ＮＲＤＦに登録する入力データの原著者によるチェックシステムの考察 *)

北星学園大学　　片山　敏之
北海道大学理学部　　加藤　幾芳

1．はじめに

データベースに登録されるデータ内容をできるだけ正確なものにすることは、データ量とともに、データベースの有用性を左右する重要な条件である。荷電粒子核反応データファイル（ＮＲＤＦ）の場合、ＮＲＤＦの入力データ（＝コード化されたソースデータ）の採録対象となる実験データは論文を１単位としてコーディングされている。論文の大部分はレフリー付きの学術雑誌に掲載されたものである。構文チェックは当初から行われており[1]、意味チェックの試み[2]もなされている。しかし、コーディングをする人が実験データの記述に直接責任のある人（＝原著者、以下では単に「著者」と記す）ではないために、入力データの記述にあいまいさが生じることがあるという、より基本的な問題が次の課題として残されている。

ＮＲＤＦの管理運営委員会では、１９８７年度に『国内の加速器を用いて生産された荷電粒子核反応の実験データを完全に登録する』[3] という新しい方針を立て、それから今年度までの間は１９８０年以降に出版公表された論文の実験データを採集・登録している。このことは国内にいる実験データの著者に、入力データの記述について協力を依頼するための体制が作りやすくなったことを示唆する。しかし、著者に直接コーディングを依頼することは、後で述べるように現実的ではない。小論では、コーディングされた入力データの内容のチェックまたは査読を著者に依頼することを可能にするための体制およびそのために必要なチェック支援システムの計画について報告する。

以下では、現状のシステムに対する問題、提起された問題についての調査・分析、新システムの必要性の検討と目的や用途など、新システムの開発にあたって

---

*) An Analysis of the Database Referee Support System for Coded data-entry of the NRDF (Nuclear Reaction Data File)
by Toshiyuki Katayama (Hokusei-Gakuen University, Sapporo 004), Kiyoshi Kato (Hokkaido University, Sapporo 060)

の基本的な要件について考察し、そして、システム基本計画とシステム分析の順に述べている。

2．入力データの内容に関する問題

研究のひとつのまとまりを表している原著論文がＮＲＤＦのコーディングされた入力データの１単位となる。入力データのコーディング作業にはＮＲＤＦのコーディング文法と原子核実験物理学の両方の知識が必要である。現在、コーディング作業は２つの原子核実験研究所にいる理論系または実験系の若手研究者が数人で分担して行っている。しかし、コーディングをする人が実験データの著者ではないために、コーディングされた入力データの内容に関して次のような問題があることが指摘されている[4]。

(1) 論文の中にコーディングに必要な情報が記述されていない。例えば、入射ビームや測定器などの実験装置について明記されていないものが多い。
(2) 論文には記述されているが、コーディングをする人が知識を持っていないために正確にコーディングができない。例えば、測定器のセットや測定グループの略称のみが書かれていてその説明がない。

これらは実験条件とデータ解析（誤差）に関する内容である。このような問題点は論文の著者に直接コーディングを依頼できれば解決すると考えられるが、その場合も次のような困難な問題が横たわっている。
(3) ＮＲＤＦのコーディングの様式は学術雑誌のキーワード付加のように単純ではないので、コーディングの誤りや不統一な記述が増大する。
(4) 二次的に導出された物理量の取捨選択にもバラツキが入り込む。
更に、コーディングにはある程度の訓練が必要でかつ記述量も多いので、論文の著者にコーディングを依頼しても実際に協力してもらえるかどうかも問題である。従って、著者に直接コーディングを依頼することは現実的ではない。

しかし、ここでリストアップした問題はいずれも入力データの内容に関するもので、なんらかの方法で著者のチェックまたは査読があることが望ましい項目である。小論では上の問題(1)と(2)に対する方策として、チェック支援システムと著者によるチェック体制の導入を考える。

3．入力データの著者によるチェックが可能か

１９８９年末に全国の原子核実験の研究者を対象に『ＮＲＤＦに関するアンケート調査』を実施したことがある[5]。その目的のひとつは「ＮＲＤＦに入力したデータの著者によるチェックが可能かどうか」を調査することであった。まず、この点に直接に関係するアンケートの結果を表１に再録しておく。なお質問文は原文のままで、有効な回答数は１６７である。

