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

Langzeitarchivierung von Forschungsdaten

Eine Bestandsaufnahme

Kapitel 14
Astronomie und Astrophysik

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

14 Astronomie und Astrophysik

Harry Enke, Joachim Wambsganß

14.1 Einführung in den Forschungsbereich

Die Astronomie ist eine der ältesten Wissenschaften. Schon vor vielen tausend Jahren haben alle Hochkulturen den Lauf der (Fix- und Wandel-) Sterne verfolgt. Von der Antike bis ins Mittelalter war die Astronomie eine sehr angewandte und praxisnahe Wissenschaft. Die Positionsbestimmung anhand der Sterne war für die Navigation, insbesondere in der Seefahrt, unverzichtbar. Aber auch für die Landwirtschaft war die Vorhersage der Jahreszeiten (und damit etwa des Nil-Hochwassers) von sehr großer Bedeutung. In der Neuzeit hat sich die Astronomie in enger Verbindung mit der Physik und der Mathematik weiterentwickelt.

Im Gegensatz zu den meisten anderen Naturwissenschaften kann die Astronomie keine Experimente durchführen. Von früh an sammelten die Astronomen daher (Beobachtungs-) Daten und versuchten, sie zu verstehen, zu interpretieren und Vorhersagen zu machen. Der älteste bekannte Sternkatalog geht zurück auf Claudius Ptolemäus (2. Jh. n. Chr.). Er hat die Positionen von etwa 1000 Sternen festgehalten und veröffentlicht. Bis ins Mittelalter gab es kaum wesentliche Fortschritte. Galileo Galilei war im Jahre 1609 der erste, der das neuentdeckte Teleskop gen Himmel richtete und damit „neue Welten“ eröffnete. Die Kepler'schen Gesetze der Planetenbewegung (1609 bzw. 1618) und das Newtonsche Gravitationsgesetz (1686) stellten die Astronomie dann auf feste mathematisch-physikalische Grundlagen. Im Jahre 1860 entwickelten der Chemiker Robert Bunsen und der Physiker Gustav Kirchhoff in Heidelberg die Spektralanalyse, die eine völlig neue Dimension in der Astronomie eröffnete: Die Nutzung der Spektroskopie gestattete es, neben den „klassischen“ Eigenschaften von Sternen – Helligkeit und Position – nun auch chemische

und physikalische Eigenschaften zu messen, wie etwa Elementhäufigkeiten, Druck und Temperatur. Dies war die Geburtsstunde der „Astrophysik“. Nun konnte man tatsächlich auch physikalische Parameter der Sterne bestimmen.

Bis zur Entdeckung der Fotografie und ihrer Anwendung auf Sterne Ende des 19. Jahrhunderts wurden astronomische Messungen und Beobachtungen als Zeichnungen und Tabellen überliefert. Seither wurden Sternpositionen und Helligkeiten mithilfe der Astrofotografie auf Fotoplatten festgehalten. Durch lange Belichtungszeiten konnte man auch schwächere Sterne nachweisen und untersuchen, als das Auge beim Blick durch ein Teleskop erkennen konnte. In den 1980er-Jahren wurde dann die analoge Fotografie durch die CCD-Chips abgelöst, zunächst mit heute winzig klein anmutenden 512×512 Pixel-Sensoren (die mit 0,25 Megapixeln deutlich kleiner waren als jede Handy-Kamera heute). Die Astronomie war allerdings eine der treibenden Kräfte hinter der Entwicklung der CCDs. Heutige Teleskope arbeiten mit Gigapixel-Kameras.

Neben den beobachtenden Astronomen gab es immer auch die Theoretiker. Früher waren dies oft Mathematiker, heute eher computer-affine Astrophysiker. Theoretische Astrophysiker haben schon in den 1950er-Jahren die Entwicklung der Computer genutzt und sogar vorangetrieben, um die physikalischen Bedingungen etwa im Inneren von Sternen durch Computersimulationen besser verstehen und modellieren zu können. Auch die Modellierung von Supernova-Explosionen oder die sogenannten „Large-Scale-Structure“-Simulationen, die die Entwicklung der großräumigen Struktur im Universum nachvollziehen, sozusagen vom Urknall bis heute, nutzen alle verfügbaren Computerresourcen und produzieren enorm große (astronomische) Datenmengen. So war und ist die Astronomie immer eine daten-getriebene und daten-dominierte Wissenschaft.

14.2 Kooperative Strukturen

Da die Zahl der Astronomen in jedem Land jeweils relativ klein ist, und da man früh erkannt hat, dass es am effizientesten ist, die weltbesten Teleskop-Standorte gemeinsam zu nutzen, war die Astronomie von je her

eine sehr internationale Wissenschaft. Schon früh haben sich kooperative Strukturen entwickelt und ganz natürlicherweise wurden Konzepte zur Standardisierung von Datenformaten und Speicherung von astronomischen Daten vorgeschlagen und entwickelt.

Als nationale Organisation im deutschsprachigen Raum existiert zum einen die 1863 gegründete Astronomische Gesellschaft (AG), zum anderen seit den 1960er-Jahren der Rat deutscher Sternwarten (RDS). Der RDS „vertritt die gemeinsamen Interessen aller deutschen astronomischen Institute gegenüber Förderinstitutionen, Regierungen und internationalen Organisationen."[1] Gegenwärtig gibt es Bestrebungen, AG und RDS miteinander zu verschmelzen.

