Universität Rostock
Fachbereich Informatik
Institut für Computergrafik

Metadatengewinnung und -spezifikation für Visualisierungsentscheidungen

DIPLOMARBEIT

vorgelegt von

Thomas Nocke
geboren am 27.09.1975
in Dresden

Betreuer: Prof. Dr. Heidrun Schumann
Dipl.-Inf. Petra Schulze-Wollgast

Abgabedatum: 28. Februar 2000

Danksagung

Vor allem möchte ich mich bei meiner Betreuerin Prof. Dr. Heidrun Schumann bedanken, die mich in aufopferungsvoller Arbeit bei der Problemanalyse, beim Schreiben und bei der Korrektur meiner Arbeit unterstützt hat. Auch Petra Schulze-Wollgast gilt mein Dank für die Unterstützung bei Korrektur, Programmdesign und in Diskussionen. Weiterhin hat mir die Unterstützung von Friedrich Wagner und Jens Miehe bei Diskussionen und Problemen in Latex weitergeholfen. Auch Dr. Holger Theisel gilt mein Dank, der mich durch verschiedene praktische Anregungen unterstütze. Desweiteren hat mir der praktische Rat von Uwe Rauschenbach und Matthias Kreuseler bei der Implementation des Programmes viel Arbeitszeit erspart. Auch Herrn Prof. Dr. Dietmar Jackèl gilt mein Dank, der so unkompliziert die Bewertung meiner Arbeit übernahm. Weiterhin danke ich Susanne Lange, deren Untersuchungen zu Metadaten im Visualisierungsumfeld eine Basis geschaffen haben, auf der diese Arbeit aufbauen konnte.

Vor allem danke ich auch meiner geliebten Ivonne, die mich durch ihre Aufmunterung, durch viele Anregungen und durch die Korrektur der Arbeit unterstützt hat, diese Diplomarbeit zu vollenden.

Zusammenfassung

Die Analyse von Datenbeständen unterschiedlicher Art ist ein wichtiges Forschungsgebiet. Insbesondere durch die Visualisierung der Daten können diese intuitiv analysiert und Zusammenhänge erkannt werden. Die Entscheidung, welche Visualisierungstechnik für eine Datenmenge besonders geeignet ist, ist von den Eigenschaften der Datenmenge abhängig.

In der vorliegenden Arbeit wird eine allgemeingültige Spezifikation von sogenannten Metadaten entwickelt, um Visualisierungsentscheidungen anhand der Eigenschaften der Datenmenge zu unterstützen. Dazu werden für die Visualisierung wichtige Metdaten definiert. Desweiteren wird ein Konzept zur Gewinnung dieser Metadaten entworfen, in dem neben einer geeigneten Reihenfolge auch eine Einbeziehung des Nutzers integriert wird.

Im Anschluß an die konzeptionellen Vorarbeiten wurde ein interaktives Programm zur Integration des Nutzers in die Metadatenerfassung entwickelt. Dieses Programm unterstützt den Nutzer in Abhängigkeit von dessen Profil durch automatische Analyseverfahren und graphische Darstellungen bei der Bestimmung von Metadaten für unterschiedliche Datenmengen.

Abstract

The analysis of data of various kind is an important field of research. Particularly the visualization of data allows to analysize it intuitively and to find out cohoerences. The desicion, which technique of visualization is especially suited to which data set, depends on the properties of the data set.

This paper conceptualizes an universal spezification of so-called metadata, to support visualisation decisions due to the properties of the data set. Therefore important metadata for visualisation will be defined. Furthermore a concept for extraction of these metadata will be designed, which integrates a suitable order and the inclusion of the user.

In conjunction with these conceptual projects, an interactive programme, which integrates the user into the extraction of metadata, was developed. The intention of this programme is to support the user in dependency of his profile. Therefore automatic techniques of analysis and graphical representations are integrated to determine metadata of different types of data.

CR-Klassifikation

E, E.1, E.4, G.1, G.1.0, G.1.3, G.3, G.4, H.1.1, H.2.8, H.3.3, I.3, I.3.6, I.5.3

Key Words

Visualisation, Metadata, Visualisation decision, Data analysis, Data mining, Databases, Statistics, Classification, Regions of interest, Information theory

Inhaltsverzeichnis

Kapitel 1

Einleitung und Motivation

Eine wachsende Informationsmenge in den unterschiedlichsten Anwendungsbereichen bringt die Notwendigkeit mit sich, diese großen Datenmengen zu überblicken und zu verwalten. Dafür bedarf es der Entwicklung neuer Konzepte und Verfahren zur Analyse. Mit den gewonnenen Erkenntnissen kann ein Wissensbestand über die Daten und dadurch über deren reale Entsprechungen aufgebaut werden. Mit Hilfe dieses Wissensbestandes können wichtige praktische Entscheidungen unterstützt werden.

Die Hauptziele bei der Analyse von Datenmengen sind die Gewinnung eines Überblicks über den Datenbestand und das Aufdecken versteckter Strukturen. Ansätze zur Lösung dieser Probleme finden sich in der mathematischen Statistik, im Gebiet der Datenbanktechnologien und in vielen Anwendungswissenschaften und zunehmend auch in den Gebieten visuelles Data-Mining und wissenschaftliche Visualisierung. In letzteren werden die Daten analysiert und je nach Kontext in eine geeignete und aussagekräftige graphische Darstellung überführt.

Da visuelle Darstellungen anwenderfreundlich und intuitiv sind, sind die Ansätze aus dem Gebiet der Visualisierung von besonderem Interesse, um die obengenannten Probleme zu lösen. Das Problem, aber auch die Chance der Visualisierung ist die große Menge an vorhandenen Visualisierungstechniken und deren Parameter. Aufgrund der zahlreichen Auswahlmöglichkeiten ist es für den Nutzer schwer, geeignete Visualisierungsentscheidungen zu treffen.

Bisherige Ansätze haben bei automatisch unterstützten Visualisierungsentscheidungen überwiegend spezielle Einflußfaktoren oder Darstellungstechniken betrachtet (vgl. z.B. [GLdCS97], [Mül98] und [Jun98]). Einer der Haupteinflußfaktoren auf die Wahl von Visualisierungstechniken sind die Daten und deren Eigenschaften. Dieser Aspekt wurde in den genannten Ansätzen nicht in seiner Gesamtheit integriert. Das Problem dabei ist, die Balance zwischen einer möglichst vollständigen Beschreibung der Dateneigenschaften und deren praktischer Handhabbarkeit zu halten.

Für diese Arbeit sollen Technologien verschiedener Bereiche integriert werden, um ein Konzept zu entwickeln und umzusetzen, mit dem anhand der Datencharakteristik die Entscheidung unterstützt wird, welche Visualisierungsmethoden und -techniken angewandt werden sollten. Dabei geht es vor allem darum, eine möglichst

umfassende Beschreibung in Form von sogenannten Metadaten zu entwickeln.

Die Definition und Erfassung von Metadaten ist ein aktuelles Forschungsgebiet. Beispiel hierfür ist die Entwicklung von Methoden zur Generierung von Metadaten über Dokumente, um vor allem im Internet eine Strukturierung nach Inhalten durchführen zu können. Dabei werden für jedes Dokument Informationen über Autor, Inhalt u.a. in sogenannten Metadaten gespeichert. Ansatz dieser Arbeit ist es, ein Konzept von Metadaten für die Visualisierung von Datenbeständen zu entwickeln.

Die hier vorgelegte Arbeit geht über bestehende Ansätze hinaus. Bisher wurden in der Visualisierung hauptsächlich so bezeichnete *beschreibende* Metadaten eingesetzt. Das sind Metadaten, die den Zugriff auf die Datenmenge und die Grundstruktur der Daten festlegen. Diese sollen aufgegriffen, zusätzlich aber um so bezeichnete *abgeleitete* Metadaten erweitert werden. Diese stellen aus Berechnungen erhaltene zusätzliche Informationen zu Struktur und Eigenschaften der Daten dar. Ziel ist es außerdem, bei der Erfassung und Berechnung von Metadaten einerseits automatisch und interaktiv bestimmbare Metadaten zu trennen und sie andererseits in einem interaktiven Programm zur Bestimmung von Metadaten in geeigneter Weise zu gewinnen.

Zuerst werden in Kapitel 2 wichtige Begriffe festgelegt und die Problemstellung genauer dargestellt. Kapitel 3 beschreibt ein Metadatenkonzept für die wissenschaftlich technische Visualisierung. Dabei werden allgemeine Eigenschaften der Metadaten festgelegt (Abschnitt 3.1) und relevante Metadaten spezifiziert (Abschnitt 3.2). In Kapitel 4 werden dann Methoden für die praktische Erfassung der Metadaten diskutiert. In Kapitel 5 wird die Umsetzung des Metadaten-Konzeptes im Werkzeug „Metadatum“ vorgestellt. Kapitel 6 präsentiert die Analyseergebnisse dreier praktischer Datenmengen und bewertet die Analysen. Kapitel 7 beschreibt Anregungen und Ideen für weitere Untersuchungen, und in Kapitel 8 werden die Ergebnisse der Arbeit noch einmal zusammengefaßt.

Kapitel 2

Begriffe und Problemstellung

Bevor die Konzepte und deren Umsetzung vorgestellt werden, ist es wichtig, einige grundlegende Begriffe zu klären und die Problemstellung zu umreißen.

2.1 Begriffe

In diesem Abschnitt sollen die *allgemein* für die Arbeit wichtigen Begriffe vorgestellt werden. *Spezielle* Begriffe bei der Metadaten-Definition werden in Kapitel 3 bei der Vorstellung der einzelnen Metadaten direkt erklärt. Bei den verwendeten Begriffen im Visualisierungsumfeld erfolgt eine Orientierung an den Begriffsdefinitionen aus [SM00].

Visualisierungspipeline Die Metapher Visualisierungspipeline faßt die Hauptschritte bei der Erzeugung einer visuellen Darstellung zusammen. Dies sind die Schritte Vorverarbeitung, Mapping und Rendering.

Visualisierungsentscheidung Das Treffen einer Visualisierungsentscheidung beinhaltet die Auswahl einer geeigneten Visualisierungstechnik und deren Parameter und beeinflußt damit in entscheidendem Maße den Mappingschritt. Die Eignung einer Technik hängt von verschiedenen Faktoren ab (vgl. Begriff Visualisierungskontext). Beispielsweise sollten die Daten eines Strömungsfeldes mit einer adäquaten Strömungsvisualisierungstechnik dargestellt werden.

Visualierungskontext Der Begriff Visualisierungskontext umfaßt die Gesamtheit aller Einflußfaktoren für das Treffen von Visualisierungsentscheidungen. Metadaten sind ein spezieller Kontext, der für die Wahl und die Parametrisierung von Visualisierungstechniken von Bedeutung ist. Weitere Kontexte sind z.B. das Nutzerprofil, die zu Verfügung stehende Hardware u.a.

Visualisierungsparameter Neben der Visualisierungsentscheidung ist weiterhin die Wahl der zugehörigen Parameter für eine aussagekräftige Darstellung wichtig. Hierunter fallen die Abbildung auf die Koordinatenachsen, die Festlegung von Farben, Formen u.a.

Datenmenge Der Begriff Datenmenge umfaßt die Gesamtheit aller in die Metadatengewinnung einfließenden Daten.

Beobachtungsraum In [SM00] wird der „Raum, in dem die Daten erhoben werden, als Beobachtungsraum“ ([SM00] S. 29) bezeichnet. Dabei wird bewußt davon abstrahiert, ob es sich um einen konkreten physikalischen oder einen abstrakten Beobachtungsraum handelt. Die Dimensionen des Beobachtungsraumes werden auch als unabhängige Variable bezeichnet. Diese können je nach Art des Raumes Ortskoordinaten, Zeitachsen und/oder abstrakte Dimensionen sein.

Beobachtungspunkt Ein Beobachtungspunkt ist ein Punkt des Beobachtungsraumes, für den Daten vorliegen.

Datensatz Ein Datensatz ist die Menge aller Daten eines Beobachtungspunktes.

Merkmal Merkmale sind Größen, die im Beobachtungsraum erhoben werden. Sie werden auch als abhängige Variable bezeichnet.

Metadaten sind „Daten über Daten“. Das bedeutet, daß sie Daten näher beschreiben bzw. Zusatzinformationen über sie darstellen. Metadaten lassen sich nach [RH97] klassifizieren in:

- beschreibende,
- abgeleitete und
- historische Metadaten.

Beschreibende Metadaten bezeichnen dabei grundlegende Attribute der Daten und legen fest, wie auf die Metadaten zugegriffen werden kann. Abgeleitete Metadaten werden durch Analyse der Daten gewonnen. Dies können z.B. statistische Daten sein. Historische Metadaten geben Aufschluß über die Entstehung der Daten. Beispiel für historische Metadaten ist die Fehlerbehaftung von Daten. Unter diesem Gesichtspunkt ist von Interesse, ob und welche Fehler bei der Messung der Daten aufgetreten sind.

Datenklassen Der Begriff der Datenklasse stellt ein Schema für Daten dar. Unterschiedliche Datenklassen haben spezifische Eigenschaften, welche in der Visualisierung spezifisch dargestellt werden sollten. Anhand der Ausprägungen der Metadaten können bestimmte Datenmengen auf bestimmte Datenklassen abgebildet werden. In der Visualisierung werden üblicherweise die Datenklassen Multiparameterdaten, Volumendaten, Strömungsdaten und Scattered-Data unterschieden.

2.2 Problemstellung

Aufbauend auf den vorangegangenen Begriffsdefinitionen kann nun die eigentliche Problemstellung erläutert und die Metadatengewinnung in die Visualisierung eingeordnet werden.

Hauptziel einer Metadatengewinnung ist die Unterstützung von Entscheidungen innerhalb der Visualisierungspipeline. Dabei geht es darum, anhand der speziellen Dateneigenschaften die Aussagekraft der visuellen Darstellung zu erhöhen.

Abbildung 2.1 zeigt die Metadatengewinnung im Visualisierungskontext.

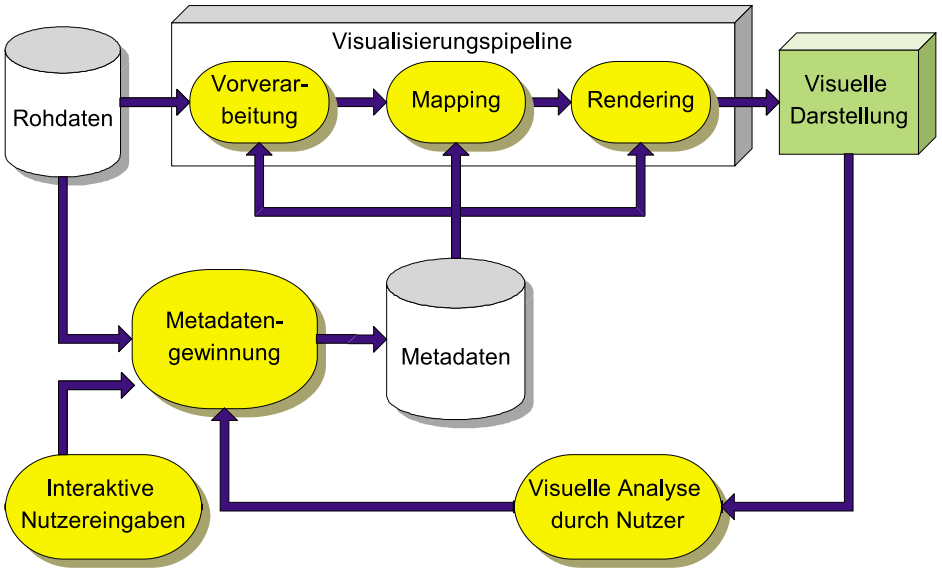

Abbildung 2.1: Metadatengewinnung im Visualisierungskontext

Aus den Rohdaten werden mit Hilfe des Nutzers und unter Verwendung automatischer Analyseverfahren Metadaten generiert. Diese können in den drei Hauptschritten der Visualisierungspipeline zur Unterstützung von Entscheidungen und zur Parametrisierung eingesetzt werden. In der Vorverarbeitung können bsw. Datenverbesserungsalgorithmen unter Nutzung der Metadaten die Ausgangsdaten in Abhängigkeit ihrer Spezifika aufbereiten. Beispiel für die Unterstützung des Mapping ist die Auswahl von geeigneten Visualisierungstechniken. Beim Rendering können z.B. regions of interest in höherer Präzision in die visuelle Darstellung überführt werden.

Ziel der Einbeziehung von Metadaten in die drei Schritte der Visualisierungspipeline ist in jedem Fall die Erhöhung der Effektivität einer visuellen Analyse. Dieser Prozeß kann iterativ ablaufen. Durch ein erstes Bild können bestimmte Eigenschaften einer Datenmenge durch den Nutzer aufgedeckt werden, welche dann zu einer verbesserten Erfassung der Metadaten beitragen.

Kapitel 3

Metadaten für Visualisierungsentscheidungen

In diesem Kapitel werden die der Metadatengewinnung zugrunde liegenden Konzepte erläutert. Beim Erstellen der Konzepte spielen neben der Relevanz von Metadaten für die Visualisierung und der Allgemeingültigkeit der erhobenen Daten für unterschiedliche Datentypen auch die Anwenderorientiertheit und die praktische Effizienz eine wichtige Rolle. Das Ziel dabei ist, durch Eingrenzung auf wichtige Datenklassen und Verfahren zur Metadatengewinnung ein praktisch umsetzbares Konzept zu entwickeln.

In Abschnitt 3.1 werden die Grundkonzepte zur Definition von Metadaten entwickelt. Daran schließt sich in Abschnitt 3.2 die Spezifikation der Metadaten an.

3.1 Allgemeine Konzepte

Bei der Erstellung eines Metadatenkonzeptes und der Sammlung von Metadaten sind die Sichten wichtig, aus denen heraus die Metadaten erhoben werden:

1. **Aus Sicht der Daten:**
   - Welche Information beinhalten die Daten?
   - Welche Speicherungsstruktur haben sie und was kann aus dieser Struktur abgeleitet werden?
   - Welche Strukturen sind in der Datenmenge enthalten?
   - Welche Qualität haben die Daten und wie aussagekräftig können Analysen sein?
2. **Aus Sicht der Visualisierung**
   - Welche Metadaten unterstützen welche Vorverarbeitungen?
   - Welche Metadaten unterstützen welche Visualisierungsentscheidungen?

- Welche Metadaten unterstützen welche Parametrisierungen von Visualisierungstechniken?

3. **Aus Sicht der Verfahren zur Metadatengewinnung**
   - Welche Verfahren gibt es, um Daten zu analysieren und Metadaten zu erheben?
   - Welche Verfahren liefern visualisierungsrelevante Ausgaben?
   - Welche Verfahren benötigen ihrerseits Metadaten als Eingabe, um sinnvolle Ergebnisse liefern zu können?
   - Wie ist die Qualität der Verfahrens-Ergebnisse einzuschätzen?
   - Wie können die Verfahren effektiviert werden?

3.1.1 Die Datensicht

Aus Sicht der Daten geht es vor allem darum, für unterschiedliche Datenklassen die Metadaten zu vereinheitlichen. So sollen allgemeine Eigenschaften unabhängig von der speziellen Datenklasse gemeinsame Informationen zusammenfassen. Ziel dabei ist, eine möglichst effektive Speicherung und Verwaltung der Metadaten durchführen zu können.

Es gibt verschiedene Ansätze zur Datenspezifikation, wie sie z.B. in [B+92], [BG89] und [WB97] durchgeführt wird. Diese Ansätze enthalten aber nur Teilaspekte und abstrahieren teilweise von wichtigen Eigenschaften.

Ein allgemeiner Ansatz für die Datenspezifikation, wie er in [The94] vorgestellt wird, ist für die praktische Speicherung von Daten- und verschiedensten Metadatenarten zu ineffizient. Dort werden die Daten und deren räumlicher Bezug in einer Matrix gespeichert, was zum Auftreten von großen Redundanzen führen kann.

Dieser Arbeit soll die Datencharakteristik aus [GLdCS97] zugrunde gelegt werden, da diese viele wichtige Eigenschaften der Datenmenge integriert. Dort wird die Struktur der Datenmenge durch den formalen Ausdruck

$$A\, c\, {}^{(t_1,o_1,u_1,d_1),\ldots,(t_k,o_k,u_k,d_k)}_{n,g,w,t} \quad \text{mit} \tag{3.1}$$

A	Anzahl der Beobachtungsfälle
c	Qualität der Datenmenge
t_i	Typ der Meßdaten
o_i	Existenz einer Ordnungsrelation über dem Wertebereich der Meßdaten (ja/nein)
u_i	Umfang des Wertebereichs
d_i	Anzahl der Merkmale mit Charakteristik(t_i, o_i, u_i)
n	Dimensionalität des Raumes
g	Art des Zusammenhangs der Meßpunkte (z.B. 3D-reguläres Gitter)
w	Wirkungsbereich der Meßpunkte (Punkt, lokal, global)

und $t = (t_a, t_{dis}, t_e)$ mit

t_a	Anfangszeitpunkt
t_{dis}	Größe der Intervalle (gleich/ungleich)
t_e	Endzeitpunkt

beschrieben. Interessant an diesem Ansatz ist sowohl die explizite Beachtung des Zeitbezuges als auch die Beachtung spezieller Eigenschaften des Beobachtungsraumes. Die speziellen Teilaspekte dieser Spezifikation werden in Abschnitt 3.2 in das Metadatenkonzept integriert.

Die betrachtete Metadatenspezifikation ist formal gegeben. Wichtige Frage für ein praktisches System ist, wie mit ihr gearbeitet werden kann und wie sowohl die Daten als auch die Metadaten gespeichert werden.

Eine Möglichkeit zur Lösung dieses Problems ist das NetCDF-Format (vgl. [RDE93]). Dort werden sowohl die Daten als auch deren Metadaten gemeinsam abgelegt. Das NetCDF-Format wird im Visualisierungsumfeld häufig eingesetzt, jedoch kann man davon ausgehen, daß die Daten im allgemeinen nicht in diesem Format erhoben werden.

Ziel dieser Arbeit ist, einen praktikablen Mittelweg zwischen einer zu allgemeingültigen und einer zu speziellen Speicherung zu finden. Der Focus liegt weder darauf, ein allgemeingültiges Datenformat zu schaffen, noch darauf, beliebige Eingabeformate einlesen zu können. Ziel ist vielmehr eine plausible und praktikable Arbeitsgrundlage zu schaffen, in welcher die verschieden Formen von Rohdaten und Metadaten integrierbar sind. Für das Format zur Speicherung der Metadaten soll gelten, daß es aus dem aufgestellten Metadatenkonzept in geeigneter und nutzbarer Weise abgeleitet wird.

Weiter wird davon ausgegangen, daß die Rohdaten in Form von Tabellen vorliegen. Dies hat vor allem den Grund, daß in praktischen Anwendungen um die 90% der Daten in relationalen Datenbanken und damit in Tabellenform gespeichert werden. Das Konzept soll jedoch so ausgearbeitet werden, das beliebige Rohdatentypen wie implizite Raumdaten eingelesen und verarbeitet werden können. Darauf wird an späterer Stelle noch genauer eingegangen.

Für die Konzeption der Metadaten liegt der Schwerpunkt auf beschreibenden und abgeleiteten Metadaten, da historische Metadaten meist stark von der speziellen Anwendung abhängen.

Hauptkonzept für die Spezifikation von Metadaten ist die *Beschreibung der internen Strukturierung* der Daten in Abhängigkeit von deren Eigenschaften. Dabei geht es vor allem darum, die Daten zu klassifizieren und interessante Teilmengen zu extrahieren.

Ein wichtiges Unterkonzept der Strukturbeschreibung ist die Ausweisung von *Bereichen von besonderem Interesse* (regions of interest) im Beobachtungsraum, z.B. Bereiche mit hohem Informationsgehalt. Um Visualisierungen aussagekräftiger

zu machen, sollten solche Regionen mit besonderer Qualität und Auflösung dargestellt werden. Ziel der Metadatenspezifikation ist deswegen, aufgrund von speziellen Dateneigenschaften und mit Hilfe des Vorwissens des Nutzers solche Bereiche auszuweisen.

Wichtiges Konzept für Metadaten ist die Beachtung ihrer *Qualität*. Die Qualität bezeichnet, wie genau und wie sicher eine Aussage, die sich auf das entsprechende Metadatum bezieht, sein kann. Ungenauigkeiten können bsw. bei der Messung der Daten auftreten. Diese müssen entsprechend berücksichtigt werden.

3.1.2 Die Visualisierungssicht

Grundsätzlich sollen Metadaten zur Unterstützung von jedem der drei Schritte der Visualisierungspipeline erhoben werden. Hauptaugenmerk liegt dabei auf der Spezifikation von Metadaten für das Mapping.

Metadaten zur Unterstützung von *Vorverarbeitungsentscheidungen* sind z.B. Metadaten zur Bestimmung von regions of interest, deren Verteilung zum einen Visualisierungsentscheidungen, zum anderen aber auch vorverarbeitende Reduktionen der Datenmenge unterstützt.

Hauptentscheidung für die Spezifikation von Metadaten zur Unterstützung des *Mappingschritts* ist, daß sie vor allem die Auswahl von Visualisierungstechniken unterstützen sollen. Beispiel hierfür ist der bereits erwähnte Fall, daß für die Daten eines Strömungsfeldes eine adäquate Strömungsvisualisierungstechnik gewählt werden sollte. Darüber hinaus können jedoch auch Metadaten erhoben werden, die bei ausgewählter Visualisierungstechnik die Abbildung der Daten auf die Elemente der visuellen Repräsentation beeinflussen. Beispielsweise sollten ordinale Daten auf ordinale Bildmerkmale wie z.B. Helligkeit abgebildet werden.

Den *Renderingschritt* können z.B. Metadaten von Bereichen von Interesse unterstützen. Denkbar ist hier bsw., in Abhängigkeit des Interesses von Teilmengen die Viewing-Parameter wie Kamerapositionen oder Beleuchtungen zu variieren.

3.1.3 Verfahrenssicht

Es existiert eine große Menge von Datenanalyseverfahren aus Mathematik, Informatik u.a. Ziel dieses Konzeptes kann und soll es nicht sein, alle möglichen Ergebnisdaten aus diesen Verfahren zu integrieren. Die Fragestellung lautet vielmehr, welche Ergebnisse von Analysen zum Treffen von Visualisierungsentscheidungen von Bedeutung sind.

Zusätzlich zu diesem Kriterium ist es sinnvoll, auch Metadaten in das Konzept aufzunehmen, die andere Metadatenerhebungen erst möglich machen[1]. Beispiel hierfür sind z.B. Skalentypen von Variablen, welche bei späteren Erhebungen die speziellen Eigenschaften der Daten nutzbar machen[2].

[1] vor allem beschreibende Metadaten

[2] Anm.: Natürlich können Skalentypen auch direkt zu Visualisierungsentscheidungen führen.

Auch der oben beschriebene Aspekt der Qualität von Metadaten speziell in Abhängigkeit vom gewählten Verfahren ist für die Bewertung dieser Verfahren und der gewonnenen Metadaten von Bedeutung. Dabei ist wichtig festzustellen, ob durch die Metadatengewinnung Unsicherheiten enstehen oder bestehende Unsicherheiten verstärkt werden.

Ein weiterer Bestandteil des Metadatenkonzeptes soll die Trennung von Metadaten, die für die gesamte Datenmenge gelten und von Metadaten, welche nur für gewisse Teilmengen der Daten gelten. Diese Unterscheidung zu machen ist wichtig, weil die Reduktion einer Datenmenge zu anderen Eigenschaften dieser Datenmenge und damit zu anderen Metadatenausprägungen führen kann. Dieses Konzept dient unter dem Blickpunkt der praktischen Effizienz vor allem der Zeitkontrolle der Metadatengewinnungsverfahren, kann jedoch auch zur Beschleunigung von Visualisierungsschritten verwendet werden. Die Spezifikation spezieller Metadaten für Teilmengen ist vor allem unter dem Gesichtspunkt sinnvoll, daß in praktischen Datenmengen häufig hochkorrelierte Informationen vorliegen. Führt man dann eine Reduktion dieser Korrelationen durch, können möglicherweise bei nicht vollständiger Korreliertheit auch gewisse Teilinformationen verloren gehen. Die Abschätzung von dadurch auftretenden Fehlern ist für die Bewertung der gewonnenen Metadaten wichtig.

3.2 Relevante Metadaten

In diesem Abschnitt werden die speziellen Metadaten spezifiziert. Wichtig dabei ist zum ersten, Metadaten zu den Variablen (Variablen-Metadaten) zu entwickeln. Dazu gehören Metadaten, die allgemeine Eigenschaften von Variablen enthalten und Metadaten, welche die speziellen Eigenschaften der abhängigen Variablen (Merkmals-Metadaten) beinhalten. Weiterhin werden Metadaten speziell zu den unabhängigen Variablen spezifiziert. Da die unabhängigen Variablen die Eigenschaften des Beobachtungsraumes festlegen, werden diese im folgenden auch als Beobachtungsraum-Metadaten bezeichnet.

Zum zweiten werden Metadaten definiert, welche die Eigenschaften der gesamten Datenmenge beschreiben. Diese sollen im folgenden als Datenmenge-Metadaten bezeichnet werden. Zu ihnen gehören die allgemeinen Eigenschaften der Datenmenge, die sich vor allem auf die Beziehungen der Variablen untereinander beziehen (allgemeine Datenmenge-Metadaten). Weiterhin gehören zu ihnen die Eigenschaften der sogenannten Beobachtungsfälle (Beobachtungsfall-Metadaten).

Weil die Einteilung in unterschiedliche Datenklassen von großer Bedeutung für die Visualisierung sind, werden zum dritten die sogenannten Datenklassen-Metadaten definiert. Diese beinhalten spezifische Eigenschaften von Daten bestimmter Datenklassen.

Ziel bei der Definition dieser speziellen Metadaten ist, möglichst viele Eigenschaften, die häufig zu speziellen Datenklassen zugeordnet werden, zu verallgemeinern,

um diese allgemeiner nutzbar zu machen. Weiteres Anliegen bei der Definition der Metadaten ist es, deren Relevanz für die Visualisierung nachzuweisen.

3.2.1 Variablen-Metadaten

Die Variablen legen die grundlegende Struktur der Daten fest. Eigenschaften wie die Abhängigkeit der Variablen sind für die Visualisierung von großer Bedeutung. Auf den Metadaten über die Variablen liegt deswegen das Hauptaugenmerk.

3.2.1.1 Allgemeine Variablen-Metadaten

Die allgemeinen Variablen-Metadaten enthalten die wichtigsten Eigenschaften der einzelnen Variablen der Datenmenge. Im folgenden werden zuerst wichtige beschreibende und abgeleitete Metadaten vorgestellt und diese im Anschluß im einzelnen erläutert bzw. definiert. Anschließend wird die praktische Relevanz dieser Metadaten anhand von Beispielen untermauert.

Zu den beschreibenden Variablen-Metadaten zählen

1. der Variablenname,
2. der Skalentyp,
3. die Abhängigkeit sowie
4. die semantischen Informationen.

Weiterhin ist eine Vielzahl von abgeleiteten Metadaten definierbar. Da diese häufig von der speziellen Anwendung abhängen, werden hier beispielhaft zwei allgemeine abgeleitete Metadaten spezifiziert. Dazu zählen

1. der Informationsgehalt und
2. die Qualität der Variablen.

Sind in Abhängigkeit von einer speziellen Anwendung weitere Metadaten erforderlich, können diese hinzugefügt werden.

Das Metadatum **Skalentyp** basiert auf der Definition aus [SM00]. Der Skalentyp einer Variable kann entweder *qualitativ* oder *quantitativ* sein (vgl. Abbildung 3.1). Bei den quantitativen ist im Gegensatz zu den qualitativen Variablen eine Metrik über dem Wertebereich definiert, durch die jedes Paar von Ausprägungen direkt vergleichbar wird. Die qualitativen Variablen untergliedern sich in *nominale*[3] und

[3]Die Werte sind ungeordnet.

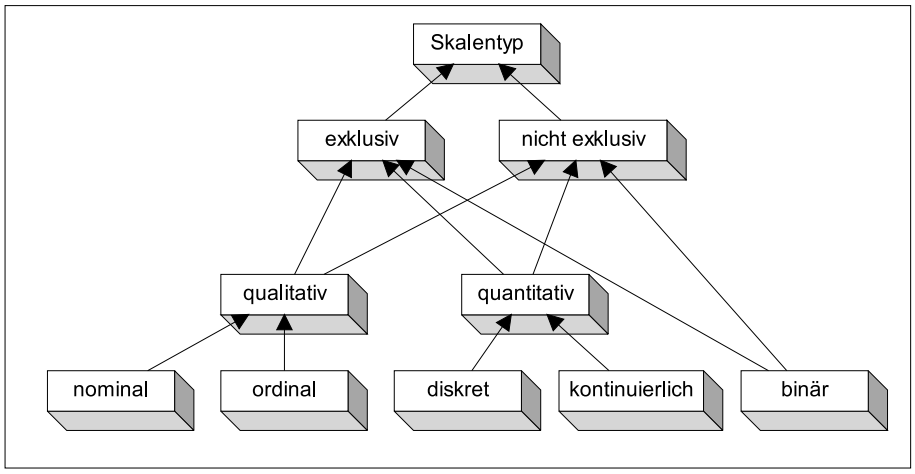

Abbildung 3.1: Skalentypen

ordinale[4] Variable. Die quantitativen Variablen teilen sich in *diskrete*[5] und *kontinuierliche*[6] auf. Zusätzlich dazu wurde ein *binärer* Skalentyp eingeführt[7]. Binäre Variable können zwar einem der anderen Skalentypen zugeordnet werden, jedoch kann dann nicht deren spezielle Eigenschaft der *Symmetrie* bzw. *Nichtsymmetrie*[8] genutzt werden. Für binäre symmetrische und binäre nicht-symmetrische Variable existieren spezielle Vergleichsmaße, welche deren spezielle Eigenschaften beachten (vgl. [Boc74] S. 48ff). Weiterhin umfaßt der Skalentyp die *Exklusivität* bzw. die *Nichtexklusivität* von Variablen. Nicht-exklusive Variable können im Gegensatz zu exklusiven Variablen mehrere Ausprägungen pro Datensatz haben. Ein praktisches Beispiel hierfür ist das gleichzeitige Auftreten mehrerer Farben bei einem Edelstein. Die Nicht-Exklusivität zu beachten schien vor allem deswegen wichtig, weil in praktischen Daten und vor allem in Tabellen häufig Mehrfachausprägungen auftreten.

