CFD Data Sets on the WWW for Education and Testing

Al Globus[1]
Computer Sciences Corporation

Abstract

The Numerical Aerodynamic Simulation (NAS) Systems Division at NASA Ames Research Center has begun the development of a Computational Fluid Dynamics (CFD) data set archive on the World Wide Web (WWW) at URL http://www.nas.nasa.gov/NAS/DataSets/. Data sets are integrated with related information such as research papers, metadata, visualizations, etc. In this paper, four classes of users are identified and discussed: students, visualization developers, CFD practitioners, and management. Bandwidth and security issues are briefly reviewed and the status of the archive as of May 1995 is examined. Routine network distribution of data sets is likely to have profound implications for the conduct of science. The exact nature of these changes is subject to speculation, but the ability for anyone to examine the data, in *addition* to the investigator's analysis, may well play an important role in the future.

Introduction

Current scientific publications do not include the raw data from CFD experiments because these data are too large and paper is only marginally machine readable. New forms of electronic publication, such as the WWW [Berners92], limit publication size only by available disk space and data transmission bandwidth, both of which are improving rapidly. It is now possible to go beyond publication of the author's analysis of a CFD experiment to provide the raw data, as many full color images as desired, animations, related wind-tunnel data, and even scripts to control visualization systems so a user can examine the data interactively with the author supplying a starting point [Clucas94].

The WWW is much like a library on the internet. To understand the WWW, consider the following: people often organized information by writing and/or drawing on 3x5 cards and collecting related cards into a shoe box. The cards can be numbered so that relationships between cards can be expressed by writing, for example, "see card 42." Now imagine spreading electronic cards, called pages, around the internet (which consists millions of interconnected computers). This is the WWW. The 'shoe boxes' (computers) handing out copies of pages are called servers. The computers and software requesting pages are called clients or browsers. Each page has a unique address called its URL (universal resource locator). URLs allow any page to point to any other page in the WWW. It is important to note that any computer on the internet can be a server. Thus, thousands of people are independently and simultaneously expanding the WWW even as you read these words -- with no central control.

1. This work is supported through NASA contract NAS2-12961.

The WWW shares many characteristics with the scientific literature. To a first approximation, the literature consists of journal issues containing papers. Each paper contains text and graphics along with references to other papers. Similarly, the WWW consists of web sites containing pages. Each HTML (hyper-text markup language) page may contain text and graphics along with pointers to other pages. Other pages may be HTML, PostScript, images, digital video, digital sound, computer programs [Gosling95], and data sets. Finally, the scientific literature and the WWW are both constantly expanding.

Simply placing data sets on the web is of marginal value. To be useful, the data's format and method of collection must be well documented. Many CFD studies lead to papers and these should be readily accessible, preferably with a click of the mouse. Script files for visualization systems (e.g. FAST [Bancroft90], PLOT3D [Buning92], AVS [Upson89]) should be provided to create visualizations on the user's local workstation that elucidate the key points of a numerical study. Users can then interact with the data starting where the investigator left off. All this presupposes user access to visualization systems. In some cases, particularly with unsteady data, servers should isolate the data of interest within the data set (e.g., pressure on a crucial cut plane) and let client software handle graphic representation (e.g., choice of color map). This has yet to be implemented; see [Globus95] for a related discussion.

There are at least four classes of data set users: students, visualization developers, CFD practitioners, and management. Students (of any age) can use the data to study fluid physics, compare results from different solvers, learn about meshing techniques, etc. leading to better informed individuals. For these users it is particularly important that visualization be integrated into data set archives. Visualization developers can use the data sets to test algorithms and techniques, leading to better visualization systems. CFD practitioners can use data set archives to better understand previous work, 'mine' solutions for new information (have you ever read a paper and thought "I wonder what the helicity density looks like?"), compare new techniques with older results, collaborate with remote colleagues, and for validation. Management can use the data to understand the reality behind the viewgraphs. All users will benefit from fast, easy, convenient, and more-or-less anonymous access to CFD data sets.

That access is more or less anonymous (WWW servers can tell which machine access comes from, but not which user) raises security issues. Several levels of security are possible. First, an access file can be used to limit access to data in a directory to specific machines (e.g., mymachine.nas.nasa.gov) or classes of machines (e.g. all machines in the U.S.). A knowledgeable, determined cracker can get around this. Second, new 'secure' WWW servers and browsers are coming on the market that encrypt communications. These are as secure as the encryption scheme used. In the last analysis, of course, classified and the most sensitive data simply should not be on public networks. Happily, there is no particular reason to restrict access to many data sets once the author has published.

