

VisInfo – Visueller Zugang zu Forschungsdaten

Foto: privat



Oliver Koepler

Foto: TIB Hannover



Irina Sens

neue Wissensobjekte
ergeben neue Frage-
stellungen für Suchen,
Finden, Präsentieren

»Big data« und »Data-driven Science« sind Schlagwörter, die an forschende Bibliotheken neue Herausforderungen stellen. Große Mengen Forschungsdaten zu visualisieren und visuell suchbar zu machen, ist in dem Projekt »Visueller Zugang zu Forschungsdaten« untersucht worden. Zentrales Element des VisInfo Prototyps ist der Visuelle Katalog, der nach Erzeugung des Suchraums über die Forschungsdaten einen Überblick über charakteristische Kurvenpattern in der Datenkollektion ermöglicht. Projektpartner sind das Fraunhofer-Institut für Graphische Datenverarbeitung und der Fachbereich Informatik der TU Darmstadt, Fachgebiet Graphisch-Interaktive Systeme.

»Big data« and »data-driven science« are key words that in researching libraries face new challenges. To visualize large amounts of research data and visually make searchable has been studied in the project »Visual access to research data«. The central element of the prototype is the Visual Visinfo catalog that allows after generating the search space on the research data an overview of characteristic curve pattern in the data collection. Project partners are the Fraunhofer Institute for Computer Graphics and the department of computer science at the TU Darmstadt, Group of Interactive Graphics.

EINLEITUNG

»Eine Gesamtstrategie für die wissenschaftlichen Informationsinfrastrukturen in Deutschland muss sich zuvorderst an dem Ziel ausrichten, Wissenschaftlerinnen und Wissenschaftlern sowie Studierenden das von ihnen benötigte und verfügbare Wissen sowie die für ihre Arbeit bzw. Bildung und Ausbildung erforderlichen Daten, Informationen und Informationsressourcen umfassend und schnell zur Verfügung zu stellen. Dies muss in einer Form geschehen, die den sich schnell wandelnden Anforderungen der unterschiedlichen Forschungsformen [...] wie auch der wissenschaftlichen Querschnittsfunktionen gerecht wird und dabei alle Bereiche der Wissenschaft – Forschung, Studium, Lehre, Nachwuchsförderung, Technologieentwicklung und Wissenstransfer – einbezieht.«¹

Die Möglichkeiten des Internets führen zu einem Wandel in der wissenschaftlichen Kommunikation – insbesondere in den datengetriebenen (Natur)Wissenschaften. Mündete die wissenschaftliche Wertschöpfungskette in der Regel darin, einen Artikel zu veröffentlichen und wurden die zugrundeliegenden Messdaten selten hinterlegt, so werden heute parallel zum Artikel die zugehörigen Forschungsdaten in einem zertifizierten Datenzentrum oder ein Video des zugrundeliegenden Experimentes auf einem Videoportal wie beispielsweise Beilstein TV² publiziert. Je nach Fachgebiet werden Erkenntnisse und Ergebnisse über chemische Strukturformeln, animierte Prozess-Visualisierungen (Technik) oder 3D-Modelle (Ar-

chitektur) kommuniziert. Wissenschaftliche Kommunikation wird dadurch transparenter und verständlicher. Die Registrierung von Forschungsdaten, aber auch von Videomaterialien mit Persistent Identifiers wie DOIs ermöglicht die langfristige Verfügbarkeit, Zitierbarkeit und vor allem die eindeutige Verknüpfung unterschiedlichster Dokumenttypen zu einer Informationseinheit.

Mit der Diversifizierung der publizierten Wissensobjekte ergeben sich neue Fragestellungen für das Suchen, Finden und Präsentieren. Das Suchen in bibliographischen Metadaten wird erweitert um inhaltsbasiertes Suchen, d.h. die Suche erfolgt innerhalb eines Objektes.

Die Technische Informationsbibliothek (TIB) als Zentrale Fachbibliothek für Technik sowie Architektur, Chemie, Informatik, Mathematik und Physik hat diese Entwicklungen frühzeitig erkannt und gestaltet sie aktiv mit. Mit dem Kompetenzzentrum für nicht-textuelle Materialien (KNM) stellt sich die TIB den Herausforderungen vielschichtiger nicht-textueller Dokumenttypen und den damit verbundenen Anforderungen an eine moderne, leistungsfähige wissenschaftliche Bibliothek. Das Probado-Portal³ für 3D-Modelle aus der Architektur oder das sich in der Entwicklung befindliche AV-Portal⁴ für Videomaterialien sind Dienstleistungen, die neue Wege bei der inhaltlichen Erschließung, Indexierung, Suche und Präsentation von nicht-textuellen Materialien gehen. Im Verbund mit den klassischen Diensten rund um Textdokumente bietet die TIB mit ihrem Fachportal GetInfo⁵ einen umfassenden Zugang zu den neuen Dokumenttypen der wissenschaftlichen Kommunikation.

AUSGANGSSITUATION UND MOTIVATION

Forschungsdaten stehen am Anfang der bereits erwähnten wissenschaftlichen Wertschöpfungskette und stellen das Fundament des wissenschaftlichen Erkenntnisprozesses dar. Forschungsdaten untermauern Hypothesen, Theorien und Interpretationen. Sie sind damit ein sehr wertvolles Gut für die Wissenschaft, das bewahrt werden muss. So gibt es in einigen Disziplinen, insbesondere in den »Data-driven Sciences« etablierte Forschungsdaten-Infrastrukturen mit Datenzentren, die die langfristige Speicherung und Verfügbarkeit von Forschungsdaten sicherstellen. Versehen mit einer DOI sind solche Forschungsdaten damit

auch zitier- und wiederfindbar.⁶ Viele dieser Datenzentren stellen ihre Forschungsdaten über Webdatenbanken den Wissenschaftlerinnen und Wissenschaftlern zur Verfügung. Paradebeispiel sind die Weltdatenzentren. 1957 richtete die Vorläuferorganisation des Internationalen Wissenschaftsrats (ICSU) die Weltdatenzentren für die Sammlung und Verarbeitung der Daten aus dem Internationalen Geophysikalischen Jahr ein. Heute archivieren die weltweit verteilten Zentren Daten aus allen Bereichen der naturwissenschaftlichen Grundlagenforschung. Bei sehr gut organisierten Datenzentren wie der PANGAEA Datenbibliothek⁷ des Alfred-Wegener-Instituts (AWI), Bremerhaven für Forschungsdaten aus den Erd- und Umweltwissenschaften sind die Forschungsdaten mit umfangreichen Metadaten beschrieben und suchbar.