Eine Variableneigenschaft, die eng mit dem Beobachtungsraum gekoppelt ist, ist die **Abhängigkeit** der Variablen. Die Abhängigkeit legt fest, ob es sich um eine *abhängige* oder *unabhängige* Variable – d.h. um ein Merkmal oder eine Dimension des Beobachtungsraumes – handelt. Im Falle einer Dimension kann diese zusätzlich entweder *abstrakt* oder *konkret* sein (siehe Beobachtungsraum-Metadaten in Abschnitt 3.2.1.3).

[4] Auf dem Wertebereich ist eine Ordnung definiert.

[5] Der Wertebereich entspricht den ganzen Zahlen.

[6] Der Wertebereich entspricht den reellen Zahlen.

[7] Binäre Variable haben genau zwei Ausprägungen. Eine Variable sollte genau dann als binär klassifiziert werden, wenn die beiden Alternativen sich im Sinne von „Datensatz hat Eigenschaft“ oder „Datensatz hat Eigenschaft nicht“ gegenüberstehen.

[8] Ein binäres Merkmal ist symmetrisch genau dann, wenn die beiden Alternativen genau gleich gewichtet sind. Hat jedoch die Aussage, daß wenn beide Datensätze die Eigenschaft besitzen eine größere bzw. kleinere Bedeutung als die Aussage, daß sie sie nicht besitzen, so wird das Merkmal als unsymmetrisch bezeichnet.

Unter den Metadaten über **semantische Informationen** werden weitere zur Spezifikation der Datenwerte relevante Eigenschaften gesammelt. So wird hier spezifiziert, ob die Datenwerte einer Variable als *numerische Werte* und/oder als *Zeichenketten* in die Visualisierung integriert werden sollten. Handelt es sich um Zeichenketten, so sollten diese bsw. in der Visualisierung als Beschriftung auftreten. Sind es Werte, sollten diese identifizierbar sein. Desweiteren zählt zu semantischen Informationen von Variablen, ob es sich um eine *Referenz* in eine andere Datenmenge handelt, welche die Datensätze dieser Datenmenge mit einer anderen verbindet.

Im folgenden werden die spezifizierten abgeleiteten Metadaten definiert und erläutert.

Der **Informationsgehalt** einer Variable sagt im allgemeinen aus, wie interessant bzw. wie uninteressant die Variable in einem bestimmten Kontext ist. Speziell wird der Informationsgehalt z.B. nach Shannon bestimmt. Er ist um so größer, je geringer die Wahrscheinlichkeit für das Auftreten einer der speziellen Werte-Verteilung ist.

Die **Qualität** beinhaltet, wie sicher Aussagen über die Variable und deren einzelne Ausprägungen sind. Hierunter fallen u.a. die *Anzahl fehlender Werte* und die *mittlere Fehlerbehaftung* der einzelnen Werte.

Die Relevanz dieser Metadaten soll im folgenden an einigen Beispielen nachgewiesen werden:

1. Relevanz für die Vorverarbeitung

   - Beispiel für die Unterstützung von Vorverarbeitungen ist die Berechnung von Ähnlichkeiten von Datenwerten und Datensätzen in Abhängigkeit von ihrem Skalentyp.

2. Relevanz für Visualisierungsentscheidungen

   - Der Skalentyp ist bei vielen Visualisierungsentscheidungen von Bedeutung. So sollten bsw. Daten bestehend aus quantitativen Variablen auf Liniendiagramme abgetragen werden, während auf qualitativen Variablen beruhende Daten besser in Balkendiagrammen verdeutlicht werden.
   - Die Qualität spielt bei der Wahl von Visualisierungstechniken ebenfalls eine Rolle. Einerseits ist es bei stark unterschiedlich verteilten Fehlern wichtig, eine Visualisierungstechnik mit Fehlerdarstellung auszuwählen. Andererseits sollte bei einer hohen Fehlerzahl keine Visualisierungstechnik gewählt werden, deren Interpretierbarkeit beim Auftreten von Fehlern oder fehlenden Werten stark abnimmt. Beispiel hierfür sind Streckenzugdarstellungen. Bei dieser Darstellungsart werden die Beziehungen von Werten benachbarter Variablen visualisiert. Dabei ist es problematisch, wenn eine zu hohe Anzahl von fehlenden Werten auftritt. Dies kann die Interpretation der Beziehungen stark erschweren.
   - Die Abhängigkeit bzw. Unabhängigkeit von Variablen und deren jeweilige Art und Anzahl ist für vielfältige Entscheidungen ausschlaggebend. So

sollten z.B. unabhängige Variable auf die Abszisse und abhängige Variable auf die Ordinate eines Diagramms abgebildet werden, um den Abhängigkeitszusammenhang zu verdeutlichen.

3. Relevanz für Verfahren zur Metadatengewinnung

   - Der Skalentyp legt fest, wie Ausprägungsdifferenzen bzw. -ähnlichkeiten in weiteren Metadatenerhebungen bestimmt werden können.
   - Die Qualität, und im besonderen die Anzahl fehlender Werte, wird ebenfalls bei der Korrelationsberechnung eingesetzt[9]. Weiterhin vermindert sich der Informationsgehalt bei geringer werdender Qualität.
   - Der Informationsgehalt von Variablen kann zur Entscheidung führen, die Variable als redundant zu kennzeichnen und sie zur Beschleunigung des Rechenprozesses bei zeitaufwendigen Berechnungen auszuschließen (vgl. Konzept redundanter Metadaten aus Abschnitt 3.1).

3.2.1.2 Merkmals-Metadaten

Für Variablen wurden die allgemeinen Variablen-Metadaten definiert. Zusätzlich zu diesen haben Merkmale spezielle Eigenschaften, die durch entsprechende Metadaten repräsentiert werden. Im folgenden werden wie oben Metadaten zunächst vorgestellt und dann ihre praktische Relevanz anhand von Beispielen nachgewiesen. Die Metadaten für Merkmale sind

1. der Datentyp
   als beschreibendes Metadatum und

2. die Histogrammeigenschaften,
   welche ein Beispiel für abgeleitete Metadaten speziell für Merkmale darstellen.

Ein wichtiges beschreibendes Metadatum ist der **Datentyp** eines Merkmals (vgl. [SM00] S. 35-36). Übliche Datentypen im Visualisierungsumfeld sind *Skalar*, *Vektor* und *Tensor n-ter Ordnung* ($n \geq 2$).

Ein abgeleitetes Metadatum zur Beschreibung von Merkmalseigenschaften sind die **Histogrammeigenschaften**. Diese beinhalten Eigenschaften der Wertebereiche der Merkmale wie Kardinalität, Minimum, Maximum, Varianz, Mittelwert, Verteilungsschätzung u.a.

Im folgenden wird die Relevanz dieser Metadaten anhand von Beispielen nachgewiesen:

1. Relevanz für Visualisierungsentscheidungen

   - Analog zum Skalentyp ist auch die Wahl von speziellen Ähnlichkeitsmaßen in Abhängigkeit vom Datentyp notwendig.

[9] Nur die Korrelation von tatsächlich vorliegenden Werten ist interessant.

2. Relevanz für Visualisierungsentscheidungen

   - Aus dem Datentyp von Variablen kann bsw. geschlossen werden, ob eine spezielle Darstellung von Vektoren oder Tensoren durchgeführt werden muß.
   - Aus den Histogrammeigenschaften können Entscheidungen zur Parametrisierung von Visualisierungstechniken abgeleitet werden. Beispiel hierfür ist, anhand von Werteverteilungen die Farbwerte in der Darstellung zu spreizen. So können Bereiche, in denen viele Werte vorliegen, mit wesentlich mehr Farbnuancen dargestellt werden, um deren Unterscheidbarkeit zu erhöhen.

3. Relevanz für Verfahren zur Metadatengewinnung

   - Histogrammeigenschaften wie Varianz und Mittelwert sind Voraussetzung für die Korrelationsberechnung (vgl. 3.2.2) von Merkmalen.

3.2.1.3 Beobachtungsraum-Metadaten

Beobachtungsraum-Metadaten fassen die speziellen Eigenschaften der Dimensionen zusammen. Diese spannen den Beobachtungsraum auf. Hier wird davon abstrahiert, ob es sich um einen abstrakten oder um einen konkreten Raum handelt.

Auch hier sollen beschreibende und abgeleitete Metadaten unterschieden werden.

Zu den beschreibenden Beobachtungsraum-Metadaten gehören

1. die Art des Beobachtungsraumes und
2. seine Dimensionalität.

Zusätzlich werden im Fall, daß Raum- und/oder Zeitdimensionen vorhanden sind, folgende Metadaten definiert:

1. Wirkungskreis und
2. Verbund der Beobachtungspunkte sowie
3. Ausweisung der Raum- und
4. Zeitvariablen.

Ein Beispiel für abgeleitete Metadaten, das aufgrund seiner Allgemeingültigkeit in diese Spezifikation aufgenommen wird, sind Segmentierungen des Raumes zur Ausweisung von Teilraumeigenschaften. Zu den Eigenschaften von Teilräumen gehören z.B.

1. ihre Verteilungseigenschaften,
2. ihre Heterogenität,

3. ihre Qualität sowie

4. ihre Relevanz.

Im folgenden werden die vorgestellten Metadaten definiert und erläutert.

Mit der **Art des Beobachtungsraumes** soll unterschieden werden, ob die Dimensionen *räumliche und/oder zeitliche Größen* verschlüsseln, oder ob sie *abstrakte Größen* sind. Diese Unterscheidung ist für die Visualisierung sehr wichtig, weil es jeweils spezielle Darstellungstechniken gibt und die Art des Beobachtungsraumes adäquat abgebildet werden sollte.

Bei Daten in einem räumlichen und/oder zeitlichen Bezugssystem sind die Eigenschaften des Wirkungskreises und des Verbunds der Beobachtungspunkte von Bedeutung. Der **Wirkungskreis** legt fest, wie groß der Wirkungsbereich der Beobachtungspunkte ist. Bei *punktuellem* Wirkungskreis hat der Wert eines Beobachtungspunktes nur genau an dem Punkt Gültigkeit. Ist er *lokal*, so gilt der Wert des Punktes in einer gewissen Umgebung. Bei *globalem* Wirkungskreis hat der Wert des Punktes Einfluß auf den gesamten Beobachtungsraum.

Beobachtungspunkte sind häufig auf einem Gitter angeordnet. **Verbund**-Metadaten beschreiben ein solches Gitter, bzw. weisen aus, daß ein solches Gitter nicht existiert. Man unterscheidet zwischen *regelmäßigen*, *blockstrukturierten*, *strukturierten*, *unregelmäßigen* und *hybriden* Gittern. (vgl. [SM00] S. 33-35).

Weiterhin ist es wichtig, die **Raum-** und die **Zeitdimensionen** des Beobachtungsraumes gesondert auszuweisen. Diese zusammen bilden die Menge der unabhängigen Variablen.

Anhand der Ausweisung von **Segmentierungen des Raumes** können Teilräume unabhängig von der speziellen Datenklasse beschrieben werden. Hiermit sollen die speziellen Eigenschaften von Teilräumen spezifiziert und für den Visualisierungsprozeß nutzbar gemacht werden.

Wichtige Teilraumeigenschaften sind Verteilungseigenschaften, Heterogenität, Qualität und Relevanz von Teilräumen. Unter **Verteilungseigenschaften** fallen dabei Informationen wie minimale und maximale Werteausprägungen, Mittelwerte und Varianzen der Merkmale sowie die Anzahl der zugehörigen Beobachtungspunkte. **Heterogenität** faßt diese Eigenschaften in der Aussage zusammen, wie gleichmäßig oder wie ungleichmäßig ein Raumbereich ist. Die **Qualität** überträgt das allgemeine Konzept der Qualität auf Teilräume. Unter gewichteter Auswertung der anderen Teilraumeigenschaften kann so für jeden Raumbereich eine **Relevanz** festgelegt werden. Diese legt fest, wie wichtig die Information eines Raumbereichs ist. Die Relevanz eines Teilraumes kann z.B. umso höher sein, je höher seine Anzahl an Beobachtungspunkten, je größer seine Heterogenität und je größer seine Qualität ist.

Im folgenden soll die Relevanz der Beobachtungsraum-Metadaten für die Visualisierung an einigen Beispielen nachgewiesen werden:

1. Relevanz für die Vorverarbeitung

   - Im Falle höherdimensionaler abstrakter Räume kann in Abhängigkeit von der Dimensionalität des Beobachtungsraumes eine vorverarbeitende Projektion der unabhängigen Variablen erfolgen.
   - Die Segmentierungen von Teilräumen ermöglichen die vorverarbeitende Reduktion von Raumbereichen mit geringer Relevanz zur Beschleunigung von Interaktionen. Im Sinne der Trennung von für die gesamte Datenmenge geltenden Metadaten und lediglich für Teilmengen gültige Metadaten können diese Informationen auch benutzt werden, um die Analyse von speziellen Datenklassenanalysen zu beschleunigen.
   - Weiterhin können Segmentierungen von Teilräumen in der Vorverarbeitung genutzt werden, um Bereiche von besonderem Interesse zu bestimmen. Dazu müssen die relevanten aneinandergrenzenden Teilbereiche zusammengefaßt werden. Die erhaltenen Bereiche von Interesse können z.B. benutzt werden, um die Auswahl von Focusing- und Linking-Techniken zu steuern.

2. Relevanz für Visualisierungsentscheidungen

   - Die Art des Raumbezuges ist entscheidend für die Darstellung des Beobachtungsraumes.
   - Die Dimensionalität des Beobachtungsraumes und die Anzahl von Raumdimensionen wirkt sich direkt auf die Dimensionalität der Darstellung aus. So kann entweder eine 2D- oder eine 3D-Technik ausgewählt werden.
   - Raum- und/oder Zeitvariablen nehmen eine besondere Stellung bei der Visualisierung ein, weil sich hierfür bestimmte Visualisierungsmethoden durchgesetzt haben, wie beispielsweise geographische Karten oder Animationen.
   - Die Ergebnisse der Segmentierung können direkt genutzt werden, um die räumliche Verteilung von Gebieten von Interesse zu visualisieren.

3. Relevanz für Verfahren zur Metadatengewinnung

   - Die meisten Metadaten des Beobachtungsraumes sind Voraussetzungen für die Berechnung von Datenklassen.
   - Der Wirkungskreis der Beobachtungspunkte beeinflußt z.B., ob ein nicht regelmäßiges Gitter in ein regelmäßiges Gitter durch Interpolation überführt werden kann.

3.2.2 Datenmenge-Metadaten

Zu den Datenmenge-Metadaten zählen alle Metadaten, die die Datenmenge in ihrer Gesamtheit beschreiben bzw. wichtige Strukturierungen der Datenmenge er-

fassen. Für die Metadaten-Spezifikation wurden sie unterteilt in die allgemeinen Datenmenge-Metadaten und in die sogenannten Beobachtungsfall-Metadaten.

3.2.2.1 Allgemeine Datenmenge-Metadaten

In den allgemeinen Datenmenge-Metadaten werden alle die Metadaten spezifiziert, die Informationen über die gesamte Datenmenge zusammenfassen.

Im speziellen sind das vor allem Metadaten, welche die Beziehung der Variablen untereinander ausdrücken.

Hierzu gehören:

1. die Anzahl der Variablen,
2. die Anzahl der Datensätze,
3. die Variablenstrukturierung mit
   (a) Schlüsselinformationen,
   (b) gemeinsamen Informationsgehalten,
   (c) gemeinsamen Korrelationen sowie
   (d) Variablenhierarchien,
4. die Widerspruchsfreiheit der Datenmenge,
5. die Menge redundanter Merkmale und
6. der mittlere Informationsgehalt der Datenmenge.

Die Anzahl der Variablen und die Anzahl der Datensätze sind beschreibende Metadaten, wogegen die anderen Beispiele für abgeleitete Metadaten sind.

Unter **Variablenstrukturierung** wurden alle direkten Beziehungen von Variablen oder Variablenteilmengen zusammengefaßt. Die zugehörigen Metadaten werden im folgenden spezifiziert und erläutert.

Nicht immer sind in einer Datenmenge abhängige und unabhängige Variable explizit getrennt. Es wurde aber schon darauf hingewiesen, daß diese Trennung bei der Visualisierung eine wichtige Rolle spielt. Aus diesem Grund soll an dieser Stelle die Definition von **Schlüsselinformationen** als Teil der Variablenstrukturierung Verwendung finden. Die Schlüsselanalyse zur Vorverarbeitung von zu visualisierenden Daten zu benutzen wurde z.B. in [Lan97] vorgeschlagen. Bezug nehmend auf den Begriff Schlüssel aus der Datenbanktechnologie (vgl. z.B. [HS95] S. 63ff.) sollen für eine gegebene Datenmenge die Mengen von Schlüsselvariablen[10] gefunden werden. Jede Menge von Schlüsselvariablen bildet einen Schlüssel. Eigenschaft von Schlüsseln ist, daß ein Tupel von Schlüsselausprägungen in einer Tabelle genau einen

[10] In [HS95] werden diese als „Schlüsselattribute“ bezeichnet. Da in der Visualisierung jedoch zwischen Dimensionen und Attributen (Merkmalen) getrennt wird, wird hier der allgemeinere Begriff der „Schlüsselvariablen“ verwendet.

Datensatz identifiziert. Damit definieren sie eine Abbildung, mit welcher ein Tupel von Ausprägungen unabhängiger Variablen[11] eindeutig ein Tupel von Ausprägungen abhängiger Variablen festlegt. Ziel der Integration der Schlüssel in das Konzept ist es, diese Abbildungseigenschaft auszunutzen und den Nutzer durch Vorschlagen verschiedener Schlüssel bei der Trennung der Variablen in abhängige und unabhängige zu unterstützen.

Gibt es für die existierende Datenmenge keinen Schlüssel oder definiert lediglich die Gesamtmenge der Variablen einen Schlüssel, so liegt die Vermutung nahe, daß die Datenmenge nicht widerspruchsfrei ist. Beispiel hierfür ist z.B. das Auftreten von identischen Datensätzen. Diese Eigenschaft wird im Metadatum **Widerspruchsfreiheit** widergespiegelt. Die Widerspruchsfreiheit beeinflußt die Gesamtqualität der Datenmenge.

Gemeinsame Informationsgehalte beinhalten, inwieweit die Informationen in zwei oder mehreren Variablen miteinander verknüpft sind (vgl. z.B. [The95]). Eine hohe gemeinsame Information einer Menge von Variablen bedeutet, daß für Ausprägungen bei einer Variablen häufig bestimmte Ausprägungen bei den anderen Variablen vorliegen. Sie sind also miteinander gekoppelt. Da die gemeinsamen Informationsgehalte, wie sie in [The95] definiert werden, sich lediglich auf Ausprägungen beziehen, sind sie in erster Linie für nominale Daten verwendbar. Ordnungen bzw. Abstände der einzelnen Ausprägungen werden nicht beachtet (zum mathematischen Hintergrund vgl. Abschnitt 5.3.2.4).

Wie bereits vorgestellt, lassen sich gemeinsame Informationsgehalte vorwiegend auf Daten mit nominalen Skalentyp anwenden. Um auch die Ähnlichkeiten von kontinuierlichen Merkmalsausprägungen einzubeziehen, wurden zusätzlich die Korrelationen von Merkmalen in das Konzept aufgenommen. Diese beinhalten neben der Stärke der Abhängigkeit auch die Art der Abhängigkeit. Beispiel hierfür ist beim Ansatz der linearen Korrelation die direkte und die indirekte Proportionalität.

Eine weitere wichtige Strukturbeziehung von Variablen sind ihre hierarchischen Abhängigkeiten. Deshalb wird die Metadatenbeschreibung von **Variablenhierarchien** durchgeführt. Zum Beispiel wird in [Rob90] eine Hierarchie von Variablen mit dem Ziel vorgestellt, voneinander abhängige Variable nicht gleichrangig zu behandeln. Beispiel für hierarchische Variablen ist das Auftreten der Variablen Jahr, Monat und Tag. In der Hierarchie sollte die Information vorliegen, daß Monat eine Teileinheit von Jahr und Tag eine Teileinheit von Monat ist. Die Beachtung von hierarchischen Abhängigkeiten von Variablen ist wichtig, um Informationsverzerrungen durch Nichtbeachtung von wichtigen Zusammenhängen[12] zu vermeiden.

Weiterhin denkbar ist, mit Hilfe der Faktorenanalyse voneinander unabhängige Faktoren zu bestimmen. Korrelierte Variable werden zusammengefaßt, um lediglich die unabhängigen Einflußfaktoren zu beachten (vgl. z.B. [BEPW96]). Integriert man die Faktoren als übergeordnete Variable und die eigentlichen Variablen als untergeordnete in eine Hierarchie, wird ebenfalls eine Hierarchisierung durchgeführt. Diese

[11] Dies sind die Schlüsselvariable.

[12] In diesem Beispiel würde die Information dadurch verzerrt, daß drei Zeitachsen statt einer auftreten.

ist dann im Gegensatz zum beschreibenden Hierarchie-Metadatum aus [Rob90] eine abgeleitetes Hierarchie-Metadatum.

Die **Menge redundanter Merkmale** beinhaltet die Definition von für den weiteren Verlauf der Metadatengewinnung nicht zu beachtenden Merkmalen. Ein Merkmal kann beispielsweise als redundant ausgewiesen werden, wenn sein Wertebereich nur eine einzige Ausprägung besitzt oder es sehr stark mit einem anderen Merkmal korreliert ist. Ein Weglassen von solchen Merkmalen kann eine wesentliche Beschleunigung der folgenden Metadatenbestimmungen ermöglichen, ohne daß dabei wichtige inhaltliche Beziehungen verlorengehen.

Der **mittlere Informationsgehalt** beschreibt, wie groß die Information der gesamten Datenmenge im Mittel ist.

Die Relevanz dieser Metadaten soll im folgenden an Beispielen nachgewiesen werden:

1. Relevanz für die Vorverarbeitung

   - Im Falle, daß die Datenmenge nicht widerspruchsfrei ist, sollten in der Vorverarbeitung Datenverbesserungen vorgenommen werden.

2. Relevanz für Visualisierungsentscheidungen

   - Die Anzahl der Datensätze und die Anzahl der Variablen ist für verschiedene Visualisierungsentscheidungen von Bedeutung. Zum Beispiel ist es für verschiedene Techniken wichtig, wie hoch die Übersichtlichkeit der Darstellung bei einer bestimmten Datenmengengröße sein kann. So kann beim Überschreiten einer Größe bsw. die Entscheidung zu einer vorverarbeitenden Selektion oder zur Wahl einer Technik mit Verträglichkeit einer entsprechend großen Datenmenge getroffen werden.
   - Die Tupellänge eines minimalen Schlüssel kann die Anzahl der in Frage kommenden Visualisierungstechniken einschränken. Weiterhin erlaubt das Auffinden eines Schlüssels einer gewissen Tupellänge die plausible Abbildung der zugehörigen Daten-Variablen auf Visualisierungsvariablen (vgl. [Lan97] S. 15/16).
   - Über Anzahl und Stärke von gemeinsamen Informationsgehalten und Korrelationen kann entschieden werden, welche Variablen in einer visuellen Darstellung zusammen dargestellt werden müssen, um diese Zusammenhänge sichtbar zu machen.
   - Existiert eine Variablenhierarchie, läßt sich dies insbesondere für Fokus- und Kontext-Techniken gut ausnutzen.
   - Der mittlere Informationsgehalt einer Datenmenge dient dem Vergleich mehrerer Datenmengen in bezug auf deren Inhalte.

3. Relevanz für Verfahren zur Metadatengewinnung

- Die Schlüsselinformationen bieten eine Hilfestellung zur Bestimmung von Abhängigkeit und Unabhängigkeit von Variablen. Weiterhin unterstützen sie die Bestimmung der Widerspruchsfreiheit.
- Die Menge redundanter Merkmale beeinflußt, welche Variablen im weiteren in die Verfahren zur Metadatengewinnung einzubeziehen sind.
- Variablenhierarchien sichern ab, daß bei nachfolgenden Metadatengewinnungen nur gleichwertige Variablen miteinander verglichen werden.

3.2.2.2 Beobachtungsfall-Metadaten

Ein Beobachtungsfall beinhaltet „die in einem k-dimensionalen Schnitt des Beobachtungsraumes enthaltenen Merkmalsausprägungen“ ([SM00] S. 39). Beobachtungsfall-Metadaten enthalten die Menge der interessanten *Beobachtungsfälle* und können dem allgemeinem Konzept der regions of interest zugeordnet werden. Damit ist unter Beobachtungsfall-Metadaten zu verstehen, daß in Abhängigkeit von der Datenmenge „interessante“ Beobachtungsfälle zusammengefaßt werden. Diese Unterräume des Beobachtungsraumes können in der Visualisierung je nach Art und Dimensionalität gesondert visualisiert werden, um die Interpretierbarkeit von speziellen Teilbereichen zu verbessern.

Im speziellen gehören zu den Beobachtungsfall-Metadaten

1. die Anzahl der Beobachtungsfälle und
2. für jeden Beobachtungsfall
   (a) Anzahl der Dimensionen,
   (b) Anzahl der Merkmale und
   (c) seine speziellen Eigenschaften.

3.2.3 Datenklassen-Metadaten

Anhand der speziellen Ausprägungen von Variablen- und Datenmenge-Metadaten lassen sich Datenklassen definieren. Die speziellen Eigenschaften dieser Datenklassen sind für die Visualisierung sehr wichtig und müssen deswegen betrachtet werden.

In den folgenden Abschnitten werden die Metadaten zu den einzelnen Datenklassen vorgestellt. Sie unterteilen sich in

- Strömungsdaten-Metadaten,
- Volumendaten-Metadaten und
- Multiparameterdaten-Metadaten.

3.2.3.1 Metadaten für Strömungsdaten

Strömungsdaten-Metadaten beinhalten alle für die Datenklasse der Strömungsdaten spezifischen Charakteristika. Grundeigenschaft von Strömungsdaten ist ein Strömungsfeld auf einen strukturierten Gitter. Das Strömungsfeld wird durch Vektoren festgelegt, die an jedem Gitterpunkt definiert sind. Die Vektoren eines Vektorfeldes und der Beobachtungsraum können 2- oder 3-dimensional sein. Ist das Strömungsfeld zeitveränderlich, spricht man von einer instationären, andernfalls von einer stationären Strömung.

In dieses Konzept integriert wurden als beschreibende Metadaten

1. die Zeitabhängigkeit und
2. Informationen zu Dimensionalität der Vektoren.

Unter der Vielzahl von möglichen abgeleiteten Metadaten wurden

1. die Strömungscharakteristik mit den speziellen Eigenschaften
   (a) der kritischen Punkte,
   (b) der Stoßfronten sowie
   (c) der Wirbel und
2. weitere allgemeine Feldeigenschaften

als wichtige Metadaten identifiziert.

Dabei weist die **Zeitabhängigkeit** aus, ob es sich um eine stationäre oder eine instationäre Strömung handelt. Die **Dimensionalität der Vektoren** legt fest, ob es sich um 2D- oder 3D-Vektoren handelt[13].

Unter **Strömungscharakteristik**-Metadaten fallen spezielle Eigenschaften des Strömungsfeldes. Das sind zum ersten die **kritischen Punkte**. Kritische Punkte sind die Punkte, an denen alle Komponenten des Geschwindigkeitsvektors gleich Null sind, was bedeutet, das an dieser Stelle keine Strömung vorliegt. In Abhängigkeit der Eigenschaften der partiellen Ableitungen der Strömung ergeben sich unterschiedliche Typen. Grundsätzlich unterscheidet man Senken, Quellen und Umströmungen nach der *Art der Flußrichtung*. Weiterhin ergeben sich Sattelpunkte, Wirbel, Knoten und Strudel nach deren *Form* (vgl. z.B. [Frü97] S.45). Die Lage, Anzahl und Art der kritischen Punkte eines Feldes legen dessen *Topologie* fest. Weiterhin von Bedeutung für die Interpretation von Strömungsfeldern sind deren **Wirbel- und Stoßfronten**. Dabei befinden sich Wirbel i.a. in Bereichen mit hoher Wirbelstärke und Stoßfronten an Diskontinuitäten von Druck und Dichte (vgl. [Frü97] S. 37 und S.48/49).

[13] Die Dimensionalität des Beobachtungsraumes ist nicht automatisch gleich der Dimensionalität der Strömungsvektoren: Ein 3D-Beobachtungsraum legt nicht zwingend fest, daß es sich um 3D-Vektoren handelt. Diese können bsw. auch die Strömung innerhalb einer dreidimensionalen Fläche repräsentieren und sind entsprechend zweidimensional.

Unter Metadaten zu **weiteren allgemeinen Feldeigenschaften** ist die Ausweisung von vor allem skalaren Eigenschaften des Feldes zu verstehen. Diese sind häufig auch direkt mit den Vektoren des Feldes korreliert oder von ihnen abhängig. Darunter fallen z.B. die Eigenschaften des Druckes, der Temperatur, der kinetischen Energie, der Krümmung innerhalb des Feldes u.a. (vgl. hierzu [Frü97] S. 37 und [The96]).

Die Relevanz dieser Metadaten für die Visualisierung wird im folgenden anhand von Beispielen nachgewiesen:

1. Relevanz für die Vorverarbeitung
   - Die Reduktion der Datenmenge[14] sollte in Abhängigkeit der Strömungscharakteristika vorgenommen werden. Sie sollte vor allem außerhalb der topologisch wichtigen Bereiche erfolgen.
2. Relevanz für Visualisierungsentscheidungen
   - Die Art der Veränderlichkeit unterstützt die Entscheidung, ob eine statische oder dynamische Strömungsvisualisierung gewählt werden sollte.
   - Grundlegendes Anliegen bei der Strömungsanalyse ist das Erkennen der topologischen Eigenschaften des Feldes. Art und Anzahl von bestimmten topologisch wichtigen Strömungscharakteristika können Visualisierungsentscheidungen dahingehend unterstützen, ob und wie diese dargestellt werden sollten[15].
   - Bei interessanten Verteilungen von allgemeinen Feldeigenschaften sollten diese gesondert dargestellt werden.

Aufgrund der hohen Spezialisierung sind die Strömungsdaten-Metadaten keine Voraussetzungen für weitere Metadatenerhebungen.

3.2.3.2 Metadaten für Volumendaten

Volumendaten-Metadaten fassen für Volumendaten spezifische Eigenschaften zusammen. Allgemein wird unter dem Begriff Volumendaten ein 3-dimensionales regelmäßiges Gitter mit einem quantitativen Wert an jedem Gitterpunkt verstanden. Diese Kategorisierung kann jedoch für die Metadaten-Gewinnung auch auf nicht 3-dimensionale regelmäßige Gitter übertragen werden.

Ein Großteil der Eigenschaften von Volumendaten wurden bereits bei den Beobachtungsraum-Metadaten erfaßt. Beispiele für abgeleitete Metadaten für Volumendaten sind

1. die spezifischen Teilraumeigenschaften mit

[14] z.B. durch Verkleinerung des Gitters

[15] Reicht eine implizite Darstellung der Charakteristika aus oder sollten sie explizit z.B durch Ikonendarstellung von kritischen Punkten visualisiert werden (vgl. [SM00] S. 326)

   (a) dem Funktionsverhalten sowie

   (b) den Körpereigenschaften der Teilräume und

2. die Eigenschaften des Gradientenfeldes.

Die **Teilraumeigenschaften**-Metadaten ordnen sich in die Ausweisung von Strukturen und von Bereichen von Interesse für Volumendaten ein. Hierbei werden Teilbereiche des Raumes spezifiziert und ihnen aus Sicht der Volumenvisualisierung interessante Aspekte zugeordnet. Darunter fallen das **Funktionsverhalten** der 3D-Funktion, die durch die Datenmenge definiert wird. Beispielsweise ist das Monotonieverhalten dieser Funktion zur Unterscheidung von gleichmäßigen und ungleichmäßigen Raumbereichen von Interesse. Darauf basierend können die ungleichmäßigen Bereiche in Vorverarbeitung und Visualisierung speziell verarbeitet werden (in Arbeit [Köl00]). Weitere Teilraumeigenschaften sind die Eigenschaften von durch die Volumendaten **Körpern**. Dazu zählen z.B. die Art, Form und Anzahl von entsprechenden Körpern.

Als weiteres Metadatum für Volumendaten sollen die **Eigenschaften des Gradientenfeldes** ausgewiesen werden. Hiermit werden Schwankungen in den Datenwerten geeignet beschrieben. Das Gradientenfeld ist ein Vektorfeld und kann entsprechend den Strömungsdaten-Metadaten spezifiziert werden. Somit können dann für das Gradientenfeld Strömungsvisualisierungstechniken angewandt werden.

3.2.3.3 Metadaten für Multiparameterdaten

Multiparameterdaten-Metadaten beinhalten alle für die Datenklasse der Multiparameterdaten spezifischen Charakteristika. Bei Multiparameterdaten handelt es sich allgemein um Daten, bei denen an jedem Beobachtungspunkt mindestens zwei skalare Merkmale vorliegen. Die Beobachtungspunkte müssen auf keinem speziellen Gitter angeordnet sein.

