(A23) Follow-up of WPEC Subgroup 30 activities on quality improvement of the EXFOR database

E. Dupont, N. Soppera, M. Bossant, N. Otsuka[*]

OECD NEA Data Bank
[*] IAEA Nuclear Data Section

The WPEC Subgroup 30 (Sg30) was established in 2007 to improve the accessibility and the quality of the EXFOR database. During the lifetime of Sg30, several batches of error reports have been sent to the NRDC [1-3]. The whole collection of suspicious entries reported by Sg30 is available on the IAEA-NDS website together with NRDC comments [4]. Sg30 was established as a short-term subgroup and was closed in June 2010. The final Sg30 report [5] describes the translation process from EXFOR into more user-friendly tabulated data files, as well as the various methods developed to verify and correct the contents of the database. Sg30 contributed to the long-term objective to establish EXFOR as an easily accessible and correct database. However, Sg30 achievements were only the "end of the beginning" and Sg30-like activities should be continued, e.g.

- Distribution of EXFOR in C4 computational format (V. Zerkin, IAEA-NDS).
- Automatic test of C4 data and comparison with TALYS/TENDL (A. Koning, NRG).
- Development of a C4 correction system to collect and apply evaluator's feedback (e.g. normalisation) on C4 data (V. Zerkin, IAEA-NDS).
- Compilation and monitoring of coding mistakes reported by users and NRDC staff [6,7] (N. Otsuka, IAEA-NDS).
- Implementation of Sg30 methods and recommendations (all).

In September 2010, Arjan Koning made available another EXFOR/TALYS comparison with a wealth of statistical information on EXFOR problems. However, analysis of this comparison still has to be completed (a partial analysis was presented during the NRDC2010 meeting [3]).

More recently, a statistical approach based on ideas developed within the Sg30 framework was implemented at the NEA Data Bank. An overview of the method is given in Appendix and will be published in NEA News [8]. Results of these statistical tests have been analysed for cross-section data from NEA DB entries and two lists of suspicious data were sent to IAEA-NDS in February 2011 for cross-checking with original articles. The following table summarises the results for the 44 subentries reported as suspicious. Coding mistakes have been registered on the IAEA-NDS feedback list [6].

	Area 2	**Area O**	**Areas 2 and O**
Not in error	4 *(15%)*	5 *(28%)*	9 *(21%)*
Error	18 *(70%)*	9 *(50%)*	27 *(61%)*
Not resolved yet	4 *(15%)*	4 *(22%)*	8 *(18%)*
Total	26 *(100%)*	18 *(100%)*	44 *(100%)*

As for the EXFOR/TALYS comparison, this type of analysis is useful but time consuming and results for cross section data from other areas remain to be analysed, as well as results for other quantities (DA, DE, FY, etc.). As mentioned in Sg30 final report, the future quality of EXFOR depends strongly on the resources that NRDC can invest in correction activities.

References

[1] “NRDC action to the list of EXFOR outliers”, NRDC working paper WP2008-3 (2008).

[2] “EXFOR Outliers (Parts 4 and 5)”, NRDC working paper WP2010-10 and memo CP-D/623 (2010).

[3] “Automatic test of EXFOR with TALYS”, NRDC working paper WP2010-11 and memo CP-D/627,633 (2010).

[4] www-nds.iaea.org/nrdc/error/exfor_err3.html.

[5] Nuclear Energy Agency (NEA), “Quality improvement of the EXFOR database”, *International Evaluation Co-operation*, Vol. 30, NEA/WPEC-30, OECD/NEA, Paris (2011). See www.oecd-nea.org/science/wpec.

[6] www-nds.iaea.org/nrdc/error/exfor_err1.html.

[7] “Review of feedback list with new flags”, NRDC working paper WP2011-16 and memo CP-D/697 (2011).

[8] “Statistical methods for the verification of databases – Application to the international database of experimental nuclear reaction data (EXFOR)”, to be published in NEA News (2011). See Publications at www.oecd-nea.org.

