Heike Neuroth, Stefan Strathmann, Achim Oßwald,
Regine Scheffel, Jens Klump, Jens Ludwig [Hrsg.]

Langzeitarchivierung von Forschungsdaten

Eine Bestandsaufnahme

Kapitel 13
Teilchenphysik

vwh Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

Universitätsverlag Göttingen

Langzeitarchivierung von Forschungsdaten

Eine Bestandsaufnahme

nestor

Heike Neuroth, Stefan Strathmann, Achim Oßwald,
Regine Scheffel, Jens Klump, Jens Ludwig [Hrsg.]

Langzeitarchivierung von Forschungsdaten

Eine Bestandsaufnahme

GEFÖRDERT VOM

Bundesministerium
für Bildung
und Forschung

Förderkennzeichen: 01 DL 001 B

Universitätsverlag Göttingen

vwh
Verlag Werner Hülsbusch
Fachverlag für Medientechnik und -wirtschaft

D-GRID GmbH

Langzeitarchivierung von Forschungsdaten. Eine Bestandsaufnahme
hg. v. Heike Neuroth, Stefan Strathmann, Achim Oßwald, Regine Scheffel, Jens Klump, Jens Ludwig
im Rahmen des Kooperationsverbundes nestor – Kompetenznetzwerk Langzeitarchivierung und Langzeitverfügbarkeit digitaler Ressourcen für Deutschland
nestor – Network of Expertise in Long-Term Storage of Digital Resources
http://www.langzeitarchivierung.de/

Kontakt: editors@langzeitarchivierung.de
c/o Niedersächsische Staats- und Universitätsbibliothek Göttingen,
Dr. Heike Neuroth, Forschung und Entwicklung, Papendiek 14, 37073 Göttingen

Die Herausgeber danken Anke Herr (Lektorat) und Sonja Neweling (Redaktion) sowie Martina Kerzel und Lajos Herpay (Gestaltung und Montage) für ihre unverzichtbare Unterstützung bei der Fertigstellung des Handbuchs.

Bibliografische Information der Deutschen Nationalbibliothek
Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet unter http://www.d-nb.de/ abrufbar.

Die Inhalte des Buches stehen auch als Onlineversion
(http://nestor.sub.uni-goettingen.de/bestandsaufnahme/)
sowie über den Göttinger Universitätskatalog
(http://www.sub.uni-goettingen.de) zur Verfügung.
Die URN lautet: http://nbn-resolving.de/urn:nbn:de:0008-2012031401.

Die digitale Version dieses Werkes steht unter einer Creative Commons Namensnennung-Nicht-kommerziell-Weitergabe unter gleichen Bedingungen 3.0 Unported Lizenz.

Einfache Nutzungsrechte liegen beim Verlag Werner Hülsbusch, Boizenburg.
© Verlag Werner Hülsbusch, Boizenburg, 2012
http://www.vwh-verlag.de
In Kooperation mit dem Universitätsverlag Göttingen.

Markenerklärung: Die in diesem Werk wiedergegebenen Gebrauchsnamen, Handelsnamen, Warenzeichen usw. können auch ohne besondere Kennzeichnung geschützte Marken sein und als solche den gesetzlichen Bestimmungen unterliegen.

Druck und Bindung: Kunsthaus Schwanheide
Printed in Germany – Als Typoskript gedruckt –

ISBN: 978-3-86488-008-7

13 Teilchenphysik

Volker Gülzow, Yves Kemp

13.1 Einführung in den Forschungsbereich

Die Teilchenphysik ist eine Disziplin der Physik und beschäftigt sich mit der Erforschung der kleinsten Bestandteile der Materie, ihren Eigenschaften und Wechselwirkungen. Die theoretische Teilchenphysik erarbeitet neue Modelle zum Verständnis der Zusammenhänge und verfeinert bestehende Theorien. Die experimentelle Teilchenphysik überprüft diese Vorhersagen mithilfe von äußerst komplexen Experimenten, die an großen Beschleunigeranlagen angesiedelt sind. Dort werden Teilchen wie Elektronen oder Protonen aber auch schwere Ionen beschleunigt bevor sie miteinander kollidieren. Als Folge der durch den Zusammenstoß freigesetzten Energie können andere Teilchen erzeugt werden. Diese Teilchen sind sehr kurzlebig und zerfallen binnen Bruchteilen von Sekunden. Um diese Zerfallsprodukte optimal vermessen zu können, werden um den Kollisionspunkt Detektoren aufgebaut. Durch die Analyse der Zerfallsprodukte können die Eigenschaften der durch die Kollision erzeugten Teilchen bestimmt werden.

Der größte und energiereichste Beschleuniger ist zur Zeit der Large Hadron Collider (LHC) am CERN in der Nähe von Genf. Die dort betriebenen Experimente sind ATLAS, CMS, LHCb und ALICE. In Deutschland sind DESY in Hamburg und die GSI in Darmstadt Standorte von Beschleunigern und Experimenten für die Teilchenphysik. Forscherinnen und Forscher an deutschen Instituten sind weiterhin auch an Experimenten am Fermilab (Chicago), SLAC (Stanford) und KEK (Japan) beteiligt.

Beschleuniger für diese Forschungen erfordern Investitionen in Höhe von Milliarden Euro und können nur noch in weltweiten Kollabo-

rationen realisiert werden. Die Betriebskosten der Beschleuniger sind immens, die Experimente sind sprichwörtlich einmalig und viele physikalische Phänomene sind extrem selten, so dass die genommenen Daten besonders kostbar und bewahrenswert sind.