表１　ＮＲＤＦに関するアンケート調査結果の一部[5]

問：現在、収録したDATAを著者にチェックしてもらっていませんが、チェックが必要と考えますか？

必要である	必要でない	分からない	その他	無回答
59人(35.3%)	32人(19.2%)	48人(28.7%)	12人(7.2%)	16人(9.6%)

問：収録したDATAを著者にチェックをお願いした時、ご協力してもらえますか？

協力できる	協力できない	その他	無回答
126人(75.4%)	9人(5.4%)	14人(8.4%)	18人(10.8%)

上の問の直前には、『あなたの発表した論文の実験データはＮＲＤＦに収録されていますか？』という質問に対して、『されている』が１１人(6.6%)、『されていない』が２８人(16.8%)、『分からない』が１１３人(67.7%)、『無回答』が１５人(9.0%)であった。その問いに引き続いて、『あなたの実験データが正しく収録されていますか？』に対しては、『されている』が６人、『されていない』が２人、『分からない』が１７人であった。

このアンケートの回収率は２５％と決して高くはないが、回答者の７５％以上の研究者が著者によるチェックに協力できると答えており、大変に協力的であるといえる。先に述べたように８７年度からのＮＲＤＦの登録には対象を国内で生産された荷電粒子核反応データに限定しているので、入力データの記述について著者によるチェック体制が作りやすい条件にある。その際、上のアンケートの回答からどのような要望や条件を考慮できるかが我々の課題である。大部分の原子核実験研究者はＮＲＤＦのデータベース活動を評価しているものと考えられるが、上のアンケートの結果は「多くのデータ生産者が自分のデータがＮＲＤＦに収録されているかどうかを知っておらず（67.7%)」、従って「収録データの著者チェックが必要と感じている（35.3%)」ことを示している。他方、著者チェックの必

要性については『分からない』、『必要でない』という回答も多い（28.7%＋19.2%）ので、著者チェックの方法については十分な検討が必要である。

著者による入力データのチェックは通常の論文のレフリーのように確立した制度ではないので、この「データ内容の査読」への理解と協力をえることが前提条件である。今後もこのような意見の交換、論文のデータを登録した著者名のリストや索引の公表または年報"NRDF ANNUAL REPORT"の配付など、情報提供の努力が必要と考えている。

4．著者によるチェックの手順の組み入れ体制

入力データの原著者によるチェック支援システムの導入に際して、入力データの作成作業の流れの中で、どこに著者査読用の原稿を作成するチェック支援システムを位置づけるかについて考察する。

まず、ＮＲＤＦの入力データの現在の作業手順は次のようになっている。

現在の作業手順

(1) 論文の選定
(2) 入力データの記述(データの採集とコーディング)
(3) データ入力
　テキストデータ(書誌情報や数値データ)の入力 ---> CPNDMTD.DATA(@@Dnnn)
　グラフデータのディジタイザによる入力 ------> DGTABLE.DATA(Dnnn)
　テキストデータとグラフデータの併合 ---------> CPNDMT.DATA(Dnnn)
(4) 入力データの文法チェック ----------> CPNDMON.DATA(Dnnn)
　入力データの修正
　文法チェックを通過した入力データ ----------> CPNDMT.DATA(Dnnn)
(5) 入力データの登録(ＮＲＤＦの検索用ファイルに追加)
　----------> GROUP#.NRDF
(6) 入力データの保守(バックアップ作業)

チェック支援システムの入力データ

著者に入力データのチェックを依頼する目的はデータベースとしてのデータ内容の正確さ（良質さ）を保つことである。コーディングやデータ入力作業の過程で生じる記入の誤りやタイプミスが残っていては言語道断であろうから、著者査

読用の原稿としては、文法チェックを通過した後の入力データか、または、ＮＲＤＦの検索システムが画面端末に表示するものと同じ書式の入力データが望ましい。この段階では、未だ、査読用原稿の媒体を何にするかについては問わないものとする。