Statistical methods for the verification of databases – Application to the international database of experimental nuclear reaction data (EXFOR)

E. Dupont [1], B. Beauzamy [2], H. Bickert [2], M. Bossant [1], C. Rodriguez [2], N. Soppera [1]

[1] OECD Nuclear Energy Agency Data Bank, Issy-les-Moulineaux, France
[2] Société de Calcul Mathématique SA, Paris, France

Large databases often contain a significant percentage of missing or erroneous data. There might be various reasons for that, e.g. failures in the measuring instruments, human errors, lack of budget. In such cases, despite significant efforts to collect the data, the overall value of the database may still be doubtful.

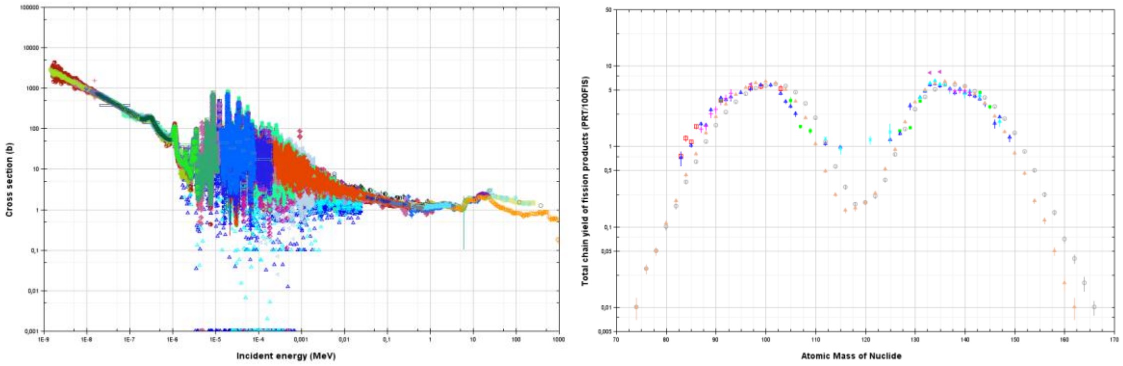
The NEA undertakes significant efforts to maintain the highest level of completeness and quality in its databases. The NEA continuously collects new data to keep its database up-to-date and devises verification procedures to check the quality and integrity. We describe below an example of such efforts; it deals with a probabilistic procedure to detect aberrant data, whether these data are isolated or come as a cluster. This procedure was developed to check the international database of experimental nuclear reaction data (EXFOR), but the same method should apply to other numerical databases.

The EXFOR database

As a member of the international network of Nuclear Reaction Data Centres (NRDC), the NEA contributes to the international database of experimental nuclear reaction data (EXFOR) and makes it available to scientists and engineers through the web and the JANIS software (NEA, 2010). Experimental nuclear reaction data have been compiled in the EXFOR database for more than 40 years. At present, EXFOR is by far the most important and complete experimental nuclear reaction database with more than 130 000 data sets from about 19 000 experiments performed since 1935. The database mainly contains numerical data on low to medium energy experiments for incident neutron, photon and various charged-particles induced reactions on a wide range of isotopes, natural elements and compounds. Different nuclear reaction quantities are available in the database, e.g. cross sections, angular distributions, energy and energy-angle distributions, resonance parameters and fission-fragment yields. These data are widely used for nuclear science and technology. However, the automation of nuclear reaction codes and plotting software, together with the increase in computer speed and direct access to these data have made their reliability an even more important issue than it was before. For this reason, the NEA Working Party on international nuclear data Evaluation Co-operation (WPEC) founded a new group of experts, Subgroup 30 (Sg30), with the aim to establish EXFOR as a more easily accessible and reliable database (NEA, 2011). The work described here is a follow up of Sg30 activities with the aim to implement Sg30 recommendations at the NEA Data Bank.

Some examples of data available in EXFOR for neutron-induced reactions are given in Figure 1. In these figures, experimental data are grouped together into clusters containing the results of measurements of the same quantity for the same reaction. The basic idea of the present verification method is to detect isolated data (also known as outliers), which are not part of any cluster and which may then be flagged as suspicious, but not always necessarily wrong. Of course, the dispersion of the cluster as well as the uncertainties on the data must be considered as well.