Die PANGAEA Datenbibliothek enthält ein breites Spektrum an Forschungsdaten aus den Bereichen Wasser, Sedimente, Eis oder Atmosphäre, die Messungen zu Parametern wie Temperatur, Luftfeuchtigkeit oder Sonneneinstrahlung über die Zeit enthalten. Ein Datensatz besteht in sehr vielen Fällen aus einer einfachen ASCII-Datei mit numerischen Daten in Tabellenform, ergänzt mit Metadaten zu Zeitpunkt, Ort und Bedingungen der Erhebung und anderem mehr. In den Metadaten eines Forschungsdatensatzes kann so z. B. nach Ort, Zeitpunkt, Parameter, Autor oder Projekt gesucht werden. Einmal gefunden, kann ein Datensatz meist in seiner originären Form vom Datenzentrum heruntergeladen werden. Schaut man einer Wissenschaftlerin oder einem Wissenschaftler über

die Schulter, wie sie/er mit solchen Forschungsdaten weiter verfährt, so ist festzustellen, dass die umfangreichen Zahlenkolonnen über Softwarewerkzeuge als Kurven, Graphiken oder in sonstiger Form visualisiert werden, bevor sie in den wissenschaftlichen Analyseprozess Eingang finden. Die Visualisierung, die Präsentation von Messdaten als Kurven ist ein entscheidender Schritt in der Datenanalyse. Die Bedeutung der visuellen Analyse von Daten verdeutlicht Abbildung 1. Bei der Betrachtung der Zahlentabelle links ist die charakteristische Aussage, die in den Daten steckt, nur schwer zu erkennen. Ein kurzer Blick auf die Visualisierung der Tabelle als Kurvenverlauf erlaubt eine sofortige Erfassung der Daten mit einem hohen Peak von 35°C nach sechs Stunden.

Gehen wir einen Schritt zurück und betrachten das Informationsbedürfnis von Wissenschaftlern zu Beginn ihrer Suche in einer Forschungsdaten-Bibliothek. Eine Fragestellung mag die nach Daten aus einer bestimmten Region oder einem bestimmten Zeitabschnitt sein. Dies lässt sich durch eine Suche in den Metadaten realisieren. Was aber ist mit Fragestellungen, die sich mit den Daten selber beschäftigen? Wie könnte man eine Suche absetzen, die beispielsweise Datensätze mit einem starken Temperaturanstieg in den frühen Morgenstunden eines Tages findet? Wie lässt sich eine Suche direkt in den Daten, also eine inhaltsbasierte Suche formulieren und darstellen? Bei Textdokumenten sprechen wir in solchen Fällen von der Volltextsuche, mit der sich das Auftreten von Worten oder Phrasen in Textdokumenten ermitteln lässt.

Visualisierung numerischer Forschungsdaten

inhaltsbasierte Suche in Forschungsdaten

Zeit [h]	T [°C]
1	12
2	13
3	12
4	12
5	13
6	35
7	17
8	11
9	10
10	12
11	13
12	13
13	12
14	12
15	12
16	11
17	11
18	10
19	10
20	11
21	11
22	10
23	12
24	12

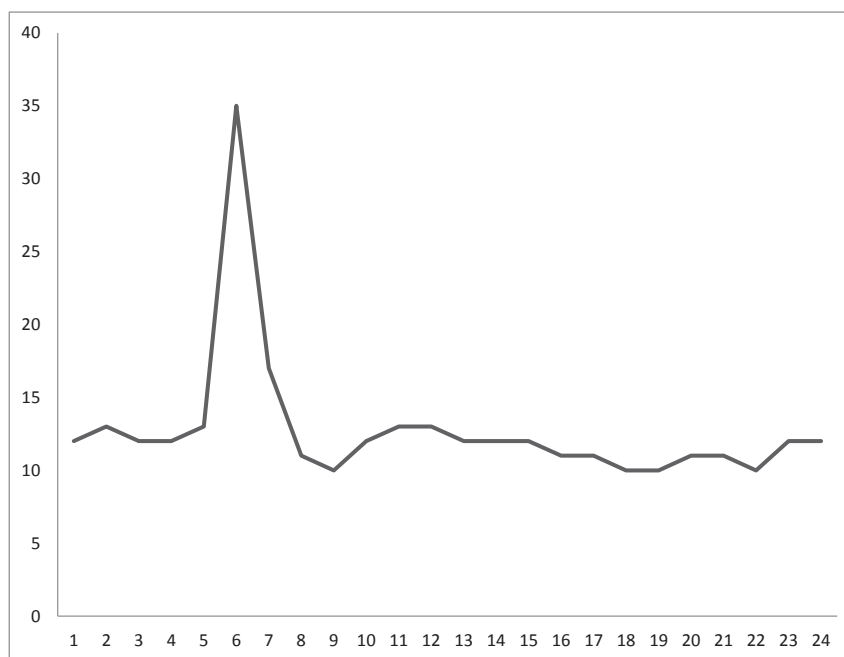


Abb. 1: Die Visualisierung numerischer Daten ermöglicht das intuitive Erkennen von Kernaussagen

Lässt sich dieses Konzept auf die Suche in numerischen Forschungsdaten übertragen? Reicht die Suche nach 35°C um 10 Uhr aus, um obige Frage zu beantworten? Wohl eher nicht.

