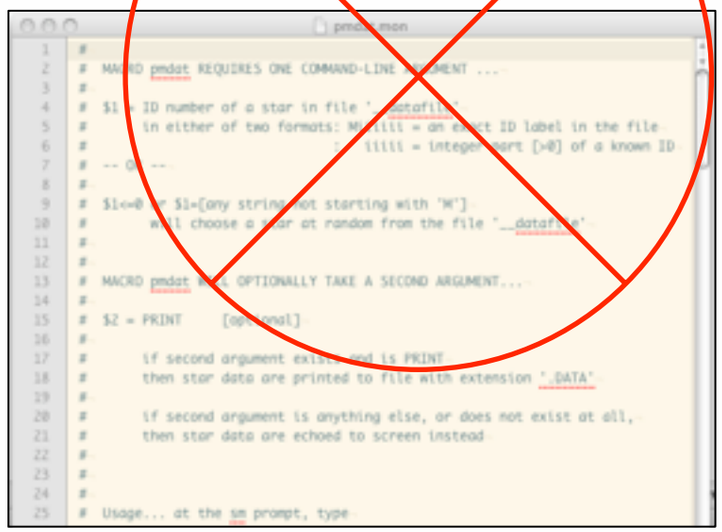
VizieR

Tables, Tables in tar file

FITS Files



Code in tar file

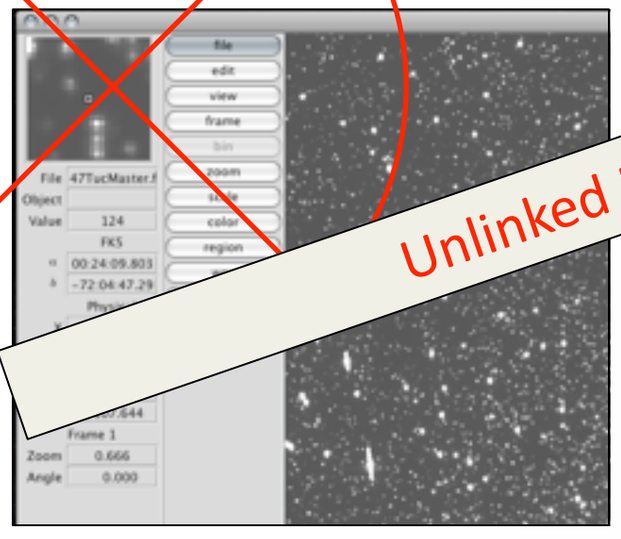


References: McLaughlin et al. 2006; http://adsabs.harvard.edu/abs/2006ApJS..166..249M