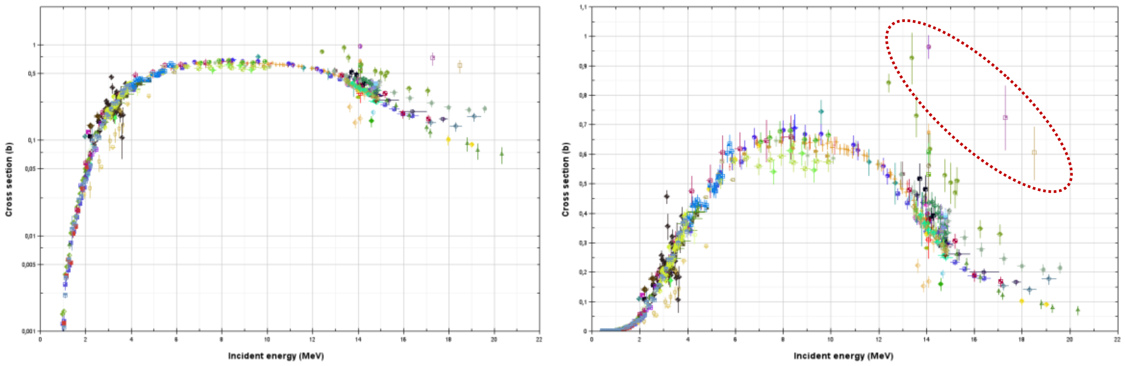
Figure 1: Experimental data available in the EXFOR database for the ^{235}U fission cross section (left) and for the fission-fragment mass distribution from the ^{238}U fission reaction induced by 14 MeV neutrons (right).



Representation of the data

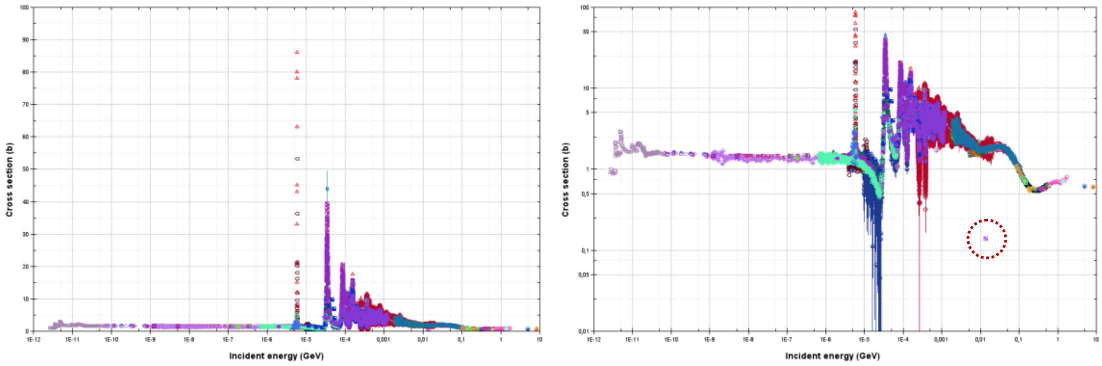
One essential point to address when looking for outliers is the representation of the data. Indeed, as can be seen on Figure 2-left (logarithmic scale) and Figure 2-right (linear scale), the scale may affect the perception of isolated data. In this example, the plot in logarithmic scale reveals only scattered measurements at high incident energy. However, the same plot in linear scale shows that a few measurements at 14 MeV and around 18 MeV are overestimated compared to other data (See the red circle in Figure 2-right).

Figure 2: Experimental data available in the EXFOR database for the ^{58}Ni(n,p) reaction cross section. Data are displayed using logarithmic (left) and linear (right) scales.



In the case of nuclear reaction quantities spanning over several orders of magnitude, the importance of the choice of a proper scale is even more obvious. As an illustration, the same data are displayed on Figure 3-left (linear scale) and Figure 3-right (logarithmic scale). In this example, a suspicious data point could be easily identified using the logarithmic scale only (See the red circle in Figure 3-right).

Figure 3: Experimental data available in the EXFOR database for the ^{27}Al total cross section. Data are displayed using linear (left) and logarithmic (right) scales.



The choice of the scale applies to all axes. In the present work, the optimal scales were chosen independently for every axis among polynomial scales of different degrees (a polynomial degree of 1 would correspond to a linear scale, whereas large degree values would give results similar to a logarithmic scale). The optimal scale is one which allows plotting the data as evenly spaced as possible. Hence, every value, small or large, is represented with the same precision.

