様式 C-19

# 科学研究費補助金研究成果報告書

平成２３年　５月３１日現在

機関番号：８２１１８
研究種目：基盤研究(C)
研究期間：2008～2010
課題番号：２０５４０３０２
研究課題名（和文）　衝突ビーム型加速器を用いた素粒子実験のデータ公開方法の研究

研究課題名（英文）　Study on Open Access to Data from Particle-Physics Experiments with Colliding-Beam Accelerators

研究代表者
上原　貞治（UEHARA SADAHARU）
大学共同利用機関法人高エネルギー加速器研究機構　素粒子原子核研究所・講師
研究者番号：７０１７６６２６

研究成果の概要（和文）：
衝突ビーム型の高エネルギー加速器を用いた大規模な素粒子実験におけるデータ公開の技術的問題について検討した。このようなデータ公開は、これまでほとんど行われていない。実際のデータ収集から最終的な物理結果を得るまでの全解析過程を検討し、公開において、困難な部分や特別な注意が必要な部分を指摘した。また、天文学分野で行われているデータ公開の手法、高エネルギー物理学の解析ツール、将来にわたって解析できる状態でデータを保存する「データ維持」との関連についても検討を行った。

研究成果の概要（英文）：
I have studied on technical issues which are raised when we are going to release data from a large-scale particle experiment with a high-energy colliding accelerator for open access; in this field, such an open access to data is so far hardly realized. I have investigated every stage of data processing, from data collections to physics analyses, and have clarified possible difficulties against the open access. I have also surveyed methods for open access in the field of astronomy, versatile analysis tools for high-energy physics, and a relation to “data preservation”, which means storage of data for future analyses.

交付決定額

（金額単位：円）

| | 直接経費 | 間接経費 | 合　計 |
|---|---|---|---|
| 2008 年度 | 700,000 | 210,000 | 910,000 |
| 2009 年度 | 300,000 | 90,000 | 390,000 |
| 2010 年度 | 400,000 | 120,000 | 520,000 |
| 総　計 | 1,400,000 | 420,000 | 1,820,000 |

研究分野：素粒子物理学実験
科研費の分科・細目：素粒子物理学実験
キーワード：衝突型加速器　データ解析　アーカイブ　データ維持

１．研究開始当初の背景

(1) 衝突ビーム型の高エネルギー加速器を用いた大規模な素粒子実験では、通常、汎用的な測定器を設置しており、様々な物理プロセスのデータ収集が同時に可能となっている。そして、広い範囲にわたる物理課題に対して、衝突事象のかたちで実験データを提供している。これを用いて、研究者は幅広い範囲の素粒子物理学のテーマに関して研究をすることが可能である。高エネルギー加速器研究機構で国際共同実験としておこなっている Belle 実験は、その典型的な例である。通常これらの実験データは、すべて実験を行ったグループに所属している研究者によって解析されている。しかし、得られたデータをより有効に活用する観点に立てば、実験データを広くグループ外の研究者にも公開し、それぞれで独自の解析をして、より多様な研究成果が挙げられることが望ましい。

(2) 一方、観測天文学（人工衛星からの観測、宇宙線による天体物理学の研究を含む）の分野では、以前より、データ公開が行われている。

(3) 素粒子実験分野におけるデータ公開の必要性を考える研究者は個別にはいるであろうが、その方法の系統的な研究成果の発表されたものはこの時点では目にしていなかった。それが、本研究の動機となった。

２．研究の目的

(1) 衝突型加速器を用いた実験に代表される大規模な素粒子実験で得られたデータをより有効に活用の観点に立てば、実験データを広くグループ外の研究者にも公開し、それぞれで独自の解析をして、研究成果が挙げられることが望ましい。本課題では、衝突型加速器を用いた素粒子実験で得られる事象毎のデータを実験グループ外に公開し、グループに属さない研究者がそれを解析して結果を得られるようにするためには、どのような範囲のデータをどのような形で公開すればよいかということを、実際の衝突実験のデータや解析手法に基づいて明らかにすることをめざす。

(2) さらに、データ解析の過程に着目し、これまでデータ公開を困難にしていた技術的な問題をあぶり出す。そして、その解決方法について検討する。また、同時に、データ公開を取り巻く、素粒子実験データ解析環境の状況についても考察する。

