

Statistische Daten

Die Erhebung statistischer Daten hat eine lange Tradition. Daten über Bevölkerungszahlen werden beispielsweise schon seit Jahrhunderten gesammelt. Eine zentrale Rolle fällt dabei den Behörden zu, die mit der Erhebung statistischer Daten beauftragt sind. In nahezu jedem Land, in manchen Ländern auch in einzelnen Regionen, gibt es hierfür entsprechende Institutionen, die diese wichtige staatliche, und gesellschaftliche Funktion wahrnehmen. Aber auch Nichtregierungsorganisationen, Verbände und Forschungseinrichtungen erfassen statistische Daten. Nicht zuletzt sammelt jedes Unternehmen innerhalb seiner CRM- oder ERP-Systeme eine Vielzahl von Daten, die für Steuerung, Organisation und Entscheidungsfindung wichtig sind. Statistische Daten haben darüber hinaus einen hohen Stellenwert bei Sachentscheidungen in Politik und Wirtschaft. In der Medizin und in den Sozialwissenschaften sind sie Grundlage für Forschung und wissenschaftliche Arbeit, beispielsweise in der Epidemiologie.

Der breiten Bevölkerung werden statistische Daten über die Medien zugänglich gemacht. Heute kommen Zeitungen oder Online-Publikationen kaum mehr ohne die Darstellung statistischer Daten in Form von Tabellen oder Diagrammen aus - ein Trend, der sich in den vergangenen 20 Jahren verstärkt hat.

Informationsvisualisierung - grafische Darstellung statistischer Daten

Die tabellarische Darstellung größerer Datenmengen ist wenig effektiv. Das menschliche Auge kann Trends oder Verteilungen wesentlich besser erfassen, wenn Daten grafisch dargestellt werden. Die Disziplin, die sich mit der grafischen Aufbereitung numerischer Daten beschäftigt, ist die Daten- oder Informationsvisualisierung. Die geläufigsten Darstellungen sind thematische Karten, Balkendiagramme oder Liniengrafen, die Zeitreihen repräsentieren. Weniger bekannte Darstellungsarten sind Korrelationsgrafiken oder Verteilungsgrafiken.

Obwohl diese sehr aussagekräftig und leicht verständlich sind, werden sie von den Medien selten genutzt. Häufig anzutreffen sind hingegen so genannte Infografiken, meist Kombinationen aus den oben genannten klassischen Darstellungsarten, die durch Bilder und redaktionelle Texte ergänzt werden.

Existierende Software

Es existieren verschiedene Spezialanwendungen zur Auswertung und Weiterverarbeitung statistischer Daten, die überwiegend im wissenschaftlichen Bereich verwendet werden. In Unternehmen übernehmen diese Aufgabe CRM- oder ERP-Systeme, die über integrierte Reporting Module verfügen. Für isoliert vorliegende Daten bieten Tabellenkalkulationsprogramme die Möglichkeit einer schnellen grafischen Darstellung und Analyse. All diesen Systemen ist gemein, dass es sich dabei üblicherweise nicht um Internetanwendungen handelt. Wenn Visualisierungsmöglichkeiten vorhanden sind, ist die Veröffentlichung der Ergebnisse im Internet nur als Export von Bildern möglich.

Im Internet gibt es verschiedene Ansätze zur Veröffentlichung statistischer Daten: kleinere Anbieter stellen oft Dokumente auf Ihre Website, in denen Tabellen und Grafiken mit redaktionellen Beiträgen gemischt sind. Die deutschen statistischen Ämter verfügen über ein Web-Frontend namens GENESIS Online. Hier können individuelle Tabellen angezeigt und heruntergeladen werden. Es handelt sich hierbei um eine HTML-Anwendung, die einige statische Grafikfunktionen bietet. Eurostat bietet ein großes Downloadportal für Tabellen und Dokumente, ebenfalls mit einer HTML-Oberfläche.

In den USA ist die Situation weit heterogener. Jedes Amt unterhält sein eigenes Internetportal und innerhalb der Portale müssen die Informationen häufig an verschiedenen Stellen zusammengesucht werden.

