*"Invisible Web", "Deep Web", "Suchwerkzeuge-Web", "singuläre Datenbanken", "Proprietäres Web"*

# Weltregionen des Internet: Digitale Informationen im WWW und via WWW

**von Wolfgang G. Stock**

## "Invisible Web", "Deep Web", "Search Engine Web", "Single Databases", "Proprietary Web": The World Regions of the Internet. Digital Information on and via the WWW.

*Abstract.* Concerning the discussions about the "invisible web" resp. the "deep web" we undertake an analytical view towards the Internet and its possibilities to search for and find out digital information in separated "internet regions". We have to make a distinction between information on the Web and information accessible via the Web. Search engines and Web catalogues cover big parts of the information on the Web. Via the Web users can find both (mostly free of charge) databases produced by public institutions and companies and proprietary content of the fee-based information suppliers. Connections between our analytically separated Internet world regions are not only possible (and already realized in some cases), they are beneficial to information providers and users.

*Zusammenfassung.* Im Anschluss an die Diskussionen über das "unsichtbare Web" bzw. die "Tiefenregionen des Web" versuchen wir eine analytische Sicht auf das Internet und dessen Möglichkeiten, digitale Informationen in den unterschiedlichen "Internetweltregionen" zu suchen und zu finden. Zu unterscheiden sind die Informationen, die im Web liegen, von solchen, die via Web erreichbar sind. Große Teile der Informationen im Web werden von den Suchwerkzeugen (Suchmaschinen und Webkatalogen) erreicht. Via Web ansprechbar sind sowohl (in der Regel kostenfreie) Datenbanken von öffentlichen Einrichtungen und Unternehmen als auch die proprietären Inhalte der kommerziellen Informationsanbieter. Verbindungen zwischen den analytisch separierten Internetweltregionen sind nicht nur möglich und werden bereits ansatzweise geschaffen, sie sind aus Anbieter- und Nutzersicht auch sinnvoll.

## Oberfläche und Tiefenregionen des Internet

Mit dem Buch "The Invisible Web" von Chris Sherman und Gary Price wurde für viele Laien klar, was für Information Professionals wohl kaum außer Frage stand: Suchwerkzeuge im Internet - Suchmaschinen wie Google oder Alltheweb sowie Webkataloge wie Yahoo! und das Open Directory - finden nicht alle digitalen Informationen, die weltweit im und über das Internet eigentlich erreichbar wären. Diesen Teil des Internet bezeichnen Sherman und Price als das **unsichtbare Web.** "In a nutshell, the Invisible Web consists of material that general-purpose search engines either can not, or perphaps more importantly, **will not** include in their collections of Web pages", so Sherman und Price 2001. Da unsere Autoren in ihrem Buch nun aber ausführliche Listen von Einstiegsseiten in das "Invisible Web" vorlegen, kann dieses wohl kaum so unsichtbar sein, wie der Name suggeriert. Insofern ist die Bezeichnung etwas irreführend, so dass die Rede vom "Deep Web", den **Tiefenregionen** des Internet wahrscheinlich treffender ist. Den Begriff des "Deep Web" verwendet das Unternehmen BrightPlanet, das eine Retrievalsoftware zum "Abfischen" von Oberflächen- und von Tiefenregionen des World Wide Web offeriert.

Das **Oberflächenweb** wird mehr oder minder komplett von den Suchwerkzeugen, den automatisch indexierenden Suchmaschinen und den intellektuell auswertenden Webkatalogen erfasst. Untersuchungen (z.B. Stock / Stock 2000) legen nahe, dass zumindest für gängige Datenformate beim parallelen Ansprechen mehrerer der großen Web-Retrievalsysteme (etwa Google, Alltheweb und AltaVista) große Teile (bis zu 90%) des World Wide Web findbar sind. (Zumindest, wenn man großzügig unterstellt, dass ein "Treffer" auch noch zählt, wenn man ein bis zwei Clicks nach dem angezeigten Treffer fündig wird.) Vollständigkeit zu erwarten, wäre allerdings maßlos übertrieben. Mit der Geschwindigkeit, in der neue Seiten generiert und alte modifiziert werden, können die Suchwerkzeuge nur moderat mithalten.

