

Webble World を用いた新たな核データベース利用システム に向けて

Toward a new utilization system of nuclear databases with the Webble World

北海道大学知識メディア・ラボラトリー

大木 平、椿原 康介

北海道大学大学院理学研究院

合川 正幸、加藤 幾芳

OOGI Taira, TSUBAKIHARA Kohsuke

Meme Media Laboratory, Hokkaido University

AIKAWA Masayuki, KATŌ Kiyoshi

Faculty of Science, Hokkaido University

Abstract

We have been preparing for a new utilization system of nuclear databases with the Webble World developed by Meme Media Laboratory at Hokkaido University. With this preparation, we have also been developing an extended nuclear reaction database formatted with XML. In this article, we report recent status of R&D in these businesses.

1 はじめに

荷電粒子核反応データファイル (Nuclear Reaction Data File: NRDF) は、1980 年代から日本国内で生産された荷電粒子入射実験を採録するために開発されてきたフォーマット、及びそれを用いたデータベースの呼称であり、現在はその採録範囲を広げ、光核反応実験の数値データも蓄積されてきている。ただし、NRDF の先進的なデータ構造に比べ、旧来からの一般的なテキストファイルを用いた採録形態は近年の様々なコンピュータ環境の進展から見るとかなり古く、利用しにくいものになってきている。こうした核反応データベースを取り巻く問題意識は国際原子力機関 (International Atomic Energy Agency: IAEA) と、協力関係にある核データセンター網が整備している EXFOR(EXchange FORmat) においても共通のものとなっている。

また、これまでに JCPRG においては、Web 上で動作する核反応データ採録用のエディタである、HENDEL 及び、それを利用した採録活動が大きな成果を収めてきた。HENDEL は採録作業時に NRDF フォーマットと EXFOR フォーマットのデータを同時に生成できるという大きな特長をもつ。しかし

ながら、核データ利用環境のさらなる向上のための改善の要望も少なくない。要望は、大きく (1) エディタそのものの機能の向上、(2) エディタ機能を含めた、データ入力・利用環境の充実という 2 つに分類される。これらの要望を満たした新たな核データベース利用システムの構築のためには、より使いやすいユーザーインターフェースの開発や、それに伴う最適なデータフォーマットの構築など、現行システムを大きく変更する必要がある。

このような状況を踏まえ、我々は現在、国際的な普及を目指した「インタラクティブエディタ」の開発および、原子核反応データの新フォーマットの構築を目指し活動を行っている。特に、新たなツール「Webble World」を導入することにより、インタラクティブな利用環境、つまり、集積されているデータを我々が提供しようとするフォーマットで引き出すだけでなく、検索者が Web 上に用意されたツールを組み合わせ、そのデータを様々な利用することが出来る環境の整備を実現しようと考えている。

加えて、データ利用に関する観点から、新しく開発されるデータ利用環境においては、Web 上での情報技術と親和性の高いデータベースが用いられるべきである。XML(eXtended Markup Language) は例えば HTML のようなマークアップ言語を定義するためのメタ言語であり、XML の記述法に則することでデータベースの要素を記述するのに必要なマークアップ言語を定義できる。特に “Well-formed”、つまり入れ子構造に矛盾のないことを要求するため、そうした入れ子構造、換言すると樹状構造を持ったデータを表現するのに適している。また、XML の持つ各種情報技術との互換性から、インターネット上での検索やその結果の望む形への変換など、様々な形での利用に優れるという特徴を持つ。

本稿では、インタラクティブエディタの開発についての現在までの活動を報告するとともに、これから XML を用いて今まで JCPRG が整備・運用してきた DB フォーマットである NRDF をどの様に拡張するべきか、またどの様な準備が必要かといった点について議論する。

2 新たなツールの導入 –Webble World–

2.1 Webble World とは

Webble World は、北海道大学知識メディア・ラボラトリーが開発を続けてきた IntelligentPad の後継にあたるソフトウェアである。Webble World においては、アプリケーションを個別の機能をもった「部品」の集合体として構成し、部品の追加や削除、改良をしながら多種多様なアプリケーションを作り出すことが可能である。