Viele grundlegende Eigenschaften, die häufig Multiparameterdaten zugeordnet werden, konnten bereits in den Variablen-Metadaten und Datenmenge-Metadaten verallgemeinert werden. Deswegen ist die Anzahl an Metadaten vergleichsweise gering, die ausschließlich für diesen Datentyp spezifiziert wird. Insbesondere brauchen hier keine beschreibenden Metadaten separat definiert werden.

Beispiele für abgeleitete Multiparameterdaten-Metadaten sind

1. Ausreißerdatensätze,

2. gleiche Merkmalsausprägungen und

3. die Strukturierung der Datensätze.

Ausreißer sind Datensätze, deren Merkmalsausprägungen sich stark von anderen Datensätzen unterscheiden. Diese sollten in einer Visualisierung gegebenenfalls speziell hervorgehoben oder sogar separat dargestellt werden.

Gleiche Merkmalsausprägungen treten auf, wenn Datensätze in allen abhängigen Variablen gleiche Ausprägungen haben. In der Visualisierung mit räumlichem Bezug können bsw. Gebiete gleicher Merkmalsausprägungen speziell hervorgehoben werden.

Die **Strukturierung** von Datensätzen ist eine Strukturbeschreibung speziell für Multiparameterdaten. Mit ihrer Hilfe kann allgemein festgestellt werden, wie die Datensätze im Merkmalsraum strukturiert sind. Für die Visualisierung von Bedeutung ist dabei, ob die spezielle Struktur der Daten visualisiert werden muß und wie dies erfolgen kann. Beispiel für eine Strukturierung ist eine Klassifikation der Datensätze, wie sie z.B. in [Noc99] durchgeführt wird. Mit einer solchen Klassifikation können Anzahl, Homogenitäten und Verteilungen von Datensatzklassen ausgewiesen werden. Dies kann zum einen in der Vorverarbeitung eine detailliertere Analyse von Klassifikationsstrukturen unterstützen und dabei z.B. die Zusammenfassung von Datensätzen zu Clustern in einer Datensatzreduktion ermöglichen. Zum anderen können signifikante Strukturen zur Entscheidung führen, diese in der Visualisierung direkt darzustellen. Weiterhin kann darüber die Entscheidung über die Anwendbarkeit von Focusing- und Linking-Techniken unterstützt werden.

3.2.4 Zusammenfassung

Grundziel bei der Aufstellung des Metadatenkonzeptes war die Spezifikation möglichst vieler visualisierungsrelevanter Metadaten vor allem unter dem Aspekt der Visualisierungsentscheidung. Dabei wurden viele in der Literatur isoliert betrachteten Aspekte, die meist nicht mit Blick auf Metadatendefinitionen sondern in einem anderen Umfeld eingeführt wurden, in ein einheitliches Konzept eingebunden. Zusammenfassend kann man sagen, daß ein Vielzahl von Metadaten identifiziert werden konnte. Solch eine umfassende Beschreibung in Bezug auf Visualisierungsentscheidungen liegt bisher noch nicht vor. Trotzdem sind Erweiterungen denkbar.

Bisher wurden die drei wichtigsten Datenklassen Multiparameter-, Volumen- und Strömungsdaten betrachtet. Weitere Datenklassen wie GIS[16] oder Scattered-Data[17] wurden hier nur indirekt über den Beobachtungsraum einbezogen, deren spezielle Eigenschaften jedoch nicht separat aufgenommen. Ihre Einbindung in das Konzept wurde jedoch offengehalten.

Desweiteren wurde das Konzept der Beobachtungsfall-Metadaten nur in Ansätzen skizziert. Eine Erweiterung ist jederzeit möglich und weiteren Arbeiten vorbehalten.

Ein Hauptziel bei der Metadatenspezifikation war die Allgemeingültigkeit der Metadaten. Zusammenfassend kann man sagen, daß viele Eigenschaften der Datenklassen in den Beobachtungsraum-Metadaten zusammengefaßt werden konnten. Weiterhin konnten üblicherweise zu den Multiparameterdaten gehörende Eigenschaften in die Variablen- und Datenmenge-Metadaten integriert werden, so daß die entsprechenden Eigenschaften auch für die anderen Datenklassen nutzbar sind.

[16] Geographische Informationssysteme
[17] Gestreute Daten

An späterer Stelle muß noch darüber diskutiert werden, ob die Vereinheitlichungen sinnvoll einsetzbar und praktikabel sind (vgl. Abschnitte 5.4 und Kapitel 6).

Kapitel 4

Steuerungs- und Ablaufkonzept zur Gewinnung von Metadaten

Dieses Kapitel hat zum Ziel, ein Konzept zur Gewinnung der im vorangegangenen Kapitel 3 vorgestellten Metadaten zu entwerfen. Dazu muß eine Abfolge entwickelt werden, in der die einzelnen Metadaten nach ihrer Abhängigkeit voneinander in einer geeigneten Reihenfolge bestimmt werden. Weiteres Ziel ist, mögliche Nutzeranforderungen an die Metadatengewinnung einzubeziehen und den Nutzer bei der Gewinnung der Metadaten zu unterstützen.

Abschnitt 4.1 stellt die allgemeinen Konzepte vor und Abschnitt 4.2 beinhaltet das entwickelte Ablaufschema. In Abschnitt 4.3 werden die Ergebnisse der Konzeption der Metadatengewinnung zusammengefaßt.

4.1 Allgemeine Grundlagen

Verschiedene Metadaten können unterschiedlich gut[1] und unterschiedlich schnell erhoben werden. Daraus ergibt sich, daß der Nutzer in Abhängigkeit von der Art des Metadatums stärker oder schwächer in den Bestimmungsprozeß eingebunden werden muß. Hierbei lassen sich folgende Stufen unterscheiden:

1. interaktives Festlegen der Metadaten
2. interaktives Entscheiden zur Steuerung des Bestimmungsprozesses
3. halbautomatische Bestimmung bzw.
4. automatische Bestimmung der Metadaten.

Ist ein Metadatum nicht algorithmisch berechenbar, muß es *interaktiv festgelegt* werden. *Interaktive Entscheidungen* dienen der Steuerung des Bestimmungsprozesses.

[1] Gut meint in diesem Zusammenhang, wie sicher die Aussage eines bestimmten Metadatum ist. Beispielsweise kann eine als nominal klassifizierte Variable häufig ordinal sein, was algorithmisch schwer feststellbar ist.

Dabei wird bestimmt, welche Arten von Metadaten im weiteren erhoben werden sollen. *Automatische* Metadatengewinnung erfordert keine Interaktion und bei *halbautomatischer* Gewinnung können vorberechnete Metadaten durch den Nutzer abgeändert werden.

Bei der Festlegung von Metadaten für eine gegebene Datenmenge können alle vier Stufen zur Anwendung kommen. Um eine einheitliche Arbeitsweise zu garantieren, wird davon ausgegangen, daß die Metadaten in so bezeichneten Bestimmungsmodulen erhoben werden. Bestimmungsmodule können demnach Berechnungen durchführen und/oder interaktive Nutzerabfragen auswerten. Ziel ist, sie in ein interaktives Werkzeug zur Metadatenerhebung zu integrieren (vgl. Kapitel 5).

Die interaktive Festlegung von Metadaten ist gerade für ungeübte Nutzer nicht immer einfach. Denkbar wäre hier in Abhängigkeit eines *Nutzerprofils* Unterstützung zu geben. Mögliche Profiltypen wären hierbei Anfänger, Fach-Spezialist und Visualisierungsexperte. Die Bestimmungsmodule könnten an den jeweiligen Nutzer unterschiedlich angepaßt werden, d.h. entsprechende Standardbelegungen vorgeben oder Hilfestellungen einblenden. Je erfahrener der Nutzer ist, um so mehr kann er in den Metadatenbestimmungsprozeß eingebunden werden. So könnte weiterhin durch Nutzung des Vorwissens eines Nutzers z.B. die Datenklassenauswahl gesteuert und so entschieden werden, welche speziellen Datenklassen-Metadaten erhoben werden sollen.

Ein weiterer wichtiger Punkt ist die Beachtung von *Effektivität* und *Zeitaufwand* bei der Metadatenerhebung. Hiermit soll erreicht werden, daß anhand von Voruntersuchungen schnell entschieden wird, ob für eine spezielle Datenmenge ein bestimmtes Metadatum effektiv erhoben werden kann oder nicht. Daraus wird die Entscheidung abgeleitet, ob eine Metadatengewinnung ausgeführt oder auf sie verzichtet wird.

Allgemein soll jedoch gelten, daß die Gewinnung von automatischen Metadaten durchaus etwas länger dauern darf, solange keine ständige Interaktion mit dem Nutzer erforderlich ist. Dies ist sinnvoll, weil zum einen durch entsprechende Metadaten die Visualisierung und vor allem die Interaktion während der Visualisierung beschleunigt werden kann. Zum anderen können so ungeeignete Darstellungstechniken, die zu Fehlinterpretationen führen könnten, vermieden werden.

Wichtig für das Verständnis und die Nutzbarkeit eines Werkzeugs zur Erfassung von Metadaten ist, jederzeit den aktuellen Stand der erhobenen Metadaten darzustellen. Dabei soll vor allem erkenntlich sein, welche Daten bisher erhoben wurden und wozu die gemachten Entscheidungen geführt haben.

4.2 Ablauf der Metadatengewinnung

In diesem Abschnitt soll das Ablaufkonzept im einzelnen vorgestellt werden. Dazu wird eine geeignete Reihenfolge mit dem Ziel entwickelt, Abhängigkeiten der Metadatenerhebungen zu beachten und diese entsprechend anzuordnen, um die Bestimmung von Metadaten optimal zu unterstützen.

Im speziellen wurde für dieses Ablaufkonzept das Vorliegen der Rohdaten in Form von Tabellen zugrunde gelegt, um eine praktikable Arbeitsgrundlage zu schaffen. Vor allem die Bestimmung von beschreibenden Metadaten, bei der bsw. Zeichenkettenanalysen durchgeführt oder die Eigenschaften des Gitters analysiert werden, ist speziell auf die Tabellenform zugeschnitten. Diese Festlegung bedeutet jedoch keine große Einschränkung, denn es besteht die Möglichkeit, auch andere Formen von Rohdaten zu integrieren. Dafür müssen teilweise nur bestimmte Module übersprungen werden, insbesondere dann, wenn die Speicherung gewisser Metadaten bereits implizit gegeben ist.

Hauptablaufschema Abbildung 4.1 zeigt die allgemeine Prozeßfolge für die Gewinnung der Gruppen von Metadaten[2] aus Abschnitt 3.2. Zuerst werden dabei die Rohdaten eingelesen. Anschließend erfolgt die Bestimmung der unterschiedlichen Metadaten und abschließend deren Speicherung.

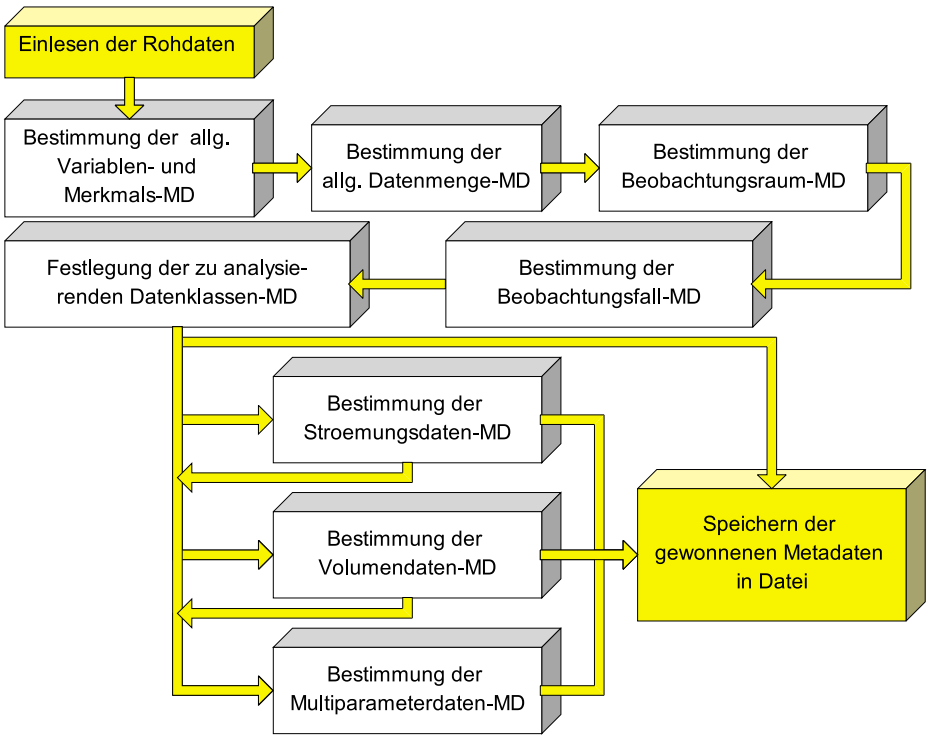

Abbildung 4.1: Prozeßkette der Metadatengewinnung

Die Bestimmungen der allgemeineren Metadaten Variablen-, Merkmals-, Beobachtungsraum-, Beobachtungsfall- und Datenmenge-Metadaten werden vor den datenklassen-spezifischen Metadaten Strömungs-, Volumen- und Multiparameterdaten durchgeführt.

[2]Mit Gruppe von Metadaten sollen im folgenden die in Abschnitt 3.2 spezifizierten übergeordneten Metadaten wie z.B. die allgemeinen Variablen- oder die Beobachtungsraum-Metadaten verstanden werden.

Weil bei den allgemeinen Variablen-Metadaten und der Merkmals-Metadaten überwiegend beschreibende Metadaten erhoben werden, werden diese vor den anderen Gruppen von Metadaten bestimmt. Desweiteren werden diese beiden Gruppen vereinfachend zusammen erhoben.

Die Datenmenge-Metadaten-Gewinnung trennt abhängige und unabhängige Variablen voneinander und ist deswegen vor der Beobachtungsraum-Metadaten-Gewinnung angesiedelt. Die Gewinnung der Beobachtungsfall- und der Datenklasse-Metadaten basiert dagegen auf den Beobachtungsraum-Metadaten und wird deswegen erst danach durchgeführt.

Im Anschluß an die Beobachtungsfall-Metadatengewinnung wird festgelegt, welche speziellen Datenklassen-Metadaten erhoben werden können und sollen. Davon abhängig können dann entweder gar keine spezielle Datenklassenanalyse, nur eine der drei, zwei oder alle drei Analysen durchgeführt werden. Entspricht die Datenmenge keiner der drei Datenklassen kann auch auf die Analyse von speziellen Datenklassencharakteristika verzichtet werden.

Im folgenden werden die Prozeßabläufe im einzelnen vorgestellt:

Gewinnung der Variablen- und Merkmals-Metadaten Abbildung 4.2 zeigt die allgemeine Prozeßfolge für die Gewinnung der allgemeinen Variablen-Metadaten und der Merkmals-Metadaten aus den Abschnitten 3.2.1.1 und 3.2.1.2.

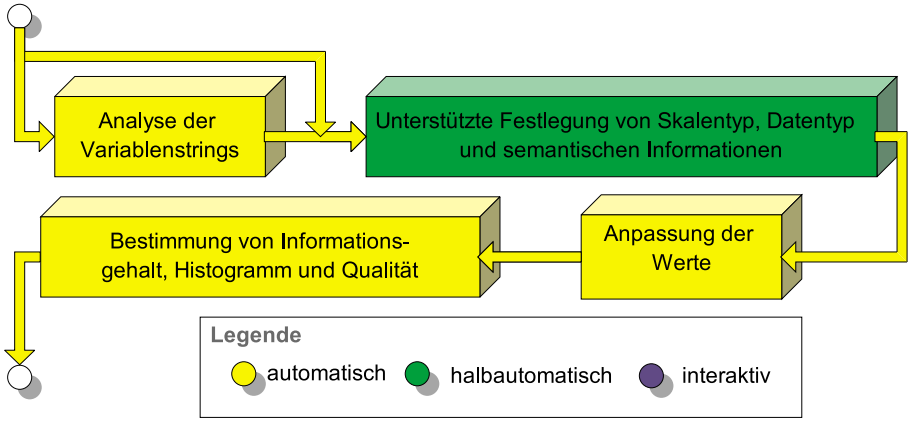

Abbildung 4.2: Allgemeine Variablen- und Merkmals-Metadaten-Gewinnung

Im Falle einer expliziten Speicherung der Rohdaten erfolgt eine Analyse der Variablenstrings. Diese kann bei impliziter Speicherung übersprungen werden, weil Art und Typen der Variablen dort bereits vorgegeben sind. Im Anschluß erfolgt die unterstützte Festlegung des Skalentyps, der semantischen Informationen für alle Variablen und des Datentyps speziell für die Merkmale. Anschließend werden die eingelesenen Werte in ein internes Format umgewandelt. Dies ist notwendig, da sie die Durchführung von internen Berechnungen eine einheitliche Datenstruktur vorliegen muß. Das kann z.B. bedeuten, daß nominale oder ordinale Zeichenketten in ein

internes Format konvertiert werden oder daß bei impliziter Speicherung die Werte der Merkmale eingebunden werden. Zuletzt erfolgt die Bestimmung von Informationsgehalten, Histogrammeigenschaften und Qualitäten der einzelnen Variablen, weil diese von Datentyp, Skalentyp und den konvertierten Werten abhängen.

Gewinnung der allgemeinen Datenmenge-Metadaten Abbildung 4.3 zeigt das Ablaufschema für die Gewinnung der allgemeinen Datenmenge-Metadaten aus Abschnitt 3.2.2.

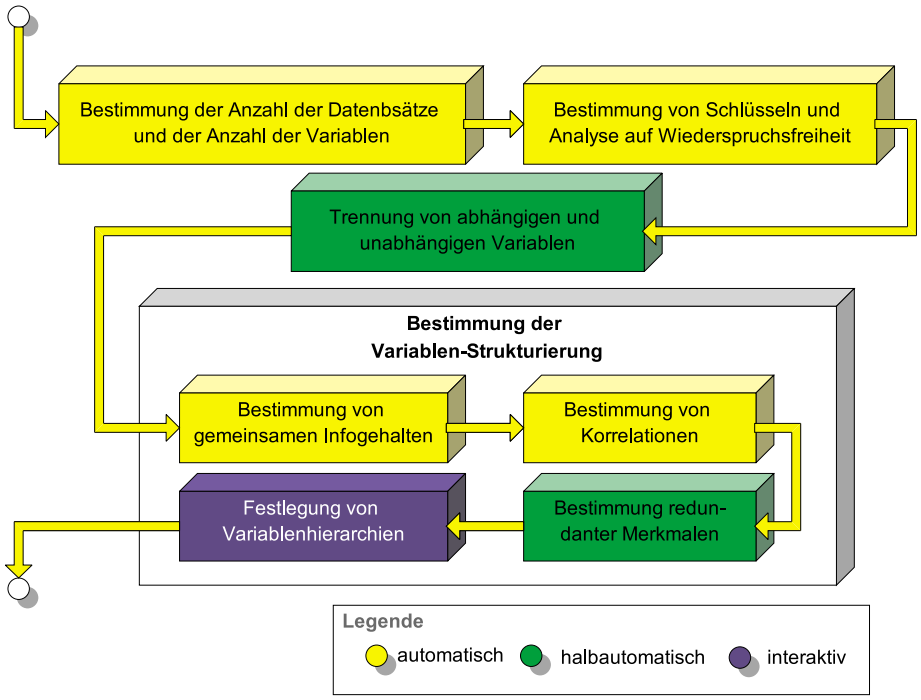

Abbildung 4.3: Allgemeine Datenmenge-Metadaten-Gewinnung

Erster Schritt bei der Gewinnung der allgemeinen Datenmenge-Metadaten ist die Bestimmung der beschreibenden Metadaten Anzahl der Datensätze und Anzahl der Variablen. Diese sind Voraussetzung für die folgenden Analysen. Im Anschluß erfolgt die Bestimmung von Schlüsseln und die Analyse auf Widerspruchsfreiheit. Unterstützt durch die Schlüsselanalyse werden dann abhängige und unabhängige Variablen von einander getrennt. Der dritte Schritt ist die Bestimmung der Variablen-Strukturierung. Zuerst werden hierbei die gemeinsamen Informationsgehalte und die Variablen-Korrelationen bestimmt. Diese unterstützen dann die Auswahl redundanter Merkmale[3]. Letzte Strukturierung stellt die Festlegung von Variablenabhängigkeiten in einer Hierarchie dar. Hierbei können die als redundant bestimmten Merk-

[3]Zum Beispiel könnte bei sehr hohen Korrelationen zweier Merkmale ein Merkmal als redundant eingestuft werden. Weiterhin sollen Merkmale mit redundanten Merkmalsausprägungen und einem entsprechenden Informationsgehalt gleich 0 im folgenden vereinfachend auch als redundant bezeichnet und dementsprechend behandelt werden. Zu überlegen ist, ob die Bestimmung der gemeinsamen Informationsgehalte in der praktischen Umsetzung nach der Bestimmung der

male beachtet werden, indem für diese bsw. keine hierarchischen Abhängigkeiten eingefügt werden.

Gewinnung der Beobachtungsraum-Metadaten Abbildung 4.4 zeigt das Ablaufschema für die Gewinnung der Beobachtungsraum-Metadaten aus Abschnitt 3.2.1.3.

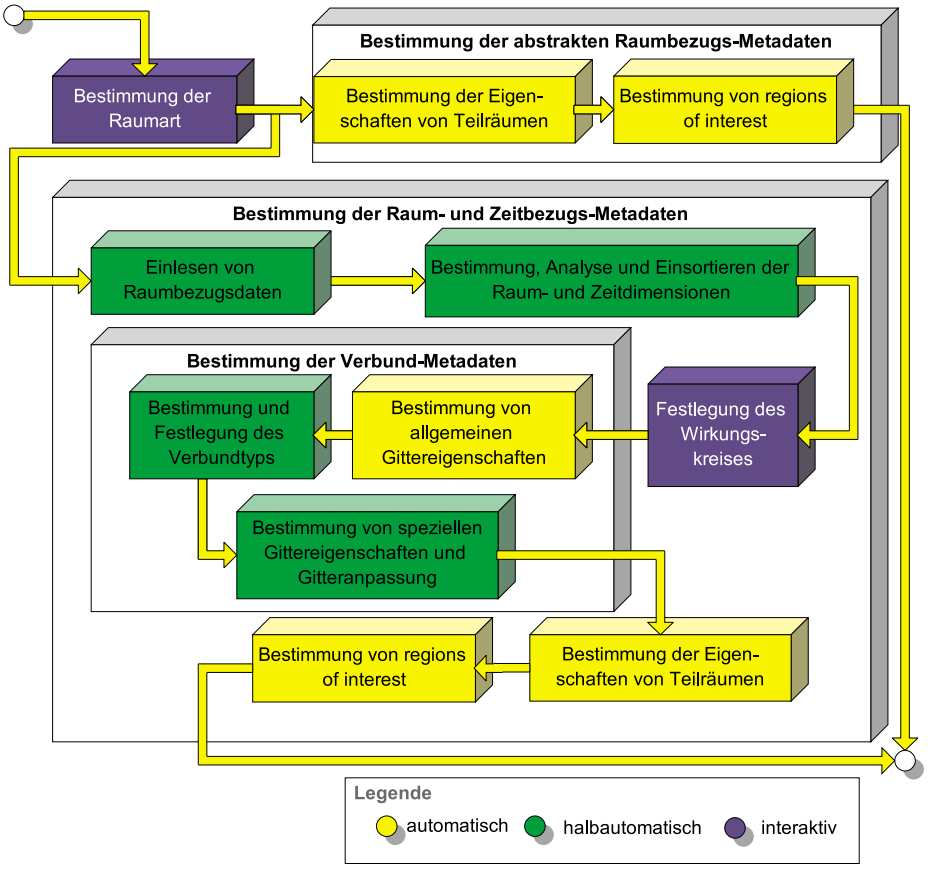

Abbildung 4.4: Beobachtungsraum-Metadaten-Gewinnung

Bei der „Bestimmung der Raumart“ wird entschieden, ob es sich entweder um Raum- bzw. Zeitdimensionen und/oder um abstrakte Dimensionen handelt. Im Falle der abstrakten Dimensionen werden die Eigenschaften von Teilräumen und die daraus resultierenden Bereiche von Interesse bestimmt.

Im Falle des Vorliegens von Raum- und Zeitdimensionen erfolgt eine komplexe Bestimmung der speziellen Eigenschaften des Raumes. Als erstes besteht dazu die Möglichkeit, die Raum- und/oder Zeitdaten aus einer separaten Tabelle einzulesen. Diese werden dann als Raum- oder als Zeitvariable einsortiert und deren

redundanten Merkmale erfolgen sollte, um die Geschwindigkeit zu erhöhen. Dies ist bei den beiden beschriebenen Redundanzkriterien möglich, weil der einfache Informationsgehalt bereits in der Variablen-Metadaten-Gewinnung bestimmt wurde. Bei einem Redundanzkriterium in Abhängigkeit von den gemeinsamen Informationsgehalten wäre dies nicht möglich.

Eigenschaften analysiert[4]. Im Anschluß daran erfolgt die interaktive Festlegung des Wirkungskreises und die Bestimmung der Verbund-Metadaten. Bei der Verbund-Bestimmung wird zuerst auf allgemeine Gittereigenschaften wie Regelmäßigkeit, Blockstrukturiertheit u.s.w. analysiert. Diese sollen den Nutzer bei der anschließenden Festlegung des Verbundtyps unterstützen. Im Anschluß erfolgt die Analyse der speziellen Gittereigenschaften wie Gitterabstände, Gitterhöhen, Vollständigkeit[5] u.s.w. In Abhängigkeit von dieser Analyse kann dann eine Gitteranpassung durchgeführt werden. Beispielsweise kann bei lokalem oder globalen Wirkungskreis eine Gitteranpassung durch Werteinterpolation durchgeführt werden. Abschließend werden dann analog zum abstrakten Bezug Eigenschaften von Teilräumen bestimmt und daraus regions of interest abgeleitet.

Gewinnung der Beobachtungsfall-Metadaten Die Gewinnung der interessanten Beobachtungsfälle wurde als Konzept für diese Arbeit nicht genauer ausgebaut. Vorgesehen ist in diesem Bestimmungsmodul, den Nutzer beim Finden interessanter Beobachtungsfälle zu unterstützen.

Festlegung der Datenklassen-Metadaten Abbildung 4.5 zeigt die Abfolge im Bestimmungsmodul zur Festlegung, welche speziellen Datenklassen-Metadaten (vgl. Abschnitt 3.2.3) erhoben werden sollen.

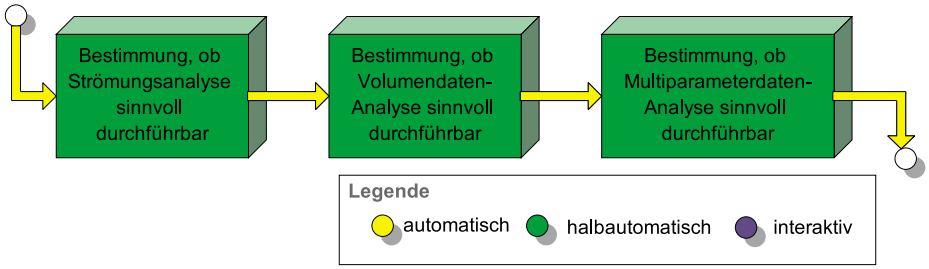

Abbildung 4.5: Festlegung der Datenklassen-Metadaten-Gewinnung

Dazu wird hintereinander bestimmt, ob auf der Datenmenge Analysen für die einzelnen Datentypen sinnvoll sind oder nicht. Beispielsweise könnte eine Strömungsdatenmenge mit Strömungs- und Multiparameteranalysen untersucht werden. Genauso kann auch in einer Multiparameterdatenmenge ein Merkmal als Volumenmerkmal aufgefaßt und entsprechende Volumenanalysen durchgeführt werden (vgl. [GLdCS97] S. 10ff.).

Die Reihenfolge der Module ist in diesem Fall nicht notwendigerweise festgelegt. Praktisch kann aus der Entscheidung, ob eine Datenklassen-Metadaten-Gewinnung durchgeführt wird, die Entscheidung zur Durchführung oder zur Nichtdurchführung einer anderen Datenklassen-Metadaten-Gewinnung abgeleitet werden. Die Module

[4] z.B. kann eine allgemeine Variablen-, eine Merkmals- und eine Datenmenge-Metadaten-Gewinnung durchgeführt werden.

[5] Mit Vollständigkeit ist gemeint, ob für alle Gitterpunkte z.B. in einem regelmäßigen Gitter Werte vorliegen oder ob einige Werte fehlen.

können aber auch unabhängig voneinander sein.

Gewinnung der Strömungsdaten-Metadaten Abbildung 4.6 zeigt das Ablaufschema für die Gewinnung der Strömungsdaten-Metadaten aus Abschnitt 3.2.3.1.

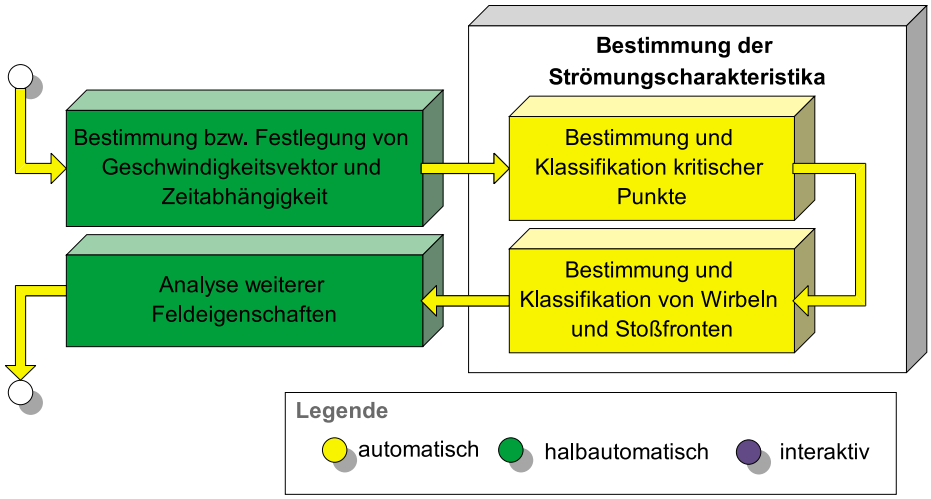

Abbildung 4.6: Strömungsdaten-Metadaten-Gewinnung

Zuerst werden in diesem Modul der Geschwindigkeitsvektor[6] und die Zeitabhängigkeit des Strömungsfeldes anhand etwaiger in den Beobachtungsraum-Metadaten vorhandenen Zeitvariablen festgelegt. Die Hauptströmungsanalyse erfolgt dann im Modul zur Bestimmung der Strömungscharakteristika. Dort können dann z.B. die Lage und die Art von kritischen Punkten, Wirbeln und Stoßfronten bestimmt werden. Abschließend erfolgt die Analyse weiterer vor allem skalarer Feldeigenschaften.

Die festgelegte Reihenfolge ist hier von Bedeutung. Zum einen sind die beschreibenden Metadaten Geschwindigkeitsvektor und die Zeitabhängigkeit Voraussetzungen für die Bestimmung der anderen abgeleiteten Metadaten. Desweiteren kann die Bestimmung der Strömungscharakteristika die Analyse der weiteren skalaren Feldeigenschaften beeinflussen. Denkbar wäre z.B., die skalaren Feldeigenschaften lediglich in der Umgebung der kritischen Punkte zu erheben.

Gewinnung der Volumendaten-Metadaten Abbildung 4.7 zeigt das Ablaufschema für die Gewinnung der Volumendaten-Metadaten aus Abschnitt 3.2.3.2.

Dabei werden zuerst die volumenspezifischen Teilraumeigenschaften Funktionsverhalten der 3D-Funktion und Innenkörpereigenschaften bestimmt. Anschließend werden unter Nutzung der Strömungsdaten-Metadaten-Gewinnung die Eigenschaften des Gradientenfeldes analysiert.

Bei der Gewinnung der Volumendaten-Metadaten spielt die Reihenfolge der Bestimmungsmodule keine Rolle.

[6] Bei expliziter Speicherung können mehrere skalare Merkmale als Vektorkomponenten interpretiert und zusammengefaßt werden.

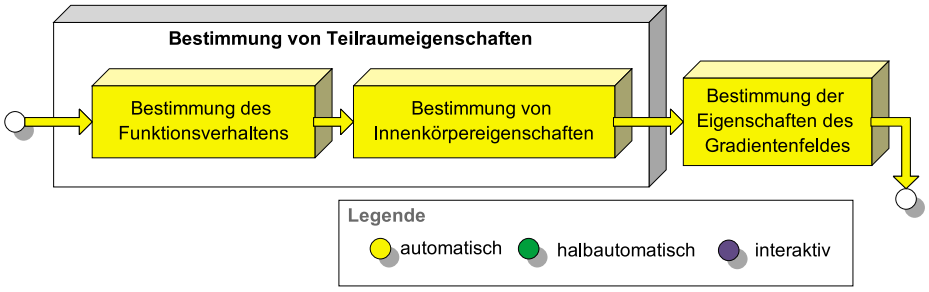

Abbildung 4.7: Volumendaten-Metadaten-Gewinnung

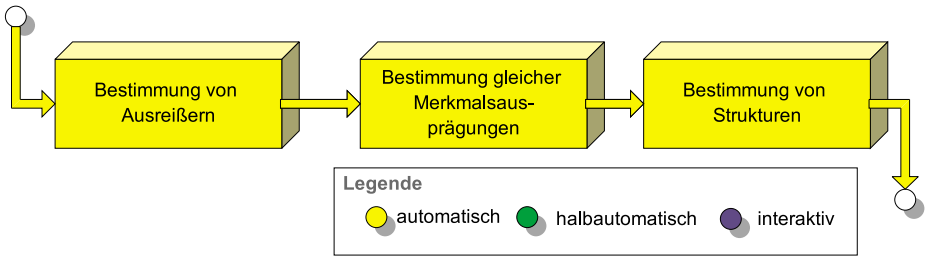

Abbildung 4.8: Multiparameter-Metadaten-Gewinnung

Gewinnung der Multiparameterdaten-Metadaten Abbildung 4.8 zeigt das Ablaufschema für die Gewinnung der Multiparameterdaten-Metadaten aus Abschnitt 3.2.3.3.