Einige Technologieanbieter haben sich auf Softwarelösungen für die Distribution von Daten im Internet spezialisiert. Diese Technologien sind bei verschiedenen

statistischen Ämtern weltweit im Einsatz. Hier sind Beyond 20/20 und PX-Axis zu nennen. Diese Programme sind jedoch durch ihre HTML-Oberflächen in Bedienung und Funktionsumfang limitiert. Weiterhin gibt es Technologieanbieter, die sich auf eine bestimmte Darstellungsart spezialisiert haben. So bietet der französische Softwarehersteller Geoclip ein in Flash realisiertes Kartenframework an, mit dem Daten visualisiert werden können. Hier ist zwar eine gute Interaktivität gegeben, die Anwendung baut sich aber um die Karten auf und nicht um die Daten.

Moderne interaktive Informationsvisualisierung mit DataDiver

Moderne leistungsfähige Computer ermöglichen heute eine ganz neue Herangehensweise an die Datenvisualisierung. Ist man bei einem Printmedium auf die Darstellung einzelner Bilder beschränkt, kann auf einem Computer das Zeichnen eines Diagramms oder einer Karte unmittelbar durch die Auswahl entsprechender Funktionen in der Anwendung erfolgen, auch innerhalb einer Web-Anwendung. Delegiert man das Zeichnen an die Clientsoftware, kann dies auch ohne die Latenz einer Serverabfrage erfolgen. Dadurch lassen sich auch Realtime Manipulationen von Visualisierungen durchführen. Für den Anwender bedeutet dies, dass er zum Beispiel die Veränderungen von Daten über einen Zeitraum hinweg, durch das Verschieben eines Reglers animieren und sichtbar machen kann. Auch die Navigation durch die Daten kann kontextbezogen erfolgen, wenn man zum Beispiel von einer Darstellungsart zu einer anderen wechseln möchte, oder ein ähnlicher Indikator für eine andere Region gewählt werden soll. Diese Funktionalität war letztendlich namensgebend für DataDiver, da mit dieser Anwendung all diese modernen Möglichkeiten konsequent umgesetzt werden. Entscheidend war daher auch, dass DataDiver als Browseranwendung konzipiert wurde. Nur dadurch wird es möglich, dass man als Anwender in Echtzeit und ortsunabhängig durch große Datenmengen navigieren kann, spontan Parameter verändert und sich auf sehr

einfache und schnelle Art aussagekräftige interaktive Visualisierungen für die untersuchten Daten erzeugen lassen kann. Nicht zuletzt deshalb kann man sagen, dass DataDiver nicht nur die Sichtweise auf statistische Daten revolutioniert, sondern auch die Möglichkeit der schnellen Analyse, Verfügbarkeit, Visualisierung und Weiterverwendung.

Viele Datenquellen – ein Portal

Neben den funktionalen Aspekten, war die Zentralisierung des Zugriffs ein Kerngedanke bei der Entwicklung von DataDiver. Es sollte ein Portal geschaffen werden, das dem Anwender für nahezu jede Fragestellung spontan die gesuchten Daten liefert. Dabei sollten vielfältige und performante Suchmöglichkeiten verfügbar sein sowie ein müheloser Übergang vom Suchergebnis zur Visualisierung mit maximal drei Mausklicks. Mit Hilfe einer einfach zu bedienenden Stichwortsuche und der Möglichkeit, Daten über einen Themenbaum zu suchen, wird diese Kernanforderung optimal erfüllt.

Ein Portal, das Antworten auf möglichst alle Fragestellungen liefert, erfordert die Bereitstellung von Daten aus vielen Datenquellen. Mit der Übernahme der Daten der deutschen statistischen Ämter und von Eurostat wird für Deutschland und Europa bereits zum Start von DataDiver eine sehr hohe Abdeckung erzielt. Auch weltweite Daten sind mit den Datenkontingenten der FAO, WHO, des World Factbook, U.S. Census, CDC, Federal Reserve, Homeland Security, SAMHSA oder der PENN World Tables verfügbar. Durch die sukzessive Erweiterung des Datenbestandes sind auch detaillierte Daten aus anderen Regionen der Welt und Bereichen, wie Demografie, Gesundheit, Justiz, Kriminalstatistik, Wahlen, Arbeitsmarkt und vielen anderen Themen bereits in der Startversion verfügbar oder kommen laufend hinzu. Um dies zu gewährleisten, wird der Datenbestand von DataDiver wöchentlich aktualisiert. Auf diese Weise entsteht dem DataDiver Nutzer kein Aktualitätsnachteil gegenüber den Portalen der Originalanbieter.

