

Versioning information goods in digital libraries

Seminar aus Informationswirtschaft (SS 2004)



Marian Formanko

Hagenmüllergasse 33
A-1030 Vienna, AUSTRIA

E-mail: marian [at] formanko . com

Phone: +43 650 921 7752

under supervision of

Univ.-Ass. Mag. Dr. Michael Hahsler

Univ. Prof. Dr. Wolfgang Panny

Department of Information Business - Institute for Information Processing

Vienna University of Economics and Business Organization

Augasse 2-6

A-1090 Vienna, AUSTRIA

Research findings of Versioning information goods in digital libraries: Development of a versioning strategy

Keywords: versioning information goods, bundling, network effects, digital libraries, digital content, product lines

Abstract

This research paper focuses on versioning information goods in digital libraries, describes the overall functioning of digital libraries and explains why it is necessary to implement a versioning strategy. Moreover, product bundling as a special form of versioning and its consequences are discussed in the main part of the paper. It also includes a case study of the CiteSeer digital library and possible applications of the versioning strategies in these scenarios.

Table of Contents

1. Introduction.....	1
1.1 Introductory note.....	1
1.2 Basic properties of information.....	1
1.3 Information as a good in electronic commerce.....	2
2. Digital libraries today.....	4
2.1 Theoretical insight.....	4
2.2 Open Archival Information System (OAIS).....	5
2.3 Digital library in break-even.....	7
3. Versioning information.....	9
3.1 General concepts of versioning information.....	9
3.2 Methods of versioning information.....	10
3.3 Economic aspects of information versioning.....	13
3.4 Bundling information products.....	16
4. Case studies.....	20
4.1 CiteSeer.....	20
5. Conclusion.....	23
References.....	24

Table of Figures and Tables

Figure 1: OAIS Model.....	6
Figure 2: Costs/Profit development.....	7
Figure 3: Four possible Precision/Recall scenarios.....	11
Figure 4: Demand curves for low-WTP and high-WTP classes.....	13
Figure 5: Self-selection problem and consumer's surplus.....	14
Figure 6: Downgrade quality shift.....	15
Figure 7: Strategic Framework for digital services.....	20
Table 1: Main properties of the CiteSeer search engine.....	21
Table 2: Implementation of versioning strategies in CiteSeer.....	22

1. Introduction

1.1 Introductory note

It is inevitable that modern information technology has obviously a great influence on both economic and social aspects of information, its creation and delivery. Internet as a channel of information exchange has drastically reduced the costs of transportation. Examples of such information include software, news coverage, multimedia or electronic journals. Much of information goods are stored in virtual databases called digital libraries. This type of data repositories has recently undergone a very intense stage of development as many organizations and companies share the opinion, that not only information itself has some specific value, but also the underlying instruments that lead to locating and filtering necessary data may increase the value of such information[Arms98].

For these reasons it has turned out to be necessary to divide this paper into three main parts. The introductory part deals with the very elementary unit of digital libraries which is naturally a piece of information. The main characteristics of information will be described, whereas it will also be necessary to analyze how it is possible to perceive information as a trade good. In the first part, the characteristics of digital libraries in internet will be discussed and analyzed. It will give a detailed insight on how such data repositories may be build, how they work and what are their primary purposes. The second part introduces general concepts of versioning information that are relevant for digital libraries, reasons for such strategies and possible economic outcomes of these activities in framework of these libraries. The last part of the paper discusses the status quo of the CiteSeer digital library whereas possible application of versioning strategies will be dealt with.

1.2 Basic properties of information

Information as a trade good seems to have three main properties that may cause difficulties for market transactions. First of all, it has to be pointed out that information is an experience good. This means that a customer has to experience the information itself to be able to estimate its subjective value. For this reason marketers have developed different strategies such as free sample, promotion pricing or testimonials to help customers learn about new goods [Shap99]. Previewing and browsing maybe suitable choices in case of multimedia goods generated by music or film industry [Rait97]. It appears more difficult to preview textual information as it is practically impossible to reveal the whole content of document without letting potential customers exploit its value for free. In such cases publishers tend to provide customers with short abstracts, table of contents, number of citations including their sources or opinions of prestigious academic bodies[Shap99].

Secondly, a typical feature of information goods is the fact that they usually have large fixed costs of production (e.g. content generation), and small variable costs of reproduction (i.e. content delivery). This cost structure can lead to difficulties for competitive markets since the costs are not just fixed but they are also sunk [Digi04]. Sunk costs must be incurred prior to production and are not recoverable in case of failure (low demand for the good) [Shap99]. Due to this fact competitive markets tend to push price to marginal cost, which, in the case of information goods is close to zero [Varia97]. For this reason it would appear illogical to implement any cost-based pricing strategies in this context [Varia97]. Hence value based pricing i.e. estimating the information value from the viewpoint of different customer groups is a approach that has become standardized in business of information goods.

Third of all, information goods are in their nature public goods that are both nonrival and nonexcludable [Varia97]. Nonrival refers to the fact that one customer's consumption does not diminish the amount available to others, while nonexcludable means that one person cannot exclude another person from consuming the good in question [Varia97]. Due to all the above described properties it would be quite difficult to compare trading with digital goods to the market of "traditional" physical goods. On the one hand, producers and distributors of digital goods might profit from new chances and possibilities provided by the nature of digital goods, whereas on the other hand they have to deal with new threats and challenges [Wim01]. Nevertheless not only these unique characteristics of information goods make it possible for a free interplay in the network economy to take place but there have to be also suppliers of this information (content creators, or intermediaries such as digital libraries) and potential consumers to generate the demand for the information and services in question .

1.3 Information as a good in electronic commerce

The main difference between traditional firms and companies in electronic commerce is based on the subject of the trade [Deeg02]. The old economy within the world of physical goods deals with tangible goods, whereas electronic commerce is focused on digital information and service. Naturally this new trend has introduced completely new field of economy where traditional economic rules must not always apply. For instance traditional economy is more or less restricted to certain regions and therefore is region-oriented [Jone02]. As it will be discussed further in the paper, digital economy enables its participants to interact without having to consider the parameter of proximity to the respective business partner and therefore opens opportunities of becoming a global player.