An dieser Stelle setzt die Idee des VisInfo Projektes an. Was wäre, wenn man die Visualisierung bei der Präsentation von Messdaten auf eine Visualisierung bei der Formulierung der Suche in den Messdaten überträgt? Die Nutzerin, der Nutzer zeichnet über ein graphisches Interface ihre/seine »Suchanfrage« als Kurvenverlauf, initiiert damit eine Suche in Forschungsdaten und erhält als Ergebnis eine Liste mit relevanten Kurven zurück. Eine inhaltsbasierte, visuelle Suche in Forschungsdaten, kombiniert mit der visualisierten Präsentation der Treffer einer solchen Suche wäre auf diese Weise Realität. Natürlich ist die Kombination dieser visuellen Suche mit den bekannten Metadaten vorstellbar. So kann die Suche im Vorfeld auf bestimmte geographische Regionen oder Zeiträume eingeschränkt werden oder im Nachhinein die Trefferliste nach geographischen Regionen oder Zeiträumen gefiltert werden. Ein solches Suchkonzept ist für viele Neuland: für Entwickler, Bibliotheken, aber insbesondere auch für die Nutzerschaft.

VISUELLE SUCHE, VISUAL ANALYTICS, FORSCHUNGSUMFELD

Das VisInfo Projekt ist ein gemeinsames Forschungsprojekt des Fachgebietes Graphisch-Interaktive Systeme (GRIS) an der TU Darmstadt, des Fraunhofer-Instituts für Graphische Datenverarbeitung (IGD), Darmstadt und der Technischen Informationsbibliothek (TIB), Hannover. Die Tätigkeitsschwerpunkte der Projektpartner lagen in der Forschung und Entwicklung visueller Such- und Präsentationsverfahren für zeitbasierte Forschungsdaten (GRIS), der Entwicklung eines prototypischen Such- und Präsentationsystems (Fraunhofer IGD) und der Implementierung und Evaluierung des Systems in die Angebote einer Bibliothek (TIB).

Ziel des VisInfo Projektes war eine schon frühe Einbindung der anvisierten Nutzergruppen in den Entwicklungsprozess. Hierfür konnte das Alfred-Wegener-Institut für Polar- und Meeresforschung (AWI) in Bremerhaven gewonnen werden. Das AWI betreibt, wie oben erwähnt, die Forschungsdaten-Bibliothek

PANGAEA für Daten aus dem Bereich Erd- und Umweltwissenschaften. Für das VisInfo Projekt wurde mit einem Datensubset des Baseline Surface Radiation Networks (BSRN) gearbeitet. Für den konkreten Fall BSRN besteht die Zielgruppe aus Wissenschaftlerinnen und Wissenschaftlern, die auf der einen Seite schon jetzt selbst Forschungsdaten produzieren und in etablierte Forschungsdaten-Infrastrukturen wie die PANGAEA Datenbibliothek einbringen. Auf der anderen Seite nutzen diese Wissenschaftlerinnen und Wissenschaftler die verfügbaren Datenbanken auch zur Suche nach Forschungsdaten für eigene Forschungsarbeiten. Die Zielgruppe vereint daher sowohl die Produzenten- als auch die Konsumentensicht bei der Bewertung des VisInfo Systems.

DATENAUFBEREITUNG

Für die Entwicklung des VisInfo Prototypen kamen Forschungsdaten aus dem Baseline Radiation Network⁸ (BSRN), einer Teilmenge der PANGAEA Datenbibliothek zum Einsatz. Die BSRN-Daten wurden aufgrund der hohen Datenqualität und -quantität, der Verfügbarkeit von normierten Metadaten zu jedem Datensatz und der aktiven Forscher-Community als Testdaten ausgewählt. Das Daten-Set bestand aus 14.331 Daten-Dateien. Die Datentabellen bestanden aus bis zu 100 Spalten (Zeitreihen) und bis zu 50.000 Mess-Einträgen. Die BSRN-Datensätze enthalten überwiegend Messreihen zur Strahlung (short-wave, long-wave, diffuse, direct), Temperatur, Luftfeuchtigkeit und Wind (Geschwindigkeit und Richtung). Jede Daten-Datei deckte eine Zeit von einem Monat ab.

Für erste Untersuchungen wurden ca. 2.500 Datensätze mit Strahlungsdaten aufbereitet und mit einem Daten-Parser in das VisInfo System importiert. Die interne Datenstruktur des VisInfo Systems ist so angelegt, dass zukünftig weitere Daten-Parser für andere Datenformate aus weiteren Quellen modular entwickelt werden können. Beim Preprocessing der Daten werden fehlende Messwerte durch Interpolation ergänzt und Zeitangaben auf eine globale, standardisierte Zeitachse angepasst. Für die Fragestellungen rund um Zeitserien von Strahlungsdaten wird eine Betrachtung von Tagespattern als sinnvoll erachtet und die Daten entsprechend in 24-Stunden-Zeitintervalle, die jeweils einen Tagesverlauf repräsentieren, aufge-



Abb. 2: Prozess des Datenimports von Forschungsdaten für die visuelle Suche

Visualisierung bei der Formulierung der Suche

Forschungsprojekt von TIB, GRIS und Fraunhofer IGD

splittet. Diese Aufteilung ergab über 70.000 Tagesverläufe unterschiedlichster Messreihen.

Die weitere Datenaufbereitung beinhaltet die Berechnung der Feature-Vektoren (Deskriptoren), für den Prototypen kam hier der Piecewise Aggregate Approximation (PAA) Deskriptor⁹ zum Einsatz. Die extrahierten Deskriptoren werden in einem Index gespeichert. Dieser Index ist in der Regel um ein vielfaches kleiner als die Originaldaten. Im letzten Schritt wird noch das Ähnlichkeitsmaß definiert. Für erste Untersuchungen kommt die euklidische Distanz zum Einsatz. Neben den eigentlichen Daten wurden auch zugehörige Metadaten wie Autor, Projekt, Messparameter, Jahr und andere importiert und indiziert.

Die kombinierte Suche in Daten und Metadaten stellt einen erheblichen Mehrwert des Systems dar und ermöglicht eine granulare Treffererschließung im Suchprozess.

