



Using DataSpace Archives to Support Long Term Stewardship of Remote and Distributed Data

Robert Grossman, Steve Eick, David Hanley,
Xinwei Hong & Parthasarathy Krishnaswamy

www.ncdm.uic.edu

NASA/IEEE MSST 2004

12th NASA Goddard/21st IEEE Conference on
Mass Storage Systems & Technologies

The Inn and Conference Center

University of Maryland University College

Adelphi MD USA

April 13-16, 2004



DataSpace Archives

- ❑ Use web services' WSDL and UDDI for *discovery* of data
- ❑ Use Data Set Icons and DataSpace services for *exploration* of remote data
- ❑ Use XML/SOAP for *transporting small data* and metadata
- ❑ Use specialized high performance web services (SOAP+) for *transporting large data*
- ❑ Use DataSpace distributed keys for *integrating* data
- ❑ Provide *long term storage* of data using Internet Backplane Protocol (IBP)

Example: UCI KDD Archives

DataSpace - Microsoft Internet Explorer

File Edit View Favorites Tools Help

Address <http://ncdm144.lac.uic.edu/ucibrowser-dev/index.php?sPage=uci> Go Links Norton Antivirus

Google Search Web Blocking popups Options

Software & Standards

Down Load

Testbeds

DSTP

SABUL

Papers

News

FAQ

Staff

LAC

NCDM

DMG

Open UCIKDD
Preliminary Browser

[Anon_MSweb_Attrib_Data](#)
(Size: 0.0112 MB)

[Anon_MSweb_Attrib_Test](#)
(Size: 0.0112 MB)

[Anon_MSweb_Data](#)
(Size: 2.8225 MB)

Open UCIKDD
Preliminary Browser

[Anon_MSweb_Test](#)
(Size: 0.43461 MB)

[Australian_Sign_Language_Data](#)
(Size: 58.954 MB)

[COIL_Analysis_Data](#)
(Size: 0.0282 MB)

Open UCIKDD
Preliminary Browser

[COIL_Eval_Data](#)
(Size: 0.0121 MB)

[COIL_Results_Data](#)
(Size: 0.0197 MB)

Open UCIKDD
Preliminary Browser

[Census_Income_Data](#)
(Size: 99.0624 MB)

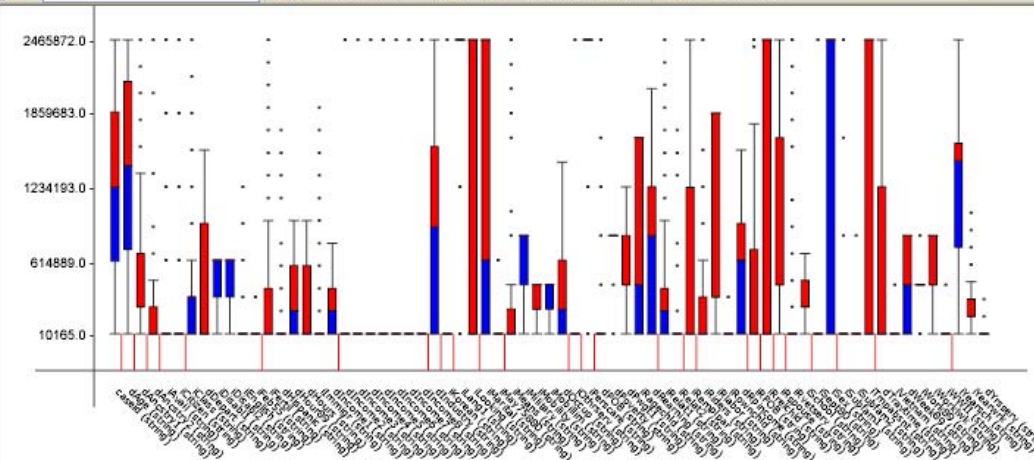
Internet

start _ODG Inbox - Micro... Microsoft PowerPoint ... DataSpace - Microsof... 11:14 AM

Example: UCI KDD Census Data Set

OpenDMIX - Metadata Browser - Microsoft Internet Explorer

Address: <http://ncdm144.lac.uic.edu/ucibrowser/out/MetaBrowser.html>



Name : caseid
 Max Value : 2465872.0 -
 Q3 : 1859683.0 -
 Median : 1234193.0 -
 Q1 : 614889.0 -
 Min Value : 10165.0 -