データベースのユーザに見える部分は本来、わかりやすさを重視して設計されているはずであるが、ＮＲＤＦの検索結果の画面表示はラインプリンタへの印字イメージと同じであり（しかも大文字のみ)[6]、今日の観点からは読みやすいとは言えない。この画面表示をそのまま採用する必要はない。ただし、ＮＲＤＦの検索システムの改良は今回のチェック支援システムの範囲を越える問題である。このＮＲＤＦの検索結果の画面表示は、一部のフレームやカラム付けやコメントの組み込みを除けば、文法チェックを通過した後の入力データが格納されているファイルのメンバ"CPNDMT.DATA(Dnnn)" の書式と同じになっていることを注意しておく。

いづれにせよ読み易い原稿に変換する必要がある。著者査読用の原稿を作成する新システムは、ＮＲＤＦの検索システムと独立した一つのユーティリティとして試作し、早期の運用をめざしたい。一般的に辞書およびソースデータは検索システムと共有した方がよい。以上の観点から、結局、検索用ファイルに登録する前の文法チェックを通過した後の入力データが格納されているファイル、即ち、CPNDMT.DATA のメンバ(Dnnn)を著者査読用の原稿を作成するチェック支援システムのソースデータとすることが適当と考えられる。

チェック支援システムの位置づけ

現在の作業手順の流れの中では上の手順(4)と(5)の間に著者チェックを位置づけることになる。そこで著者によるチェックを行う場合の作業手順は次のように変更されるであろう。

(4) 入力データの文法チェック(ここまでは現在と同じ)
(5) 著者によるチェック
  (5.1) 著者査読用の原稿の作成　<---------- CPNDMT.DATA(Dnnn)
      (原稿の媒体変換も含む)　　---------> Dnnn.TXT
  (5.2) 著者へチェックの依頼・査読用原稿の発送
  (5.3) 著者による入力データのチェック
  (5.4) 査読済原稿の返送･受け取り
  (5.5) 査読済の入力データの修正　<--------- Dnnn.TXT
                                ----------> CPNDMT.DATA(Dnnn)

(6) 入力データの文法チェック(これから後は現在と同じ)
(7) 入力データの登録(ＮＲＤＦの検索用ファイルに追加)

チェック支援システムの体制

新たな「(5.3) 著者による入力データのチェック」の過程において、著者からの原稿の返却が遅れることがないようにするための体制、そして、遅れた場合をある程度見積もった対策を手順に組み入れて置く必要がある。

まず、著者による入力データのチェックの作業を円滑に進めるために「コーディング協力員」を荷電粒子核反応実験の施設があるすべての研究所に置くことができないであろうか？　この「コーディング協力員」は、国立大学大型計算機センターとその地域の各大学・研究機関に置かれている「プログラム指導員（相談員）」の制度をモデルにしたものである。「協力員」の人数は１研究所に１人でよいであろう。「協力員」は著者による入力データのチェックの作業の相談に応じて作業が円滑に進むようにモニターする他に、ＮＲＤＦデータベースの利用を拡大するための活動にも協力する。協力員はコーディング作業の担当者を兼ねることができる方がよい。協力員に対するＮＲＤＦデータベース講習会を定期的に開くことも必要になるであろう。

次に、査読済原稿の返送が遅れた場合の対策については、当面は「(5) 著者によるチェック」と「(7) 入力データの登録」の２つの手順を平行に進めた方が良いと考える。その理由は次の４つである。

(a) 従来の作業量と我々のマンパワーを考慮すると、著者チェックの査読用原稿を発送する時期は１２月以降になることがある。

(b) 査読済原稿の受け取りが予定どうりに進まないことがある。

(c) 予算に見合ったデータ量（約３Ｍｂ）だけは年度内に新規に登録する必要がある。

(d) 登録済の入力データに対して、後から著者による入力データのチェックに基づいて入力データの再修正・再登録が必要になった場合でも、現行のＮＲＤＦシステムのユーティリティを利用して処理できるようになっている[7]。

5．著者にチェックを依頼する入力データの内容

次に、入力データの原著者によるチェックシステムの導入に際して、入力データの記述項目の中で著者にチェックを依頼する内容または範囲を考える。内容については第２節で指摘された問題(1)と(2)の解決を中心に置くこととして、ここ