Die „Internationale Astronomische Union" (IAU), gegründet 1919, ist mit über 10000 Mitgliedern die weltweite Dachorganisation der professionellen Astronomen. Sie organisiert wissenschaftliche Zusammenkünfte und themenzentrierte Arbeitsgruppen. Eine der zwölf Abteilungen führt den Ausschuss „Dokumentation und astronomische Daten", der sowohl für die Erhaltung und Entwicklung des wichtigsten astronomischen Datenformates, des Flexible Image Transport System (FITS) zuständig ist, wie auch für die Working Group on Astronomical Data[2] (WGAD), die u.a. die Astronomie in dem interdisziplinären CODATA (Committee on Data for Science and Technology des ICSU) vertritt. Die IAU-Working Group zum Virtual Observatory (VO) hat hier ebenfalls ihren Platz[3], mit der Aufgabe, die durch

1 Vgl. RDS (2011).

2 "The Working Group on Astronomical Data (WGAD) of the International Astronomical Union (IAU) is a working group concerned with issues around the access of astronomical data. Current issues include: -Promoting open access to observatory archives; -Promoting the concept of the Virtual Observatory; -Promoting common formats for astronomical data to facilitate sharing and entry into public-domain databases; -Discussing the implications (for data centres, for the VO, for astronomers) of future terabyte astronomical databases." (WGAD (2011)).

3 "The International Virtual Observatory is one of the rare truly global endeavours of astronomy. Many projects, each with its own goals, have been set up around the world to develop the IVO. The International Virtual Observatory Alliance (IVOA) is an alliance of the VO projects, with the aims of managing communication between VO projects, defining a common roadmap, and managing propositions for the definition and evolution of IVO standards through the IVOA Working Groups. [...] Its primary role is to provide an interface between IVOA activities, in particular IVOA standards and recommendations, and other IAU standards, policies, and recommandations. [...] It is responsible for approving the standards pro-

das VO entwickelten Standards und Verfahren mit der IAU zu koordinieren und abzugleichen.

Kooperative Strukturen gab und gibt es in verschiedenen Varianten. Oft sind sie in erheblichem Maße durch die Art der Teleskope geprägt, besser gesagt: durch die Frequenzbereiche dieser Instrumente. So wurde bereits in den 1960er-Jahren die europäische Südsternwarte (ESO) gegründet. Ursprünglicher Zweck war es, den Südhimmel für die Astronomen Europas zugänglich zu machen. Daraus hat sich in 50 Jahren das modernste optische Teleskop der Welt entwickelt, das Very Large Telescope (VLT): Vier Teleskope mit jeweils 8,2m Spiegeldurchmesser, die auf dem Paranal in Chile stehen. Auch Radioastronomen haben sich früh international zusammengefunden, um gemeinsam neue Techniken zu entwickeln und zu nutzen, insbesondere für die Very Long Baseline Interferometrie (VLBI).

Um das Jahr 2002 haben einige visionäre Astronomen erkannt, dass die Zukunft der Astronomie stark davon abhängen wird, inwieweit man weltweit einheitliche Standards, Protokolle und Datendefinitionen entwickelt. Dies war die Grundlage für die IVOA, die International Virtual Observatory Alliance[4]. Die Grundidee des Virtual Observatory ist es, jedem Astronomen auf der Erde Zugang zu sämtlichen Beobachtungsdaten zu geben, sozusagen „the universe at your finger tips“. Eine wichtige Aufgabe auf diesem Weg ist es, für den „Multiwavelength-View of The Universe“ die Daten der digitalen astronomischen Archive verschiedener Wellenlängen mit standardisierten Interfaces zu versehen, die Daten dann zu publizieren und sie über einen einheitlichen Zugang zugreifbar zu machen. Die wissenschaftliche Struktur der IVOA ist multinational. Da die „Funding Structures“ von Land zu Land sehr verschieden sind, gibt es in vielen Ländern jeweils eine nationale Gruppierung, die wiederum Mitglied bei der IVOA sind. In Deutschland ist dies GAVO, das „German Astrophysical Virtual Observatory“. Mittlerweile sind 18 solcher nationalen Gruppen an der IVOA beteiligt, die jeweils ihre eigenen Beiträge zum VO einbringen und an der Erarbeitung von Standards und

posed by IVOA, after checking that there has been a process of consultancy according to the IVOA procedures, and that the proposed standards are consistent with other IAU approved standards (e.g., FITS, coordinate standards, etc.).” (IAU-IVOA (2011)).

4 Vgl. IVOA (2011).

Protokollen und deren Referenz-Implementation arbeiten. Für die Mehrzahl der astronomischen Datenarchive im optischen Bereich hat das VO bereits akzeptable Lösungen und Standards gefunden, jedoch sind andere Bereiche wie Radio- oder Gammastrahlenastronomie, noch nicht vollständig abgedeckt.

GAVO wird seit 2003 als BMBF-Projekt gefördert, getragen von astronomischen Forschungsinstituten und universitären Einrichtungen.[5] Die GAVO-Arbeit umfasst zwei Schwerpunkte. Zum einen sollen astronomische Datenarchive in Deutschland gemäß den Standards der IVOA etabliert werden. Zum anderen werden geeignete Strukturen, Metadaten und Formen erarbeitet, in denen Resultate von astrophysikalischen Simulationen im VO publiziert und als Datenquellen wie Beobachtungsdaten genutzt werden können. Hier hat GAVO für die Millennium-Simulation (Max-Planck-Institut für Astrophysik) grundlegende Arbeiten geleistet. Mittlerweile sind das Virtuelle Observatorium und dessen Arbeiten fester Bestandteil der Astronomie.[6]

Im europäischen Rahmen wurde das FP7-Projekt EURO-VO gefördert, wobei der Schwerpunkt auf der Implementation von Tools und Software liegt, die mit Standards und Protokollen der IVOA arbeiten[7]. Hauptbeteiligte sind hier Frankreich und Großbritannien, wobei insbesondere das CDS Strasbourg, die University of Edinburgh als Repräsentant des

5 Ursprünglich initiiert wurde GAVO vom Max-Planck-Institut für Extraterrestrische Physik (MPE) und vom Leibniz-Institut für Astrophysik Potsdam (AIP). Seit der zweiten Förderphase hat das Zentrum für Astronomie der Universität Heidelberg (ZAH) die Leitung inne, weitere Universitäts- und Max-Planck-Institute sind beteiligt.