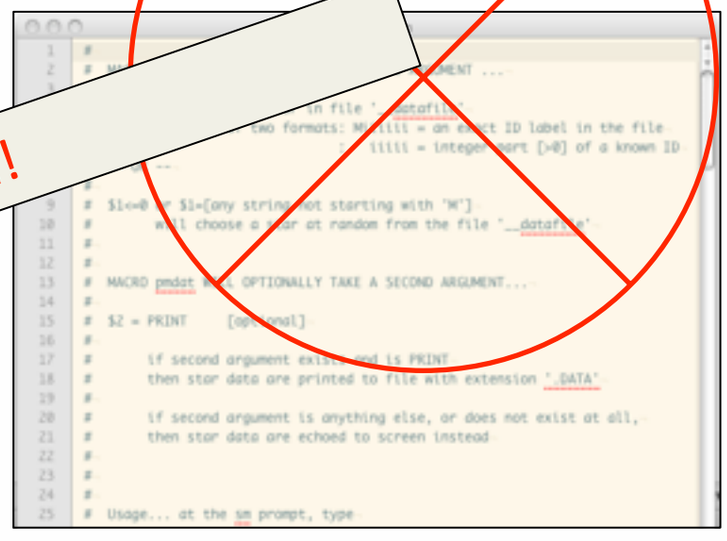
VizieR

Tables, Tables in tar file

FITS Files

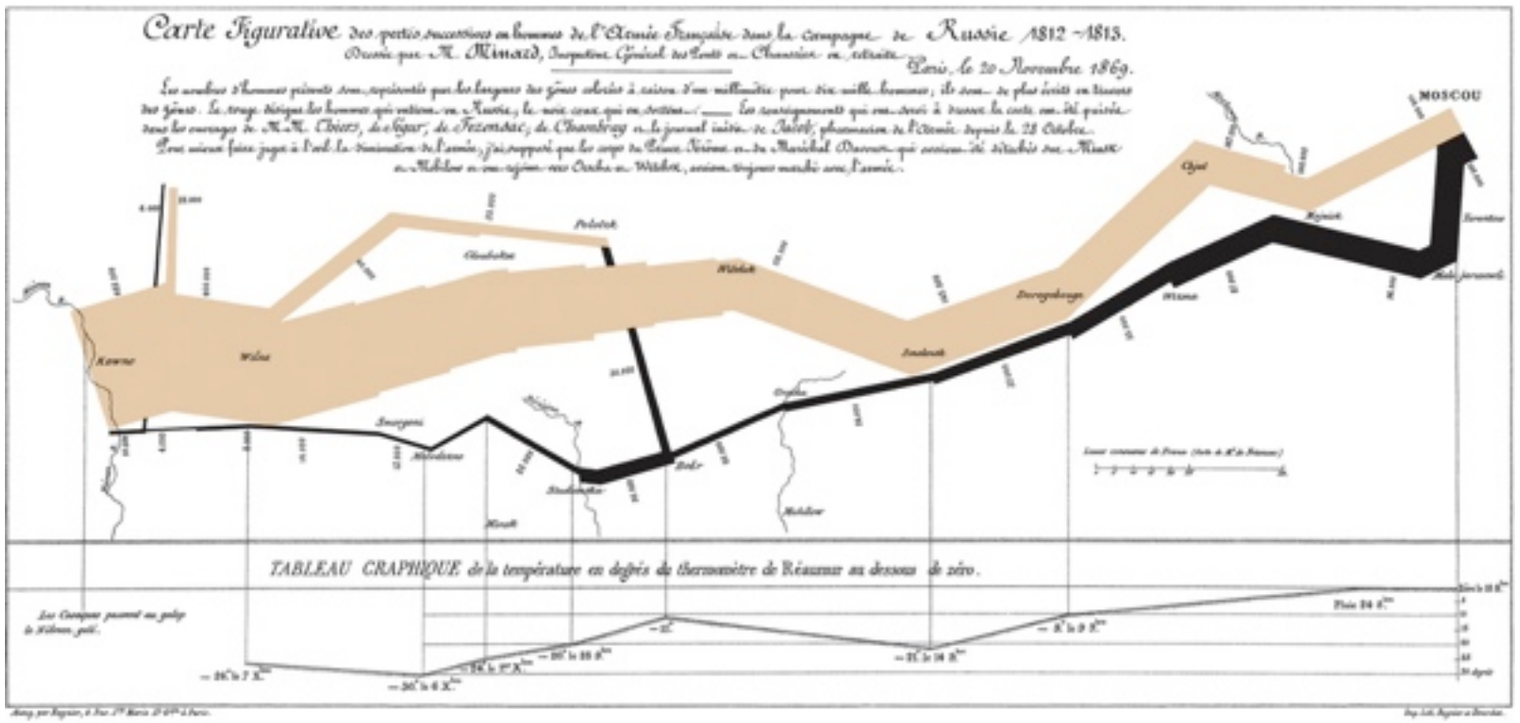


Code in tar file



Unlinked Data is LOST data!

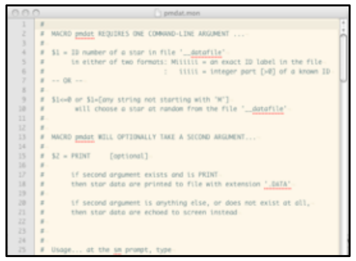
References: McLaughlin et al. 2006; http://adsabs.harvard.edu/abs/2006ApJS..166..249M

