Policy proposal on the management of atomic and molecular data for fusion modelling

2024-11-25 – 2025-07-11

Table of Contents

1. Preamble

This document was initiated as an outcome of the “Meeting on unified A&M data policies” in FZ Juelich (see https://indico.euro-fusion.org/event/3240/) triggered at the Decennial IAEA Technical Meeting on Atomic, Molecular and Plasma-Material Interaction Data for Fusion Science and Technology 2024 (AMPMI 2024 - https://conferences.iaea.org/event/384/) held at the University of Helsinki, Finland, with follow-up discussion supported by a dedicated IAEA TM (see https://conferences.iaea.org/e/GNAMPP-3). It reflects the views of the informal researcher group formed during those meetings, including indirect participants. This document is a product of those discussions and joint effort approved by consensus between the main participants listed in section "Participants".

2. Participants

The following researchers, each with long experience of working (producing, utilizing in codes and experiments, validating, etc.) have jointly developed and support the following document to be addressed to ITER, EUROfusion, IAEA, and then to all the fusion research community:

1. Dr. Dmitriy V. Borodin, Forschungszentrum Jülich GmbH, Germany
2. Dr. Xavier Bonnin, ITER Organization, France
3. Dr. Martin O'Mullane, University of Strathclyde, UK
4. Prof. Dmitry Fursa, Curtin University, Australia
5. Dr. David Coster, IPP-Garching, Germany
6. Prof. Dr. Ursel Fantz, Universität Augsburg, Germany
7. Dr. Dirk Wunderlich, IPP-Garching, Germany
8. Dr. Juri Romazanov, Forschungszentrum Jülich GmbH, Germany

9. Dr. Kalle Heinola, IAEA, Austria

It should be noted, that among those researchers are the director of ADAS (ADAS webpage), principal developers of SOLPS-ITER (ITER press release, [Ref.]), ERO2.0, EIRENE (EIRENE webpage) and YACORA (YACORA online), representative of MCCC data production (MCCC DB), long-term leader of EUROfusion AMNS activity, as well as the IAEA A&M Data Unit Head. Thus this intitiative group (involving indirectly also further colleagues) has expertise covering data production, maintenance (databases) and utilization in fusion modelling. Naturally, this group is absolutely open to further extension by any relevant expert who shares its general view and is willing to contribute.

3. Purpose of this document

This initiative has brought together data users (i.e. people responsible for modelling codes or researchers interpreting diagnostics and analyzing experiments) and data providers. It aims at identifying ways of improving the management of atomic and molecular (A&M) and plasma-material interaction (PMI) data in fusion applications. The hope is that this document will trigger a broader response in the community in developing standards for data sharing. The suggested standards are agnostic towards the codes that might use the data, data production methods and validation tools - we aim for a commonly accepted, unified approach.
The initiative group aims at reaching concrete goals in the interest of fundamental fusion research. We are open for broader cooperation and participation and count on support from ITER, EUROfusion, and the IAEA (as this initiative is strongly supported by the A&M Data Unit).

4. Executive summary

The informal initiative group, inspired by the FAIR [Ref.] principles, proposes for consideration:

1. A set of data management policies - “best practices” regarding joint work on A&M data. It is based on experience of fusion-relevant datasets development. Those policies are suggested for a broader use far beyond the initiative group.
2. A proposal for next steps to be undertaken in the establishment and extension of the fusion-relevant A&M databases, development of CRMs (collisional-radiative models), and usage inside fusion codes.
3. A procedure to assess and recommend datasets for use in fusion modelling (including validation, quality and accuracy assessment, adequate resolution, and formatting).

5. Data types addressed in this document

A&M data can vary significantly due to production methods and their intended use. For instance, the data can have critically different levels of detail: bundled or resolved by ionization states, resolved by internal states (including generalized meta-stables), resolved by Rydberg, SLJ-terms, ro-vibrational states in molecular species, etc. Often, in addition to the basic fundamental data, we need to have effective data produced for particular modelling tasks (typically requiring additional assumptions and additional, sometimes hidden, parameters). This may lead to multiple datasets for the same species and processes, which cannot be sorted just by quality as a parameter, but need additional descriptions for proper selection in view of a given application.

Below we list the most commonly used data-types used in the fusion plasma modelling.

5.1. Cross-section data

At the most fundamental level, one requires a collection of data for individual processes. We also need to deal with collision cross-section types: resolved/unresolved by states of various resolution (e.g. one can consider or neglect the resolution by vibrational or rotational states in molecules, by SLJ temms or just main metastables in atoms - moreover, sometimes even bundeling of multiple charge states is used in W and other high-Z species), total or differential. In many cases, energy-averaged collision data can be preferable to reduce the size of the extremely high resolution needed for resonance effects (e.g. R-matrix with 100,000s point per transition). It is mostly of advantage to keep datasets produced by different methods complementing alternatives in a view of e.g. method advantages at various energy ranges.

