https://doi.org/10.61092/iaea.9p7w-5msj

IAEA
International Atomic Energy Agency

INDC(NDS)- 0727
Distr. LP,NE,SK

INDC International Nuclear Data Committee

Developments in Data Exchange

Summary Report of an IAEA Consultants Meeting

IAEA Headquarters, Vienna, Austria

28-29 July 2016

Prepared by

Hyun-Kyung Chung and Bastiaan J. Braams

January 2017

IAEA Nuclear Data Section
Vienna International Centre, P.O. Box 100, 1400 Vienna, Austria

Selected INDC documents may be downloaded in electronic form from
http://www-nds.iaea.org/publications
or sent as an e-mail attachment.
Requests for hardcopy or e-mail transmittal should be directed to
NDS.Contact-Point@iaea.org
or to:
Nuclear Data Section
International Atomic Energy Agency
Vienna International Centre
PO Box 100
1400 Vienna
Austria

Printed by the IAEA in Austria

January 2017

INDC(NDS)- 0727
Distr. LP,NE,SK

Developments in Data Exchange

Summary Report of an IAEA Consultants Meeting

IAEA Headquarters, Vienna, Austria

28-29 July 2016

Prepared by

Hyun-Kyung Chung and Bastiaan J. Braams

Abstract

A Consultancy Meeting on Developments in Data exchange was held on 28 and 29 July 2016 with six external participants and IAEA staff. The meeting was called to help guide the database work of A+M Data Unit over the next 7 years or so. We see increased emphasis on very large datasets containing relatively unprocessed data all backed up by new technologies for data analysis; keywords here are machine learning, Gaussian processes and polynomial chaos. In the area of calculated A+M data this may lead us towards very different databases than we have now; for example storing full scattering matrices and not just integrated cross sections, and in addition storing enough calculated data to allow simulations to include uncertainty propagation. In the area of data for PMI we imagine developments towards providing detailed representations of material microstructure and parameterizations of interaction potentials and their uncertainties, and not just reflection coefficients and sputtering yields; and again we want to present the data in a form that permits further uncertainty propagation.

January 2017

TABLE OF CONTENTS

1. Introduction

A Consultants' Meeting (CM) on Developments in Data Exchange was held in the IAEA headquarters Vienna, from 28 to 29 July 2016. The objective of the 2-day meeting is to help guide the work of our Atomic and Molecular (A+M) Data Unit over the next 7 years or so.

There is increased emphasis on very large datasets containing relatively unprocessed data all backed up by new technologies for data analysis; keywords here are machine learning, Gaussian processes and polynomial chaos. In the area of calculated A+M data this may lead us towards very different databases than currently available databases; for example storing full scattering matrices and not just integrated cross sections, and in addition storing enough calculated data to allow simulations to include uncertainty propagation. In the area of data for plasma-material interaction (PMI) developments are foreseen towards providing detailed representations of material microstructure and parameterizations of interaction potentials and their uncertainties, and not just reflection coefficients and sputtering yields; and data should be in a form that permits further uncertainty propagation. This is considered as the next major development area for our A+M Data Unit.

For the 2-day meeting assembled was a group with a range of expertise: management of large datasets, machine learning, uncertainty assessment and uncertainty propagation (and how to handle data to allow uncertainty assessment), applications of databases in fusion plasma modelling, data integrity, and related matters. The invited participants are Dr Yuri Ralchenko from the NIST Atomic Spectroscopy Group (USA) responsible for the Atomic Spectra Database that is very important in our community, Dr Udo von Toussaint of IPP Garching (Germany), a specialist in Bayesian methods and in heterogeneous data for fusion), Dr Robert Harrison of Stony Brook University (USA) in the field of quantum chemistry and massively parallel computation, Dr Gábor Csányi of Cambridge University (UK) in the field of large datasets for parameterization of MD potentials; machine learning, and Dr Kerstin Lehnert of Columbia University (USA) in the field of geophysical data, data integration including bibliographical data, Dr Christian Hill of University College London (UK) responsible for molecular data and the Virtual Atomic and Molecular Data Centre and Dr Naohiko Otsuka of the IAEA Nuclear Data Section responsible for the EXFOR database of experimental nuclear reaction data. Participants discussed implications of present developments for A+M Data Unit's work on atomic, molecular and plasma-material interaction data for fusion applications

This report contains the proceedings of the meeting, conclusions and future work. The list of participants is provided in Appendix 1 and the meeting agenda in Appendix 2.