Schon die Planung der Beschleuniger und Experimente ist rechen- und datenintensiv: Die Eigenschaften der Beschleuniger werden simuliert, die Simulationen verschiedener Detektortechnologien werden mit Versuchsdaten verglichen und unterschiedliche Physikmodelle werden simuliert, um das Potential neuer Experimente zur Entdeckung neuer Phänomene zu bestimmen. Im Vorfeld der 2009 erarbeiteten Studie zu einem Detektor am vorgeschlagenen Internationalen Linearen Kollider wurden ca. 50 TB Simulations-Daten generiert und analysiert.[1]

Die Datennahme und die Analyse moderner Experimente der Teilchenphysik stellt große Herausforderungen an eine Computing Infrastruktur dar. Aus diesem Grund war und ist die Teilchenphysik auch federführend bei der Benutzung von Grid-Ressourcen, sei es innerhalb von D-GRID[2] oder in WLCG[3], dem weltweiten Computing Grid für den LHC. So wurden 2010 insgesamt 13 PetaByte Daten von den vier LHC Experimenten am CERN gesammelt und weltweit zur Analyse verteilt.[4]

Dank des technologischen Fortschrittes sind die Datenmengen heutiger Experimente wesentlich größer als die älterer Experimente. So wurde in den Jahren 1992–2007 von den beiden am HERA-Ring am DESY Hamburg betriebenen Experimenten H1 und ZEUS 200 TB Roh-Daten pro Jahr genommen: Zur damaligen Zeit technologisches Neuland in der Datenhaltung.[5]

Die Teilchenphysik ist daher seit Jahrzehnten geübt im Umgang mit großen Datenmengen. Aus ihrem Umfeld entstanden diverse Speichertechnologien, die mittlerweile auch erfolgreich in anderen datenintensiven Forschungsbereichen eingesetzt werden. Ein Beispiel ist dCache[6],

1 Vgl. Stoeck et al. (2009).
2 Vgl. D-Grid (2011a).
3 Vgl. WLCG (2011).
4 Vgl. Brumfield (2011).
5 Auflistung der im DESY RZ befindlichen Daten durch Dmitri Ozerov und Ralf Lueken (private Mitteilung).
6 Vgl. dCache (2011).

welches federführend am DESY entwickelt wird, und von D-GRID unterstützt wurde.

Neben den bekannten Gründen für die Langzeitarchivierung von Daten, wie z.B. den von der DFG vorgegebenen Richtlinien zur guten wissenschaftlichen Praxis[7], gibt es vier für die Teilchenphysik spezifische Argumente:

- Langzeitanalyse und Verlängerung des wissenschaftlichen Programms eines Experimentes
- Kombination von Daten unterschiedlicher Experimente
- Wiederverwendung von Daten für zukünftige Fragestellungen
- Ausbildung von Studenten und Schülern sowie „Outreach“

Eine Begründung, die Daten eines Experimentes langfristig zu archivieren, ist die physikalische Relevanz und Einzigartigkeit dieser Daten für die Zukunft. Das Beispiel der JADE Daten hat gezeigt, dass alte Daten ein hohes physikalisches Potential haben.[8] Anhand von diesem und anderen Beispielen wird geschätzt, dass bis zu 10% zusätzliche Publikationen während der Langzeitarchivierung entstehen. Da diese Publikationen mit dem bestmöglichen Verständnis des Detektors, der Daten und der Modelle erstellt werden, ist deren wissenschaftliche Relevanz sehr hoch einzuschätzen.

Wenn in der Teilchenphysik von Langzeitarchivierung gesprochen wird, so meint dieses nicht nur die Archivierung von Rohdaten. Auch wenn diese einen großen Teil der Daten darstellen, so sind doch andere Daten für deren späteres Verständnis unabdingbar und müssen ebenfalls archiviert werden. Das betrifft in erster Linie „Conditions Data“, also Daten, die die Experimentbedingungen und den Detektor beschreiben. Von besonders großer Wichtigkeit ist weiterhin das Archivieren aller Software, die zum Erzeugen, Verarbeiten und Analysieren der Daten gebraucht wird. Es ist dabei notwendig, im Rahmen der Langzeitarchivierung diese Software immer in einem Zustand zu halten, dass eine Analyse durchgeführt werden kann, eine organisatorische Herausforderung, die erhebliche Personalressourcen erfordert.

7 Vgl. DFG (1998).
8 Vgl. Bethke (2010).

Schließlich sind weitere, für die Langzeitarchivierung in Frage kommenden Daten auch alle Dokumente, die von den Mitgliedern der Kollaboration erstellt worden sind, seien es interne Dokumente, Präsentationen und Notizen oder externe Publikationen.

13.2 Kooperative Strukturen

Große Experimente der Teilchenphysik sind über Kollaborationen weltweit organisiert. Üblicherweise werden ganze Institute oder Forschungslabore durch Unterzeichnung von einem sogenannten „Memorandum of Understanding“ (MoU) in eine Experiment-Kollaboration aufgenommen und ermöglichen so die für eine lange Planungsphase und den Betrieb notwendige Beständigkeit. Die ATLAS Kollaboration z.B. bestand Anfang 2011 aus über 3000 Mitgliedern aus ca. 200 Instituten und 38 Ländern.[9]

Ein typisches Teilchenphysikinstitut ist in der Regel Mitglied in einer oder ein paar wenigen Kollaborationen, je nach Größe des Instituts. Diese Affiliationen bleiben über Jahre bestehen, teilweise über mehrere Jahrzehnte. Vor dem Bau eines Experimentes werden von diesen Instituten z.B. Konzeptstudien erarbeitet, beim Bau selber wird unterschiedliche Expertise beigetragen oder auch Hardware-Beiträge geleistet.

Verglichen mit anderen Disziplinen der Physik oder auch Disziplinen außerhalb der Physik ist die Teilchenphysik eine recht große, aber homogene Community. Sowohl innerhalb der jeweiligen Kollaborationen als auch innerhalb der Community ist ein gemeinsames, organisiertes Vorgehen üblich.

