IAEA
International Atomic Energy Agency
Atoms for Peace and Development

Concept for new database

Shin Okumura
Nuclear Data Services Unit, Nuclear Data Section, IAEA

WPEC Subgroup 50 (SG50) WebEx meeting, 14-15 September 2020

The Database Needs

The use of advanced machine learning algorithms or data mining in science is ongoing and thriving in many fields:

- Improving the throughput of experimentation
  Innovation: materials, chemical substance, and drug discovery
  Automation: automatically processable, analysable, exploitable, amenable to data mining
  Prediction: predict properties without experiments

  } worldwide

- Improving the throughput of nuclear data evaluation processes
  Innovation: unearth efficient reactions
  Automation: machine-readable format
  Prediction: model development based on nuclear physics curated experimental information

  } SG50

Unified, may be not unique, machine-readable format is strongly desired.

Other Experimental Databases; EXFOR is not the only one

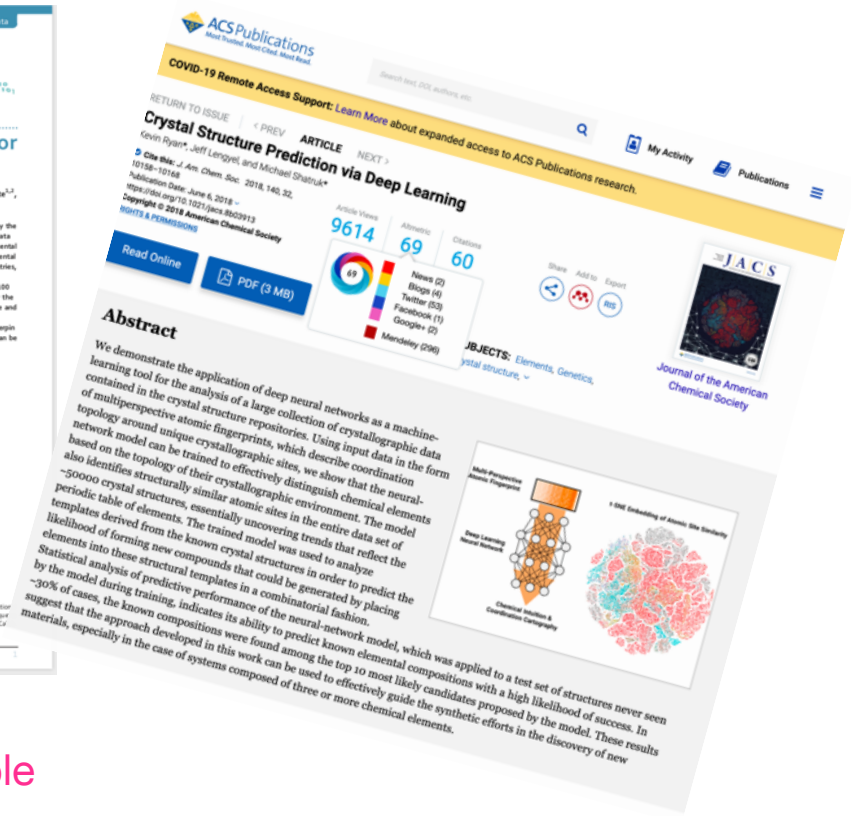
Database	Established	Number of records	User submission	Web retrival	API
Repository for publication-related High-Energy Physics data (HepData Project) https://www.hepdata.net/	(past four decades)	86,816 data tables	Yes	Yes	No
High Throughput Experimental Materials (HTEM) Database https://htem.nrel.gov/	Collected over seven years	140,000 samples	Yes (YAML format)	Yes	Yes (JSON)
Worldwide Protein Data Bank http://www.wwpdb.org/	1971	177,212 structures (2000-)	Yes	Yes	No (rsync, ftp)
The Cambridge Structural Database (CSD) https://www.ccdc.cam.ac.uk/structures/	1965	1,086,719 structures	Yes (CIF format)	Yes Limited to members	No

and more..

- Journal publishers encouraged author(s) to submit the experimental (numerical) data before manuscript submission.
- User (experimentalist) can submit the data by themselves using the unified format.
- Such unified formats can be directly used in many applications.

Machine Learning and Experimental Database

- Many attempts…e.g.
  - How to represent crystal structures for machine learning: Towards fast prediction of electronic properties [1]
  - Crystal Structure Prediction via Deep Learning [2]
  - An open experimental database for exploring inorganic materials [3]