まず、このような Webble World が開発されるに至った、IntelligentPad および Webble World の設計思想について簡単に紹介したい。基本的な思想として、知識を「遺伝子」(meme: ミーム、模倣子) ととらえ、知識はコピー (自己複製)、人による再編集 (遺伝子組み換え・突然変異)、人による評価 (自然淘汰) を経ながら進化していくものと考え。そして、インターネットを通じた知識の流通や再利用を行うことで、知識の進化を目指す。このような思想のもと、作られたシステムが IntelligentPad である。IntelligentPad では、IntelligentPad 内で定義されたメディア (Pad) を組み合わせて機能合成を行うことによって、新たな機能を持った Pad の定義が可能となる。後継である Webble World では、Microsoft 社が開発したウェブブラウザ用のプラグインである Silverlight を導入することにより、Web ブラウザ上での動作が可能となり、ユーザにとってより使いやすいツールとなった。つまり、これらの環境の中では取り扱われている情報だけではなく、個々の Pad(Webble World においては Webble) で表現されるもの自身も、先に述べたミームに相当するものであると言える。こうした「インターネットを通じた知識の流通や再利用を行うことで、知識の進化を目指す」という考え方は、

拡張性が求められる学術データベースやその利用システムのあるべき姿と合致しており、原子核反応データベース利用システムに導入するツールとしてふさわしいものといえる。

2.2 Webble World の活用法

Webble World を導入することにより、ユーザーがカスタマイズ可能な、インタラクティブな環境を実現することが可能となる。例えば、エディタ機能、検索機能、PDF ビューワー機能、グラフ作成機能を独立な「Webble」として開発することで、ユーザーが使いたい機能だけを選択してブラウザ上に表示し、利用することなどができる。さらに Webble World では、Webble 間をつなぐことができる。この機能を利用し、例えばエディタ Webble と検索 Webble をつなげることにより、採録中に既存のデータ内を検索するようなことができるだろう。これらは一例であり、今後も有効な利用法について活発に議論していくことが望まれる。

3 システム概要

3.1 サーバ全体の概要

図 1 に、インタラクティブエディタのために構築するサーバ全体の概念図を示す。各部分は MVC 型設計により解釈される。まず View として、Silverlight アプリケーションである Webble World がある。Silverlight アプリケーションはデータベースに直接アクセスすることができないため、Controller として Microsoft の別の技術である WCF サービスをおくことが必要となる。最後に、Model として核データベースをおく。新たに開発する部分は、Silverlight アプリケーションと WCF サービスである。開発には Microsoft の統合開発環境 Visual Studio 2010 Professional を用いる。開発言語は C# とする。

3.2 開発状況

開発の第一歩として、Silverlight を用いた簡単なエディタの開発を行っている。この開発経験により、Silverlight アプリケーション開発の基礎となる、XAML (eXtended Application Markup Language: XML ベースのオブジェクトやプロパティ、あるいはそれらの関係などを定義するために用いられる言語) によるインターフェースの定義や C# によるロジックの記述法を習得しつつある。また、WCF サービス開発のために、WCF のプログラミングモデルとそれを用いた具体的な Web サービスの実装法について学習している。

4 新 XML フォーマットの策定に向けて

XML について詳しい解説は過去の年次報告 [1] に譲り、本稿では要点だけを示す。XML は 1998 年に World Wide Web Consortium(W3C) により勧告された比較的新しい言語であり、様々な情報をその「意味・種類」と「実際の内容」という二つに分けてテキストにて記述するための比較的平易な方法である。つまり、マークアップ言語を組み立てる際に、従うべき規則と考えることができ、実際に XML の表記法を取り入れた XHTML が最も著名なマークアップ言語である HTML と同様に W3C によって勧告されており、他の XML 文章を埋め込むといった高度な Web 上でのデータ利用法が提

