First Meeting of the WPEC Subgroup 50 on Developing an Automatically Readable, Comprehensive and Curated Experimental Reaction Database

Amanda M Lewis (NNL)
Denise Neudecker (LANL)
Arjan Koning (IAEA)
Michael Fleming (NEA)

October 12, 2020

1 Meeting Outcomes

1.1 Establishment of Previous and Relevant Work

1.1.1 EXFOR Correction System - Viktor Zerkin

The EXFOR correction system is already in place in the EXFOR web retriever and can modify the information as it is retrieved by the user. It does not make changes to the underlying EXFOR entry, and application of the modifications must be chosen by the user. This system stores the changes that should be made rather than storing the modified data sets. There are some changes that can be done automatically, such as updating monitors and standard values. For more complicated changes suggested by experts, manual changes can be made with numerous options available including re-normalizing the data, changing uncertainties, fixing units, abundances, half-lives and monitors.

1.1.2 X4Lite - Viktor Zerkin

X4Lite is EXFOR in a SQLite database, which allows for easy filtering and searching of the data sets. The EXFOR entries can be translated into many different formats, including the widely used C4 and C5 formats and newer formats such as JSON. X4Lite can be accessed by the GUI, API or by FTP. X4Lite includes the expert changes from the EXFOR correction system, as well as other extensions such as archives of old subentries.

1.1.3 JSON Database - Georg Schnabel

The EXFOR database has been translated into JSON and is currently available both as a CouchDB database and as a MongoDB database, which have powerful search capabilities. The parser was written in R and preserves the hierarchical structure and the keywords and codes in the EXFOR format. The database is available in JSON format which users can then easily read in with their preferred language to perform modifications.

1.1.4 x4i - Dave Brown

The x4i program processes the EXFOR database into python data structures and can easily be translated into JSON. The structure, keywords and codes of the EXFOR format are preserved, and the program deals with the uncertainties that it can parse. The reaction strings are treated as a grammar, which allows for complex searches and for symbolic math to be performed on the reaction.

1.1.5 EXFORTables - Arjan Koning

EXFORTables is a database that contains the EXFOR database in a format that is similar to but simpler than the EXFOR format. The database makes comparisons between EXFOR entries and TALYS very easy to perform. EXFORTables already includes C/E scores that suggest the quality of the data set, based on comparisons of all available data sets and library evaluations for that reaction observable. Distributions can be easily constructed, such as the uncertainties of a specific reaction observable or the values for the observable for finding outliers. EXFORTables will be made available in 2020.

1.2 Determination of User Needs

The users for this database were split into two different groups. The first group is represented by "the evaluators" who are expected to understand the data sets to some degree and want access to as much information as possible. This includes experimental condition information, opinions of and corrections made by previous evaluators, and enough uncertainty information to bring all considered experiments to consistency. The second group is represented by the "Machine Learning users" who have less expertise in nuclear reaction measurements and want a curated database that can be easily fed into their automated code. They do not want to wrestle with the data prior to using it and therefore want the experts to provide a curated set of data sets that are corrected when necessary. Both groups contain more than just their representative members, and in many cases overlap, but the needs here will be described by these two groups.

Machine-readability All users would like for the database to be more machine readable than the current EXFOR database. The hierarchical format of the EXFOR entries can be parsed, but there is a lot of important information stored in free-text portions. For example, data reduction corrections and uncertainty sources are compiled into free-text portions of the EXFOR format, in the same order and using the same language as the journal article. This perpetuates the problem within the field of not having consistent methods to describe or quantify systematic uncertainties and corrections. There are also issues within the EXFOR database where formats have changed over time or have ambiguities, such as the way the year is represented in the reference. These issues have mostly been corrected in the X4Lite system.

Realistic Covariance Matrices Both types of users need realistic covariances in order to use the data effectively. There are some data sets in EXFOR that do not have any associated uncertainties and many data sets in EXFOR that have uncertainties for each data point without covariances. Many do not provide partial uncertainty sources to allow the user to reconstruct or estimate the covariance matrix. The field is improving in this regard, but there is a wealth of information contained in the older data sets that should not be lost. In order to bring the data sets onto the same level as the newer data sets, and to be able to compare them effectively, realistic covariance matrices need to be estimated for these data sets. Estimating uncertainties in this way is well outside of the philosophy of EXFOR, and therefore must be done in a new database like this one.