Teilweise ignorieren die Suchwerkzeuge aber auch "bewusst" Material, und das findet ausdrücklich Zustimmung. Hier geht es um **Spam,** Informationsmüll, den keiner haben möchte und den auch keiner braucht - einige wenige Perverse und Waffennarren vielleicht ausgenommen. (Dass an dieser Stelle latent die Gefahr einer Zensur lauert, sei zugegeben.)

Die **Tiefenregionen** des Web erschliessen sich dem Nutzer durch das Ansprechen gewisser Datenbanken. Da hier die Webseiten dynamisch erzeugt werden, können diese von den Suchwerkzeugen nicht aufgefunden werden. Michael K. Bergman betont: "Traditional search engines create their indices by spidering or crawling surface Web pages. To be discovered, the page must be static and linked to other pages. Traditional search engines cannot 'see' or retrieve content in the deep Web - those pages do not exist until they are created dynamically as the result of a specific search. Because traditional search engine crawlers can not probe beneath the surface, the deep Web

has heretofore been hidden" (Bergman 2001).

Über die Website des HWWA sprechen wir die Wirtschaftsdatenbank des Hamburgischen Welt-Wirtschafts-Archivs oder über die Site des Deutschen Patent- und Markenamtes die Datenbank DepatisNet an, wobei die Inhalte weder der HWWA- noch der DepatisNet-Datenbank bei Google und Co. recherchierbar sind. Der analoge Fall, nur hier mit der Zwischenstation bei einem Kassenhäuschen, liegt bei den kommerziellen Anbietern von Informationen, also bei Selbstvermarktern (wie z.B. Creditreform Online oder Hoppenstedt Firmendatenbank) und bei Aggregatoren (u.a. den Online-Hosts) vor. Die Informationen dieses Teils des sog. "Deep Web" haben eines gemeinsam: Sie liegen überhaupt nicht **im** Web, sondern in einer Datenbank, die **via** Web erreichbar ist.

Der Umfang des Deep Web ist gemäß Michael K. Bergman rund 400- bis 500mal größer als das Oberflächenweb. Gerechnet wird dabei nicht in der Anzahl von Dokumenten, sondern im Umfang von Speicherplatz. Da Bergman u.a. Datensammlungen wie die des "National Climatic Data Center" (mit rund 370.000 GByte) oder die der NASA (mit 220.000 GByte) mitrechnet, ist die große Zahl nicht mehr ganz so verwunderlich. Weil im Deep Web auch die Kataloge aller Online-Bibliothekskataloge vorkommen, die sich ja durchaus überschneiden, müsste die geschätzte Informationsmenge um die Dubletten bereinigt werden. Da allein ein Informationsanbieter wie Lexis-Nexis in etwa mit dem Suchmaschinen-Primus Google nach der Anzahl der Datensätze in etwa gleich zieht, ist es ganz sicher richtig, dass das Deep Web weitaus größer ist als das Oberflächenweb. Orientiert man sich an der Zahl der Datensätze, so ist Bergmans Größenfaktor wohl um eine Zehnerpotenz überschätzt.

## Eine Taxonomie der digitalen Online-Information

Das Gesamt der online erreichbaren digitalen Informationen ist zunächst in zwei Hauptklassen einzuteilen:

- Informationen, die im Web liegen (d.h. fest verlinkt sind, wobei keinerlei Kosten für den Nutzer entstehen)
- Informationen, die in Informationssammlungen (i.d.R. Datenbanken) liegen, wobei die Einstiegsseite der Informationssammlung via World Wide Web erreichbar ist.

Die Mehrzahl der Informationen im Web wird von Suchwerkzeugen - nicht von einem einzelnen, aber in der Summe der Werkzeuge - ausgewertet. Nicht immer erfasst werden Informationen mit speziellen Datenformaten, Dokumente, auf die kein Link gerichtet ist; Probleme können u.U. bei der Verwendung von Frames auftauchen - hier droht in der Tat Informationsverlust. (Weiterer Informationsverlust entsteht bei suboptimalem Relevance Ranking von Suchmaschinen: Was nutzt ein Treffer auf Platz 543, den der Nutzer nie sehen wird? Da uns hier aus-schließlich die Datenbasen und nicht die Retrievalsysteme interessieren, müssen wir solche Aspekte außer Acht lassen.) Gewünschter Informationsverlust liegt beim Ausfiltern von Spam vor.