Zuerst werden dabei Ausreißerdatensätze bestimmt. Im Anschluß werden Datensätze mit gleichen Merkmalsausprägungen gewonnen. Letzter Schritt ist die Durchführung von Strukturanalysen in der Datenmenge. In diesen Analysen können je nach Nutzer- und Algorithmusanforderungen Ausreißer und gleiche Datensätze eliminiert oder beachtet werden. Deshalb steht dieser Schritt am Ende des Ablaufschemas.

4.3 Zusammenfassung

Zusammenfassend kann festgestellt werden, daß es gelungen ist, allgemeine Grundsätze für eine Metadatengewinnung zu definieren und für spezielle Metadaten entsprechende Ablaufmodule zu entwickeln. Weiterhin wurde eine geeignete Reihenfolge der Ablaufmodule festgelegt, so daß sich die Metadatengewinnungen gegenseitig unterstützen. Auch innerhalb der einzelnen Ablaufmodule konnten die wichtigsten Metadatenarten in einer geeigneten Reihenfolge zusammengefügt werden. Dabei werden beschreibende Metadaten vor abgeleiteten Metadaten erhoben.

Mit dem entwickelten Ablaufkonzept wurde die Grundlage für die Implementation eines Programmes zur Metadatengewinnung gelegt.

Kapitel 5

„Metadatum“ - Umsetzung eines Werkzeuges zur Metadatengewinnung

In diesem Kapitel wird die Umsetzung der Konzepte aus Kapitel 3 und 4 im Programm „Metadatum“ vorgestellt. Zunächst werden die Architektur des Programmes und daraus abgeleitete Entscheidungen zur Implementationsumgebung und zu Programmierdesign-Richtlinien vorgestellt (Abschnitt 5.1). Danach erfolgt die Beschreibung der Ein- und Ausgabeschnittstelle (Abschnitt 5.2). Im Anschluß werden die verschiedenen Module, die die Metadatengewinnung durchführen, vorgestellt (Abschnitt 5.3). Dort erfolgt auch die Vorstellung wichtiger Algorithmen zur Bestimmung spezieller Metadaten. Abschließend werden die Leistungsfähigkeit und die Grenzen des Werkzeuges „Metadatum“ bewertet (Abschnitt 5.4).

5.1 Architektur und allgemeine Grundlagen

5.1.1 Architektur der Metadatengewinnung

Abbildung 5.1 zeigt die Architektur des Programms „Metadatum“. Die Metadatengewinnung überführt Eingabedaten, die entweder bereits gewonnene Metadaten oder Rohdaten[1] sind, in Ausgabedaten, welche gespeicherte Metadaten sind. Die interne Struktur, die diese Abbildung umsetzt, trennt drei Module. Jedes dieser Module ist unabhängig von den anderen und kommuniziert mit diesen über vorgegebene Schnittstellen. Speziell handelt es sich um die Module *Steuerung*, *Bestimmung* und *Interaktion*. Das Steuerungsmodul bestimmt das Nutzerprofil, liest die Daten ein, speichert diese und startet die Metadatenbestimmung. Das Bestimmungsmodul beinhaltet die reihenfolgeabhängige Gewinnung der Metadaten. Hierzu werden automatische Analyseverfahren gestartet und Nutzerinteraktionen an den entsprechenden Stellen im Ablauf angesteuert. Das Interaktionsmodul realisiert die Eingabe

[1] Die Rohdaten liegen in Tabellen vor (vgl. Abs. 5.2).

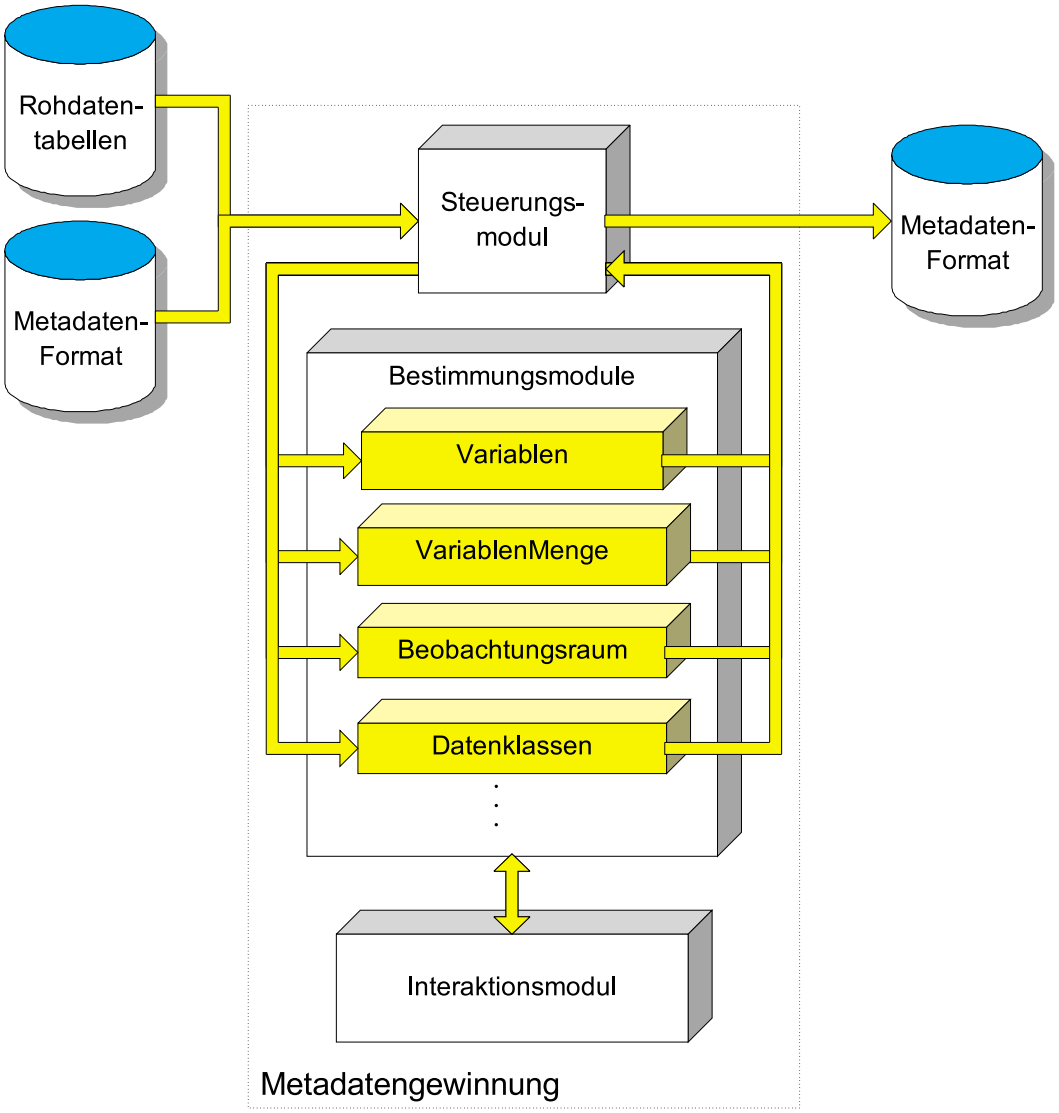

Abbildung 5.1: Architektur der Metadatengewinnung

oder die Änderung von Metadaten durch den Nutzer. Weiterhin dokumentiert es für den Nutzer den aktuellen Stand der Metadatenerhebung.

5.1.2 Umgebung

Die Implementierung erfolgte unter dem Betriebssystem „Windows NT“ in der Programmiersprache „Visual C++ 6.0“. Hauptgrund hierfür war, daß einerseits die Programmiersprache C aufgrund von Geschwindigkeit und Flexibilität gut für mathematische Aufgaben geeignet ist und in ihr bereits vielfältige Algorithmen umgesetzt sind. Diese wurden in das Programm „Metadatum“ integriert.

Durch die Untergliederung in separate Module kann die Metadatengewinnung sowohl als Commandline-Anwendung als auch als Windows-Programm übersetzt werden. Das Bestimmungsmodul ist von der jeweiligen Art der Anwendung unabhängig. Es handelt sich dabei um eine C++-Implementierung unter Nutzung der Standard-Template-Library (STL). Die Steuerung erfolgt im Commandline-Programm über die Klasse „Main“ und im Windows-Programm über entsprechende Windowsfunktionen (vgl. Abschnitt 5.3.1). Für das Windows-Programm wurde eine Multi-Document-Interface-Architektur[2] gewählt. Jede Metadatenerhebung ist so einem Dokument zugeordnet. Es können somit mehrere Metadatenerhebungen zu einem Rohdatum mit unterschiedlichen Nutzereinstellungen als auch Metadaten mit unterschiedliche Rohdaten gleichzeitig eingelesen und parallel berechnet werden.

Weiterhin wurde im Interaktionsmodul eine Windows-Schnittstelle implementiert, um die Ergebnisse der Metadatenerhebung anzuzeigen und Interaktionen zur Steuerung und Manipulation des Bestimmungsmoduls komfortabel zu gestalten. Wichtig hierbei war die saubere Trennung von Bestimmungsfluß und Interaktion, um eine größtmögliche Portabilität zu gewährleisten. Deswegen fiel die Entscheidung bei der Nutzung der internen Datenstrukturen auf die „Standard Template Library“, während die Windows-Funktionalität davon abgetrennt mit Hilfe der MFC realisiert wurde.

5.1.3 Designrichtlinien

Die Programmierung erfolgte nach der Metapher der Objektorientierung. Wiederverwendung von Funktionen und Membervariablen durch Vererbung wurden weitgehend durchgeführt. Funktionen mit ähnlichen Aufgabenbereichen wurden in Klassen zur Erhöhung der Übersichtlichkeit des Quell-Codes zusammengefaßt. Ziel war, durch Objektorientierung die Wiederverwendbarkeit und die flexible Erweiterbarkeit zu ermöglichen.

Auf die Verwendung globaler Variablen wurde weitestgehend verzichtet. Dies ist vor allem unter dem Gesichtspunkt der MDI-Architektur sinnvoll, so daß mehrere parallele Metadatenerhebungen durchführbar sind.

Weitere Richtlinie bei der Programmierung ist die Kompatibilität des Kerncodes, d.h. der Metadatengewinnung. Deswegen erfolgen die Interaktionen und Sta-

[2] Abgekürzt: MDI-Architektur

tusanzeigen durch plattformunabhängige Systemaufrufe. So kann z.B. die Funktion „Status“ im Commandline-Programm zum Schreiben auf die Standardausgabe im CMD-Fenster führen, wogegen im Windows-Programm dieselbe Information in einem Statusfenster dargestellt wird.

Bei der Definition von Klassen und bei der Nutzung von temporären Strukturen wurden entweder eigene Datenstrukturen geschaffen, STL-Datenstrukturen benutzt oder eine hybride Programmierung von eigenen Datenstrukturen unter Nutzung der STL durchgeführt. Windows-spezifische Datenstrukturen wie MFC-interne wurden aus Gründen der Portabilität nur im Windows-spezifischen Teil verwendet.

Eine weitere wichtige Frage ist die Art und Weise der internen Fehlerbehandlung. Diese erfolgt in zwei unterschiedlichen Arten. Handelt es sich um einen Programmierungsfehler, wird dieser mit einem Debug-Assert abgefangen. Zur besseren Verständlichkeit des Programmcodes und dem schnelleren Finden von Fehlern sollen Invarianten, die unbedingt gelten müssen, mit Hilfe dieser Debug-Asserts eingefügt werden. Soll am Ende des Implementierungsprozesses ein schnelles Programm erzeugt werden, werden die Debug-Asserts nicht übersetzt.

Die zweite Art sind Fehler, die aufgrund fehlerhafter Nutzereingaben oder fehlerhafter Ein- bzw. Ausgabe entstehen. Bei diesen Fehlern darf sich das Programm nicht beenden, sondern die Fehler müssen abgefangen werden. Dies geschieht mit dem C++-Konzept der exceptions.

5.2 Ein- und Ausgabeschnittstelle

Wie aus der Architektur der Metadatengewinnung ersichtlich wird (vgl. Abb. 5.1), können Metadaten aus Rohdaten und aus Metadaten-Dateien eingelesen und in Metadaten-Dateien gespeichert werden. Die Formate dieser beiden Dateitypen werden im folgenden näher beschrieben. Desweiteren wird die spezielle Umsetzung des separaten Einlesens von Raum- und/oder Zeitdimensionen vorgestellt[3].

Aus der Definition der Datenformate ergibt sich, wie die Metadatengewinnung die Eingabe in die Ausgabe umwandeln (vgl. Abschnitt 5.3).

5.2.1 Rohdaten

Grundsätzlich können die Rohdaten sowohl aus Dateien als auch aus Datenbanken eingelesen werden. Basierend auf dem in Abschnitt 3.1 vorgestellten Konzept zur Art der Eingabedaten stand bei der Implementierung die Tabellenform im Vordergrund.

Wegen der großen Vielfalt unterschiedlicher Dateiformate wurden als Eingabeformat im Dateifall ASCII-Tabellen festgelegt, in denen die einzelnen Werte bei-

[3]Separates Einlesen beinhaltet, daß die Raum- und/oder Zeitdimensionen explizit in einer anderen Tabelle vorliegen. Im Gegensatz zu einer impliziten ist bei der expliziten Speicherung die Position eines Beobachtungspunktes nicht durch die Stelle in einem Array, sondern explizit durch die in der Tabelle vorhandenen Dimensionen festgelegt. Eine weitere Möglichkeit neben dem Einlesen aus einer separaten Tabelle ist das Einlesen der expliziten Raum- und/oder Zeitdimensionen aus der Tabelle der Merkmale.

spielsweise durch Tabulator-Zeichen getrennt sind. Diese haben den Vorteil, sehr flexibel zu sein, da sie bsw. aus Internetbrowsern oder aus Excel exportiert werden können.

Das Einlesen von ASCII-Tabellen erfolgt mit Hilfe eines Scanners, in dem unterschiedliche Arten der Spaltentrennung und der Leerspaltenbehandlung integriert sind[4].

Analog zum Einlesen der Rohdaten aus ASCII-Tabellen können diese auch aus Datenbanken eingelesen werden. Wichtig für das Programm war die Integration einer Datenbankschnittstelle, weil ein Großteil der Daten in praktischen Anwendungen in Datenbanken vorliegen. Ihr Vorteil gegenüber ASCII-Tabellen ohne weitere semantische Informationen ist, daß die Typinformationen der einzelnen Datenbankspalten zur Bestimmung der Metadaten genutzt werden können (vgl. Ausblick in Kapitel 7).

Der Hauptnachteil an Tabellen ist, daß ihre Nutzung durch die explizite Definition des Raumbezuges, falls ein solcher vorhanden ist, einerseits und durch die nichtkomprimierte Speicherung von Zahlenwerten andererseits vor allem bei Volumen- und Strömungsdaten langsamer und zeitintensiver als die implizite Speicherung des Raumbezuges in Binärformat ist. Nachteil der impliziten Speicherung ist jedoch, daß sie auf ein bestimmtes einfaches Werteformat[5] eingeschränkt ist und *ausschließlich* vollständige reguläre Gitter unterstützt. Um die Geschwindigkeitsvorteile der impliziten Speicherung nutzen zu können, werden in der aktuellen Programmversion als Tabellendaten gespeicherte Volumen- und Strömungsdaten intern in die implizite Form umgewandelt.

Sinnvoll wäre als nächster Schritt, auch Daten mit einer impliziten Speicherung der Raum- und/oder Zeitvariablen einzulesen. Problem hierbei ist jedoch die Vielzahl an möglichen Formaten. Allerdings könnte man durch ein Auslesen von Headerinformationen Art und Strukturierung der Daten für bestimmte Dateitypen umsetzen und für diese Typen die Vorteile des schnelleren Einlesens und des schnelleren internen Zugriffs nutzen. Die Schnittstelle auf die Datenwerte über die interne Wertedatenstruktur müßte dann lediglich erweitert werden.

5.2.2 Metadatenformat

In diesem Abschnitt wird das Format zur Speicherung der gewonnenen Metadaten vorgestellt. Die Schaffung eines solchen Formates hat drei Hauptgründe. Zum ersten soll eine durchgeführte Metadatengewinnung, die unterbrochen werden mußte, wieder aufgenommen werden können. Dies ist vor allem unter dem Gesichtspunkt sinnvoll, daß die Analyseprozesse teilweise länger dauern können. Zum zweiten soll

[4]Der Zeichenketten-Scanner wurde von Matthias Kreuseler entwickelt und für diese Arbeit angepaßt und erweitert. Er dient dazu, unterschiedliche Tabellenformate in ähnlicher Weise einlesen zu können. So können die einzelnen Variablenausprägungen sowohl durch Anführungszeichen eingegrenzt als auch einfach durch Leerzeichen oder Tabulatoren getrennt sein. Leerspalten werden durch mehrere aufeinanderfolgende Tabulatoren beschrieben.

[5]z.B. double oder integer

so ermöglicht werden bereits bestimmte Metadaten bei einer abgewandelten Analyse der gleichen Datenmenge wieder verwenden zu können. Dies kann bsw. dann geschehen, wenn die visuelle Analyse und deren Interpretation es erfordern, die Daten unter neuem Blickwinkel zu untersuchen. Zum dritten ist der Hauptgedanke bei der Schaffung eines solchen Formates, daß unterschiedliche Visualisierungssysteme dieses Format als Grundlage nutzen sollen, um ihre Visualisierungsentscheidungen durchzuführen oder vorberechnete Metadaten anderweitig einzusetzen.

Das Metadaten-Format ist eng an die in Abschnitt 3.2 vorgestellten speziellen Metadatenarten angelehnt. Hauptziel bei der Definition dieses Formates war es, eine eindeutige Abbildung der internen Metadatenstrukturen auf die gespeicherten Daten und umgekehrt zu entwickeln, mit der für genau eine interne Struktur eine Datei und aus einer Datei genau eine interne Struktur erzeugt werden.

Das Dateiformat der Metadaten beinhaltet, daß alle Haupt-Metadatenarten nacheinander und in ihnen entsprechend Teilmetadaten vorliegen. So werden nacheinander die Variablen-, die Variablenmenge-, die Beobachtungsraum-, die Datenklasse-, die Strömungsdaten-, die Volumendaten-, die Multiparameterdaten- und die Werte-Metadaten in das Dateiformat eingefügt oder ausgelesen. Abbildung 5.2 zeigt das Grundschema einer Metadaten-Datei im ASCII-Format.

```
Variablenmetadaten:
  Variable: Variablenname1
    Skalentyp: ordinal, exklusiv
    Informationsgehalt: 0.7821
    ...
  Variable: Variablenname2
  ...
Beobachtungsraummetadaten:
  ...
Datenmengemetadaten:
  ...
...
```

Abbildung 5.2: Ausschnitt aus einer Metadaten-Datei im ASCII-Format

In Anhang C befindet sich eine praktisches Beispiel für eine Metadaten-Datei.

Die Implementierung der Speicherung erfolgt über die internen Metadatenstrukturen. Bei Umsetzung der speziellen Konzepte wurden für die einzelnen Teilmetadaten entsprechende Datenstrukturen geschaffen und jede dieser Metadatenarten kann in eine Datei geschrieben oder aus einer Datei gelesen werden. Die Ein- und Ausgabe beinhaltet, daß für jede der Gruppe von Metadatenarten entweder die Funktion „LeseAusDatei“ oder „SchreibeInDatei“ angestoßen wird. Beim Schreiben werden dabei die internen Datenstrukturen abgelegt, so daß diese bei Wiedereinlesen zur Erzeugung interner Strukturen genutzt werden können.

Grundsätzlich sind hierbei zwei Lese- und Schreib-Modi vorgesehen. Zum einen

wurde eine Speicherung im ASCII-Format implementiert. Vorteil dieser Speicherung ist, daß die Metadaten in der Datei mit einen Texteditor angeschaut und manuell modifiziert werden können. Weiterhin kann ein die Metadaten nutzendes Programm durch Suche nach entsprechenden Schlüsselwörtern schnell die interessierenden speziellen Metadaten finden. Nachteil dieser Speicherung ist der hohe Speicheraufwand und eine relativ hohe Einlese- und Schreibdauer. Zum zweiten wurde in die Implementation deswegen bereits die Speicherung als Binärformat eingebettet, jedoch noch nicht vollständig umgesetzt. Vorteil des Binärformates sind dessen komprimierte Datenspeicherung. Nachteil ist entsprechend das kompliziertere Ein- und Auslesen sowie eine erschwerte Modifikation.

5.2.3 Einlesen von separatem Raum- und/oder Zeitbezug

Als eine weitere Verbesserungsmöglichkeit für nicht regelmäßig strukturierte Gitter mit Dimensionen in Zeit und Raum wurde das Konzept der separaten Speicherung des Raum- bzw. Zeitbezuges umgesetzt. Das bedeutet, daß über eine Id in der Haupt-Rohdatentabelle ein Raum- oder Zeitbezug zu einer zweiten Rohdatentabelle[6] hergestellt wird, aus der die Dimensionsausprägungen ausgelesen werden können. So wird die redundante Speicherung von Dimensionen, die beim Speichern aller Dimensionen in einer Tabelle auftreten würde, vermieden. Diese Vorgehensweise lohnt sich dann, wenn entweder viele gleiche Meßorte oder Meßzeiten auftreten oder mögliche komplexe Zusatzinformation an den Raumbezugspunkten gespeichert sind.

Für diese zweite Rohdatentabelle werden ebenfalls die allgemeinen Variablen-, die Merkmals- und die allgemeinen Datenmenge-Metadaten erhoben, um in gleicher Weise auch auf die Daten dieser Tabelle zugreifen zu können. Von den Beobachtungsraum-Metadaten werden dann sowohl die Dimensionen der Ausgangs- als auch die Dimensionen der neuen Tabelle verwaltet. Dafür wurde eine Schnittstelle geschaffen, für die im späteren Verlauf der Berechnungen auf alle Dimensionen einheitlich zugegriffen werden kann, ohne wissen zu müssen, ob eine separate Raum-Bezugstabelle vorliegt oder nicht. In der Implementierung wurde weiterhin das Einlesen weiterer Dimensionstabellen vorgesehen, jedoch noch nicht vollständig implementiert.

Ein Beispiel für eine separate Raumtabelle befindet sich in Anhang B.2

5.3 Module zur Metadatengewinnung

Nachdem im vorangegangen Abschnitt die Ein- und Ausgabeschnittstelle vorgestellt wurde, sollen nun die Module im einzelnen vorgestellt werden, welche die Rohdaten in das Ausgabeformat umwandeln.

[6] sogenannte Raumbezugstabelle

5.3.1 Steuerung

Das Hauptmodul der Metadatengewinnung ist die Steuerung. Die Steuerung

1. verwaltet die Nutzerprofile sowie

2. die Ein- und Ausgabeschnittstelle und

3. startet die speziellen Metadatenbestimmungen.

Vom Konzept der Nutzerprofile wurden exemplarisch die Automationsgrade umgesetzt. Im Programm gibt es die vier einstellbaren Grade „interaktiv“, „halbautomatisch“, „automatisch“ und „maximal“. Bei „interaktivem“ Automationsgrad kann der Nutzer *alle* interaktiv veränderbaren Metadaten modifizieren. Im Gegensatz dazu führt der „automatische“ Automationsgrad dazu, daß nur die wichtigsten Entscheidungen und Eingaben vom Nutzer durchgeführt werden müssen; der Rest erfolgt über Standardbelegungen (vgl. Anhang A). Bei „halbautomatischem“ Automationsgrad werden nicht alle, jedoch mehr Interaktionen als bei automatischem Grad durchgeführt. Bei „maximalem“ Automationsgrad werden *keine* Interaktionen durchgeführt.

Zusätzlich wurden im Sinne des Nutzerprofilkonzeptes die Alternativen schnelle Berechnung oder ausführliche Berechnung integriert. Diese setzen das Konzept zu Effektivität und Zeitintensität von Metadatenerhebungen (vgl. Abschnitt 4.1) um. Im Falle der schnellen Berechnung werden bei großen Datenmengen entweder die Parameter so modifiziert, so daß die Berechnung trotzdem relativ schnell ausgeführt werden kann, oder es wird ganz auf die Bestimmung von zeitintensiven Metadaten verzichtet. Im ausführlichen Fall werden alle Metadaten mit den Standardparametern berechnet.

Der Ein- und Ausgabemechanismus wurde bereits in Abschnitt 5.2 beschrieben. Die Steuerung stellt die Funktionalität zur Verfügung, das Lesen und Schreiben von Metadaten und das Einlesen der Rohdaten aus ASCII-Tabellen und Datenbanken zu verwalten. Dazu müssen Dateinamen und Pfade bestimmt, Datenbanken geöffnet und Dateien für das Lesen und Schreiben vorbereitet werden. Im Windows-Fall erfolgt die Speicherung in Abhängigkeit von Windows-Dokumenten (MDI-Architektur). Jedes Dokument beinhaltet dabei die entsprechende Menge von gewonnenen Metadaten, welche durch die Lese- und Schreib-Funktion „serialize“ des Dokuments eingelesen oder geschrieben wird.

Desweiteren startet die Steuerung die jeweiligen Bestimmungsmodule. Im Commandline-Programm erfolgt der Aufruf der Bestimmungsmodule sequentiell. Im Windows-Programm kann der Nutzer explizit das zu berechnende Metadatum auswählen oder auch einfach die sequentielle Abfolge durchlaufen lassen. Bei der expliziten Auswahl der Gewinnung eines bestimmten Metadatums müssen jedoch die Metadaten, von denen seine Bestimmung abhängt, bereits gewonnen worden sein. Von ihm abhängige Metadaten werden gelöscht.

5.3.2 Bestimmung

Die Umsetzung der Bestimmungsmodule orientiert sich weitgehend an der Konzeption zur Gewinnung der Metadaten (vgl. Kapitel 4). Für jede Gruppe von Metadaten aus Abschnitt 3.2 wurde ein zugehöriges Bestimmungsmodul entworfen und umgesetzt, indem eine zugehörige C++-Klasse definiert und implementiert wurde. Nicht sinnvoll war der Entwurf von Bestimmungsmodulen für jedes einzelne Metadatum. Dies hat den Grund, daß sich Metadaten oft im Zusammenhang bestimmen lassen. Außerdem würde das Programm durch eine große Anzahl von Bestimmungsmodulen unnötig komplexer.

Die C++-Klassen der Bestimmungsmodule enthalten das Nutzerprofil, die Hauptfunktion „calculateMetadatum“ und die die Bestimmung beinflussenden Metadaten. Weiterhin beinhalten diese C++-Klassen für die Bestimmungen der einzelnen Metadaten zugehörige Funktionen sowie Hilfsfunktionen und Hilfsdatenstrukturen zur Unterstützung der Bestimmungen.

Jedes Objekt einer solchen Bestimmungsmodul-Klasse erhält bei seiner Konstruktion die benötigten Metadaten und das Nutzerprofil. Konzeptionell kann das Nutzerprofil also vor jeder Berechnung einer Gruppe von Metadaten verändert werden[7]. Im Anschluß wird die Bestimmung durch Ausführen der „calculateMetadatum“ Funktion gestartet.

Die Schnittstelle zum Interaktionsmodul wurde durch eine Menge von Kommunikationsfunktionen umgesetzt. An den entsprechenden Stellen im Bestimmungsmodul werden die Funktionen „Status“, „MDAendern“, „bestimmeDateiNameUndPfad“, „FehlerMeldung“ u.a. aufgerufen. Je nach Interaktionsschnittstelle werden diese Funktionen unterschiedlich verarbeitet (vgl. Abschnitt 5.3.3).

Da dieses Programm viele unterschiedliche und auch bereits vorhandene Algorithmen nutzt, besteht der Bedarf, diese effektiv einzubinden. Dazu gibt es zwei entgegengesetzte Möglichkeiten. Zum einen können die vorhandenen Algorithmen in den Quellcode des entsprechenden Bestimmungsmodul integriert werden. Dazu müssen die Datenstrukturen des eingebundenen Algorithmus an die Metadatenstruktur angepaßt werden. Alternative Möglichkeit ist die Nutzung der Algorithmen in Bibliotheken. Hierbei müssen die Eingabe-Metadatenstrukturen in das entsprechende Algorithmen-Datenformat überführt werden bzw. eine funktionale Datenzugriffsschnittstelle geschaffen werden. Nachteil der ersten Variante ist ein erhöhter Programmieraufwand. Das Kopieren der Daten in die Algorithmen-Datenstruktur (Variante 2) kann hingegen bei großen Datenmengen sehr zeit- und speicheraufwendig sein. Beste Variante ist, über Schnittstellen-Funktionen den Zugriff auf die Metadaten auszuführen. Ob dies möglich ist, hängt vor allem von der Programmierweise des Algorithmus ab. Die Entscheidung wurde in den speziellen Fällen unterschiedlich gefällt.

In den folgenden Unterabschnitten werden wichtige implementierte Algorithmen

[7] Soll beispielsweise die große Zeit erfordernde Bestimmung des Datenmenge-Metadatum beschleunigt werden, werden die anderen Metadaten im ausführlichen und nur dieses Metadatum im schnellen Modus gewonnen.

und gegebenenfalls dafür entwickelte Metadatenstrukturen vorgestellt.

5.3.2.1 Analysen von Zeichenketten zur Bestimmung der Skalentypen und der Umwandlung der Werte in ein internes Format

Die in einer Tabelle vorliegenden Zeichenketten werden u.a. analysiert, um die Skalentypen der Variablen zu bestimmen und darauf aufbauend die Zeichenketten in ein internes Format umzuwandeln. Dazu werden die Zeichenketten zuerst in ein Zeichenkettenhistogramm einsortiert, um mit ihnen im folgenden effektiver arbeiten zu können. Anschließend erfolgt die Analyse auf mögliche Trennzeichen in den Zeichenketten, um die Bestimmung von nicht-exklusiven Variablen vorzubereiten. Beispielsweise könnten die unterschiedlichen Ausprägungen einer Variablen durch ein Komma voneinander abgetrennt sein. Dann wird bestimmt, ob Kommas in den Zeichenketten vorliegen. Falls dies der Fall ist, wird eine Aufspaltung der Zeichenketten durchgeführt. In Abhängigkeit von den entstehenden Teil-Zeichenketten-Mengen wird vorberechnet, ob es sich um exklusive Variable handeln könnte oder nicht.

Anschließend werden die Zeichenketten der Variablenausprägungen in ein internes Zahlenformat umgewandelt[8]. Dies ist wichtig, damit der Nutzer in der anschließenden interaktiven Änderung der Skalentypen für ordinale Variable eine Ordnung festlegen kann. Weiterhin legt der Nutzer dort fest, ob die vorberechnete Exklusivität korrekt ist.

Anschließend werden je nach Skalentyp alle Variablen in das internes Zahlenformat umgewandelt und als Zeiteinsparung gleichzeitig in die Histogramm-Metadaten eingefügt. Hier erfolgt auch die endgültige Trennung von nicht-exklusiven Variablen. Bei der Speicherung in das interne Zahlenformat kann resultierend für einen Variablen-Wert eines Datensatzes entweder kein, ein oder mehrere Werte eingetragen werden, je nachdem, ob es sich um einen Fehlwert, einen Wert einer exklusiven oder einen Wert einer nicht-exklusiven Variable handelt.

5.3.2.2 Bestimmung von Schlüsseln

Für die Bestimmung von Schlüsseln wird im Programm „Metadatum“ ein von Susanne Lange speziell für Visualisierungszwecke umgesetzter Algorithmus verwendet (vgl. [Lan97]). Dieser bestimmt minimale Schlüssel[9] einer in Tabellenform gegebenen Datenmenge.

Weil praktische Datenmengen sehr groß sein können, kann eine Schlüsselanalyse dort sehr zeitaufwendig sein. Deswegen beschränkt sich der von ihr umgesetzte Algorithmus auf die Anzahl von maximal drei Schlüsselvariablen. In den meisten getesteten Beispiel-Datenmengen war dies für die Metadatengewinnung ausreichend. Wenn die Anzahl der unabhängigen Variablen einer Datenmenge die Zahl drei übersteigt, findet der Algorithmus keine Schlüssel. In diesen Fall muß die interaktive

[8] Binäre, ordinale und nominale Variable werden auf die Werte $0.0, 1.0, 2.0, \ldots$ abgebildet.

[9] Ein minimaler Schlüssel ist ein Schlüssel, für den alle Kombinationen von Teilmengen seiner Schlüsselvariablen keine Schlüssel sind.

Trennung der abhängigen und der unabhängigen Variablen erfolgen.

Das Grundschema des Algorithmus soll hier vorgestellt werden:

1. Tupellänge 1: Analysiere für alle Variable, ob doppelte Ausprägungen vorliegen. Ist dies nicht der Fall, wird die Variable als Schlüssel eingefügt.