VISINFO SYSTEM

Das VisInfo System bietet der Nutzerschaft zwei grundsätzliche Einstiege in den visuellen Information Retrieval Prozess:¹⁰ Auf der einen Seite der Query Editor mit einem graphischen Suchinterface, das das Zeichnen eines Kurvenverlaufs und die Suche danach ermöglicht; auf der anderen Seite der Visuelle Katalog, der einen interaktiven, explorativen Zugang zu dem verfügbaren Datenraum bietet.

Der Visuelle Katalog

Der Visuelle Katalog greift Shneiderman's Information Visualization Mantra¹¹ »overview first, zoom and filter, then details on demand« auf und ermöglicht es der Nutzerschaft, sich einen ersten Überblick über charakteristische Kurvenverläufe in dem zu durchsuchenden Datenraum zu verschaffen.

Für diesen ersten Überblick über eine große Menge von Daten muss eine Reduktion der zu visualisierenden Zeitserien-Datensätze und eine gesamtheitliche, intuitive Darstellung gefunden werden. Die Antwort auf diese Herausforderung ist die Anwendung eines visuellen Cluster-Algorithmus in einer sogenannten Self-Organizing Map (SOM). Der SOM-Algorithmus¹² ist im Bereich der Visuellen Analyse großer Datenmengen und deren Reduktion auf wesentliche Aussagen gut etabliert. Basierend auf den während der Datenaufbereitung extrahierten Deskriptoren der Zeitserien werden aus beliebig vielen Datensätzen charakteristische Kurven-Pattern berechnet und in einer Zellen-Matrix dargestellt. Die Zellen der Matrix werden wiederum nach den Ähnlichkeiten ihrer Kurvenpattern sortiert und angeordnet und ergeben dann den Visuellen Katalog.

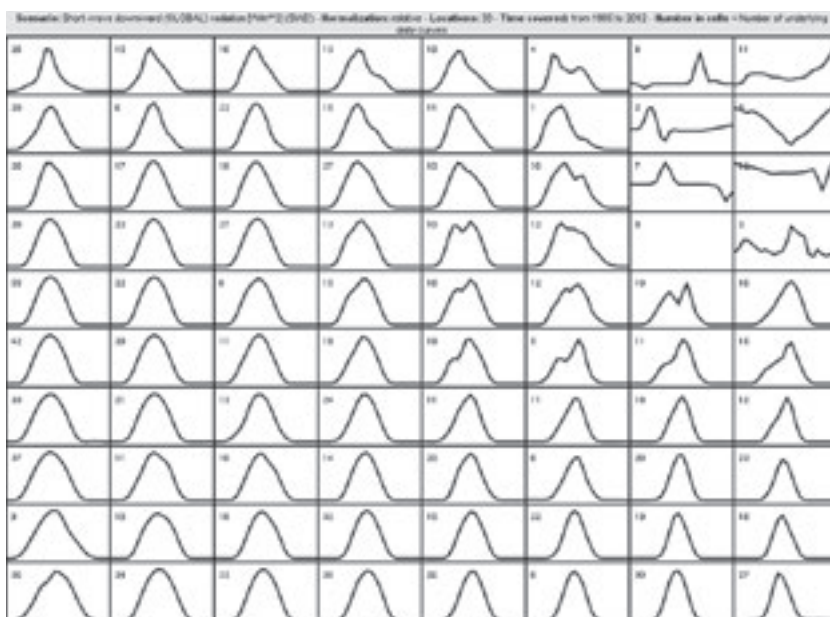


Abb. 3: Visueller Katalog mit Gesamtübersicht des indizierten Datenbestandes

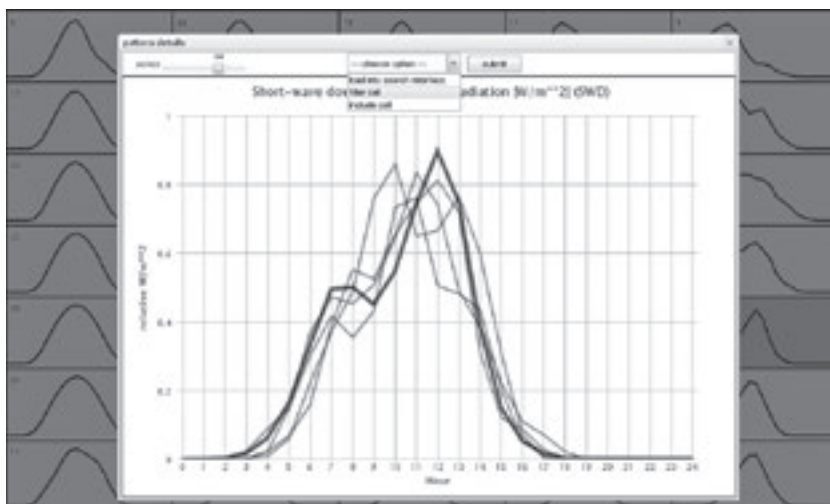


Abb. 4: Detailansicht einer Zelle des visuellen Katalogs mit repräsentativem Kurvenpattern und tatsächlichen Datenkurven

Abbildung 3 zeigt einen solchen Visuellen Katalog mit Zeitserien von BSRN-Daten für eine inhaltsbasierte, explorative Suche. Dargestellt werden hier zeitbasierte Messungen der short-wave downwards (SWD)-Strahlung im Zeitraum 1995 bis 2012. Jede Zelle enthält ein charakteristisches Kurvenpattern eines Tagesverlaufs (0–24 h), das eine beliebige Anzahl von Datensätzen repräsentiert. Die Anzahl der Forschungsdatensätze, die durch eine Zelle repräsentiert werden, wird durch die Zahl links oben in der Zelle angezeigt, so enthält die erste Zelle links 26 Forschungsdatensätze.