2. Proceedings of the Meeting

Arjan Koning, the head of Nuclear Data Section (NDS), welcomed the participants and shared an interest in statistics, big data and Baysean ideas of nuclear data science, which is the topic of the present meeting. He encouraged participants to come forward with new ideas for atomic and molecular data unit activities, which may become a common thing in 5 to 10 years.

Bas Braams, the head of Atomic and Molecular Data Unit (AMU) described the meeting objectives, described NDS activities in EXFOR database, RIPL and FENDL data libraries as well as AMU data activities such as ALADDIN, AMBDAS databases and GENIE search engine. Activities of AMD are quite diverse in order to serve the data needs of fusion applications and there are heterogeneous databases as well as a wide variety of data formats. An internationally evaluated and recommended data are needed but not achieved yet. Currently the uncertainty quantification (UQ) of AM data is a top priority in AMD activities.

The agenda was adopted and participants introduced themselves. The meeting proceeds with presentations of participants, as summarized in this section. Presentations are available at the Unit's website at https://www-amdis.iaea.org/meetings/DataExchange2016/.

Data exchange in the Earth sciences
Kerstin Lehnert

Interdisciplinary Earth Data Alliance (IEDA) is a community-based data facility funded by the US National Science Foundation (NSF) to support, sustain, and advance the geosciences by providing data services for observational solid earth data from the Ocean, Earth, and Polar Sciences. It promotes data exchange of broader earth science communities, transparency and reproducibility of data, use of data creating new science. It is important to give a credit for researcher producing data, which can lead to a funding. Publishers and journals, funders and researchers as well as repositories are involved in this activity. FAIR Guiding principles for scientific data management and stewardship promote data to be Findable, Accessible, Interoperable and Reusable. For this domain repositories will set up data curation standards for identification and persistence (findable), protection and protocols (accessible) and domain-specific data standard for context and provenance (re-usable) and harmonized and machine-readable format (integrated/interoperable).

IEAD facility covers a wide range of data ranging from data of geochemistry, marine geophysics, marine geology and geochronology data to sensor data and data from sample-based observations and experiments and various of sources of data such as field data, lab data, processed data, samples. It offers comprehensive services: trusted repository services, domain-specific data curation, science-specific user interfaces and software tools, data products, cross-disciplinary data access, programmatic, standards-based interfaces as well as user support and training. Recently, publishers, repositories and funders gather to agree on guidelines of data projects. TOP (transparency and openness promotion) guidelines are adopted by 538 journals. COPDESS.org (Coalition on Publishing Data in the Earth and Space Sciences) endorsed by most publishers and repositories in the Earth and space sciences. Joint Declaration of Data Citation Principles is endorsed by 109 organizations including most major publishers.

There are challenges of data integration due to heterogeneous data formats database schema, data types and metadata standards. EarthCube Geolink projects via Semantic Web (linked data) are launched to find and integrate resources across repositories. Data synthesis, data mining, machine learning for data analysis are big efforts and require a collaboration to learn from other communities and disciplines.

Reliable data from theory and computation – art vs science
Robert Harrison

Theory and computational chemistry, once subjects of basic research, are now central to many other fields of discovery, design, and engineering. To realize the full broad potential of theory and simulation from many areas of chemistry and physics we must develop new techniques to manage data and associated uncertainty, and radically change communal approaches to sharing and publication.

Modern computational chemistry historically driven by basic science is changing and materials and processes are designed with computers. A national strategic computing initiative (presidential) is launched with a vision of simulation rapidly transferring basic science and engineering knowledge and enabling new technologies. Challenges exist due to theoretical and computational gaps such as chemistry at interfaces, time and length scales, limitations of current density functionals, speed and accuracy of many-body methods, accurate treatment of excited states and scattering, coupling of electronic and nuclear motion as well as spectroscopy. They are all affected by missing rigorous uncertainty quantification (UQ), little data sharing and collaboration, explosion of data and lack of their utility and undefined workflow to capture data over the entire research life cycle.