Standardized Systematic Uncertainties Many users who fall into the "evaluators" group would like to know what the sources of uncertainty are for the data sets, beyond just a realistic covariance matrix. This allows them to understand the data set itself in more detail, and may help to uncover issues with experimental biases and correlations with other data sets used in the analysis. As explained above, within the field of nuclear data measurements it is common for the same source of systematic uncertainty to be referred to by different terms, calculated by different means, or separated out into different components for different experiments. The EXFOR compilation contains the sources, values and terms that were used in the journal article, consistent with the philosophy of EXFOR. In order to have standardized systematic uncertainties, this subgroup will need to define the sources and their names for the new database. This way, the same source of uncertainty will be coded in the same way for all experiments, regardless of the terms used in the paper. This change, similarly to estimating uncertainties, falls outside of the philosophy of EXFOR and is therefore more appropriate in this new database.

Important Experimental Conditions For evaluators, it is important to include all experimental information that is necessary to understand the experimental data. For example, the evaluation of total cross sections in the

Resolved Resonance Region requires forward modeling of the experiment to match the experimental observable, transmission. This forward modeling can only be done accurately if experimental information like the resolution function is included with the experimental data set itself. This information is usually stored in the EXFOR entry, but consistent and standardized formats need to be developed if the users are to be able to extract this information from the database without having to consult the primary references.

Important Data Reduction Corrections For evaluators, corrections applied to the data in the process of data reduction are very important as well. For example, multiple scattering and secondary neutron production in the experiment can cause biases in the measured cross sections in the fast region. The best way to correct for both effects is with detailed, iterative Monte Carlo simulations. This solution has only recently become widespread and accurate and if there is no information in the database about how (or whether) the corrections were performed, then the evaluators will have to go back to the primary reference. Corrections such as these can provide information about discrepancies between different data sets and insight about whether presented uncertainties are realistic.

A Curated Database For "Machine Learning users" a curated database is desired. These users are usually putting the experimental data into automated codes and should not have to spend time cleaning up and understanding the experimental data sets. For these users, the database should have some set of quality indicators, whether they are flags or numerical scores, that allow them to automatically select trusted, corrected data sets with realistic uncertainties. For the evaluators, this type of quality indication is also very helpful. A major goal of this database is knowledge management within the evaluation field. Too often, evaluators will spend weeks or even months studying, testing and correcting data sets that they use in their evaluations, and when the evaluation is published that information is lost. This curated database would be created from the corrected data sets that expert users including evaluators would submit to the database, thus preserving their work. Future evaluators will not have to discover problems with data sets again on their own, but rather will be able to build on the work of the previous evaluator.

1.3 Creation of Sub-subgroups

This project involves many tasks that are varied in their purpose and required expertise. For this reason, six sub-subgroups were proposed to handle specific issues. The sub-subgroups are explained below, along with a list of potential members. These descriptions are current to the time of preparation of this document, early October 2020.

1.3.1 Metadata/uncertainty keywords

This subgroup will define the minimum metadata and uncertainty information needed for all types of experiments that will be considered here, and will produce a requirements document. Smaller groups of experts (both experimentalists and evaluators) in particular observables will create lists of the metadata and uncertainties needed to fully document each experiment type. Then, the sub-subgroup will combine all of the lists together with consistent terminology, and assign keywords and codes to each. Alias tables will be constructed to improve the efficiency of automatic translation of EXFOR files into this database. The sub-subgroup will iterate on the requirements with feedback from the other sub-subgroups.

- Jesse Brown
- Roberto Capote
- Stefan Kopecky
- Amanda Lewis
- Denise Neudecker
- Mark Paris
- Peter Schillebeeckx
- Vladimir Sobes

- Don Smith

1.3.2 Coordination with NRDC

This sub-subgroup will define feedback mechanisms to the International Network of Nuclear Reaction Data Centres (NRDC) throughout the whole process. The new keywords and information that are added to the format for this database can potentially be included in the EXFOR compilation process, thus improving the ability to translate EXFOR to Layer 1 over time. This subgroup will also provide feedback to the keywords sub-subgroup regarding the ability of the compilers to add all of the new information to the compilation process.

- Amanda Lewis
- Shin Okumura
- Naohiko Otsuka
- Boris Pritychenko
- Viktor Zerkin

1.3.3 Database creation and structure

This sub-subgroup will decide where to store the database and a database type. They will create a specifications document to follow the requirements document. The structure will be defined for all three layers, including the best way to store correlations between data sets and which information should be stored and which should be generated on-the-fly.