Aus wissenschaftlichen und wirtschaftlichen Gründen wurden und werden häufig mehrere Experimente an einem Beschleuniger aufgebaut. So wurden am LEP Ring am CERN die vier Experimente ALEPH, DELPHI, L3 und OPAL betrieben, am Tevatron in den USA CDF und DØ, und auch am LHC werden vier Experimente betrieben. Am HERA-Ring bei DESY waren dies neben den schon genannten H1 und ZEUS Experimenten auch HERMES und HERA-B. In Fällen, in denen nur ein Experiment an einem Beschleuniger betrieben wird, war in der Vergangenheit

9 Vgl. ATLAS Personen Datenbank (ATLAS-interne Webseite).

manchmal ein ähnliches Experiment an einem anderen Beschleuniger aktiv, wie z.B. BaBar am SLAC und Belle am KEK.

Die unterschiedlichen Experimente überprüfen ihre Resultate wechselseitig. Während der Laufzeit der Experimente sind Kooperationen und Austausch von Daten zwischen den verschiedenen Experimenten aber sehr selten, teilweise sogar von den Kollaborationen verboten. Es ist hingegen üblich, dass es am Ende der Laufzeit und besonders nach Beendigung der Datennahme zu Kombinationen der Daten kommt. Besonders sinnvoll sind solche Kombinationen bei ähnlichen Detektoren an einem Beschleuniger, wie z.B. zwischen H1 und ZEUS.[10]

Die Rohdaten der Experimente werden im Normalfall von IT-Spezialisten der Institute gespeichert, die den Beschleuniger beherbergen. Daraus ergibt sich, dass z.B. DESY-IT die Daten aller bei HERA beteiligten Experimente speichert und für eine Verfügbarkeit innerhalb eines bestimmten Zeitraums verantwortlich zeichnet.

Aus diesen beiden Punkten wird deutlich, dass dem Labor, das den Beschleuniger betreibt, eine herausragende Rolle bei der Speicherung der Rohdaten zukommt. Dieses Institut ist somit bestens geeignet, auch LZA-Dienste zu übernehmen. Da es nur wenige solcher Labore gibt, sollten diese sich untereinander austauschen und ihre Anstrengungen koordinieren. Es ist dabei wichtig, dass alle Schritte in einem internationalen Kontext gesehen werden.

Um die Planung und den Bau zukünftiger Beschleuniger zu koordinieren wurde das „International Committee for Future Accelerators“, ICFA, etabliert. Im Jahr 2009 rief die ICFA eine „Study Group on Data Preservation and Long Term Analysis in High Energy Physics“ (DPHEP) ins Leben, um einige schon bestehende Anstrengungen auf diesem Gebiet abzustimmen und zu verstärken.[11] Es ist also sicher gestellt, dass die unterschiedlichen Labore wie z.B. CERN, DESY oder FNAL und die HEP Experimente eine Plattform haben, auf der sie sich austauschen und ihre Arbeiten koordinieren können.

Das Stanford Linear Accelerator Center (SLAC) beherbergt seit Ende der 1960er-Jahre eine Datenbank mit Literatur aus der Teilchen-

10 Vgl. HERA (2011).
11 Vgl. DPHEP (2011).

physik. Seit dieser Zeit sind andere Institute diesem Projekt beigetreten und haben den Datenbestand stetig vergrößert. Heute ist dieses System, SPIRES, ein wichtiges Archiv der Teilchenphysik, und eine unersetzliche Quelle für Informationen.[12] CERN, DESY, Fermilab und SLAC arbeiten an einem Nachfolger, INSPIRE, der die bestehenden Publikationsdatenbanken und Archive noch besser vernetzten wird.[13]

Für die Langzeitarchivierung von internen Dokumenten, Präsentationen und Notizen sowie Veröffentlichungen jeglicher Art aus der Teilchenphysik ist somit das INSPIRE Projekt bestens aufgestellt.

Archivierung von Software ist in der Teilchenphysik bisher parallel zur Archivierung von Rohdaten gehandhabt worden, insofern sind auch hier die Labore bestens platziert, um diese zu archivieren. Auch hier ist eine Vernetzung der Labore sinnvoll und wird auch von der DPHEP Arbeitsgruppe untersucht.

13.3 Daten und Metadaten

In der Teilchenphysik wird sowohl mit Experimentdaten als auch mit simulierten Daten gearbeitet. Obwohl die Herkunft dieser beiden Kategorien unterschiedlich ist, werden sie im Verlauf einer Analyse ähnlich behandelt. Die Analyse selber verwendet unterschiedliche Datenmodelle: Man spricht auch von einem Data-Tier[14]. Beispielhaft soll der Verlauf anhand der Experimente ATLAS[15] und CMS[16] gezeigt werden. Andere Experimente haben prinzipiell ähnliche Vorgehensweisen, aber eventuell eine andere Namensgebung.

Am Anfang einer Analysekette stehen die Rohdaten des Experimentes oder ihr Äquivalent aus simulierten Daten. In der ATLAS- und CMS-Nomenklatur werden diese RAW-Daten[17] genannt und am CERN, dem beherbergenden Labor, gespeichert und dort mindestens für die Dauer des Experimentes vorgehalten. Diese RAW-Daten werden am CERN im

12 Vgl. SPIRES (2011).
13 Vgl. INSPIRE (2011).
14 Vgl. CMS (2011).
15 Vgl. ATLAS Collaboration (2005).
16 Vgl. CMS Collaboration (2005).
17 ATLAS Collaboration (2005); CMS Collaboration (2005).

Tier-0 Rechenzentrum ein erstes Mal mit den Detektor beschreibenden Daten kombiniert und daraus dann physikalisch relevante Größen rekonstruiert. Das Ergebnis der Kombination und Rekonstruktion sind sogenannte RECO-Daten und sind in etwa ähnlich groß. Diese Daten werden dann im weltweiten LHC-Grid an die Tier-1 Zentren verteilt. Im Laufe der Zeit werden die Rekonstruktionsalgorithmen verbessert und verfeinert. Periodisch werden die RAW-Daten deshalb neu rekonstruiert. Es kann also mehrere Versionen von RECO-Daten geben.