Brookhaven National Laboratory (BNL) Center for DATA Driven Discovery (C3D) focuses BNL's diverse expertise in scientific data on the defining challenges, leverage multiple research facilities and BNL's international leadership and collaborations and integrates the entire laboratory into an instrument for scientific discovery. Its vision is to build application-driven integrated data solution center incorporating both measurement data and theory and modeling and constructing computer and data processing tools.

One should be aware that parallelism is now the only path to increased performance in this post Moore's law era. Science and computer science students should be prepared for the new trend.

Machine learning the Born-Oppenheimer potential energy surface
Gábor Csányi

Material simulations are affected by data of first principle calculations. Interatomic potentials (IP) used for molecular dynamics have different requirements depending on applications. For example, for biomolecular force fields (simulations of proteins) transferability is important while accuracy is important for small molecules in gas phase (vibrational spectra) and reactiveness for solid state materials (bond breaking). Ingredients of a potential are representation of atomic neighbourhood, interpolation of functions and database of configuration.

Traditional functional forms based on quantum mechanics are not the correct functions with limited accuracy and not systematic and the new potential should be given by quantum mechanics using a concept called SOAP (Smooth Overlap of Atomic Positions). The new way is to fit IP with kernels so called as Gaussian Process Regression, a probability approach. Kernel requirements for IP construction are physical symmetries (rotation, translation, permutation), continuity with respect to changes of neighbourhood , uniqueness, multiresolution which is refinable to precise atomic resolution and smoothness (deformation stability). One may construct a good kernel by hand or use simple descriptors that respect some symmetries and optimization.

Database of first principles calculations should be generated for Gaussian process potentials. However, as tools and technologies are not existent for data exchange as well as provenance is rather non-existent, data exchange has been problematic. Currently available are NoMaD (Novel materials Discovery) repository and openKIM (Knowledgebase of Interatomic Models). There are outstanding problems such as accuracy on database (or accuracy in properties), data contents such as region of validity or global error bounds, how to validate potentials.

Integrated data analysis and uncertainty quantification in fusion
Udo von Toussaint

The scientific progress in areas like magnetic confinement fusion depends on the joint evaluation of many different diagnostics, which not only probe different physics aspects but at the same time are also mutually dependent. The commonly applied iterative evaluation process hampers an assessment of the uncertainty of the obtained results. This can be overcome using an integrated and simultaneous evaluation scheme.

Examples for the integrated data analysis (IDA) approach at ASDEX Upgrade are presented. The incoorporation of uncertainty quantification (UQ) into the integrated tokamak modelling concept and the present challenges are discussed. For plasma-wall-interaction problems with a low- to medium number of parameters the polynomial chaos expansion method provides a suitable tool for efficient UQ. This is demonstrated with PWI-simulations based on the binary collision code SDTrim.

EXFOR: Exchange of Experimental Nuclear Reaction Data
Naohiko Otsuka

The IAEA Nuclear Data Unit organised the first meeting of Four Centres (Brookhaven, Saclay, Obninsk, Vienna) in 1966 to discuss international exchange of experimental neutron-induced reaction data. After several meetings, EXFOR (Exchange Format) was adopted as the official format for data exchange among the Four Centres in 1969, and data exchange with the EXFOR format was started in 1970. The format was then extended to accommodate charged-particle and photonuclear reaction data in addition to neutron-induced reaction data. Currently 13 centres from 8 countries and 2 international organisations (China, Hungary, India, Japan, Korea, Russia, Ukraine, USA, NEA, IAEA) are collaborating for data exchange as members of the International Network of Nuclear Reaction Data Centres (NRDC) [1].

When Four Centres started data exchange, they maintained their own formats (SCISRS in Brookhaven, NEAUDADA in Saclay, DASTAR in Obninsk and Vienna). Though it is still possible for data centres to define centre-specific storage formats [2], almost all centres now adopt EXFOR as the primary format. On the other hand, various database systems and online and offline retrieval systems are developed and maintained by centres for their customers.

The data exchange is currently performed by the following steps [3]:

1. The originating centre assembles EXFOR entries (1 EXFOR entry = 1 experimental work) in a single ASCII file ("tape"), and submit it to other centres ("preliminary transmission").
2. Other centres review the submitted EXFOR entries, and send comments to the originating centre within one month.
3. The originating centre revises the tape, and submits the finalized tape to other centres ("final transmission").
4.

