On-line Atomic Data Access

David R. Schultz

Physics Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831-6373

Jeffrey K. Nash

Lawrence Livermore National Laboratory, Livermore, CA 94551

The need for atomic data is one which continues to expand in a wide variety of applications including fusion energy, astrophysics, laser-produced plasma research, and plasma processing. Modern computer database and communications technology enables this data to be placed on-line and obtained by users over the INTERNET. Presented here is a summary of the observations and conclusions regarding such on-line atomic data access derived from a forum held at the *Tenth APS Topical Conference on Atomic Processes in Plasmas.*

I. INTRODUCTION

The use of data for a wide variety of applications has gone on practically since the earliest beginings of atomic physics. In particular, present applications such as interpreting astrophysical environments (e.g. solar and stellar flares, supernovae, and the interstellar medium), developing fusion energy reactors (e.g. plasma composition and transport diagnostics, divertor chamber modeling), laser-produced plasma research (e.g. temperature and density modeling, x-ray production diagnosis), and the processing of material by plasmas (e.g. etching and deposition of semiconductor surfaces), require large databases of atomic collisional and spectroscopic data. Major new applied research projects are on the horizon, such as new laser facilities (e.g. the National Ignition Facility), the International Thermonuclear Experimental Reactor (ITER), various space missions (e.g. AXAF, SOHO), and many industrial programs utilizing atomic processes to aid in manufacturing.

With the advent of the INTERNET, the linking of computers world wide has recently provided the means to rapidly exchange data. Furthermore, the World Wide Web (WWW or simply the 'Web'), adoption of a standard for hypertext communication (HTML - *H*yper*T*ext *M*arkup *L*anguage), and the widespread availability of so-called Web Browsers such as Mosaic and Netscape has allowed a very easy to use, yet powerful way for non-expert users to search for and retrieve data in a variety of formats. Thus, as a community of atomic data producers, collectors, and users, we have an unprecedented opportunity to preserve and make readily accessible needed atomic data.

II. FORUM

During the *Tenth APS Topical Conference on Atomic Processes in Plasmas* a one evening forum was held in order to

- Initiate discussion of issues relating to this opportunity to place atomic data on-line,
- Identify the consitutencies interested in this activity and discuss their roles, and
- Help to define some of the problems, open questions, and parameters for undertaking this endeavor.

The speakers included Douglass Post (ITER), Peter Smith (Harvard-Smithsonian CfA), Norman Bardsley (LLNL), Richard Lee (LLNL), Takako Kato (NIFS), Peter Mohr (NIST), and the present authors. Each speaker tried to address each of these three objectives especially as pertains to their own speciality. For example, Bardsley related his experience with forming an APS Forum on Low Temperature Plasma Science, and in particular the interaction of industry and the atomic physics community in linking producers and users of data in the semiconductor industry. Also, a number of the speakers who represent data centers which already have data available over the WWW provided demonstrations of these sites at the forum and, in fact, throughout the conference. Yuri Ralchenko (Weizmann Institute) has surveyed a large portion of the atomic data sites on the Web and has listed their WWW Uniform Resource Locator (URL) addresses along with a brief summary of their content. This manuscript in PostScript format may be viewed using a WWW browser helper such as "xdvi" at http://plasma-gate.weizmann.ac.il/~fnralch/app.dvi.

As an example of how to reach these resources, and how they are becoming linked, you can start by using a Web browser to connect to the Weizmann Institute site (http://plasma-gate.weizmann.ac.il) and from there navigate by following the hypertext links to a number of other sites. Similarly, one could begin at the home page for the Controlled Fusion Atomic Data Center at http://www-cfadc.phy.ornl.gov/ and follow the links page to the Weizmann Institute WWW page, or other atomic data sites such as the NIST Atomic Spectroscopic Database (http://aeldata.phy.nist.gov/nist_beta.html). Here we make no attempt to summarize the details of each speaker's presentation or the content of their Web site, rather we seek only to provide a synthesis of the ideas exchanged. Thus, the responsibility for the conclusions drawn here and any statements of fact rests with the present authors rather than the speakers.

III. GOALS AND METHODS FOR ON-LINE ACCESS

Storing atomic data electronically has many benefits. For example, data printed graphically or in tabular form in journals presents the end-user of that data with the need to locate the appropriate reference, obtain the particular volume, and to digitize or scan and recognize the graphs or tables, thereby converting them to a format usable in his application. Simply finding the appropriate reference is a step full of difficulty for a user who wishes not simply to find all references on a particular reaction but rather find a recommended cross section or rate coefficient valid over a wide range of energy. Another advantage of electronic storage of atomic data is the fact that many data sets can be stored and categorized in order that an expert can evaluate the data and make a recommendation regarding it. Finally, since the data, and hopefully the recommended data, are stored electronically, they can easily be transmitted to other sites over the INTERNET.