2. Tupellänge 2: Analysiere alle Kombinationen von 2 Variablen, ob Ausprägungspaare doppelt vorliegen. Ist dies nicht der Fall und bildet keine der beiden Variablen einen Schlüssel, füge das Paar als Schlüssel ein.

3. Tupellänge 3: Analysiere alle Kombinationen von 3 Variablen, ob Ausprägungstripel doppelt vorliegen. Ist dies nicht der Fall und bildet keine der Teil-Kombinationen aus den drei Variablen einen Schlüssel, füge das Tripel als Schlüssel ein.

Weiterhin wurde der Algorithmus u.a. durch Anwendung einer Hashing-Strategie beschleunigt.

Für das Programm „Metadatum“ wurde die Schnittstelle der Schlüsselanalyse so angepaßt, daß die Werte direkt aus den Werten der Metadatengewinnung ausgelesen werden können. Dies hat die Gründe, daß das Kopieren der internen Werte in die vorgegebene Datenstruktur der Schlüsselanalyse bei großen Datensätzen zeit- und vor allem speicherplatzaufwendig gewesen wäre und die Schnittstellenanpassung einfach durchzuführen war.

5.3.2.3 Bestimmung von Korrelationen

Für das Programm „Metadatum“ wurde die lineare Korrelation von zwei Variablen beispielhaft umgesetzt. Die lineare Korrelation zweier diskreter Variable j und k wird über den empirischen Korrelationskoeffizienten ϱ_{jk} berechnet:

$$\varrho_{jk} := \frac{\sigma_{jk}}{\sigma_j \cdot \sigma_k} \tag{5.1}$$

Hierbei sind σ_j und σ_k die Varianzen der Variablen j und k mit

$$\begin{aligned} \sigma_i &:= \sqrt{\frac{1}{N}\sum_{k=1}^{N}(x_{ki} - \bar{x}_i)^2} \quad \text{und} \\ \bar{x}_i &:= \frac{1}{N}\sum_{k=1}^{N} x_{ki} \quad \text{(Mittelwert).} \end{aligned} \tag{5.2}$$

σ_{jk} ist die Kovarianz der Variablen j und k, welche durch

$$\sigma_{jk} := \frac{1}{N}\sum_{i=1}^{N}(x_{ji} - \bar{x}_j) \cdot (x_{ki} - \bar{x}_k) \tag{5.3}$$

definiert ist. N bezeichnet dabei die Anzahl der gemeinsamen Datensätze und x_{ji} die Ausprägung der Variable j im i-ten Datensatz.

Beim Vorliegen von fehlenden Datenwerten in den Variablen werden bei der Berechnung des empirischen Korrelationskoeffizienten nur die Datensätze einbezogen, in denen beide Werte vorliegen.

5.3.2.4 Bestimmung der gemeinsamen Informationsgehalte

Basierend auf der Definition der gemeinsamen Informationsgehalte aus [The95] wurde ein Algorithmus entwickelt und realisiert. Die gemeinsame Information $I(X_1, \ldots, X_q)$ eines Tupels von q diskreten Zufallsgrößen $X_1, \ldots, X_q$ wird in [The95] durch

$$I(X_1, \ldots, X_q) := \begin{cases} I'(X_1, \ldots, X_q), & \text{falls } I'(X_1, \ldots, X_q) > 0 \\ & \text{und } I(X_2, \ldots, X_q) > 0 \\ & \text{und } I(X_1, X_3, \ldots, X_q) > 0 \\ & \vdots \\ & \text{und } I(X_1, \ldots, X_{q-1}) > 0 \\ 0, & \text{sonst.} \end{cases} \tag{5.4}$$

definiert, wobei $I'(X_1, \ldots, X_q)$ durch

$$\begin{aligned} I'(X_1, \ldots, X_q) := \quad & H(X_1) + \ldots + H(X_q) \\ - \quad & H(X_1, X_2) - H(X_1, X_3) - \ldots - H(X_{q-1}, X_q) \\ + \quad & H(X_1, X_2, X_3) + H(X_1, X_2, X_4) \\ & + \ldots + H(X_{q-2}, X_{q-1}, X_q) \\ & \vdots \\ \pm \quad & H(X_1, \ldots, X_q) \end{aligned} \tag{5.5}$$

festgelegt ist. Hierbei ist $H(X_1, \ldots, X_q)$ die Entropie der diskreten Zufallsgrößen $X_1, \ldots, X_q$. Diese kann z.B. nach Shannon berechnet werden:

$$H(X_1, \ldots, X_q) := \sum_{i_1=1}^{N_1} \sum_{i_2=1}^{N_2} \ldots \sum_{i_q=1}^{N_q} p\left(x_1^{i_1}, x_2^{i_2}, \ldots, x_q^{i_q}\right) \cdot log_2 \frac{1}{p\left(x_1^{i_1}, x_2^{i_2}, \ldots, x_q^{i_q}\right)} \tag{5.6}$$

Dabei ist $p\left(x_1^{i_1}, x_2^{i_2}, \ldots, x_q^{i_q}\right)$ die Wahrscheinlichkeit, mit der ein Ausprägungstupel $(x_1^{i_1}, x_2^{i_2}, \ldots x_q^{i_q})$ auftritt. Die $x_k^{i_k}$ sind hierbei die i_k-ten Ausprägungen der Zufallsgrößen X_k. N_k bezeichnet die Anzahl unterschiedlicher Ausprägungen einer Zufallsgröße X_k.

Basierend auf dieser mathematischen Grundlage wurde eine Datenstruktur und ein Berechnungsalgorithmus für die Berechnung der gemeinsamen Informationsgehalte $I(X_1, \ldots, X_q)$ und der zugrunde liegenden gemeinsamen Entropien $H(X_1, \ldots, X_q)$ für Variablen entworfen. Die Variablen wurden dafür als diskrete Zufallsgrößen X_k interpretiert.

Ziel der Umsetzung war es, diese möglichst zeit- und speicherungseffektiv zu gestalten und bereits berechnete Größen wiederzuverwenden. Deswegen wurde u.a. eine iterative Implementation der Rekursion durchgeführt.

Es wurde desweiteren eine Relevanzschranke im Intervall (0..1) für die gemeinsamen Informationsgehalte eingeführt[10]. Will der Nutzer lediglich alle gemeinsamen Informationen über einem Wert bestimmen, legt er die Relevanzschranke entsprechend fest. Will er alle gemeinsamen Informationen bestimmen, wird die Schranke auf den Wert 0 gesetzt. Vorteil der Benutzung einer Relevanzschranke ist, daß die Berechnung wesentlich beschleunigt werden kann.

Um die redundante Speicherung der Variablen-Kombinationen für jeden gemeinsamen Informationsgehalt zu vermeiden, werden die gemeinsamen Informationen in einem Baum gespeichert. Mit der Position im Baum ist dort eindeutig die zugehörige Variablen-Kombination bestimmt (vgl. Abbildung 5.3).

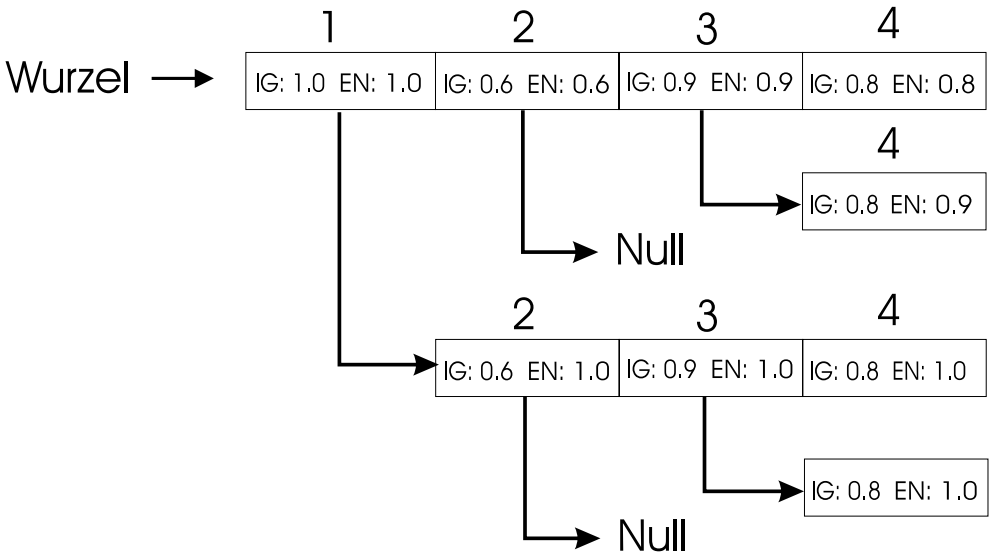

Abbildung 5.3: Beispiel eines Baumes zur Speicherung von Informationsgehalten und Entropien für vier Variablen mit einer Relevanzschranke von 0.7

Jeder Pfad in diesem Baum, der durch die schwarz gezeichneten Verweise der Vaterknoten auf die Sohnknoten definiert wird, legt eine Kombination von Variablen[11] fest.

Für die Berechnung der gemeinsamen Informationsgehalte werden zuerst die Informationsgehalte und Entropien in die Wurzel eingetragen[12]. Dann werden für jede Anzahl an Kombinationen (≥ 2) von Variablen iterativ die Kombinationen bestimmt[13] und neue Hierarchiestufen in den Baum eingetragen, falls die zugrunde

[10] Diese werden entsprechend auch auf das Intervall (0..1) normiert.

[11] ein ungeordnetes Tupel von Variablen

[12] Diese sind in der Wurzel natürlich gleich. Jedoch macht es keinen Sinn, für die Wurzel einen separaten Knotentyp zu definieren um diese relativ geringen Redundanzen zu vermeiden.

[13] Dies geschieht mit Hilfe eines Iterators, der über alle eingetragenen Kombinationen einer Länge iteriert.

liegenden Informationsgehalte der Teilmengen der Kombinationen alle die Relevanzschranke überschreiten. Ist dies nicht der Fall, braucht der neue Informationsgehalt nicht berechnet und eingetragen zu werden, weil er zwangsläufig kleiner gleich dem minimalen Teilinformationsgehalt ist. Entsprechende bereits berechnete Entropien und Informationsgehalte müssen nicht neu berechnet, sondern lediglich ausgelesen werden. Ist die Berechnung abgeschlossen, entsteht ein Baum, wie er in Abbildung 5.3 dargestellt ist.

5.3.2.5 Segmentierung des Beobachtungsraumes zur Bestimmung allgemeiner Beobachtungsraum-Eigenschaften und der Eigenschaften von Volumendaten

Zur Umsetzung von Segmentierungen des Beobachtungsraumes wurden eine Datenstruktur und ein Algorithmus entwickelt, der analog zu Quad- und Octrees allgemeine n-dimensionale Räume gleichmäßig unterteilt. Diese Verallgemeinerung ist sinnvoll, um z.B. auch höherdimensionale Räume mit abstrakten Dimensionen untersuchen zu können. Im speziellen werden damit zum einen allgemeine Eigenschaften des Beobachtungsraumes und zum zweiten spezielle Eigenschaften von Volumendaten bestimmt[14].

Die gleichmäßige Segmentierung soll lediglich als Arbeitsgrundlage für weitere Analysen des Raumes dienen. Ob die Anwendung bsw. bei einem nicht-strukturierten Gitter sinnvoll ist, muß sich in der praktischen Anwendung zeigen. Jedoch können mit diesem Ansatz verschiedene allgemeingültige Aussagen getroffen werden, die unabhängig von den speziellen Verbundeigenschaften sind. Ist für spezielle Analysefälle eine andere Segmentierungsart sinnvoller, kann diese in das Programm integriert werden.

Weiteres Anliegen bei der Implementation der Segmentierung des Beobachtungsraumes war die Integration von Dimensionen aus unterschiedlichen Tabellen[15]. Um eine komfortable Benutzung von Dimensionen sowohl für die Segmentierung als auch für weitere Datenklassen-Analysealgorithmen zur Verfügung zu haben, wurde eine Klasse definiert und implementiert, welche die unterschiedlichen Tabellen kapselt und eine einheitliche Schnittstelle für den Zugriff auf die Dimensionen zur Verfügung stellt.

Um die Segmentierung durchzuführen zu können, wurde eine Baumdatenstruktur entwickelt. In ihr repräsentieren die Knoten entsprechende „quaderförmige“ n-dimensionale Teilräume mit zwei n-dimensionalen Eckpunkten. Jeder Knoten enthält im Fall, daß der Teilraum weiter unterteilt wird, eine Liste von 2^n Knoten. Diese repräsentieren die ihn enthalten Teilräume. Desweiteren existiert für jedes Blatt des Baumes eine Liste der enthaltenen Beobachtungspunkte.

Unabhängig von der speziellen Segmentierung enthält jeder Teilraum Teilrau-

[14] Dies sind nur zwei Beispiele für die Anwendung von Segmentierungen in der Metadatengewinnung. So ist z.B. auch eine Segmentierung von Strömungsfeldern denkbar.

[15] Im speziellen sind das die Ausgangstabelle und die Tabelle, die zusätzliche räumliche Dimensionen enthält.

meigenschaften. Diese wurden mit Hilfe von Templates implementiert. Zum ersten enthalten die Teilraumeigenschaften eine Funktion zu ihrer Berechnung und eine Funktion, die bei speziellen Teilraumeigenschaften entscheidet, ob eine weitere Unterteilung erfolgen soll oder nicht. Desweiteren kann eine maximale Baumtiefe angegeben werden, um die Berechnungszeit und den Speicheraufwand zu steuern.

Bei der Segmentierung zur Bestimmung allgemeiner Beobachtungsraumeigenschaften werden für jeden Teilraumbereich die Minimal- und die Maximalwerte der quantitativen Merkmale[16], die Heterogenität in Abhängigkeit von den Merkmalsextrema, die Anzahl fehlender Werte sowie die Relevanz des Teilraumes bestimmt. Die Relevanz eines Raumbereiches ist umso größer, je größer die Dichte der Beobachtungspunkte, je größer die Heterogenität der Merkmale und je geringer die Anzahl fehlender Werte in diesem Raumbereich ist.

Für die Segmentierung von Volumendaten wurde exemplarisch das Monotonieverhalten der Volumendaten analysiert. In jedem Teilraum wird dabei festgestellt, ob sich die Werte entlang von achsenparallelen Geraden durch die Beobachtungspunkte des Gitters entweder monoton wachsend, monoton fallend, konstant oder nicht monoton verhalten.

Problematisch bei dieser Vorgehensweise war die Überlagerung des regelmäßigen Gitters mit der Segmentierung. Diese funktioniert bei strengem halbieren des Raumes nur dann optimal, wenn die Anzahl der Gitterpunkte in jeder Dimension durch $2^{\text{Segmentierungstiefe}}$ teilbar ist. Da dies in praktischen Anwendungen nicht vorausgesetzt werden kann, könnte als Alternative eine nicht-gleichmäßige Segmentierung durchgeführt werden, die jeweils nur solche Ebenen des Raumes teilt, in denen sich Gitterpunkte befinden.

5.3.2.6 Bestimmung kritischer Punkte

Bei der Analyse von Strömungsfeldern wurde exemplarisch die Bestimmung der Lage und der Klassifikation von kritischen Punkten[17] implementiert, weil diese für die visuelle Analyse von Strömungsfeldern von großer Bedeutung sind. Da insbesondere die Klassifikation kritischer Punkte für 3-dimensionale Strömungsfelder relativ komplex ist, wurde ihre Bestimmung beispielhaft für 2-dimensionale Strömungsfelder implementiert. Im folgenden soll das umgesetzte mathematische Verfahren kurz vorgestellt werden.

Voraussetzung für die Suche nach kritischen Punkten ist, daß die Beobachtungspunkte auf einem regelmäßigen Gitter vorliegen. Die kritischen Punkte können sich darin entweder auf den Gitterpunkten oder zwischen ihnen befinden. Der Algorithmus zur Findung kritischer Punkte durchläuft dann alle Zellen des Gitters[18] und überprüft, ob in ihr kritische Punkte enthalten sind. Unter der Voraussetzung einer einfachen bilinearen Form eines 2-dimensionalen Vektorfeldes ergeben sich die

[16] Als Beispiel für Verteilungseigenschaften

[17] Anm.: Kritische Punkte sind Punkte im Beobachtungsraum, an denen alle Komponenten des Strömungsvektors gleich 0 sind.

[18] Im 2-dimensionalen Fall ist eine Gitterzelle durch 4 benachbarte Punkte definiert. Diese sind dort in einem Rechteck angeordnet.

kritischen Punkten analytisch aufgrund des einfachen Gleichungssystems

$$\begin{aligned} \overrightarrow{v} &:= (1-x)\cdot(1-y)\cdot\overrightarrow{v_{00}} + (1-x)\cdot y\cdot\overrightarrow{v_{01}} \\ &\quad + x\cdot(1-y)\cdot\overrightarrow{v_{10}} + x\cdot y\cdot\overrightarrow{v_{11}} \\ &= \overrightarrow{0}. \end{aligned} \tag{5.7}$$

Hierbei sind $\overrightarrow{v_{00}}$, $\overrightarrow{v_{00}}$, $\overrightarrow{v_{00}}$ und $\overrightarrow{v_{00}}$ die 2-dimensionalen Eckvektoren der Gitterzelle, in deren Abhängigkeit das 2-dimensionale Geschwindigkeitsfeld $\overrightarrow{v}$ mit den Koordinaten $x \in [0,1]$ und $y \in [0,1]$ der Gitterzelle definiert ist. Sucht man die Stellen, an denen die Strömung gleich Null ist, ergeben sich zwei Gleichungen mit den zwei Unbekannten x und y. Es existieren jeweils zwei Lösungen dieses Gleichungssystems in x und y, die jeweils durch eine quadratische Gleichung errechnet werden können. Es sind dann alle die Punkte kritische Punkte, welche reelle Lösungen dieser Gleichung sind und sich innerhalb der Gitterzelle befinden. Die Koordinaten der Punkte werden dann analog zur bilinearen Interpolation der Geschwindigkeitsvektoren über die Eckpunkte und die erhaltenen Parameter für x und y bestimmt.

Um dann die Art der gewonnenen kritischen Punkte zu bestimmen, werden jeweils die partiellen Ableitungen der Geschwindigkeitsvektoren in den kritischen Punkten bestimmt. Damit ist die Matrix

$$M := \begin{pmatrix} v_x & u_x \\ v_y & u_y \end{pmatrix} \tag{5.8}$$

für jeden kritischen Punkt definiert, in der die Komponenten der beiden partiellen Ableitungen stehen. Nach Bestimmung ihrer beiden komplexen Eigenwerte kann anschließend die Art des kritischen Punktes bestimmt werden. Je nach Imaginär- und Realteil der beiden Eigenwerte ergibt sich, ob es sich um eine Senke, eine Quelle oder um eine Umströmung sowie zugehörig um einen Sattelpunkt, einen Wirbel, einen Knoten oder um einen Strudel handelt (vgl. z.B. [Frü97] S.45).

5.3.2.7 Bestimmung der Eigenschaften des Gradientenfeldes

Zur Bestimmung der Eigenschaften der Gradientenfeldes in Volumendaten werden diese exemplarisch in ein Strömungsfeld umgewandelt und dessen Eigenschaften analysiert. Die **Umwandlung** kann sowohl für normale 3-dimensionale Volumendaten, als auch für 2-dimensionale Bilddaten als auch allgemein für n-dimensionale Daten erfolgen. Einzige Beschränkung bei der **Analyse als Strömungsfeld** ist, daß dessen kritische Punkte nur für ein 2-dimensionales Gitter bestimmt werden können, und damit auch nur für 2-dimensionale Daten als ein Strömungsfeld analysiert werden können.

In der Umwandlung erfolgt zuerst die Schätzung der Anstiege in jedem Beobachtungspunkt. Für einen Punkt $P_i = (x_1^i, x_2^i, \ldots, x_n^i)^T$ wird für jede Komponente x_k der mittlere Anstieg Δx_k^i mit

$$\Delta x_k^i := \frac{1}{2}\cdot((x_k^i - x_k^{i-1}) + (x_k^{i+1} - x_k^i)) = x_k^{i+1} - x_k^{i-1} \tag{5.9}$$

bestimmt. Dabei sind x_k^{i-1} und x_k^{i+1} die Werte der Komponente x_k der beiden zu P_i in Richtung x_k benachbarten Punkte. Bei Randpunkten erfolgt die Schätzung des Anstieges lediglich über eine Differenz. Für Δx_k^i gilt dann entsprechend

$$\Delta x_k^0 := x_k^1 - x_k^0 \tag{5.10}$$

am unteren Rand bzw.

$$\Delta x_k^{max} := x_k^{max} - x_k^{max-1} \tag{5.11}$$

am oberen Rand.

Das berechnete Gradientenfeld wird anschließend in eine Datenmenge in Form einer Strömungsdatenmenge überführt und als solcher berechnet.

5.3.2.8 Bestimmung von Ausreißer-Datensätzen

Oft ist es in praktischen Datenmengen von Interesse, stark vom Durchschnitt abweichende Teildaten zu bestimmen und zu visualisieren. Beispiel hierfür sind Datensätze, die wesentlich von den anderen Datensätzen abweichen. Diese werden auch Ausreißerdatensätze genannt.

Bei der Berechnung von speziellen Eigenschaften von Multiparameterdaten im Programm „Metadatum“ werden solche Ausreißerdatensätze bestimmt. Dazu wird die Distanz- bzw. Ähnlichkeitsmatrix aller Datensätze bestimmt. Für diese Berechnung wird für jedes Datensatzpaar die Distanz- bzw. Ähnlichkeit der Merkmalsausprägungen berechnet und eingetragen. Die Wahl von Maßen zur Berechnung der Werte erfolgt in Beachtung der entsprechenden Skalentypen der Merkmale, in dem für gewisse Skalentypen passende Maße automatisch ausgewählt werden (vgl. [Noc99] und [Boc74]).

Mit Hilfe dieser Matrix werden für alle Datensätze die mittleren Distanzen bzw. Ähnlichkeiten zu allen anderen Datensätzen berechnet. Resultierend werden dann alle die Datensätze, deren mittlere Distanzen- bzw. Ähnlichkeiten über einen gewissen Grad hinaus abweichen, als Ausreißerdatensätze klassifiziert.

5.3.2.9 Bestimmung von Klassifikationen

Zur Analyse von Multiparameterdaten wurde die Durchführung einer Standardklassifikation der Datensätze integriert. Dabei wird der in [Noc99] umgesetzte Algorithmus unter Nutzung der dort verwendeten Ähnlichkeits- und Distanzmaße verwendet und mit Standardparametern angesteuert.

Durch Umwandlung der Distanz- bzw. Ähnlichkeitsstruktur der durch die Distanz- bzw. Ähnlichkeitsmatrix aus Abschnitt 5.3.2.8 festgelegten internen Struktur der Datensätze wird in diesem Algorithmus eine Klassifikation aufgebaut.

Ergebnis ist ein Klassifikationsbaum, in dem die Daten zu Klassen auf mehreren Hierarchiestufen angeordnet sind. An den Blättern dieses Baumes befinden sich die zu klassifizierenden Datensätze.

Über die Verteilungen der Datensätze in diesem Hierarchiebaum kann ein Überblick über die Strukturierung der Datensätze und über die Homogenität bzw. Heterogenität der Datensätze gewonnen werden.

5.3.3 Interaktion und Präsentation

In diesem Abschnitt wird die Interaktionsschnittstelle beschrieben. Wie bereits mehrfach vorgestellt, wurde die Interaktion von den anderen Modulen abgetrennt, um maximale Portabilität und Flexibilität zu erreichen. Zum Interaktionsmodul gehören neben der interaktiven Nutzereingabe von Metadaten die Präsentation des aktuellen Analysestatus und des aktuellen Standes der Metadaten-Information sowie die kontextsensitive Hilfe. Die folgenden Ausführungen werden überwiegend die Windows-Interaktion und Oberfläche zum Thema haben, da die Commandline-Interaktion und Präsentation äußerst gering ist.

Interaktion Wichtige Anforderung bei der Entwicklung und Implementation des Programmes war es, den Nutzer bei der Eingabe und Änderung von Metadaten zu unterstützen. Deswegen wurde eine dialogbasierte Interaktionsschnittstelle in Form eines „Windows-Propertysheets“ entwickelt (vgl. Abbildung 5.4).
Für jede Gruppe von Metadaten gibt es hier ein Register. Vorteil der Anordnung in Registern ist, daß der Nutzer eine Übersicht darüber hat, welche Metadatenerhebungen er bereits durchgeführt hat und welche ihn noch erwarten.

Je nachdem, welches Bestimmungsmodul und welche spezielle Metadatengewinnung gerade ausgeführt wird, kann er das aktuell zu erhebende Metadatum eingegeben oder die Standardbelegung verändern. In Abbildung 5.4 ist bsw. das Register zur interaktiven Unterstützung der allgemeinen Variablen- und Merkmalsmetadaten aufgeschlagen. Dort können der Skalentyp und die semantischen Informationen der Variablen der Datenmenge geändert werden. Weiterhin kann für ordinale Variable eine Reihenfolge der Variablenausprägungen festgelegt werden.

Ist die Metadateneingabe abgeschlossen und soll die aktuelle Metadatenbestimmung fortgesetzt werden, erfolgt das Betätigen des Schalters Ok. Soll die aktuelle Metadatengewinnung abgebrochen werden, wird der Schalter Abbrechen betätigt. Um eine Hilfe über zur Metadateneingabe zu erhalten, muß der Schalter Hilfe betätigt werden. Ist die Metadatengewinnung einer Gruppe von Metadaten[19] abgeschlossen, kann über die Schalter Zurück und Weiter entweder die letzte Metadatengewinnung erneut angestoßen oder die Metadatengewinnung für eine weitere Gruppe fortgesetzt werden.

Allgemein erfolgt die Ausführung einer interaktiven Metadatenfestlegung durch den Aufruf der Funktion „MDAendern“, die in einer der Funktionen der dem aktuellen Bestimmungsmodul zugehörigen C++-Klasse aufgerufen wird. Über entsprechende Parameter wird dort eine Interaktion in einem der Register des „Windows-Propertysheets“ aktiviert. Die Interaktionen werden mit Standard-Windows-Elementen wie „Buttons“, „List-Boxes“ u.a. durchgeführt.

[19] hier: Variablen-Metadaten

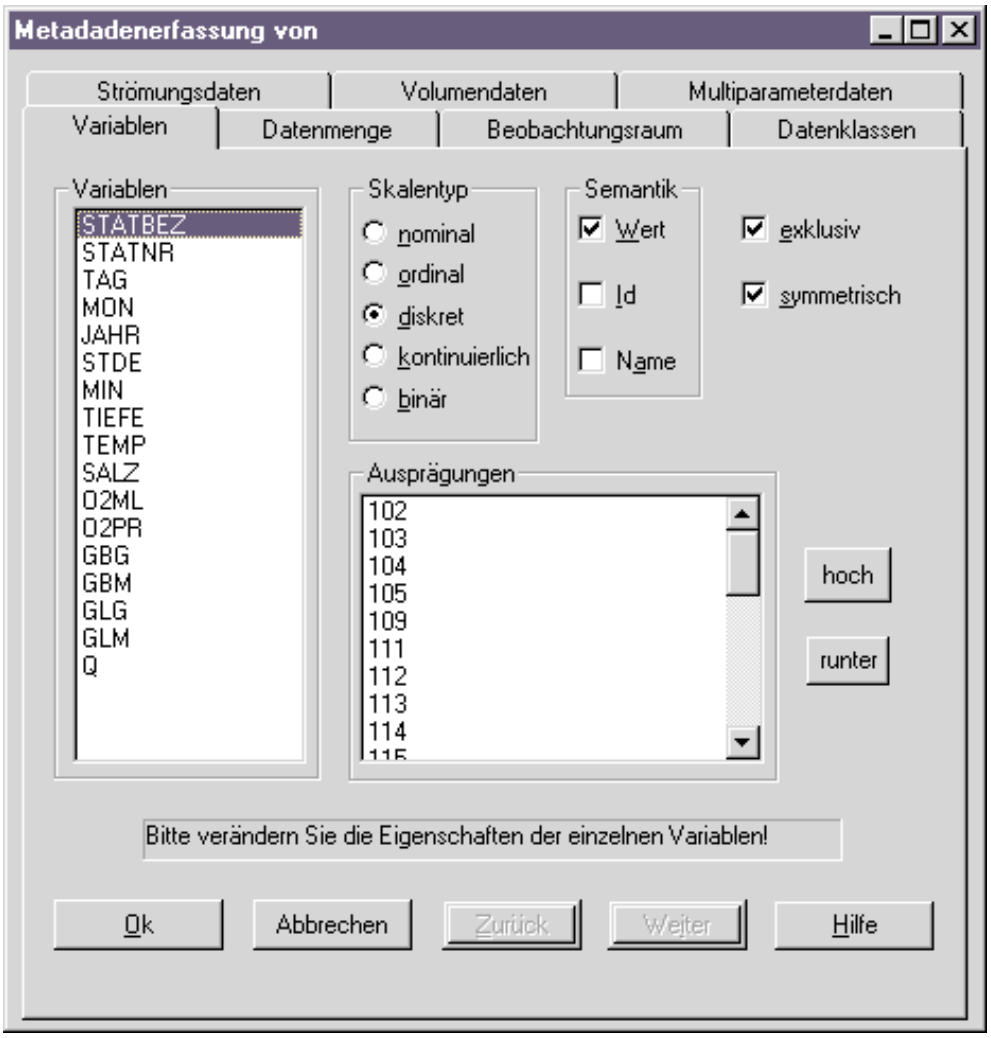

Abbildung 5.4: Dialog zur Ausführung von Interaktionen am Beispiel der Ostseedatenmenge (vgl. Abschnitt 6.3)

Im Falle des Commandline-Tool erfolgen die Nutzer-Eingaben über die Standardeingabe.

Status Weiterhin wurde eine Statusanzeige implementiert. Diese informiert den Nutzer, an welcher Stelle des Bestimmungsmoduls sich das Programm befindet und wie lange die Berechnungen noch dauern werden. Im Windows-Programm erfolgt die Anzeige eines Statusfensters, welches einen Statusbalken und mehrere Anzeigebereiche enthält. Abbildung 5.5 zeigt den Statusdialog in bei der Berechnung der gemeinsamen Informationsgehalte.

Im Commandline-Programm erfolgt lediglich eine Ausgabe des Berechnungsstandes auf die Standardausgabe.

Aktueller Stand der Information Der aktuelle Stand der Information ermöglicht es dem Nutzer, einen Überblick über die bisher gewonnenen Metadaten zu erlangen. So werden in einer dem Metadatendokument zugehörigen Ansicht die

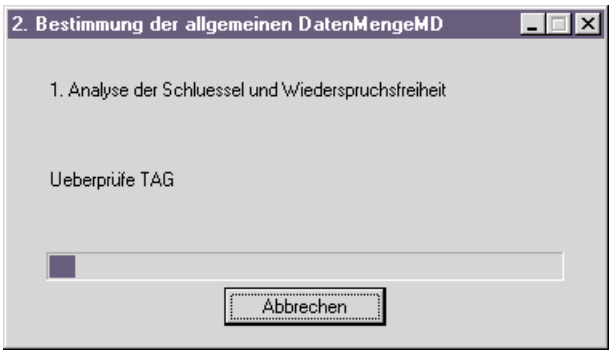

Abbildung 5.5: Statusdialog während der Berechnung der gemeinsamen Informationsgehalte

wichtigsten Metadaten visuell dargestellt. Abbildung 5.6 zeigt einen Ausschnitt aus dem Ansichtsfenster der Variablen-Metadaten.

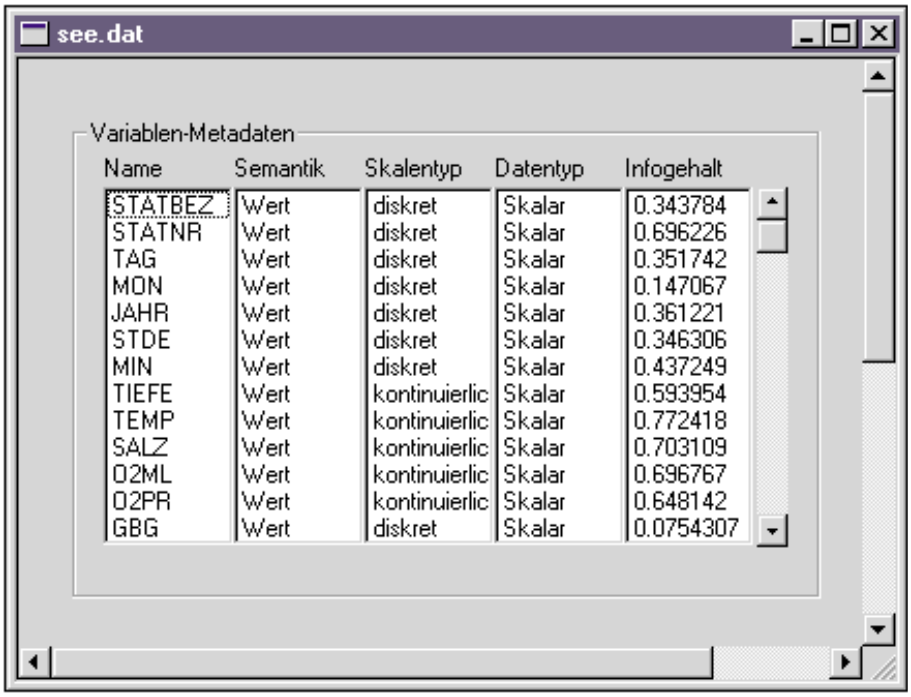

Abbildung 5.6: Ausschnitt des Ansichtsfensters der Variablen-Metadaten

Kontextsensitive Hilfe Da sich der Nutzer im internen Programmablauf oder in der Begriffswelt der Visualisierung nicht notwendigerweise auskennt, muß das Programm ihn beim Verständnis von Begriffen, Darstellungen und Interaktionen unterstützen. Deswegen wurden zum einen mit Hilfe des Windows-Hilfesystems eine kontextsensitive Hilfe eingefügt. So kann der Benutzer auf jedem Interaktionsregister und in der Metadatenansicht passende inhaltliche Hilfe bekommen. Zum anderen wurde eine Direkthilfe eingebaut, die aktiviert wird, wenn der Nutzer mit der Maus eine gewisse Zeit über einem Interaktions- oder Ansichtsobjekt verweilt.