Speed and ease of access for data sets are a function of the network bandwidth between the server storing the data and the user’s client computer. The largest data set currently available in the NAS archive (an impinging jet) exceeds seven gigabytes. This would take approximately 22.5 days to download using a 28.8K baud modem, or 1.5 hours on a very efficient ethernet. However, billions of dollars are being spent to deliver interactive TV to homes, classrooms, and businesses. The high bandwidth necessary for interactive TV will make data transfers in the multiple gigabyte range quite rapid. As data are downloaded they must be stored, but with one gigabyte disks costing $200-300 and dropping the future looks bright.

The NAS has begun building an archive of CFD and other aerospace data sets on the WWW. The URL is http://www.nas.nasa.gov/NAS/DataSets/. At present there are only a handful of data sets on line, but all NAS users (about 400 projects this year) are asked to submit non-sensitive and non-classified data sets generated on the NAS super-computers to the archive. In time, we hope to provide a rich source of data to the nation and the world. The rest of this paper describes the technology in use and some of the lessons learned so far: the user interface icon, diversity of data sets, special treatment for unsteady data sets, and application of feature extraction algorithms.

Technology

Since the archive is expected to have a long life, only the simplest and most standard technology available to do the job is used. All data and other information are stored as files in a UNIX file system. Data and visualization scripts are stored in the format provided by data providers. All other documents use standard WWW formats: HTML, GIF, JPEG, etc. HTML documents are used to 'glue' all the other documents together in a hyper-media web.

We currently use the NCSA WWW server software, which is freely available over the net for noncommercial use. This could be replaced by any WWW compliant server. Server hardware is an SGI workstation with 12 gigabytes of disk space. Since the data sets would quickly fill these disks, part of the NAS mass storage disk system is mounted on the SGI. This provides access to 100 gigabytes of disk space backed up by many terabytes of tape robot storage. Seldom used files are automatically migrated to tape and back by the robots and the NAStore software [Tweten90]. Thus, when users access seldom used data sets there may be short delays (about a minute at present).

User Interface Icon

When a user scans a list of data sets, they would like to see a great deal of information in a small space. We have developed a data set icon that incorporates substantial amounts of information about a given data set using very little vertical space in an HTML page. The icon uses HTML tables, so only the most up-to-date browsers can display the icon. Table 1 is a particular data set icon from the NAS archive. The fields can be hyper-text

Table 1: Pulliam Impinging Jet [Pulliam95]

Dr. T. Pulliam, Ph.D.	Data	unsteady	alpha: 0 degrees
See Also	PLOT3D, multi-zone, iblank	periodic	Re: 3 x 10^5
May, 1995	Software	Navier-Stokes	
Fluids	Visualization	mach-number: 0.4	

links to additional information. Table 2 is a generic data set icon with explanations of the fields.

Table 2: Title (click for abstract with pointer to paper)

Author	Data	Property 1[a]	Property 5	...
See Also[b]	Data Format	Property 2	Property 6: value	...
Date	Software[c]	Property 3: value[d]	Property 7	...
Discipline	Visualization[e]	Property 4	Property 8: value	...

a. These refer to properties of the solution.
b. This links to pointers to related data sets; e.g., wind-tunnel results.
c. This links to information about grid generators and solvers used.
d. Some properties may have values associated with them.
e. This links to image galleries, animations, and visualization scripts. There is also text describing appropriate visualization techniques for some of the data sets.

Diversity of Data

At this early stage it does not seem appropriate to attempt to enforce a data format. Therefor, contributors are free to submit data in any format. As might be expected, even though only a handful of data sets have been archived and all are from the same facility, commonality is not as great as could be desired. Most of the data are in various forms of the PLOT3D format, but there is one unstructured data set in a format determined by the number of processors used to generate the solution.

Not only are data formats diverse, but visualization techniques are as well. The archive already contains PLOT3D and FAST scripts and one VISUAL3 [Haimes91] executable. A plan is under consideration to encourage commercial visualization vendors to provide command scripts so that their products can be used to examine data in the archive.

At some point it may be appropriate to standardize the data format or choose a database management system for this archive. At that time it will be necessary to convert the existing data. Unfortunately, there is currently no generally accepted data format to standardize on, so we can only document the formats used. If we were to try to standardize the data format, we would risk reducing the number of data sets submitted. At this time, we believe that expanding the contents of the archive is more important than format standardization. One can always convert a well documented data set when a good standard emerges, but there must be data to convert. If adding data to the archive is too difficult, scientists will not do it.