Weiterverwendung von Visualisierungen

Ein Portal, auf dem man Daten visualisieren kann, ist uns zu wenig. Was nützt die Recherche und die Erstellung einer aussagekräftigen Visualisierung, wenn sie nicht exportiert oder in eine Webseite eingebunden werden kann? DataDiver stellt hierfür eine Vielzahl von Exportmöglichkeiten zur Verfügung.

Neben Druckfunktionen können DataDiver Runtimeversionen für einzelne Visualisierungen heruntergeladen und in die eigene Homepage oder in das Intranet eingebunden werden. Sollte beispielsweise bei einem Meeting kein Internet verfügbar sein, kann eine Offline-Runtimeversion erzeugt werden, die auf jedem Notebook oder Rechner läuft, ohne Verbindung zum DataDiver Server. Unsere Hotlink-Funktionalität ermöglicht das Einbetten von DataDiver Visualisierungen in eine Homepage schlicht durch das Einfügen eines kurzen HTML-Fragments (Code Snippet), das sich einfach per Cut & Paste übertragen lässt. Die Visualisierung wird dann direkt vom DataDiver Server geladen.

Internationalisierung – konsequent umgesetzt

DataDiver kommt aus Deutschland, wurde aber für den weltweiten Einsatz konzipiert. Wir unterstützen nicht nur verschiedene Sprachversionen sondern auch die unterschiedlichen Regionalschemas. Bei der Anmeldung kann ein Regionalschema ausgewählt werden, das festlegt, wie Zahlen- und Datumswerte formatiert werden. Auch beim Export von Visualisierungen und Präsentationen können Sprache und Regionalschema festgelegt werden. Zum Start von DataDiver sind die Sprachen Deutsch und Englisch verfügbar, Erweiterungen auf Französisch und Spanisch sind bereits in Arbeit.

Technische Umsetzung - Verwaltung von Metadaten

Nach Formulierung der Leistungsziele für die Plattform wurde schnell klar, dass nur eine Trennung von Metadatenverwaltung und Live-System die Anforderungen erfüllen konnte.

Es war ein Verwaltungssystem zu realisieren, das folgende Anforderungen erfüllt:

- Verwaltung von Metadaten wie Indikatoren, Variablen und Ausprägungen
- Lokalisierungswerkzeuge
- Beschreibung und Klassifizierung der Metadaten anhand der Visualisierungsmöglichkeiten und Rechenoperationen, die auf dem Live-System verfügbar sein sollten.
- Beschreibung der Formate der zu importierenden Originaldaten
- Generierung, Zuweisung und Indizierung von Suchbegriffen
- Kompilieren von Daten für das Livesystem aus den Originaldaten und der Metadatenbeschreibung. Dabei sollten verschiedene Kompilate nach Datenquellen und Sprachen möglich sein, um Runtime-Datensätze für verschiedene Plattformen erzeugen zu können.
- Beschreibung der Konvertierungsregeln für die Kompilierung der Originaldaten
- Multiuserfähigkeit

Die Verwaltungssoftware wurde als generischer Windows Client implementiert, der via HTTP mit einer PostgreSQL Datenbank kommuniziert.

Technische Umsetzung - Client

Bei der Wahl der Technologie für den Client waren neben der Grundanforderung, eine plattform-übergreifende Anwendung, die in allen Browsern läuft zu schaffen, zwei weitere Aspekte zu berücksichtigen.

Zum einen mussten die für die interaktive Datenvisualisierung erforderlichen Funktionen implementierbar sein. Dies erforderte insbesondere die Verfügbarkeit einer hochperformanten clientseitigen Grafik-Engine. Zum anderen sollte ein GUI-Framework implementiert werden, das für die Such- und Verwaltungsfunktionen in DataDiver verwendet werden kann. Dieses GUI-Framework sollte aber auch für die schnelle Realisierung anderer Projekte, wie zum Beispiel CMS-Systeme verwendet werden können. Das Erscheinungsbild der Anwendungen sollte dabei dem einer Desktop-Anwendung entsprechen. Zudem war einfache Wartbarkeit ein wichtiges Kriterium.

Ein gewöhnliches HTML-Interface schied schon wegen des Mangels an clientseitiger Grafikfunktionen aus. Die Zeichen- und Renderingvorgänge an einen Server zu delegieren, hätte Zugeständnisse an Interaktivität und Performance bedeutet. Die Alternative, ein HTML-Interface mit eingebettetem SVG-Code unter Verwendung von ECMAScript auf der Clientseite, brachten beim Testen nicht die gewünschten Resultate in Bezug auf Performance und Wartbarkeit.