Abbildung 4 zeigt eine detaillierte Ansicht einer Zelle des Visuellen Katalogs. Der breit gezeichnete Kurvenverlauf stellt das abstrahierte, charakteristische Kurvenpattern der Zelle dar. Die dünn gezeichneten Kurvenverläufe stellen den tatsächlichen Kur-

explorative Suche mit dem Visuellen Katalog

venverlauf der Zeitserien-Datensätze dar, die durch das Kurvenpattern der Zelle repräsentiert werden. Bei sehr typischen Kurvenverläufen wie beispielsweise einer Gauß'schen Verteilung liegen die Kurvenverläufe des charakteristischen Kurvenpatterns und der repräsentierten Datensätze naturgemäß nah übereinander.

Der Query-Editor

Zeichnen einer Suchkurve mit dem Query Editor

Der zweite Sucheinstieg in das VisInfo System ist der Query-Editor (Abb. 5), der die unmittelbare, interaktive Formulierung von visuellen Suchanfragen als gezeichnete Kurve durch die Nutzerin, den Nutzer ermöglicht. Dargestellt ist hier ein Suchszenario mit der sogenannten short-wave downwards radiation (relative Werte) und einer Betrachtung von Tagespattern mit einer Auflösung von einer Stunde. Der Query-Editor ermöglicht über das Zeichenpanel die freie Definition einer Suchkurve für eine inhaltsbasierte Suche. Neben der dargestellten freien Formulierung einer Suchkur-

ve besteht auch die Möglichkeit, den Visuellen Katalog und den Query-Editor kombiniert zu nutzen. So kann ein Kurvenpattern aus dem Visuellen Katalog in das Zeichenpanel übertragen werden und in modifizierter oder unmodifizierter Form als Suchkurve benutzt werden.

Die Suche lässt sich durch Einbeziehung der Metadaten zu den Forschungsdatensätzen weiter konkretisieren. Über Geo-Filter, Time-Filter oder Metadata-Filter kann die Suche nach einem Kurvenverlauf auf eine geographische Region, ein Zeitintervall oder einen Wert eines Metadatenfeldes (z.B. Autor) eingeschränkt werden. So ist es beispielsweise möglich, die Suche auf Daten zu beschränken, die in den Jahren 1990 bis 2010 in Südamerika erhoben wurden.

Treffer einer inhaltsbasierten Suche

Die Ergebnisse einer inhaltsbasierten Suche in den Forschungsdaten werden wiederum visualisiert; nämlich als Trefferliste von Kurvenverläufen. Abbildung 6 zeigt die Trefferliste einer Suche, die im Query-Editor formuliert wurde. Das Fenster der Trefferliste ist aufgeteilt in drei Bereiche. In der Mitte sind die Forschungsdatensätze als Kurvenverläufe visualisiert. Wie eingangs erwähnt, umfasst ein BSRN-Datensatz Messungen von 30 Tagen, die hier in ihrer Gesamtheit dargestellt werden. Als Balken markiert sind nun die Treffer in Tagesverläufen (die Suche bezog sich auf einen 24-stündigen Zeitabschnitt), wobei die Farbsättigung die Güte bei der Übereinstimmung zwischen Suchkurve und Trefferkurve repräsentiert.

Rechts neben den Kurvenverläufen werden die Metadaten des Forschungsdatensatzes mit Informationen zu Creator, Contributor, Publisher, Publication Year und mehr angezeigt. Das verwendete Metadatenschema ist eng angelehnt an das Schema von PANGAEA. Über die Angabe der DOI ist ein unmittelbarer Zugriff auf den originären Forschungsdatensatz beim entsprechenden Datenzentrum möglich. Das VisInfo System versteht sich als reines Suchsystem, die originären Forschungsdaten werden über das verantwortliche Datenzentrum zur Verfügung gestellt.

Links neben den Kurvenverläufen werden unterschiedliche Facetten für eine weitergehende Treffererschließung angezeigt. So können Treffer, ganz ähnlich den Filtern vor Beginn der Suche, im Nachhinein nach geographischen Regionen oder Zeit eingegrenzt werden.

Über die Map View oder Time View besteht ferner die Möglichkeit, die Trefferliste nach anderen Merkmalen zu visualisieren. Die Map View zeigt die Verteilung der Treffer nach ihrer Herkunft auf einer Weltkarte an. Die Time View präsentiert die Verteilung der Treffer auf



Abb. 5: Eingabemaske Query Editor zum Zeichnen einer visuellen Suche

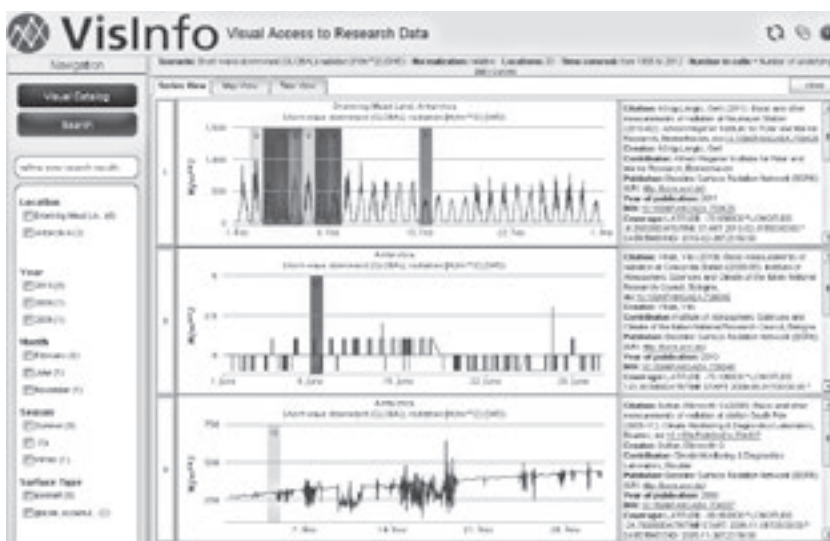


Abb. 6: Trefferliste einer visuellen Suche, ebenfalls als Kurven visualisiert

einer Zeitmatrix und zeigt z. B. an, dass besonders viele Treffer im Frühjahr eines Jahres gemessen wurden.