３．研究の方法

(1)本課題では、３つの方面から研究を行った。一つめは、すでにデータ公開が進んでいる天文分野の中から比較的研究手法が素粒子実験に似通っていると思われる人工衛星を利用した天体・宇宙線観測における状況を調べること、二つめは、高エネルギー素粒子実験の国際的な研究所で研究されている Data Preservation(以下、「データ維持」と訳すことにする)の内容を研究すること、そして、三つめに、国内最大の衝突型加速器実験である Belle 実験のデータ解析の手法を、要素に分け、それぞれの段階を検討することで、最後のものが本研究の中心的な部分となった。この三つの調査・研究を統合して結論を導き出した。

(2)Belle 実験のデータ解析を一つの例として検討をする際に、実際のデータフォーマット、データ量、データベース、解析手法等について、具体的にデータを変換し、記録し、貯蔵し、そのデータ処理の全過程をフォローすることによって、データ公開がなされる場合を推定して検討を行った。（なお、Belle 実験はアップグレードされることになっており、今のところデータ公開の予定はない。）

４．研究成果

(1) 衝突型加速器を用いた素粒子実験分野において、事象レベルでのデータ公開が進んでいない（現状で行われていない）最大の理由は、技術的な問題にあると考えられる。政治的、ヒューマンパワー的な要因もあるかもしれないが、それだけでは、データ公開が進んでいる天文学分野との相違が説明できない。

研究者がグループを作って研究している、建設費が高価である、国際共同研究で進められている、などの点は素粒子実験分野と天文学分野で共通しているので、これらが要因として重要であるにしても、ここに両分野で差が出ている原因を求めることはできない。

したがって、技術的な問題を考えることは、この問題の本質を掘り下げる上で大いに意味がある。

(2) 天文学分野のデータ公開の枠組みや全体的な流れは、素粒子実験のデータ公開にも適用できそうなものである。解析の手順も共通している部分が多い。従って、天文学分野のデータ公開は素粒子実験分野のモデルになりうる。

しかし、流れの個別の要素において、大規模な素粒子実験に特有の困難な箇所、いわば高いハードルとなる項目が存在することがわかった。

(3)図１に、高エネルギー加速器研究機構にある電子陽電子衝突型加速器 KEKB を使って行われている Belle 実験で用いられている測定器を示す。おもに７種類の検出器コンポーネントから収集されるデータを用いて、素粒子物理の解析がなされている。

また、図２に、Belle 実験でのデータの処理と解析の流れを表すブロックダイアグラムを示す。実験で得られた「生」のデータ(Raw Data)は、各検出器コンポーネントの較正情報などを利用して、位置、エネルギー、運動量、速度等の基本的な粒子情報に変換される。そして、異なる種類の検出器情報を総合して、粒子識別等を行い、MDST (mini-data-summary tape)と呼ばれる物理解析に良く用いられる情報を要約した便利に使用できるデータセットにまとめられる。

さらに、この MDST に含まれる事象ごとのデータは、物理テーマに基づく目的ごとに大きく分類かつ選別され、何種類かのスキムファイルと呼ばれるデータセットの集団にコピーされる。多くの場合、個々の研究者は、このスキムファイルを出発点にして解析を行う。

Belle 実験に限らず、同種の大規模な素粒子実験においては、ほぼ同様の手法が取られている。データ公開を考える際は、このスキムファイルを公開することが、その基本になる。したがって、図２の行程は、スキムファイルがたびたび改訂されることがなければ、適度な補正を提供すればよく、データ公開において大きな問題になることはない。