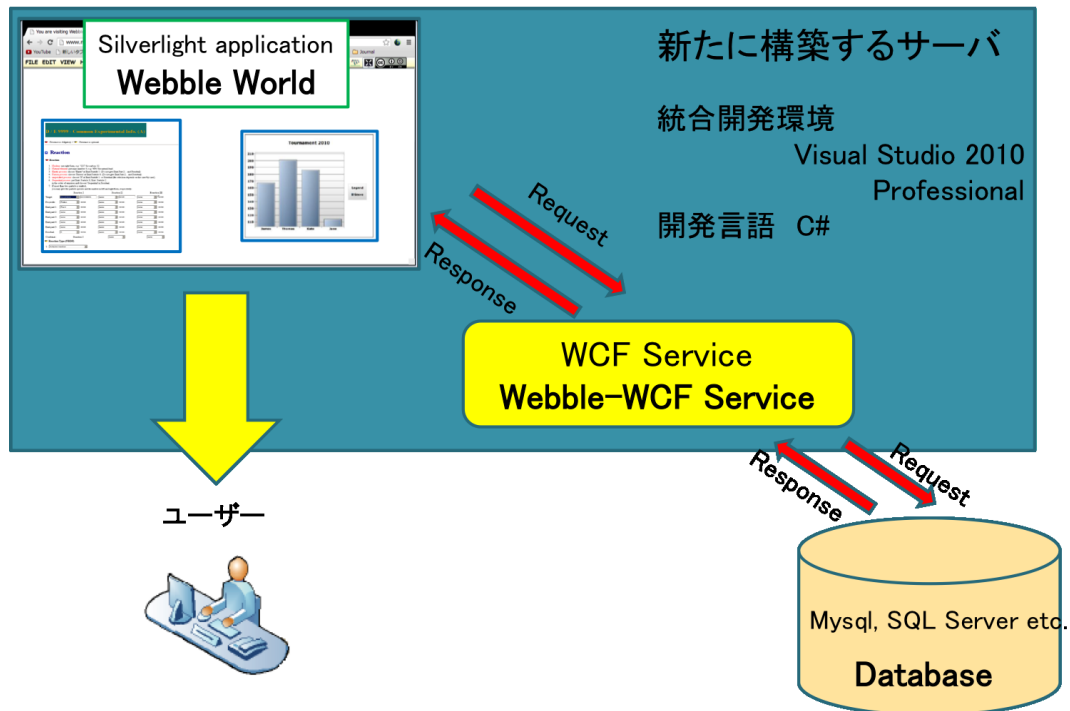


図 1: サーバ全体の概念図

案されている。よって、XML に基づき、採録すべき情報にラベル付けを行い、計算機上での処理に適した形を構成することで、インターネット上での利用・応用が利く文書データを作ることが出来る。加えて、先に述べた “Well-formed” である階層構造の恩恵により、従来のリレーショナル・データベース (RDB) では表現する事が難しい、データ間の包含関係を自然に定義することができる。また、データベース運用上の利点もある。RDB では、一度データテーブルの構造が決まると変更が難しいため、まずしっかりとした構造を決めてから、データベースを利用するアプリケーションを開発する必要がある。一方 XML を使ったデータベース (XML-DB) では、アプリケーションでの管理上では全体の検索テーブルを構築し直す必要はなく、思ったところに自由に内容を追加することが可能である。また、検索結果を XML 化したテキストとして取り出せば、他の XML エディタで編集することも、適当なツールによって望ましい形式に変換し、読みやすい文章を構成することも可能であろう。一方で、MySQL など RDB を取り扱う計算機環境は長い開発の歴史を持ち、ソフトウェアとしては安定したパフォーマンスが望めるが、XML-DB を取り扱う環境は開発の途中にあるものも多く、検索効率としては RDB にどうしても劣ることになる。以上の内容から、RDB は用途・採録対象がしっかりと決まった (正規化が済んだ)、この先変更する可能性の低いデータベースに用い、学術的データベースのような拡張性を持たせるべきデータベースには XML-DB を用いると、開発者や採録者、そして利用者の全員にとって恩恵が大きい。つまり、RDB は定型文書、XML-DB は部分的に定型文書を含む非定型文書 (半定型文書) を再録するのに向いているといえる。ただし、XML-DB は RDB に比べ、データの転送量や階層構造の深さなどからくるサーバ側への負担も大きくなるため、データベース上の情報を全て XML 化すれば良いとは限らない。その点で、現行の NRDF を XML 形式に単純に置き換えることは、タグを付与する規則を完備すれば、一義的には難しいものではない。但し、例えば将来的に、核データ採録用のエディタを開発する際など、核データ開発に関するソフトウェア開