5.4 Leistungsfähigkeit und Grenzen

Zusammenfassend kann man sagen, daß es gelungen ist, ein leistungsfähiges und den Nutzer unterstützendes Programm zu entwickeln, welches viele der entwickelten Konzepte umsetzt. Die Eigenschaften und Grenzen dieses Programmes sollen in diesem Abschnitt zusammengefaßt werden.

Durch die strenge Trennung von Interaktion, Steuerung und Bestimmung war es möglich, ein flexibles Programm umzusetzen, welches sowohl unter einer Windows-Oberfläche als auch als Commandline-Tool übersetzbar und damit in verschiedenen Anwendungsszenarien verwendbar ist.

Das Programm „Metadatum“ integriert dabei eine Vielzahl von Metadatengewinnungsverfahren, um Datenmengen in Tabellenform zu analysieren. Dafür wurde neben der Integration einer allgemeinen Datei-Schnittstelle auch die Integration einer Datenbank-Schnittstelle[20] untersucht. Weiterhin ist es gelungen, ein Metadatenformat zu entwerfen, das es erlaubt, komplexe Metadaten zu speichern. Die Analyse von großen Datenmengen, welche auf einem regelmäßig strukturieren Gitter gegeben sind, ist jedoch für Daten mit z.B. in der Volumenvisualiserung üblichen Größenordungen im Tabellenformat nicht praktikabel. Deswegen ist eine Erweiterung um entsprechende Datenformate angedacht.

Ein weiteres wichtiges Ergebnis ist die Umsetzung des Einlesens von Raumbezugstabellen, welche räumliche Dimensionen zur Reduktion von Redundanzen enthalten. Dabei kann in der bisherigen Implementation genau eine solche Tabelle eingelesen werden. Das Einlesen weiterer Tabellen wurde in der Datenstruktur offengehalten.

Wichtige Analyseverfahren wie die Segmentierungen des Raumes oder die Bestimmung gemeinsamer Informationsgehalte wurden unabhängig von der speziellen Anzahl der Variablen bzw. Dimensionen mit einer hohen Berechnungsgeschwindigkeit umgesetzt. Andere Analyseverfahren wurden jedoch nur beispielhaft implementiert. Dazu gehören die Bestimmung von kritischen Punkten (bisher nur für 2-dimensionale Strömungsfelder) und die Bestimmung von Schlüsseln (maximale Schlüssellänge ist drei).

Auch das Konzept zur Ausweisung von Variablenhierarchien wurde mit einer rein interaktiven Bestimmung nur exemplarisch umgesetzt. Außerdem werden die Variablenhierarchien im weiteren Metadatengewinnungsprozeß bisher nicht gesondert beachtet. Denkbar für die Bestimmung von Variablenhierarchien ist neben einer Data-Dictionary-Analyse auch die Durchführung einer Faktoren-Analyse. Für den weiteren Bestimmungsprozeß ist z.B. die Zusammenfassung mehrerer in der Hierarchie angeordneter Variable zu nur einer einzelnen Variable sinnvoll. Beispiel hierfür wäre der bereits genannte Fall des Auftretens der drei Dimensionen Jahr, Monat und Tag. Diese könnten z.B. zu einer Variablen Zeit zusammengefaßt und ihre Wertebereiche entsprechend aneinander angepaßt werden.

[20] ODBC-Schnittstelle

Die Metadatengewinnungen von datenklassenspezifischen Metadaten wurden vor allem unter dem Blickwinkel der Visualisierungsentscheidung ausgewählt. Erweitert man den Kontext um die Bestimmung von Visualisierungsparametern, eröffnet sich hier ein breites Feld zur Definition und Umsetzung von Metadaten.

Kapitel 6

Fallbeispiele

In diesem Abschnitt wird anhand von drei Beispielen die Nutzung des Werkzeugs „Metadatum“ vorgestellt. Es wurden Beispiele aus den drei Datenklassen Strömungsdaten, Volumendaten und Multiparameterdaten ausgewählt, um die Metadatenerhebung unter unterschiedlichen Gesichtspunkten vorstellen und bewerten zu können. Bewußt wurden Datenmengen aus der Anwendung in den üblichen Größenordnungen gewählt, um das Zeitverhalten und die Machbarkeit der Algorithmen praktisch zu testen.

6.1 Eine Strömungsdatenmenge

6.1.1 Vorstellung der Datenmenge

Als Beispiel einer Strömungsdatenmenge soll hier für das stationäre, elektrostatische 2D-Vektorfeld eines Wassermolekülseine Metadatengewinnung durchgeführt werden (vgl. Abbildung 6.1).

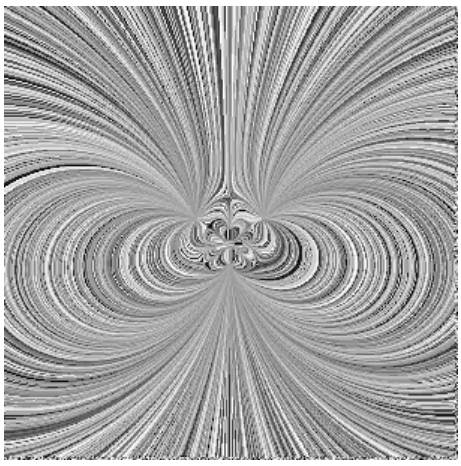

Abbildung 6.1: Integrate and Draw-Darstellung des elektrostatischen Feldes eines Wassermoleküls; Erhoben vom Konrad-Zuse-Institut

Diese Datenmenge enthält die 4 Variablen x, y, vx und vy und hat eine Größe von 102×102 Datensätzen[1].

Abbildung 6.1 läßt vermuten, daß sich die interessanten Regionen und vor allem die kritischen Punkte in der Mitte des Vektorfeldes befinden. Um Aufschluß über die genaue Beschaffenheit dieser Region zu erhalten, wurde eine Metadatenanalyse durchgeführt[2].

6.1.2 Metadatenbestimmung

1. Bei der **Bestimmung der allgemeinen Variablen- und Merkmals-Metadaten** wurden für die vier Variable x, y, vx, und vy zugehörige Variablen-Eigenschaften bestimmt. Als Skalentyp wurde der kontinuierliche für vx und vy und der diskrete für die Variablen x und y festgelegt. Desweiteren wurden der Informationsgehalt der einzelnen Variablen und die Verteilungen der Werte bestimmt. Auffällig hierbei ist die Gleichverteilung der Variablen x und y, die bereits an dieser Stelle das Vorhandensein eines regelmäßigen Gitters vermuten läßt. Fehlende Werte liegen nicht vor.

2. In der anschließenden **Bestimmung der allgemeinen Datenmenge-Metadaten** wurden sowohl x und y als auch vx und vy als Schlüsselpaar erkannt. Da die Standardbelegung für die Trennung der unabhängigen und der abhängigen Variablen die Variablen des ersten minimalen Schlüssels als unabhängig auswählt, wurden x und y als unabhängig, vx und vy aber als abhängig klassifiziert. Dies ist hier natürlich sinnvoll, so daß keine interaktive Korrektur erforderlich ist. Aus der Analyse der gemeinsamen Informationsgehalte und der Korrelationen geht hervor, daß x und y keine gemeinsamen Informationen tragen und unkorreliert sind. Im Gegensatz dazu konnte für vx festgestellt werden, daß es mit vy stark negativ korreliert ist (Korrelationskoeffizient = -0.923839). Dies bedeutet, daß für hohe Werte von vx an einem Beobachtungspunkt meist geringe Werte von vy vorliegen und umgekehrt. Weiterhin wird die Stärke der Beziehung dieser beiden Merkmale durch einen hohen gemeinsamen Informationsgehalt von 0.957806 untermauert. Zwischen den Merkmalen x bzw. y und den Merkmalen vx bzw. vy liegen keine signifikanten gemeinsamen Informationen vor.

   Auch Hierarchien von Variablen liegen nicht vor.

3. Bei der **Bestimmung der Beobachtungsraum-Metadaten** wurden x und y als Raumkoordinaten eines 2D-Raumes festgelegt. Ein Einlesen des Raumbezuges aus einer separaten Tabelle war nicht notwendig. Die Analyse der

[1] Da das Einlesen der Daten aus einer Tabelle erfolgt, ist eine Angabe der Anzahl der Datensätze hier sinnvoll. Vor allem dient sie dem Vergleich der Größe der drei hier vorgestellten Datenmengen.

[2] In diesem Fall wurde die Visualisierung vorangestellt, um dem Leser eine Vorstellung der Datenmenge zu vermitteln. Praktisch erfolgt die Metadatengewinnung im allgemeinen jedoch vor einer speziellen Visualisierung, um die Wahl einer geeigneten Technik zu unterstützen. So ist es bei der vorgestellten Datenmenge z.B. nicht a priori klar, daß es sich um ein Vektorfeld handelt und die Methode Integrate and Draw anwendbar ist.

Segmentierung zeigte keine Konzentration von Bereichen von Interesse. Sie sind relativ gleichmäßig in den Dimensionen verteilt.

Dies hat den Grund, daß in das Maß zur Bestimmung der Relevanz von Teilräumen nicht im speziellen Werte in der Nähe des Nullpunktes eingehen. Dort wird lediglich eine starke Heterogenität der Merkmalsausprägungen in einem Raumbereich zugrundegelegt. Eine hohe Heterogenität ist in dieser Datenmenge jedoch an keiner Stelle des Beobachtungsraumes gesondert ausgeprägt, so daß die Relevanz der Teilräume in etwa überall gleich ist. Die extrahierten Bereiche von Interesse weichen in ihrer Relevanz nur geringfügig von den anderen Bereichen ab.

Desweiteren wurde als Verbund aufgrund der Verteilungseigenschaften der Dimensionen x und y ein vollständiges regelmäßiges Gitter erkannt. Unter Annahme eines lokalen oder globalen Wirkungskreises erfolgte dann die Entscheidung, eine Erhebung von Strömungs-Metadaten durchzuführen. Andere Datenklassen-Metadatenerhebungen wurden als nicht den Daten entsprechend eingestuft.

4. In der **Bestimmung der Strömungsdaten-Metadaten** wurden als erstes die beiden Merkmale vx und vy als Vektor zusammengefaßt. Um die Charakteristik des Strömungsfeldes zu bestimmen, erfolgte im Anschluß eine Analyse der kritischen Punkte. Als Ergebnis dieser Analyse ergab sich eine Anzahl von 12 kritischen Punkten. Lage und Art seien an 4 Beispielen dargestellt:

   1. Kritischer Punkt:
      - Koordinaten: $(42.5881, 47.4624)$
      - Eigenwerte: $(79.6174, 0), (29.375, 0)$
      - Art: Knoten, Quelle
   2. Kritischer Punkt:
      - Koordinaten: $(44.3501, 56.8317)$
      - Eigenwerte: $(0.927545, 0), (-0.99645, 0)$
      - Art: Sattel, Umströmung
   3. Kritischer Punkt:
      - Koordinaten: $(46.8823, 53.9142)$
      - Eigenwerte: $(-64.5354, 0), (-5736.71, 0)$
      - Art: Knoten, Senke
   4. Kritischer Punkt:
      - Koordinaten: $(57.3989, 47.0373)$
      - Eigenwerte: $(16.81, 6.8158), (16.81, -6.8158)$
      - Art: Strudel, Quelle

   ⋮

Alle kritischen Punkte befinden sich in der Mitte des Definitionsbereiches in einem rechteckigen Bereich mit den Eckpunkten $(42.5881, 42.304)$ und $(57.3989, 57.2992)$.

6.1.3 Interpretation

Die erhobenen Metadaten für das statische Vektorfeld eines Wassermoleküls geben einen umfassenderen Einblick in die Datenmenge, als es eine reine Visualisierung (vgl. Abb. 6.1) vermag. Die anhand des Bildes getroffene Feststellung, daß sich die interessanten Charakteristika in der Mitte befinden, hat sich bestätigt. Zusätzlich konnten jedoch durch Ausführung der Metadatengewinnung auch quantitative Aussagen über die Segmentierung des Raumes und über die Lage und Art der kritischen Punkte abgeleitet werden.

Aus den erhobenen Metadaten können die folgenden Schlüsse im Hinblick auf eine anschließende Visualisierung gezogen werden:

1. Es ist sinnvoll, die inneren Bereiche in einer Visualisierung zu vergrößern.
2. In den äußeren Bereichen könnte die Darstellung in einer geringeren Auflösung erfolgen, falls dies erforderlich ist.
3. Bei der vorliegenden großen Anzahl an kritischen Punkten sollten diese in einer Visualisierung geeignet erkennbar sein. Sinnvoll wäre hier z.B., eine Ikonendarstellung[3] mit einer Stromliniendarstellung zu kombinieren, um durch mehrere Ansichtsarten das Verständnis der Topologie des Vektorfeldes zu verbessern.

Weiterhin ist diese Datenmenge ein Beispiel für eine fast vollständig automatische Gewinnung der Metadaten. Bei Auswahl eines geringen Interaktionsgrades bedarf es lediglich einer einzigen Unterstützung durch den Nutzer, da für alle anderen Metadaten-Bestimmungen Standarbelegungen und -analysen ausreichend sind. Der Nutzer muß lediglich die Art des Raumes festlegen, so daß die beiden automatisch bestimmten Dimensionen x und y als Raumkoordinaten interpretiert werden.

6.2 Eine Volumendatenmenge

6.2.1 Vorstellung der Datenmenge

Volumendaten sind definitionsgemäß auf einem 3-dimensionalen, regelmäßigen Gitter gegeben, in dem an jedem Gitterpunkt mindestens eine Merkmalsausprägung vorliegt. Die Metadatengewinnung für Volumendaten wurde jedoch so konzipiert und umgesetzt, daß sie auch für höher- und niederdimensionale Räume durchgeführt werden kann. Da die Ergebnisse der Metadatengewinnung im 2D-Raum am besten

[3] Explizite Darstellung der Art der kritischen Punkte an den zugehörigen Positionen im Vektorfeld

veranschaulicht werden können, soll hier als Beispiel einer Volumendatenmenge für eine 2-dimensionale Bilddatenmenge eine Metadatengewinnung durchgeführt werden. Im allgemeinen ist natürlich mit dem Programm „Metadatum“ auch eine Analyse von 3-dimensionalen Volumendaten möglich.

Im speziellen handelt es sich bei der Datenmenge um den Schnitt durch den Zellkern einer Maus (vgl. Abbildung 6.2), welche die 3 Variablen x, y und Value enthält.

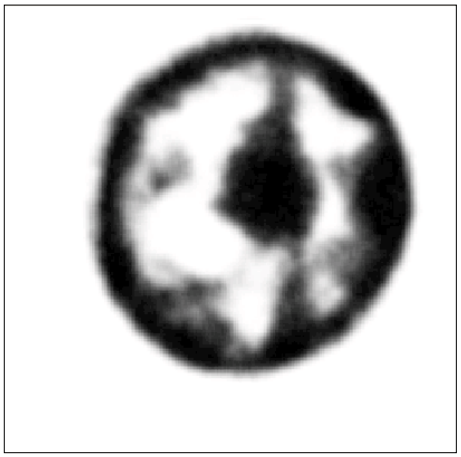

Abbildung 6.2: Schnitt durch den Zellkern einer Maus - dunkle Bereiche repräsentieren hohe Werte und helle Bereiche niedrige Werte[4]

Sie enthält 258×255 Datensätze.

Man sieht bereits an dieser einfachen Darstellung, daß für eine große Fläche mit weißer Farbe (Wert 0) keine relevante Information vorliegt. Desweiteren befindet sich der Zellkern nicht in der Mitte der räumlichen Koordinaten. Was in dieser Darstellung schlecht sichtbar wird, ist das genaue Verhalten der Daten in Bereichen mit einem Übergang von hohen zu niedrigen Werten und umgekehrt. Die Kontraste zwischen geringen und hohen Werten sind hier nur schwer zu erkennen. Aufschluß über diese Bereiche und eine resultierende Unterstützung der Visualisierung soll eine Metadatengewinnung geben.

6.2.2 Metadatenbestimmung

1. Bei der **Bestimmung der allgemeinen Variablen- und Merkmals-Metadaten** wurden für die drei Variablen x, y und Value zugehörige Variablen-Eigenschaften bestimmt. Für alle Variablen wurde der diskrete Skalentyp festgelegt. Desweiteren wurden der Informationsgehalt der einzelnen Variablen und die Verteilungen der Werte bestimmt. Auch in dieser Datenmenge deutet die Gleichverteilung der Variablen x und y auf das Vorhandensein eines regelmäßigen Gitters hin. Um bsw. eine Kontrastverbesserung des Bildes vor-

[4]Der Koordinatenursprung mit $x = 0$ und $y = 0$ befindet sich links unten.

nehmen zu können, ist insbesondere das Histogramm der Variablen Value von Interesse (vgl. Abb. 6.3).

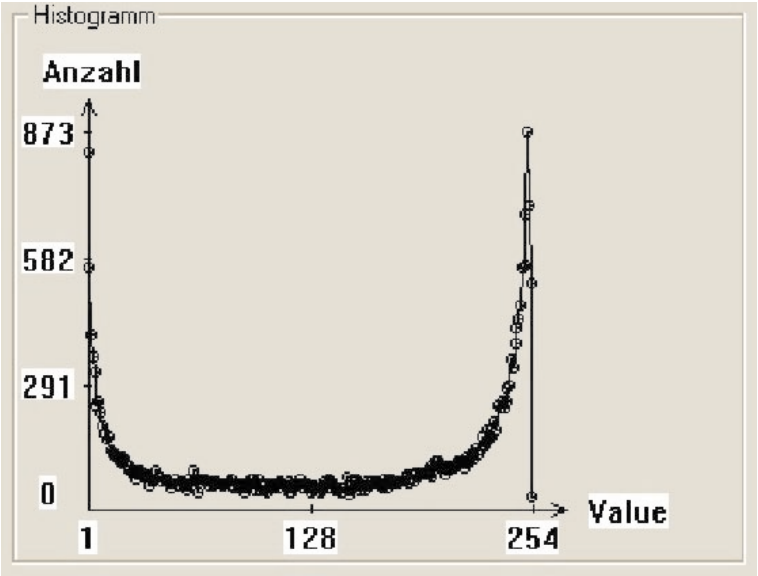

Abbildung 6.3: Histogramm der Variablen Value aus dem Programm „Metadatum“; der Wert 0 wurde weggelassen, da dessen große Häufigkeit die Darstellung zu stark verzerren würde

Auffällig bei dieser Darstellung ist, daß vor allem in den Bereichen nahe den Extrema 0 und 254 die Häufigkeit stark zunimmt. Um diese Bereiche besser sichtbar zu machen, könnte dort ein größerer Kontrast in einer Visualisierung gewählt werden.

2. In der anschließenden **Bestimmung der allgemeinen Datenmenge-Metadaten** wurden x und y als Schlüsselpaar erkannt. Daraus schlußfolgernd wurden x und y als unabhängig, Value aber als abhängig klassifiziert. Aus der Analyse der gemeinsamen Informationsgehalte und der Korrelationen geht hervor, daß x und y keine gemeinsamen Informationen tragen und unkorreliert (Korrelationskoeffizient von 0) sind. Im Gegensatz dazu konnte für die Variable Value festgestellt werden, daß sie leicht mit x und y positiv korreliert ist. Dies bedeutet, daß hohe Werte der Variablen Value eher bei hohen Werten von x und y vorliegen. Dies untermauert die über die Abbildung 6.2 getroffene Aussage, daß das Innere des Zellkerns (dunkle Bereiche) leicht nach rechts oben verschoben ist.

3. Bei der **Bestimmung der Beobachtungsraum-Metadaten** wurden x und y als Raumkoordinaten eines 2D-Raumes festgelegt. Ein Einlesen des Raumbezuges aus einer separaten Tabelle war nicht notwendig. Die Analyse der Segmentierung zeigte, daß Regionen von besonderem Interesse sich in den Übergangsbereichen von hohen zu niedrigen Werten des Merkmals Value befinden (vgl. Abbildung 6.4).

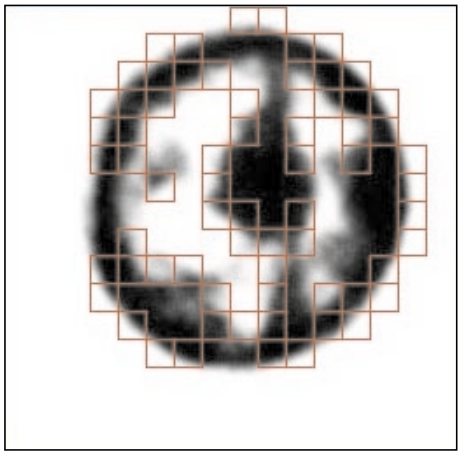

Abbildung 6.4: Bereiche von Interesse in der Datenmenge „Zellkern einer Maus“; Die Quadrate kennzeichnen die Bereiche von Interesse

Desweiteren wurde auch hier aufgrund der Verteilungseigenschaften der Dimensionen x und y ein vollständiges regelmäßiges Gitter erkannt. Unter Annahme eines lokalen oder globalen Wirkungskreises und des Vorliegens nur eines Merkmals erfolgte dann die Entscheidung, eine Erhebung von Volumen-Metadaten durchzuführen. Andere Datenklassen-Metadatenerhebungen wurden als nicht den Daten entsprechend eingestuft.

4. In der **Bestimmung der Volumendaten-Metadaten** wurde als erstes eine Segmentierung des 2D-Raumes zur Analyse der volumendatenspezifischen Teilraumeigenschaften durchgeführt. Im speziellen wurde dabei die Monotonie der Variablen Value in den Dimensionen x und y untersucht. Dabei stellte sich heraus, daß die äußeren Bereiche konstant sind. In Bereichen mit hohem Wert des Merkmals Value verhält sich die Funktion überwiegend leicht monoton wachsend oder fallend. Im Gegensatz dazu ist diese in den Übergangsbereichen entweder stark monoton fallend oder steigend. Dadurch wurden auch hier die Übergangsbereiche als relevant ausgewiesen.

   Die Analyse des Gradientenfeldes[5] durch Bestimmung kritischer Punkte führte in dieser Datenmenge zu keinen aussagekräftigen Ergebnissen, da in den Gebieten mit konstanten Werten (z.B. $= 0$) die Gradienten in den Beobachtungspunkten gleich 0 werden. Entsprechend ist in diesen Gebieten jeder Punkt ein kritischer Punkt. Sinnvoll wäre die Analyse auf kritische Punkte also nur in nicht-konstanten Gitterzellen. Diese Einschränkung wurde jedoch bisher noch nicht umgesetzt.

[5] Anmerkung: Natürlich weisen auch die Monotonien in den einzelnen Raumbereichen Eigenschaften des Gradientenfeldes aus.

6.2.3 Interpretation

Die gewonnen Metadaten für den Datensatz des Zellkerns einer Maus bestätigen Vermutungen, die Anhand der Abb. 6.1 aufgestellt wurden. So konnten die interessanten Bereiche herausgefiltert und quantitative Aussagen über sie abgeleitet werden.

Aus den erhobenen Metadaten können die folgenden Schlüsse im Hinblick auf eine anschließende Visualisierung gezogen werden:

1. Es ist sinnvoll, die Gebiete im Übergangsbereich von hohen zu niedrigen Werten des Merkmals Value speziell zu vergrößern.

2. Die äußeren Bereiche, in denen der Wert des Merkmals Value 0 ist, können in einer Vorverarbeitung reduziert werden, um die Komplexität der Datenmenge zu verringern.

3. Durch eine geeignete Transformation der Farbskala anhand des Histogramms sollte der Kontrast erhöht werden.

4. Möglicherweise ist eine separate Darstellung der Übergangsbereiche sinnvoll.

Wie bei der Datenmenge des elektrostatischen Vektorfeldes eines Wassermoleküls kann auch hier die Analyse weitgehend automatisch erfolgen.

6.3 Eine Multiparameterdatenmenge

6.3.1 Vorstellung der Datenmenge

Als drittes Beispiel soll hier eine im Visualisierungsumfeld üblicherweise zur Klasse der Multiparameterdaten gehörige Datenmenge vorgestellt werden. Im speziellen handelt es sich hierbei um Messungen an verschiedenen Meßpunkten in der Ostsee. An jedem Meßpunkt werden dabei 12 Merkmale erhoben. Dazu gehören z.B. Temperatur, Salz- und Sauerstoffgehalt. Abbildung 6.5 zeigt eine mögliche Visualisierung dieser Datenmenge.

Ausgewählt wurde diese Datenmenge u.a., um an ihr das Einlesen der Raumdimensionen aus einer separaten Tabelle vorzuführen. Die Variable STATNR in der Haupttabelle verweist auf eine Tabelle, in der die Beobachtungspunkte[6] separat aufgeführt werden. Diese Tabelle enthält die Variablen STATNR, gradbrei, minbrei, gradlaen, minlaen und andere Eigenschaften der Stationen. Hierbei legen gradbrei und minbrei die geographische Breite und gradlaen und minlaen die geographische Länge der Stationen fest. In der Haupttabelle befinden sich die Dimension STDE[7], die Zeitdimensionen Jahr, Monat und Tag, welche die Meßzeit festlegen, und 12 Merkmale, welche die gemessenen Daten repräsentieren.

[6] Die Beobachtungspunkte entsprechen den Meßstationen.

[7] bezeichnet die Meßtiefe

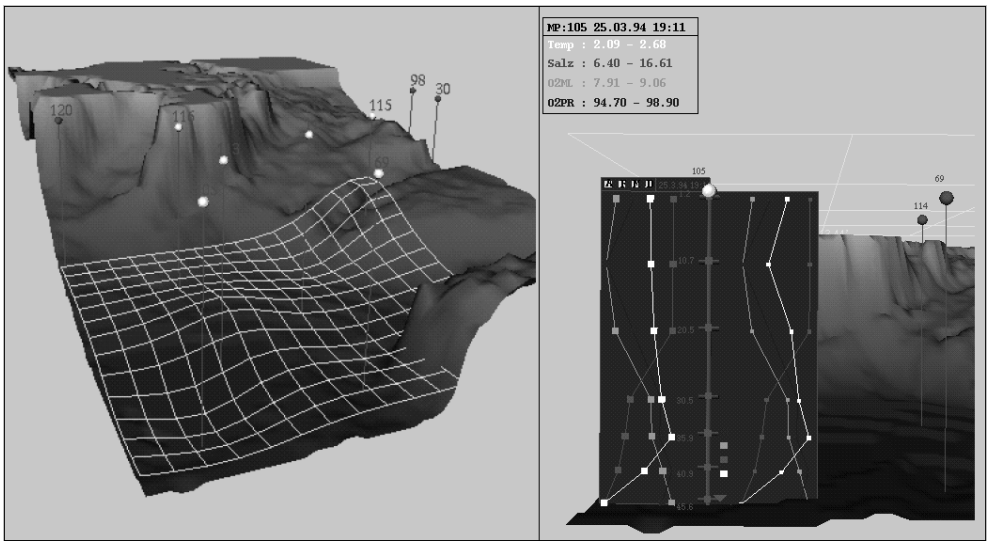

Abbildung 6.5: Darstellung des Ostsee-Datenmenge aus [Kre96]

Insgesamt umfaßt diese Datenmenge zwei Tabellen mit einer Anzahl von 9097 Datensätzen und 17 Variablen in der Haupttabelle (vgl. Anhang B.1) und 48 Datensätzen und 11 Variablen in der Tabelle der Beobachtungspunkte (vgl. Anhang B.2). Bei der Komplexität dieses Datenmenge ist die Interpretierbarkeit einer einzigen Darstellung relativ gering. Deswegen soll im folgenden eine Metadatengewinnung durchgeführt werden.

6.3.2 Metadatenbestimmung

1. Bei der **Bestimmung der allgemeinen Variablen- und Merkmals-Metadaten** für die Haupttabelle wurden für die 17 Variablen zugehörige Variablen-Eigenschaften bestimmt. Alle Variablen konnten als diskret oder als kontinuierlich klassifiziert werden. Nicht-exklusive Variable liegen nicht vor. Desweiteren wurden der Informationsgehalt der einzelnen Variablen und die Verteilungen der Werte bestimmt. Interessant hierbei sind die Variablen BGB und Q. Diese besitzen lediglich zwei Ausprägungen und haben resultierend die geringsten Informationsgehalte von 0.0754307 und 0.00290142. Die höchsten Informationsgehalte haben die beiden Variablen TEMP (0.772418) und SALZ(0.703109). Diese besitzen mit 1438 und 1274 auch die größte Kardinalität der Wertebereiche aller Variablen. Interessant bei der Analyse der Wertebereiche ist, daß bei verschiedenen Merkmalen Werte wie 99 oder 999 häufig auftreten, die außerhalb der normalen Werte liegen. Diese sind Fehlwerte und wurden als solche klassifiziert.

2. In der anschließenden **Bestimmung der allgemeinen Datenmenge-Metadaten** konnte keine Kombination aus Variablenkombinationen als Schlüssel identifiziert werden. Dies hat vor allem den Grund, daß mit den drei Variablen Jahr, Monat und Tag sowie STATNR und STDE die reale minimale

Schlüssellänge gleich fünf ist. Da der eingesetzte Algorithmus nur Schlüssel von 3 Schlüsselvariablen erkennen kann, muß die Trennung der abhängigen und unabhängigen Variablen in diesem Falle interaktiv vorgenommen werden. Entsprechend wurden STATNR, STDE, Jahr, Monat und Tag als unabhängig, die anderen Variablen aber als abhängig klassifiziert.

Interessantes Ergebnis der Analyse der gemeinsamen Informationsgehalte ist, daß bei einer Relevanzschranke von 0.8 keine relevanten gemeinsamen Informationen festgestellt werden konnten. Dies liegt vor allem daran, daß die gemeinsamen Informationsgehalte vor allem für die Analyse von qualitativen Variablen geeignet sind. Bei quantitativen Variablen ohne Ausführung von Gruppierungen werden die Entropien und damit die gemeinsamen Informationsgehalte i.a. zu gering bestimmt. Das führt in diesem Fall dazu, daß die maximale Entropie aller Variable mit dem Wert 0.772418 in der Variablen TEMP bereits unter der Relevanzschranke von 0.8 liegt. Wollte man auch gemeinsame Informationsgehalte unter diesem Wert bestimmen, müßte die Relevanzschranke entsprechend herabgesetzt werden.

Im Gegensatz zur Analyse der gemeinsamen Informationsgehalte brachte eine Korrelationsanalyse der Daten signifikante Ergebnisse. Bei der Untersuchung aller Paare von Variablen stellte sich zwar heraus, daß die meisten einen Wert nahe 0 besitzen und damit relativ unkorreliert sind. Jedoch fanden sich auch Paare mit hoher gemeinsamer Korrelation. Besonders hoch korreliert sind die Variablenkombinationen

- GBG und GBM (-0.888622),
- SALZ und O2PR (0.717924) sowie
- STATBEZ und GLG (0.964653).

Interessant hierbei ist, daß für die Kombinationen GBG und GBM ein umgekehrt proportionaler Zusammenhang erkennbar ist.

Aufgrund dieser Ergebnisse wurden bei der Festlegung redundanter Merkmale das Merkmal GLG als wegen seiner hohen Korrelation zu STATBEZ als redundant festgelegt.

Besonders wichtig für eine sinnvolle Interpretation dieser Datenmenge ist die Festlegung der Abhängigkeiten der Zeitvariablen. Um zu vermeiden, daß in späteren Analysen und in der Visualisierung drei Zeitachsen vorliegen, wurden in der Variablenhierarchisierung die Variable Monat der Variable Jahr und die Variable Tag der Variablen Monat untergeordnet.

3. Bei der **Bestimmung der Beobachtungsraum-Metadaten** erfolgte nach der Festlegung eines konkreten Raumes das Einlesen des Raumbezuges aus der separaten Tabelle der Stationen. Für diese Tabelle erfolgte eine Bestimmung der allgemeinen Variablen- und der Merkmals-Metadaten sowie der allgemeinen Datenmenge-Metadaten. Entsprechend wurden die Skalentypen der Variablen bestimmt und deren Informationsgehalte berechnet. Speziell war es

hierbei wichtig, die Abhängigkeiten der Raumdimensionen gradlaen und minLaen sowie gradbrei und minbrei[8] in der Variablenhierarchie zu beachten.

Im Anschluß erfolgte die unterstütze Festlegung des Id-Paares STATNR aus der Haupt- und aus der Tabelle der Beobachtungspunkte, um den Zusammenhang der beiden Tabellen festzulegen. Dann wurden die Variablen STDE, Jahr, Monat und Tag aus der Haupttabelle sowie gradlaen, minLaen, gradbrei und minbrei aus der Tabelle der Beobachtungspunkte als Raum- bzw. als Zeitdimensionen festgelegt.

Die Analyse der Segmentierung zeigte, daß vor allem die Bereiche zwischen 54° und 55° n.B. und zwischen 12° und 15° ö.L. von besonderem Interesse sind. Die interessanten Meßzeiten liegen vor allem in den Jahren 71 bis 76, 80 bis 86 und 90 bis 95. Als relevante Meßmonate konnten die Zeiträume Februar bis März und September bis Oktober ausgewiesen werden. Bei der Analyse der relevanten Meßtiefen ergaben sich die Bereiche von 4.3m bis 5.8m, von 15.8m bis 18.7m, von 9.3m bis 10.0m und von 20.1m bis 21.6m als besonders interessant (vgl. Anhang C.2).

