Stewarding systems: database curation and preservation

UNIVERSITY
LIBRARIES

Karl Nilsen
SAA 2014
2014-08-15
knilsen@umd.edu

Research Data Services

lib.umd.edu/data
lib-research-data@umd.edu

Data management, curation, publishing, preservation, and related schol. comm.

Extragalactic Distance Database

Compiled from various data sources, both literature and observations

MySQL, file system, PHP

Roughly 500GB, 118000 files

OPTIONAL: Enter Galaxy Name: Display only tables with info on this galaxy Reset Search

Redshift Catalogs

LEDA
on
Entries: 100631

2MRS K<11.75
on
Entries: 43526

- All
- ID_2MASXJ
- RAJ
- DEJ
- Glon
- Glat
- SGL
- SGB
- K_c
- H_c
- J_c
- K_tc

2MASS K<11.25 V
on
Entries: 24746

2M++
on
Entries: 64745

- All
- ID_2MASXJ
- J2000
- Glon
- Glat
- SGL
- SGB
- Ag
- Ks
- Vhel
- Vls
- Vcmb

Submit All Reset

Summary Distances

Cosmicflows-2 Distances
on
Entries: 8163

- All
- Dist
- DM
- eD
- C
- T
- L
- M
- S
- N
- H
- F

EDD Distances
on
Entries: 3529

Quality Distances
on
Entries: 658

- All
- objname
- grpname
- Mod_mean
- Mod_SBF
- Mod_ceph
- Mod_TRGB
- Mod_other
- Source

Cosmicflows-1 Distances
on
Entries: 1797

SFI++
on
Entries: 5780

Virgo/Fornax SBF

PGC	J2000	Name	g-z	e_gz	m_sbf	e_msbf	DM	e_dm	dist	e_d	Altname
			mag	mag	mag	mag	mag	mag	Mpc	Mpc	
12636	J032222.7-372351	FCC19	1.066	0.025	29.258	0.036	31.532	0.074	20.2	0.7	ESO301-08
12651	J032241.7-371230	FCC21	1.368	0.007	29.676	0.020	31.607	0.065	21.0	0.6	NGC1316
12691	J032337.3-354642	FCC26	0.830	0.025	28.974	0.055	31.491	0.139	19.9	1.3	ESO357-25
12825	J032602.2-325340	FCC43	1.154	0.007	29.283	0.039	31.483	0.073	19.8	0.7	ESO358-01
12848	J032632.2-354249	FCC47	1.298	0.013	29.271	0.040	31.314	0.075	18.3	0.6	NGC1336
12878	J032718.0-343135	FCC55	1.248	0.008	29.492	0.051	31.598	0.080	20.9	0.8	ESO358-06
12917	J032806.6-321710	FCC63	1.373	0.029	29.548	0.019	31.470	0.083	19.7	0.8	NGC1339
12923	J032819.6-310405	NGC1340	1.314	0.007	29.583	0.028	31.603	0.068	20.9	0.7	NGC1344
13028	J033035.0-345114	FCC83	1.363	0.017	29.482	0.020	31.422	0.071	19.2	0.6	NGC1351
13058	J033108.2-361724	FCC90	1.013	0.047	29.126	0.144	31.443	0.193	19.4	1.7	
13084	J033124.8-351952	FCC95	1.262	0.013	29.385	0.037	31.475	0.073	19.7	0.7	
13097	J033147.6-350305	FCC100	1.105	0.011	29.324	0.048	31.566	0.078	20.6	0.7	
13146	J033247.6-341419	FCC106	1.186	0.017	29.320	0.025	31.491	0.068	19.9	0.6	
13177	J033333.9-333424	FCC119	1.182	0.018	29.363	0.077	31.538	0.100	20.3	0.9	
13230	J033429.5-353247	FCC136	1.218	0.020	29.248	0.038	31.387	0.075	18.9	0.7	
13252	J033459.2-351016	FCC143	1.273	0.035	29.350	0.041	31.427	0.086	19.3	0.8	NGC1373
13266	J033516.8-351556	FCC148	1.225	0.009	29.367	0.037	31.499	0.072	19.9	0.7	NGC1375
13267	J033516.6-351335	FCC147	1.376	0.014	29.543	0.023	31.459	0.070	19.6	0.6	NGC1374
13277	J033531.0-342650	FCC153	1.262	0.009	29.498	0.034	31.588	0.071	20.8	0.7	ESO358-26
13281	J033533.1-322754	FCC152	1.125	0.011	29.130	0.021	31.355	0.065	18.7	0.6	ESO358-25
13318	J033627.6-345834	FCC167	1.394	0.019	29.750	0.021	31.632	0.075	21.2	0.7	NGC1380
13321	J033631.7-351743	FCC170	1.376	0.019	29.790	0.028	31.706	0.076	21.9	0.8	NGC1381
13335	J033647.5-344423	FCC177	1.257	0.009	29.412	0.019	31.509	0.065	20.0	0.6	NGC1380A

Delimiter for download:
○ XML (VOTable) ○ comma ◉ pipe ○ tab ○ space ○ fixed format

Download
Download rows 1 to 200

Researchers not interested in curation formalities

Ingest workflow

Larger IT/dev role for curators

Assessing long-term value

Intellectual value of a database is in ad hoc combinations of data from multiple tables (joins and selections)

How to preserve?

file-centric preservation

Export the database contents to files for preservation and future access

AIP: database files, file system, application code, documentation

Native dump formats: sql, csv, xml

Software Independent Archiving of Relational Databases (SIARD)

Database Preservation Toolkit

RDF

http://www.bar.admin.ch/dienstleistungen/00823/00825/index.html?lang=en
http://keeps.github.io/db-preservation-toolkit/
Stefanova, Silvia, and Tore Risch. "Scalable Long-Term Preservation of Relational Data through SPARQL Queries." Semantic Web Journal, 2012. http://www.semantic-web-journal.net/content/scalable-long-term-preservation-relational-data-through-sparql-queries

What is the DIP?

Disseminate files

Reconstruct application

Alternative query system

system-centric preservation

Provide continuing access to the system in a way that satisfies the needs and expectations of the designated community

Emulation: preserve executability of original application software

Risk: emulator support and maintenance

System evolution: maintain access to working application as preservation (not necessarily original software)

Risk: can be resource-intensive

File-centric or system-centric?

Not either/or

Demand for access

Value of data

Resources available

Fidelity of representation

Access to versions

Metadata granularity and sources

Thank you

UNIVERSITY
LIBRARIES

Karl Nilsen
SAA 2014
2014-08-15
knilsen@umd.edu