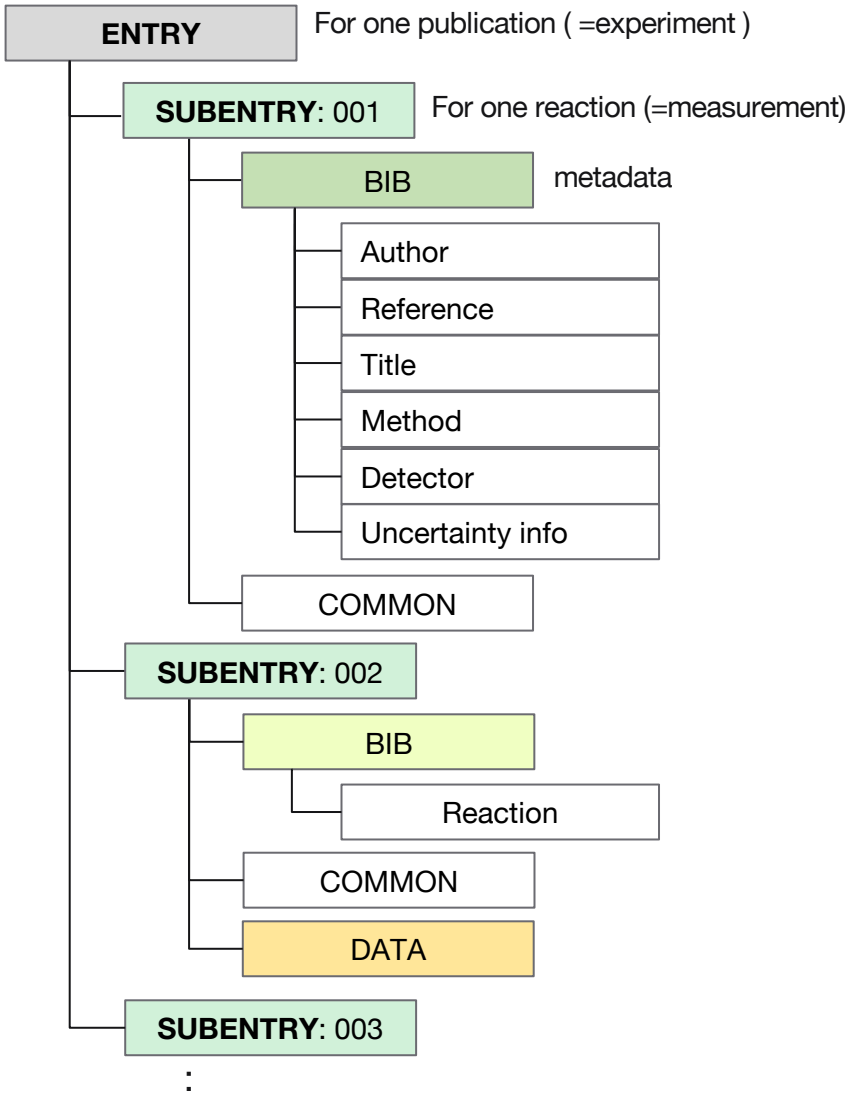


- A large data set with a unified and machine-readable data format has a good compatibility to machine learning.
- The EXFOR database covers much diverse data.

[1] K. T. Schütt, et al., Phys. Rev. B 89, 205118 (2014).
[2] K. Ryan et al., *J. Am. Chem. Soc.*, 140, 32, 10158–10168 (2018).
[3] A. Zakutayev et al., Scientific Data, 5, 180053 (2018).

EXFOR is the EXchange FORmat

```
ENTRY          13388      891220              20050926        0000
SUBENT      13388001      891220              20050926        0000
BIB                9          11
INSTITUTE  (1USALAS)
REFERENCE  (J,PR,99,730,5508)
AUTHOR     (A.C.WAHL)
TITLE      FISSION OF U-235 BY 14-MEV NEUTRONS: NUCLEAR CHARGE
            DISTRIBUTIONS AND YIELD FINE STRUCTURE
METHOD     (RCHEM)
DETECTOR   (PROPC)
ERR-ANALYS (DATA-ERR) STANDARD DEVIATION OF THE AVERAGE OF THE
            RESULTS
STATUS     (RIDER)
HISTORY    (891212C) VM
ENDBIB            11
NOCOMMON           0           0
ENDSUBENT         14
SUBENT      13388003      891220              20050926        0000
BIB                6          12
REACTION   (92-U-235(N,F)ELEM/MASS,IND,FY)
FACILITY   (CCW)
INC-SOURCE (D-T)
MONITOR    ((MONIT1)92-U-235(N,F)42-MO-99,CUM,FY)
           ((MONIT2)92-U-235(N,F)42-MO-99,CUM,FY,,SPA)
           ((MONIT3)92-U-235(N,F)ELEM/MASS,CUM,FY,,SPA)
DECAY-DATA ((1.)53-I-131,8.07D,B-)
           ((2.)53-I-132,2.3HR,B-)
           ((3.)53-I-133,20.9HR,B-)
           ((4.)53-I-134,52.5MIN,B-)
           ((5.)53-I-135,6.7HR,B-)
STATUS     (DEP,13388002)
ENDBIB            12
COMMON             4           3
EN         EN-NRM     MONIT1     MONIT2
MEV        EV         PC/FIS     PC/FIS
 14.        0.0253       5.17       6.14
ENDCOMMON          3
DATA               5           5
MASS       ELEMENT    DATA       MONIT3     DECAY-FLAG
NO-DIM     NO-DIM     PC/FIS     PC/FIS     NO-DIM
131.        53.         4.47       3.02      1.
132.        53.         5.03       4.49      2.
133.        53.         5.36       6.62      3.
134.        53.         5.20       8.00      4.
135.        53.         4.35       6.31      5.
ENDDATA            7
```



To use DATA in subentry:001, it needs to read metadata from many places.