Detection and quantification of suspicious data

The approach used here is probabilistic and allows the identification of isolated suspicious data, as well as a cluster of suspicious data. The method is based on the discretization of the parameter space (energy, angle, temperature, etc.) and of the functions of these parameters (e.g. cross section, angular and energy distributions). In every discrete element of the parameter space (e.g. every energy bin), the histogram of the measured data (e.g. cross sections) is built. When uncertainty information is available, the histogram is built assuming that data are uniformly[1] distributed over their uncertainty range. In the case of suspicious data, the histogram will reveal one or more discontinuities in the distribution of the data. The width of these discontinuities will give an indication of the probability for the suspicious data to be an actual mistake. In addition, the systematic detection of a discontinuity in consecutive histograms also indicates the presence of a suspicious data set.

To illustrate the method, let us consider the data displayed in Figure 4. The visual inspection of the data plotted on the left hand side reveals a suspicious data set a factor of 10 too low as compared to other mutually consistent measurements (See the red circles in Figure 4-left). The discretization of the X and Y axis in a constant 19 x 19 grid is shown on the right hand side, where data are displayed using an optimal scale (close to log-log scale in this case). Histograms for each of the first 16 X-slices are shown in Figure 5. The distribution of the data is continuous on the first four histograms (labelled X1 to X4 in Figure 5). A first discontinuity is seen in histogram number 5, where two clusters of data are observed. The first one in the grid-element (X5,Y11) and the second ones in the grid-elements (X5,Y14) to (X5,Y15). The method tentatively determines which data set is the most suspicious according to the number of measurements present in these clusters. On Figure 4 (right), the lower grid-element (X5,Y11) contains two measurements only and is therefore the most suspicious. Other similar discontinuities are observed in histograms X10 to X16 of Figure 5 and corresponding elements of the grid are highlighted in Figure 4.

[1] This is an approximation compared to the use of a normal distribution, but it does not affect the detection of suspicious data and makes the procedure easier to implement.

Figure 4: Experimental data available in the EXFOR database for the ^{nat}Lu(n,γ) cross section (left). The plot on the right hand side displays the same data with an optimal scale. The grid shows the discretization of the plan. Elements containing suspicious data are highlighted in orange, whereas elements coloured in yellow illustrate discontinuities in the data.

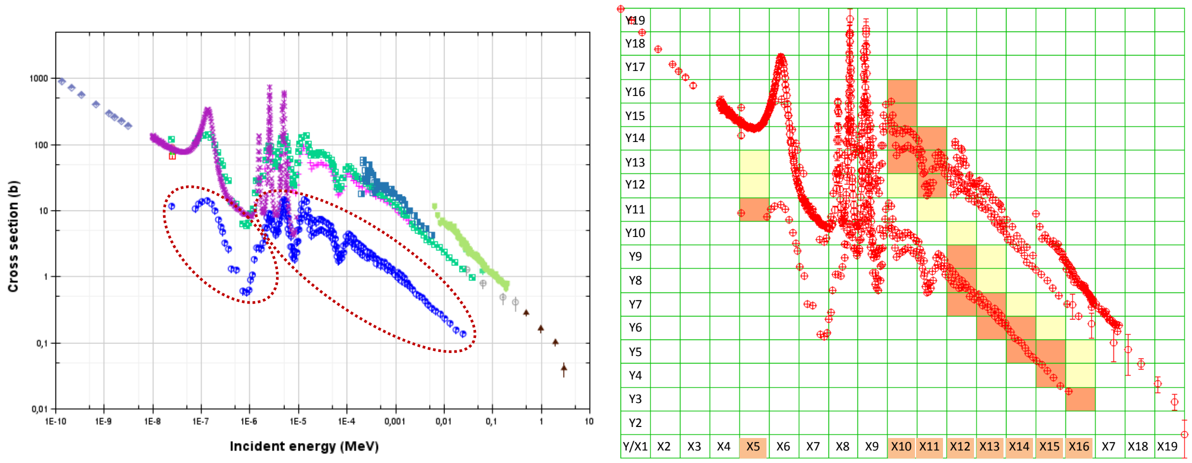
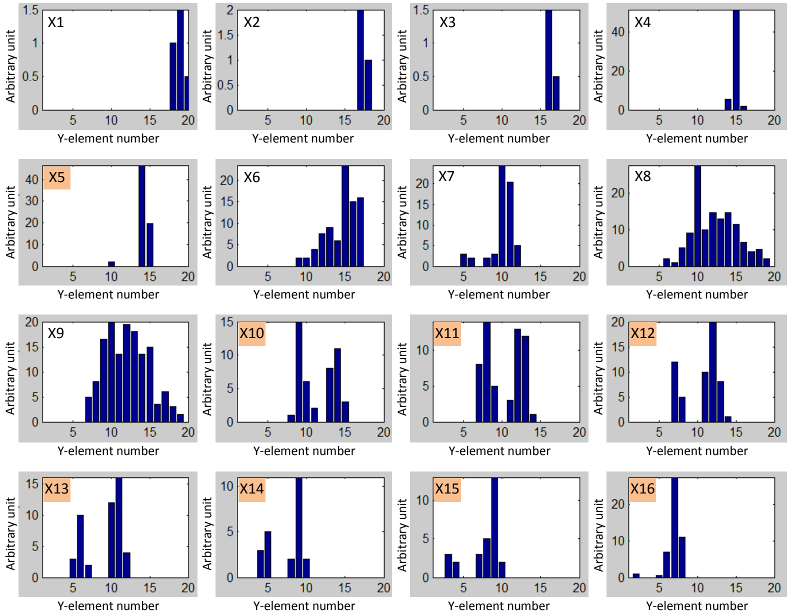