表 1: データベースと型

ネイティブ型	BaseX(GPL、非商用)、Cyber Luxeon(サイバーテック、商用)、 EsTerraXSS(メディアフュージョン、商用)、 NeoCore XMS(米 Xpiori →サイバーテック、商用・非商用あり)、 TX1(東芝ソリューション、商用)、Tamino(独 Software AG、商用) など
ハイブリッド型	DB2(Ver.9 から、米 IBM、商用・非商用あり)、 Oracle XML DB(米 Oracle、商用) など

発を考えた時、必ずしも単純な置き換えたものが効果的ではないことが予想される。また、NRDF、EXFOR 共々、各種実験施設や物理量、その他の実験情報について冗長化を避けるためのコード化を行っている。これは計算機の環境が現在ほど潤沢ではない時代の名残りと言えるが、採録者や、その採録された結果を確認する実験者、また核反応データを検索し実際に利用する側からすれば、簡単に理解することが出来るとは言い難いものになりつつある。さらに、理化学研究所のもつ RIBF など、今までの実験とは異なり不安定核の領域に大きく足を踏み入れる実験を行うことが出来る施設が増えてきたため、今後は現状の NRDF、EXFOR 形式での採録が難しい核データが増えてくることもある。それに対して、なるべく拡張性の広いフォーマットを整備し、新しいデータに対応しうる余力を残してあるようなものをベースに採録活動を行っていくべきであると考えられる。

これらの状況を鑑みて、我々は今年度、NRDF、EXFOR を基盤とした新たなデータベース形式開発を目指し、議論および開発活動を行ってきた。特に NRDF は先に述べたように、核データの採録という特定の目的ではあったが、フォーマットとして XML と同様の汎用性を本来持っているため、そのコンセプトを受け継ぐような形式を確立する事を目的としている。

5 本年度における XML フォーマットに関する議論について

5.1 使用する DB ソフトウェアの候補

現在、XMLDB における検索に関しては、方式として XPath、XQuery で行う XML:DB 規格が W3C によって策定され、これらを用いた検索が様々な DB ソフトウェア上で実装されている。RDB における SQL にあたる XQuery を用いることにより、DB ソフトウェア固有の検索命令を修得する必要がなく、導入する環境やそのパフォーマンスで選ぶことが出来る。また、同じ XML データに対応したソフトウェアの中でも、大まかに分けて XML データの格納を専門に作られたネイティブ型と、元々あった RDB 向けのソフトウェアに XML データ専用の機能を加え、両者を統合して扱えるようにしたハイブリッド型というバリエーションがあるため、いかに XML 形式に拡張性があるとはいえ、大まかな方向性に関する議論、例えば、先に述べたように定型化されている部分を抜き出して RDB 型とし、残りの部分を XML 型とするか、全ての内容を XML 型としてデータベースにするかといった観点で使用する DB の選択について議論する必要がある。ここに例としてネイティブ型、ハイブリッド型のそれぞれに属する DB を挙げる。このように、ネイティブ型、ハイブリッド型それぞれに商用・非商用の別を含め、さまざまなソリューションが考えられ、それぞれの DB ソフトウェアで得意とする処理も異なる。ゆえに、データベース仕様の方向性が固まり次第、適宜、それを動作させる DB ソフトウェアも選ぶことが重要になるであろう。

