



Gruppe 2: Invisible Web

Unter dem sog. *Invisible Web* (auch *Deep Web*) versteht man denjenigen Teil des Web, der von Suchmaschinen nicht erfasst wird. Dafür kann es unterschiedliche Gründe geben; neben technischen Hürden, die es den Suchmaschinen unmöglich machen, diesen Teil des Web zu erschließen, gibt es von den Inhalte-Anbietern selbst erstellte Barrieren oder solche Dokumente, die die Suchmaschinen willentlich von der Erschließung ausschließen. Sherman und Price definieren das Invisible Web wie folgt:

„Text pages, files, or other often high-quality authoritative information available via the World Wide Web that general-purpose search engines cannot, due to technical limitations, or will not, due to deliberate choice, add to their indices of Web pages" (Sherman u. Price 2001, 57). [...]

Die hauptsächliche Unterscheidung liegt in der Erreichbarkeit der Informationen. Während die von den Suchmaschinen erschlossenen Informationen im Web erreichbar sind, sind die Inhalte des Invisible Web nur über das Web erreichbar, d.h. es bestehen zwar Schnittstellen im Web, die dahinter liegenden Inhalte sind jedoch nicht direkt erreichbar. Besonders bedeutend ist der Bereich der kommerziellen Informationsanbieter: die Menge der hier erschlossenen Dokumente kann bei einem einzelnen Anbieter durchaus die Menge der von den größten Suchmaschinen erschlossenen Dokumente erreichen (vgl. Lexis-Nexis 2004). Dies mag verdeutlichen, dass heutige Suchmaschinen (entgegen von den Anbietern vorgetragenen Behauptungen, die dies implizieren) nicht in der Lage sind, alle online verfügbaren relevanten Informationen zu erschließen. [...]

Sherman und Price (2001, 70ff.) schlagen eine Unterteilung des Invisible Web in unterschiedliche Ebenen vor. Diese sind das Opaque Web, das Private Web, das Proprietary Web und das Truly Invisible Web.

Das *Opaque Web* („undurchsichtige Web“) besteht aus Seiten, die von den Suchmaschinen technisch erfasst werden könnten, die aber aufgrund bestimmter Restriktionen auf Seiten der Suchmaschinen nicht erfasst werden. Beschränkungen der Suchmaschinen bestehen in Bezug auf die Tiefe des Crawlings (Websites werden nur bis



zu einer bestimmten Ebene bzw. nicht vollständig erfasst), die Crawl-Frequenz (die Indizes der Suchmaschinen werden nicht oft genug aktualisiert, um mit der Aktualisierungsfrequenz manche[r] URLs mithalten zu können), die Maximalzahl der angezeigten Ergebnisse (es wird zwar eine Trefferzahl angegeben, die tatsächlich über die Trefferlisten der Suchmaschinen zugänglichen Dokumente beschränkt sich aber in der Regel auf etwa 1.000 Dokumente) und das Problem der *disconnected pages* [d.h. Seiten, auf die von keiner anderen Seite verlinkt wird]. [...]

Das *Private Web* („privates Web“) besteht aus Seiten, die von ihren Autoren bewusst von der Indexierung durch Suchmaschinen ausgeschlossen wurden, sei es durch eine Passwort-Abfrage, durch die Nutzung des „noindex“-Metatags oder durch den Einsatz einer Robots-Exclusion-Datei („robots.txt“). Nur die erste Methode garantiert allerdings, dass die Seiten nicht erfasst werden; Anweisungen an die Robots der Suchmaschinen werden zwar in aller Regel befolgt, stellen aber tatsächlich nur eine freiwillige Einschränkung der Suchmaschinen dar.

Das *Proprietary Web* („proprietäres Web“, „geschütztes Web“) ist für die Suchmaschinen nicht zugänglich, da für seine Nutzung die Zustimmung zu bestimmten Nutzungsbedingungen notwendig ist. Dies kann eine Registrierung mit den persönlichen Daten sein, hierunter fallen aber auch die kostenpflichtigen Inhalte, die zunehmend angeboten werden (Lewandowski 2003, 35). Die technischen Beschränkungen bestehen in der ersten Linie aus einer Passwort-Abfrage wie im Fall vieler Seiten des Private Web, aber auch Einschränkungen aufgrund eines IP-Adressbereichs sind denkbar.

Das *Truly Invisible Web* („wirklich unsichtbares Web“) besteht aus Seiten bzw. Sites, die für die Suchmaschinen aufgrund technischer Gegebenheiten nicht indexierbar sind. Welche Dokumente zum wirklich unsichtbaren Web gehören, verändert sich aufgrund der Weiterentwicklung der Suchmaschinen natürlich ständig (Sherman, Price 2001, 74). Als die heute noch bedeutendsten Bereiche sind einerseits die dynamisch generierten Seiten und andererseits - und dies macht den bedeutendsten Teil aus - die Inhalte von Datenbanken anzusehen. Zwar gibt es einzelne Ansätze, diese zu erfassen (vgl. z.B. Hamilton 2003; siehe auch Kapitel 12.8), diese müssen aber wenigstens zur Zeit noch als Experimente angesehen werden, die noch weit vor der allgemeinen Durchsetzung stehen.



75 Betrachtet man die Bereiche des Invisible Web hinsichtlich ihrer Bedeutung für die
weitere Erschließung des Web, so lässt sich folgendes feststellen: Die Erschließung
des *Opaque Web* ist weit fortgeschritten; in diesem Bereich sind die Probleme am ge-
ringsten. Eine Ausnahme bilden die *disconnected pages*, die den Suchmaschinen
schlicht unbekannt bleiben. Hier zeichnet sich zur Zeit auch keine Lösung ab. Dass
das Private Web durch Suchmaschinen nicht erschlossen wird, wird sich nicht ändern
lassen und sollte auch nicht durch die Umgehung von De-facto-Standards (Robots
Exclusion) umgangen werden. Im Bereich des *Proprietary Web* zeichnen sich Lösun-
80 gen ab, registrierungs- oder auch kostenpflichtige Inhalte zu erschließen. [...] Bei [den
entsprechenden] Methoden ist die Zustimmung bzw. aktive Mitwirkung der Inhaltean-
bieter notwendig. [...]

85 Stock (2003, 27) schätzt die Größe des Invisible Web [auf das etwa Vierzig- bis Fünf-
zigfache], Sherman (2001) auf das etwa Zwei- bis Fünfzigfache des Visible Web.

90 Suchmaschinen werden das Web nicht vollständig abdecken, da ökonomische und
vor allem technische Hindernisse dem entgegenstehen. Hier entsteht ein grundle-
gendes Dilemma der Informationsrecherche mittels Suchmaschinen: einerseits liefern
diese in der Regel lange Trefferlisten, also zu viele Dokumente, als dass der Nutzer
diese alle begutachten könnte. Auf der anderen Seite erreichen die Suchmaschinen
in ihren Nachweisen bei weitem keine Vollständigkeit, liefern also zu wenige Doku-
mente. [...]

95 Textauszug aus: Lewandowski, Dirk (2005): *Web Information Retrieval*. Frankfurt am Main:
DGI-Schrift (Informationswissenschaft 7), S. 51-57.