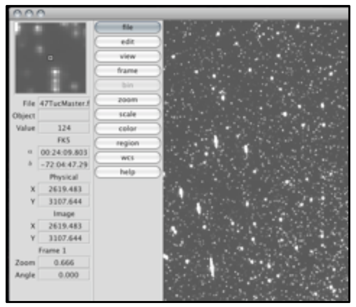


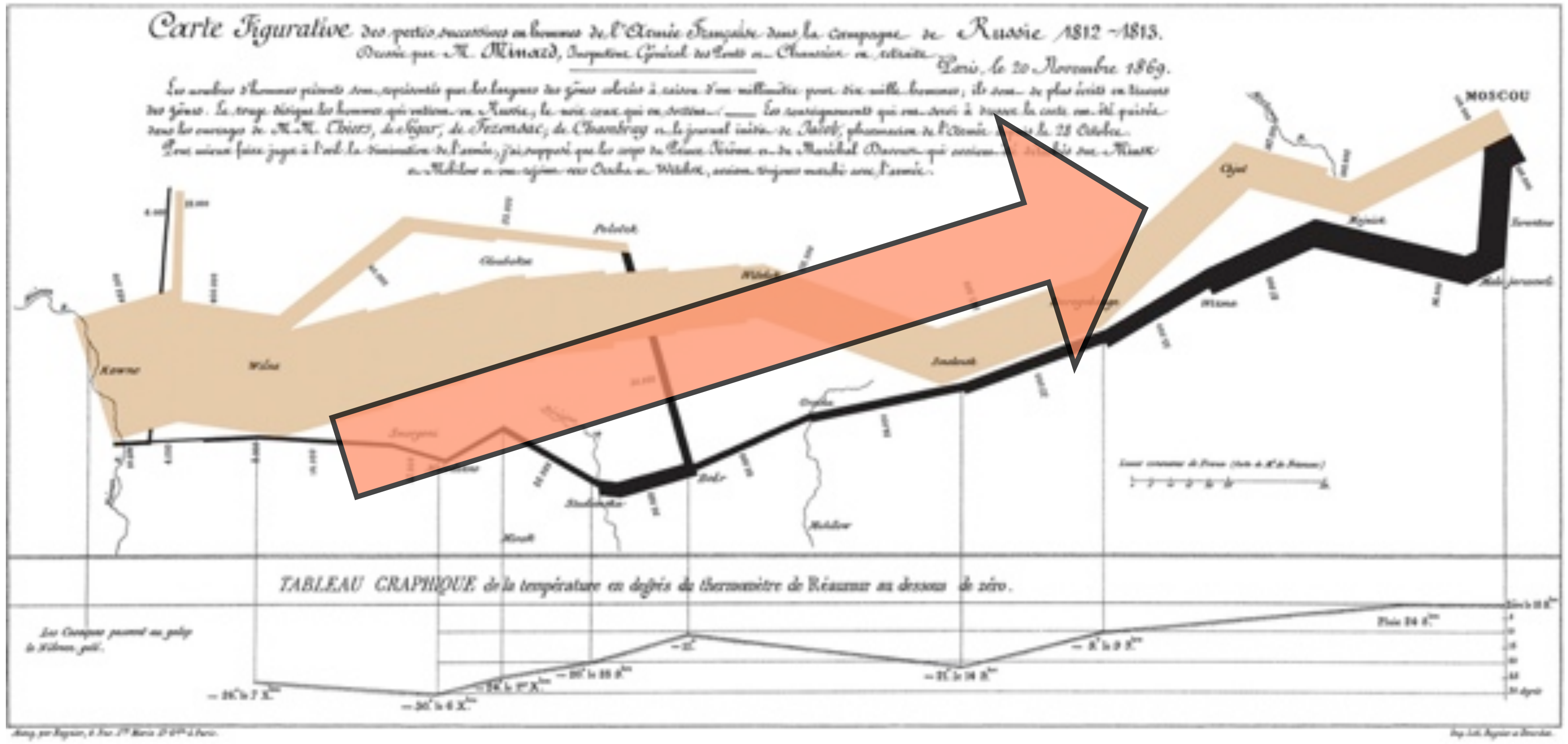
And now for a remix...

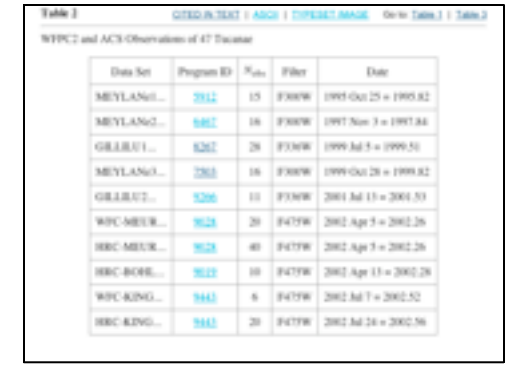
Consider Minard's "infographic" charting the demise of Napoleon's army on its roundtrip to Moscow...except instead of losing soldiers, we ask about ***losing data behind or in a paper...***