6 Die Jahrestagung der Astronomischen Gesellschaft in Heidelberg im September 2011 etwa hatte den Titel: “Surveys and Simulations – The Real and the Virtual Universe”.

7 “The EURO-VO project aims at deploying an operational VO in Europe. Its objectives are the support of the utilization of the VO tools and services by the scientific community, the technology take-up and VO compliant resource provision and the building of the technical infrastructure.” (EuroVO (2011)).

AstroGrid[8] UK, sowie die ESO zu nennen sind. Die Universität Heidelberg ZAH) ist im Euro-VO als Vertreter der deutschen Astronomie beteiligt.[9]

Während das VO sich vor allem mit der Standardisierung der Datenformate und Software-Tools sowie der Publikation von Beobachtungsarchiven befasst, wurden andere Probleme der IT-basierten Infrastruktur durch das AstroGrid-D-Projekt adressiert.[10] Ziel von AstroGrid-D war es, die kollaborativen IT-Infrastrukturen des Grid für die Astronomie nutzbar zu machen und die schon vorhandenen Computer- und Datenspeicher-Ressourcen der beteiligten Institute und der Rechenzentren „in eine bundesweit einheitliche Forschungs-Organisationsstruktur für verteiltes kollaboratives Arbeiten“[11] einzubinden. Darüber hinaus wurde auch die Einbeziehung von Instrumenten als Ressourcen in das Grid am Beispiel der robotischen Teleskope angestrebt.[12] Einige der vom AstroGrid-D errichteten Dienste und Infrastrukturen werden nach Ende der Förderung in 2009 von beteiligten Instituten aufrechterhalten, wie z.B. die VO (Virtuelle Organisation), AstroGrid und deren Management als Basis für die

8 Das AstroGrid UK hat, im Gegensatz zur Namensgebung, nur wenig die Nutzung der Grid-Technologie als IT-Infrastruktur befördert, jedoch sehr wichtige Beiträge zur VO-Technologie erbracht.

9 Ein wichtiges Instrument der Planung und Lenkung von Förderung gemeinsamer europäischer Projekte ist das Astronet im Rahmen von ERA-Net. “ASTRONET was created by a group of European funding agencies in order to establish a comprehensive long-term planning for the development of European astronomy. The objective of this effort is to consolidate and reinforce the world-leading position that European astronomy has attained at the beginning of this 21st century.” (AstroNet (2011)) Vom Astronet wurde beispielsweise die Gründung des Astronomical Software Laboratory initiiert, welches in Zusammenarbeit und Ergänzung zum Virtual Observatory die Förderung von den in der Community am meisten benutzten Software-Tools zur Aufgabe hat.

10 Das AstroGrid-D wurde als German Astronomy Community Project (GACG) gestartet, die Namensänderung erfolgte kurz nach Projektstart. Neben den schon in GAVO aktiven Einrichtungen AIP und MPE haben sich MPA, AEI, ZAH als Verbundpartner und LMU, MPIfR und MPIA von den astronomischen Forschungseinrichtungen beteiligt, als Partner aus der Informatik waren es ZIB und TUM sowie die UP als assoziierter Partner, als Rechenzentren waren LRZ, RZG und FZK assoziiert. Das AIP, das schon in GAVO die Möglichkeiten der Grid-Infrastruktur evaluiert hatte, übernahm die Leitung des Projekts.

11 Das AstroGrid-D, dessen Aufbau und Implementierung auf ca. sechs bis sieben Jahre konzipiert war, wurde, aufgrund des generellen Wechsels in der Förderkonzeption von Grid-Projekten, nur vier Jahre gefördert.

12 Eine ausführlichere Darstellung von AstroGrid-D findet sich in Enke et al. (2010).

Teilnahme der Astronomie-Community am D-Grid oder die Unterstützung bei der Nutzung von D-Grid-Ressourcen durch astronomische Projektgruppen.

Europaübergreifend gibt es neben Astronet, das für die EU eine ähnliche Funktion wie der RDS für Deutschland hat, die bereits erwähnte ESO, die in der chilenischen Atacama-Wüste zwei Beobachtungsstandorte betreibt und damit die bedeutendsten bodengestützten europäischen Beobachtungsinstrumente betreibt. Darüber hinaus betreibt die ESA (European Space Agency) ein Programm mit Satelliten-gestützten Teleskopen verschiedener Wellenlängenbereiche. ESO wie ESA haben im Bereich der Datenhaltung einen nicht unerheblichen Einfluss auf die in der Community üblichen Verfahren. Für den Bereich der Radio-Astronomie bestehen mit IRAM und JIVE zwei europäische Einrichtungen. Die FP7-Aktivitäten der Radio-Astronomie werden über RadioNet koordiniert. Die Astroteilchen-Physik hat sich durch ASPERA (ERA-NET)/ApPEC organisiert, die Infrarot-Astronomie in OPTICON, die Helioseismologie im HELAS-Netzwerk. Nicht zuletzt durch das Virtual Observatory sind hier auch erste Ansätze zu weiterer Interoperabilität der Archive und Datenspeicher zu sehen.

14.3 Daten und Metadaten

Die astronomischen Beobachtungsdatenarchive

Eine konzise Charakterisierung der Motive und der Erkenntnisse, die zur Virtual-Observatory-Initiative in der Astronomie geführt haben (hier in Bezug auf die genutzten Koordinaten-Systeme) gibt A. Rots in „Space-time Coordinate Metadata“:

> “In the past, most data archives have not been extremely concerned with dotting every i and crossing every t, in a meticulous specification of all the details of all coordinate axes. Such context-dependent defaults are quite acceptable: issues are generally well-defined, obvious, and clear for single-observatory observations, even when not all is explicitly specified. However, there are no global defaults in the VO. All implicit assumptions need to be made explicit since they will not be obvious anymore. One must be able to transform the coordinates of two observations to a common coor-

dinate system; this means that every little tidbit of information needs to be documented in the metadata."