Unsteady Data

Unsteady data sets, with dozens, hundreds, or even thousands of time steps, present a unique challenge. The data sets are too large to provide as a single file (the largest data set currently in the archive has approximately seven gigabytes of data in 400 time steps, each in a separate file). Therefor, access to the data is provided in two forms. First, each file has a hyper-text link. This is adequate when a user only wants a few files, but is cumbersome when accessing hundreds of time steps. Second, HTML forms are used to specify a set of files to deliver. The files specified are archived together using the UNIX tar utility and sent to the user. Data compression techniques are under consideration.

Feature Extraction Algorithms

Many flows are characterized by a set of features; for example, vortices, shocks, and separations. Extracting these features is often a batch operation. Therefor, in most cases where a feature extractor is available, the extractor is run and the results made available as part of the archive. At present, only the vortex core extractor developed at MIT [Sujudi95] is used. Other feature extractors are welcome.

Conclusions

We are moving into an era of research publication dominated by digital networks. This opens many possibilities. Among these is the chance to transcend the limitations of paper and make available the full breadth of data generated in the course of CFD research. In particular, the data generated by numerical simulations can be easily made available to anyone anywhere in the world who is interested. The impact of such availability on research is difficult to predict, but is almost certain to be profound.

Acknowledgments

I would like to thank Kris Miceli, David Lane, and Bonnie Klein for reviewing this paper.

References

[Bancroft90] G. Bancroft, F. Merritt, T. Plessel, P. Kelaita, R. McCabe, and A. Globus, “FAST: A Multi-Processing Environment for Visualization of CFD,” *Proceedings Visualization ‘90*, IEEE Computer Society/ACM SIGGRAPH, San Francisco, 1990.

[Berners92] T. J. Berners-Lee, R. Cailliau, J. F. Groff, B. Pollermann, CERN, "World Wide Web: The Information Universe," *Electronic Networking: Research, Applications and Policy*, Vol. 2 No 1, Spring 1992, pp. 52-58, Meckler Publishing, Westport, CT, USA.

[Buning92] P. P. Walatka, P. G. Buning, *PLOT3D User’s Manual,* NASA Technical Memorandum 101067, NASA Ames Research Center. See URL http://www.nas.nasa.gov/NAS/Docs/uguides/Visual/html/vis3.html#REF50225.

[Clucas94] J. Clucas, V. Watson, "Interactive Visualization Of Computational Fluid Dynamics Using Mosaic," *Second International WWW Conference,* Chicago, IL, Oct 1994. URL http://www.nas.nasa.gov/NAS/FAST/FASTtreks/paper.html.

[Globus95] A. Globus, "A Software Model for Visualization of Large Unsteady 3-D CFD Results," AIAA 95-0115, *33rd Aerospace Sciences Meeting and Exhibit*, January 9-12, 1995, Reno, NV. URL http://www.nas.nasa.gov/NAS/TechReports/RNRreports/aglobus/RNR-92-031/alUnsteady.html.

[Gosling95] J. Gosling, URL http://java.sun.com/.

[Haimes91] R. Haimes and M. Giles, "Visual3: Interactive Unsteady Unstructured 3D Visualization," AIAA Paper 91-0794, *29th Aerospace Sciences Meeting and Exhibit*, Reno, NV, Jan. 1991. See URL http://raphael.mit.edu/visual3/visual3.html.

[Pulliam95] T. Pulliam, personal communication. See URL http://www.nas.nasa.gov/NAS/DataSets/MassStorage/Pulliam.1/index.html.

[Sujudi95] D. Sujudi and R. Haimes, "Identification of Swirling Flow in 3-D Vector Fields," AIAA Paper 95-1715, San Diego, CA, June 1995. URL http://raphael.mit.edu/pv3/AIAA-95-1715.ps.

[Tweten90] D. Tweten, "Hiding mass storage under UNIX: NASA's MSS-II architecture.," in Digest of Papers, pages 140-145. *Proc. Tenth IEEE Symposium on Mass Storage Systems*, May 1990, Monterey, CA.

[Upson89] C. Upson, T. Faulhaber, D. Kamins, D. Laidlaw, D. Schlegel, J. Vroom, R. Gurwitz, A. van Dam, “The Application Visualization System: A Computational Environment for Scientific Visualization,” *IEEE Computer Graphics and Applications*, July 1989, pp. 30-41.