5.2. Atomic data

For many practical purposes, the data is only required at a coarser description level. This can be e.g. Maxwell-averaged rate coefficients (effective data calculated for particular transitions depending only on electron temperature (Te) or even higher level “effective rates” depending on electron density (ne) as well using a local thermodynamic equilibrium (LTE) assumption). Effective rates may also have more parameters (e.g. neutral density, initial conditions, resolution level, assumptions on bundling, ion to electron temperature ratio, etc.). Thus, such higher-level datasets may be depend on multiple parameters specific for the case at hand, requiring to produce them upon request as a pre-processing step before running a simulation code. This is for example the current practice with ADAS.

5.3. Surface data

There is a close connection between surface data and boundary conditions at the plasma-material interface, in particular for molecular data, as the surface state (temperature, roughness, crystallographic orientation, etc.) can affect the ro-vibrational state distribution of molecules emitted from that surface, as well as other internal states. A more general discussion of surface data needs to be undertaken, but is not covered in this document. Many of the points raised here with regard to metadata and data formats are just as relevant for surface data.

Surface data is not limited to sputtering yields, but also includes reflection, surface recombination, implantation, etc, with dependencies not only on incident particle but also on surface state and composition/history.

6. Best practices for working with A&M data

All data should be properly documented, for which we recommend to use schemas and other similar technologies, rather than just describing the data with text.

Each data file (or group of files) should contain a metadata block in an agreed format (we recommend JSON with a schema), mandatory, and containing at least:

1. data origin (DOI, reference, etc...) and date of production
    a. unique ID (at least within the original database)
    b. clearly indicate the licensing associated with the data
2. data type
    a. data acquisition method (calculated, compiled, measured, etc).
    b. data nature type (cross-section, rate, etc.)
3. base process (e.g. base reaction or reaction type)
4. reference to a detailed data description document (files structure, units, etc).

5. general description of the data (probably at least as a comment, best containing the link to detailed description).
6. at least a general statement about the data validity range and specifying extrapolation methods
7. at least a general statement on the data accuracy and validation
8. sub-layered metadata for all included data (whenever possible) and optionally containing further data description, including links to previous instances of the same dataset and indicating the significance of the changes from the previous version.

In addition to the mandatory points listed above, we recommend to extend the metadata with various optional points, e.g.:

9. using the CollisionDB ontology (CollisionDB webpage) to identify species, reaction type classification (IAEA defined reaction types), etc.
10. indicating the best interpolation method for data tables.

We refer to the provided examples of the JSON metadata files described in the dedicated subsection "Metadata examples" as implementation proposals.

These metadata blocks can then be leveraged with the following practices:

1. All data users are recommended to keep the list of the data in use (and also the history of it) based on that set of recommendations. The unified assessment of the data from this group should use that information and regularly release the unified recommendation list (of course with additional checks and considerations). This could take the form of a higher-level document with overview about different available data versions for a specific question, and comments regarding quality, and contact persons, such as for instance the OpenADAS 'appxa' documentation (see e.g. https://open.adas.ac.uk/man/appxa-11.pdf).
2. Make data processing as automatic as possible and make the routines available (API approach is a good practice) as open source software with necessary documentation.
3. I/O routines should be open source; they should be universally applicable to all files of that format (versioning of any format is a must).
4. Possible use the pyvalem toolbox (https://github.com/xnx/pyvalem) to standardize the conversion of the data description to the metadata blocks.
5. Establish a set of standardized inter- and extrapolation routines (open source).
6. All data should be licensed and all data provided openly should remain as such. With regard to licensing, it is strongly remommended to use one of the well-established sets of licenses as e.g. Creative Commons (CC) license list. We recommend to use less restrictive licenses e.g. the creative commons CC BY-SA (Attribution-ShareAlike) which was suggested to allow use by commercial entities (i.e. all private fusion companies), but we understand that it may be necessary in some cases to limit the availabilty of the data. It should be noted that using specific types of licenses may restrict the applicability of data as input of codes producing effective data that itself is intended to be distributed. For example, CC BY-ND prohibits the distribution of material built upon data using this license. Nontheless the metadata should be provided for all cases (see point 9), separatelly licensed (as open as possible).
7. Motivate and assist towards proper referencing of the data:
    - provides (if possible) a DOI that can be used to refer to the data source and a DOI for one or more publications describing the data
    - provides a DOI that relates to the validation method of the data

    - provides a list of references to be cited when the data is used
    - provides references to use cases of the data with indication of success.
    - provides references to validation cases (if available) including the validation category.
8. Following the metadata format decided above, make any necessary changes to the IMAS Data Dictionary as it relates to such matters.
9. In accordance with FAIR, keep metadata open even for datasets with restricted access. Also make it available even in case when the actual data is no longer accessible.