Die Menge an RECO-Daten ist zu groß, als dass jede Analyse RECO Daten benutzen könnte. Im LHC-Analysemodell gibt es deshalb ein Format, welches für die Analyse optimiert ist: AOD, Analysis Object Data. In der Konvertierung zu AOD-Datensätzen werden für die jeweilige Physikgruppe uninteressante Ereignisse weggelassen und Informationen aus den Ereignissen zusammengefasst. Hieraus folgt, dass es für jede Physikgruppe unterschiedliche AOD Datensätze geben kann. Auch die AOD Datensätze profitieren davon, dass sich im Laufe der Zeit die Rekonstruktionsalgorithmen verbessern, es kann also mehrere Versionen der AOD-Daten geben.

Es ist üblich, dass ein Teil der Analyse mit AOD-Daten gemacht wird, ab einem bestimmten Zeitpunkt aber zu noch kleineren, kompakteren und spezialisierteren Formaten übergegangen wird. Das können dann einfache Ntuple sein, oder aber auch komplexere PAT-Tuples (CMS) oder D3PD (ATLAS). Auch hiervon gibt es mehrere Varianten und Versionen.

Neben den Experiment-Daten sowie simulierten Daten gibt es noch die sogenannten „Conditions Data“, welche den Zustand des Experimentes beschreiben. Diese sind erstmal separate Daten, fließen aber in das Erzeugen von Sekundär- und Tertiärdaten ein. Ob und wie man „Conditions Data“ speichert, hängt davon ab, ob man Primär-, Sekundär- oder Tertiärdaten in ein LZA übernehmen möchte.

Weiterhin fallen Daten in Form von internen Dokumenten, Präsentationen und Notizen sowie Veröffentlichungen jeglicher Art an. Auch diese gilt es zu sichern.

Daten sind untrennbar verwoben mit der Software, um diese zu analysieren und Steuerdateien um die Software zu steuern. Somit ist Software zum Simulieren, Rekonstruieren und Analysieren mitsamt den Steuerdaten selbst ein Datenbestand, der für eine LZA betrachtet werden muss.

Es ist finanziell nicht möglich, alle diese Daten in unterschiedlichen Varianten und Versionen langfristig vorzuhalten. Eine erste Einschränkung kann gemacht werden durch Löschen von älteren Versionen eines Datensatzes. Welche anderen Einschränkungen man macht, hängt von den finanziellen Möglichkeiten der Archivare ab, aber auch vom späteren Einsatzzweck der Daten. So kann man sich folgende Szenarien vorstellen:[18]

- *Modell 1:* Bereitstellung zusätzlicher Dokumentation: Sinnvoll, wenn später nur wenig Informationen über Publikationen hinausgehend gesucht werden sollen.
- *Modell 2:* LZA der Daten in einem vereinfachtem Format: Sinnvoll für den Einsatz in der Lehre und für „Outreach".
- *Modell 3:* Analyse-Software und Analyse-Daten speichern (AOD in LHC-Nomenklatur): Ermöglicht komplette Analysen basierend auf der existierenden Rekonstruktion.
- *Modell 4*: Rekonstruktions- und Simulationssoftware sowie Roh-Daten speichern: Erhält das volle Potential der Physikdaten.

Diese Modelle schließen sich nicht gegenseitig aus: So kann z.B. Modell 1 unabhängig von den anderen drei implementiert werden.

Wie es unterschiedliche Arten von Daten in der Teilchenphysik gibt, so gibt es auch unterschiedliche Kulturen im Umgang mit diesen und mit dem Publizieren der Daten. Software z.B. wird teilweise unter freien Lizenzen wie der GPL veröffentlicht und ist somit für andere Experimente nutzbar. Auch bei der Nutzung von Dokumentservern wie bei SPIRES ist die Teilchenphysik sehr weit fortgeschritten bei der Nutzung von LZA.