ではその具体的項目について考察する。

ＮＲＤＦの入力データは大きく３つのセクション（ＢＩＢ，ＥＸＰ，ＤＡＴＡ）に区分されている。ＤＡＴＡセクションは一般に複数のサブセクションをもち、それぞれのＤＡＴＡサブセクションには実験データ（数値データ、グラフデータ）が表を単位として格納されている。

その内、入力データのコーディング作業の過程でデータ内容の正確さに関して問題点があるとされているのは、ＥＸＰセクションの記述である。ＥＸＰセクションには次のような事項の形式で実験条件が記述されている。

(1) 個別の核反応式　(2) 標的核の情報　(3) 加速器系
(4) 入射ビームの情報　(5) 測定器系　(6) 測定した物理量　(7) 解析法

ＤＡＴＡセクションにも多少の問題が指摘されている[4]。ＤＡＴＡセクションには、入射粒子のエネルギーや反応の終状態など実験データに直接的な説明を与える情報が記述されるが、ここでは入力データのコーディングに当たって誤差解析に関する情報の不足が指摘されている。

ＢＩＢセクションについては問題はないが、これは論文の書誌情報なので他のセクションを関連づけ、個々のＤＡＴＡセクションを識別する意味でも、著者チェックのために必要であろう。

従って、著者によるデータチェックのためにもコーディングされた入力データのほぼ総てが必要であるということになる。ただし、ＤＡＴＡサブセクションの校正は入力データを作成する我々の責任で行うべきである。特に、論文のグラフをディジタイザで読み取ったグラフデータの数値はグラフ表示に再生できる許容範囲でのみ意味を持つものである。著者には再生したグラフ[8]を見ていただくか、あるいは、そのグラフデータのもとの数値を著者から提供してもらうことを考えるべきであろう。

以上の議論より、著者にチェックを依頼する入力データの内容として次の項目に整理される。

a．文法チェックが済んだ後の入力データのすべてを送る。
b．ＥＸＰセクションの記述を中心に査読をお願いする。
c．ＤＡＴＡセクションについては、コーディングの際に疑問が出されたものに限り査読をお願いする。
d．ＤＡＴＡサブセクションの査読や校正はお願いしない。

その他に参考資料として、ＮＲＤＦの入力データの様式またはコーディングマニュアルの概略、コード名に関する辞書などを利用しやすい形にしておく必要がある。この辞書の件については後でまた取り上げる。

6．チェックを依頼する査読用原稿の分量

新たに著者によるチェック支援システムを作成しようとするとき、そのインプットとアウトプットの量を決めておくことは、このシステム作成の規模を決めるために重要である。インプットについては第４節で、アウトプットの内容については第５節でそれぞれ調べた。著者にＮＲＤＦの入力データのチェックを依頼するとき、著者にお渡しする原稿の分量が多くなり過ぎては今回考えているような著者によるチェック体制はうまくいかないであろう。ここではアウトプットの量を見積もっておくことにする。

アウトプットとなる査読用原稿の分量を見積もるためにサンプルとして、１９９１年度末でＮＲＤＦに登録済となっている[9] 入力データの中から、データ番号の新しい順に１６の論文を対象にする。これは一年間に登録される論文数でみ

表２　サンプル入力データの分量

データ番号	テーブル数	文字数(ﾊﾞｲﾄ)	単語数	行数	ページ数	転送時間(秒)	ＥＸＰ ｾｸｼｮﾝ数
D1193	197	55824	3899	2645	41	418	6
D1194	56	89215	7380	2716	42	669	10
D1195	25	18624	1560	634	10	139	4
D1198	13	8635	701	315	5	64	8
D1199	8	6888	606	232	4	51	1
D1200	14	14885	1251	474	8	111	3
D1201	14	18321	1466	649	10	137	8
D1206	98	88285	6426	2475	38	662	4
D1208	117	87209	6362	2628	40	654	1
D1211	20	19589	1395	569	9	146	1
D1212	10	7887	627	240	4	59	1
D1213	19	5465	371	285	5	40	3
D1216	5	1747	127	91	2	13	1
D1217	32	23454	1772	663	11	175	7
D1223	47	34786	2741	1121	17	260	9
D1225	86	188880	12761	4738	72	1416	1
平均	47.6	41856	3090	1280	19.9	313.4	4.3
標準偏差	51.1	48338	3369	1300	19.8	362.6	3.2