- Amanda Lewis
- Denise Neudecker
- Georg Schnabel
- Viktor Zerkin

1.3.4 Code creation

This sub-subgroup will be in charge of producing the codes needed for the construction and use of the database. Codes are needed to translate EXFOR entries into the new database, and to apply objective and subjective corrections to the data sets in Layers 2 and 3 and to reconstruct covariance matrices on-the-fly for some data sets. In addition, an API and codes are needed to translate from this database to the other commonly used data formats (such as C4, C5, JSON) that many users rely on for their own analysis codes. The functionality to translate from EXFOR into the other formats already exists in the X4Lite code, and the majority of the work for this subgroup will be the API and the codes for the corrections and the generation of covariance matrices on-the-fly.

- Dave Brown (advisory role)
- Amanda Lewis
- Gustavo Nobre
- Georg Schnabel

1.3.5 Correction procedures and documentation

This sub-subgroup will work to define the procedures and documentation for applying corrections and flags to the data sets. The list of corrections and flags that are considered "objective" (appropriate for Layer 2) needs to be decided on, as well as the standards for making such corrections. In addition, this sub-subgroup will define the process by which the corrections and flags are applied, such as whether it can be done automatically on a large scale. For the "subjective" corrections allowed in Layer 3, this sub-subgroup will define the process for experts to

submit the corrected data sets and how they should be documented, including the format for submission. GNDS v2.0 will have a mechanism to record what data sets were used in the evaluation, and the documentation in this database should complement this. ENDF is implementing a peer review system for new evaluations that should be mirrored by this database as well. This sub-subgroup should define the method by which new corrected data sets submitted to Layer 3 can be peer-reviewed. If possible, the peer-review for new evaluations should include this data set review as well. Finally, this sub-subgroup will decide what options the user is given when they search for a reaction. They will decide if the user should be presented with all options or if a default choice should be given.

- Dave Brown
- Roberto Capote
- Mike Herman
- Arjan Koning
- Stefan Kopecky
- Amanda Lewis
- Denise Neudecker
- Gustavo Nobre
- Mark Paris
- Boris Pritychenko
- Georg Schnabel
- Vladimir Sobes

1.3.6 Testing formats

This sub-subgroup will take the example files created for each layer and test them in evaluation and machine learning uses. The feedback from this sub-subgroup will inform the other sub-subgroups and allow them to iterate on their decisions. For example, this sub-subgroup will give feedback to the corrections sub-subgroup about whether the flags are helpful and usable in their applications. In addition, this sub-subgroup will test the submission format and method for corrected data sets. This rapid user testing and feedback is crucial for creating a database that will continue to be populated and used after the subgroup is closed.

- Denise Neudecker
- Henrik Sjöstrand
- Vladimir Sobes
- Kyle Wendt

1.4 Decisions

The scope of the subgroup will cover:

- The CIELO isotopes (^{1}H, ^{16}O, ^{56}Fe, ^{235,238}U, ^{239}Pu) for in-depth correction procedures and translation schemes and files for testing.
- Neutron and possibly charged particle-induced reactions for global translation schemes.
- A subset of observables that have an MFMT number in the libraries, including cross sections and some angular distributions.

Other decisions that have been made:

- Object-oriented databases will be used.

- There needs to be some sort of quality score or flag for the data sets, and all versions and comments should be available to users.
- Large covariance matrices should be stored in separate pieces and reconstructed on-the-fly where possible.
- Consistency with SG-49 must be prioritized.
- Layer 3 will have tags for data sets as used in evaluations.
- Journal articles and reports should be stored if possible.
- The NEA github will be used for codes in progress.
- We can start from X4Lite for the conversion codes.

Decisions that still need to be made:

- Which database to use (MongoDB vs CouchDB).
- Which corrections are objective or subjective.
- Which information to store and which to reconstruct (covariances, corrections).
- How to store the corrections and quality indicators.
- How to represent a single entry with multiple subentries in Layer 1.
- How to ensure that the database continues to be populated after the subgroup ends.
- How the users will interact with the database.
- How expert users will submit versions to Layer 3.
- How to deal with uncertainty sources that are presented together, rather than separated out.
- How to store complex experimental setup information such as resolution functions, and where to keep the documentation that explains them.
- Whether to include thermal scattering data, based on conversations with SG-49.
- How to distribute the codes when they are complete.

2 Deliverables

The work of this subgroup has been split up into four main deliverables:

1. Create a requirements document based on the needs of the users of the database and the experts in the experiment types. This will include the types of metadata needed for different observables.
2. Create a specifications document based on the requirements and the type of database chosen.
3. Create codes to populate the database and translate into many formats.
4. Produce example files for each layer and test them.

The whole process will allow for iteration, and the last two steps will be repeated for each layer of the database.

Mini-meetings will be held with small groups of experts to define the metadata needed for each observable, and a draft of the requirements document should be produced by the May 2021 meeting.