18 Vgl. South et al. (2010).

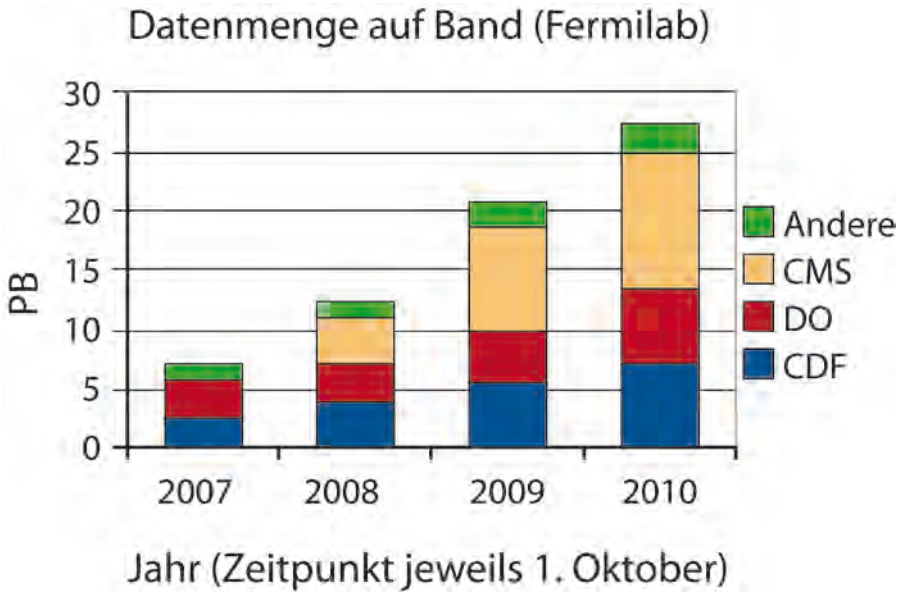

Abb. 1: Am Fermilab auf Band gespeicherte Datenmenge

Publizieren von „Conditions Data" ist nur zusammen mit den Rohdaten sinnvoll. Rohdaten selber werden mindestens während der gesamten Dauer der Datennahme gespeichert, was durchaus mehrere Jahrzehnte dauern kann. Während dieser Zeit werden die Daten selber nicht publiziert. Nach der Zeit der Datennahme besteht eine Kollaboration noch einige Zeit weiter, während dieser Zeit werden noch Analysen fertig gestellt und somit die Daten weiter gespeichert. Während des Bestehens der Kollaboration entscheidet diese, wer Zugang zu den Daten erhält, normalerweise nur Mitglieder der Kollaboration. Nach dem Ende einer Kollaboration ist die „Ownership" der Daten häufig unklar. In der Vergangenheit war das Bewusstsein für ein Archivieren und Publizieren der Daten nach dem Ende der Kollaboration nur selten ausgeprägt. Eine erfolgreiche Ausnahme stellt die JADE-Kollaboration dar, die mittels alter Daten und neuer Methoden die Messung der Kopplungskonstanten α_S verbesserte. In den letzten Jahren ist, auch dank der DPHEP-Initiative, das Bewusstsein für diese Fragestellungen in die Breite des Forschungsfeldes vorgedrungen. Das Projekt WissGrid[19] widmet sich unter anderem auch der Thematik Langzeitarchi-

19 Vgl. WissGrid (2011b).

vierung. HEP Grid ist eine Community innerhalb von WissGrid, somit wird die Thematik auch über diesen Weg in die HEP hinein gebracht.Nicht alle Daten, die von einem Experiment genommen werden, sind langfristig von Interesse. So gibt es zum Beispiel die sogenannten „cosmic runs", in denen die Datennahme bei ausgeschaltetem Beschleuniger läuft. Statt Kollisionen wird kosmische Höhenstrahlung gemessen. Diese ist zur Kalibrierung und Testen des Datennahmesystems sinnvoll, wird aber nur einige Stunden bis wenige Monate gespeichert. Auch gibt es Zeiten, zu denen eine Detektorkomponente während der Datennahme ausgeschaltet ist. Auch hier werden die Daten üblicherweise nach kurzer Zeit gelöscht. Ähnliches gilt für simulierte Daten: Sobald Daten mit einer verbesserten Simulation vorliegen werden die älteren Daten zur Löschung frei gegeben. Die Qualität der Forschungsdaten wird also schon während der Laufzeit des Experimentes überprüft und sichergestellt. Ein LZA-Projekt muss diese Entscheidung nicht mehr treffen. Die in diesem Beitrag genannten Datenmengen sind auch immer schon nach Abzug von nicht benötigten Daten zu verstehen.

Die großen Datenmengen in der Teilchenphysik bedurften immer schon ausgereifter Verwaltungsmechanismen und Metadatenkataloge. Als Beispiel seien PhEDEx (CMS) und DQ2 (ATLAS) genannt. Um eine LZA erfolgreich durchzuführen, muss geklärt werden, in welchem Umfang diese Systeme aufrecht gehalten werden müssen oder Informationen in andere Verwaltungssysteme übertragen werden sollen.

Schon jetzt ist die Menge an Daten von beendeten Experimenten recht groß, je nach Modell müssen zwischen einem halben PetaByte und ein paar wenigen PetaByte gespeichert werden. Dies sind Datenmengen, die in heutigen Rechenzentren prinzipiell gespeichert werden können. In dieser Abschätzung sind die LHC Experimente nicht einbezogen. Diese produzieren zusammen pro Jahr die Datenmenge von 13 PetaByte und stellen somit die zuzeit größten Produzenten von Daten in der Teilchenphysik dar. Auch die Daten aus den CDF- und D0-Experimenten am Fermilab sind in dieser Abschätzung nicht inbegriffen, diese sind aber durchaus substantiell, wie in Abb.1[20] sichtbar.

Der Begriff „Langzeitarchivierung" sagt erstmal nichts über die konkrete Dauer der Archivierung aus. Innerhalb der HEP gibt es Diskus-

20 Vgl. Steven Wolbers; Gene Oleynik der Fermilab Computing Division (private Mitteilung).

sionen darüber, wie lange „lang“ sein soll. Das kann einige wenige Jahre sein bis hin zu einer permanenten Archivierung. Diese Spanne lässt sich durchaus begründen: Forscher am still gelegten BaBar-Experiment wissen, dass mit Super-B Ende dieses Jahrzehnts ein Experiment starten wird, welches die BaBar-Ergebnisse reproduzieren kann. Man muss also nur die Zeit überbrücken bis zur Verfügbarkeit der neuen Daten und eventuell noch erste vergleichende Analysen anstellen. Ganz anders sieht die Situation z.B. für die HERA-Experimente aus: Es gibt momentan keine Planungen für einen neuen Elektron-Proton-Beschleuniger. Man muss also vorerst davon ausgehen, dass die HERA-Daten permanent archiviert werden müssen. Die physikalische Relevanz der HERA-Daten ist auf jeden Fall gegeben: HERA-Daten ermöglichen unter anderem einen sehr genauen Blick in das Innere des Protons, was bei weiteren Beschleunigern wie dem LHC für Analysen von großer Wichtigkeit ist.

Das Format der Daten ist üblicherweise auf das jeweilige Experiment zugeschnitten, es gibt in der Teilchenphysik keine experiment-übergreifenden Formate für Rohdaten. Neuere Experimente wie die LHC-Experimente verwenden Dateiformate welche auf ROOT[21] aufbauen, aber auch hier benutzt jedes Experiment sein eigenes Format.