ると約３０％に相当する。これらの入力データのソースデータである"CPNDMT.DATA(Dnnn)"の個々のメンバー（１つの論文に対応）について、それぞれに含まれるデータテーブル数、文字数、単語数、行数、ページ数、ファイル転送時間を調べた。その結果を表２に示す。ここで、データテーブル数とはＤＡＴＡサブセクションの数であり、文字数は半角文字の数で単位はバイト、単語は空白で区切られた文字列を意味する。また、行数は１行に半角７２文字で計算した場合、ページ数は１ページ当たり６６行で計算した場合の数である。ファイル転送時間はソースデータを１２００ｂｐｓの通信速度でファイル転送した場合に要する時間で単位は秒である。

著者に査読を依頼する部分は、ＢＩＢセクションが１ページと異なる核反応毎に１ページ程度のＥＸＰセクションであるから、原稿チェックのために行う著者の仕事量を決める因子はデータテーブル数ではなく、ＥＸＰセクションの数である。そこで上の一覧表にはひとつの入力データの中の異なるＥＸＰセクションの数もあわせて書いておいた。

この調査結果では分散が大きすぎてあまり参考にならないように見える。ところで１９８８年度末までの累積データで見ると、論文件数が６９１で累積データ量が４５．９６ＭＢであるから、論文１件当りの平均データ量は６６．５ＫＢとなる。この平均データ量を文字数に換算すると上の結果の約１．５倍であるから、今回のサンプルにはデータ量の少ないものが入り過ぎたため大きな分散がでていると思われる。この点に注意することを条件にして上の結果を利用する。

さて、文字数とデータテーブル数との間の相関係数は０．５９なのでここでは相関があるとみなす。この相関係数は D1225を除いたときには０．７２となる。データテーブル数と行数（またはページ数）の間には強い相関があるが（相関係数＝０．７４４または０．７４１）、データテーブル数（またはページ数）とＥＸＰセクションの数の間には相関がない（相関係数＝０．０８または０．０２７）。行数とページ数、文字数と転送時間は当然のことながら比例している。以上のデータに直線回帰計算を適用すると、データ量が６６．５ＫＢの平均的論文に対

表３　平均的論文に対する入力データの分量

データ番号	テーブル数	文字数	行数	ページ数	転送時間（秒）	ＥＸＰセクション数
Dxxxx	62.9	66500	1926	34.8	584	5.68
	86.5	66500	2101	38.2		7.81

して、それに含まれるデータテーブル数、文字数、行数、ページ数、ファイル転送時間および（テーブル数に比例すると仮定したときの）ＥＸＰセクション数は次の表３ようになる。答えが２つあるのは統計計算の誤差であり、上がサンプルのすべてを含めたときで、下が D1225を除いたときである。ただし、この表でのページ数は紙に印刷したときの上下の余白の１０行を含めて１ページに６６行として計算してある。

著者による査読用原稿の場合はチェック箇所を示すための行が追加されることになるが、表３から平均的論文に対して作成されるアウトプットの分量を推定できる。紙に印刷したときページ数は３５ページ、フロピィーディスクにテキストファイルとして記録したときで６６．５ＫＢとかなりの量である。その中で著者に特に注意深くチェックをお願いする部分はＢＩＢセクションとＥＸＰセクションであるから、その分量は６ページ程度、今回の調査結果からも多くて１０数セクションで１０ページ程度におさまるものと推定されることが分かった。やはり、著者チェックの体制づくりとチェック支援システムの作成のいづれにもかなりの努力が要求される。

7．チェック支援システムの開発計画

ここではシステム開発の作業段階のうち、前節までの検討に基づいて、システム基本計画とシステム分析までを述べる。

システム基本計画

・新システムの必要性の検討　　（第２、３節）
・目的　　（第２節）
　　ＮＲＤＦのデータ内容の論文の著者による査読作業を支援し、査読による修正内容を入力データにフィードバックする。
・実現性の検討　　（第３、４節）
・開発期間　　１年間
・開発資源　　大型汎用計算機、高性能パソコン及び外部メモリと通信装置
・開発コスト　（開発要員や資源に依存する）