Das FITS[13] (Flexible Image Transport System) ist von der IAU als standardisiertes Format angenommen worden. Es ist ein Datei-Format, das in den 1980er-Jahren entwickelt wurde.[14] Nicht nur die sehr beschränkte Auswahl an festgelegten Metadaten für das FITS, auch die Möglichkeit, eigene Key/Values hinzuzufügen, ohne dass eine Systematik vorgegeben ist, machen deutlich, dass dieses Format eine nicht hinreichende Lösung für die notwendigen Metadaten zur Beschreibung der astronomischen Daten ist. Das FITS bleibt jedoch das Basiselement für digitale astronomische Beobachtungsdaten, das auch von astronomischer Software und Instrumenten, von vielen genutzten Script-Sprachen wie perl oder python und den genutzten Bildverarbeitungsprogrammen gelesen, interpretiert und geschrieben werden kann.[15]

Neben den FITS-Dateien sind in den letzten Jahren verstärkt Datenbanken als Speichermedium in Erscheinung getreten. Dies gilt insbesondere für die effiziente Ablage und Nutzung tabellarischer Informationen. Hierzu zählen raumzeitliche Koordinaten, Geschwindigkeiten, Helligkeiten und andere Messgrößen sowie Referenzen auf Dateien, die komplexe-

13 "This document, hereafter referred to as the 'standard', describes the Flexible Image Transport System (*FITS*) which is the standard archival data format for astronomical data sets. Although *FITS* was originally designed for transporting image data on magnetic tape (which accounts for the 'I' and 'T' in the name), the capabilities of the *FITS* format have expanded to accommodate more complex data structures. The role of *FITS* has also grown from simply a way to transport data between different analysis software systems into the preferred format for data in astronomical archives, as well as the online analysis format used by many software packages." (FITS (2010a), S. 1).

14 Es besteht aus einem Header, der Key/Value/Comment Triplets mit max. 80 Zeichen im ASCII-Format enthält. Der Key darf maximal acht Zeichen lang sein. Der Header beschreibt durch einen Satz von Metadaten den Datei-Inhalt. Dieser Inhalt kann sowohl aus einfachen Tabellen in textueller Form als auch aus binären Daten beliebiger Art, meist Bildinformationen, bestehen. Ein kleiner und selbst nicht zum IAU-Standard gehörender Satz von feststehenden Key/Value Paaren (mandatory/reserved) des Headers wird durch eine Vielzahl von Key/Value Paaren erweitert, die spezifisch für verschiedene Beobachtungsarten, Teleskope usw. sind. Es hat große strukturelle Ähnlichkeiten zum NetCDF-Format, das beispielsweise in der Klimafolgenforschung häufig Verwendung findet.

15 Es gibt eine FITS-Extension-Registry, die derzeit 15 von der IAU Working Group akzeptierte Erweiterungen das FITS Standards aufweist, sowie drei weitere vorgeschlagene, die noch in der Prüfungsphase sind. (Vgl. FITS (2010b)).

re Informationen bereithalten (Images, Spektren etc). Die SDSS-Datenbank z.B. enthält mehr als sechs Terabyte an tabellarischer Information über mehrere hundert Millionen astronomische Objekte.[16] Die von den Katalogen referenzierten Spektren und Images (weitere ca. 60 Terabyte) werden in FITS-Dateien in einem regulären Dateisystem vorgehalten.

Um die Interoperabilität der Datenarchive (von Beobachtungsdaten) sicherzustellen, hat das VO standardisierte Formate und Metadaten für einfache physikalische Messgrößen, raum-zeitliche Koordinaten sowie für eindimensionale Spektren erarbeitet. Weitere Standardisierungen von Formaten und Diensten werden laufend hinzugefügt, da längst nicht alle Datentypen erfasst sind. Die Metadaten-Systematik ist in „Ressource-Metadata“[17] festgehalten, welches den Ausgangspunkt für weitere, auf Teilbereiche bezogene Standards bildet. Für die generischen Metadaten wird, soweit als möglich, der Dublin Core-Metadatenstandard herangezogen, wobei die Dublin Core-Elemente „Language“ und „Relation“ nicht genutzt werden.

Ein im VO entwickeltes semantisches Klassifikationssystem von Metadaten für physikalische Größen sind die Unified Content Descriptors (UCD). Die UCD bestehen aus einer formalen (von einem Komitee aus Wissenschaftlern) kontrollierten Wortliste und wenigen sehr einfachen Regeln, um mit diesen (atomaren) Wörtern eine kurze Typisierung einer physikalischen Größe zu erstellen, die dann z.B. in einer Tabelle aus einem astronomischen Katalog verwendet wird. Die UCD sind sowohl für einen Menschen lesbar als auch für Computerprogramme verwertbar, da es nur sehr begrenzte zulässige Kombinationen von Worten gibt. Komplexere Systeme, z.B. „UTypes“ werden entwickelt, da die UCD für kompliziertere Datenmodellierung nicht trennscharf genug sind, ihre Nutzung ist jedoch beschränkt auf wenige Anwendungen.

---

16 “We strongly recommend that the archive data be hosted in a DBMS unless
- the archive has a large software development budget (unusual for astronomical archives)
- there are specific and unusual requirements (such a high security or mission critical data), or
- the data access patterns are unusual and restricted (for example, in some particle physics data set, the main type of query is searching for a few rare events in a huge data set).” (Thakar et al. (2008)).