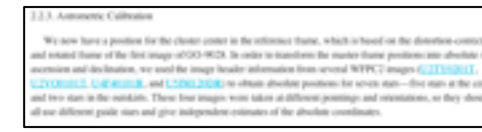
References: Charles Minard (1781-1870) (see upload log) [Public domain], via Wikimedia Commons







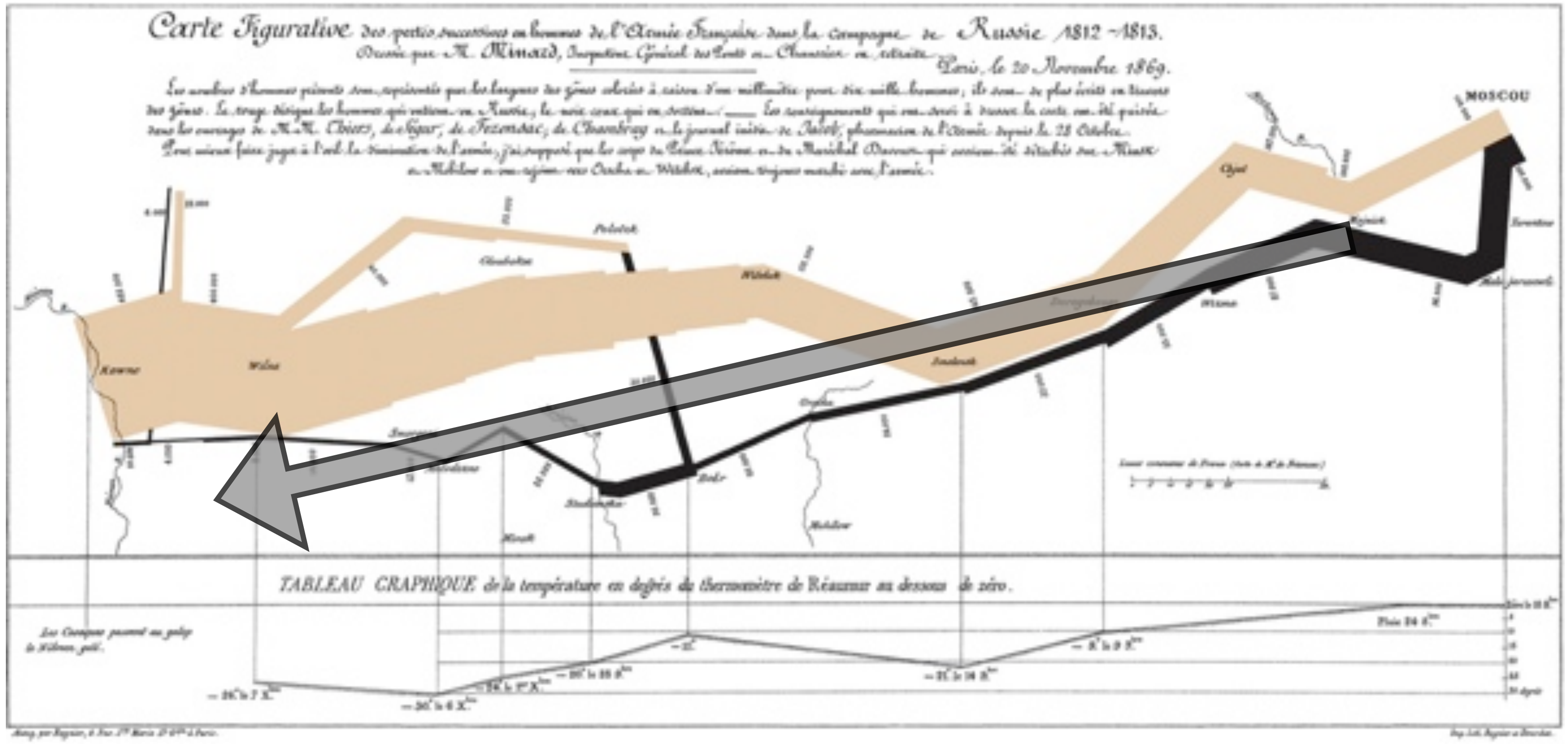


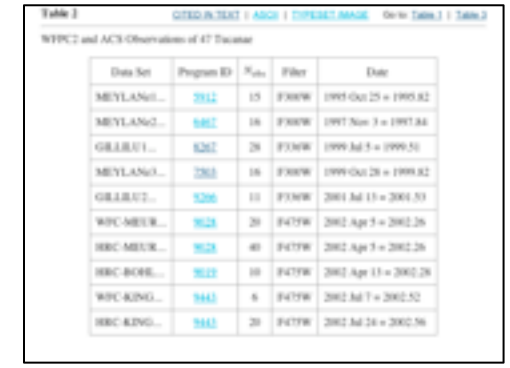


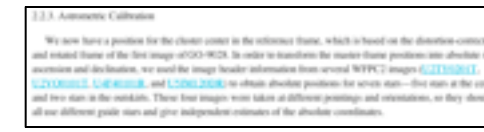
Losses from Data to Literature

- 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







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™ Project