Typically 70 to 80 tapes are submitted by data centres per year. In the past, data exchange was done on an NDS ftp server. However the ftp server is no longer available due to IT security issues. Now NDS receives tapes from originating centres by emails, and upload them to the website for distribution to data centres.

Figure 1 and 2 show size distributions of EXFOR entry and tape. Usually one EXFOR entry and tape are smaller than 1 Mb and 10 Mb, respectively, and they are reasonable sizes for email exchanges. There are a lot of blanks in EXFOR entries and tapes, and zipping reduces their sizes very effectively.

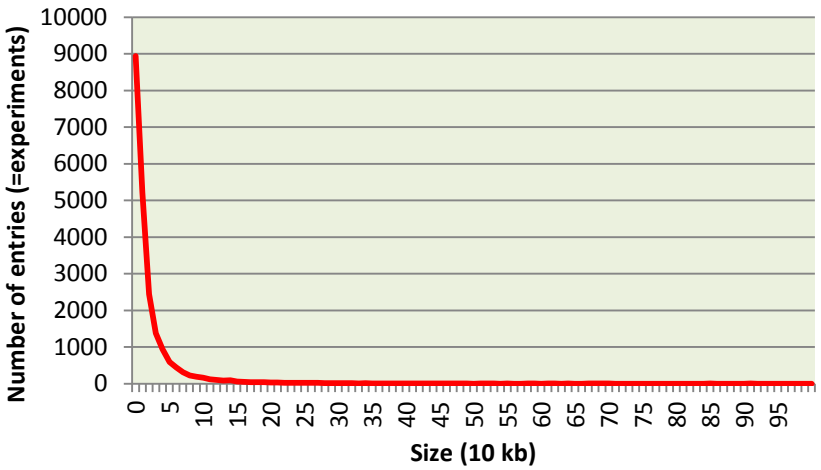

Fig. 1. Size distribution of one EXFOR entry (1 entry = 1 experimental work). Note that about 1% of EXFOR entries are larger than 1 Mb (The largest one is about 60 Mb).

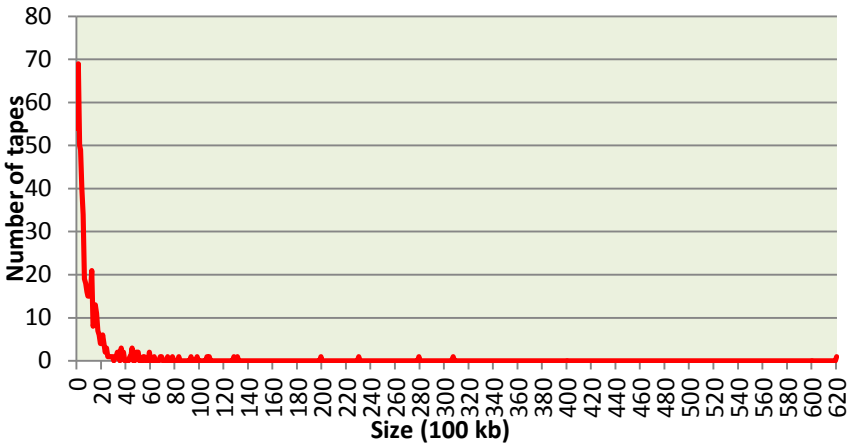

Fig. 2. Size distribution of one EXFOR tape.