Die zweite Hauptklasse digitaler Online-Informationen liegt nicht selber im Web, sondern ist nur über das Web erreichbar. Natürlich lassen sich die Einstiegsseiten solcher Systeme auch bei den Suchmaschinen und Webkatalogen finden, nicht aber die gespeicherten Datensätze. Eine über Suchwerkzeuge vermittelte Suche verläuft indirekt. Zunächst suche ich die Einstiegsseite der Informationssammlung, um im zweiten Schritt beim betreffenden Retrievalsystem mein konkretes Suchargument zu formulieren. Diese via Web erreichbaren Informationen zerfallen wiederum in zwei Unterklassen, und zwar nach der Kostenpflichtigkeit der Angebote. Es gibt singuläre kostenlose Datenbanken (Beispiele: Amazon, eBay, HWWA, Medline, diverse Datenbanken nationaler Patentämter, Bibliothekskataloge) und es gibt kostenpflichtige Angebote. Letztere haben als Spezies die Datenbanken der Selbstvermarkter, die genau so singulär dastehen wie die alleinstehenden kostenlosen Datenbanken (nur dass hier halt zu zahlen ist) und als weitere Spezies die Aggregatoren, die verschiedene Informationssammlungen unter einer Oberfläche vereinigen. Aggregatoren können sowohl bei Märkten nicht-digitaler Güter (Beispiele: diverse Shopping Malls), bei Verlagsprodukten (Beispiele: Xipolis als aggregiertes Angebot des Brockhaus und weiterer Nachschlagewerke, Springer Link oder ScienceDirect) als auch in der Form von Online-Hosts auftauchen.

## Querweltein-Suchen - Hybridsysteme als Lösung

Michael K. Bergman fordert völlig zu Recht: "Clearly, simultaneous searching of multiple surface and deep Web sources is necessary when comprehensive information retrieval is needed" (Bergman 2001). Ob dies nun stets innerhalb genau einer Retrievaloberfläche für alle Informationen vonstatten gehen sollte, kann bezweifelt werden, verliert der Suchende doch dabei die Vorzüge und Extras der einzelnen Datenbanken - das bekannte Problem wie bei den Meta-Suchmaschinen. Eine erste Alternative wäre, aus einem einheitlichen System heraus die einzelnen Internetwelten bzw. Datenbanken einzeln anzusteuern und die Dokumente in einen einheitlichen Warenkorb abzulegen. Die zweite Alternative wird darin bestehen, fachlich begrenzte Ausschnitte zu wählen, über alle Dokumente aller infragekommenden Welten ein kontrolliertes Vokabular zu legen und auch alle Welten mittels dieses Vokabulars abzusuchen. Wie auch immer die Lösung aussehen wird: Stets ist ein Suchen in mehreren unterschiedlichen Internetwelten nötig, also ein "Querweltein-Retrieval". Wenn wir quer durch die Internetwelten Recherchen anbieten, wollen wir von "Hybridsystemen" reden. Ansätze, Hybridsysteme zu kreieren, gibt es bereits vielfach. Je nach der Herkunft des Unternehmens unterscheiden wir folgende Fälle:

- "reiner" Hybrid (unsere Beispiele: Northern-Light, BrightPlant)
- Suchwerkzeuge im Web mit Querweltein-Ergänzungen (AltaVista.de, Google.com)
- Singuläre Datenbank mit Querweltein-Ergänzungen (HWWA)
- Kommerzielle Informationsanbieter mit Querweltein-Ergänzungen (Factiva).