Thus, electronic storage of data facilitates

- locating needed data,
- preservation of data in a readily accessible format,
- collection of available data sets in centralized locations,
- the ability to compare and evaluate data sets so that a recommended set of values can be determined, and
- the means to share the fundamental and recommended data with others over the INTERNET.

Recognizing these benefits, it is next important to consider the means by which these data will be stored and disseminated. Typically, each research group or data center will have their own specific platforms and database products which they have judged to be most suitable for their applications or user groups. For example, atomic data may be stored on personal computers, mainframes, or UNIX workstations. Furthermore, it may be stored in a wide variety of formats, such as ASCII tables, relational databases (such as Lotus, ORACLE or INGRES), or some other specialized format such as ALADDIN. ALADDIN, for example, has been adopted by the International Atomic Energy Agency as a standard format for exchanging atomic data within the fusion energy research community. Since it does not seem practical to dictate the format used by individual groups, what is necessary is some means through which data can be exchanged.

Owing to its widespread use and the relative ease with which applications can be developed for it, the WWW appears to be the best choice for on-line atomic data access. Some experience exists with locally developed data interfaces written in C or FORTRAN, but performance of these tools over the INTERNET on a wide range of different accessing platforms lacks flexibility and often simply does not work on terminals or computers that they do not recognize. In particular, even within the framework of the X/Motif standard windowing environment, screen movements or character sizing are not controlled in a sufficiently uniform way across the different vendor's platforms to allow this approach to find much success. On the other hand, standards for Web browsers and the hypertext language they recognize are such that practically every computer on the INTERNET now has these tools. They alleviate the need of the research group or data center to develop the full interface to their data. Almost any database product, or FORTRAN, C, or ASCII file structure can be interfaced with the WWW in a relatively straightforward manner. The complexity and utility of these interfaces can range from simple file lists which allow ftp'ing of the data files, to graphical displays in GIF or PostScript format. Search engines are also easy to write so that facilities such as on-line bibliographies can be implemented. Thus, the WWW is the presently preferred medium for data exchange, and allows for the freedom of each group to decide on internal storage formats and data management products.

IV. ISSUES

A number of issues relating to the efforts to place data on-line were also identified at the forum. We comment on the following points in some detail:

- unrestricted data access,
- attribution of authorship,
- proper and effective data usage, and
- standardized data formats.

For example, it was the view of the large majority of the participants that access to atomic data should be open, on-line, and non-commercial. That is, data should be placed into the public domain rather than sold or distributed only to narrow interest groups. This reflects the general consensus of the scientific community regarding publication of their work in the open literature. Also, if data could be obtained on-line, rather than through, say receiving diskettes in the mail, or even worse in paper copies of tables or graphs, this would be the best way to efficiently disseminate data. On-line repositories of data would also allow a potential user to browse the available data, selecting for downloading only what they decided they needed after viewing it. Experience with bibliographic search requests often shows that the initial request of the users does not yield exactly what they are looking for, and that subsequent iterations are needed to refine the parameters of the search. Therefore, using mediums such as facsimile, electronic mail, or postal mail to transmit search results, sufficiently slows the overall process so as to compromise the utility of the search. Finally, it was generally agreed that users should not be required to purchase specific commercial products simply to obtain data. For example, only standard, commonly available, free Web browsers should be required in order to gain access to data.

Another important issue, as yet not completely resolved, is one of proper attribution. If a user downloads data from a Web site, it was agreed that there should be some mechanism to assure that the originator of that data is properly credited. Often the measure of a scientist's performance is judged by the citations that his or her work receives, and if placing data on the Web circumvents normal procedures for obtaining that attribution, then data producers might be quite reluctant to release their work. In addition, without attribution, the vital contacts between data users and producers breaks down.

Related to this issue is the perceived need to create mechanisms which would help assure that data is properly used. That is, all the information that a user would require about a particular data set needs to be included in the downloaded file. For example, is the data one set or a number of theoretical and experimental results, or is it evaluated or recommended data? What is the range of validity for the data if it is a theoretical result? What are the experimental uncertainties of the data? What precisely are the units and, of course, what is the full citation for the work (i.e. where can all the details be found, and who to contact for full information since this could differ from the lead author on a citation or since the data may not be formally published)?