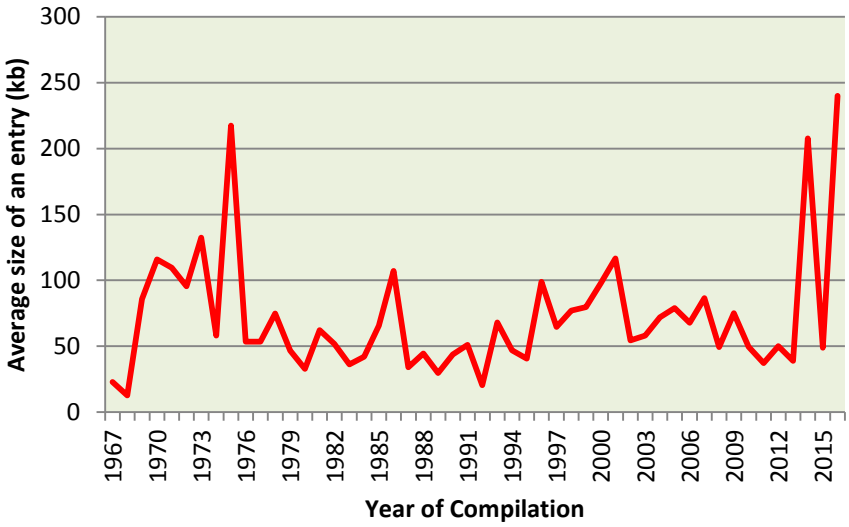

Fig. 3. Average size of one EXFOR entry in each year of compilation.

Figure 3 shows average size of EXFOR entries in each year of compilation. There is no clear time dependence except for 1975 and 2013 to 2016 where we can see sudden increases. This is due to systematic compilation of neutron time-of-flight spectra (transmissions and reaction yields) which are essential for evaluation of resonance parameters for neutron fission energy system application. They were not compiled systematically in the past because they are voluminous and also not always available from authors. However it is not an issue nowadays to exchange an EXFOR entry exceeding 10 Mb, and data centres are making significant effort to collect time-of-flight spectra in both old and new experiments in strong collaboration with experimental facilities like GELINA (EC-JRC IRMM), n_TOF (CERN) and ORELA (ORNL). Experimental and facility specific information is often required to derive physical quantities (*e.g.*, resonance parameters, cross sections) from time-of-flight spectra. For example, it is necessary to know the response function of the facility to establish the relation between time-of-flight and neutron energy. Sample description is also important to correct data measured with relatively thick samples to correct time-of-flight spectra for finite sample size effects (*e.g.*, self-shielding, multiple scattering). We organized a meeting in 2013 to identify such key information required for proper use of time-of-flight spectra compiled in EXFOR [4], and encourage experimentalists to submit their time-of-flight spectra with experimental information filled in the template agreed in the meeting.

References

[1] N. Otsuka et al., Nucl. Data Sheets 120 (2014) 272, and references therein.

[2] O. Schwerer (ed.), Report IAEA-NDS-207 Rev. 2015/08 (2015).
[3] N. Otsuka (ed.), Report IAEA-NDS-215 Rev. 2014/05 (2014).
[4] F. Gunsing, P. Schillebeeckx, V. Semkova (ed.), Report IAEA(NDS)-0647 (2013).

Data exchange in atomic physics; past and future
Yuri Ralchenko

An overview of the Atomic Data Program at National Institute of Standards and Technology (NIST) was presented with emphasis on the data exchange issues. Collection and critical evaluation of data is an important part of the NIST mission, and the atomic reference data program is active in all its components. The standard reference databases include critically evaluated spectroscopic data for all elements of the periodic table up to Z=110. These databases are connected with a number of international projects and thus a reliable exchange of atomic data is of high priority. The talk also presented a well-developed strategy to evaluation of uncertainties for basic atomic data.

Online database services for atomic and molecular data
Christian Hill

The following atomic and molecular data resources were presented: VAMDC standards, HITRAN, ExoMol, ALADDIN and Quantemol. Much work has been done in writing and promoting standards for the representation and distribution of spectroscopic and collisional data in the form of: transition line lists (HITRAN, ExoMol – with a huge number of lines); storage requirements and data structures. spectroscopic cross sectional data (HITRAN, ExoMol – with an online service for binning down high-resolution cross sections and for generating "k-coefficient tables") collisional cross sectional data (ALADDIN, Quantemol – with the ability to generate rate parameters for custom energy profiles, e.g. Arrhenius parameters from the Maxwell-Boltzmann distribution) generic parameterized data (Quantemol – Relational database structures for the representation of generic physical quantities with names, units, uncertainties, references, etc.).

It has been noted that it is not easy for users to adopt new standards and sustainability is crucial. It is important to provide software tools for interconversion between data formats that are actually used. Visualization is helpful for more use of data and provenance of data should be provided for better citation practice.