As suggested by Porter, a firm in traditional economy may exercise competitive advantages if it either introduces cost leadership or differentiation of its products [Shaw00]. Producers of digital goods trying to gain advantages from a

strategy of differentiation will fail, since digital information can be easily transformed or varied. Thus if a producer offer an innovative digital product, every other producer can imitate this good, failing to gain competitive advantage by differentiation. If pursuing a cost leadership strategy, firms have to compete in prices. However in the framework of electronic commerce this may have fatal consequences, since the neglectable marginal costs of production and the winner-takes-all properties of electronic markets might lead prices down to zero according to microeconomic theory[Shaw00].

For this reason it can be concluded that competitive advantage in the network economy can neither be gained through product differentiation (in respect to other competitors), nor by pursuing a cost leadership strategy[Schack02]. However, there are other approaches such as massive product customization or quality discrimination within own product lines that could be a basis for a winning strategy.

2. Digital libraries today

2.1 Theoretical insight

Before stepping into an analysis on how businesses and organizations implement digital libraries, it is substantial to analyze the question, why there has been such striving for development of digital libraries. With the establishment of modern information infrastructure, the need for an interconnecting interface between vast data resources and end users has become a crucial to the existence of complex virtual libraries[Rait97]. The fundamental reason behind building such intelligent data repositories is the belief that they would provide better delivery of information that was possible in the past[Tehng03]. Potential customers are able to review and consult materials that are stored on computers all around the world. In other words, the question of proximity to the information source does not play any role in the information age, as it is possible to reach out virtually anywhere in the world and seek for information there. On the other hand, there has been an intense discussion on whether digital libraries can substitute ordinary libraries in whole. Many critics recognize that printed materials are so much a part of civilization that their dominant role in storing and conveying information cannot change except gradually[Tehng03].

Perhaps the major difference between printed and electronic materials is the fact that with the latter one there is a separation of data from the interface or delivery mechanism[Lee02]. Naturally, this difference introduces many factors that have to be considered in digital libraries (as they are a part of the network economy). One of the most interesting factors of digital libraries is the fact the information can be shared. Placing digital information on a network makes it available to everybody. Many digital libraries or electronic publications are maintained at a single central site, perhaps with a few duplicate copies strategically placed around the world. This is a vast improvement over expensive physical duplication of little used material, or the inconvenience of unique material that is unobtainable without traveling to the location where it is stored[Arms98]. For this reason the main advantage of digital libraries is that there is no such a term as an original issue and a copy, or being more specific all of the entities possess the identical "hierarchical" value and there is no need to distinguish between these two. This issue will be discussed further in the paper, as it introduces considerable advantages over physical libraries.

It is obvious that one of the factors that influence the value of information goods is their validity. As it has already been said printed materials are difficult to update, since the entire document must be reprinted and all copies of the old version must be tracked down and replaced. Digital updating and duplicating is a great improvement that makes it possible to provide customers with the latest information available. Digital library is online 24 hours per day and virtually accessible for any users. Such 24-hour availability is a strong facilitator of a

global presence, overcoming time differences between different regions in the world [Shaw00].

Achieving interoperability is difficult, as it requires all participants such as resource creators, and users to agree on the development of standards and formats for information interchange. Miller suggests that there are six aspects to interoperability to be considered: technical, semantic, intercommunity, political, legal and international[Mill04]. Technical interoperability is in many ways the easiest to achieve. The widespread adoption of XML may server as a good example. On the one hand, communities wish to share information more widely than ever before, given the ease of electronic information interchange, but the differences in languages can be difficult to resolve. Cross-language information access has always been a problem, as differences in terminology can create real barriers to understanding[Mill04]. While these differences man usually be resolved in the world of face-to-face communication, electronic communication requires more precision to enable information exchange. There may also be legal or political barriers to information access but this issue is beyond the scope of this paper.

2.2 Open Archival Information System (OAIS)

The OAIS is a relatively simple reference model currently under review of the ISO that contains the means of adoption of a long-term digital information library[Deeg02]. The following figure shows the six functional entities as described by the draft recommendations:

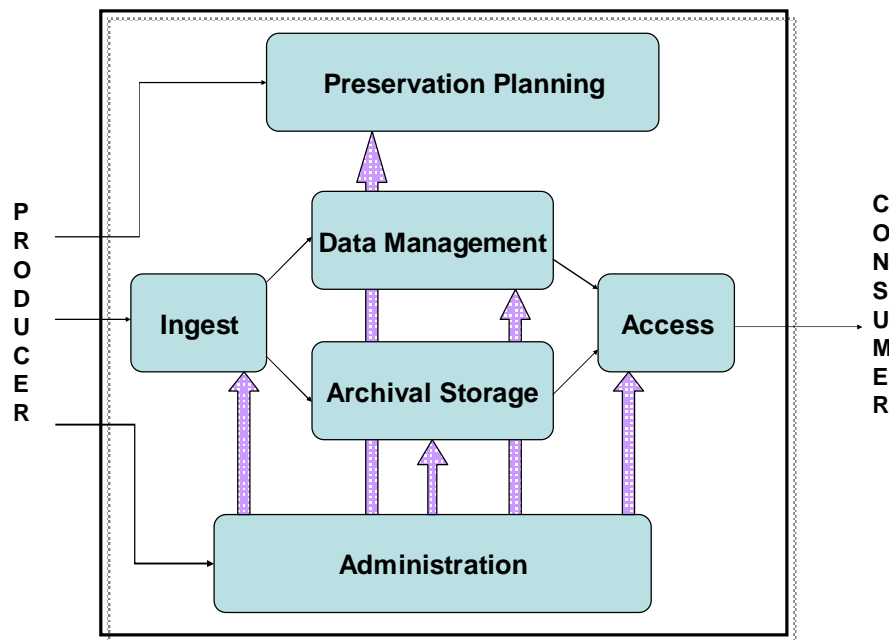


Figure 1: OAIS Model[Deeg02]

The ingest modules provides those services and functions that allow digital content to be submitted to the system for storage and management. Its purpose is to offer a means of collecting, and generating data to other processes within this framework. The archival storage module is the area concerned with the storage, maintenance and retrieval of data once it has been accepted by the ingest module. It is also in archival storage that data is gathered together in semantically similar collections to optimize end-user retrieval from the system. Data is managed from within the data management module. This entity mainly stores and retrieves metadata and administers the database functions, such as maintaining the Meta schemes. Other duties of the management module include cataloguing, access control information, authenticity and integrity control. By using these separate modules to store content, the OAIS resolves the principle of keeping content and the underlying architecture separate[Deeg02].

The overall operation of the system is managed from within the administration function. It provides essential management information through monitoring and quality control of other components. The preservation planning module evaluates the contents of the archive and provides regular reports and

recommendations for updates. Finally the Access module supports the end-used access to the content of the system. The keys here are to enable the user to identify resources, and to have all the necessary information regarding the availability of information easily to hand.