Es gilt jedoch nicht nur, die Experimentdaten zu speichern. Wir haben schon über die Bedeutung der „Conditions Data“ gesprochen, sowie der internen Dokumentation. Die Software, die notwendig ist um die Daten zu prozessieren und analysieren, ist sehr umfangreich. So umfasst die CMS Software ca. 5 Millionen Zeilen Code (inklusive Kommentar), hinzu kommen noch notwendige externe Programmpakete wie z.B. das ROOT Framework. In dieser Software steckt sehr viel Wissen über die Daten und den Detektor, es ist also wichtig, dass die Software auch archiviert wird. Es muss allerdings überlegt werden, wie ein geeignetes Langzeitarchiv für Software im Allgemeinen und für Software diesen Umfangs im Speziellen aussehen kann. Analog zu der periodischen Überprüfung der Richtigkeit von Daten in einem Langzeitarchiv muss auch die prinzipielle Funktionsfähigkeit und Korrektheit der Resultate der Software periodisch überprüft werden. Diese Thematik muss unbedingt von den jetzt bestehenden Kollaborationen aufgegriffen werden. Die DPHEP-Initiative hat eine Diskussion darüber in die

21 Vgl. ROOT (2011).

Wege geleitet. Erste Ergebnisse eines Projektes der HERA-B-Kollaboration zeigen, dass es möglich ist, die wichtigsten Teile der Experimentsoftware auf einen aktuellen Status zu portieren. So stammte die letzte offizielle Version der HERA-B-Software von 2003. Nach zwei Monaten Arbeit konnte 2010 eine Version erstellt werden, die das Lesen und Analysieren von Daten auf einem modernen ScientificLinux-5-System ermöglicht.[22]

Daten in der Teilchenphysik unterliegen nicht dem Datenschutz, d.h. es gibt keine rechtlichen Vorgaben, die bei einer Langzeitarchivierung beachtet werden müssten. War bisher das Recht, Experimentdaten zu untersuchen und Ergebnisse zu veröffentlichen eng an die Mitgliedschaft in der Kollaboration geknüpft, muss dies für die Langzeitarchivierung überdacht werden. Wer kann, Jahre nach Beendigung einer Kollaboration, entscheiden, welche Daten von wem untersucht werden? Wer garantiert die kollaborationsinterne Qualitätskontrolle von Resultaten vor einer Publikation, wenn es keine Kollaboration mehr gibt? Das bisher erfolgreichste Beispiel für Publikationen von Resultaten nach einer Langzeitarchivierung, JADE, zeigt diese Problematik auf, umgeht eine Beantwortung der prinzipiellen Frage aber durch die Feststellung, dass die Daten noch immer am DESY liegen, und „das für die Analyse notwendige Wissen nur am DESY und am MPI für Physik in München ist, dadurch die Daten noch immer sozusagen im Besitz der Kollaboration sind, und kein genereller „Open Access“ der Daten erfolgen musste“[23]. Hier kann die Teilchenphysik durchaus von anderen Bereichen wie z.B. der Astrophysik lernen, die mittlerweile sehr erfolgreich Daten veröffentlichen.

Beschleuniger werden üblicherweise von Generation zu Generation energiereicher, deshalb verschiebt sich auch das Physikprogramm. Für das Verständnis aktueller Daten sind aber die Daten von Vorgängerexperimenten enorm wichtig. So stammt z.B. die zur Zeit präziseste Messung der Z-Boson-Masse von den vier LEP-Experimenten, und ist somit über zehn Jahre alt.[24] Es ist nicht zu erwarten, dass diese Präzision von den Tevatron-Experimenten oder dem LHC in absehbarer Zukunft erreicht wird. Gruppen

22 Vgl. Ozerov et al. (2010).

23 “[...] knowledge is currently maintained at the Max-Planck-Institute for Physics at Munich and at DESY. The data and usage thereof is stil ‘owned’ by the original JADE collaboration, such that no general ‘open access’ to these data is granted” (Bethke (2010)).

24 Vgl. Nakamura et al. (2010).

wie die „Particle Data Group“ sind spezialisiert darin, notwendige Daten von Experimenten mit theoretischen Modellen zu kombinieren, um solche wichtigen Resultate für die Allgemeinheit der Teilchenphysiker aufzuarbeiten und bereitzustellen.

Die „Particle Data Group“ macht allerdings keine neue Analyse von alten Daten, sondern kombiniert nur schon publizierte Resultate. Insofern würde die PDG indirekt von einer LZA profitieren, indem neuere und präzisere Analyen von alten Daten bereitstehen könnten.

Im Rahmen des europäischen Projektes PARSE.Insight[25] wurden experimentelle und theoretische Teilchenphysiker zur Langzeitarchivierung befragt.[26] Insbesondere wurde die Frage gestellt, ob eine Langzeitarchivierung überhaupt als sinnvoll erachtet wird. Überwiegend wurde von beiden Gruppen eine LZA als „sehr wichtig“ bis „äußerst wichtig“ erachtet. Erstaunlicherweise wurde die Wichtigkeit von Theoretikern höher eingestuft als von Experimentalphysikern. Hieraus kann man sehen, dass besonders die theoretischen Physiker ein sehr vitales Interesse haben, alte Daten mit neuen Methoden zu analysieren oder Daten von unterschiedlichen Experimenten zu vergleichen. Es ist nicht auszuschließen, dass eine LZA der Daten langfristig das Verhältnis zwischen Theorie und Experiment verschiebt: Wenn die Rahmenbedingungen stimmen, erhalten Theoretiker einen viel breiteren Zugang zu Daten als bisher üblich. Dadurch können sie selber Daten analysieren und diese auch verstärkt publizieren. Die Grenzen zwischen Theorie und Experiment könnten weiter verschwimmen.

