Chapter 10

Search Tools and Content Aggregators

10.1 Typology of Search Tools and Content Aggregators

In Chapters 7 through 9, we got to know various specialized information products, each of them catering to specialized markets. In the **Deep Web** (Bergman, 2001)–also called "Invisible Web" (Sherman & Price, 2001)–there are thousand of databases, which offer highly specialized information. Added to this are the billions of pages in the **Surface Web** (Stock, 2007, 108-111). The Surface Web comprises all digital documents that are *within* the Web (and are generally interlinked), while the Deep Web summarizes all digital documents that are integrated in to their own respective information collections (databases), the start pages of which are accessible *via* the WWW (Stock & Stock, 2004c, 3-13).

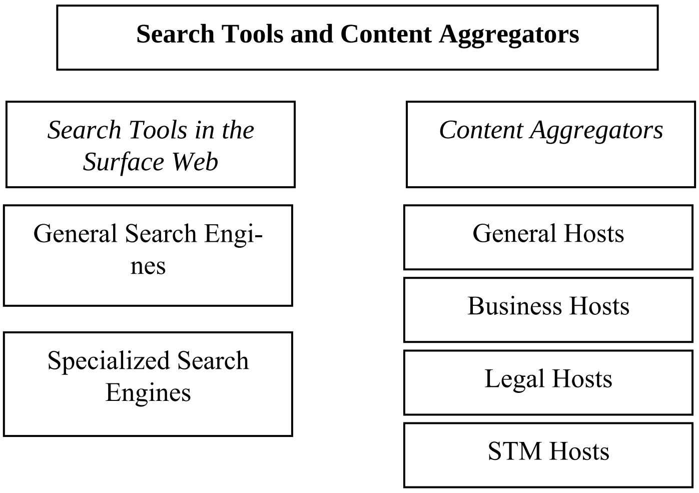
Search Tools and Content Aggregators
Search Tools in the Surface Web
Content Aggregators
General Search Engines
Specialized Search Engines
General Hosts
Business Hosts
Legal Hosts
STM Hosts

Figure 10.1: Classification of Search Tools and Content Aggregators.

Search and retrieval in the Surface Web is performed via **search engines**, which are either aligned to Web contents in general (like Google), or which retrieve specific documents (such as Google News or Google Scholar). The variety of databases in the Deep Web is bundled via Content Aggregators. Such so-called **hosts** summarize (anywhere between hundreds and thousands of) individual databases under one single retrieval system and one single user interface. Depending on the content on offer, we distinguish between general hosts (with no thematic emphasis) and–analogously to information services (Chapters 7-9)–hosts for economic, legal and STM information.

Web search engines cater to mass markets and offer their services free of charge, securing their funding by marketing customers' attention via adverts. Hosts act on (sometimes very small) niche markets. As a critical mass of attention that could bind advertising customers is seldomly reached here, the hosts sell both digital content, i.e. the full texts, bibliographic citations or fact documents they provide, and their services of searching and retrieving content. It can occasionally be observed that operators of Deep Web databases (e.g. JSTOR) deposit their documents for search (but not for display) with search tools in the Surface Web (here: in Google).

10.2 Online Search Engines

In pretty much every country of the world, the market for general Web search engines follows an inverse power law: one single company dominates the market in question, the competitors following some distance behind. In the U.S.A., around two thirds of all Web searches (around 15bn in total) are conducted via Google, with the closest competitor (Yahoo!) accounting for 17% (source: comScore, data for February 2010). In Germany, the distance between the market leader (again Google, this time with 89%) and the second-placed player (T-Online, 3%) is even more extreme (source: Webtrekk, data for June 2009). In China, we can observe the same form of distribution, but with different players: here, 61% of all searches are performed via Baidu, with Google.cn coming in second place with 27% (source: Internet World Business, data for September 2009). The market for search engines thus very impressively demonstrates the "winner takes all" principle. For companies (and all other parties whose websites are meant to be retrieved on the internet), this means that they have to safeguard their sites' visibility with the respective search engine market leader. This is done in two ways via **search engine marketing** (von Bischopinck & Ceyp, 2007):

- Search engine optimization (SEO),
- Sponsored Search.