Conclusively it should be said that the OAIS model is not a specific strategy for implementing a digital library, but rather a general model stating the functions of the fundamental elements that have to be considered in order to build a digital library.

2.3 Digital library in break-even

Measuring the relationship between profits and costs remains a difficult issue and with digital services still developing and the benefits often not fully recognized, the point of breaking even seems to be indefinable[Jone02]. It would be a great advantage if one was able to predict when or whether a digital library project will make its transition from loss into the profit stage. It is necessary to point out that the two terms “loss” and “profit” do not only refer to monetary results of a digital library but include other rather embedded properties such as improved user interface, overall service enhancement and organizational prestige[Deeg02]. Hence it could be concluded that loss is the area where the investment exceeds benefits, and profit is the area where benefits are greater than the cost of maintaining the service online. In all projects and services so far viewed by the authors, the benefits have been initially slow to arrive but once achieved, they grow steeply[Deeg02]. The following chart describes a typical progress of a digital library project:

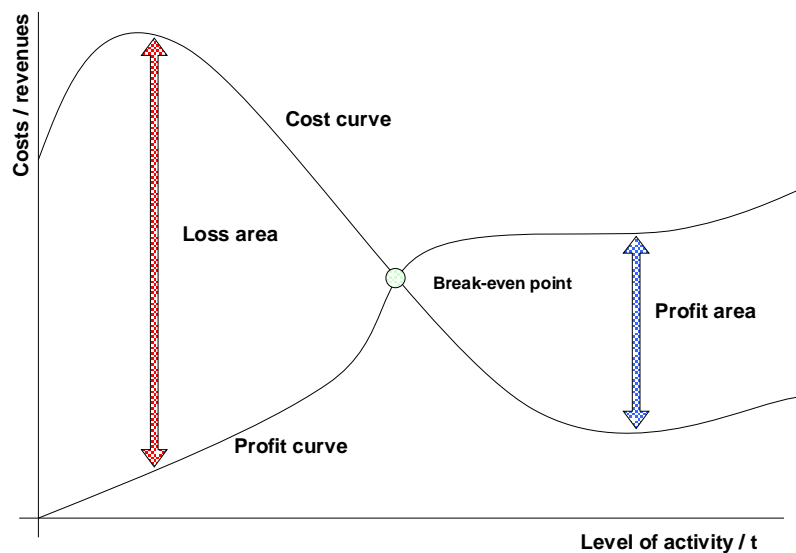


Figure 2: Costs/Profit development[Deeg02]

As it can be read out from the graph there are quite high initial costs that are accounted for by all the tasks that have to be carried out in order to develop the basic structure of a digital libraries. This also includes either programming or purchasing the whole front-end engine, customizing its features to the particular project, connection fees, etc. Costs incurred by purchasing or generating documents are not taken into consideration. As the level of activity increases (usually leading to a broader usage of the services offered) the revenues tend to rise and can be used to cover running costs. The differentials between the cost and revenue curves after reaching the break-even point might be small but they are not limited and their extension into the future has potential to repay initial investment[Jone02].

When calculating costs and profits of such a project, opportunity costs should also be taken into consideration. In the digital world today, these opportunity costs are extremely difficult to perceive in advance, and have long-term consequences[Ersh02]. For libraries these may include loss of user base for instance by not meeting the growing user desires in the information age.

3. Versioning information

3.1 General concepts of versioning information

The term information versioning or quality discrimination refers to the activity pursued by a producer of information goods (e.g. digital library) by which several versions of some specific information are offered to customers at different prices. This strategy enables customers to self-select the version that meets their expectations in the most satisfying manner whereas strategies and tools of versioning information do not have to necessarily apply only to the goods themselves, but also to the underlying business conduct entities such as the market place or search instruments. In other words potential customers sort themselves out to different groups according to their willingness to pay.

As it has been already mentioned, the key aspect of pricing information goods is to use value-based pricing, which means selling the same product at different prices to different customers, according to how may each customer or customers group is willing to pay for it. In some literature this approach is referred to as price discrimination. According to Pigou there are three types of differential pricing, which can be called, first, second and third degree discrimination[Shap99].

The first type of discrimination implies to a strategy where seller sells his/her digital products to each user at a different price. Under such circumstances the seller has to have perfect information about the potential customers so that it is possible to specifically determine how much each of them is willing to spend for a good. Different consumers have different preferences and levels of purchasing power and thus the amount they would be willing to pay for a good often exceeds a single competitive price[Ruby04]. The difference between what a consumer is willing to pay and the price actually paid is known as consumer's surplus. Hence a seller engaging in first price discrimination aims to extract the most of consumer's surplus as profits.

Second degree price discrimination often called quality discrimination stands for producing multiple versions of one product and pricing them accordingly to their quality. The fundamental question is therefore how to distinguish specific versions in order to maximize profits resulting from sales of these versions. Conclusively, third degree discrimination provides methods to divide customer segments into multiple groups by precisely defined criteria such as age or income levels and offering products to these groups at different prices. In general there are four main reasons that speak for selling rather to groups than directly to individuals. Price sensitivity within a specific group of customers tends to move more or less in the same direction among the individuals within one group[Shap99]. Students or senior citizens may be a prime example. Second of all, digital products have network externalities whose economic consequences are not fully accounted for by price or a market system. These could be either

positive network effects or their counterpart – negative, harmful effects. For instance the (informative) value of an online discussion forum increases as more people join in for the discussion and are willing to share their knowledge in respect to some specific topic with other users, expanding the amount and diversity of the information base offered[Shap99]. Such a positive externality is not reflected in the price of such a digital product (membership fee in the case of the discussion forum). Software products being market leaders in the digital economy have network externalities since their value increases as more customers decide to buy and use that specific piece of software. Once again, this is mainly due to the fact, that one specific group of the customers using the same technology widens and therefore offers greater possibilities for information and experience exchange[Kosk98e].

As suggested by Metcalfe[Bako97], if there are N people in a network, and the value of the network to each of them is proportional to the number of other users, then the total value of the network (to all the users) is proportional to

$$V(n) = n(n-1) = n^2 - n \quad (1)$$

However, the value of a network does not have to be necessarily influenced only by its size n , but other factors such as network lock-ins have to be taken into consideration. Lock-ins as they usually take place in the field of technology, represent also the third reason that speaks for group pricing. The term refers to a situation in which a customer becomes dependent on a seller of products (be it some technology, or digitized information) and cannot move to another competitor without incurring substantial switching costs[Leyd97].