3. Discussions and Recommendations

Participants had a brainstorming discussion on current data issues such as big/large data, data management, and data exchange for atoms, molecules and solids and recommended future activities for the atomic and molecular data unit (AMU). Discussions and recommendations are summarized.

Data Exchange

The AMU has supported the development of a new XML-based standard for exchange of atomic, molecular, and particle-solid-interaction data (XSAMS) since 2003. It is recommended that most current version v1.0 of the XSAMS standard that has been adopted by VAMDC (Virtual Atomic and Molecular Data Consortium) should be formally ratified by the IAEA and the Schema and documentation placed on the AMDIS website. These are the documents currently at http://standards.vamdc.eu/ -- the VAMDC has not made (or at least not approved) any substantial changes to XSAMS since 2012, but is planning a meeting in January which may make recommendations for a further iteration of the Schema.

A common data format of XSAMS allows data sets of several databases to be compared at once as demonstrated by VAMDC services. It helps checking the quality and integrity of data sets and a visual

tool to compare data will be very helpful. It makes sense to automate visualized comparisons of a large number of sets, which is effective to identify any errors or mistakes (for example, errors in state assignment in HITRAN database). If physics is well understood, it is rather straightforward to develop automated tools to visualize and evaluate spectroscopic data and XSAMS allows development of visualization and evaluation tools. It is recommended that IAEA should develop a Consultants' meeting to develop tools for spectroscopic data from available online resources (NIST, VALD, SPEC-W3 etc).

It is also recommended that IAEA should develop and maintain a standard data format, standardized provenance information as well as recommended data sets. If IAEA decides on how to categorize and store data in the IAEA databases, it may affect the organization of data within integrated tokamak modeling (ITM) development currently led by fusion communities. XSAMS was developed by atomic and molecular physicists based physics and generalized notation. Data exchange standards developed by users may not be based on physics consideration and hence may not be the best practice. On the other hand, data formats used by users accessing databases with XSAMS protocols should be easy to use. One should note, however, that such automatic visualization comparisons turn the relatively small size of data problem into a large data problem.

A light-weight set of standards and tools for identifying and describing atomic and (particularly) small molecule species and states is currently investigated. The benefit of this approach is that it would provide validation of data in ways that are difficult to achieve through an XML Schema alone (e.g. valid ranges of quantum numbers, specification of symmetries, etc.).

With respect to PMI (Plasma-Matter Interaction) processes and surface description, it would not be wise to extend XSAMS into these areas at present. The description of solids and surfaces is a sufficiently large and diverse area that the development of standards in this field requires the input of several groups who have not been actively involved in XSAMS and/or do not currently use it for their own purposes -- these span the radiation damage community, the solid spectroscopy community (e.g. SSHADE database), astrophysics (e.g. GhoSST), as well as the plasma-modellers, heterogeneous chemists, etc. -- it seems as though data standards in this area could be the subject of a coordinated research project.

Data Provenance, Curation and Quality of Databases

As the number of citation of a paper becomes a key parameter to judge the quality of work and hence for future funding opportunity, it becomes an issue how to give a credit to researchers who provide data sets for applications. For example, a Chianti database (http://www.chiantidatabase.org/) provides a comprehensive set of atomic data composed from many different sources. It provides a series of papers by the developers of the database on evaluated data sets, however, references of individual data sets are provided only in the database. Users may choose to give references to individual data sets and in practice, it is impossible since there are hundreds of papers for a single Chianti data file.

There is an active discussion between scientists and publishers to develop an infrastructure which give a proper credit to all contributors. Supplementary information for extra references is considered. Databases such as HITRAN (http://hitran.org/) provide a custom bibliography as a query produces a bibtex document for selected data sets. However, this approach will not apply to all databases.

Recently, a long-term stability and maintenance of a database is not guaranteed due to the lack of funding and personnel. Most databases of VAMDC consortium are not stable and they can disappear along with data sets. One solution is to distribute a database to interested parties with an infrastructure and software to support the distribution and data updates. This may reduce the probability of databases to disappear. However, it requires resources and the voluntary contribution from the parties to host and maintain a database for access as well as a policy about data maintenance. The success of this approach depends on the culture and need in each community. For example, the bitcoin database is

one of the most successful and safest distributed databases. However, this approach may apply only to relatively stable databases of moderate sizes since data transfer for large data sets is costly. It is different from the idea of mirror sites which were popular in 1990s as the distributed data sharing is a de-centralized database.