SEO serves to construct a website in such a way that it will land as high up in the hit list as possible (ideally in first place) if certain search arguments are being used. Sponsored Search (as part of online advertising) pursues the goal of leading potential customers to one's own Web presence via short advertising texts that are displayed, context-specifically, for the search arguments that are used (see Chapter

15). SEO requires technical and content-related measures to be applied to one's own website, Sponsored Search requires financial means (next to the best possible advertising copy and the acquisition of the right search arguments). Whether via SEO or advertising, the central goal of companies is to get their websites (with their products, services, self-projection etc.) retrieved and displayed for the suitable search arguments.

SEO can be performed in the company itself; however, there are also external service providers that specialize in search engine optimization. We distinguish between **on-site optimization** (measures applied to one's own site, e.g. using the correct terminology in the continuous text as well as in the title, number and distribution of crucial terms in the text or in subheadings, the folder structure for the entire site or the placement of internal links) and **off-site optimization** (links to one's own site from external sources and their anchor texts, as well as the number of clicks to one's own site). All optimization measures require detailed knowledge of computer and information science, both for the measures applied in information linguistics as well as the search engines' sorting algorithms used. Only the method of on-site optimization is fully subject to the optimizers' control, off-site methods requiring the help of others. Here, one can very quickly encounter dubious practices (such as the managing of link farms) that are considered spam (Stock, 2007, 125-128) and–if recognized–result in a deletion of the websites by search engines.

Operators of search engines (in most countries at the moment Google) pursue the task of constructing and expanding the broadest possible mass of users for their **advertising customers** (which are, after all, their sole source of profit). All products, be it general search engines (Google.com or, specifically for Germany, Google.de), specialized search engines (Google Scholar, Google News, Google Books etc.) or additional offers (such as Gmail or Google Earth) serve the sole purpose of binding the search engine's users to this research tool in the long term. This is achieved by satisfying the users' information needs via sophisticated search technology and the right content–without charging the users a Cent. In Google's annual report (2009, 1), we read:

> We will do our best to provide the most relevant and useful search results possible, independent of financial incentives. Our search results will be objective, and we do not accept payment for search result ranking or inclusion.
>
> We will do our best to provide the most relevant and useful advertising. Advertisements should not be an annoying interruption. If any element on a search result page is influenced by payments to us, we will make it clear to our users.
>
> We will never stop working to improve our user experience, our search technology, and other important areas of information organization.
>
> We believe that our user focus is the foundation of our success to date.

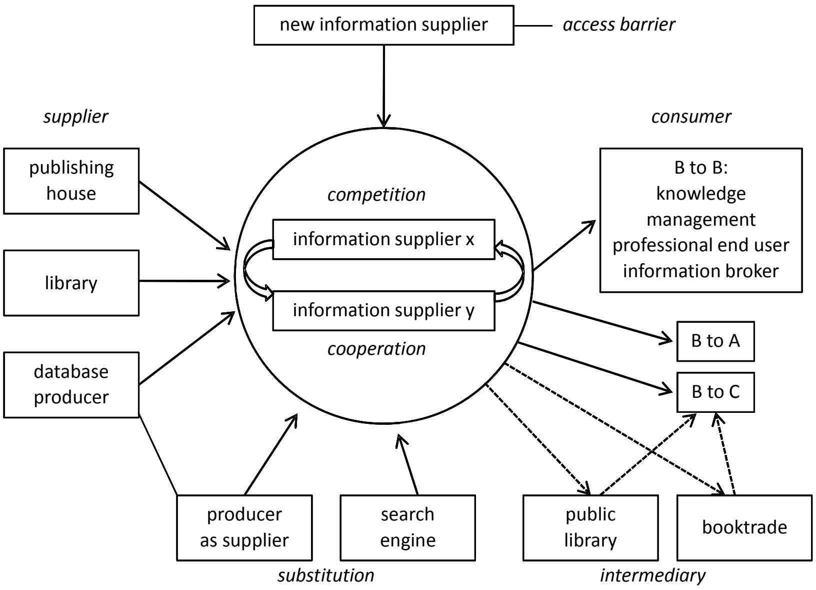

Figure 10.2: Industry Structure of Content Aggregators. Source: Stock & Stock, 2004b, 20.