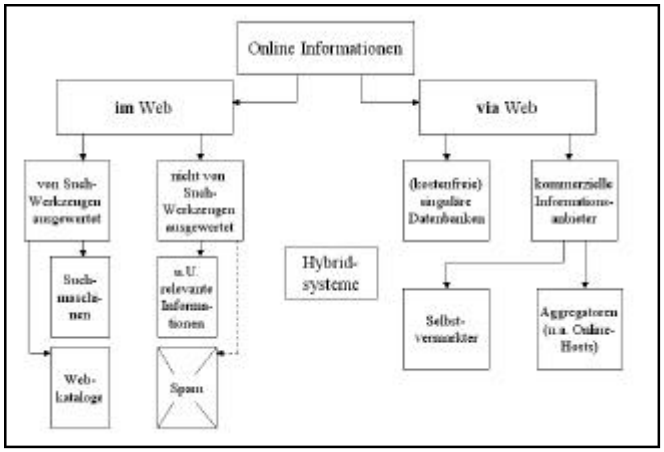


*Abbildung 1. Welten der Online Informationen im World Wide Web.*


*Abbildung 2. Hybridsystem Northern Light.*

Ein vom Ansatz her eindeutiger Hybrid ist Northern Light. Dokumente im Web werden konsequent mit proprietärem Inhalt innerhalb eines Systems verknüpft. Abbildung 2 zeigt die Suchoberfläche von NLResearch. Webdokumente und "Special Collection & Premium Content" werden innerhalb einer Datenbank angeboten, zusätzlich gibt es Links zu weiteren einschlägigen Informationssammlungen.

Einen ähnlich hybriden Ansatz verfolgt BrightPlanet, indem - kostenpflichtig - in ansonsten kostenfreien Datenbanken, die via Web angesprochen werden, innerhalb eines Retrievalschrittes gesucht wird. BrightPlanet bietet demnach eine datenbankübergreifende Suche in diversen (mehreren zehntausend) Datenbanken an. Systeme der kommerziellen Informationswirtschaft werden dabei nicht beachtet.

AltaVista liefert uns ein Beispiel für eine Web-Suchmaschine, die gewisse Datenbanken in die Suche mit einbezieht. Abbildung 3 zeigt eine Trefferliste beim deutschen AltaVista, in der wir (vor die AltaVista-Inhalte sortiert) den Link zur passend zur Suchfrage generierten Trefferliste von Wer liefert Was? finden. Unten auf der Seite liegen Links u.a. zu eBay und Amazon. Hier ist jedoch die Suche noch nicht ausgeführt worden, so dass u.U. auch null Treffer herauskommen können.

Ähnlich geht (die us-amerikanische Version von) Google vor. Verfolgt man den Link zu Dictionary.com (im Google.com-Anzeigebildschirm bei der Angabe des unterstrichenen Sucharguments), kommt man zu weiteren Datenbanken, darunter via eLibrary auch zu kostenpflichtigen Volltexten dieses Aggregators.

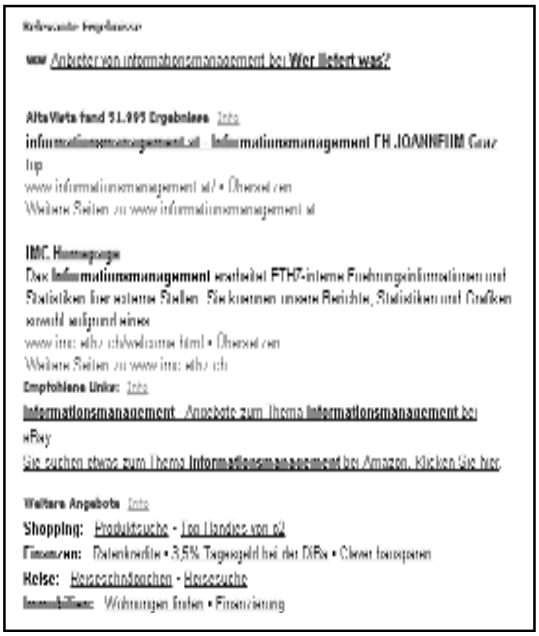


*Abbildung 3. Suchmaschine AltaVista (deutsch) mit Querweltein-Ergänzungen: Links zu den Datensatz-Treffern singulärer Datenbanken wie Wer liefert Was? nebst Links zu eBay oder Amazon*

Das HWWA versorgt uns mit einem Beispiel der Verknüpfung einer singulären Datenbank mit dem Oberflächenweb. Fachspezifische digitale Dokumente werden analog zu gedruckten Dokumenten indexiert und als Katalogkarte gespeichert. Die Links der HWWA-Datenbank verweisen sowohl auf proprietäres Material (zu PDF-Volltexten diverser Zeitschriften) als auch (unsere Abbildung 4) auf Dokumente im freien Oberflächenweb.