13.4 Interne Organisation

Die Labore, an denen die Experimente stattfinden, sind während der aktiven Analysephase für die Speicherung, Archivierung und Bereitstellung der Rohdaten verantwortlich. Die aktive Analysephase dauert typischerweise 3–5 Jahre länger als die eigentliche Datennahme. Zusammen mit der Bereitstellung der Daten werden auch Analyseumgebungen zur Verfügung gestellt. Die Finanzierung und notwendige Manpower ist im

25 Vgl. PARSE (2011).
26 Vgl. Holzner; Igo-Kemenes; Mele (2011).

Budget der Labore eingerechnet. Auf Seite der Experimente sind in dieser aktiven Analysephase noch erfahrene Leute. Allerdings muss auch beachtet werden, dass ein Abgang von Personal aus den Experimenten hin zu neueren Experimenten stattfindet, der nur bedingt durch neue Manpower in den bestehenden Experimenten kompensiert werden kann. Am Ende der Analysephase kann durchaus der Status quo erhalten bleiben, große Neuerungen sind aber nicht zu erwarten.

Hieraus kann man für die Langzeitarchivierung nach der Analysephase mehrere Punkte herausgreifen, die die interne Organisation betreffen:

- Die Organisation und Finanzierung eines Langzeitarchivs nach der aktiven Analysephase muss sichergestellt werden. Dies ist eine Aufgabe des beherbergenden Labors.
- Bereits während der aktiven Phase des Experimentes, am besten schon von Beginn des Experimentes an, muss auf die Phase der LZA hingearbeitet werden. Dies ist eine Aufgabe der Kollaborationen.
- Es ist sinnvoll, die LZA dort durchzuführen, wo die Experimente durchgeführt wurden, und wo somit das größte Know-how besteht. Es muss aber langfristig damit gerechnet werden, dass sich Forschungsschwerpunkte von Laboren verschieben und kein Interesse mehr am Archivieren von alten Daten besteht. Für diesen Fall sollte es ein zentrales Archiv geben, das diese Daten und deren Pflege übernehmen kann. Das Modell besteht also aus einigen verteilten Archiven und einem zentralen Archiv.
- Die Zusammenarbeit zwischen den einzelnen archivierenden Laboren muss intensiviert werden. Bisher war eine Zusammenarbeit auf informeller Ebene vorhanden, z.B. im Kontext der HEPiX Tagungen[27]. Im HEPiX Rahmen werden allgemeine IT-Themen von den IT Abteilungen der großen HEP-Institute und -Labore behandelt. Es ist die Aufgabe der ICFA und der von ihr ins Leben gerufenen DPHEP-Initiative diese Koordination aufzubauen und entsprechende Strukturen zu bilden. Dabei ist klar, dass diese Strukturen international sein müssen. Nationale Projekte müssen, um Erfolg zu haben, in einen größeren Kontext eingebettet sein.

27 Vgl. HEPIX (2011).

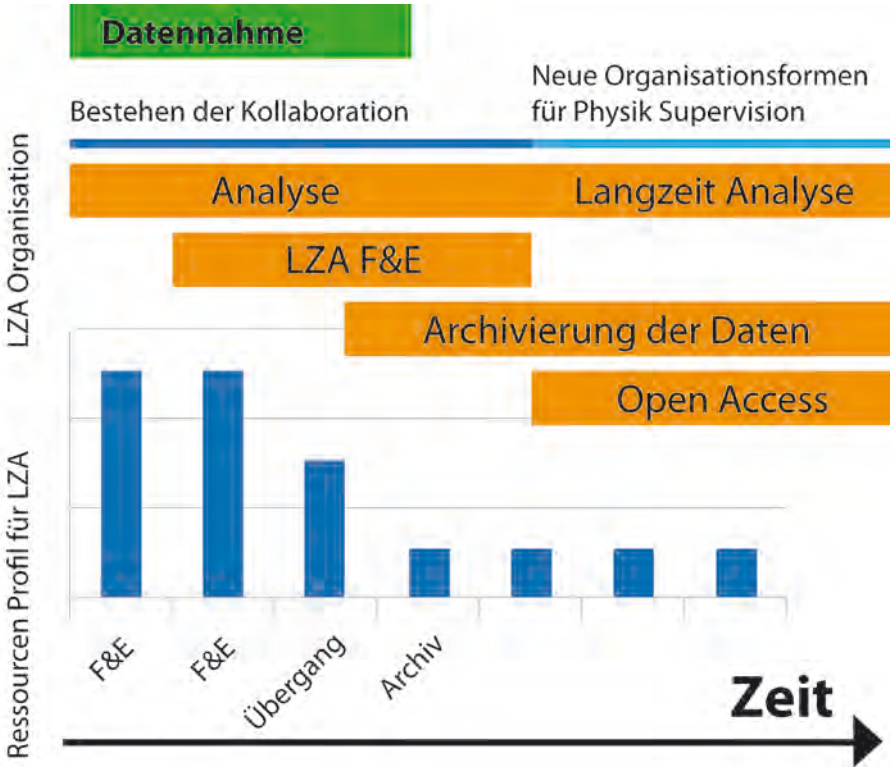

Abb. 2: Ein mögliches Organisationsmodell und Ressourcenprofil für LZA in der HEP[28]
In diesem Beispiel steht Physik Supervison für alle Aspekte, die den Zugang zu Daten, die Durchführung der Analyse, die Kontrolle und Veröffentlichung der Resultate betreffen.