5.2 現在用いられているデータ構造の解析

近年、我々JCPRGは新XMLフォーマットのテストとして、文献情報や、実験のセットアップに関する情報に関する部分をXML化する試みを行った。しかし、本来、こういった文献情報自体は、十分RDBで採録しうる定型化されたものである。ゆえに、XMLフォーマットを議論する上で、(1)完全に定型化しうる部分はないか(2)ネイティブ型、ハイブリッド型のどちらが有効なのかといった点について議論する必要があるといえる。また今回の場合、単一の文献をReferenceとして採録する今までの方式に則れば、RDBに収録される形になるが、複数の文献をReferenceとして持ちうるように、例えば、博士論文やProceedingsでの登録と、本論文として出版された際の採録を併記できるように拡張することを考えれば、この部分もXML化しておくとも便利である。このように、単純に拡張性だけを念頭におけば、全ての内容をXML化するのが簡便であるともいえる。例に挙げた文献の場合、その登録だけを考える際、登録した文献に通し番号を付与することさえXMLにおいて必要はなく、同じ場所に現れる、同じタグの内容を出てくる順番によって検索でフィルタを掛けることが出来るからである。勿論、この文献情報をデータベース上で他の場所から引用するつもりがあるなら話は変わり、何らかの通し番号を付与するべきであろう。

ここで、今現在、我々JCPRGが核データ活動に用いている2つのデータベース、NRDFとEXFORにおいて用いられているフォーマットの分析を行なってみたい。ここで構造を記述するのにUMLに則り描かれた、クラス図を用いた。また、簡便のため、多少内部の詳細なデータ構造に関しては省いている部分がある。

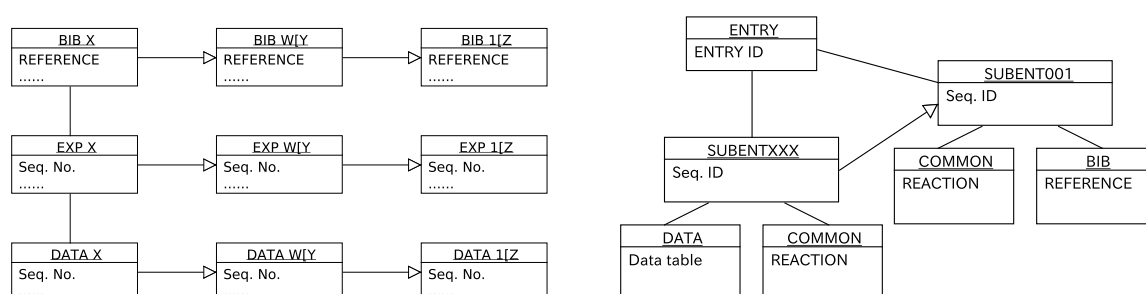


図 2: NRDF、EXFOR に対応するデータ構造のクラス図

まず、左の図がNRDFのデータ構造を模式化したものになる。大きく分けると書誌情報を含んだBIB要素、実験データを計測した際の状況を示すEXP要素、そして実際のデータが入力されるDATA要素の3つからなる。BIB、EXP、DATA要素にはそれぞれの中で部分・全体を問わず共通のデータがあれば、まとめて一つの要素として採録することが出来、それらは全体共通を表すBIB1[Z、EXP1[Z、DATA1[Zや、部分共通を表すBIBX[Y、EXPX[Y、DATAX[Yといったシートに含まれている。ここで、各BIB、EXP、DATAの後に示されている部分が、共通部分として記述される範囲として定義されており、全体共通 → 部分共通 → 個別として、各内容を各BIB、EXPやDATAにて継承し、各段階での共通情報を追加できるようになっている。また、同じエントリー番号を含むBIB、EXP、DATA要素間には対応関係があり、あわせてZ個の実験データをDATA要素として採録していることになる。このデータ構造自体は非常にわかりやすく、複数採録しうるデータに対して通し番号を設けておくなどの参照周りの仕組みもしっかりしているように思える。また、包含関係(言い換えれば階層構造)がしっかりしているため、そのままXML化すること自体は難しくなく見える。一方、右に示したものがEXFOR形式におけるデータ構造を模式化したものになる。この形式では、まず一

