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

2024-11-25 – 2024-11-27

This document was started at the “Meeting on unified A&M data policies” in Juelich. It reflects the views of the participants in this meeting (see <https://indico.euro-fusion.org/event/3240/>).

The authors of this document are ready to continue work on this document and further steps in the direction as an informal group until the initiative will get a more solid support. The IAEA is ready to host the document at its website; we will also create a Mattermost channel for further communication.

# Purpose of this document

This meeting brought together data users (i.e. people responsible for modelling codes) and data providers and aimed at identifying ways of improving the process of producing and managing atomic and molecular (A&M) and plasma-material interaction (PMI) data in fusion modelling codes. 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 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, in particular, the IAEA (as the initiative was triggered by the discussions at IAEA Decennial Meeting on A&M and PMI data),

This meeting was triggered by discussions at the Decennial IAEA Technical Meeting on Atomic, Molecular and Plasma-Material Interaction Data for Fusion Science and Technology 2024 (AMPMI 2024) held at the University of Helsinki, Finland and aims at reaching concrete goals in the interest of fundamental fusion research.

# Executive summary

The meeting has triggered a creation of informal initiative group which proposes 2 documents for consideration:

1. Policy document - “Best practices” regarding the joint work on A&M data to be applied at least to own work first, but later on (after reaching maturity by that experience) can be suggested for a broader use.
2. Proposal for next steps to be undertaken in the interest of joint development of A&M data bases, CRMs and related entrees to the fusion codes.

The initial draft of the documents formulated, the work should continue after the meeting.

We aim to seek more solid support from…

We aim at developing a procedure to assess and recommend for use in fusion modelling the datasets (including validatiion, other quality assesment, adequate resolution and format).

#

# Data types addressed in this document

## Atomic data

### Rate coefficients (basic data calculated for particular transitions depending only on Te as well as higher level - “effective rates” depending on ne as well). Effective rates may also have more parameters (e.g. ,resolution level, assumptions on bundling, Te and Ti relation etc.) So, it may dependent on multiple parameters specific for the case at hand, which is also the current practice with ADAS - to produce such datasets with varios resolution by states on request. This is largely satisfied by ADAS, mainly ADF11 and ADF15 datasets

Process resolved data is available from D.Fursa group (own database…) but it is recommended to take the data via CollisionDB. This is fundamental cross section data (including differential ones). This data may partially be already available in ADAS; if missing, but higher level data is needed the participants of this group are welcome to take the effort to put the data (after reasonable checks) to ADAS.

###

Expansion of ADF15 line transitions list 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.

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)

For ADAS formats, and particularly those used by major fusion modelling codes, add in a metadata block (following FAIR definition) to include summary of contents, ranges, dependencies and units. This should be separate from the existing comments.

### Cross section data

Requires collection of data from different atomic/molecular data providers (ADAS, IAEA database, NIFS, NIST etc.)

Collision cross-section types: state resolved/unresolved, total and differential.

Energy averaged collision data as a preferred format to reduce the size of the extremely high resolutions needed for resonance effects (eg R-matrix with 100,000s point per transition).

New data requests should be prioritized. At present, e+W excitation, H^+(D,T) + W - CX

Atomic data:

Data with variable resolution (bundled or resolved by ionisation states, resolved by internal states (generalized metastables), resolved by Rydberg/SLJ …. Etc. states

##

## Molecular data

Many of the existing (European) codes use molecular data based on the AMJUEL suite of data files. These files represent data as a function of one or two parameters.

All fusion-relevant molecules (H/D/T incl. HT, DT, also W, B, Be, Li, N containing molecules - also with isotopes)

Mixed molecule (DT) data? Yes

In order of general importance is the proper treatment of the hydrogen molecule with differing isotopes including the full break-even chain in edge plasma conditions. This data should be rotationally and vibrationally resolved (of particular importance for spectroscopy), and reduced versions of this data should also be produced for plasma modelling. Much of the modelling for ITER has been done with SOLPS-ITER using a version of AMJUEL/… data which doesn’t resolve vibrational states.

Data for negative ions and molecular ions such as H- and H3+ are also of interest.

## Surface data

~~We leave that generally for a dedicated discussion, but clear that it should be as much in line with the A&M data~~

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

There is a close connection between the surface data and the boundary condition for the molecular data as the surface can affect the ro-vib state of the molecule starting at the surface. The more general discussion of surface data needs to be considered, but is not covered in this document. Many of the points with regard to metadata and data formats are also expected to be important for surface data.

# Data formats

Many of the data providers already have a format for “their” data. This section recommends some additions and/or changes that will be of benefit to the community.

## Addition or inclusion of metadata

Metadata is an important aspect of providing FAIR data.

Recommendation #: That any A&M data have associated metadata that

* use the CollisionDB ontology (<https://db-amdis.org/collisiondb/>) to identify species, reactions (<https://amdis.iaea.org/databases/processes/>), etc.
	+ Perhaps use the pyvalem toolbox (<https://github.com/xnx/pyvalem>) to standardise
* clearly indicate range of validity of data, recommended extrapolation methods
* should indicate the best interpolation method for data tables
* clearly indicates the licensing associated with the data
	+ with regard to licencing, a good starting point might be the creative commons set of licenses
	+ CC BY-ND (Attribution-NoDerivatives) was suggested to allow use by commercial entities (ie all private fusion companies)
	+ List of licences https://creativecommons.org/share-your-work/cclicenses/
* 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
* Provide references to use cases of the data with indication of success.
* Provide reference to validation cases (if available) including the validation category.