Fluctuations in EXFOR

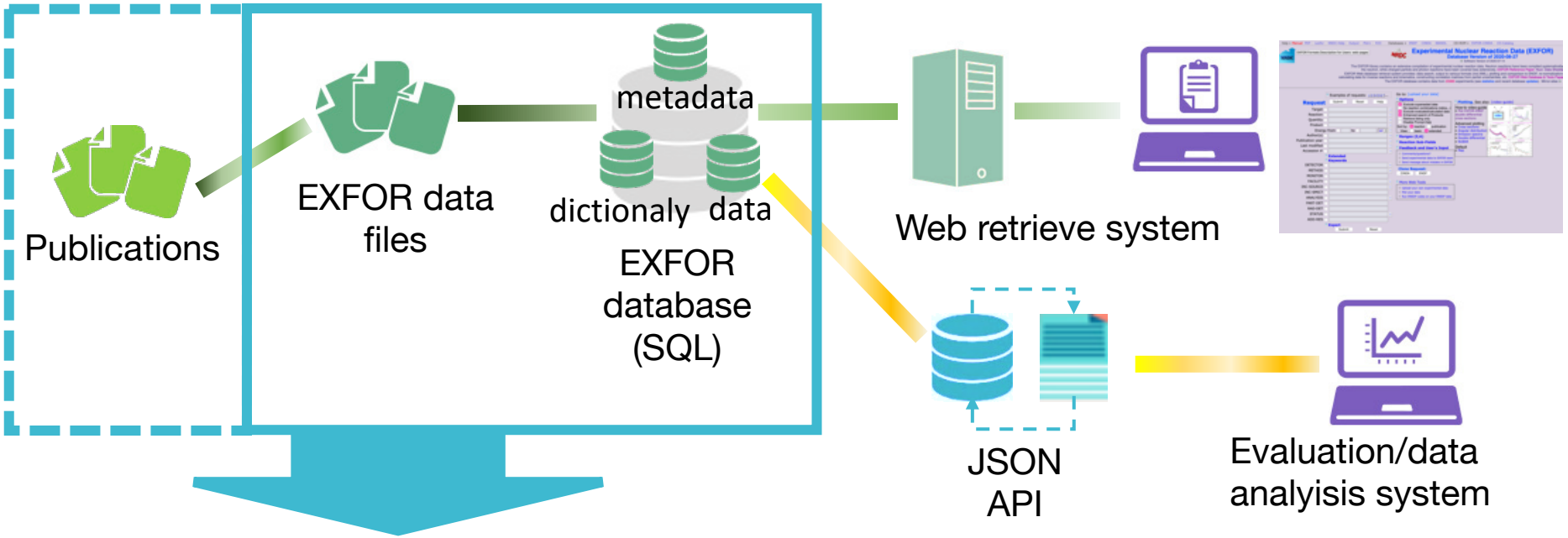
- Metadata
  - "Year" format in Reference

Case	Example	Number of data
Recommended format	J,YK,1982,(2/46),9,1982	18,033
2 digit	J,JCP,17,653,49	1,673
4digit YYMM	J,PR/C,4,723,7109	2,656
6digit not YYYYMM	J,NP/A,329,1,141,791008	56
6digit with YYYYMM	J,NSE,66,24,197804	5,798

  - Author's name
    - Augustyniak --- Augustynyak
    - Brinkmoeller --- Brinkmoller
  - Incompleteness of information as a database
    - One of the redundant information (PART-DET (G) and DETECTOR (HPGE)) are deleted during compilation.
  - Much free text
- Data
  - Uncertainty information in ERR-* fields
    - ERR-1 -- ERR-10 not always the same meaning or interpretation
  - Based on the old standards, decay data,..etc

- Some of these fluctuations are "**absorbed**" by IAEA's SQL & web retrieval system.
- However, more automatically processable and analysable format (scheme) is required.

Three Layers of Database



Layer 1

- Build the database from automatically readable format → Candidate for a novel EXFOR format

Layer 2

- Adopt objective and subjective corrections: updating standards, flagging outliers..etc

Layer 3

- Curated evaluation database

Summary

- The use of advanced machine learning algorithms require the unified and curated data.

- EXFOR, EXchange FORmat, is sufficient for compilation and data exchange but not at the level of automated use by a machine learning system or evaluation system. Such machine readability would be useful for improving the throughput of nuclear data evaluation processes.

- Layer 1 database may be implemented from the current EXFOR system which is assisted by SQL database and even by the information from original publisher or PDF.

- Layer 1 data format (scheme) would be useful for the idea of the next EXchange FORmat as well.

Thank you for your attention!

IAEA
International Atomic Energy Agency
Atoms for Peace and Development