Perhaps the most difficult issue to address is that of encouraging standard formats for on-line data. Some groups have already agreed within their community to distribute data in particular formats so as to minimize the difficulty in deciphering exchanged data. For example, one group might always fit their data using Chebyshev polynomials, whereas another always lists it as pairs of columns of raw numbers. The only way to resolve these issues seems to be by engaging in further discussions in future workshops and forums. Certainly, within a particular community, such as astrophysics, it might be possible to agree on some standards for atomic data exchange. As mentioned above, this has been done within the fusion community, but it is widely accepted that the current choice of an exchange format is not at all optimal.

Finally, it was emphasized that no single data access system seems appropriate for all needs. For example, some users require extremely large sets of data which could only be obtained by running large production-oriented codes, whereas other users require only a narrow range of data which must be of the highest accuracy possible. Thus, one user may seek to first remotely run a production code, then download hundreds of megabytes of spectroscopic line energies and transition probabilities, while another may want only a single number or bibliographic reference. Data repositories should also, ideally, allow for the interactive browsing and intercomparison of available datasets to facilitate choosing appropriate data. A distinction should also be made between facilities giving access to data archives versus those which include certain modeling or computing capabilities.

V. CONSTITUENCIES AND ROLES

The forum also helped to identify key constituencies each having specific roles to play in the quest to bring atomic data on-line. The most obvious identification sees individuals as either producers, users, or archivers of atomic data. Clearly, communication between users and producers should be improved so that producers know what users want, and users know what is available from producers. The Web sites which act as repositories of atomic data can facilitate this communication by posting information about data needs, for example. In fact, data centers play a very important role in that their missions most closely fit that goal of placing data on-line.

However, funding for data centers is particular difficult to come by. One of the diverse priorities of the communities wishing to either place data on-line or use it, should be to help increase the awareness of funding agencies of the importance of data centers. Such repositories form a kind of "community memory" by storing data. If allowed to shut-down, the hard work of collecting, managing, evaluating, and disseminating data is lost. In addition, traditional data center activities have focussed on collecting data sets which each individually consist typically of only a modest number of values, but often, researchers require very large, comprehensive data sets at whatever level of accuracy is possible. This highlights the prevailing focus on critically evaluated data, which might not best serve all users.

Related to this means of preserving information in data centers is the role that journals can play in on-line atomic data access. All journals should eventually be placed on-line, so as to realize the goal that the data contained in the articles can be electronically accessed. Much of the development along these lines also depends on the economies of electronic publishing. Archival journals such as Atomic Data and Nuclear Data Tables and the Journal of Physical and Chemical Reference Data are also important as repositories of needed data, often recommended or "best available" data, and should also be placed on-line.

Much of the business of coordinating data production, collection, and dissemination can be aided by networks and organizations such as the IAEA Atomic and Molecular Data Center Network, the Committee on Data for Science and Technology (CODATA), and the Atomic Data and Analysis Structure consortium. These networks are becoming quite aware of the need for on-line data access and are actively encouraging it. Similarly, meetings and workshops can provide forums for the discussion of issues relating to on-line data access. In particular, the interests present at the Atomic Processes in Plasmas Conference and the Gaseous Electronic Conference include particularly those influenced by these issues. It was noted that further workshops are already being organized specifically aimed at improving on-line atomic data access.

VI. CONCLUSIONS

Thus, the forum on on-line atomic data access showed that already much has been done to convert the interfaces to existing data archives to be compatible with the WWW, but that a number of key issues remain to be resolved. In particular, further coordination is needed to facilitate communicating data needs to producers, and the electronic distribution of results. Future workshops and conferences should also address issues regarding standardization of public data formats to ease exchange of data. It was noted that the focus on short-term results by both private and public sector funders is causing a great deal of pressure on data centers and individual data producers which are the source of the data needed for on-line access to be achieved. Particular difficulties which the community should seek to remediate include methods for easily locating existing data matching a user's needs, ways in which users can solicit new calculations and experiments, procedures for obtaining evaluated or recommended data in addition to or rather than primitive data, methods for assuring that data is used correctly and effectively, and that attribution is given to the originators of the data which is retrieved on-line.

ACKNOWLEDGMENTS

We wish to thank the speakers and the audience at the forum for their interest and contributions to the discussions. We are also grateful to the organizers of the *Tenth APS Topical Conference on Atomic Processes in Plasmas* for the opportunity to hold the forum and for their support of it. DRS wishes to express his thanks for support of the Controlled Fusion Atomic Data Center from the Office of Fusion Energy, US DOE through contract No. DE-AC05-96OR22464 with Oak Ridge National Laboratory, managed by Lockheed Martin Energy Research Corp. JKN acknowledges support provided under the auspices of the US Department of Energy by Lawrence Livermore National Laboratory under Contract No. W-7405-ENG-48.