3.2 Methods of versioning information

Generally speaking, versioning may not apply only to the information itself but can also be embodied in services that surround information resources. For instance, time and actuality play important role when considering specific types of information such as new coverage or stocks data. The age of information is negatively correlated with its value. In other words, the older a piece of information is the less relevance and value it has for its users. Hence time and eagerness for fresh information is the first parameter that divides customers to separate groups. Some companies such as PAWWS Financial Network are willing to incur extra costs of delaying stocks information on its way to the customers just to be able to introduce multiple subscription plans and let users self-select the one that suits their needs[Shap99].

Product versioning through access interfaces is another possibility how to extract the consumer's surplus from high-paying customers. Simple search capabilities such as keyword or full-text based searching are usually enough for the most customers. However, there may be a certain segment of users who are

willing to pay for advanced search capabilities. This could include regular expression based search capabilities or searching based on semantic similarity of multiple documents. The primary purpose of these techniques is to enable users to produce very specific search definitions so that the relevance of the retrieved information is as high as possible. The following figure shows four possible recall/precision scenarios that may arise:

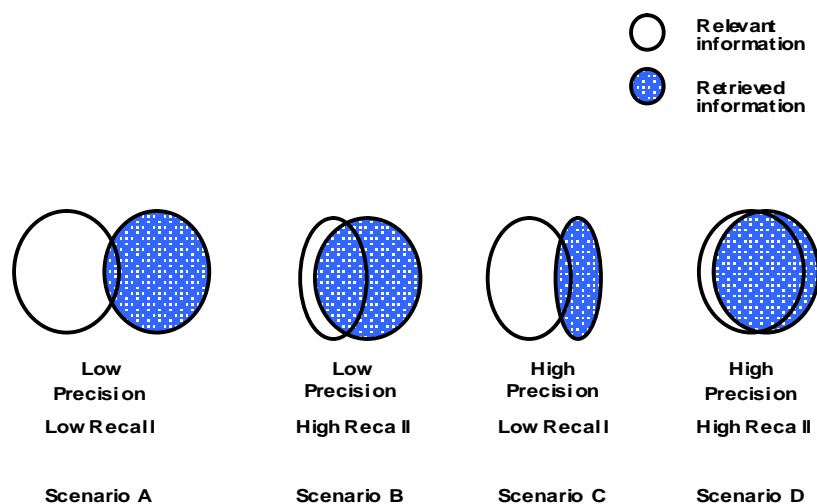


Figure 3: Four possible Precision/Recall scenarios[Goble03]

Let us examine the four possible scenarios. Search capabilities in the first two scenarios provide users only with basic search capabilities. In this case the low-recall rate of the retrieved information is a consequence of a low-precision search. Users will find difficult to locate particular document although there are present in the resource repository. Such a problem arises especially when the number of documents reaches several hundreds documents in the repository. On the other hand even low-precision searches may lead to a high-recall rate of the retrieved information. Once again, if the information base of a digital library includes several thousands documents, then it is quite possible that such a search query will deliver highly relevant results. However, the occurrence of a situation cannot be influenced by any relevant parameters and therefore it can be considered as a mere coincidence. The third example depicts a situation in which users are able to produce a high-precision search query. Nevertheless this does not automatically lead to high-recall rate of the results. Obviously, the reason is insufficient presence of search related documents in the repository. In other words, even a most precise search query cannot retrieve documents that are not stored in the library. The figure D represents a winning scenario, in which a high-precision search generates high-recall results with high relevance. Naturally, both sellers and buyers of digital information strive for systems that would exhibit such rates.

Speed of operation may be considered as another method of producing different versions of the offered information. For instance Wolfram Research sells Mathematica, a computer program that does symbolic and graphical mathematics[Shap99]. In the student version of this software the floating-point coprocessor was disabled, slowing down mathematical and graphical calculations. Even though disabling FPC incurred additional costs to itself, this version of Mathematica was sold for a cheaper price to enable access to low-income but not uninteresting segments such as students. Another well suited example could be CHIP, a German computer and technology magazine, which artificially slows down downloading of software published on its web site to non-paying customers. Managers of a digital library may decide to restrict access to certain content of the information repository to a specific group of low-paying customers in order to introduce multiple versions of their products. As Varian suggest, this strategy that is closely related to bundling of information goods, is called capability versioning and will be dealt with in a separate chapter.

Naturally it is not possible to version documents such as documents on the basis of their scientific value or quality. Removing certain pages or similar methods would probably ruin the whole concept of digital libraries[Tehng03]. However, in certain cases it is possible to degrade quality of other embedded elements, such as graphics, or audio to produce different quality versions of documents without discarding their original value. For this reason image resolution is an important method when versioning information goods containing graphical content. This method is broadly used since it is possible to carry out this versioning strategy automatically using technology that converts high-resolution graphics into medium- or even low-resolution images.

The final dimension that needs to be considered is technical support. Some customers may be willing to purchase products that include technical support for additional price, whereas others may be rather reluctant and purchase just the product itself. However this strategy is somewhat dangerous because technical support is very costly to provide[Shap99]. This can either make the projects unprofitable or the high-quality products may end up being too expensive for the most of the prospective customers.

3.3 Economic aspects of information versioning

As it has been already shown the main two issues regarding versioning information are differentiating products (i.e. creating multiple quality versions) and setting appropriate prices for each of the version. Let us now examine a simple example. Suppose that there are two groups that differ only in their willingness to pay for quality. In the first case it can be assumed that the information goods producer can perfectly distinguish between high- and low-willingness customers and identify them during purchasing. The following figure shows the demand curves for both of the groups:

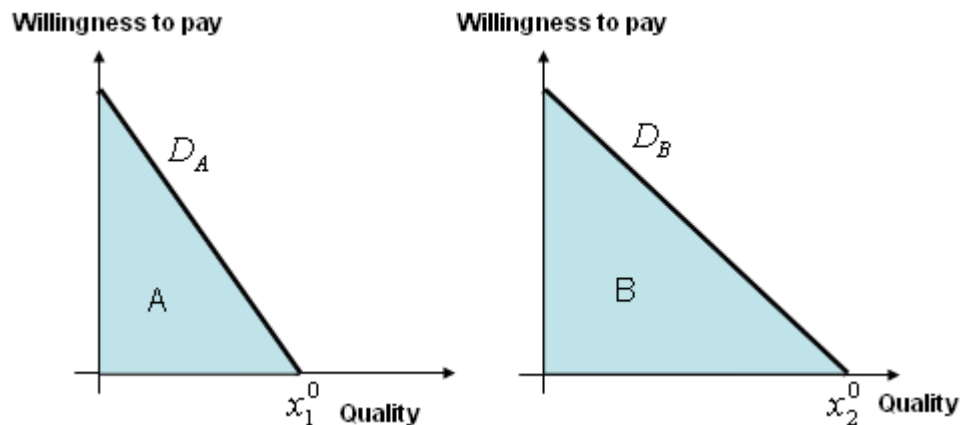


Figure 4: Demand curves for low-WTP and high-WTP classes[Varia97]

Since the producer can perfectly identify the type of the customer, he will price to good in question so as to extract the entire consumer's surplus. Naturally, if the producer is able to identify the type, and in this way extract the whole surplus, he will choose the quality product in order to maximize this surplus[Varia97]. Under this scenario the producer would set the quality for the two types to x_1^0, x_2^0 respectively and charge

$$r_1(x) = \int_0^{x_1^0} D_A (2), \quad r_2(x) = \int_0^{x_2^0} D_B (3)$$

where D are demand curves for each of the types. Conclusively it should be noted equilibrium allocation of any pricing solution with the producers being able to identify the customer's willingness to pay, is (strongly)[EconL04] Pareto efficient, which means there is no way to make the consumers better off without making the producer worse off[Varia97].

In the above mentioned example, the feasibility of price discrimination was guaranteed by ability to distinguish between the different customers segments, which obviously does not have to apply to all scenarios. Producers of information goods are frequently possess only statistical data based on

information provided during registration. This information may include age, zip code, profession or even income levels. For this reason it can be only assumed that a fraction p exhibits the high willingness to pay (high-WTP), and the complementary segment $1-p$ (the rest of the population) is of the low willingness to pay (low-WTP) type. The profit function for both segments would be

$$q_A = pr_2 \quad (4), \quad q_B = pr_1 + (1-p)r_1 = r_1 \quad (5)$$

The producer chooses whichever strategy yield larger profit [Varia97]. Once again selling to both types is Pareto efficient. On the other hand if profits resulting from selling only the high-quality product are greater than selling to both types, the producer may decide to restrict selling low-quality products, which would lead to Pareto inefficiency. The other possibility would be selling the low-quality product for marginally small penetration-price or even for free in order to widen the overall customer base. However this strategy may be exercised only for a short period of time since low-quality product would cannibalize on high-quality product which would reduce the profit made on the high-WTP type [Gabs97]. In other words new customers would very likely tend to shift from the high-quality to the low-quality segment to extract additional consumer's surplus resulting from the substantial difference between r_1 and r_2 .

Let's now examine an example in which the producer cannot base prices of his products on an exogenous observable characteristic such as membership in some social group, but can price different versions on an endogenous or intrinsic characteristic such as the quality choices based on the purchase history. Again, the intention here is to settle prices and qualities in a manner that would drive customers to a self-select process so that the seller can extract the whole customer's surplus. The following figure depicts a possible strategy for the self-selection problem:

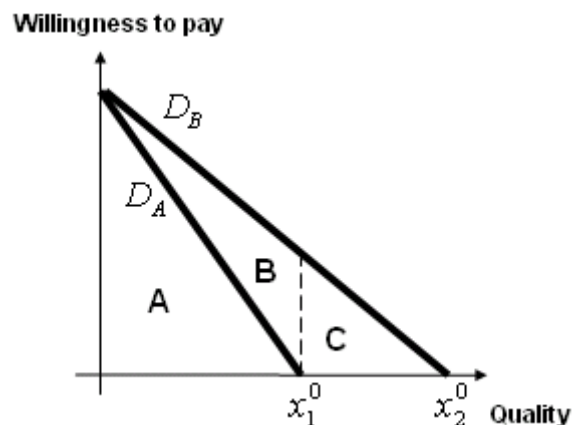


Figure 5: Self-selection problem and consumer's surplus

If the seller produces and sells two qualities (x_1^0, x_2^0) and sells them at prices $r_1^0 = A$ and $r_2^0 = B + C$ respectively. However this strategy does not satisfy the self-selection problem since high-WTP customers may decide to purchase the version for low-WTP customers and therefore achieve a positive consumer's surplus. More specifically, a positive surplus equaling the area B on the figure would be achieved:

$$CS(\text{area } B) = \int_0^{x_1^0} D_B - \int_0^{x_1^0} D_A = \int_0^{x_1^0} (D_B - D_A) \quad (6)$$

Hence, the seller's profit would equal to $r_1^0 = A$, which again is not an optimal result. In order to prevent this inefficient self-selection, the seller could set a price of $A+C$ for x_2^0 [Gabs97] so that high-WTP customer willing to pay r_2^0 for the high quality version would still receive a consumer's surplus of B but would pay a higher price ($A+C > A$). This pricing strategy induces semi optimal self-selection since it yields profits of $p(A+C) + (1-p) = A + pC$ which is definitely higher than A [Varia97].

Setting the right price was the main issue in all of the above analyzed models. The question is whether it is possible to induce shifts between the customer segments so some additional profits may be achieved. Naturally, every producer aims to induce positive shifts i.e. low-WTP customers purchasing versions originally intended for high-WTP customers. The following figure describes the quality adjustment approach:

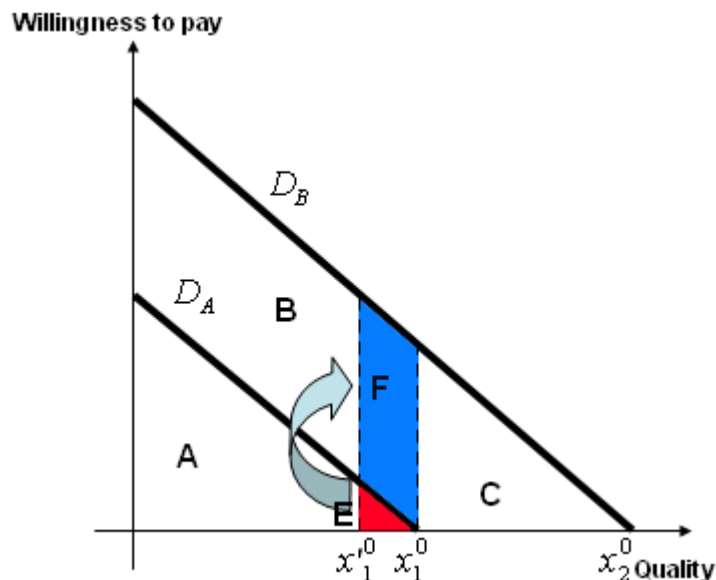


Figure 6: Downgrade quality shift

Under the circumstances that were described in the previous models, low-WTP customers purchase the quality x_1^0 for the price r_1^0 . However if the nature of the good allows even further quality downgrading of an already low-quality product then additional profits may be achieved. To be more specific, we assume a downgrade quality shift:

$$x_1^0 \rightarrow x_1'^0, \text{ so that } x_1^0 > x_1'^0$$

The red area **E** represents the loss of customers (=profits) that were willing to purchase (x_1^0, r_1^0) but will not buy $(x_1'^0, r_1^0)$, because the lowered quality renders the product unusable in respect to customers' needs. Another reason could be simply the fact that the price-utility ratio fell under the minimum expectations of this particular segment of customers. However the loss **E** in the area **A** leads to a considerable gain in the upper blue part labeled as **F**. This gain represents the increase of profit caused by the above mentioned quality shift. Thus it can be concluded that this shift produces a profit amounting to:

$$q_{A-E} + q_{C+F} = \int_0^{x_1^0} D_A + \int_{x_1'^0}^{x_2^0} D_B - 2 \int_{x_1'^0}^{x_1^0} D_A > q_A + q_C \quad (7)$$

Therefore it can be concluded that by decreasing the quality of the low-quality bundle, the seller achieves gain in the high-quality bundle. The seller will continue to reduce the quality of the low-quality bundle until the marginal reduction in revenues from the low-WTP consumers just equals to the marginal increase in revenues of the high-WTP consumers[Varia97].

3.4 Bundling information products

Bundling is a special form of versioning in which two or more distinct are offered as a package at single price[Kosk98]. This strategy allows the seller to extract maximum consumer surplus. The main advantage behind this idea is that the seller is not required to identify the different types of customers and price discriminates accordingly. Instead the seller offers bundled products that provide him the greatest utility. It is obvious that bundling reduces product diversity and therefore should be used in markets where customers' preferences are heterogeneous.

Let's suppose a theoretical example of a digital library that offers access to database of journal articles divided into three different sections. Each section includes a certain number of articles regarding some specific topic. There are

three different subscription plans available. The first one gives access only to one of the sections selected by the customers. The second plan grants access to two sections and naturally the third open enables to browse system wide through the three journal sections. This approach is often called as quantity bundling. For simplicity purposes it is assumed that the digital library is the only supplier of those articles and therefore acts as a monopolist firm on the market. Following parameters will be analyzed[Kosk98]:

- p_i is the purchase price for a bundle including i information goods(number of accessible sections)
- d_i partial demand for a bundle containing i information goods(number of accessible sections)
- u_i quality measurement parameter of a bundle whereas it can be supposed that there are no identical articles within one section. A bundle containing more information (access to more sections) has a higher quality (i.e. $u_1 < u_2 < u_3$)
- k refers to the type of customers. As it has already been pointed out in the previous chapter, consumer preferences are assumed to be heterogeneous i.e. there is not a single pair of customers that would assign the same value to a particular bundle. For simplicity reasons consumer types are uniformly distributed within the range $[0,1]$

Further constraints will be taken into consideration:

$$\forall d_i \in D; d_i < d_{i+1} \text{ if } p_{i+1} \leq p_i \quad (8)$$

- (A) where D is aggregated demand for all bundles. This constraint states that any of the customers prefers more quality for the same price (i.e. subscription plan with access to multiple sections will be favored over the one with a single section if and only if price for these two plan is equal.

$$CS_{k,i} = ku_i - p_i \quad (9)$$

- (B) Consumer's surplus CS for a customer k that purchases a bundle containing access to i sections. A higher type consumer is willing to pay more for a bundle of a given quality.

In this case, consumer surplus may be referred to as the amount that a consumer with quality expectations q benefits by purchasing a bundle i for the price p_i that is less than they would be willing to pay. Under these model circumstances and constraints consumers will be willing to purchase a bundle if the surplus they obtain is greater than zero. There will also be consumers who will be indifferent between buying and not buying due to extrinsic reasons which do not have to be explored in this model. This consumer type can be found by solving:

$$CS_{k,0} = CS_{k,1} \quad (10)$$

As it has already been said, the consumer surplus obtained by not buying is zero and therefore:

$$ku_1 - p_1 = 0 \Rightarrow k_{0,1} = \frac{p_1}{u_2} \quad (11)$$

where $k_{0,1}$ is the consumer type being indifferent between buying and not buying a bundle containing access to only one section. Analogically, it is possible to find a consumer type who is indifferent between buying subscription plan with one section and two sections and can be found by solving:

$$CS_{k,1} = CS_{k,2} \Rightarrow ku_1 - p_1 = ku_2 - p_2 \Rightarrow k_{1,2} = \frac{p_2 - p_1}{u_2 - u_1} \quad (12)$$

Rewriting this equation parametrically gives:

$$k_{i,i+1} = \frac{p_{i+1} - p_i}{u_{i+1} - u_i} \quad (13)$$

which leads to the proof of the two fundamental bundling and versioning prerequisites. There must be a positive difference in price between a higher class bundle and a lower class bundle (i.e. $\Delta p > 0$) and also between the respective quality values (i.e. $\Delta u > 0$). Moreover it is possible to determine partial demand functions for each of the three bundles if one supposes for possible market segments: $S_0: [0, k_{0,1}]$, $S_1: [k_{0,1}, k_{1,2}]$, $S_2: [k_{1,2}, k_{2,3}]$ and $S_3: [k_{2,3}, 1]$.