Ähnlichkeit von Kurven

Bei der Suche in Textdokumenten sind die Kriterien für »ähnliche Dokumente« vertraut, Maße wie Term-Frequency und Inverse-Term-Frequency zur Beurteilung der Relevanz von Dokumenten bekannt. Für die visuelle, inhaltsbasierte Suche in Forschungsdaten ist die Betrachtung und Definition von Ähnlichkeitsmaßen notwendig. Wie ist eine Suche zu gestalten, um eine Treffermenge mit »ähnlichen« Kurven zu erhalten? Was charakterisiert diese Ähnlichkeit, die Relevanz eines Treffers? Abbildung 7 zeigt unterschiedliche Interpretationsmöglichkeiten von »ähnlichen« und damit relevanten Kurvenverläufen. Ist eine Kurve noch relevant, wenn sie den gewünschten Verlauf der Werte zeigt, aber zeitlich verschoben ist? Ist ein Kurvenverlauf von Interesse, wenn er den gewünschten Verlauf zeigt, aber auf der Werteachse verschoben ist? Wie sieht es mit zeitlich gestreckten Kurvenverläufen aus?

Für die Betrachtung der Ähnlichkeitsmaße im VisInfo System kommen die bereits erläuterten Feature-Vektoren bzw. Deskriptoren zur Anwendung, die aus den zugrundeliegenden Zeitserien berechnet und extrahiert werden. Die Ähnlichkeit von zwei Zeitserien (Kurven) ergibt sich dann aus dem berechneten Abstand der beiden Feature-Vektoren. Für Zeitserien lassen sich unterschiedliche Deskriptoren berechnen, die

dann zu unterschiedlichen Ähnlichkeitsmaßen führen. Für den VisInfo Prototyp wurde ein Ähnlichkeitsmaß basierend auf der euklidischen Distanz eingesetzt. Die Ergebnisse der Evaluierungen deuten aber bereits an, dass zukünftig differenzierte Suchszenarien mit variabel anpassbaren Ähnlichkeitsmaßen angeboten werden sollten.

PROTOTYP EVALUIERUNG, USABILITY TESTS

Online-Umfrage

Die Zusammenarbeit mit dem AWI hatte schon früh gezeigt, dass Diskussionen über das Thema Visuelle Suche in Forschungsdaten und den VisInfo Prototypen um einiges leichter geführt werden konnten, wenn diese visuell aufbereitet wurden. Da die anvisierte wissenschaftliche Nutzergruppe des Baseline Surface Radiation Network weltweit verteilt ist, wurde für eine Einführung und Präsentation des VisInfo Projektes bei den BSRN-Nutzern ein Screencast erstellt, der die wesentlichen Ideen der Visuellen Suche und die angedachte Funktionsweise des VisInfo Prototypen mit Visuellem Katalog, Query Editor und visueller Präsentation der Trefferliste in Bild und Ton darstellte. Neben dem Screencast wurde optional eine Software-Dokumentation mit einer Beschreibung des Systems zur Verfügung gestellt. Gemeinsam mit dem AWI wurden Wissenschaftlerinnen und Wissenschaftler, die

Ähnlichkeitsmaße bei Kurven

Nutzerbefragung mit Screencast

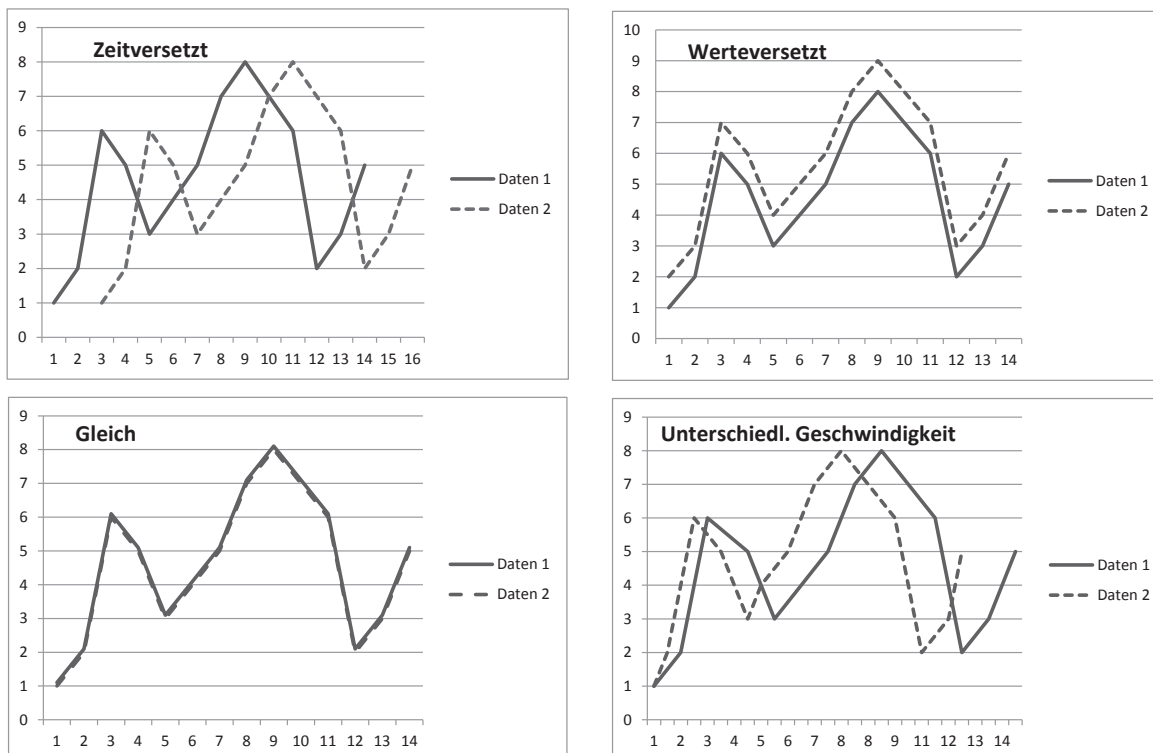


Abb. 7: Was bedeutet Ähnlichkeit bei der Suche nach Kurven?