*Abbildung 4. Datensatzbeispiel der HWWA-Wirtschaftsdatenbank. Verlinkung eines Katalogisats innerhalb einer singulären Datenbank ins Oberflächenweb*

Unser letztes Beispiel führt in die kommerzielle Informationswirtschaft. Factiva bietet nicht nur proprietären Content an, sondern zusätzlich Dokumente, die im Oberflächenweb liegen (hier meist Newsservices).

## Fazit

Querweltein-Retrieval ist nicht nur nötig angesichts der unterschiedlichen Internetwelten, es gibt bereits erste Ansätze, die vormals getrennten Weltregionen gemeinsam zu durchsuchen. Die Bestrebungen der singulären Datenbanken sowie der kommerziellen Informationsanbieter verfolgen dabei einen fachspezifischen Zugang. Dieser hat den Vorteil einer recht zielgenauen Suche, können doch dokumentarische Werkzeuge wie Thesauri bzw. Klassifikationssysteme eingesetzt werden. Die Bestrebungen der Suchmaschinen weiten den Horizont, indem sie die an sie gerichteten Suchargumente an weitere (kostenpflichtige wie kostenlose) singuläre Datenbanken weiterreichen. Hier liegt der Vorteil in der Breite der Datenbasis, wobei die Qualität der zielgenauen Suche leidet. Strategie der Suchmaschinenfirmen sollte sein, die Anzahl der durchzuschaltenden singulären Datenbanken zu maximieren, zumindest aber solche Datenbanken zu erreichen, bei denen ein "allgemeines Interesse" vorausgesetzt werden kann. Dass an gewissen Stellen dann ein Kassenhäuschen auftritt, dürfte kaum schaden - der Nutzer entscheidet, ob ihm die volle Information Geld wert ist. (Eine andere Frage ist die der fairen Abrechnung, insbesonders im grenzüberschreitenden Vertrieb. Es hat derzeit den Anschein, als würden Suchmaschinen nicht global agieren wollen, sondern stets national, wenn sie singuläre Datenbanken in ihr Programm aufnehmen.) Die Strategie der singulären Datenbankanbieter müsste sich eigentlich mit der der Suchmaschinen decken, da sie wohl daran interessiert sind, ihre Informationen soweit wie möglich zu streuen. Der große Vorteil fachspezifischer Datenbanken liegt darin, mit kontrolliertem Vokabular zu arbeiten und enge Nutzergruppen anzusprechen. Hier muss es Strategie sein, alle Informationsressourcen, also auch die des Oberflächenweb, mittels einheitlicher Methoden und Werkzeuge auszuwerten. Mit der Erkenntnis, die Welt der Informationen im Internet mit der Welt der Informationen, die via Internet erreichbar werden, querweltein zu verbinden, dürfte eine Umsetzung des gemeinsamen Zugriffs auf alle digitalen Online-Informationen realistisch erscheinen. Dass die Retrievalsysteme und insbesondere das Relevance Ranking vor neuen Herausforderungen stehen, ist klar, denn die Menge der Datensätze wird massiv erweitert.

## Literatur

Michael K. Bergman: The Deep Web: Surfacing Hidden Value. - In: The Journal of Electronic Publishing 7 (2001) Iss. 1 - URL: www.press.umich.edu/jep/07-01/bergman.html.

Chris Sherman; Gary Price: The Invisible Web. - Medford: Information Today, 2001.

Mechtild Stock; Wolfgang G. Stock: Internet-Suchwerkzeuge im Vergleich. Teil 1: Retrievaltest mit Known Item Searches. - In: Password Nr. 11 (2000), S. 23-30.