Das Gitter der Meßpunkte wurde als unregelmäßig klassifiziert.

In der anschließenden Festlegung der anzuwendenden Datenklassen-Metadatenbestimmungen wurde die Gewinnung von Metadaten für Multiparameterdaten ausgewählt.

4. In der **Bestimmung der Multiparameter-Metadaten** wurden als erstes eine hierarchische Klassifikation der vorhandenen Datensätze durchgeführt. Problematisch dabei war, daß die Anzahl der Datensätze bereits eine Proximitätsmatrix[9] von der Größe 9000×9000 erzeugen würde, was zu einer praktischen Größe von etwa 309 MB entspricht[10]. Bedenkt man, daß auch die entstehenden Klassifikationsdatenstrukturen im Speicher gehalten werden müssen, ist die Durchführung einer Klassifikation für eine Datenmenge dieser Größe auf der aktuellen Rechentechnik nicht praktikabel. Deswegen wurde die Metadatengewinnung der Klassifikation automatisch übersprungen. Eine Lösung dieses Problems ist, eine repräsentative Untermenge der Gesamtdatenmenge zu bilden und diese zu klassifizieren.

   Deswegen wurde für eine Untermenge von 563 Datensätzen eine Klassifikation durchgeführt. Diese zeigte, daß sich die Datensätze in der obersten Hierarchiestufe in 2 Klassen aufspalten. Dabei ist die Verteilung der sich bildenden Klassen ungleichmäßig.

   In Klasse 1 fallen 392 Datensätze. Sie untergliedert sich relativ stark in weitere Unterklassen. In die zweite Klasse fallen hingegen lediglich 171 Datensätze. Diese ist mit einer Unterteilung in drei relativ gleichmäßig starke Klassen ziemlich homogen.

[8] Diese Variablen repräsentieren jeweils die Gradzahlen und die Minuten der entsprechenden Meßstationen.

[9] Diese wird bei jeder Klassifikation erzeugt.

[10] unter Annahme der Speicherung von double-Werten in einer Dreiecksmatrix

Um genauere Aussagen über die Eigenschaften der Teilklassen ableiten zu können, müßte eine detailliertere Analyse der Verteilung der Datensätze auf die einzelnen Klassen z.B. in einer Visualisierung durchgeführt werden.

Weiterhin wurden in dieser Untermenge[11] keine gleichen Datensätze und keine Ausreißer identifiziert.

6.3.3 Interpretation

Die gewonnen Metadaten für die Ostsee-Datenmenge haben interessante Informationen über die Datenmenge aufgedeckt und sie für eine Visualisierung vorbereitet. Interessant war vor allem, genauere Angaben über Lage der relevanten Regionen in Zeit, Meßposition und Meßtiefe zu erhalten. Aus der Klassifikation eines Ausschnitts der Datenmenge konnten nur bedingt allgemeingültige Aussagen abgeleitet werden.

Wichtig für die Visualisierung war die Hierarchisierung der Raum- und der Zeitdimensionen, um dort eine Darstellung von geeigneten Koordinatenachsen durchführen zu können.

Weitere Entscheidungsunterstützungen im Hinblick auf eine anschließende Visualisierung sind:

1. Die redundante Variable GLG sollten gesondert dargestellt werden.
2. Die Bereiche von besonderem Interesse sollten hervorgehoben werden.
3. Aufgrund des Vorliegens von drei Raumdimensionen ist die Wahl einer 3-dimensionalen Darstellungstechnik sinnvoll.
4. Auch der Zeitbezug sollte visualisiert werden.
5. Da alle Variablen quantitativ sind, sollten deren Werte identifizierbar sein.

Um eine geeignete Metadatengewinnung durchführen zu können, mußte bei dieser Datenmenge das Wissen des Nutzers maximal eingebunden werden. Entsprechend war ein hoher Interaktionsgrad erforderlich.

6.4 Zeitverhalten und Machbarkeit

Die Zeiten für die Durchführung der rechenintensiven internen Berechnungsalgorithmen auf einem PC mit einem Pentium 3 mit 450MHz und 128MB RAM ist in Tabelle 6.1 dargestellt.

[11] Die Bestimmungen der gleichen Datensätze und der Ausreißer basieren auf der in der Klassifikation erzeugten Proximitätsmatrix und sind deswegen von der durchgeführten Klassifikation abhängig.

Algorithmus	Datenmenge		
	Wassermolekül	Zellkern	Ostsee
Schlüsselanalyse	1min	27s	3min 10s
Bestimmung der gemeinsamen Informationen (Relevanzschranke 0.8)	7s	3s	1s[12]
Segmentierung von Teilräumen	1s	2s	20s
Klassifikation der Datensätze	-	-	nicht praktikabel
Analyse des Monotonieverhaltens	-	5s	-

Tabelle 6.1: Zeitverhalten von Metadatenanalysen für die Datenmengen des elektrostatischen Feldes eines Wassermoleküls, des Zellkerns einer Maus und der Messungen in der Ostsee

Hieran wird deutlich, daß die Gewinnung der Metadaten der Ostsee-Datenmenge bereits an die Grenzen von dem, was auf der aktuellen Rechentechnik durchführbar und für den Nutzer zumutbar ist, stößt. Die im Programm umgesetzte Unterstützung bei diesem Problem ist die Auswahl des schnellen Berechnungsmodus, in dem auf zu zeitintensive Metadatengewinnungen verzichtet wird oder diese durch geeignete Parameterwahl beschleunigt werden.

Allerdings sind auch die Analysen des elektrostatischen Feldes eines Wassermoleküls und des Zellkerns einer Maus verhältnismäßig langsam, bedenkt man, daß es sich hierbei um 2-dimensionale Datensätze von den Größen 102×102 und 258×255 handelt. In praktische Anwendungen sind Volumen- und Strömungsdatenmengen im 3-dimensionalen Raum sind jedoch um ein Vielfaches größer. Diese haben bsw. Gittergrößen von $256 \times 256 \times 128$. Eine Möglichkeit zur Lösung dieses Problems ist die interne Verwaltung von strukturierten und regelmäßigen Gittern durch implizite Speicherung der Raumkoordinaten in einem Array.

[12]Dieser geringe Wert kommt ist darauf zurückzuführen, daß die Informationsgehalte aus den Variablen-Metadaten bei der Berechnung der gemeinsamen Informationsgehalte höherer Ordnung wiederverwendet werden. Da die einfachen Informationsgehalte aber alle bereits unter der Relevanzschranke liegen, brauchen keine gemeinsamen Informationsgehalte berechnet zu werden.

Kapitel 7

Erweiterungsmöglichkeiten von Metadaten

Die in den Kapiteln 3 und 4 vorgestellten Konzepte konnten aufgrund ihrer Vielschichtigkeit nur teilweise umgesetzt werden. Vor allem während der „Metadatenaquisitionsphase“ ergaben sich Ideen, die den Rahmen der Arbeit gesprengt hätten. Auf Erweiterungen und Verfeinerungen der Konzepte und Verbesserungen des Werkzeugs „Metadatum“ soll in diesem Kapitel eingegangen werden.

In 7.1 werden Ideen im Rahmen der Parametrisierung von Visualisierungstechniken vorgestellt. 7.2 beschreibt die Möglichkeiten des Einsatzes neuer Metadaten und Algorithmen. In 7.3 werden Erweiterungen für das Programm „Metadatum“ vorgestellt.

7.1 Visualisierungsparameter

In dieser Arbeit lag der Schwerpunkt bei der Definition und Implementation von Metadaten auf der Unterstützung von Visualisierungsentscheidungen vor allem bei der Wahl einer geeigneten Visualisierungstechnik. Ein nächster Schritt wäre die Einbindung von Metadaten, die für eine gegebene Visualisierungstechnik die Abbildung der Daten auf geeignete Visualisierungsprimitive und die Wahl von entsprechenden Parametern unterstützen. Hierfür ist eine Vielzahl von unterschiedlichen Metadaten denkbar. Im folgenden werden beispielhaft einige solcher Metadaten und deren Nutzen für die Parametrisierung von Visualisierungstechniken vorgestellt:

- Die Metadaten zur Ausweisung von Verteilungen der Wertebereiche könnten erweitert werden, um z.B. relevante Intervalle für die Isoflächenextraktion in der Volumenvisualisierung vorschlagen zu können. Dazu ließe sich zum einen die Verteilung automatisch untersuchen. An den Stellen, an denen die Häufigkeit signifikant abfällt, könnten automatisch Intervallgrenzen eingefügt werden. Zum anderen ist eine anwendungsspezifische Vorgabe von Schwellwertintervallen denkbar. Zum Beispiel könnten in medizinischen Datenmengen Übergänge von einem Gewebetyp zum anderen immer gleichen Schwellwerten entsprechen

und entsprechend zur Festlegung von Intervallen genutzt werden.

- Ein weiteres Beispiel für die Unterstützung von Volumenvisualisierungen ist, erhaltene Isoflächen auf deren spezielle Eigenschaften hin zu untersuchen. So könnten z.B. Krümmungen und Formen von Isoflächen analysiert werden, um diese entweder direkt zu visualisieren oder eine geeignete Abbildung auf die Primitive der Darstellung durchführen zu können. So könnten bsw. Transparenzwerte in und Kugelgrößen entsprechenden Volumenvisualisierungs-Techniken vorgeschlagen werden, um bei vorgegebenem Datenmenge eine große Interpretierbarkeit der Datenmenge zu erreichen.

- Bisher wurden mit der Bestimmung der gemeinsamen Informationsgehalte, der Korrelationen und anderer Strukturbeziehungen zwischen den Variablen allgemeine Beziehungen zwischen Variablen bestimmt. In speziellen Visualisierungstechniken ist es jedoch häufig von Nutzen, eine sogenannte optimale Reihenfolge der Merkmale zu bestimmen. Diese ordnet die Merkmale in einer Liste. Ziel dieser Anordnung kann es z.B. sein, daß sich in dieser Liste hochkorrelierte Merkmale nebeneinander, schwach korrelierte Merkmale aber voneinander entfernt befinden. Mit einer solchen Anordnung sind die Beziehungen zwischen den Merkmalen z.B. bei einer Streckenzugdarstellung besser zu erkennen.

- Für die Parametrisierung von Strömungsvisualisierungstechniken ist z.B. die Bestimmung charakteristischer Strömungslinien von Interesse. So könnten in der Umgebung von kritischen Punkten die Verläufe der Strömungslinien analysiert werden. Dies kann zum einen Aufschluß über deren Verhalten geben und zum anderen durch Extraktion interessanter Strömungslinien eine Reduktion der Komplexität der Darstellung ermöglichen.

- Ein Metadatum, das unabhängig von einer speziellen Datenklasse erhoben werden könnte, ist die Stabilität der Datenmenge auf Interpolationstechniken verschiedener Ordnung. Grundziel bei der Interpolation von Datenwerten im Beobachtungsraum ist, eine möglichst geringe Ordnung zu verwenden, da diese wesentlich schneller berechenbar ist. Interpolationen geringer Ordnung erzeugen jedoch in der Visualisierung Aliaseffekte, aufgrund derer z.B. die Gitterstruktur erkennbar wird. Weiterhin wichtig bei der Wahl einer Interpolationsmethode ist vor allem, daß die Topologie des entsprechenden Feldes erhalten bleibt. Diese unterschiedlichen Anforderungen wirken teilweise entgegengesetzt.

  Es läßt sich vermuten, das in Abhängigkeit von den speziellen Eigenschaften der Daten unterschiedliche Ordnungen von Interpolationsverfahren unterschiedliche Qualitäten einer resultierenden Visualisierung erzeugen.

  Zur Ausweisung eines geeigneten Interpolationsverfahrens könnten Metadaten erhoben werden.

7.2 Weitere Algorithmen und Metadaten

Im folgenden werden Ideen zu weiteren Metadaten und Algorithmen zu deren Bestimmung vorgestellt:

- Bei der Definition und Umsetzung des Metadatenkonzeptes erfolge eine Konzentration auf die drei für die Visualisierung wichtigsten Datenklassen. Denkbar wäre allerdings, auch für andere Datenklassen Metadaten zu definieren, welche die Spezifika dieser Datenklassen repräsentieren und zugehörige Visualisierungen unterstützen. Beispiele hierfür wäre die Einbeziehung der Datenklassen GIS und Scatterd-Data in das Metadatenkonzept.

- In der Spezifikation der Metadaten wurden Metadaten für Beobachtungsfälle lediglich angedeutet. Die Definition und Bestimmung von Metadaten für interessante Schnitte durch den Beobachtungsraum scheint vielversprechend, da Bereiche von Interesse häufig nicht in einfacher Weise – z.B. achsenparallel – vorliegen. Aufgrund der Bestimmung von interessanten Beobachtungsfällen könnte die Aussagekraft von Visualisierungen signifikant erhöht werden.

- Auch eine Erweiterung der Hierarchisierung von Variablen über eine rein interaktive Eingabe hinaus erscheint sinnvoll. So könnten analog zu den in [Rob90] vorgestellten Tabellen die Strukturierung von Variablen eingelesen werden. Ein anderer bereits vorgestellter Ansatz zur Hierarchisierung von Variablen wäre die Durchführung einer Faktorenanalyse. Denkbar ist weiterhin, mit Hilfe von semantischen Analysen z.B. mit Hilfe eines Data-Dictionarys Beziehungen zwischen Variablen aufgrund deren Namen und deren Ausprägungen zu bestimmen.

- Allgemein könnten analog zu den Segmentierungen von Teilräumen weitere Metadaten definiert und erhoben werden, die sowohl der Beschleunigung von Interaktions- und Berechnungsprozessen als auch der Unterstützung von Datenverbesserungen und Bestimmung von Strukturen in der Datenmenge dienen. Vorrangiges Ziel bei der Spezifikation solcher Metadaten wäre es, die Reduktion großer Datenmengen geeignet zu unterstützen, ohne daß dabei wichtige Informationen verloren gehen.

- Ein weiterer interessanter Punkt bei der Erhebung von Metadaten ist die Beachtung der Unsicherheit von automatischen Metadatengewinnungen. Die Frage hierbei ist, wie aussagekräftig Metadaten sein können, bei deren Berechnung z.B. numerische Fehler gemacht werden. Entsprechend müssen dann auch Visualisierungen bewertet werden, die auf solchen unsicheren Metadaten basieren.

- Neben der Unsicherheit der Verfahren zur Metadatengewinnung können auch die eingelesenen Werte fehlerbehaftet sein. Bisher wurden im speziellen lediglich fehlende Werte eingebunden. Die Einbeziehung von Fehlern von Eingangswerten in das Konzept scheint ebenfalls von großer Bedeutung zu sein, um die

Qualität von abgeleiteten Metadaten einschätzen und resultierende Visualisierungen bewerten zu können. Weiterhin könnten entsprechende Fehler auch separat dargestellt werden.

- Neben der Hierarchisierung von Variablen läßt sich Vermuten, daß die Analyse von semantischen Informationen unter Auswertung von Datensatz- und Variablennamen sowie der Zeichenketten der Ausprägungen qualitativer Merkmale interessante Aspekte einer Datenmenge ausweisen kann. Vorbild hierfür wären Szenarien im Umfeld Knowledge Discovery in Datenbanken.

- Interessant wäre weiterhin zu untersuchen, ob aufgrund von spezifischen Größen von Strömungsfeldern wie Krümmung, Divergenz, Druck u.a. und den Beziehungen dieser Größen untereinander relevante Metadaten abgeleitet werden könnten. Entsprechende Metadaten könnten versuchen, folgende Fragen zu beantworten:

  1. Welche spezifischen Größen des Vektorfeldes beinhalten große Information?
  2. Welche Paare von Größen weisen Korrelationen auf, und bei welchen liegen erwartete Korrelationen nicht vor?

  Entsprechend könnten signifikante Metadaten in eine Visualisierung integriert werden.

- Bei der Bestimmung von interessanten Teilräumen des Beobachtungsraumes ist eine Vielzahl von Analysen zu deren Ausweisung denkbar. Eine Idee hierfür ist, die Eigenschaften der frequenztransformierten Teilräume zu bestimmen. Das aufgrund einer Fourier-Transformation bestimmte Frequenzspektrum könnte dann Aufschluß darüber geben, wie homogen bzw. wie heterogen die Teilräume sind.

- Eine weiterer Ansatz zur Analyse von 3-dimensionalen Volumendaten wäre die Nutzung von Methoden der Bildverarbeitung zur Analyse sowohl von 2D-Schnitten als auch der gesamten Volumendatenmenge. Hierbei müßte geklärt werden, welche Metadaten unter dem Gesichtspunkt der Unterstützung der Visualisierung abgeleitet werden könnten.

- Ein weiterer Punkt ist die Einbeziehung historischer Metadaten in Konzept und Programm. Denkbar wäre z.B., anwendungsspezifisch historische Metadaten zu spezifizieren. Neben dem bereits genannten Beispiel der Fehlerbehaftung der Eingangswerte sind z.B. auch Metadaten-Gewinnungen denkbar, welche während der Datenmessung oder -berechnung fortlaufend spezielle Eigenschaften der Datengewinnung wie Datenveränderungen oder aktuelle Stände ausweisen.

7.3 Implementationserweiterungen

Hier sollen einige Erweiterungen, die sich speziell auf das Programm „Metadatum“ beziehen, aufgelistet werden:

- Bisher wird bei Einlesen einer Datenbank diese behandelt wie eine Rohdatendatei. Unter Nutzung der speziellen Eigenschaften von Datenbanken könnten allgemein verschiedene Metadatenerhebungen unterstützt werden. So könnten beispielsweise aufgrund der Spaltentypen die Skalentypen von Variablen schneller bestimmt werden. Desweiteren liegt bei den meisten Datenbanken bereits ein Schlüssel vor. Auf eine zeitaufwendige separate Schlüsselanalyse könnte entsprechend verzichtet werden.

- Das Einlesen der Daten sollte um neue Datenformate erweitert werden, um deren spezielle Vorteile nutzen zu können.

- Die Speicherung des Metadatenformates ist in ASCII-Format zwar relativ übersichtlich und leicht manipulierbar, jedoch können bei großen Datenmengen verhältnismäßig große Dateien entstehen. Entsprechend dauert das Einlesen dieser Dateien auch eine vergleichsweise lange Zeitdauer. Um diesen Problemen entgegenzuwirken, könnte eine Speicherung der Metadaten in das Binär-Format umgesetzt werden. Diese wurde im Programmcode bereits vorgesehen, ist jedoch noch nicht vollständig umgesetzt.

  Der Nutzer könnte dann wahlweise entscheiden, welche der beiden Speicherungsmethoden er verwenden möchte.

- Bisher kann eine Datenmenge maximal aus zwei Tabellen bestehen. Liegen in praktischen Anwendungen jedoch komplexere Zusammenhänge vor, könnte dies nicht mehr ausreichen. Eine Erweiterung der Datenmenge um weitere Tabellen wurde deswegen in Ansätzen integriert, jedoch nicht vollendet.

- Eine Verbesserung des Algorithmus zur Bestimmung von gemeinsamen Informationsgehalten ist die Ausführung von Gruppierungen der Ausprägungen der Variablen. Diese beschleunigt die Berechnung und führt bei „guten“ Gruppierungen zur besseren Veranschaulichung der gemeinsamen Informationsgehalte.

Kapitel 8

Zusammenfassung

Im Laufe der Arbeit ist es gelungen, ein erweiterbares Konzept zur Spezifikation und Erfassung von Metadaten zu entwickeln, um wichtige Eigenschaften von Datenmengen aus Sicht der wissenschaftlichen Visualisierung erfassen zu können. Dabei wurde eine umfassende Definition von beschreibenden Metadaten zur Unterstützung von Visualisierungsentscheidungen durchgeführt. Eine solche liegt bisher noch nicht vor. Weiterhin konnte eine Vielzahl von abgeleiteten Metadaten spezifiziert werden, um interessante Aspekte der Dateneigenschaften zu beschreiben. Wie in Kapitel 7 ausgeführt wurde, gibt es bei Definition und Umsetzung von abgeleiteten Metadaten noch eine breite Palette an Erweiterungsmöglichkeiten. Weiterhin konnten bei der Konzeption von Metadaten sowohl allgemeine Eigenschaften der Datenmenge als auch Eigenschaften spezieller Datenklassen ausgewiesen werden. Dabei ist es gelungen, Metadaten, die üblicherweise speziellen Datenklassen zugeordnet werden, zu verallgemeinern.

Aufgrund der mit Bedacht sehr allgemein formulierten Metadaten wurde dann ein Konzept zur Erfassung dieser Metadaten entwickelt, welches speziell für Datenmengen in Tabellenform konzipiert wurde. Damit wurde die Grundlage für eine praktikable Umsetzung der allgemeinen Konzepte gelegt. Im speziellen wurde zum einen eine geeignete Reihenfolge der Metadatengewinnungen festgelegt, so daß sich voneinander abhängige Metadatengewinnungen optimal unterstützen. Zum anderen wurde durch die Konzeption von Bestimmungsmodulen eine Unterstützung des Nutzers bei der Metadatengewinnung integriert. Dabei wurden spezielle Nutzerprofile einbezogen, um den Nutzer intuitiv, je nach dessen Erfahrung und dessen Anforderungen, in die Gewinnung von Metadaten einzubinden.

Aufbauend auf dem Konzept zur Erfassung der Metadaten wurde dann das Werkzeug „Metadatum“ entwickelt und implementiert. Dabei wurde eine Eingabeschnittstelle für Tabellen, die Metadatengewinnung und eine Ein- und Ausgabeschnittstelle für Metadaten entwickelt. Durch Integration verschiedener Analysealgorithmen konnte eine Metadatengewinnung umgesetzt werden, die mit Hilfe von Standardbelegungen und Nutzerprofilen den Nutzer in geeigneter Weise unterstützt.

Eine mögliche Anwendung der entwickelten Metadatengewinnung wäre zum einen die Einbindung in eine Visualisierungshilfe. Hier könnten unter Integration der Me-

tadaten mit anderen Kontexten wie z.B. den Nutzerinterpretationszielen die Metadaten genutzt werden, um Visualisierungsentscheidungen zu treffen. Zum anderen ist jedoch auch eine direkte Nutzung der Ergebnisse durch spezielle Visualisierungsanwendungen denkbar. Diese könnten die im Metadatenformat gespeicherten Metadaten auslesen und in Abhängigkeit ihrer speziellen Erfordernisse benutzen.

Abschließend bleibt festzustellen, daß die Ausweisung und die Gewinnung von Metadaten breite Anwendungsperspektiven in der Visualisierung verspricht.

Literaturverzeichnis

[B+92] K. W. Brodlie et al., *Scientific Visualisation*, Springer-Verlag, Berlin, 1992.

[BEPW96] Klaus Backhaus, Bernd Erichson, Wulff Plinke, and Rolf Weber, *Multivariate Analysemethoden. Eine anwendungsorientierte Einführung.*, 8 ed., Springer, 1996.

[BG89] R.D. Bergeron and G.G. Grinstein, *A Reference Model for the Visualisation of Multidimensional Data*, Proceedings Eurographics '89, 1989.

[Boc74] Hans Hermann Bock, *Automatische Klassifikation*, Vandenhoeck & Ruprecht, Göttingen, 1974.

[Frü97] T. Frühauf, *Graphisch-Interaktive Strömungsvisualisierung*, Springer-Verlag, Berlin, 1997.

[GLdCS97] K. U. Graw, S. Lange, N. Lopez de Chavez, and H. Schumann, *Konzept und Realisierung einer intelligenten Visualisierungshilfe*, Preprint, Universität Rostock, August 1997.

[HS95] A. Heuer and G. Saake, *Datenbanken, Konzepte und Sprachen*, 1 ed., International Thomson Publishing GmbH, 1995.

[Jun98] V. Jung, *Integrierte Benutzerunterstützung für die Visualisierung in Geo-Informationssystemen*, Dissertation, TU Darmstadt, Fachbereich Informatik, Fachgebiet GRIS , Fraunhofer IRB Verlag, Stuttgard, 1998.

[Köl00] M. Köller, *Schwellwertunabhängiges Preprocessing für Marching Cubes*, Diplomarbeit, Universität Rostock, Fachbereich Informatik, 2000.

[Kre96] M. Kreuseler, *Visualisierung maritimer Umweltdaten in ihrem geographischen Kontext unter Einbeziehung der Navigation im Darstellungsraum und in der Zeit*, Diplomarbeit, Universität Rostock, Fachbereich Informatik, 1996.

[Lan97] S. Lange, *Ermittlung von Wertebereichs- und Schlüsseleigenschaften für Visualisierungsentscheidungen*, Universität Rostock, Preprints aus dem Fachbereich Informatik (1997).

[Mül98] P. C. Müller, *Zur automatischen Segmentierung von Knochen des Hüftgelenks in Volumendaten für ein Operationsplanungssystem*, Dissertation, Universität Hildesheim, Fachbereich für Mathematik, Informatik und Naturwissenschaften, 1998.

[Noc99] T. Nocke, *Konzeption und Realisierung einer flexiblen Pipeline zur numerischen Vorverarbeitung in der Informationsvisualisierung*, Studienarbeit, Universität Rostock, Fachbereich Informatik, 1999.

[RDE93] R. Rew, G. Davis, and S. Emmerson, *Net CDF user's Guide. An Interface of Data Access*, UCAR, 1993.

[RH97] P. K. Robertson and M.A. Hutchins, *An Approach to Intelligent Design of Color Visualisation*, In: G. Nielson, H. Hagen, H. Müller: Scientific Visualisation. IEEE Computer Society, Los Alamitos (1997), S. 179–190.

[Rob90] P. K. Robertson, *A Methodology for Scientific Data Visualisation: Chosing Representations Based on a Natural Scene Paradigm.*, Procedings Visualisation '90, IEEE Computer Society Press, Los Alimatos (1990), S. 114–123.

[SM00] H. Schumann and W. Müller, *Visualisierung, Grundlagen und allgemeine Methoden*, 1 ed., Springer-Verlag, Berlin Heidelberg, Januar 2000.

[The94] H. Theisel, *Automatische Auswahl geeigneter Visualisierungstechniken für allgemeine wissenschaftliche Datensätze*, Diplomarbeit, Universität Rostock, Fachbereich Informatik, 1994.

[The95] H. Theisel, *Analyse und Visualisierungshilfe für mehrdimensionale wissenschaftliche Daten*, Informatik, Forschung und Entwicklung **10** (1995), S. 91–98.

[The96] H. Theisel, *Vector Field Curvature and Applications*, Dissertation, Universität Rostock, Fachbereich Informatik, 1996.

[WB97] P. C. Wong and R. D. Bergeron, *30 Years of Multidimensional Multivariate Visualisation*, In: G. Nielson, H. Hagen, H. Müller: Scientific Visualisation. IEEE Computer Society, Los Alamitos (1997), S. 3–33.

Abbildungsverzeichnis

Tabellenverzeichnis

Anhang A

Standardbelegungen

Metadatum	Standardbelegung
Skalentyp	nominal, diskret oder kontinuierlich
Art des Raumes	abstrakte Raumdimensionen
Trennung abhängiger und unabhängiger Variable	Auswahl der Variablen des ersten minimalen Schlüssel
Ausweisen redundanter Merkmale	Auswahl aller Merkmale mit einem Informationsgehalt $= 0$ oder Wahl des zweiten Merkmals bei Merkmalspaaren mit einer Korrelation ≥ 0

Tabelle A.1: Standardbelegungen ausgewählter Metadaten

Anhang B

Beispieltabellen

B.1 Ausschnitt aus der Haupttabelle der Ostsee-Datenmenge

STATBEZ	STATNR	TAG	MON	JAHR	STDE	MIN	TIEFE	TEMP	SALZ	O2ML	O2PR	GBG	GBM	GLG	GLM	Q
40	52	26	3	68	20	10	0.5	2.95	10.77	9.79	111.1	54	29	12	3	1
40	52	26	3	68	20	10	5	2.95	10.81	9.7	110.2	54	29	12	3	1
40	52	26	3	68	20	10	10	2.5	11.82	9.24	104.5	54	29	12	3	1
33	55	27	3	68	2	42	0.5	2.44	8.08	9.37	103	54	36	12	19	1
33	55	27	3	68	2	42	5	2.43	8.08	9.33	102.5	54	36	12	19	1
33	55	27	3	68	2	42	10	2.26	8.15	9.33	102.3	54	36	12	19	1
33	55	27	3	68	2	42	17	2.14	10.64	8.81	97.9	54	36	12	19	1
33	55	27	3	68	2	42	0.5	2.44	8.08	9.37	103	54	36	12	19	1
33	55	27	3	68	2	42	5	2.43	8.08	9.33	102.5	54	36	12	19	1
33	55	27	3	68	2	42	10	2.26	8.15	9.33	102.3	54	36	12	19	1
33	55	27	3	68	2	42	17	2.14	10.64	8.81	97.9	54	36	12	19	1
31	59	30	3	68	7	38	0.5	3.68	8.28	9.66	109.8	54	40	12	33	1
31	59	30	3	68	7	38	5	4.4	8.59	9.55	110.8	54	40	12	33	1
31	59	30	3	68	7	38	10	2.05	9.09	9.3	102.3	54	40	12	33	1
31	59	30	3	68	7	38	15	1.61	9.72	9.11	99	54	40	12	33	1
30	60	30	3	68	9	45	0.5	2.9	7.76	9.7	107.8	54	43	12	47	1
30	60	30	3	68	9	45	5	2.57	7.81	9.6	106	54	43	12	47	1
30	60	30	3	68	9	45	10	2.03	7.95	9.37	102	54	43	12	47	1
30	60	30	3	68	9	45	20	1.64	8.24	9.24	99.6	54	43	12	47	1
115	61	30	3	68	11	55	0.5	2.82	7.94	9.62	106.9	54	47	13	3	1
115	61	30	3	68	11	55	5	2.77	7.94	9.64	106.9	54	47	13	3	1
115	61	30	3	68	11	55	10	2.55	8.01	9.64	106.4	54	47	13	3	1
115	61	30	3	68	11	55	20	2.25	8.19	9.36	102.6	54	47	13	3	1
115	61	30	3	68	11	55	26	1.92	10.95	8.86	98.2	54	47	13	3	1
114	62	30	3	68	13	34	0.5	3.25	7.88	9.82	110.3	54	51	13	16	1
114	62	30	3	68	13	34	5	3.01	7.94	9.8	109.3	54	51	13	16	1
114	62	30	3	68	13	34	10	2.21	8.03	9.45	103.4	54	51	13	16	1
114	62	30	3	68	13	34	20	1.85	8.33	9.35	101.5	54	51	13	16	1
114	62	30	3	68	13	34	30	1.35	10.12	9.04	98	54	51	13	16	1
114	62	30	3	68	13	34	41	1.81	15.14	8.23	93.7	54	51	13	16	1
113	63	30	3	68	15	40	0.5	2.97	7.67	9.59	106.4	54	55	13	30	1
113	63	30	3	68	15	40	5	2.91	7.72	9.63	106.9	54	55	13	30	1

...

B.2 Tabelle zur Speicherung des separaten Raumbezuges der Meßstationen der Ostsee-Datenmenge

SATZNR	STATBEZ	gradbrei	minbrei	gradlaen	minlaen	quad	tiefe	bmp	iby	monitor
1	011	54	24.80	11	37.00	1	25.0	-	-	-
2	012	54	18.90	11	33.00	1	25.0	M2	-	M
25	030	54	43.40	12	47.00	1	22.0	K8	-	M
23	031	54	40.20	12	33.70	1	17.0	-	-	-
13	033	54	36.30	12	19.90	1	20.0	-	-	-
155	040	54	29.30	12	3.90	1	13.0	-	-	-
28	069	55	0.00	13	18.00	1	46.0	K7	-	M
39	102	55	9.30	13	56.50	1	45.0	-	-	-
40	103	55	3.80	13	59.30	1	47.0	-	-	-
35	104	55	4.10	13	48.80	1	46.0	-	-	-
34	105	55	1.50	13	36.40	1	45.0	-	-	-
41	109	55	0.00	14	5.00	1	47.0	K4	-	M
42	111	54	53.40	13	58.10	1	44.0	-	-	-
43	112	54	48.20	13	57.50	1	40.0	-	-	-
29	113	54	55.50	13	30.00	1	47.0	K5	-	M
27	114	54	51.60	13	16.60	1	45.0	-	-	-
26	115	54	47.70	13	3.50	1	29.0	-	-	-
30	116	54	47.40	13	29.50	1	44.0	-	-	-
31	120	54	43.30	13	42.20	1	40.0	-	-	-
44	121	54	42.60	13	56.80	1	29.0	-	-	-
36	144	55	15.00	14	30.40	1	46.0	-	-	-
37	145	55	10.00	14	15.00	1	47.0	-	-	-
45	150	54	36.70	14	2.60	1	22.0	-	-	-
52	160	54	14.40	14	4.10	1	13.0	-	-	-
53	162	54	8.40	14	12.50	1	13.0	-	-	-
54	202	54	42.00	15	15.00	1	65.0	-	6B	-
55	204	54	50.70	15	22.50	1	70.0	-	-	-
63	213	55	15.00	15	59.00	1	91.0	K2	-	M
56	215	55	0.00	15	30.00	1	76.0	-	-	-
7	044	54	12.90	12	5.10	1	11.0	-	-	-
8	043	54	15.20	12	4.10	1	14.0	-	-	-
9	042	54	19.20	12	3.70	1	18.0	-	-	-
10	041	54	24.40	12	3.70	1	19.0	-	-	-
147	057	54	5.81	12	7.12	1	0.0	-	-	-
149	055	54	7.09	12	5.88	1	0.0	-	-	-
151	053	54	8.32	12	5.78	1	0.0	-	-	-
153	051	54	9.60	12	5.91	1	0.0	-	-	-
154	050	54	10.73	12	5.79	1	0.0	-	-	-
130	093	54	52.50	12	37.00	1	17.8	-	-	-
131	095	54	49.00	12	21.00	1	22.1	-	-	-
132	094	54	48.50	12	27.00	1	20.7	-	-	-
133	096	54	46.20	12	32.70	1	17.0	-	-	-
134	097	54	43.40	12	37.40	1	19.0	-	-	-
135	098	54	39.70	12	46.30	1	13.0	-	-	-
136	005	54	37.00	12	39.50	1	0.0	-	-	-
137	003	54	35.60	12	32.70	1	17.7	-	-	-
138	080	54	36.80	12	10.50	1	20.0	-	-	-

Anhang C

Ausschnitte aus einer Metadatendatei

Im folgenden werden Ausschnitte aus der durch die Analyse der Ostsee-Datenmenge erzeugten Metadatendatei vorgestellt.