10.3 Content Aggregators (Hosts)

Online hosts bundle the content of various different databases under one surface and using one retrieval system. For the user, this bears the advantage of having all the important information collections in front of one at a single glance, and only having to speak one retrieval language. However, such search languages are not always easy to use, which is why hosts offer both relevant courses and a help desk for any urgent questions.

The Content Aggregators' suppliers are

- publishers (with their digital content),
- libraries with their document delivery services (for content that is not available digitally and must thus be acquired in the form of a print copy),
- providers of bibliographical information services.

Hosts can be separated into **general information providers** without any thematic restrictions (such as DIALOG or–with an emphasis on magazines' full texts–EBSCO*host*) and **specialist providers**. The latter act in the area of either economic, market and press (e.g. Factiva, Nexis, Profound or–with a particular emphasis on the German economy–GENIOS) (Stock & Stock, 2003), legal (Lexis, Westlaw and–for German law–Juris) (Kremer, 2004) or STM information (Stock & Stock, 2005). Among STM hosts, there are once more, apart from general STM providers (such as STN International or Thomson Reuters with its product Web of Knowledge), specialists, e.g. DIMDI and Ovid for medical information or Questel for information of commercial legal protection (Stock & Stock, 2006). Hosts act on niche markets, which makes it very difficult for new providers to successfully establish themselves on the information market. The market has been in the hands of the established players for years–the online hosts' roots go back to the year 1972 (Stock, 2007, 43-46).

A problem of many hosts is that the **suppliers** also market their information services themselves, thus binding many possible customers. Search engines are regarded as a threat with some justification: Google News is a competitor on the market for press information, Google Scholar, for court decisions (i.e. "Legal Opinions and Journals"), is at least a competitor of the American legal hosts, and Google Scholar (in the segment "Articles and Patents") competes with STM hosts.

On the **customers' side**, B-to-B business models dominate, i.e. companies act as customers. Here, three strategies are pursued in the context of operational knowledge management (see Chapter 7 above): end user research, installation of an information retrieval service or a mixed form of both strategies. Particularly in the area of legal information, but also for resort-specific information (e.g. medical information from DIMDI for the German Federal Ministry of Health), we can find B-to-A business models, in which public administrations act as customers. Due to the lack of end users' willingness to pay, B-to-C business models are hardly realizable. Attempts to incorporate public libraries or stationary book trade into the value chain as a further sales intermediary (Bieletzki & Roth, 1998) must be deemed failures.–An overview of the industry structure of content aggregators is provided by Figure 10.2.

For the **pricing models**, many online hosts prefer subscriptions–either to their entire offer or to individual databases. However, it is also an option for registered customers to selectively access hosts' offers after paying a basic fee, and then paying for them on an individual basis. Thus, the host STN International charges €120 for one hour's access to the database *Compendex*, or €475 for *World Patents Index* (as of 2010). For each bibliographic citation, *Compendex* charges €2.85; viewing the display of a patent document in the *World Patent Index* costs €7.91. Searches to survey a thematic profile (SDI; Selective Dissemination of Information; Stock, 2007, 154-156) are an important product of hosts. Weekly SDI searches in *Compendex* cost €3.50, and €57.60 in the *World Patents Index* (displayed documents are charged additionally). Special commands lead to charges shown separately. The command ANALYZE (for up to 50,000 data pools to be processed), important for informetric analyses (Stock, 2007, Ch. 11), costs €43.90 in STN. Some online hosts (such as GENIOS) do not charge users' access time, which means that only the documents users view generate costs. To safeguard the transparency of these (not inconsiderable) costs, GENIOS shows the fee that is incurred before any document is displayed.

Due to the competition between (free) Web search engines and (commercial) Content Aggregators, it was suggested (Bock, 2000) to use **certification marks** in order to effectively designate the latter as quality information, signaling users that online hosts provide a different kind of information–of higher quality. Highly specialized technical information in particular always represent credence goods for laymen, as they will not be able to exhaustively determine the quality of these economic goods before or after the purchase. Certification marks (e.g. registered as a collective mark) have not (yet) been able to assert themselves for online hosts. How to operationalize the quality of digital information services in such a way that they can be registered via quantitative characteristic values, leaving us able to actually drawing a clear line between quality information and all the rest, is an unresolved problem.