Es blieb die Option eine Technologie zu verwenden, die clientseitig auf einer Virtual Machine läuft. Zur Auswahl standen Java oder Flash. Java hatte den Vorteil, eine etablierte Technologie zu bieten, für die eine Vielzahl guter Entwicklungswerkzeuge existiert. Zudem sind generische Bibliotheken für Graphik- und GUI-Funktionalität integriert. Andererseits ist die Java Unterstützung gerade auf Windows Plattformen schlecht und das Installieren der Java VM ist für den unerfahrenen Anwender ein nicht ganz einfacher Prozess.

Flash hat dagegen eine sehr hohe Verbreitung und bietet hervorragende Graphikperformance, insbesondere für den Zweck der Datenvisualisierung als Vektorgrafiken. Als Entwicklungswerkzeuge wären die Adobe Technologie Flex und das Open Source Framework OpenLaszlo verfügbar gewesen. Die Entscheidung fiel zugunsten von Flash, da die hohe Verbreitung und die einfache Installierbarkeit des Flash Players die Vorteile von Java überwogen. Die verfügbaren Frameworks Flex und OpenLaszlo überzeugten uns jedoch nicht in Hinblick auf die gestellten

Anforderungen.

Wir entwickelten daraufhin zwei verschiedene Frameworks. Das erste abstrahiert die Zeichenfunktionen, die benötigt werden, um die interaktiven Visualisierungen implementieren zu können. Das zweite Framework implementiert ein komplettes Fenstermanagement mit klassischen GUI-Elementen, wie Buttons, Listen, Tabs etc. Dieses Framework wurde daraufhin optimiert, dass auch beliebige andere Applikationen wie beispielsweise CRM-Systeme in kürzester Zeit in der Qualität von Desktop-Anwendungen realisiert werden können - jedoch als Browseranwendungen.

Eine weitere Herausforderung war die Server-Kommunikation, da bei der Datenvisualisierung große Datenmengen anfallen, aus denen in kürzester Zeit beliebige Teilmengen extrahiert werden müssen. Zudem erfordern diese Datenmengen clientseitige Persistenz. Der klassische Weg, XML-Dateien in clientseitige Objekte zu überführen, war in Bezug auf die Performance des Random Access hinreichend, jedoch sehr speicherintensiv. Wir implementierten daher ein eigenes Memory-Management für Flash, das hinsichtlich der Parameter, der Übertragungs- und Verarbeitungsgeschwindigkeit sowie des Speicherverbrauch optimiert ist. Zum Kompilieren der Flash Action Script Quellcodes wählten wir den Open Source Compiler MTASC.

Ausblick

Schon zum Start in Q2/2010 bieten wir DataDiver mit einem riesigen Datenbestand an. Dabei ergibt der Grundstock von fast eine Milliarde Datenfelder durch die verschiedenen Rechen- und Auswahloperationen eine nicht quantifizierbare Menge möglicher Visualisierungen. Neben der ständigen Aktualisierung des Datenbestandes werden wir laufend neue Datenquellen hinzunehmen und die bestehenden ausbauen. Unser Augenmerk richten wir dabei besonders auf eine thematische und

regionale Verbreiterung des Datenbestandes. Dazu werden wir gezielt auch kleinere Anbieter ansprechen, deren Daten für die Öffentlichkeit von großem Interesse sind. Insbesondere wollen wir das Datenangebot für bestimmte Regionen erweitern.

Auch die Funktionalität wird erweitert, die DataDiver letztlich zum universellen Visualisierungsframework für das Internet machen soll. So wird es demnächst für Jedermann möglich sein, die eigenen, beispielsweise innerhalb der Firma anfallenden Daten, mit DataDiver zu visualisieren. Über einfache Importmöglichkeiten mit CSV-Dateien oder Spreadsheets können eigene Runtimeversionen von DataDiver für die eigene Website, für das Internet oder für Offline-Präsentationen generiert werden. Die Daten können mit Daten aus DataDiver kombiniert werden, es können Normierungen oder Anteile berechnet werden und selbstverständlich stehen auch die kompletten Internationalisierungsfunktionen zur Verfügung.