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

Created at IQSS at Harvard University

- ☑ 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.

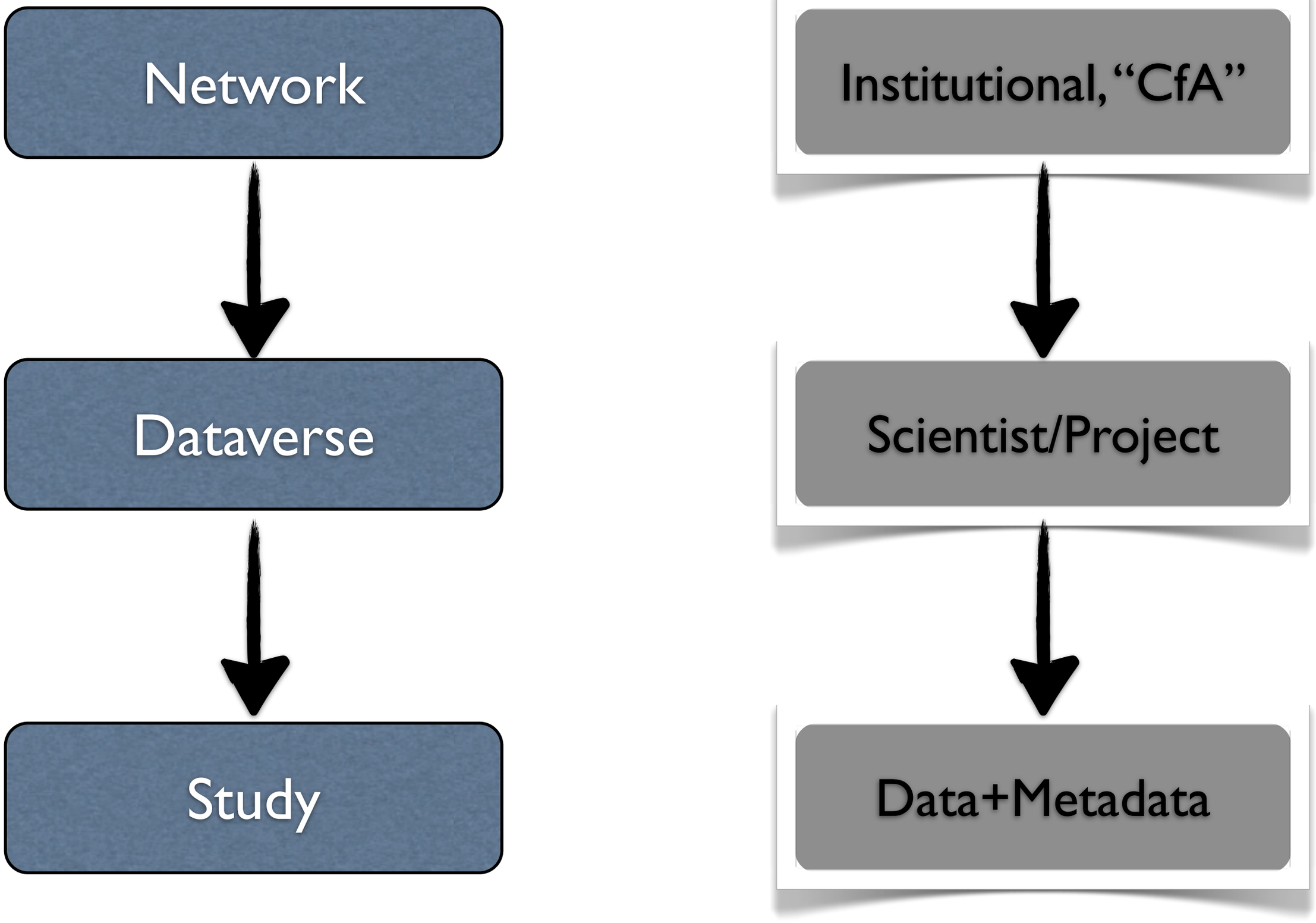
The Dataverse Network™ Project

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

Created at IQSS at Harvard University

- ☑ 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"