Recommendation #: Following the metadata format decided above, make any necessary changes to the IMAS Data Dictionary as it relates to such matters.

Amongst the recommendations are:

1. Data quality
	1. Metadata must
	2. Metadata
2. Versioning (generation, DOIs)
	1. Metadata to include how to refer to the data in publications for proper credit

## Format of data

1. “Carrier”-formats: encapsulated, up-to-date: JSON (with Schema), XML, HDF5 - compatible with IMAS
	1. Preference given to self-describing formats?
	2. Standardized units?
2. Data documentation, examples, standard i/o libs (with APIs)
3. Compatible with main simulation packages and
	1. Get EIRENE to be able to fetch data using the AMNS library
4. Data uncertainty
	1. If we want to be able to perform uncertainty quantification, having a representation of uncertainty in the data including cross correlations would be useful.
	2. We should separate systematic errors and precision uncertainties

## Data repositories

1. Data acceptance policies, including licensing
	1. One should set standards to accept the data to database, also review procedure.
	2. There should be a policy on the accumulated data revisions.
2. Data repository, access
	1. IAEA databases: CollisionDB, Aladdin, etc…
	2. Long term data repositories so DOIs do not end up pointing to dead links
	3. Data versions to be preserved for reproduction of runs (even incorrect data is valuable!). May be not every data should be allowed at this stage, see “Data acceptance”.
	4. Data variability (by parameters, history, licence)
	5. Data conversions. Should we have 1 base format (convert to it whatever comes in) or keep as submitted?.. Or ease the interchange between formats because providers will not necessarily want to change their own formats.

Key words not to forget: data dictionary, data mining

* [DF] Integral cross sections versus differential cross sections
	+ Too much data?
	+ Can provide momentum transfer cross-sections
	+ Or viscosity cross sections
	+ Are there methods for reducing the amount of data for differential cross-sections?
	+ Needed for cases with particles with comparable velocities
	+ Issues with resonances in e-molecule scattering (typically assume fixed nuclear distances which results in artificial resonances)
	+ Proper treatment of resonances would be very expensive
	+ MCCC produces data in adiabatic approximation averaged over resonances – very happy with this
* [MoM] atomic systems have resonances and procedures are in place to average over resonances which are compatible with later maxwellian averaging to produce rates
* [DF] MCCC currently produces text files with a header and then X and Y columns
	+ Would like to know what information should be in the heading
		- Need to agree on a format!
* More generally: data files (as delivered) should have some metadata in a standard format [JSON]
	+ Can we agree on some standard labels? [Schema, Data Dictionary]
	+ Example from CollisionDB is a good starting point
		- TRIM data files are a good example of how not to do this
* [DF] Would like to have a publicly accessible page that identifies missing data, priorities for data [IAEA agrees to host this]
	+ Would need to have a mechanism for adding requests
	+ DCN could regularly provide feedback on this list
	+ Annotate the list once work has been done with DOIs of publications that addressed the request
* [MoM] Need a section in the document on concrete actions(e.g. New AMJUEL file from UF; XB then runs the H ITER standard cases; publication about the changes in results; ADAS should add a metadata block [see above])
* CH: need a mechanism for communicating about further outcomes of this meeting, and also a mechanism to discuss the priorities for new data as mentioned above
* MO’M promises from the participants are (1) to add sufficient metadata to existing/legacy formats and (2) that edge codes use the new data.

BEST PRACTICES with working on A&M data

1. Each data file (or group of files) should contain metadata in an agreed format (JSON with schema) containing
	1. data origin, date of production and unique DOI
	2. data on validation including use cases not dedicated to validation
2. All data users are recommended to keep the list of the data in use (and also the history of it) with based on that list of recommendations. The unified assesment of the data from this group should use that information and regularly release the unified recommendation list (of course with additional checks and considerations).
3. Do data processing as automatic as possible and make the routines available as open source with necessary documentation.
4. All data should be properly documented, moreover we recommend to use schemas and other similar technologies rather than just describing the data as a text. 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).

LIST of RECOMMENDED ACTIONS int the in interest of fusion-relevant A&M data development

Recommendation 1: That the IPP group lead by Ursel Fantz generate a new set of AMJUEL-like files based on the MCCC data as used by YACORA; that this file / these files be passed to Xavier Bonnin at ITER where he will rerun the hydrogen fueling scan and document any differences that might arise; these results will appear in a conference presentation and a paper. After this the data should be made more generally available (e.g. at PLOUTOS meaning as a standard input for EIRENE, SOLPS and other related packages).

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

Recommendation 2. As part of the IMAS-ification activity within the TSVV-5 EUROfusion activity, have EIRENE and/or its ancillary program be able to read data in using the AMNS library from IMAS. Longer term, encourage other codes to support that format.