17 Vgl. IVOA (2007).

Neben der Standardisierung des Content ist ein umfangreicherer Satz an Dokumenten des VO vorhanden, der sich auf die Applikationen und Formen des Austausches dieser Daten über die interoperablen Schnittstellen des VO bezieht. Dieses sind Dienste wie ConeSearch (Suche nach Objekten durch Angabe von Koordinaten), Image Access und Spectral Access (jeweils spezielle auf einen Content abgestimmte Dienste) und andere. Diese Dienste werden in einer Registrierung (VO-Registry) erfasst, deren Schema durch VO-Resources und VO-Identifiers festgelegt wird. Um die Interoperabilität der Dienste zu sichern, gibt es Web- und Grid-Schnittstellen (z.B. das TableAccessProtocol (TAP) oder den Grid-WorkerService (GWS)).

Die Implementierungen dieser Dienste beginnen jetzt in der Community partiell akzeptiert und genutzt zu werden. Dies ist jedoch stark davon abhängig, inwieweit sie einen Effizienzgewinn in der wissenschaftlichen Arbeit darstellen. Viele der über das Internet verfügbaren Daten-Archive bieten meistens eine, in der Funktion noch eingeschränkte, VO-Komponente an. Insbesondere Institute mit größeren Datenarchiven unternehmen gezielte Anstrengungen, die VO-Kompatibilität zu erhöhen. Ein schrittweiser Übergang ist auch deswegen möglich, weil das VO nicht eine Umorganisation der Daten selbst, sondern nur eine Zugriffsschicht erfordert, die es möglich macht, die Datenorganisation intern nach anderen Kriterien vorzunehmen als für die Publikation erforderlich.

Die astrophysikalischen Simulationsdatenarchive

Astrophysikalische Phänomene und Vorgänge können in irdischen Labors oder auch mit Großgeräten nicht erzeugt werden. Durch Simulationen in einer Virtuellen Realität können jedoch Theorien und Erklärungen dieser Effekte dargestellt und untersucht werden. Solche Simulationen sind zu einem unverzichtbaren Instrument herangewachsen. Sie können als virtuelles Labor betrachtet werden, in dem die Interpretationen von Beobachtungen und Theorien auf den Prüfstand gebracht werden. Damit stellen sie ein essentielles Bindeglied zwischen Theorie und Beobachtung dar.

Die Durchführung einer Simulation erzeugt jedoch zunächst Rohdaten, die – wie bei Beobachtungsdaten – durch eine intensive Analyse in wissenschaftlich verwertbare Informationen über Strukturen, Wechsel-

wirkungen usw. verwandelt werden. Diese Analyse besteht in der Bildung von Statistiken, von Korrelationen, der Verfolgung der zeitlichen Entwicklung von physikalischen Messgrößen und nicht zuletzt der Visualisierung dieser Entwicklung. Hier sind extrem „langzeitliche“ Entwicklungen (vom Urknall zur Jetztzeit ~14,3 Mrd. Jahre) mit „kurzzeitigen“ Vorgängen wie der Bildung von Galaxien, von Sternentstehung und -explosionen, deren Dynamik auf völlig verschiedenen Zeitskalen angesiedelt ist, zu kombinieren.

Die Datenformate der Simulationen sind weit weniger vereinheitlicht als die der Beobachtungen.[18] Getrieben von der Bestrebung, Beobachtungs- mit Simulationsresultaten zu vergleichen, sind jedoch auch hier erste Ansätze zur Standardisierung von Dateiformaten entstanden. Hinzu kommt, dass sich auch für Simulationsresultate die Erfassung in Datenbanken als effizient erweist. Es lassen sich dann auch viele wissenschaftliche Fragestellungen in einer gemeinsamen Sprache (SQL) abfassen und die Vergleichbarkeit von Fragestellung und Resultat wird erhöht. Auch diese Entwicklung ist über das VO entschieden beschleunigt worden. Hinzu kommen Metadaten, die die bei den Simulationen genutzte Software und Hardware-Architektur beschreiben. Und nicht zuletzt müssen die (oft verteilten) Daten der Simulationen verwaltet werden. Hierfür sind erst in Ansätzen geeignete Metadaten-Systeme vorhanden, die das gemeinsame Management von Simulationsdaten und Software unterstützen.

Die derzeit umfangreichsten Simulationen haben 2048^3=8,6 Mrd. Teilchen[19] und umfassen bis zu 150 Snapshots, die jeweils etwa 0,5 Terabyte

18 Die rasche Entwicklung der Hardware (von 8Bit zu 64 Bit, von FloppyDisk zu FlashSpeicher), unterschiedliche Architekturen der Großrechner und der aus Commodity-Hardware gebauten Parallel-Rechner und damit verbundene Entwicklung von Dateiformaten ist einer der hierfür bestimmenden Faktoren. Und natürlich ist die Tendenz eines jeden, der ein Simulations-Programm schreibt oder nutzt, das ihm bequemste Format zu wählen.

19 Eine kosmologische Vielteilchen-Simulation, auf die hier Bezug genommen wird, benötigt für jedes dieser Teilchen Raum- und Geschwindigkeitskoordinaten, Masse, ggf. elektrische Ladung usw. Ein Satz von acht oder mehr Größen für jedes Teilchen wird für jeden Zeitschritt gespeichert. Die Teilchenmassen liegen bei ca. 10^8 Sonnenmassen, zwischen gespeicherten Zeitschritten (Snapshots) liegen ca.100 Millionen Jahre. Neben den Simulationen, die Teilchen interagieren lassen, gibt es Gittercodes mit feststehenden Zellen, in denen die zeitlichen Änderungen von physikalischen Größen wie elektromagnetischer Felder oder gravitativer Felder hoher Dichte untersucht werden, die sich besser für die Untersuchung z.B.

Speicherplatz benötigen. Die Rohdaten einer solchen Simulation werden in der Regel nicht an einem einzigen Supercomputer erzeugt. Die benötigte Rechenzeit von vielen Millionen CPU-Stunden muss verteilt über verschiedene Rechenzentren in der ganzen Welt eingeworben werden.