Overall Statistics

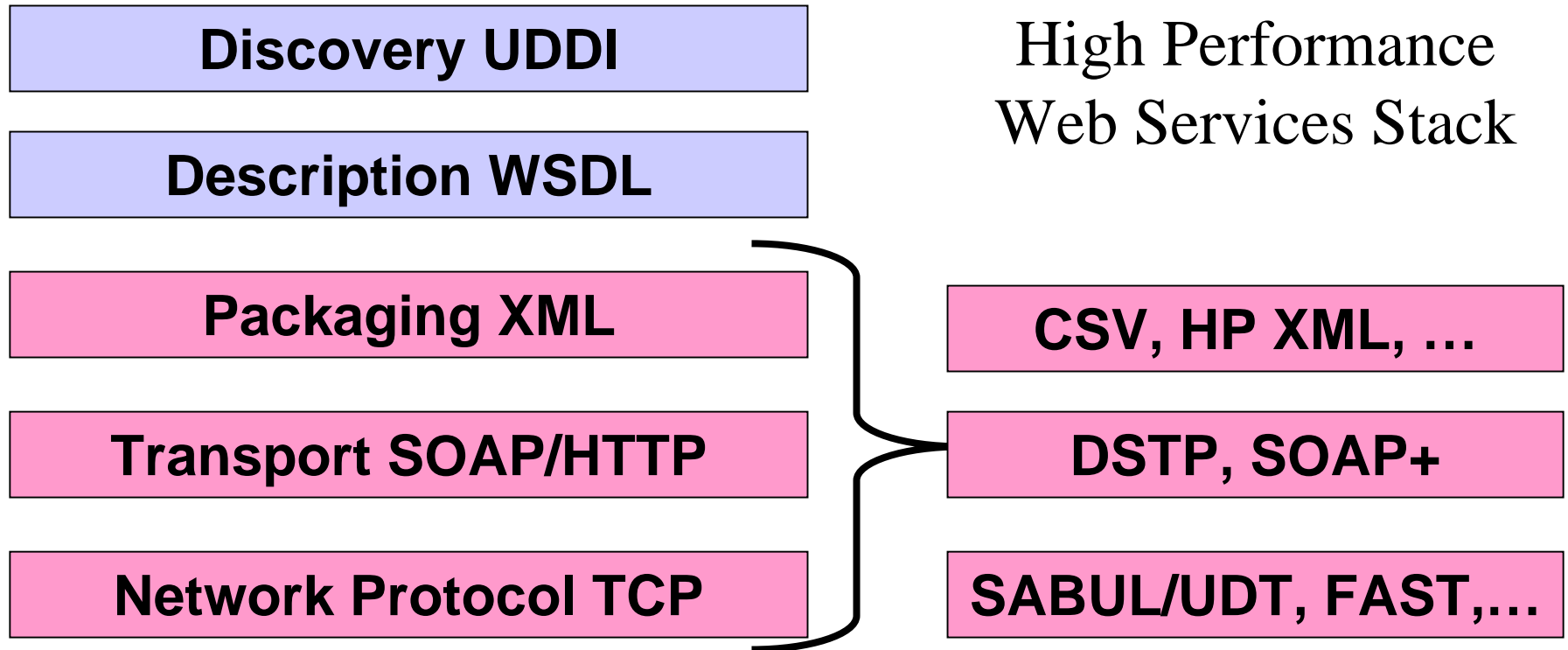
Attributes : 69
 Rows : 2,458,286
 Size(MB) : 344.6047

Most Frequent Values

caseid			dAge			dAncstry1			dAncstry2			iAvail			iCitizen		
Value	Count	Percent	Value	Count	Percent	Value	Count	Percent	Value	Count	Percent	Value	Count	Percent	Value	Count	Percent
802489.000	1	0.10	1.000	188	18.27	1.000	578	56.17	1.000	712	69.19	0.000	1005	97.67	0.000	957	93.00
805469.000	1	0.10	4.000	165	16.03	11.000	183	17.78	2.000	260	25.27	4.000	20	1.94	4.000	38	3.69
807858.000	1	0.10	3.000	155	15.06	0.000	100	9.72	3.000	30	2.92	3.000	2	0.19	3.000	24	2.33

start | kdd04 | OpenDMIX ... | UIC 02 - Mi... | notGNU/32... | Microsoft P... | 5:45 AM

High Performance Web Services (SOAP+)



Data Transfer using XML/SOAP vs. XML/SOAP+

Number Records	XML/SOAP (sec)	SOAP+ (sec)
10,000	0.65	0.21
50,000	177	0.72
150,000	673	125
375,000	3078	301
1,000,000	21121	823