C.1 Variablen- und Datenmenge-Metadaten

```
Variablenmetadaten: Berechnet
 Anzahl_Variable: 17
 Variable:  STATBEZ
  ReferenzNummer: 0
  Wertebereich:
   Skalentyp: diskret, exklusiv
   Histogramm:
    Anzahl_Datensaetze_Ohne_Fehlende: 9097
    Mittelwert: 122.275
    Varianz: 2981.09
    Verteilung:
     Anzahl_Auspraegungen: 27
      ( 12 : 378 ) ( 30 : 272 ) ( 31 : 194 ) ( 33 : 237 ) ( 40 : 77 )
      ( 69 : 275 ) ( 102 : 433 ) ( 103 : 403 ) ( 104 : 330 ) ( 105 : 339 )
      ( 109 : 288 ) ( 111 : 337 ) ( 112 : 310 ) ( 113 : 1024 ) ( 114 : 345 )
      ( 115 : 238 ) ( 116 : 264 ) ( 120 : 246 ) ( 121 : 316 ) ( 144 : 332 )
      ( 145 : 322 ) ( 150 : 252 ) ( 162 : 122 ) ( 202 : 395 ) ( 204 : 20 )
      ( 213 : 951 ) ( 215 : 397 )
  Qualitaet:
   Anzahl_Fehlender_Werte: 0
  Abhaengigkeit: ja
  Abstraktion: ja
  Semantik:
   Wert: ja
   Name: nein
   Id: nein
  RaumBezugsReferenz: nein
  DatenTyp: 0
  Informationsgehalt: 0.343784
 Variable:  STATNR
```

```
 ReferenzNummer: 1
 Wertebereich:
  Skalentyp: diskret, exklusiv
  Histogramm:
   Anzahl_Datensaetze_Ohne_Fehlende: 9097
   Mittelwert: 414.964
   Varianz: 128519
   Verteilung:
    Anzahl_Auspraegungen: 713
     ( 1 : 14 ) ( 3 : 5 ) ( 4 : 11 ) ( 5 : 10 ) ( 6 : 4 )
     ( 7 : 4 ) ( 8 : 10 ) ( 9 : 14 ) ( 10 : 10 ) ( 11 : 17 )
     ( 12 : 16 ) ( 13 : 15 ) ( 14 : 9 ) ( 15 : 18 ) ( 16 : 14 )
     ...
     ( 2034 : 7 ) ( 2035 : 9 ) ( 2040 : 6 ) ( 2041 : 7 ) ( 2077 : 5 )
 Qualitaet:
  Anzahl_Fehlender_Werte: 0
 Abhaengigkeit: nein
 Abstraktion: nein
 Semantik:
  Wert: ja
  Name: nein
  Id: nein
 RaumBezugsReferenz: nein
 DatenTyp: 0
 Informationsgehalt: 0.696226
Variable:  TAG
 ReferenzNummer: 2
 Wertebereich:
  Skalentyp: diskret, exklusiv
  Histogramm:
   Anzahl_Datensaetze_Ohne_Fehlende: 9097
   Mittelwert: 19.8257
   Varianz: 86.7757
   Verteilung:
    Anzahl_Auspraegungen: 31
     ( 1 : 335 ) ( 2 : 384 ) ( 3 : 212 ) ( 4 : 279 ) ( 5 : 106 )
     ( 6 : 133 ) ( 7 : 79 ) ( 8 : 71 ) ( 9 : 80 ) ( 10 : 123 )
     ( 11 : 135 ) ( 12 : 191 ) ( 13 : 193 ) ( 14 : 239 ) ( 15 : 229 )
     ( 16 : 201 ) ( 17 : 136 ) ( 18 : 49 ) ( 19 : 92 ) ( 20 : 136 )
     ( 21 : 265 ) ( 22 : 376 ) ( 23 : 535 ) ( 24 : 499 ) ( 25 : 671 )
     ( 26 : 879 ) ( 27 : 545 ) ( 28 : 509 ) ( 29 : 410 ) ( 30 : 604 )
     ( 31 : 401 )
 Qualitaet:
  Anzahl_Fehlender_Werte: 0
 Abhaengigkeit: nein
 Abstraktion: nein
 Semantik:
  Wert: ja
  Name: nein
  Id: nein
 RaumBezugsReferenz: nein
 DatenTyp: 0
 Informationsgehalt: 0.351742
Variable:  MON
```

```
 ReferenzNummer: 3
 Wertebereich:
  Skalentyp: diskret, exklusiv
  Histogramm:
   Anzahl_Datensaetze_Ohne_Fehlende: 9097
   Mittelwert: 6.85995
   Varianz: 12.527
   Verteilung:
    Anzahl_Auspraegungen: 6
     ( 2 : 10 ) ( 3 : 3070 ) ( 4 : 1417 ) ( 10 : 3119 ) ( 11 : 1455 )
     ( 12 : 26 )
 Qualitaet:
  Anzahl_Fehlender_Werte: 0
 Abhaengigkeit: nein
 Abstraktion: nein
 Semantik:
  Wert: ja
  Name: nein
  Id: nein
 RaumBezugsReferenz: nein
 DatenTyp: 0
 Informationsgehalt: 0.147067
Variable:  JAHR
 ReferenzNummer: 4
 Wertebereich:
  Skalentyp: diskret, exklusiv
  Histogramm:
   Anzahl_Datensaetze_Ohne_Fehlende: 9097
   Mittelwert: 81.7152
   Varianz: 61.8955
   Verteilung:
    Anzahl_Auspraegungen: 28
     ( 68 : 242 ) ( 69 : 342 ) ( 70 : 366 ) ( 71 : 304 ) ( 72 : 187 )
     ( 73 : 365 ) ( 74 : 288 ) ( 75 : 373 ) ( 76 : 342 ) ( 77 : 436 )
     ( 78 : 312 ) ( 79 : 138 ) ( 80 : 302 ) ( 81 : 298 ) ( 82 : 335 )
     ( 83 : 340 ) ( 84 : 406 ) ( 85 : 312 ) ( 86 : 326 ) ( 87 : 348 )
     ( 88 : 403 ) ( 89 : 355 ) ( 90 : 358 ) ( 91 : 417 ) ( 92 : 446 )
     ( 93 : 413 ) ( 94 : 269 ) ( 95 : 74 )
 Qualitaet:
  Anzahl_Fehlender_Werte: 0
 Abhaengigkeit: nein
 Abstraktion: nein
 Semantik:
  Wert: ja
  Name: nein
  Id: nein
 RaumBezugsReferenz: nein
 DatenTyp: 0
 Informationsgehalt: 0.361221
Variable:  STDE
 ReferenzNummer: 5
 Wertebereich:
  Skalentyp: diskret, exklusiv
  Histogramm:
```

```
    Anzahl_Datensaetze_Ohne_Fehlende: 9097
    Mittelwert: 11.1223
    Varianz: 47.4154
    Verteilung:
     Anzahl_Auspraegungen: 24
      ( 0 : 432 ) ( 1 : 316 ) ( 2 : 338 ) ( 3 : 416 ) ( 4 : 525 )
      ( 5 : 666 ) ( 6 : 346 ) ( 7 : 239 ) ( 8 : 310 ) ( 9 : 345 )
      ( 10 : 354 ) ( 11 : 405 ) ( 12 : 393 ) ( 13 : 369 ) ( 14 : 411 )
      ( 15 : 440 ) ( 16 : 350 ) ( 17 : 345 ) ( 18 : 399 ) ( 19 : 372 )
      ( 20 : 303 ) ( 21 : 316 ) ( 22 : 358 ) ( 23 : 349 )
  Qualitaet:
   Anzahl_Fehlender_Werte: 0
  Abhaengigkeit: nein
  Abstraktion: nein
  Semantik:
   Wert: ja
   Name: nein
   Id: nein
  RaumBezugsReferenz: nein
  DatenTyp: 0
  Informationsgehalt: 0.346306
 Variable:  MIN
  ReferenzNummer: 6
  Wertebereich:
   Skalentyp: diskret, exklusiv
   Histogramm:
    Anzahl_Datensaetze_Ohne_Fehlende: 9097
    Mittelwert: 27.7844
    Varianz: 315.679
    Verteilung:
     Anzahl_Auspraegungen: 60
      ( 0 : 592 ) ( 1 : 152 ) ( 2 : 119 ) ( 3 : 182 ) ( 4 : 97 )
      ( 5 : 200 ) ( 6 : 142 ) ( 7 : 115 ) ( 8 : 215 ) ( 9 : 79 )
      ( 10 : 232 ) ( 11 : 141 ) ( 12 : 147 ) ( 13 : 75 ) ( 14 : 130 )
      ...
      ( 55 : 215 ) ( 56 : 83 ) ( 57 : 116 ) ( 58 : 132 ) ( 59 : 67 )
  Qualitaet:
   Anzahl_Fehlender_Werte: 0
  Abhaengigkeit: ja
  Abstraktion: ja
  Semantik:
   Wert: ja
   Name: nein
   Id: nein
  RaumBezugsReferenz: nein
  DatenTyp: 0
  Informationsgehalt: 0.437249
 Variable:  TIEFE
  ReferenzNummer: 7
  Wertebereich:
   Skalentyp: kontinuierlich, exklusiv
   Histogramm:
    Anzahl_Datensaetze_Ohne_Fehlende: 9097
    Mittelwert: 23.6965
```

```
    Varianz: 358.31
    Verteilung:
     Anzahl_Auspraegungen: 661
      ( 0 : 3 ) ( 0.1 : 3 ) ( 0.2 : 6 ) ( 0.3 : 9 ) ( 0.4 : 13 )
      ( 0.5 : 208 ) ( 0.6 : 17 ) ( 0.7 : 14 ) ( 0.8 : 16 )
      ( 0.9 : 22 ) ( 1 : 672 ) ( 1.1 : 24 ) ( 1.2 : 29 )
      ...
      ( 90.1 : 2 ) ( 90.5 : 2 ) ( 90.8 : 1 ) ( 91 : 1 ) ( 92.3 : 1 )
  Qualitaet:
   Anzahl_Fehlender_Werte: 0
  Abhaengigkeit: ja
  Abstraktion: ja
  Semantik:
   Wert: ja
   Name: nein
   Id: nein
  RaumBezugsReferenz: nein
  DatenTyp: 0
  Informationsgehalt: 0.593954
 Variable:  TEMP
  ReferenzNummer: 8
  Wertebereich:
   Skalentyp: kontinuierlich, exklusiv
   Histogramm:
    Anzahl_Datensaetze_Ohne_Fehlende: 9097
    Mittelwert: 6.64923
    Varianz: 34.3712
    Verteilung:
     Anzahl_Auspraegungen: 1438
      ( -0.56 : 1 ) ( -0.52 : 1 ) ( -0.51 : 1 ) ( -0.46 : 1 ) ( -0.43 : 1 )
      ( -0.42 : 1 ) ( -0.41 : 2 ) ( -0.4 : 1 ) ( -0.39 : 1 ) ( -0.38 : 2 )
      ( -0.37 : 2 ) ( -0.36 : 5 ) ( -0.35 : 4 ) ( -0.34 : 9 ) ( -0.33 : 12 )
      ...
      ( 14.83 : 1 ) ( 14.96 : 1 ) ( 15.05 : 1 ) ( 15.06 : 1 ) ( 99.99 : 18 )
  Qualitaet:
   Anzahl_Fehlender_Werte: 0
  Abhaengigkeit: ja
  Abstraktion: ja
  Semantik:
   Wert: ja
   Name: nein
   Id: nein
  RaumBezugsReferenz: nein
  DatenTyp: 0
  Informationsgehalt: 0.772418
 Variable:  SALZ
  ReferenzNummer: 9
  Wertebereich:
   Skalentyp: kontinuierlich, exklusiv
   Histogramm:
    Anzahl_Datensaetze_Ohne_Fehlende: 9097
    Mittelwert: 11.4596
    Varianz: 125.314
    Verteilung:
```

```
     Anzahl_Auspraegungen: 1274
      ( 1.7 : 1 ) ( 5.45 : 1 ) ( 5.46 : 4 ) ( 5.49 : 2 ) ( 5.5 : 1 )
      ( 5.52 : 8 ) ( 5.53 : 1 ) ( 5.56 : 1 ) ( 5.58 : 1 ) ( 5.61 : 4 )
      ( 5.62 : 3 ) ( 5.67 : 1 ) ( 5.9 : 2 ) ( 5.92 : 1 ) ( 5.94 : 1 )
      ...
      ( 22.89 : 1 ) ( 23.05 : 1 ) ( 23.79 : 1 ) ( 25.49 : 1 ) ( 99.99 : 132 )
  Qualitaet:
   Anzahl_Fehlender_Werte: 0
  Abhaengigkeit: ja
  Abstraktion: ja
  Semantik:
   Wert: ja
   Name: nein
   Id: nein
  RaumBezugsReferenz: nein
  DatenTyp: 0
  Informationsgehalt: 0.703109
 Variable:  O2ML
  ReferenzNummer: 10
  Wertebereich:
   Skalentyp: kontinuierlich, exklusiv
   Histogramm:
    Anzahl_Datensaetze_Ohne_Fehlende: 9097
    Mittelwert: 8.86706
    Varianz: 1410.52
    Verteilung:
     Anzahl_Auspraegungen: 1093
      ( -0.88 : 1 ) ( -0.86 : 2 ) ( -0.79 : 2 ) ( -0.74 : 2 ) ( -0.69 : 1 )
      ( -0.67 : 1 ) ( -0.61 : 2 ) ( -0.53 : 1 ) ( -0.46 : 1 ) ( -0.37 : 1 )
      ( -0.26 : 2 ) ( -0.25 : 3 ) ( -0.24 : 1 ) ( -0.22 : 2 ) ( -0.21 : 1 )
      ...
      ( 12.91 : 1 ) ( 13.11 : 1 ) ( 13.13 : 1 ) ( 14.09 : 1 ) ( 999.99 : 13 )
  Qualitaet:
   Anzahl_Fehlender_Werte: 0
  Abhaengigkeit: ja
  Abstraktion: ja
  Semantik:
   Wert: ja
   Name: nein
   Id: nein
  RaumBezugsReferenz: nein
  DatenTyp: 0
  Informationsgehalt: 0.696767
 Variable:  O2PR
  ReferenzNummer: 11
  Wertebereich:
   Skalentyp: kontinuierlich, exklusiv
   Histogramm:
    Anzahl_Datensaetze_Ohne_Fehlende: 9097
    Mittelwert: 112.542
    Varianz: 20228.6
    Verteilung:
     Anzahl_Auspraegungen: 1102
      ( 0 : 6 ) ( 0.1 : 1 ) ( 1.1 : 1 ) ( 1.4 : 1 ) ( 1.5 : 2 )
```

```
      ( 1.6 : 2 ) ( 1.7 : 1 ) ( 1.8 : 3 ) ( 2 : 2 ) ( 2.2 : 4 )
      ( 2.4 : 2 ) ( 2.5 : 2 ) ( 2.6 : 1 ) ( 2.8 : 3 ) ( 2.9 : 1 )
      ...
      ( 138.2 : 1 ) ( 140.2 : 1 ) ( 140.3 : 1 ) ( 150.3 : 1 ) ( 999.9 : 223 )
  Qualitaet:
   Anzahl_Fehlender_Werte: 0
  Abhaengigkeit: ja
  Abstraktion: ja
  Semantik:
   Wert: ja
   Name: nein
   Id: nein
  RaumBezugsReferenz: nein
  DatenTyp: 0
  Informationsgehalt: 0.648142
 Variable:  GBG
  ReferenzNummer: 12
  Wertebereich:
   Skalentyp: diskret, exklusiv
   Histogramm:
    Anzahl_Datensaetze_Ohne_Fehlende: 9097
    Mittelwert: 54.4474
    Varianz: 0.24726
    Verteilung:
     Anzahl_Auspraegungen: 2
      ( 54 : 5027 ) ( 55 : 4070 )
  Qualitaet:
   Anzahl_Fehlender_Werte: 0
  Abhaengigkeit: ja
  Abstraktion: ja
  Semantik:
   Wert: ja
   Name: nein
   Id: nein
  RaumBezugsReferenz: nein
  DatenTyp: 0
  Informationsgehalt: 0.0754307
 Variable:  GBM
  ReferenzNummer: 13
  Wertebereich:
   Skalentyp: diskret, exklusiv
   Histogramm:
    Anzahl_Datensaetze_Ohne_Fehlende: 9097
    Mittelwert: 27.3621
    Varianz: 414.497
    Verteilung:
     Anzahl_Auspraegungen: 29
      ( 0 : 960 ) ( 1 : 332 ) ( 3 : 281 ) ( 4 : 459 ) ( 8 : 118 )
      ( 9 : 433 ) ( 10 : 180 ) ( 11 : 142 ) ( 14 : 10 ) ( 15 : 1263 )
      ( 18 : 260 ) ( 19 : 128 ) ( 25 : 4 ) ( 29 : 77 ) ( 35 : 3 )
      ( 36 : 409 ) ( 37 : 77 ) ( 40 : 194 ) ( 41 : 8 ) ( 42 : 596 )
      ( 43 : 617 ) ( 47 : 433 ) ( 48 : 387 ) ( 50 : 16 ) ( 51 : 245 )
      ( 52 : 104 ) ( 53 : 331 ) ( 55 : 1013 ) ( 56 : 17 )
  Qualitaet:
```

```
   Anzahl_Fehlender_Werte: 0
  Abhaengigkeit: ja
  Abstraktion: ja
  Semantik:
   Wert: ja
   Name: nein
   Id: nein
  RaumBezugsReferenz: nein
  DatenTyp: 0
  Informationsgehalt: 0.31753
 Variable:  GLG
  ReferenzNummer: 14
  Wertebereich:
   Skalentyp: diskret, exklusiv
   Histogramm:
    Anzahl_Datensaetze_Ohne_Fehlende: 9097
    Mittelwert: 13.3632
    Varianz: 1.04178
    Verteilung:
     Anzahl_Auspraegungen: 5
      ( 11 : 378 ) ( 12 : 789 ) ( 13 : 4844 ) ( 14 : 1323 ) ( 15 : 1763 )
  Qualitaet:
   Anzahl_Fehlender_Werte: 0
  Abhaengigkeit: ja
  Abstraktion: ja
  Semantik:
   Wert: ja
   Name: nein
   Id: nein
  RaumBezugsReferenz: nein
  DatenTyp: 0
  Informationsgehalt: 0.140221
 Variable:  GLM
  ReferenzNummer: 15
  Wertebereich:
   Skalentyp: diskret, exklusiv
   Histogramm:
    Anzahl_Datensaetze_Ohne_Fehlende: 9097
    Mittelwert: 35.193
    Varianz: 337.895
    Verteilung:
     Anzahl_Auspraegungen: 36
      ( 0 : 45 ) ( 1 : 7 ) ( 2 : 175 ) ( 3 : 356 ) ( 4 : 36 )
      ( 5 : 243 ) ( 11 : 4 ) ( 12 : 118 ) ( 13 : 13 ) ( 14 : 8 )
      ( 15 : 554 ) ( 16 : 241 ) ( 17 : 246 ) ( 18 : 275 ) ( 19 : 158 )
      ( 20 : 79 ) ( 22 : 20 ) ( 29 : 264 ) ( 30 : 1703 ) ( 31 : 9 )
      ( 33 : 495 ) ( 34 : 102 ) ( 36 : 339 ) ( 37 : 6 ) ( 38 : 7 )
      ( 42 : 246 ) ( 46 : 6 ) ( 47 : 266 ) ( 48 : 273 ) ( 49 : 57 )
      ( 50 : 9 ) ( 55 : 14 ) ( 56 : 650 ) ( 57 : 392 ) ( 58 : 364 )
      ( 59 : 1317 )
  Qualitaet:
   Anzahl_Fehlender_Werte: 0
  Abhaengigkeit: ja
  Abstraktion: ja
```

```
  Semantik:
   Wert: ja
   Name: nein
   Id: nein
  RaumBezugsReferenz: nein
  DatenTyp: 0
  Informationsgehalt: 0.316024
 Variable:  Q
  ReferenzNummer: 16
  Wertebereich:
   Skalentyp: diskret, exklusiv
   Histogramm:
    Anzahl_Datensaetze_Ohne_Fehlende: 9097
    Mittelwert: 0.995933
    Varianz: 0.00405118
    Verteilung:
     Anzahl_Auspraegungen: 2
      ( 0 : 37 ) ( 1 : 9060 )
  Qualitaet:
   Anzahl_Fehlender_Werte: 0
  Abhaengigkeit: ja
  Abstraktion: ja
  Semantik:
   Wert: ja
   Name: nein
   Id: nein
  RaumBezugsReferenz: nein
  DatenTyp: 0
  Informationsgehalt: 0.00290142
Variablenmengemetadaten:  Berechnet
 Anzahl_Datensaetze: 9097
 Anzahl_Variable: 17
 Strukturierung:
  Schlüssel:
   Anzahl_Schluesselkombinationslaengen: 3
   Anzahl 1 er: 0
   Anzahl 2 er: 0
   Anzahl 3 er: 0
 Gemeinsame_Information:
  Relevanz_Schranke: 0.8
  Vielfachheit: 17
   EndeInformationsgehalt
 Korrelationen:
  Korrelationsschranke: 0.8
  Anzahl_Korrelationswerte: 136
   -0.888622 : 12 13
   -0.275598 : 13 14
   -0.165442 : 0 13
   -0.146082 : 7 13
   -0.136424 : 0 2
   -0.124392 : 2 14
   -0.111114 : 6 9
   -0.109427 : 13 15
   -0.10925 : 9 14
```

```
-0.094847 : 9 12
-0.0903705 : 0 9
-0.0814994 : 6 11
-0.0744873 : 5 15
-0.0708864 : 4 15
-0.069319 : 11 12
-0.0643363 : 2 7
-0.0558148 : 3 4
-0.0546006 : 2 15
-0.0538397 : 3 10
-0.05346 : 4 8
-0.0485089 : 4 11
-0.0470867 : 5 13
-0.0465935 : 15 16
-0.0457206 : 6 13
-0.0457165 : 2 11
-0.0439324 : 5 9
-0.0423491 : 8 10
-0.0412335 : 11 14
-0.0402976 : 0 16
-0.0381807 : 14 16
-0.0366013 : 4 13
-0.036551 : 0 5
-0.036271 : 1 10
-0.0359458 : 0 4
-0.0359256 : 7 11
-0.0354176 : 1 12
-0.0349168 : 5 7
-0.0347651 : 4 5
-0.0310003 : 8 14
-0.0302301 : 4 9
-0.0294671 : 5 16
-0.0288099 : 7 16
-0.0268292 : 6 16
-0.0260682 : 0 11
-0.0255834 : 0 8
-0.0253722 : 9 15
-0.0236237 : 5 14
-0.0217755 : 5 11
-0.0216549 : 8 12
-0.0216483 : 7 10
-0.017137 : 2 13
-0.0165989 : 0 6
-0.0154205 : 4 6
-0.0145399 : 1 6
-0.0141421 : 4 14
-0.0138766 : 4 12
-0.0136268 : 10 15
-0.0128527 : 10 12
-0.0120983 : 13 16
-0.0114294 : 11 15
-0.0113757 : 6 10
-0.0106067 : 6 7
-0.00961361 : 1 16
```

```
-0.00961134 : 3 13
-0.00886287 : 2 12
-0.00805669 : 2 16
-0.0078541 : 6 15
-0.00716429 : 6 14
-0.00552893 : 2 3
-0.0053162 : 8 15
-0.0045419 : 2 4
-0.00304001 : 2 9
-0.0015499 : 12 16
0.000212137 : 2 10
0.000725793 : 10 16
0.00336769 : 1 14
0.00452008 : 3 15
0.00578937 : 1 15
0.00695172 : 9 16
0.00724952 : 3 11
0.00828061 : 1 7
0.00832215 : 11 16
0.00990703 : 0 10
0.0108088 : 4 7
0.0108182 : 14 15
0.0109169 : 0 1
0.0110559 : 3 12
0.0113527 : 2 6
0.0114701 : 3 7
0.0117849 : 10 14
0.0123448 : 1 4
0.012626 : 10 13
0.0210624 : 8 13
0.02513 : 5 10
0.0259198 : 3 14
0.0264262 : 5 12
0.0271443 : 1 5
0.0281792 : 3 5
0.0289256 : 0 3
0.0313599 : 7 8
0.0321894 : 3 6
0.0331171 : 6 12
0.0334592 : 2 5
0.0358912 : 1 13
0.0366039 : 1 11
0.0383027 : 4 16
0.0397154 : 5 6
0.0421487 : 8 16
0.0493273 : 6 8
0.0576155 : 4 10
0.0584229 : 5 8
0.0648177 : 3 16
0.0801994 : 2 8
0.0811138 : 9 10
0.091207 : 8 9
0.0947807 : 11 13
0.0963619 : 9 13
```

```
  0.102799 : 1 2
  0.108346 : 0 15
  0.113945 : 3 9
  0.134232 : 7 9
  0.140258 : 1 9
  0.159117 : 8 11
  0.1709 : 7 15
  0.240142 : 10 11
  0.26317 : 12 15
  0.272496 : 7 12
  0.367329 : 0 12
  0.385538 : 7 14
  0.390066 : 1 8
  0.397235 : 0 7
  0.467406 : 12 14
  0.626967 : 3 8
  0.637598 : 1 3
  0.717924 : 9 11
  0.964653 : 0 14
Hierarchien:
 Dimensionen:
  Anzahl: 2
  ( 3 2 )
  ( 4 3 )
 Merkmale:
  Anzahl: 0
Redundante_Merkmale:
 Anzahl: 1
 14
Widerspruchsfreiheit: ja
```

C.2 Bereiche von Interesse

Die folgenden Bereiche von Interesse aus dem Metadatenformat sind folgendermaßen zu lesen: In runden Klammern sind jeweils die beiden Eckpunkte des Raumquaders im Format ($Dimension_1 \ldots Dimension_n$) gegeben. M : [] faßt die Menge an interessanten Raumquadern eines Bereiches von Interesse zusammen.

1. Bereiche von Interesse nach Länge und Breite der Messung (gradbrei minbrei gradlaen minlaen):

```
ROIS: Berechnet
  Anzahl: 12
     1 : [ ( 54.75 13.875 14 45.125 ) , ( 55 27.75 15 59.3 ) ]
     1 : [ ( 54.875 6.9375 13 52.2125 ) , ( 55 13.875 13.5 59.3 ) ]
     1 : [ ( 54.875 6.9375 12.5 52.2125 ) , ( 55 13.875 13 59.3 ) ]
     1 : [ ( 54 48.5625 13 52.2125 ) , ( 54.125 55.5 13.5 59.3 ) ]
     1 : [ ( 54 48.5625 12.5 52.2125 ) , ( 54.125 55.5 13 59.3 ) ]
     1 : [ ( 54 48.5625 13 23.8625 ) , ( 54.125 55.5 13.5 30.95 ) ]
     1 : [ ( 54 48.5625 12.5 23.8625 ) , ( 54.125 55.5 13 30.95 ) ]
     1 : [ ( 54.875 0 13 52.2125 ) , ( 55 6.9375 13.5 59.3 ) ]
     1 : [ ( 54.875 0 12.5 52.2125 ) , ( 55 6.9375 13 59.3 ) ]
     1 : [ ( 54 48.5625 13 9.6875 ) , ( 54.125 55.5 13.5 16.775 ) ]
```

```
      1 : [ ( 54 48.5625 12.5 9.6875 ) , ( 54.125 55.5 13 16.775 ) ]
      1 : [ ( 54 41.625 13 23.8625 ) , ( 54.125 48.5625 13.5 30.95 ) ]
```

2. Bereiche von Interesse nach dem Meßjahr und dem Meßmonat

```
ROIS: Berechnet
  Anzahl: 22
     1 : [ ( 93.3125 2.9375 ) , ( 94.1563 3.25 ) ]
     1 : [ ( 93.3125 10.75 ) , ( 95 11.375 ) ]
     1 : [ ( 74.75 3.875 ) , ( 75.5938 4.1875 ) ]
     1 : [ ( 80.6563 9.8125 ) , ( 81.5 10.125 ) ]
     1 : [ ( 85.7188 9.8125 ) , ( 86.5625 10.125 ) ]
     1 : [ ( 75.5938 9.8125 ) , ( 76.4375 10.125 ) ]
     1 : [ ( 84.875 10.75 ) , ( 85.7188 11.0625 ) ]
     1 : [ ( 90.7813 9.8125 ) , ( 91.625 10.125 ) ]
     1 : [ ( 77.2813 2.9375 ) , ( 78.125 3.25 ) ]
     1 : [ ( 81.5 2.9375 ) , ( 82.3438 3.25 ) ]
     1 : [ ( 90.7813 2.9375 ) , ( 91.625 3.25 ) ]
     1 : [ ( 82.3438 2.9375 ) , ( 83.1875 3.25 ) ]
     1 : [ ( 76.4375 9.8125 ) , ( 77.2813 10.125 ) ]
     1 : [ ( 92.4688 2.9375 ) , ( 93.3125 3.25 ) ]
     1 : [ ( 73.0625 9.5 ) , ( 74.75 10.125 ) ]
     1 : [ ( 81.5 9.8125 ) , ( 82.3438 10.125 ) ]
     1 : [ ( 82.3438 10.75 ) , ( 83.1875 11.0625 ) ]
     1 : [ ( 73.0625 2.625 ) , ( 74.75 3.25 ) ]
     1 : [ ( 92.4688 10.75 ) , ( 93.3125 11.0625 ) ]
     1 : [ ( 71.375 2.625 ) , ( 73.0625 3.25 ) ]
     1 : [ ( 71.375 9.5 ) , ( 73.0625 10.125 ) ]
     1 : [ ( 74.75 2.625 ) , ( 76.4375 3.25 ) ]
```

3. Bereiche von Interesse nach der Meßtiefe

```
ROIS: Berechnet
  Anzahl: 6
     1 : [ ( 4.3125 ) , ( 5.75 ) ]
     1 : [ ( 16.5313 ) , ( 17.25 ) ]
     1 : [ ( 15.8125 ) , ( 16.5313 ) ]
     1 : [ ( 9.34375 ) , ( 10.0625 ) ]
     1 : [ ( 20.125 ) , ( 21.5625 ) ]
     1 : [ ( 17.25 ) , ( 18.6875 ) ]
```

Erklärung

Hiermit erkläre ich, daß ich die vorliegende Arbeit selbständig und nur unter Vorlage der angegebenen Literatur und Hilfsmittel angefertigt habe.

Rostock, d. 28. 2. 2000

Thesen

1. Metadaten als Daten über eine Datenmenge können Visualisierungsentscheidungen unterstützen und dazu beitragen, daß aussagekräftige, für die Datenmenge geeignete Darstellungen erzeugt werden.

2. Es ist nicht möglich, eine alle Eigenschaften der Daten umfassende Metadatenspezifikation zu beschreiben, die in ihrem Umfang praktikabel umsetzbar ist. Beschränkt man sich jedoch auf wichtige Teilaspekte (in diesem Falle die Unterstützung von Visualisierungsentscheidungen und die Beschränkung auf Tabellen als Eingabeformat der Rohdaten), wird die Gewinnung der spezifizierten Metadaten praktikabel.

3. Metadaten lassen sich teilweise automatisch anhand von Berechnungen aus einer Datenmenge ableiten. Andere Metadaten erfordern eine interaktive Festlegung durch den Nutzer. Durch die Kombination von automatischen Metadatenbestimmungen und der Einbeziehung des Nutzers in die Metadatengewinnung kann eine maximale Aussagekraft der Metadaten und eine Effektivierung deren Bestimmung erreicht werden.

4. In dieser Arbeit wurden 8 Gruppen von Metadaten identifiziert und spezifiziert. Dies sind

   - die allgemeinen Variablen-Metadaten,
   - die Merkmals-Metadaten,
   - die Beobachtungsraum-Metadaten,
   - die allgemeinen Datenmenge-Metadaten,
   - die Beobachtungsfall-Metadaten,
   - die Strömungsdaten-Metadaten,
   - die Volumendaten-Metadaten und
   - die Multiparameterdaten-Metadaten.

5. Durch die Festlegung einer geeigneten Reihenfolge der Metadatengewinnungen können diese effektiv bestimmt werden.

6. Es wurde ein Werkzeug entwickelt, das die Gewinnung von Metadaten unter Einbeziehung des Nutzers unterstützt. Praktische Tests anhand realer Datensätze haben gezeigt, daß die Metadatengewinnung mit den zur Verfügung stehenden Ressourcen in aktzeptabler Zeit durchgeführt werden kann.

7. Das entwickelte Programm stellt eine praktikable Arbeitsgrundlage dar. Die Spezifikation und Erhebung weiterer Metadaten wurde vorgesehen. Vielfältige Erweiterungsmöglichkeiten sind damit denkbar.