Datenvolumina und Wachstumsraten

Sowohl für Beobachtungsinstrumente als auch für Simulationen halten die Datenmengen und deren Wachstumsraten Schritt mit dem technisch gerade Machbaren. Planungs- und Bauzeit eines Instrumentes für ein Teleskop oder einer Satellitenmission liegen im Bereich von zehn Jahren, die Betriebszeiten dann zwischen drei und zehn Jahren. Die Datenaufnahme und deren Auswertung ist fester Bestandteil der Planung für das Instrument. Während es bis in die 1990er-Jahre durchaus gängige Praxis war, die genommenen Daten und deren Archivierung der jeweiligen Wissenschaftler-Gruppe anheimzustellen und der Betreiber des Instruments keinerlei Archivierung vornahm, ist mit dem Aufkommen von systematischen Himmelsdurchmusterungen (Surveys) – bei den Satellitenmissionen ohnehin – ein Umdenken erfolgt. Das derzeitige Datenarchiv der ESO, welches (bereits reduzierte) Daten der ESO Teleskope- und Satelliteninstrumente aufnimmt, enthält derzeit 65 TB, mit einer Wachstumsrate von 15 TB pro Jahr. Die ESO erwartet eine Steigerung auf 150 TB pro Jahr in naher Zukunft.[20] Neben den ESO-Archiven werden von einzelnen Kollaborationen ausgedehntere Archive aufgebaut, die neben den reduzierten Rohdaten ausgedehntere Objekt- und Auswertungskataloge enthalten. Ein Beispiel ist die mit dem 2013 in Betrieb gehenden MUSE (Multi Unit Spectrographic Explorer) Instrument verbundene Kollaboration, welche 250 TB an Daten archivieren wird. Etwa ein Drittel dieser Daten wird von der ESO in ihr Datenarchiv aufgenommen.

von sehr kurzzeitigen Vorgängen (einige Jahre bis einige Sekunden) beim Zusammenstoß von Schwarzen Löchern, Supernovae usw. eignen.

20 Die Vorgaben der ESO für die Archivierung sind sehr restriktiv und nicht ausreichend für eine vollständige Erschließung des wissenschaftlichen Gehalts. (Vgl. ESO-ARCH (2011)).

Jahr	Survey	Datenmenge	Inhalt
1994	Digitzed Sky Survey (DSS)	~73 Gigabyte	Digitialisierte Fotoplatten
2001	Catalog of DSS	~16 Gigabyte	89 Mio. Objekte
1997–2001	2 Micron All Sky Survey (2MASS)	~200 Gigabyte	300 Mio. Punktquellen und 1 Mio. ausgedehnte Quellen
2000–2006	Sloan Digital Sky Survey (SDSS)	~6 Terabyte Katalog, 70 Terabyte Images u. Spektren	ca. 200 Mio. Objekte und 1 Mio. Spektren
2011–	LOFAR	~2 Terabyte / Tag	Radio Daten (Images)
2013–	GAIA (Satelliten-Instrument)	50 Gigabyte / Tag	1 Milliarde Objekte
ab 2014–2024	LSST	30 Terabyte / Nacht	60 Petabyte, 20 Millarden Objekte

Tab. 1: Eine Auswahl von verfügbaren digitalen astronomischen Archiven.[21]

Da nahezu alle Instrumente von internationalen Kollaborationen gebaut werden, sind nationale Archivierungsstandards von vornherein ausgeschlossen. Die jetzt in Bau- oder Planungsphase befindlichen astronomischen Instrumente haben als selbstverständlichen Bestandteil einen Daten-Management-Plan. Es ist jedoch noch längst keine weitergehende, übergreifende Standardisierung bei Metadaten erreicht. Die Archive der Kollaborationen sind verteilt über viele Institutionen, wobei sich als Trend auch feststellen lässt, dass sich in jedem Land einige Institute beim Datenmanagement besonders engagieren. Die Situation ist sehr verschieden mit der experimentellen Hochenergie-Physik, da die Datenquellen wesentlich vielfältiger sind, wobei die Datenmengen jedoch durchaus ähnliche Größenordnungen haben. Der Weg, der vom VO gewählt wurde, um die verteilten und unter-

21 Für die GAIA Mission werden derzeit 10^{21} Floating Point Operations (Computer-Rechenschritte) geschätzt. Die derzeitig verfügbaren Höchstleistungsrechner würden drei bis vier Jahre benötigen.

schiedlichen Datenarchive durch standardisierte Interfaces und Verfahren der Community zur Verfügung zu stellen, ist dieser Situation angemessen. Über die bei den Simulationen vorliegende Situation sagt S. Gottlöber: „Wenn man 20 Jahre zurückschaut, stellt man fest, dass auch die numerischen Simulationen der gravitativen Strukturbildung dem Mooreschen Gesetz folgen. Während die ersten Simulationen in den achtziger Jahren etwa 32^3 Teilchen behandelt haben, sind inzwischen 1024^3 Teilchen der Standard für kosmologische Simulationen mit dunkler Materie, ein Wachstum von 2^{15} über 20 Jahre. Das Moore'sche Gesetz würde 2^{13} voraussagen. Der verbleibende Faktor von 2^2 kann mit den verbesserten Algorithmen erklärt werden, mit denen inzwischen auch eine wesentlich höhere Auflösung erzielt wird. Dabei darf nicht vergessen werden, dass der Speicherbedarf für diese kosmologischen Simulationen um den gleichen Faktor angewachsen ist."[22] Dieses wird in Tabelle 2 illustriert.