番大きな ENTRY という枠の中に、複数の SUBENT が含まれる形になっている。ただし、SUBENT 内で SUBENT001 の果たす役割は他のものと異なり、先に NRDF で述べたような文献情報と共通部分をまとめたものになる。そして次の通し番号である SUBENT002 から実際のデータを採録するような形になっている。この同じ名前かつ通し番号で管理されるものが実は違う役割を担っているという構造は、データフォーマットになれない人にとっては、可読性を多分に下げているように感じられる。また、各 SUBENT はそれぞれ BIB、COMMON、DATA という 3 つを持つことになるが、これも先に述べた SUBENT001 とその他の SUBENT では多少役割が異なる。例えば、BIB は SUBENT001 においては必ず REFERENCE を一つ含む形になるが、その他の SUBENT においては REACTION を必ず一つ以上含む形になり、同じ名前に見えて必要としているものが異なることになる。ここで、SUBENT001 を特徴付けているのは実は REFERENCE であり、また他の SUBENT を特徴付けているのは REACTION と見ることが出来るため、データ構造として BIB よりも一つ上の情報として取り扱われるべき内容のように思える。また BIB という名前から連想されるいわゆる「書誌情報」だけではなく、実験の細部に関しても内容に含まれており、検索したユーザに多少の混乱を与えるように思う。また、COMMON においては幾つかの物理量・単位・数値が与えられるようだが、SUBENT001 に全体に共通な情報が与えられ、それを継承した他の SUBENT にて独立な情報が定義される構造において、今ひとつ COMMON という名前から果たすべき役割と異なっているように思える。全体的に見て、EXFOR フォーマットを単純に XML 形式に書き換える難しさを感じる結果となった。

5.3 データ構造の候補

前節までに NRDF フォーマットにおいて、XML 形式への変換が比較的簡単であろうことについて論じた。ただし、NRDF に関して先ほどの図でモデル化したように著者情報、所属情報は BIB に直接結びつく (BIB 要素が INST-ATH、ATH 要素を「持つ」関係になっている) が、実は著者自体も一つの独立したデータベースを作りうる。また、場合によっては、大きな研究グループ (例: 重イオン反応における STAR、PHENIX、PHOBOS、BRAHMS といった検出システム毎の研究グループ) も検索対象となる可能性がある。そうならば、エンタリー全体が一つの大きなレコードとして作られ、各要素自体はそれぞれのデータベースに採録されるという構造も考えうる。例えば、上の図は一つの構