Ergebnisse der Online-Umfrage

BSRN-Daten aktiv nutzen, per E-Mail angeschrieben und gebeten, nach Betrachten des Screencasts und der Informationsmaterialien einen Online-Fragebogen zu VisInfo auszufüllen. Der Fragebogen deckte neben allgemeinen Fragen zur bisherigen Datennutzung und Schwerpunkten von Fragestellungen an Forschungsdaten insbesondere Fragen zur Anwendbarkeit des VisInfo Systems und dessen geplanten Funktionalitäten bei der Arbeit der Wissenschaftlerinnen und Wissenschaftler ab.

Die Online-Umfrage ergab eine grundsätzlich positive Resonanz auf die Darstellung des Datenraums als Visuellem Katalog, die Formulierung visueller Suchanfragen mit dem Query Editor und die visuelle Darstellung von Trefferlisten. Ferner wurde die Verknüpfung der visuellen Suche mit der bisherigen Suche in Metadaten als wertvoller Mehrwert angesehen. Die Umfrage ergab auch, dass die Fragestellungen der Wissenschaft sehr variabel sind und daher vielfältige Ausprägungen von Suchszenarien hinsichtlich interessanter Parameter (Temperatur, Luftfeuchtigkeit, Strahlung), Auflösung der Zeitachse, Ähnlichkeitsmaße etc. zu erwarten sind.

Die Ergebnisse der Online-Umfrage flossen dann in die weitere Entwicklung des VisInfo Prototypen ein.

Usability Tests VisInfo Prototyp

Usability Testing

Eine intensive Evaluierung des VisInfo Prototypen aus Sicht der Usability konnte während des 12. BSRN Workshops vom 1. bis 3. August 2012 in Potsdam durchgeführt werden. Der Workshop wurde von der Forschungsstelle Potsdam des Alfred-Wegener-Instituts organisiert. Insgesamt waren knapp über 60 Teilnehmerinnen und Teilnehmer aus dem BSRN-Umfeld vor Ort. Während des dreitägigen Workshops wurden in einem separaten Raum taskbasierte Usability Tests durchgeführt. Die Schwerpunkte lagen hier auf der Gestaltung der Benutzeroberfläche und der ausführbaren Aktionen, um die Gebrauchsfähigkeit des

VisInfo Prototypen zu testen. Zum Zeitpunkt der Usability Tests lag das VisInfo System bereits als webbasiertes Suchportal vor, das mit einem Browser genutzt werden konnte. Für den Test saß die Probandin, der Proband vor einem Laptop mit dem VisInfo System, daneben saß der Testleiter, der bei Bedarf weitere Erläuterungen geben konnte und zusätzliche Fragen stellte. Für die Tests wurden die Klickpfade der Probanden auf der Benutzeroberfläche erfasst sowie das Verhalten und die Kommentare der Teilnehmenden mit Video/Audio aufgenommen. Die Auswertung der Tests erfolgte anonymisiert. Nach einer Warm-Up-Phase zum Kennenlernen des VisInfo Systems wurden den Probanden insgesamt vier Aufgaben gestellt, die diese mit dem VisInfo System lösen sollten. Die Probanden wurden aufgefordert, nach dem Prinzip des Thinking-Aloud ihre Eindrücke in der Interaktion mit dem Suchportal zu beschreiben.

Die Zeit zum Lösen der Tasks wurde erfasst und die Erfolgsquote auf einer Skala von »erfolgreich«, »mit Problemen, aber erfolgreich«, »erfolgreich, mit Hilfe des Testleiters«, »nicht erfolgreich« sowie »Abbruch durch Testleiter« bewertet. Die Probandin, der Proband wurde im Anschluss um eine Bewertung der Schwierigkeit der Tasks auf einer Skala von -3 bis +3 sowie weiteren Kommentaren gebeten.

Insgesamt konnten während des BSRN-Workshops 21 Probanden gewonnen werden, wovon zwei den Test aufgrund größerer Schwierigkeiten und des daraus resultierenden Zeitmangels abbrechen. Offensichtlich hatten Jüngere weniger Probleme mit dem System und fanden schneller Lösungen für die gestellten Tasks. Bei Problemen suchten sie länger als ältere Kolleginnen und Kollegen nach Alternativen, was auf die vermutlich höhere Internet- und Computeraffinität zurückzuführen ist. Die Zeit für das Absolvieren eines Tasks schwankte zwischen 30 Sekunden und 8 Minuten, wobei 8 Minuten die maximal zur Verfügung stehende Zeit darstellte.