Neben den digitalisierten Daten gibt es in der Astronomie viele in analoger Form vorliegende Daten. Insbesondere für die Fotoplatten, die seit 1893 eingesetzt wurden, werden erhebliche Anstrengungen unternommen, diese in digitaler Form zugänglich und nutzbar zu machen. Damit ist auch die Perspektive der Astronomie für die längerfristige Datenarchivierung umrissen: auf der Zeitskala der Veränderungen von astronomischen Objekten sind 100 Jahre wenig, so dass Daten von weit zurückliegenden Beobachtungen kostbare Informationen enthalten. Die mit der Digitalisierung der Daten zu lösenden Probleme der langfristigen Erhaltung dieser Daten sind jedoch nur in wenigen Fällen über konzeptuelle Überlegungen hinaus gekommen.[23]

22 Stefan Gottlöber in einem Vortrag, AIP-Seminar, Potsdam 2010.

23 Hier ist wiederum der SDSS im Verbund mit der JHU-Library ein Beispiel. (Vgl. Marcum; George (2009)).

Jahr	Simulationen	Datenmenge	CPUh
1990	Zusammenstoß zweier Galaxien 32^3 Partikel	~ 20 Megabyte	10 000
2009	Zusammenstoß von Galaxien 1024^3 Partikel	~100 Terabyte	1 Mio.
2009–	Sterndynamos	~50 Gigabyte	100 000
2009–	Black Hole Collisions	~5 Terabyte	1 Mio.
2009–2011	Großräumige Strukturen im Universum 2048^3–4096^3 Partikel	~1 Petabyte	20–50 Mio.

Tab. 2: Schätzung von Datenmengen für astrophysikalische Simulationen.

Oft stützen sich derzeitige Überlegungen dazu auf das Moore'sche Gesetz, das ca. alle zwei Jahre eine Verdoppelung der Speicherkapazitäten voraussagt. Damit wird jedoch nur das rein technische Problem der Datenmengen angesprochen. Neben den schon vorhandenen Standardisierungen der Metadaten und Zugriffsprotokolle, deren notwendige Weiterentwicklung und dem Füllen von aufgezeigten Lücken, wird die Frage nach dem Lifecycle-Management für die Forschungsdaten immer wichtiger, wie auch die Entwicklung von Formen in der Community, die Erhaltungswürdigkeit von Daten über längere Zeiträume zu managen. Bei den Simulationen ist diese Frage vergleichsweise einfach, da nach spätestens zehn Jahren die Neuberechnung der Daten, falls erforderlich, viel weniger kostet als die Erhaltung der Rohdaten selbst.[24] Für Beobachtungsdaten ist dieses Problem sehr viel komplexer, zumal die zeitliche Dimension eine wachsende Bedeutung in der heutigen beobachtenden Astronomie gewinnt.

24 Diese Aussage gilt natürlich nur unter der Voraussetzung, dass eine ausreichende Langzeitarchivierung von Metadaten, Software und Anfangsbedingungen vorhanden ist.

14.4 Interne Organisation

Forschungsdaten werden in der Astronomie produziert durch:

- Einzelbeobachtungen von Forschern und kleinen Arbeitsgruppen,
- Systematische Beobachtungskampagnen und Surveys von größeren Kollaborationen und Satellitenmissionen,
- Simulationen von kleineren Rechnungen für die Anfertigung einer Dissertation, bis zu großen kosmologischen Simulationen, die von internationalen Kollaborationen durchgeführt und ausgewertet werden.

Die Verantwortlichkeit für die Speicherung der Daten und deren Archivierung obliegt entsprechend nicht einer einzelnen Institution, und eine systematische Organisation von Verantwortlichkeiten gibt es nicht.

Auf der Ebene der Einzelbeobachtung obliegt die Archivierung in der Regel dem Forscher und dem Institut. Über die DFG-Richtlinien hinaus besteht keine Regelung zur Archivierung. Es gibt Bemühungen, insbesondere durch VO und andere eScience-Projekte, Datenarchive aufzubauen und die einzelnen Forscher gezielt bei der Einbringung ihrer Daten in diese Archive zu unterstützen. Diese Arbeiten sind häufig sehr zeitintensiv und werden umso komplizierter, je länger Daten in den individuellen Systematiken der jeweiligen Forschergruppe verwaltet wurden. In der Regel ist dies auch verbunden mit der Entwicklung von passender Software und findet v.a. dort statt, wo eScience im Institut vertreten ist. Eine signifikante Verbesserung der Situation wäre hier zu erreichen, wenn die Nutzung der eScience Bestandteil der Ausbildung der Wissenschaftler würde. Die Ebene der Daten von Einzelbeobachtungen ist am ehesten einer nationalen Anstrengung zugänglich, in der einheitliche Policies für die Archivierung von Forschungsdaten festgelegt werden. Hierzu ist aber auch die Unterhaltung von Ressourcen für diese Datenarchive notwendig.

Für die internationalen Kollaborationen, die Surveys durchführen oder Instrumente bauen, ist das Datenmanagement mittlerweile fester Bestandteil des Gesamtplanes. Es ist damit jedoch keineswegs gesichert, dass der Datenarchivierung und -publizierung auch der notwendige Anteil an personellen und finanziellen Ressourcen zugeordnet wird. Wie schon beschrieben, werden

die Verantwortlichkeiten innerhalb der Kollaborationen von einzelnen beteiligten Instituten, sowie auch – für Europa – durch Organisationen wie die ESO oder die ESA wahrgenommen. Eine solche Kollaboration besteht i.d.R. für mindestens ein Jahrzehnt und übernimmt in dieser Zeit den Aufbau und den Betrieb des Datenmanagements für „ihre" Daten.[25] Mittlerweile werden Survey-Daten nach einer gewissen Zeit (meist zwischen zwei und fünf Jahren) publiziert und der Community zur Verfügung gestellt. Die publizierten Daten haben oft auch eine VO-Compliance, d.h. sie nutzen die Standards und Metadaten des VO für den Web-basierten Zugriff.[26] Es zeichnet sich ab, dass in jedem Land einige Institute das Datenmanagement als einen spezialisierten Bereich gezielt fördern. Diese Entwicklung könnte auch eine Möglichkeit bieten, die Langzeitarchivierung von Forschungsdaten in der Astronomie anzugehen und einen „institutionellen Rahmen" dafür zu schaffen.