The Content Aggregators' companies can only survive by establishing **unique selling propositions** vis-à-vis competitors on their own market as well as substitute products from other industries (Stock & Stock, 2004b). Such propositions, in the sense of critical success factors, are, for online hosts:

- exclusive content (at least a few of the host's databases are only available here),
- the "right" selection of required databases, from the customer's perspective (for reasons of time and economy, customers prefer one-stop shopping, which means that all relevant sources that are required need to be available via the host),
- the power of the retrieval system used (search and retrieval are conducted on a professional level, which means that the research options must stand out strongly against regular search engines),
- unified knowledge organization systems (thesauri, classification systems etc.) in restricted thematic areas (across the borders of singular bibliographic databases),
- synergies between bibliographical databases, full texts and facts.

Hosts bank on strategic alliances with their suppliers and, partly, with customers (which are asked for their expertise during product development), but also on **cooperation with competitors** (Stock & Stock, 2004a). Only in cooperation is it possible, in some areas of this niche market, to create marketable products in the first place. Joint venture partners, such as the FIZ Karlsruhe and the Chemical Abstracts Service (CAS), make up the STM host STN International in cooperation with the Japan Association for International Chemical Information (JAICI). FIZ Karlsruhe and CAS distribute their own respective databases via STN (apart from various third-party products), CAS with its *Chemical Abstracts* and FIZ Karlsruhe with its own smaller databases. The STN interfaces are very elaborate and address both information professionals (with STN on the Web or the client software STN Express) and professional end users (with STN Easy) at the same time. With the end user product *SciFinder*, CAS markets its *Chemical Abstracts* all over again, past STN, and thus becomes a competitor (especially of STN Easy). For the weaker partner–in this case FIZ Karlsruhe–such a combination of partner and competitor can become a serious burden.

10.4 Conclusion

Only available in the printed version.

10.5 Bibliography

Bergman, M.K. (2001). The Deep Web: Surfacing hidden value. JED–The Journal of Electronic Publishing 7(1).

Bieletzki, C., & Roth, K. (1998). Online-Hosts in Öffentlichen Bibliotheken. Neue Nutzer–neue Märkte. Köln: FH Köln. (Kölner Arbeitspapiere zur Bibliotheks- und Informationswissenschaft; 12).

Bock, A. (2000). Gütezeichen als Qualitätsaussage im digitalen Informationsmarkt, dargestellt am Beispiel elektronischer Rechtsdatenbanken. Darmstadt: STMV S. Toeche-Mittler.

Google (2009). Annual Report for the Fiscal Year Ended December 31, 2009. Google Inc. Washington, DC: United States Securities and Exchange Commission. Form 10-K.

Kremer, S. (2004). Die großen Fünf. Professionelle Online-Dienste für Juristen im Test. JurPC, Web-Dok., 205/2004.

Sherman, C., & Price, G. (2001). The Invisible Web. Medford, NJ: Information Today.

Stock, M., & Stock, W.G. (2003). Online-Hosts für Wirtschaft und News auf dem deutschen Informationsmarkt. Eine komparative Analyse. Password, N^o 7/8, 29-34.

Stock, M., & Stock, W.G. (2004a). Kooperation und Konkurrenz auf Märkten elektronischer Informationsdienste: Mit dem Wettbewerber zusammenarbeiten? Password, N^o 1, 20-25.

Stock, M., & Stock, W.G. (2004b). Kritische Erfolgsfaktoren von Anbietern elektronischer Informationsdienste. Password, N^o 4, 16-22.

Stock, M., & Stock, W.G. (2004c). Recherchieren im Internet. Renningen: Expert.

Stock, M., & Stock, W.G. (2005). Online-Hosts für Wissenschaft, Technik und Medizin auf dem deutschen Informationsmarkt. Eine komparative Analyse. Password, N^o 2, 18-23.

Stock, M., & Stock, W.G. (2006). Intellectual property information. A comparative analysis of main information providers. Journal of the American Society for Information Science and Technology, 57(13), 1794-1803.

Stock, W.G. (2007). Information Retrieval. Informationen suchen und finden. München, Wien: Oldenbourg.

Von Bischopinck, Y., & Ceyp, M. (2007). Suchmaschinen-Marketing. Konzepte, Umsetzung und Controlling. Berlin: Springer.