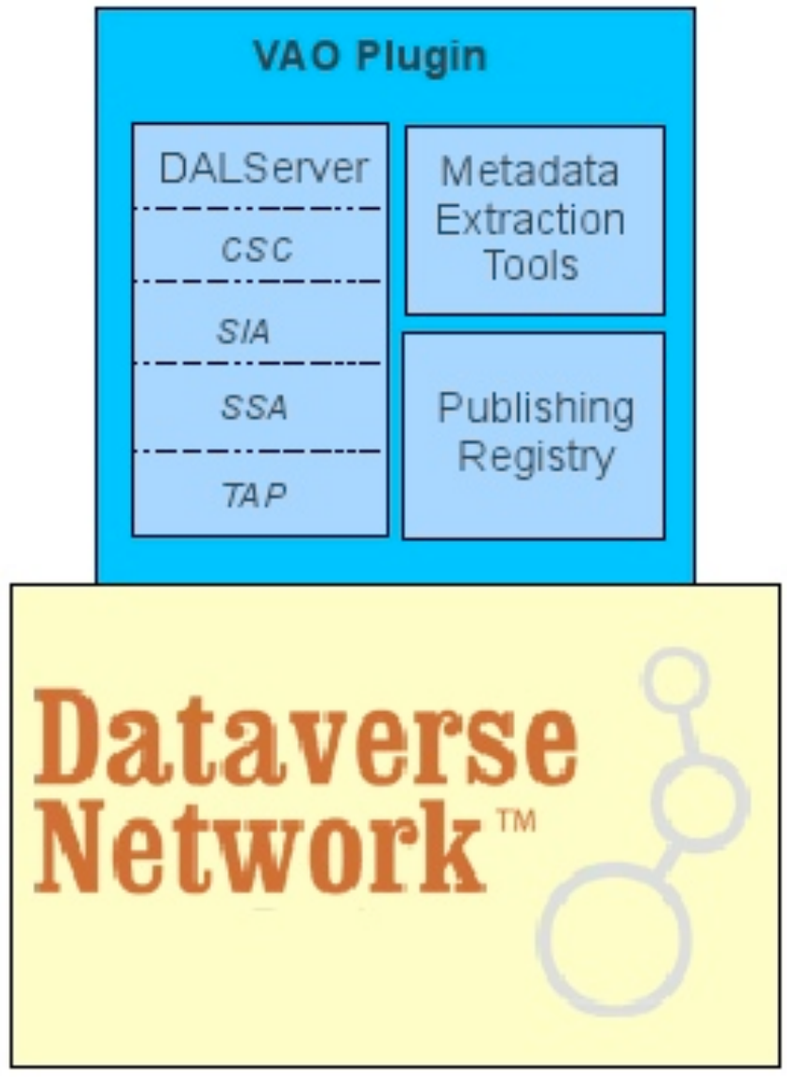


Studies, Bags, & Blobs

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."

"Currently KB, MB (reduced)"

"Terabite-ish."

"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."

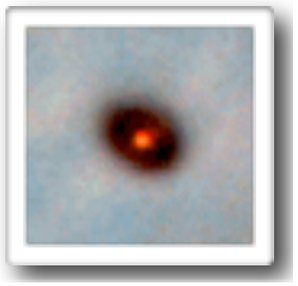
"if we were rich and organized, we would be like Sloan..."

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

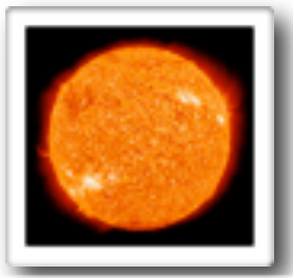
"We don't anticipate any fancy interactive data browsing capability. You just download the data and you do anything you like with it."

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.

The Dataverse Network™ Project

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