7. Next steps

1. A working group should be established to elaborate on the metadata description (mandatory and recommended parts).

2. New data requests should be prioritized. Those can come from differen atomic/molecular data providers (ADAS, IAEA database, NIFS, NIST, etc.). Often the provided data takes the final form of an ADAS dataset, mainly the ADF11 and ADF15 formats. Other process-resolved data is available from the Curtin University group (with its own database) but it is recommended to access the data via CollisionDB (IAEA). These are fundamental cross-section data (including differential ones). These data may partially be already available in ADAS; if missing, but higher level data is needed, the participants of this group welcome effort to put the data (after reasonable checks) into ADAS.

3. Expansion of ADF15 line transitions lists available for spectroscopic data comparison with codes. The fundamental data may already exist but just needs post-processing. There is some ITPA-diagnostics effort to collate desired line lists. There may be merit in curating a smaller set of files specifically for AMNS purposes. Make the tools for producing such data open source, well documented, and commonly available.

4. Some dedicated effort must be made to bring the finer data at the individual rate level to the coarser description needed by many codes (effective cooling rates, total radiation emissivity, total particle balance, total emissivity within a diagnostic-relevant wavelength range, etc.).

5. Consider providing means to document automatically the particular dataset use experience (in the codes or post-processing analysis).

6. In the longer term, it should be evaluated whether the AMJUEL format is the desired format going forward, and to identify an alternative format if appropriate.

7. The proposed initiative will require focused attention and effort to be realized. Thus, it would necessary for the stakeholders to provide resources and, equally important, provide a contact point to coordinate actions and maintain this activity in the long run.

8. Additional context

- IAEA is a neutral forum for scientific and practical discussions for all its 180 Member States. It provides platforms, such as databases, which are free of use and are publicly available. Databases follow the FAIR principle.

- IAEA databases for A+M processes (CollisionDB) and PWI processes (pwiDB) use JSON metadata which include various information, such as reaction, process categorization (3-letter codes), data

type, bibliographical reference, DOI, free comment line, fit coefficients (if any), information on the time when the data was added in the database, etc.

- This draft was circulated among participants of the IAEA Data Centres Network to get feedback and consensus:

    1. ADAS (Atomic Data and Analysis Structure), UK
    2. Bariloche Atomic Centre, Argentina
    3. CRAAMD (China Research Association of Atomic and Molecular Data), China
    4. Forschungszentrum Jülich, Germany
    5. IAEA
    6. Queens University Belfast, UK
    7. KAERI (Korea Atomic Energy Research Institute), Korea
    8. Kurchatov Institute, Russia
    9. Korea Institute of Fusion Energy (KFE), Republic of Korea
    10. National Institute for Fusion Science (NIFS), Japan
    11. National Institute of Standards and Technology (NIST), USA

- A first effort towards the standardization of the AMNS metadata was undertaken by EUROfusion as part of the Integrated Tokamak Modelling (ITM) task force activities, and later absorbed into the IMAS framework developed at ITER. The responsible officer for this activity over the years (Dr. D. Coster, IPP-Garching, Germany) is among the authors of this proposal.

9. Note on the data quality assessment

As a spin-off of the the proposed effort on the standartized metadata, we suggest the new Data Quality Experience (DQE) databases (or forums) to be created to monitor the application and validation/verification effort of the A&M data utilising the same metadata as a cross-reference.

1. One should sepate the validation type:
    a. Validation with measured data by application of the data (often as a part of the code input) to the particular experiments
    b. Verification of the data for consistency, sufficient resolution, absence of abnormalities, etc.
    c. Uncertainty quantification - relating the uncertainties of the fundamental A&M data with the confidence intervals in the fusion-relevant modelling results.
    d. Code-code validation, analysis of assumption sets.
    e. Comparison of particular datasets.
2. The filling of those DQE databes should be as much as possible voluntary by the researchers undertaking the validation/verification effort. The additional effort to summarize the experience should be minimised (at that it is anyway a part of any scientific pulication or a report on that kind of work). The obvious common good and popularisation of one's own scientific results can be a sufficiently strong motivation, however one can think of additional stimulation mechanisms.
3. The creation and maintainence effort of the DQE databases is not meant to be on the origianal data providers. It is expected to be a separate independent projects supported by the major stakeholders. Of corse the bases will most probably cover just a part of the data field in the interest of the the particular owner. Still, the standartized metadata will facilitate that they will complement each other.
4. The DQE database should contain references to the related application/validation/verification effort publications.

5. The DQE database should, where possible, contain links to the input file packages allowing reproducibility of the simulations or analysis.
6. The DQE database can contain multiple entries for each data piece, in fact that should be even encouraged. It is also only of advantage if there will be multiple such databases, provided they refer to the same metadata, which will allow mutual cross-reference.
7. The DQE databases (after being filled to a certain extent) will be very useful for any kind of dataset evaluation committee.

Policies formulated in this document do not directly lead to the data quality assessment. Still, they provide an opportunity to track the history of the data production and use based on the metadata in a standard format. That may enable following the validation experience by the data used (in particular if DQE databases will be created and flourish) in different applications and facilitate the work of any kind of evaluation committee.