It has been already shown that segment S_0 is not profitable. The indexes (1, 2, and 3) of the successive segments correspond with the number of sections purchased by respective consumer types. Thus partial demands for can be defined as:

$$d_i = (k_{i,i+1} - k_{i-1,i})N \quad (14)$$

The demand can be calculated by multiplying the customers type in a segment i (i.e. purchase tendency to a bundle i) by N which is mass of potential consumers (for unity it can be assumed that $N=1$). The total profit resulting from the purchases of each bundle type can be logically defined as:

$$\Pi = \sum_{i=1}^3 p_i d_i = p_1 d_1 + p_2 d_2 + p_3 d_3 \quad (15)$$

Substitution in (15) using (14) leads to:

$$\Pi = [p_1(k_{12} - k_{01}) + p_2(k_{23} - k_{12}) + p_3(1 - k_{23})]N \quad (16)$$

and using (13) leads to the final profit function:

$$\Pi = \arg \max [p_1 \left(\frac{p_2 - p_1}{u_2 - u_1} - \frac{p_1}{u_1} \right) + p_2 \left(\frac{p_3 - p_2}{u_3 - u_1} - \frac{p_2 - p_1}{u_2 - u_1} \right) + p_3 \left(1 - \frac{p_3 - p_2}{u_3 - u_2} \right)]N \quad (17)$$

As it can be seen the profit in this scenario is a non-linear function with three independent types of variables: the price of the bundles, their respective quality and naturally the number of customers[Kosk98].

4. Case studies

4.1 CiteSeer

CiteSeer(CS) was chosen for the application of versioning strategies, as it is one of the most popular digital libraries in the web, storing a large number of scientific papers ranging from biology to computer science and most importantly the access to this library has been free of charge ever since. CS is a specialized type of digital library since there is actually no centralized data repository that could be managed and administered by the CiteSeer themselves. Let us first start with examining the CiteSeer search engine, so it is possible to distinguish between the main properties that are crucial and modules which in a way could serve as premium services:

Property	Description
Autonomous Citation Indexing (ACI)	CiteSeer uses ACI to autonomously create a citation index that can be used for literature search and evaluation. Compared to traditional citation indices, ACI provides improvements in cost, availability, comprehensiveness, efficiency, and timeliness.
All cited documents	CiteSeer computes citation statistics and related documents for all articles cited in the database, not just the indexed articles.
Reference linking	As with many online publishers, CiteSeer allows browsing the database using citation links.
Citation context	CiteSeer can show the context of citations to a given paper, allowing a researcher to quickly and easily see what other researchers have to say about an article of interest.
Awareness and tracking	CiteSeer provides automatic notification of new citations to given papers, and new papers matching a user profile.
Related documents	CiteSeer locates related documents using citation and word based measures and displays an active and continuously updated bibliography for each document.
Similar documents	CiteSeer shows the percentage of matching sentences between documents.
Full-text indexing	CiteSeer indexes the full-text of the entire articles and citations. Full boolean, phrase and proximity search is supported.
Query-sensitive summaries	CiteSeer provides the context of how query terms are used in articles instead of a generic summary, improving the efficiency of search.
Citation graph analysis	CiteSeer analyzes the graph of citations, e.g. to provide hubs and authorities ranking (ala Kleinberg).
Powerful search	e.g. CiteSeer allows using author initials to narrow a citation search.
Autonomous location of articles	CiteSeer uses search engines and crawling to efficiently locate papers on the Web.

Table 1: Main properties of the CiteSeer search engine[CiteS]

As it can be seen, the CiteSeer offers many functionalities that enable users to carry out very precise searching. Services that are necessary for at least basic functioning of the project are marked grey, whereas the blue ones could be introduced only to paying customers. Unfortunately the nature of these premium elements would not allow a pay-per-use model, so it can be assumed that access to a complete package of the functions could be based upon a flat rate price. In scientific literature CS is often addressed as an “assistant agent” and serves three basic purposes:

1. It automates the repetitive and slow process of finding and retrieving Web based publications.
2. Once potentially relevant papers are retrieved, it guides users towards interesting papers by making them searchable.
3. When a relevant paper is found, it helps the user by suggesting other related papers using similarity measures derived from semantic features of the retrieved documents. [Bollac]

It is necessary to point out that the third feature cannot be found in all digital libraries and therefore could be introduced only to paying customers as a some kind of premium services. Similarity measuring was introduced by Porter[Bollac] and tries to calculate the word weight as follows:

$$w_{ds} = \frac{(0.5 + 0.5 \frac{f_{ds}}{f_{d \max}}) (\log \frac{N_D}{n_s})}{\sqrt{\sum_{j \in d} ((0.5 + 0.5 \frac{f_{dj}}{f_{d \max}})^2 (\log \frac{N_D}{n_j})^2)} \quad (18)$$

where the frequency of each word stem (e.g. “walk”, “walking”, “walked”) s , is f_s , the number of documents having stem s is n_s , and the highest term frequency is $f_{D \max}$.

The following table shows some possible implementations of the versioning techniques that were discussed throughout the paper.

Strategy	Description	Pros	Cons
Capability	<ul style="list-style-type: none"> Disabling searching based on similarity measuring for non-paying customers Disabling reference linking Disabling relevant documents 	<ul style="list-style-type: none"> clear quality distinction observable by all users. Likely to be a reason for switching from free to subscription plan since CS customers usually seek for highly scientific papers. Decreasing server work load 	<ul style="list-style-type: none"> Some users may switch to other digital libraries that provide these advanced search features for free.

Speed of operation	<ul style="list-style-type: none"> • CiteSeer is often overloaded, therefore limiting the number of concurrent connections from non-paying customers to decrease work load. 	<ul style="list-style-type: none"> • paying customers enjoy better reliability and general reaction time of the service 	<ul style="list-style-type: none"> • only temporary advantage, since the number of subscriptions will likely rise
File types - Resolution	<ul style="list-style-type: none"> • CiteSeer automatically converts all documents to several formats so that users may choose for themselves which version they prefer. • Disabling automatic conversions or offering documents only as TXT files without graphics 	<ul style="list-style-type: none"> • Very strong versioning principle. TXT files include much of the information, but without graphics, they may not be usable for all purposes. • Decreasing server work load. 	<ul style="list-style-type: none"> • Information of some documents can be fully destructed • Possible misunderstanding of retrieved documents
Technical support	<ul style="list-style-type: none"> • Some of the search tools are quite difficult to use. • There could be an online tutorial explaining how to get the most of the features provided. 	<ul style="list-style-type: none"> • Newcomers may find this feature very helpful. • Relevant information will be located faster. 	<ul style="list-style-type: none"> • Online tutorial: no cons • Email support can be very time consuming and costly
Delay	<ul style="list-style-type: none"> • CiteSeer agent crawl the web to find the latest scientific papers published. The papers are processed and then stored in the database. • Latest papers would not appear in the search results of non-paying customers. 	<ul style="list-style-type: none"> • clear quality distinction observable by all users. • Paying customers know what they spend their money for (comparing search results when logged-in vs. not-logged) 	<ul style="list-style-type: none"> • Some extra costs incurred by additional programming
Annoyance	<ul style="list-style-type: none"> • Currently, there is no online advertising on CS • Non-paying customers would have to click them through some advertisements related to their search queries to receive documents 	<ul style="list-style-type: none"> • Extra profit from advertisers 	<ul style="list-style-type: none"> • May become too annoying, users may switch to some other services •

Table 2: Implementation of versioning strategies in CiteSeer

5. Conclusion

The aim of this paper was to show that there are significant differences between physical and information goods. We have witnessed a strong development of digital libraries lately, not only in the technological and systematical field, but there have also been attempts to turn digital libraries into profitable businesses. For that reason it was necessary to define the nature of information goods from the view point of economics and econometrics. We have discussed the necessity of specialized strategies such as versioning and pricing that have to be taken into consideration when commencing business in this field, because they represent the only method that is able to locate, analyze and most importantly satisfy demand for such products.