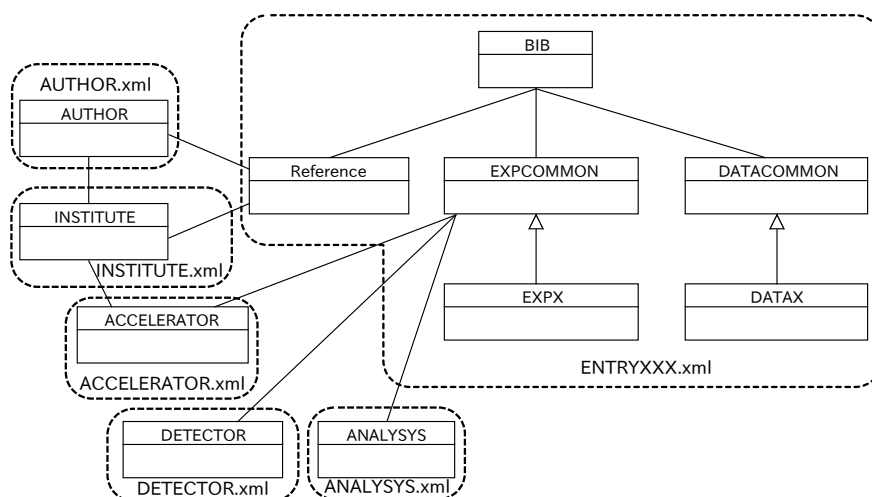


図 3: 新 XML フォーマットの構成案

成案である。この案では、代表的に他のエントリーにおいても複数回現れうる情報を切り出して、別の XML ファイル、つまり別のデータベースに格納している。お互いに参照しあう関係がわかっているならばこういった情報の切り分けも可能になり、結果的に負荷の分散などに繋がるかもしれない。よって、そういったデータベースの実装を行う前に、どのような検索が行われうるか利用者の観点から調査を行い、データベース自体の構造についてしっかりと議論を行う必要が大いにある。

また、データベースが別になっていれば、検索における負荷軽減やデータベース自体のスリム化を行う事が出来る可能性があるだけでなく、例えばエディタ開発に置いてても有用になるかもしれない。データベース上で対応する部分がしっかりと抜き出されていれば、データ検索における負荷の問題を解決して、エディタ自体にコード表を持つことなく現在登録されている中から各情報を選択するという事も出来るかもしれない。もし可能であれば、採録者に依存する各記述のふらつきを抑制することが出来るだろう。また、タイプミス等による初等的なエラーがかなり減ることも期待できる。こうした点はエディタを開発する上での設計に大きく関係するため、エディタ開発に先駆けて、データベースの概形を決定することが強く望まれる。そのためにも仔細な点にとらわれず、より大きな枠でのデータベース構造の議論を行い、深めていく必要があると考えられる。

6 おわりに

本稿では、新たな核データベース利用システムの開発を始めるに至った経緯について述べた後、新たに導入するツールである Webble World について簡単に紹介し、Webble World を用いた新システムの概要と現在までの取り組みを述べた。システム全体の開発が完了するまでにはまだ時間がかかると思われるが、開発の意義は大きく、今後もステップバイステップで取り組んでいきたい。

加えて、これから XML を用いて今まで JCPRG が整備・運用してきた DB フォーマットである NRDF をどの様に拡張するべきか、またどの様な準備が必要かといった点について議論してきた。その中で XML フォーマットへ拡張する意義や候補となりうる DB について紹介し、現在用いられている NRDF、EXFOR データベースフォーマットについてクラス図を用いてその内容を吟味し、NRDF が XML 的な階層構造をしていることを顕にした。また、EXFOR 形式においてはデータベース中の役割が一貫していないことを指摘し、XML フォーマットへの書き換えが単純ではないことを示した。最後に技術的な構成として、データベースに採録される情報の分割可能性について簡単な例を示し、問題提起を行った。総じて、データベースの構成を考える上で、現在あるフォーマットの翻訳という形で行うのも一つの進め方として認めるが、全体的に (1) どの情報が各要素の中で主たる検索対象となりうるのか (2) その上で、各要素の独立・従属関係をクラス図などで明らかにし、独立な要素としてデータベースに格納するべきものはないか調べるなど、実作業に入る前に全体の構成を考えたメタ的な考察を行う必要があると提言したい。

今後はより具体的に各核反応データの採録方法やその形式について、どのような仕様が求められるか更に R&D を進めると共に、具体的なデータベース上での利用やエディタ開発との連携について詳細を詰めていく必要がある。またこの議論を進める上で、従来の NRDF において意識されてきた事などについて、詳細な理解を進める必要がある。今後もその点について研究・議論を進め、新インタラクティブエディタや、新フォーマットの策定に活かしていきたい。

参考文献

- [1] 松本琢磨、千葉正喜、加藤幾芳, 「NRDF の XML 形式への変換 (Transformation of NRDF to an XML Format)」, 荷電粒子反応データファイル年次報告 No.23, 2 (2009) 及びその中で紹介された参考文献