- In der Vergangenheit haben die Kollaborationen einen organisatorischen Rahmen für den Zugang zu Daten und für die Freigabe von Veröffentlichungen von Analysen geliefert. Die Langzeitarchivierung bietet die Möglichkeit eines Zugriffs auf Daten nach Beendigung einer Kollaboration. Dies stellt die Forschergemeinde vor viele neue Fragen auch organisatorischer Natur. Die Antworten können von einer kompletten und uneingeschränkten Veröffentlichung der Daten bis hin zu einer Art Kuratorium ein breites Spektrum umfassen. Hier müssen die Kollaborationen selbst entscheiden, welchen Weg sie wählen. Gegebenenfalls muss z.B. unter der Obhut der ICFA

28 Vgl. DPHP (2009).

ein Kuratorium entstehen, welches den Zugang zu Daten und die Erlaubnis zur Publikation regelt. Ein besonderes Augenmerk muss dem Wegfall der kollaborationsinternen Qualitätskontrolle vor einer Veröffentlichung gelten. Dazu ist ein verstärkter Dialog mit anderen Disziplinen wie z.B. der Astrophysik notwendig, die Forschungsdaten sehr früh veröffentlichen. Die HEP kann von den Erfahrungen dieser Disziplinen lernen.

- Die verfügbare Manpower innerhalb der Experimente nimmt ab, da Experten zu neueren Experimenten abwandern. Gleichzeitig steigt auch innerhalb der IT-Abteilungen der Labore die Wichtigkeit neuerer Experimente. Es bedarf also aktiver Personalpolitik sowohl der Experimente als auch der Labore, um eine Langzeitarchivierung erfolgreich durchzuführen. Abbildung 2 zeigt ein mögliches Organisationsmodell und das damit verbundene Ressourcenprofil.[29]
- Die Langzeitarchivierung ist nur sinnvoll, wenn die Fähigkeit zur Analyse der Daten erhalten bleibt. Dieses muss vorbereitet werden, am Besten gleich von Beginn des Experimentes an. Da dies für aktuelle Experimente oder bereits abgeschlossene Experimente nicht immer möglich ist, müssen organisatorische und technische Wege gefunden werden, die eine Langzeitanalyse ermöglichen können.

13.4 Perspektiven und Visionen

Eine Langzeitarchivierung und Langzeitanalyse von Teilchenphysikdaten ist technisch, organisatorisch und finanziell herausfordernd – aber möglich und sinnvoll.

Es wird geschätzt, dass in der LZA-Phase eines Experimentes bis zu 10% weitere Publikationen erstellt werden – Publikationen die auf dem besten Verständnis von Detektor und Daten beruhen und neueste Analysetechniken und theoretische Modelle benutzen.

Es hat sich gezeigt, dass gerade theoretische Physiker an einer LZA der HEP-Daten interessiert sind. Wenn eine LZA mit einem offenerem Umgang mit den Experimentdaten einhergeht (Stichwort Open Access), so wird

29 Vgl. ebenda.

das Verhältnis zwischen Theorie und Experiment neu definiert werden. Aber auch das Verhältnis zu anderen Physikdisziplinen könnte sich ändern: Astrophysiker können z.B. Präzisionsmessungen aus einem bestimmten Energiebereich eines HEP-Experiments in ihre eigenen Messungen einbinden.

Die von der ICFA ins Leben gerufene DPHEP-Initiative hat in den bisher zwei Jahren ihres Bestehens mehrere Beiträge zu wichtigen Konferenzen der Teilchenphysik geliefert. Mehrere Treffen der DPHEP-Initiative zeigen, dass diese Problematik von vielen Laboren und Experimenten ernst genommen wird.

Das Bewusstsein für die speziellen Probleme der LZA steht erst am Anfang und es ist wichtig, dass dieses Bewusstsein gestärkt wird. Es muss klar sein, dass schon zu Beginn eines Experimentes an die Zeit nach dem Experiment gedacht wird. Gerade in der Ausbildung von Nachwuchswissenschaftlern an Unis und Forschungslabors muss dies ein wichtiger Bestandteil werden.

Das Problem der „Ownership“ der Daten muss gelöst werden. Kollaborationen müssen eine neue Kultur im Umgang mit Daten und Veröffentlichungen entwickeln, die nachhaltig in einer Phase der Langzeitarchivierung und Langzeitanalyse funktioniert.

Ein erfolgversprechendes Modell für Langzeitarchive ist, die Rohdaten an den einzelnen Laboren zu archivieren, und ein zentrales Archiv für weitere Daten wie Publikationen und Übernahme von Rohdaten anderer Labore zu benennen. Wichtig ist eine Zusammenarbeit zwischen den einzelnen Archiven, zum Beispiel unter der Führung der ICFA/DPHEP Initiative.

Die Nutzung der archivierten Daten steht und fällt mit der Qualität der Analysesoftware und der Dokumentation. Software zu archivieren und dauerhaft einsetzbar zu machen ist von einer bisher noch unterschätzten Komplexität. In diesem Bereich ist weitere Forschung notwendig.

Eine Möglichkeit, die Komplexität der Analyseproblematik zu vereinfachen ist eine (verlustbehaftete) Konvertierung der Daten in einfachere Formate. Diese ermöglichen dann nur einige wenige Analysen, sind also kein Ersatz für eine vollständige Langzeitarchivierung. Ein solches Format bietet allerdings die Chance, diese Daten für Studenten oder sogar für eine breite Öffentlichkeit aufzubereiten. Dadurch kann sowohl das Bewusstsein für die Problematik der Langzeitarchivierung bei den Studenten gesteigert

werden als auch die Sichtbarkeit der Teilchenphysik über den akademischen Bereich hinaus erhöht werden.

Die Geldgeber müssen für die Thematik sensibilisiert werden. Der Aufbau und Betrieb eines Experimentes hat große Mengen Forschungsgelder benötigt. Der verantwortungsvolle Umgang mit diesen Geldern gebietet es, die gewonnenen Daten langfristig zu archivieren und einer breiteren Öffentlichkeit zukommen zu lassen. Der Einsatz weiterer Forschungsgelder beträgt dabei ein Bruchteil dessen, was für Aufbau und Betrieb notwendig war – und garantiert interessante weitere Forschungsergebnisse.