In general, massive customization of products is the obviously the only way to attract all potential segments of customers to the offered products. Information goods can be easily unbundled to their elementary units and rebundled according to very specific customer's needs and preferences with almost no additional costs[Schack02]. Therefore bundling as a special method of versioning has been a subject to an intense discussion by many authors over the last few years. As suggested by Gabszewics[Gabs97], there should be more research carried out on different models, such as non-linear pricing, or flat vs. pay-per-view strategies, because whether we accept it or not, the price is still the most important factor even in the world of zeros and ones.

References

- [Arms98] *Arms, Y. W.*: Digital Libraries. The MIT Press, Cambridge 2000.
- [Bako97] *Bakos, Y.; Brynjolfsson, E.*: Aggregation and Disaggregation of Information Goods: Implications for Bundling, Site Licensing and Micropayment Systems. <http://www.gsm.uci.edu/~bakos/aig/aig.html>. 1997-06. Accessed on 2004-04-26.
- [Bollac] *Bollacker, D. K.; Lawrence, S.; Giles, Lee C.*: CiteSeer: An Autonomous Web Agent for Automatic Retrieval and Identification of Interesting Publications. <http://www.neci.nj.nec.com/homepages/lawrence/papers/cs-aa98/latex.html>. Accessed on 2004-05-29.
- [Carp98] *Carpenter, L.; Shaw, S.*: Towards the Digital Library. The British Library, London 1998.
- [CiteS] About CiteSeer. <http://citeseer.ist.psu.edu/citeseer.html>. Accessed on 2004-05-29.
- [Craw99] *Crawford, W.*: Being Analog: Creating Tomorrow's Libraries. American Library Association, Chicago 1999.
- [Deeg02] *Deegan, M.; Tanner S.*: Digital Futures: Strategies for the Information Age. Library Association Publishing, London 2002.
- [Digi04] Digital Economy 1. <http://www.tml.hut.fi/Opinnot/Tik-111.050/2003/kalvot/Digecon6s03f.pdf>. Accessed on 2004-05-14.
- [EconL04] Glossary: Strongly Pareto Efficient. http://www.econlinks.com/glossary/strongly_pareto_efficient.php. Accessed on 2004-05-28.
- [Ersh02] *Ershova, T.; Hohlov Y.*: Libraries in the Information Society. KG Saur, Munich 2002.
- [Flaat92] *Flaatten, P.; McCubbrey, D.; Burgess, K.*: Foundations of Business Systems. The Dryden Press, Fort Worth 1992.
- [Gabs97] *Gabszewics, J. J.*: Subscription as a Price Discrimination Device. Center for Operations Research and Econometrics, Louvain 1997.
- [Goble03] *Goble, C.; Harper, S.*: Information Retrieval, Hypermedia, and the Web. Accessed 2004-05-15.
- [Harri02] *Harris, L. E.*: Licensing Digital Content: A Practical Guide for Librarians. American Library Association, Chicago 2002.
- [Haru00] *Harum, S.*: Successes and failures of digital libraries: Library Applications of Data Processing. Graduate School of Library and Information Science, Champaign 2000.
- [Jones02] *Jones, M.*: Preservation Management of Digital Materials. British Library, London 2001.
- [Koch03] *Koch, T.*: Research and Advanced Technology for Digital Libraries. Springer, Berlin 2003.
- [Kosk98] *Koski, H.*: Economic Analysis of the Adoption of Technologies with Network Externalities. Oulu University Press, Oulu 1998.
- [Lee02] *Lee, D. S.*: Building an Electronic Resource Collection. Library Association Publishing, London 2002.

- [Leyd97] *Leydesdorff, L.*: Competing Technologies: Lock-ins and Lock-outs. <http://users.fmg.uva.nl/lleydesdorff/casys97/1997liege.pdf>. Accessed on 2004-05-11.
- [Mill04] *Miller, P.*: Interoperability: what is it and why should I want it?. <http://www.ariadne.ac.uk/issue24/interoperability/intro.html>. Accessed on 2004-05-14.
- [Rait97] *Raitt, D.*: Libraries for the New Millennium: Implications for managers. Library Association Publishing, London 1997.
- [Ruby04] *Ruby, A. D.*: Price Discrimination. http://www.digitaleconomist.com/pd_4010.html Accessed on 2004-04-14.
- [Schack02] *Schackman, J.; Link, H.*: Intermediaries for Provision of Mass Customized Digital Goods in Electronic Commerce. 2002-04. Accessed 2004-05-25.
- [Shap99] *Shapiro, C.; Varian R. H.*: Information Rules: A Strategic Guide to the Network Economy. Harvard Business School Press, Boston 1999.
- [Shaw00] *Shaw, M.; Blanning, R.; Strader, T.*: Handbook on Electronic Commerce. Springer, Austin 2000.
- [Tehng03] *Tengku, S.*: Digital Libraries: Technology and Management of Indigenous Knowledge for Global Access. Springer, Berlin 2003.
- [Varia97] *Varian, R. H.*: Versioning information goods. www.sims.berkeley.edu/~hal/Papers/version.pdf, 1997-03-17. Accessed on 2004-05-22.
- [Wile00] *Wiley, J.*: Electronic Commerce: Strategies and Models for Business-to-Business Trading. John Wiley & Sons Ltd., Chichester 2000.
- [Wils00] *Wilson, M.*: Understanding the Needs of Tomorrow's Library User: Rethinking Library Services for the New Age. Australasian public Libraries and Information Services, 2000.
- [Winn98] *Winn, J.*: Open System, Free Markets and the Regulation of Internet Commerce, Tulane Review, Tulane 1998.