Figure 5: Distribution of the data along the Y axis in the first sixteen slices of the X axis (Same data as in Figure 4).



An estimation of the probability for suspicious data to be erroneous is given for every X-slice of the grid by the width of the discontinuity, which corresponds to the distance between discrepant clusters, e.g. 2 grid-elements for the data in the X5 slice in Figures 4 and 5. Note that 17 grid-elements is the largest possible value for a discontinuity. Another indication of the presence of an anomaly is given by the number of times a suspicious data set is detected on a plot, e.g. 8 times in Figure 4. Both estimations are important. In this example, the second indication is the most useful since discrepant data are rather close (the width of the discontinuity is only 1 or 2 grid-elements). However, the first estimation is the main indication to be used in most cases, especially to detect an isolated anomaly. For a 19 x 19 discretization of the plan, a sensitivity study has shown that the probability for a suspicious data to be an anomaly was significant when one of these two estimations was higher than (or equal to) 3. In any case, this procedure is only a tool for the expert, who alone makes the final decision.

Validation of the method

About 28 000 EXFOR data sets representing a given quantity (e.g. cross section) of a given reaction (e.g. U-235 fission) were identified assuming more than 5 independent measurements or more than 10 numerical values for any data set. For validation purposes, 55 data sets were selected randomly and 58 others manually, according to their importance or because they contained anomalies to be detected. The method was validated over these 113 cases. A number of false alarms (< 5%) was observed in cases with few scattered data measured in a limited energy range due to a too fine discretization for the grid. In order to minimise the false alarm rate, the use of an adapting mesh could be considered. Nevertheless, the correct-answer rate of the method was better than 95% and all anomalies were correctly identified.

Conclusions

A first attempt to analyze the content of the EXFOR database was performed in the framework of the NEA WPEC Subgroup 30 (NEA, 2011). In continuation of these activities, the NEA Data Bank is initiating further studies to implement Sg30 recommendations with the aim of ensuring the highest level of quality for the contents of its database. The statistical methods developed to check the consistency of the EXFOR database have been proved efficient and robust with a small false alarm rate. These methods are now being applied to the whole EXFOR database to detect remaining anomalies and could be used to verify new data before inclusion in the database. These efforts will contribute to further improve the quality of the EXFOR database and to ensure that its invaluable contents can be used in modern nuclear data evaluation work. It is expected that these methods should apply to other numerical databases with the same success.

References

NEA (2010), JANIS 3, A Java-based Nuclear Data Display Program, DVD, OECD NEA, Paris. More information at www.oecd-nea.org/janis.

NEA (2011), Quality improvement of the EXFOR database, International Evaluation Co-operation, Volume 30, OECD NEA, Paris.