In order to mitigate the problem of distributed hosting and support for stability of databases, government and national institutes should be involved. A funding agency should be encouraged to promote national archives for supplementary information. For more longevity and access of data, repositories are better than databases to send data.

On a different note, an automatic update of data in the user side is a desirable feature. When data is renewed, the information is conveyed to a user who downloaded the data in the past, for example, by an e-mail. In order for a data retrieval information system to handle this update, the maintenance is likely to be costly and the IT security to host personal information of users can be an issue.

Big Data and Large Data

The exoscale computing will be the future of computing resources and computer architecture is also changing. The US government funds projects to rewrite software specifically for exoscale computing to be portable to future architecture of computers.

With the advances in computing, the volume of data from scientific computing is increasing as well. Big data refers to the large volume of data with its software stack relating small Meta data. Data from the Sky survey which is correlation of everything is a big data. In this sense, data in EXOMOL database (http://www.exomol.com/) is not big data but large data with a good structure. Generally, small data tools may manage these types of data and it is fair to say that atomic and molecular data scientists don't have issues with big data analysis. Electronic structure wave functions data are not big data and don't require correlation analysis. On the other hand, geological data, if correlated and patterned, are big data.

There is an increasing interest to save more elementary data before final products so that they can be used for other calculations. In this case, a standardization of data format is an issue and requires a tool to read/write/convert data formats. For example CHECK2XSF program provides an independent tool to convert formats for wavefunction processing. Storing a density matrix calculation can become very large. GAP (Gaussian Approximation Potentials) are available through LAMMPS (http://lammps.sandia.gov/), also ASE (Atomic Simulation Environment) Phython environment where everything that can generate potentials are stored.

Experimental data involving materials are indirect and in many cases inaccessible. Data for material sciences should deal with nanometer like structures which have micrometer size of extension from centimeter sizes of sample area and there is no standard way to describe images, surfaces or interfaces of samples. For PMI, density of vacancies is of interest as well as microscopic structures, dislocations, grain boundaries as well as element compositions. At present, atom probe tomography (APT) is used to deduce the lateral location of atoms of metals, semiconductors as well as biomaterials but not used by plasma-material interaction (PMI) research yet. The use of APT will create a huge amount of data with a few correlations that is "big data" and the community need to develop tools to process such big data.

Modeling of PMI is used to calculate damage production and defect distribution inside material. A database for hydrogen in tungsten is of interest to fusion community, however, it is a hard problem and the area is still evolving. There is an enormous need of modeling for experimental data analysis including atomic structure information inside the material. Sometimes experimentally smoothing techniques or a modeling with noise distribution is needed for experimental data analysis. Therefore a strong support from modeling and community is need to advance experimental PMI research.

Uncertainty quantification (UQ) of data becomes an issue of big data quite quickly since there need a large volume of data in order to produce uncertainties associated with data. As one wants to quantify non-trivial uncertainties, the size of data volume increases significantly. In general, UQ studies require 100 times more storage. Then the question is what should be stored for UQ studies. It may not be affordable to store covariance matrix elements.

Data Quality and Uncertainty Quantification

Data integrity is a big issue in the era of data research. For large data, the quality of all data can't be checked. The HITRAN databases hosts 10 million lines of a few of TB size, however, not every line need to be in spectroscopic accuracy. On the other hand, the completeness of lines to be included in a modeling of radiative transfer for atmospheres is more important.

American Statistical Society no longer recommends P-values as many studies claiming a measurement with a confidence of P-value > 0.5 can't be reproduced. Higgs Boson measurements with many noise values are one of the examples where it is not certain that the measurement was meaningful or a noise.

Integrated analysis with uncertainty distributions, for example, ASDEX-U profile data using diagnostics data as posterior information of Baysean analysis, makes the data analysis much more convenient and robust. It is necessary to define the best practice in doing integrated analysis.