システム分析

・対象業務の範囲　　（第４、５節）

・環境からの制約条件　　（第３、４節）

チェック支援システムを具体的に設計するためには、査読用原稿やコード辞書などの媒体とそれらを送る方法、そして著者と原稿とのインタフェースをどうするかを決める必要がある。

・入力出ファイルの同定　　（第４、６節）

原稿やコード辞書の媒体としては、著者側の便利さとＮＲＤＦ側の処理効率を考えて、現状では紙に印刷したものとフロピィーディスクに記録したものの併用で両方に対応できる方がよいであろう。

・機能仕様（要求仕様）の概略

a)ＮＲＤＦの入力データから著者査読用の原稿を印刷原稿とＭＳ－ＤＯＳのテキストファイルの２つの形で作成する。

b)原稿にはチェックの場所を示す案内を入れる。

c)原稿にはコーディングの人からのコメントや質問を入れる。

d)原稿はＮＲＤＦの特徴であるデータセット構造がわかるように作成する。

e)原稿はＭＳ－ＤＯＳ上の通常のテキストエディタで修正できること。

f)印刷原稿からの修正の場合は、現在と同じように人手でをテキストファイルまたは大型汎用計算機のファイルにキーボード入力する。

g)テキストファイル原稿からの修正の場合は、このシステムが処理する。

h)コード名に関する辞書をＭＳ－ＤＯＳのテキストファイルとして作成する。ただし、その検索機能はユーザが利用するエディタに委ねる。

i)印刷原稿にはグラフデータから再生したグラフも入れる。

テキストファイルの原稿を使っていただける場合については、著者または前述の協力員が使い慣れたパソコンのテキストエディタを利用して原稿の査読・校正ができるようにする予定である。校正された原稿と元の原稿とのファイル比較をする機能をチェック支援システムに持たせればよいと考えている。このようにすればチェック支援システムについて著者または協力員に新たに知ってもらうことはほとんどないし、このチェック支援システムの設計に当たっても著者との直接のインタフェース部分が必要ないことになる。

８．おわりに

ＮＲＤＦに登録する入力データの原著者によるチェックを取り入れるにあたって、チェックが必要な入力データの内容、著者にチェックを依頼する条件や体制とこれまでの手順との関係、チェック支援システムに対する要求仕様と入力・出

力データとその媒体や規模について述べてきた。システム開発の考え方に従うと、システム分析の次の作業は機能仕様に基づいてシステムをいかに実現するかを考えるシステム設計の段階である。

著者自身にデータ内容の査読を依頼する条件や体制について管理運営委員会での意見がまとまれば、以上の方針に沿って、著者による入力データの内容チェック支援のための新システムの設計を進める予定である。

引用文献

［１］富樫雅文・田中一ほか、荷電粒子核反応データファイルユーティリティ開発報告書（1985年3月）

［２］向井重雄・長田博泰、ＮＲＤＦ意味チェックプログラム、
NRDF ANNUAL REPORT 88 (1989年3月),pp.69-78

［３］赤石義紀、新しい段階を向かえたＮＲＤＦ、
NRDF ANNUAL REPORT 87 (1988年3月),pp.2-7

［４］野尻多真喜・手塚洋一、コーディングについて（１）、（２）、
ibid., pp.68-72

［５］加藤幾芳・吉田瞳・佐藤友美、荷電粒子核反応データファイル（ＮＲＤＦ）に関するアンケートの結果、NRDF ANNUAL REPORT 89 (1990年3月),pp.2-12

［６］荷電粒子核反応データファイル（ＮＲＤＦ）使用説明書(1983年3月)
オンライン・データベース利用ガイド第１０版(1990年11月)，全国共同利用大型計算機センターデータベース連絡会，pp.27-29

［７］片山敏之、グラフ併合・登録・保守の管理マニュアル、
NRDF ANNUAL REPORT 87 (1988年3月),p.14-29

［８］岡部成玄、ディジタイザによるグラフ読み取り変換システムの更新、
NRDF ANNUAL REPORT 90 (1991年3月),p.2-8

［９］加藤幾芳・千葉正喜、資料：ＮＲＤＦ核反応リスト、
NRDF ANNUAL REPORT 90 (1991年3月),p.47-79