Die erforderliche IT-Infrastruktur hierfür ist jedoch erst in Ansätzen vorhanden. Zu dieser gehört neben den Datenservern auch die Bereitstellung einer effizienten Arbeitsumgebung, von Zugangsstrukturen und von Netzwerken. Darüber hinaus werden auch personelle Ressourcen benötigt, um das Datenmanagement und die zugehörige Software zu implementieren. Die durch D-Grid und AstroGrid-D zur Verfügung gestellte kollaborative IT-Infrastruktur ist noch nicht annähernd weit genug ausgebaut, um beispielsweise einen effizienten Zugang zu den Daten des LOFAR-Instruments zu gewähren. Hierfür muss in Zusammenarbeit mit den großen Rechenzentren und der LOFAR-Zentrale der Zugang der Forschergruppen zu Daten über das LOFAR Netzwerk entwickelt werden.

---

25 Die übliche Finanzierung eines solchen Projektes ist komplex, auch wegen der nationalen Begrenzungen der Fördergeber, und meist werden die Kosten für ein Instrument durch die Institute aufgebracht, die im Gegenzug garantierte Beobachtungszeiten am Instrument bekommen.

26 "The VO is in the early stages of deployment, thus much of the software that is currently being written is infrastructure software. However, 84% of the facilities surveyed by Panels A, B and C (see Appendix IV) indicated plans for a public data archive, and 53% of those are committed to publishing datasets and resources to the VO. Since these include the majority of the large data providers (e.g., ESA, ESO, LOFAR, etc.), this implies VO compliance of a much higher percentage of the actual data volume. VO projects are now ramping up support to data centres as they implement VO standards, and there is little doubt that by the end of the next decade most astronomical data will be VO compliant." (o.V. (2008), S. 80).

Bei den Simulationen sind Projekte wie die Millennium-Datenbank[27], CLUES[28] und MultiDark[29] dabei, die Archivierung und Publikation von Simulationsdaten zu erforschen. Es sind weitere intensive Entwicklungsarbeiten notwendig für die Erarbeitung von Standard-Metadaten und geeigneten Formaten für die Publikation der Rohdaten sowie für deren Verwaltung. Diese Aufgabe wird durch die Erfahrungen, die bei den Beobachtungsdaten gewonnen wurden, erleichtert. Zusätzlich schlägt die Astronet Roadmap auch die Gründung des Astrophyiscal Software Laoratoy (ASL)[30] als europäische Einrichtung zur gezielten Unterstützung der Entwicklung von Simulationscodes durch die Community vor. Die Erarbeitung von Standards und Metadaten für Simulationsdaten in Kombination mit der Verbesserung und Pflege der Simulationscodes können bei Realisierung der genannten Vorhaben in wenigen Jahren zu einer signifikanten Verbesserung des Forschungsdatenmanagements in diesem Bereich führen.

Die Arbeit von eScience-Projekten wie GAVO, dem VDZ (Virtuelles Datenzentrum), der Ausbau der Services für die Community, wie z.B. das GAVO DataCenter oder der SDSS Mirror, wie auch die verstärkte Einbeziehung der eScience in die Planung und Implementierung des Datenmanagements von Instrumenten und Surveys sowie Fortführung der Entwicklungen, die mit den Simulationsdaten-Projekten Millennium, MultiDark und CLUES begonnen wurden, bildet die Basis für die Bewältigung der sich derzeit stellenden Aufgaben in der Forschungsdatenarchivierung in der Astronomie.

14.5 Perspektiven und Visionen

Gegenwärtig werden in der Astronomie einige Großprojekte durchgeführt oder vorbereitet, die extrem große Datenmengen produzieren. Meist handelt es sich um sogenannte Himmelsdurchmusterungen („Sky Surveys“), bei denen die Sterne und Galaxien am Himmel mit sehr hoher Winkelauflösung,

27 Vgl. GAVO (2011).
28 Vgl. CLUES (2011).
29 Vgl. MultiDark Database (2011).
30 Vgl. o.V. (2008), S. 85.

oft in verschiedenen Farb-Filtern oder gar mit hoher spektraler Auflösung aufgenommen werden. Schließlich wird nun auch noch die „Time Domain" als weitere Dimension eröffnet: Diese Aufnahmen werden zu verschiedenen Zeiten („Epochen") wiederholt, um die zeitliche Veränderung der physikalischen Parameter (Ort, Geschwindigkeit, Helligkeit) präzise zu bestimmen. Das Archiv des bereits im Jahre 2008 abgeschlossenen Sloan Digital Sky Survey (SDSS) enthält beispielsweise ein Datenvolumen von mehr als 100 Terabyte. Das in Vorbereitung befindliche Large Synoptic Survey Telescope (LSST) soll ab dem Jahre 2020 mit einer 3-Gigapixel-Kamera jede Nacht astronomische Daten im Umfang von 30 Terabyte produzieren. Bei einer Laufzeit von zehn Jahren entspricht das 100.000 Terabyte = 100 Petabyte.

Diese Größenordnung von Beobachtungsdaten wie auch die sich parallel dazu entwickelnden Simulationsdatenmengen haben damit eine Stufe erreicht, die eine Erweiterung, wenn nicht sogar Ablösung der klassischen Arbeitsmethoden der Astronomie erfordert. Bei Objektkatalogen, die Milliarden Einträge aufweisen, treten neben das Studium von Einzelobjekten das Data-Mining und andere statistische Methoden der Auswertung. Visualisierungstechniken werden immer wichtiger und bedeutsamer werden. Es wird einen Paradigmenwechsel geben: Nicht mehr die Vermessung eines einzelnen Sterns oder einer einzelnen Galaxie wird wichtig sein, sondern statistische Untersuchungen und Auswertungen werden immer größere Bedeutung erhalten. Entsprechend müssen neue Konzepte für verlässliche Datenarchivierung und effizienten Datenzugang entwickelt werden.