Recommendations to IAEA

- Meeting to encourage database synchronization: a step beyond mirroring
- Meeting on experimental materials micro properties characterization: AFM (Atomic Force Microscopy) etc
- Meeting that leads to storage of potential energy surfaces (PES)
- Code comparison/model comparison workshop in the area of DFT and PES for materials: Workshop to put fusion PMI problem in front of electronic structure theorists. Repository of workshop data may be published at supplementary papers and/or hosted as a database. The organizing committee identifies interesting cases and determines test cases.
- Meeting on long term data preservation: Need to encourage funding agencies to support persistence and promote national archives for information. A letter of support from IAEA to funding agencies and journals will be helpful. IAEA should encourage communities to take an initiative and write an article about the sustainability of data repositories.
- Development of databases to host DFT results, for example, million data from T. Oda of SNU (Seoul National University): data sets can be disseminated.

Appendix 1

List of Participants

Gábor CSÁNYI, University of Cambridge, Trumpington Street, CB2 1PZ CAMBRIDGE, UNITED KINGDOM

Robert HARRISON, Institute for Advanced Computational Science, Suite 230 Heavy Engineering, Stony Brook University, 11794-5250 NEW YORK, UNITED STATES OF AMERICA

Christian HILL, Department of Physics and Astronomy, University College London, Gower Street, London WC1E 6BT, UNITED KINGDOM

Kerstin LEHNERT, Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY, 10964, UNITED STATES OF AMERICA

Yuri RALCHENKO, Atomic Spectroscopy Group, National Institute for Standards and Technology, 100 Bureau Dr., Stop 8422, Gaithersburg MD 20899, UNITED STATES OF AMERICA

Udo VON TOUSSAINT, Max-Planck-Institut fuer Plasmaphysik, Boltzmannstrasse 2, 85748 GARCHING, GERMANY

Bastiaan J. BRAAMS, International Atomic Energy Agency, Atomic & Molecular Data Unit, Wagramerstrasse 5, P.O. Box 100, A-1400 Vienna, AUSTRIA

Hyun-Kyung CHUNG, International Atomic Energy Agency, Atomic & Molecular Data Unit, Wagramerstrasse 5, P.O. Box 100, A-1400 Vienna, AUSTRIA

Naohiko OTSUKA International Atomic Energy Agency, Atomic & Molecular Data Unit, Wagramerstrasse 5, P.O. Box 100, A-1400 Vienna, AUSTRIA

Appendix 2

IAEA Consultants' Meeting on

Developments in Data Exchange

28-29 July 2016, IAEA Headquarters, Vienna, Austria

Meeting Agenda

Thursday 28 July 2016 **Meeting**

Room:A23-11

09:30 – 09:45 **Arjan Koning** Opening, Introduction of Participants, Adoption of Agenda

09:45 – 10:00 **Bas Braams** "Meeting Objectives"

10:00 – 10:40 **Kerstin Lehnert** "Data exchange in Earth science"

10:40 – 11:00 *Coffee break*

11:00 – 11:40 **Robert Harrison** "Computer science for Data Research"

11:40 – 12:20 **Gábor Csányi** "Machine learning and Gaussian processes"

12:30 – 14:00 *Lunch*

14:00 – 14:40 **Udo von Toussaint** "Data exchange in plasma physics; integrated data analysis"

14:40 – 15:20 **Yuri Ralchenko** "Data exchange in atomic physics; past and future"

15:20 – 15:40 *Coffee break*

15:40 – 16:20 **Naohiko Otsuka** "Data exchange in nuclear data community"

16:20 – 17:00 **Christian Hill** "The Virtual Atomic and Molecular Data Centre and future of XSAMS"

19:00 *Social Dinner*

Friday 29 July 2016

09:00 – 10:30 Discussion on big data and the future of data exchange for atoms, molecules and solids

10:30 – 10:50 *Coffee break*

10:50 – 12:20

12:20 – 14:00 *Lunch*

14:00 – 17:00 Future Activities of IAEA Atomic and Molecular Data Unit on Data exchange

17:00 – *Adjournment of Meeting*

Nuclear Data Section
International Atomic Energy Agency
Vienna International Centre, P.O. Box 100
A-1400 Vienna, Austria

E-mail: nds.contact-point@iaea.org
Fax: (43-1) 26007
Telephone: (43-1) 2600 21725
Web: http://www-nds.iaea.org