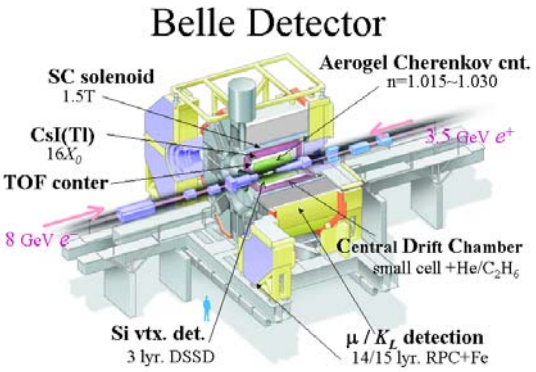


図１．Belle 測定器の概観と検出器コンポーネント。

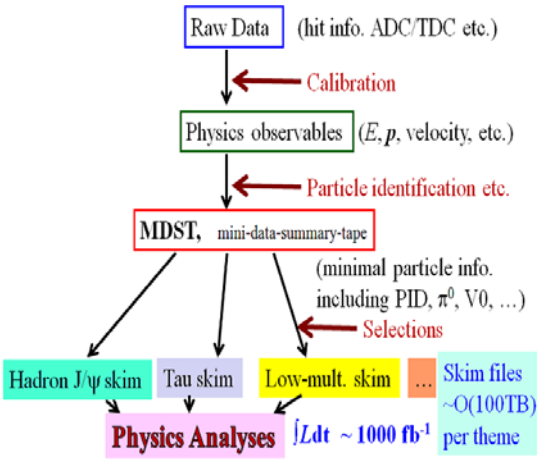


図２．Belle 実験におけるデータ処理と解析の流れを示すブロックダイアグラム。

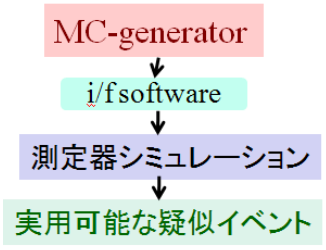


図３：解析に必要なモンテカルロ（MC）事象の発生と処理の流れ。”i/f software” は、実験グループ外の研究者が用意した MC 事象発生プログラムと、グループ内で用意された解析プログラムとの間を橋渡し(interface)するためのソフトウェアである。

(4)大規模な素粒子実験でのデータ公開において、困難を引き起こす技術的要因はいくつか考えられる。

まず、データ量が大きく、そのデータ処理時間(CPU 時間)が長いことが挙げられるが、図２にあるように、１つのスキムファイルの大きさは 100 テラバイト程度であり、これは小さくはないが、実験実施期間中における手順の工夫と計算機技術の進歩を考えると、量的に克服可能であって質的に本質的な問題とはいえない。

検討の結果、主要な問題は、データ解析に不可欠なモンテカルロ (MC) 事象の発生と、実験期間における検出器状態の較正およびその測定器シミュレーションの MC 事象への適用にあるという結論に至った。図３に MC 事象の処理の手順を示す。とくに、研究者が

測定したい信号プロセスの検出効率を求めるために、この MC 事象の手順は不可欠なものである。物理学理論に基づく MC 事象の発生部分（MC-generator）は、しばしば、データ解析を行う研究者によって準備され、この図の全工程がその研究者によってなされなければならない。また、測定器シミュレーションの部分は長年の実験の期間によって変化する部分である。

これらは、実験グループ内の研究者にとっても、困難を伴う克服すべきプロセスであるが、グループ外の研究者にとっても同様に手を抜くことができない。それは、汎用型の素粒子実験において、反応事象を解析するうえで、本質的に避けて通ることのできないプロセスであり、グループ内外の研究者において、基本的に同じ研究方法がとられねばならない。

(5) 検討の結果、データ公開をする場合に、グループの内と外に向けて別々の解析環境を用意することは、手間がかかるだけで、実質的利益がないと判断した。したがって、データ公開用に、特別な解析体制をとることは効率的ではない。ただし、データフォーマットやデータベース、解析ツールが入手しやすく広く使われているものを選ぶことには十分な価値がある。

(6)また、衝突型加速器実験で得られたデータを、実験グループの解散後に生じるかもしれない予測できない研究課題のために解析できる状態で保存（維持）しておくことが望ましい、とする「データ維持」の考えからすると、データ公開を行わなかった場合にデータ維持を行おうとすると、それは実験グループ外の研究者を想定することになるので、技術的にはデータ公開と同じことを用意しなければならない。

したがって、データ公開を行うためには、実験グループ解散後にデータ維持を行うことも含めて念頭に置き、MC シミュレーションと測定器の応答性能のドキュメンテーションのグループ外の研究者にも理解できるようなものを用意すること、汎用的なデータフォーマットやデータベース、解析ツールを用いること、が最重要である。

(7)「データ維持」についての海外の研究者の結論には、実験解析の初期段階から「データ維持」のことを念頭に置いて、それに向けた必要な仕事を並行してすすめることを推奨するものがある。しかし、実際には、ヒューマンパワーの問題や解析手法の成熟に時間がかかることから、これには困難が多いと予想する。むしろ、ある段階でグループ外の研究者にも解析を可能にする「データ公開」の観点からのツールやドキュメンテーションの充実をはかることが、将来の「データ維持」の可能性につながっていくのではないかと考える。

５．主な発表論文等
（研究代表者、研究分担者及び連携研究者には下線）

〔学会発表〕（計１件）
①上原　貞治,「大規模素粒子実験におけるデータ公開の技術的検討」日本物理学会, 2011 年 3 月 25 日, 新潟大学.　（本学会は東日本大震災の影響で中止になったが、本発表は、発表資料の公開により発表は成立した、と見なされる。）

６．研究組織
(1)研究代表者
上原　貞治（UEHARA SADAHARU）
高エネルギー加速器研究機構・素粒子原子核研究所・講師
研究者番号：70176626