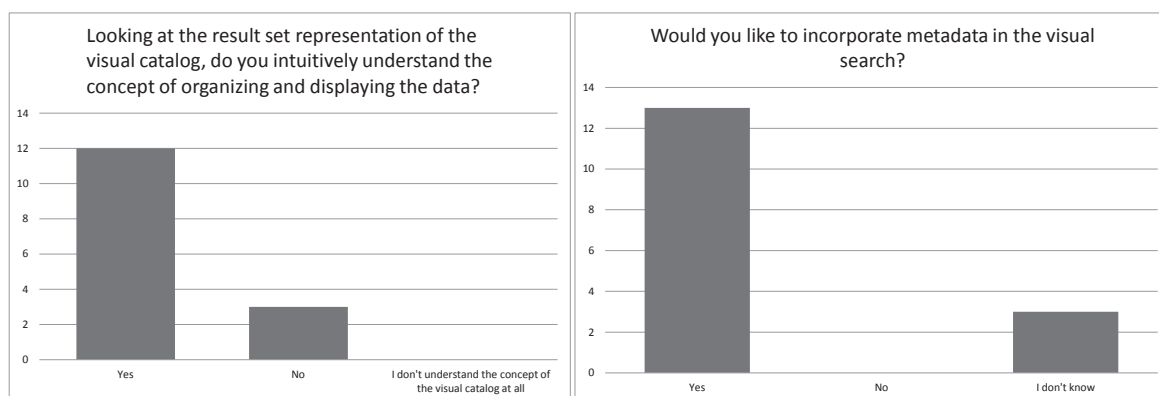


Abb. 8: Ergebnisse der Online-Umfrage

Das Konzept der visuellen Suche mittels des Zeichnens einer Kurve im Query-Editor war vielen Probanden nicht unmittelbar verständlich. So gab es immer wieder Irritationen hinsichtlich der Bedienung des Query-Editors, insbesondere was das konkrete Zeichnen einer Kurve mit der Maus betraf. Bei den Tasks kamen zahlreiche Unstimmigkeiten bei der Reihenfolge von Arbeitsschritten oder auch der Benennung und der Anordnung von Seitenelementen und Schaltflächen zu Tage. Aus diesen Ergebnissen konnte eine Reihe von Gestaltungsempfehlungen für die Benutzeroberfläche abgeleitet werden, die in die Entwicklung des nächsten Prototypen einfließen. Dieser ist im Dezember 2012 einem weiteren Usability Test unterzogen worden.

In Bezug auf die Nützlichkeit und Anwendbarkeit des Systems gab es positive und negative Reaktionen. Viele Teilnehmerinnen und Teilnehmer fanden das System grundsätzlich interessant, konnten sich aber spontan für ihre Arbeit keinen konkreten Anwendungsfall vorstellen. In späteren Gesprächen mit den Teilnehmerinnen und Teilnehmern des Workshops ließ sich jedoch feststellen, dass im Nachgang doch die eine oder andere Idee für den Einsatz des VisInfo Systems aufkam. Dies geht einher mit der grundsätzlichen Beobachtung, dass das Konzept der Visuellen Suche in Forschungsdaten bisher nicht im Repertoire der bekannten Suchwerkzeuge vorhanden ist und eine gewisse Zeit braucht, um zu »sacken«.

ZUSAMMENFASSUNG

Im Rahmen des VisInfo Projektes wurden Verfahren zur visuellen Suche in und zur Präsentation von zeitreihen-basierten Forschungsdaten entwickelt und in einem prototypischen, webbasierten Suchportal umgesetzt. Die Entwicklung des Suchportals erfolgte begleitend mit Usability-Evaluierungen, um die Anforderungen der Nutzerschaft frühzeitig mit einbeziehen zu können. Die inhaltsbasierte Suche in Forschungsdaten wird zukünftig die Suchmöglichkeiten nach Forschungsdaten im Fachportal GetInfo der Technischen Informationsbibliothek Hannover erweitern. War dort bisher die Suche nach Forschungsdaten auf Basis der Metadaten möglich, sollen diese Dienste mittelfristig um die inhaltsbasierte Suche in den Daten über eine Anbindung des VisInfo Systems erweitert werden.

Das VisInfo Projekt »Visuelle Suche in Forschungsdaten« wurde im Rahmen des Paktes für Forschung

und Innovation vom 1. Januar 2010 bis zum 31. Dezember 2012 durch die Leibniz-Gemeinschaft gefördert.

DANKSAGUNG:

Wir möchten uns bei Jürgen Bernard, Tobias Ruppert und Jörn Kohlhammer vom Fraunhofer IGD, Darmstadt für die hervorragende Zusammenarbeit im Projekt bedanken. Großer Dank gilt weiterhin Gert König-Langlo vom Alfred-Wegener-Institut, Bremerhaven sowie den Wissenschaftlerinnen und Wissenschaftlern des BSRN Netzwerkes für die Bereitschaft zu Usability Tests.

¹ Wissenschaftsrat. Empfehlungen zur Weiterentwicklung der wissenschaftlichen Informationsinfrastrukturen in Deutschland bis 2020, Drs. 2359-12. Berlin 13.07.2012, S. 34.

² www.beilstein.tv aufgerufen am 31.12.2012.

³ www.probedo.de aufgerufen am 31.12.2012.

⁴ www.tib-hannover.de/de/forschung-und-entwicklung/projekt/av-portal/ aufgerufen am 31.12.2012.

⁵ www.getinfo.de aufgerufen am 31.12.2012.

⁶ So war die TIB 2005 die weltweit erste DOI-Registrierungsagentur für Forschungsdaten und führt heute die Geschäftsstelle von DataCite, dem internationalen Konsortium für einen weltweit einheitlichen Zugang zu Forschungsdaten.

⁷ www.pangaea.de aufgerufen am 31.12.2012.

⁸ www.bsrn.awi.de aufgerufen am 31.12.2012.

⁹ Lin, J., Keogh, E., Lonardi, S., Chiu, B.: A symbolic representation of time series, with implications for streaming algorithms. In: Proceedings of ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery (2003).

¹⁰ Bernard, J., Brase, J., Fellner, D.W., Koepler, O., Kohlhammer, J., Ruppert, T., Schreck, T., Sens, I.: A Visual Digital Library Approach for Time-Oriented Scientific Primary Data. In: International Journal on Digital Libraries, Vol. 11 (2011), 2, S. 111–123.

¹¹ Shneiderman, B.: The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. In: Proceedings of the IEEE Symposium on Visual Languages, Washington. IEEE Computer Society Press, 1996, S. 336–343.

¹² Kohonen, T.: Self-Organizing Maps, 3. Aufl. Springer, Heidelberg (2001).

**neue Suchmöglichkeiten
für Forschungsdaten in
GetInfo**

DIE VERFASSER

Dr. rer. nat. Oliver Koepler ist Wissenschaftlicher Mitarbeiter im VisInfo Projekt, Technische Informationsbibliothek und Universitätsbibliothek (TIB/UB), Welfengarten 1B, 30167 Hannover, Tel.: 0511 – 7623449, E-Mail: oliver.koepler@tib.uni-hannover.de

Dr. Irina Sens ist Stellvertretende Direktorin TIB und Leiterin Produktentwicklung, Technische Informationsbibliothek und Universitätsbibliothek (TIB/UB), Welfengarten 1B, 30167 Hannover, E-Mail: irina.sens@tib.uni-hannover.de